

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Statistical Foundations and Algorithms for Modern AI: Contrastive Learning, Scaling Laws, and Unlearning

Permalink

<https://escholarship.org/uc/item/3td8f6z1>

ISBN

9798189276026

Author

Lin, Licong

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

Statistical Foundations and Algorithms for Modern AI: Contrastive Learning, Scaling Laws,
and Unlearning

by

Licong Lin

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor Song Mei, Co-chair

Professor Peter L. Bartlett, Co-chair

Distinguished Professor Bin Yu

Assistant Professor Nikita Zhivotovskiy

Spring 2026

Statistical Foundations and Algorithms for Modern AI: Contrastive Learning, Scaling Laws,
and Unlearning

Copyright 2026
by
Licong Lin

Abstract

Statistical Foundations and Algorithms for Modern AI: Contrastive Learning, Scaling Laws,
and Unlearning

by

Licong Lin

Doctor of Philosophy in Statistics

University of California, Berkeley

Assistant Professor Song Mei, Co-chair

Professor Peter L. Bartlett, Co-chair

Generative artificial intelligence has emerged as one of the most transformative developments in science and technology. Modern AI models—such as large language models (LLMs), diffusion models, and multimodal systems—have achieved remarkable success across language, vision, and cross-modal tasks. Despite this empirical progress, rigorous theoretical foundations for these models remain limited, posing challenges for understanding their capabilities and guiding the development of more powerful and trustworthy AI.

This dissertation contributes to the statistical foundations and algorithms for modern AI from three complementary angles.

First, motivated by the concept of sufficient statistics, we develop a statistical theory of contrastive pre-training and multimodal generative AI. We introduce *approximate sufficient statistics* as a measure of representation quality and show that near-minimizers of the contrastive loss are approximately sufficient, making them adaptable to diverse downstream tasks such as zero-shot classification, conditional diffusion models, and vision-language models. We further propose the Joint Generative Hierarchical Model (JGHM) for the joint image-text distribution and prove that Transformers can efficiently approximate belief propagation on JGHMs, yielding end-to-end sample complexity guarantees for multimodal contrastive learning.

Second, we study the theory of neural scaling laws in a high-dimensional linear regression setting. Under a power-law spectral assumption on the data covariance, we derive matching upper and lower bounds on the test error of one-pass SGD applied to sketched linear models. Our analysis shows that the variance error, which classical theory predicts to grow linearly with model size, is of smaller order due to the implicit regularization of SGD, thus partially

explaining the inconsistency between standard learning theory and empirically observed scaling laws.

Third, we address the problem of LLM unlearning: removing specific knowledge from a pretrained model without retraining from scratch. We propose the *Negative Preference Optimization* (NPO) loss, inspired by preference optimization. NPO mitigates the *catastrophic collapse* that results from gradient ascent by adaptively downweighting already-unlearned samples. We provide theoretical guarantees showing that NPO diverges logarithmically slower than gradient ascent, and demonstrate strong empirical performance on the TOFU unlearning benchmark.

To my parents.

Contents

Contents	ii
List of Figures	iv
List of Tables	viii
1 Introduction	1
1.1 Bibliographic Notes	3
2 A Statistical Theory of Contrastive Pre-training and Multimodal Generative AI	4
2.1 Introduction	4
2.2 Related literature	6
2.3 Statistical properties of contrastive pre-training	7
2.4 Sample-efficient learning in hierarchical models	15
2.5 Experiments	23
2.6 Conclusion	25
3 Scaling Laws in Linear Regression: Compute, Parameters, and Data	26
3.1 Introduction	26
3.2 Related work	29
3.3 Setup	30
3.4 Scaling laws	32
3.5 Experiments	36
3.6 Risk bounds under a general spectrum	36
3.7 Conclusion	38
4 Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning	40
4.1 Introduction	40
4.2 Preliminaries on Machine Unlearning	43
4.3 Negative Preference Optimization	45
4.4 Synthetic Experiments	48

4.5	Experiments on the TOFU Data	49
4.6	Conclusion	55
	Bibliography	56
A	Appendix for Chapter 2	71
A.1	Background on CLIP, ZSC, CDM, and VLM	71
A.2	Further related literature	73
A.3	Results for vision-language models	76
A.4	Proofs in Section 2.3	79
A.5	Proof of Theorem 2.10	106
A.6	Proof of Theorem 2.12	130
A.7	Proof of Theorem A.2	144
A.8	Auxiliary lemmas	168
A.9	Experimental details	184
B	Appendix for Chapter 3	194
B.1	Preliminary	194
B.2	Proofs in Section 3.4	200
B.3	Approximation error	201
B.4	Bias error	209
B.5	Variance error	218
B.6	Expected risk of the average of (SGD) iterates	219
B.7	Concentration lemmas	227
C	Appendix for Chapter 4	238
C.1	Proofs	238
C.2	The role of KL Divergence on the Forget Set	247
C.3	Ablation experiments on the role of retain loss	248
C.4	Experimental details of the synthetic experiments	249
C.5	Experiments on the TOFU dataset	250

List of Figures

2.1	Left: the JGHM used to generate the joint distribution of text and images. Right: an illustrative example of a generated text-image pair.	17
2.2	Risks (solid curves) and excess risks (dashed curves) as a function of the parameter p_{flip} for CLIP training, ZSC, CDM, and VLM. The training setups for the different curves are described in Section 2.5. Across all setups, the excess risks of Guided TF approach zero. The risks of Standard TF are close to the Bayes risk in ZSC, CDM, and VLM, demonstrating that CLIP representations can effectively adapt to downstream tasks.	24
3.1	Test error of one-pass SGD vs. effective sample size and model size, for power-law degrees $a = 1.5$ and $a = 2$	28
3.2	Test error minus irreducible risk vs. effective sample size and model size on a log-log scale.	36
4.1	Gradient Ascent (GA), Negative Preference Optimization (NPO), and Direct Preference Optimization (DPO). NPO can be interpreted as DPO without positive samples. The gradient of NPO is an adaptive weighting of that of GA, and the weight vanishes for unlearned samples.	41
4.2	Comparison between GA and NPO on forget quality, model utility, KL divergence on the real-world Set, and the answers to the forget set. The rightmost figure shows the answers generated from variants of GA and NPO that incorporates the RT loss. All figures are generated on the Forget05 task in the TOFU data, trained for 10 epochs (detailed setup in Appendix C.5.1).	45
4.3	Retain distance versus forget distance for GA and NPO with varying levels of β in the binary classification experiment with $\alpha = 1$. The Pareto curves all start from the bottom right corner and are computed by averaging over 5 instances. We observe that the NPO trajectory converges to the GA trajectory as $\beta \rightarrow 0$. More details can be found in Section 4.4.	47
4.4	Forget distance and retain distance versus optimization steps for $\alpha = 1$ (a1, a2, a3) and $\alpha = 0$ (b). Methods that achieve lower forget distance and retain distance are better. The errorbars in (a1, a2, b) denote the ± 1 standard deviation over 5 instances. The Pareto curves in (a3) all start from (1.70, 0.02), and are averaged over 5 instances.	50

4.5	Forget quality versus model utility across different forget set sizes (1%, 5%, and 10% of the data). Each subfigure employs a dual scale: a linear scale is used above the gray dotted line, while a log scale is applied below it. The values of forget quality and model utility are averaged over five seeds. Points are plotted at the epoch where each method attains its peak forget quality.	51
4.6	Evolution of forget quality and model utility across different forget set sizes (1% (left), 5% (middle), and 10% (right) of the data). Each line is averaged over 5 seeds. Each figure in the top row employs a dual scale as in Figure 4.5. In Forget01, we evaluate the performance of the unlearned model in every gradient step, while in Forget05 and Forget10, we evaluate it in every epoch.	52
4.7	Sampled response to questions in three subsets of TOFU. Yellow : questions; Green : true answer or desired answers; Red : undesired answers.	53
4.8	The evolution of the Forget KL during the unlearning process on the Forget10 task in TOFU data. Note that the KL term in GA+KL is the divergence on the retain set, not on the forget set. More experimental details are included in Appendix C.5.1.	54
4.9	Evolution of forget quality and model utility on Forget50 and Forget90 for NPO+RT with a proper componential weights between loss terms.	54
A.1	Illustration of CLIP and zero-shot classification. CLIP trains a text encoder and an image encoder by maximizing the similarity of paired image-caption representations and minimizing the similarity of non-paired representations. After pre-training, zero-shot classification predicts the label whose representation has the highest similarity with the image representation. This figure reproduces Figure 1 from Radford et al. (2021).	73
A.2	Illustration of VLMs and CDMs. VLMs use neural networks to approximate the conditional distribution of each next token, given prior tokens and image embeddings. CDMs employ neural networks to approximate the conditional expectation of a clear image, given a noisy input image and text embeddings.	74
A.3	Risks of zero-shot learning versus number of samples. The setup of Figure A.3 is almost the same as that of Figure 2.2b, except that we fix $p_{\text{flip}} = 0.2$ and vary M , where M ranges from 5 to 250. We can observe that larger numbers of samples lead to improved predictions.	190
A.4	Fitted curves for the risks in zero-shot learning. Each risk curve is fitted by the function $f(M) = A + B/M$, with parameters A and B . The fitted results are shown as dark dashed lines, while the empirical risks are plotted as solid lines. All fitted curves achieve an R^2 value above 0.98.	191

A.5	Zero-shot classification performance of a ResNet-50 model pretrained on CC12M, evaluated on the ImageNet-1k validation set. We use the standard 80 text templates from OpenAI’s CLIP paper (Radford et al., 2021). For each run, we sample a random permutation of these 80 templates and progressively increase the number of templates in that order. This procedure is repeated 16 times, and we report the mean and standard deviation of cross-entropy loss, top-1 accuracy, and top-5 accuracy. We also fit the results with functions of the form $f(M) = A + B/M$, where A and B are parameters. Fitted curves are shown as orange dashed lines, while empirical results are shown as blue solid lines. All fitted curves achieve an R^2 value above 0.98.	191
A.6	Out-of-distribution (OOD) risks (solid curves) and excess risks (dashed curves) as a function of the parameter p_{flip} for CLIP training, ZSC, CDM, and VLM. Models are trained with a fixed $p_{\text{flip}} = 0.2$. Across all setups, Guided TF closely matches the performance of Mis-spec. BP. In the CDM (A.6c) and VLM (A.6d) setups, Standard TF performs similarly to Guided TF , whereas in the CLIP training (A.6a) and ZSC (A.6b) setups, Standard TF shows a greater gap from Guided TF . This suggests that standard-trained transformers may perform closer to the belief-propagation algorithm when the in-distribution risk is smaller.	192
A.7	OOD risks and excessive risks of various transformer architectures and belief propagation for VLMs and CDMs. These figures exhibit same trends as Figures A.6c and A.6d. (a) fix the text $p_{\text{flip}} = 0.2$ but vary the image p_{flip} . (b) fix the image $p_{\text{flip}} = 0.2$ but vary the text p_{flip}	193
B.1	The expected risk (Risk) of the average of iterates of (SGD) versus the sample size N and the model size M for different power-law degrees a . The expected risk is computed by averaging over 1000 independent samples of $(\mathbf{w}^*, \mathbf{S})$. We fit the expected risk using the formula $\text{Risk} \sim \sigma^2 + c_1/M^{a_1} + c_2/N^{a_2}$ via minimizing the Huber loss as in Hoffmann et al. (2022). Parameters: $\sigma = 1, \gamma = 0.1$. Left: For $a = 1.5, d = 20000$, the fitted exponents are $(a_1, a_2) = (0.59, 0.33) \approx (0.5, 0.33)$. Right: For $a = 2, d = 2000$, the fitted exponents are $(a_1, a_2) = (1.09, 0.49) \approx (1.0, 0.5)$. Note that the values of (a_1, a_2) are close to our theoretical predictions $(a - 1, 1 - 1/a)$ in both cases.	222
B.2	The expected risk of the average of iterates of (SGD) minus the irreducible risk versus the effective sample size and model size. Parameters $\sigma = 1, \gamma = 0.1$. (a), (b): $a = 1.5, d = 10000$; (c), (d): $a = 2, d = 1000$. The errorbars denotes the ± 1 standard deviation of estimating the expected risk using 100 independent samples of $(\mathbf{w}^*, \mathbf{S})$. We use linear functions to fit the expected risk under log-log scale and report the slope of the fitted lines (denoted by k).	222
C.1	Forget KL versus optimization steps for all methods in the synthetic experiment. (a): $\alpha = 1$, (b): $\alpha = 0$. The errorbars denote ± 1 standard deviation over 5 runs.	247

C.2	The evolution of the forget quality and model utility when we adjust the weights of the NPO term and the retain loss term in NPO+RT. The experiments are performed on Forget10.	248
C.3	Statistics for NPO-based methods and baselines on the Forget01 task of TOFU.	254
C.4	Statistics for NPO-based methods and baselines on the Forget05 task of TOFU.	255
C.5	Statistics for NPO-based methods and baselines on the Forget10 task of TOFU.	256
C.6	Statistics for NPO-based methods and baselines on the Forget20 task of TOFU.	257
C.7	Statistics for NPO-based methods and baselines on the Forget30 task of TOFU.	258
C.8	Statistics for NPO-based methods and baselines on the Forget50 task of TOFU.	259

List of Tables

A.1	Learning rates and penalties for different models.	187
A.2	Top-1 and top-5 validation accuracies (%) across different choices of τ . The ResNet-50 model was pretrained on the CC3M dataset (3M samples) for 32 epochs under various τ configurations. For the trainable τ case, τ was initialized at 0.07 and constrained to be no smaller than 0.01 during training, following the design of (Radford et al., 2021). Each setting was evaluated over 3 random seeds, and results are reported as mean $\pm$ standard deviation.	189
C.1	Values of learning rate and β for different methods when $\alpha = 1$ and $\alpha = 0$ in the synthetic experiments.	249
C.2	The weights for different components in GA-based losses and IDK+RT loss. . . .	252

Acknowledgments

I would like to express my deepest gratitude to my advisors, Prof. Song Mei and Prof. Peter L. Bartlett, for their guidance, support, and mentorship throughout my Ph.D. Their insight and perspective have profoundly shaped how I think about research, from formulating meaningful questions to developing principled and rigorous approaches. I have greatly benefited from their guidance in developing both technical depth and a broader vision, as well as from their ability to use theory to better understand and guide practical problems. Beyond research, they have been incredibly generous with their time, always willing to discuss ideas, provide feedback, and offer encouragement during challenging moments. Their mentorship has played a central role in my growth as an independent researcher, and I am deeply grateful for the opportunity to learn from them.

I would also like to thank Prof. Cheng Zhang, Prof. Edgar Dobriban, and Prof. Martin J. Wainwright for their guidance at the early stages of my academic journey. Through their mentorship, I learned how to approach research problems, write papers clearly and effectively, and present ideas in a coherent and compelling way. Their advice and feedback have had a lasting influence on how I communicate and think about research, and I am deeply grateful for their support. I would also like to express my gratitude to Prof. Bin Yu for introducing me to the PCS framework and for emphasizing the importance of developing theory relevant to practice. In addition, I would like to thank Prof. Nikita Zhivotovskiy for serving on my committee and for providing valuable feedback on my research.

I would also like to thank my collaborators and mentors Koulik Khamaru, Wenlong Mou, Tijana Zrnic, Yu Bai, Jingfeng Wu, and Jason D. Lee, for many insightful discussions and rewarding collaborations. I have benefited immensely from their ideas, feedback, and support, and I am grateful for the opportunity to learn and collaborate with them. I would also like to thank my many other collaborators, without whom my work would not have been possible.

Beyond research, many of the memorable moments during my Ph.D. have come from the time spent with friends. I am especially grateful to my friends Chengzhong Ye, Yaxuan Huang, Andy Shen, Seunghoon Paik, Karissa Huang, Tiffany Ding, Austin Zane, Omer Ronen, Dan Soriano, Fangzhou Su, Drew Nguyen, Tianyu Guo, Ruiqi Zhang, Sizhu Lu, Zhexiao Lin, Arisa Sadeghpour, Xuelin Yang, Niclas Dern, and many others in the department, for the countless chats and discussions, games of foosball, trips, wind-downs, and time spent playing basketball and pickleball together. These moments brought a great deal of joy and balance to my Ph.D. journey. I am also deeply grateful to my friends Qiuyu Ren, Baihe Huang, Hengyu Fu, Yuhang Cai, Yuzheng Hu, Shange Tang, and Zijie Zhuang, for their support and for sharing so much happiness with me over the years. I would also like to thank the staff, especially La Shana Polaris, for their patience in answering my numerous administrative questions. Lastly, I would like to thank my parents for their constant support and encouragement. I truly appreciate everything they have done for me.

Chapter 1

Introduction

Modern AI models such as large language models (LLMs), diffusion models, and multimodal systems have demonstrated striking empirical success across a wide range of tasks in language, vision, and cross-modal understanding. Despite this progress, rigorous theoretical foundations for these models are still limited. This makes it challenging to understand their capabilities and limitations, and to guide the development of more powerful and trustworthy AI models.

In this dissertation, we study the *statistical foundations and algorithms* for modern AI. Drawing on tools from statistical learning theory, high-dimensional statistics, and optimization, we study the statistical foundations of architectures, algorithms, and phenomena in modern AI, and develop mathematically motivated algorithms for LLM safety and alignment.

The three main contributions of this dissertation are summarized as follows.

Contrastive learning and multimodal generative AI (Chapter 2). Multimodal generative AI models, which process and generate content across modalities such as language and vision, have become a central focus of modern AI research. These models often rely on contrastive pre-training (Radford et al., 2021) to learn joint representations across modalities, which are then adapted to downstream tasks such as zero-shot classification, vision-language modeling, information retrieval, etc.

Motivated by the idea of *sufficient statistics* from classical statistics, we introduce the concept of *approximate sufficient statistics* to characterize the quality of representations learned by contrastive pre-training. We show that this measure can be bounded by the excess risk of contrastive learning, and can be used to control the generalization error of downstream tasks such as zero-shot classification and conditional diffusion generation. In particular, near-minimizers of the contrastive (CLIP) loss (Radford et al., 2021) are approximately sufficient statistics for the joint image-text distribution, making them adaptable to diverse downstream tasks.

We further propose the **Joint Generative Hierarchical Model** (JGHM), a tree-structured graphical model to mimic the joint distribution of image and text. We prove that Transformers (Vaswani, 2017) can efficiently approximate belief propagation on the JGHM, leading to sample complexity guarantees for multimodal learning based on contrastive-pretrained representations. These results provide end-to-end statistical guarantees for mul-

timodal generative AI, offering insights on why contrastive pretraining can learn useful representations for multimodal tasks.

Theory of scaling laws (Chapter 3). Empirically, LLMs often satisfy a neural scaling law: the test error of the pretrained model decreases polynomially as model and data size increase (Kaplan et al., 2020; Hoffmann et al., 2022). However, this appears inconsistent with classical statistical theory, which decomposes test error into approximation, bias, and variance errors, with the variance error expected to grow linearly with the model size.

We provide a theoretical explanation for this contradiction in a high-dimensional linear regression setting. We consider a model with M parameters trained by one-pass SGD on N sketched data under Gaussian design, with a power-law spectral assumption on the data covariance. We derive finite-sample upper and lower bounds on the test error; when the data and model parameters have aligned power-law spectra, the bounds match and the test error exhibits the same form as the empirical scaling law. The key insight is that the variance error is of smaller order due to the *implicit regularization* of one-pass SGD, thereby giving rise to scaling laws.

LLM unlearning via negative preference optimization (Chapter 4). LLMs often memorize sensitive, private, or copyrighted information during pre-training (Carlini et al., 2021), raising significant safety and legal concerns. *LLM unlearning* aims to remove the influence of specific training data from a pretrained model while maintaining its overall utility. Traditional approaches to unlearning often require retraining the model from scratch without the undesirable data, yet it is prohibitively expensive for large models.

We propose *Negative Preference Optimization* (NPO), a new loss function for LLM unlearning inspired by preference optimization (Rafailov et al., 2024). NPO can be viewed as an adaptively weighted gradient ascent method that downweights already-unlearned data points, keeping the loss bounded and avoiding catastrophic collapse. Theoretically, we prove that in logistic regression, NPO iterates diverge logarithmically slower than gradient ascent iterates, suggesting that NPO is more stable and less prone to catastrophic collapse. Empirically, on pretrained LLaMA-2-7B models, NPO significantly outperforms GA-based methods and achieves state-of-the-art results on the TOFU dataset (Maini et al., 2024) for LLM unlearning. Since its introduction, NPO has become a standard baseline in LLM unlearning and has inspired the design of many finetuning-based unlearning algorithms.

A guiding principle throughout this thesis is that theory is only useful when it captures practical phenomena while remaining mathematically tractable. We recognize that not every theoretical model bears on practice; that said, we deliberately choose theoretical setups whose assumptions are closely connected to modern AI practice, and focus on extracting concrete lessons that may be useful in practice. The Joint Generative Hierarchical Model in Chapter 2 is motivated by the hierarchical compositional structure of natural images and text; the power-law spectral assumption in Chapter 3 is aligned with the empirically observed frequency of words in text, and the one-pass training reflects how modern LLMs are trained over massive datasets; and the unlearning analysis in Chapter 4 is motivated by the catastrophic collapse

phenomenon observed when applying gradient ascent. Each chapter aims to provide concrete takeaways: that contrastive pretraining provably produces representations that transfer broadly to downstream tasks, supporting the prevailing pretrain-then-adapt paradigm; that the implicit regularization of one-pass SGD partially explains why overparameterized LLMs trained for a single epoch do not overfit and exhibit neural scaling laws; and that one can design more stable unlearning algorithms by adaptively downweighting already-unlearned samples as in NPO (Chapter 4). Overall, we hope these results offer not only mathematical understanding but also insights for practice.

1.1 Bibliographic Notes

- Chapter 2 is based on joint work with Kazusato Oko, Yuhang Cai, and Song Mei (Oko et al., 2025), which is under major revision at the *Journal of the American Statistical Association* (JASA).
- Chapter 3 is based on joint work with Jingfeng Wu, Sham Kakade, Peter Bartlett, and Jason Lee (Lin et al., 2024b), published at the *Conference on Neural Information Processing Systems* (NeurIPS) 2024.
- Chapter 4 is based on joint work with Ruiqi Zhang, Yu Bai, and Song Mei (Zhang et al., 2024), published at the *Conference on Language Modeling* (COLM) 2024.

Chapter 2

A Statistical Theory of Contrastive Pre-training and Multimodal Generative AI

2.1 Introduction

Multi-modal generative AI systems, such as DALL-E (OpenAI, 2022) for generating images from text prompts and GPT-4V (OpenAI, 2023) for generating text based on both image and text inputs, have achieved remarkable empirical success. The training process for such systems often begins with contrastive pre-training (Radford et al., 2021; Jia et al., 2021), which learns lower-dimensional neural network representations for each modality using large-scale pretraining datasets. Subsequently, the contrastively pre-trained representations of one modality are fixed and used to guide the training of a generative model for the other modality.

To elaborate, we focus on multi-modal learning in the image-text domain¹, where the contrastive pre-training process is known as Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021). Given a dataset of paired image-text samples $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \in \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}}$, CLIP trains a pair of neural network encoders, $(\mathbf{E}_{\text{im}} : \mathcal{X}_{\text{im}} \rightarrow \mathbb{R}^p, \mathbf{E}_{\text{tx}} : \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}^p)$, by aligning paired image-texts while simultaneously pushing apart non-paired ones. This alignment is achieved by minimizing the contrastive loss defined in Eq. (2.1). The pre-trained CLIP encoders have shown exceptional performance in various downstream tasks, including:

- *Zero-shot classification* (Radford et al., 2021; Jia et al., 2021). The goal is to predict the label $y \in \mathcal{Y}$ for a new image $\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}$. Using the pre-trained encoders $(\mathbf{E}_{\text{im}}, \mathbf{E}_{\text{tx}})$, a good classifier can be constructed without the need for fine-tuning on task-specific

¹We use “image-text domain” as convenient terminology, but the theory applies universally and is not restricted to this setting. Our analysis focuses on two domains for simplicity but extends to multi-domain scenarios. Notably, in the machine learning literature, “multi-modal learning” does not necessarily require three or more domains.

data.

- *Conditional diffusion models* (OpenAI, 2022; Esser et al., 2024): The task is to generate an image $\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}$ from a text prompt $\mathbf{x}_{\text{tx}} \in \mathcal{X}_{\text{tx}}$. In these models, the text embedding $\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})$ is used in the conditional denoising function, without directly referencing the original text prompt during training.
- *Vision-language models* (Li et al., 2022; Liu et al., 2024a). The task is to generate text $\mathbf{x}_{\text{tx}} \in \mathcal{X}_{\text{tx}}$ from an image prompt $\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}$. In such models, the image embedding $\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})$ is used in the auto-regressive transformer, without directly referencing the original image prompt during training.

The empirical success of multimodal learning underscores the need for a theoretical framework to better understand this paradigm, ideally within the context of statistical learning theory. To achieve this, two key theoretical questions need to be addressed:

- 1) *Why are CLIP encoders effective representations for downstream tasks?* The statistical properties of contrastive loss minimizers have been extensively studied in the literature (Saunshi et al., 2019; Tosh et al., 2021a,b; HaoChen et al., 2021). Existing works often leverage the structure of the contrastive loss and its connection to downstream tasks to show that linear functions of learned representations perform well in these settings. However, such analyses fall short in explaining tasks like zero-shot classification, where no fine-tuning is required, as well as tasks involving conditional diffusion models and vision-language models, where linear functions of learned representations are insufficient to capture relevant functions.
- 2) *Why do the encoders and downstream functions admit efficient neural network approximations?* This question has received relatively less attention. While neural networks are universal function approximators (Barron, 1993), they can suffer from the curse of dimensionality (Bach, 2017) when dealing with general high-dimensional target functions. The primary theoretical challenge lies in constructing a tractable yet realistic statistical model for the joint image-text distribution. A Gaussian assumption, though mathematically convenient, is often overly restrictive and unrealistic, whereas a fully non-parametric approach could lead to the curse of dimensionality.

This chapter addresses the two theoretical questions outlined above. In Section 2.3, we reveal a surprisingly simple property of the near-minimizers of the CLIP loss: they are pairs of **approximate sufficient statistics**, a generalization of the classical concept of sufficient statistics. Due to their approximate sufficiency and the straightforward implications of data processing inequalities, these representations can adapt to a variety of downstream tasks, including zero-shot classification, conditional diffusion models, and vision-language models. Furthermore, when a simple “canonical representation” of the data exists, we show that it can be recovered from any near-minimizer of the CLIP loss through a simple two-layer network. This enables CLIP representations to effectively adapt to downstream tasks where the canonical representations serve as sufficient statistics.

In Section 2.4, we apply our general framework to a statistical model for the joint distribution of images and text, which we call the **Joint Generative Hierarchical Model** (JGHM). The JGHM is a graphical model consisting of two trees with a shared root, where the root node captures high-level features, and the leaf nodes represent observed images or text. We demonstrate that transformers (Vaswani, 2017) can efficiently approximate the relevant functions within JGHMs by approximating the belief propagation algorithm, thus **breaking the curse of dimensionality**. Building on this insight, we derive end-to-end sample complexity results for tasks such as zero-shot classification, conditional diffusion models, and vision-language models, all utilizing the pre-trained CLIP representations.

Numerical simulations are presented in Section 2.5 within the simulated JGHM framework. The experimental results demonstrate that transformers trained using the Adam algorithm (Kingma, 2014) can achieve near-optimal minimizers, exhibiting strong generalization performance. Additionally, out-of-distribution tests show that the minimizers obtained by Adam closely emulate the behavior of belief propagation, a result of independent interest.

2.2 Related literature

Contrastive learning and multi-modal learning. CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) leverage large-scale contrastive pretraining to extract visual and textual embeddings, relying on loss functions such as NCE (Gutmann and Hyvärinen, 2010), InfoNCE (Oord et al., 2018), and Multi-class N-pair loss (Sohn, 2016) to distinguish paired from non-paired samples. Conditional Diffusion Models, exemplified by DALL-E (OpenAI, 2022) and Stable Diffusion (Esser et al., 2024), generate realistic images from text prompts, while Vision-Language Models like Flamingo (Alayrac et al., 2022), BLIP (Li et al., 2022), and Llava (Liu et al., 2024b,a) interpret and describe images based on textual inputs. These frameworks highlight the versatility of contrastive learning in advancing multimodal understanding and generation.

Theories of Contrastive Learning and CLIP. Numerous studies have shown that InfoNCE loss (derived from the InfoMax principle (Linsker, 1988)) maximizes a lower bound on mutual information between positive sample pairs (Oord et al., 2018; Poole et al., 2019; Hjelm et al., 2018; Bachman et al., 2019; Tian et al., 2020a; Zimmermann et al., 2021; Lu et al., 2024), which aligns with Lemma 2.1 and Proposition 2.3. Theoretical analysis of the adaptation properties of contrastive learning has been investigated in a series of work (Wang and Isola, 2020; Saunshi et al., 2019; Tosh et al., 2021a,b; HaoChen et al., 2021). Our work diverges from these existing theories of contrastive learning in three key ways: (1) While many studies provide “absolute risk bounds” for downstream tasks under structural conditions, our work offers “excess risk bounds,” which require more refined statistical analysis; (2) We analyze the multimodal learning, including zero-shot prediction task, conditional diffusion models, and vision-language models, which have not been addressed in these work; and (3) We proposed a data distribution for image and text pairs and provided end-to-end statistical efficiency guarantees for multimodal learning through neural networks.

Closest to our approach are [Uesaka et al. \(2024\)](#) and [Chen et al. \(2023\)](#). The former uses point-wise mutual information to bound excess risk in downstream classification, while the latter examines CLIP’s minimizer under completeness conditions, demonstrating its strong zero-shot classification capabilities. In contrast, our work (1) adopts a sufficient statistics framework to interpret CLIP, (2) uncovers additional properties of CLIP representations, and (3) provides a unified theory for multimodal learning, including vision-language models and conditional diffusion frameworks.

Approximate sufficient statistics. The concept of approximate sufficient statistics was mentioned in [Chen et al. \(2021\)](#), which proposed an approach to find them. However, this work did not provide a formal definition of approximate sufficient statistics or explore its theoretical properties. The relationship between contrastive loss minimizers and sufficient statistics was examined in [Xu and Zheng \(2024\)](#), but the notion of approximate sufficient statistics was not considered. After an extensive review of the literature, we conclude that the definition of approximate sufficient statistics and its connection to the approximate minimizer of CLIP loss, to the best of the authors’ knowledge, is novel.

Further related works are summarized in Appendix [A.2](#).

2.3 Statistical properties of contrastive pre-training

In this section, we demonstrate that CLIP provides effective representations that can adapt to downstream tasks. In Section [2.3.1](#), we show that any near-minimizer of the CLIP risk yields a pair of near-sufficient statistics. In Section [2.3.2](#), we demonstrate that this near-sufficiency facilitates the adaptability of CLIP representations to various downstream tasks. Furthermore, in Section [2.3.3](#), we show that if the joint distribution allows for a canonical representation with certain well-posedness properties, a simple adapter (a small network) enables efficient neural network approximations for downstream tasks where the canonical representations serve as sufficient statistics.

2.3.1 Near-sufficiency of CLIP minimizers

To simplify the discussion and avoid measure-theoretic complications, we assume that both $\mathcal{X}_{\text{im}}$ and $\mathcal{X}_{\text{tx}}$ are discrete spaces. Let the image-text pair $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \in \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}}$ follow a joint distribution $\mathbb{P}_{\text{im,tx}} \in \mathcal{P}(\mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}})$. We denote the marginal distributions of $\mathbf{x}_{\text{im}}$ and $\mathbf{x}_{\text{tx}}$ as $\mathbb{P}_{\text{im}}$ and $\mathbb{P}_{\text{tx}}$, respectively, and the conditional distributions of $\mathbf{x}_{\text{im}}$ given $\mathbf{x}_{\text{tx}}$ and $\mathbf{x}_{\text{tx}}$ given $\mathbf{x}_{\text{im}}$ as $\mathbb{P}_{\text{im}|\text{tx}}$ and $\mathbb{P}_{\text{tx}|\text{im}}$, respectively. For clarity, we will omit subscripts in probability expressions when the context is clear.

In the CLIP framework, paired image-texts are used as positive samples, while unpaired image-texts are used as negative samples. Specifically, within each batch we have K i.i.d. samples $(\mathbf{x}_{\text{im},i}, \mathbf{x}_{\text{tx},i})_{i=1}^K$ from $\mathbb{P}_{\text{im,tx}}$, and we use $(\mathbf{x}_{\text{im},i}, \mathbf{x}_{\text{tx},i})_{i=1}^K$ as the paired image-text samples and $(\mathbf{x}_{\text{im},i}, \mathbf{x}_{\text{tx},j})_{i \neq j; i,j=1}^K$ as the non-paired samples.

Let $E_{\text{im}} : \mathcal{X}_i \rightarrow \mathbb{R}^p$ and $E_{\text{tx}} : \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}^p$ represent the image and text encoders, respectively, both parameterized by neural networks. Using a user-defined similarity link function $\Upsilon : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$, the similarity score for an image-text couple $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$ is given by $S_{E_{\text{im}}, E_{\text{tx}}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) := \Upsilon(E_{\text{im}}(\mathbf{x}_{\text{im}}), E_{\text{tx}}(\mathbf{x}_{\text{tx}}))$. The CLIP risk function is the expected InfoNCE loss over paired and non-paired samples:

$$\begin{aligned} \bar{R}_{\text{clip}, K}(S) &:= \mathbb{E} \left[-\log \frac{\exp(S(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},1}))}{\sum_{j \in [K]} \exp(S(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},j}))} \right] + \mathbb{E} \left[-\log \frac{\exp(S(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},1}))}{\sum_{j \in [K]} \exp(S(\mathbf{x}_{\text{im},j}, \mathbf{x}_{\text{tx},1}))} \right], \\ R_{\text{clip}, K}(E_{\text{im}}, E_{\text{tx}}) &:= \bar{R}_{\text{clip}, K}(S_{E_{\text{im}}, E_{\text{tx}}}), \end{aligned} \quad (2.1)$$

where the expectation is over $(\mathbf{x}_{\text{im},i}, \mathbf{x}_{\text{tx},i})_{i=1}^K \sim_{\text{iid}} \mathbb{P}_{\text{im}, \text{tx}}$. This risk comprises the cross-entropy losses for classifying paired and non-paired samples, based on $\text{softmax}((\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},j})_{j \in [K]})$ and $\text{softmax}((\mathbf{x}_{\text{im},j}, \mathbf{x}_{\text{tx},1})_{j \in [K]})$, respectively. The function $\bar{R}_{\text{clip}, K}$ is defined over all possible similarity scores $S : \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}$, while $R_{\text{clip}, K}$ is defined over all couples of encoders $(E_{\text{im}} : \mathcal{X}_{\text{im}} \rightarrow \mathbb{R}^p, E_{\text{tx}} : \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}^p)$.

Global minimizers of CLIP as sufficient statistics. The InfoNCE loss, first introduced by Oord et al. (2018), underpins the CLIP framework and leads to the following characterization of its global minimizers. For completeness, the proof is provided in Appendix A.4.1.

Lemma 2.1 (Global CLIP minimizer (Oord et al., 2018)). *Consider minimizing $\bar{R}_{\text{clip}, K}$ over all possible similarity scores $S : \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}$. For all $K \geq 3$, the set of global minimizers of $\bar{R}_{\text{clip}, K}$, denoted by $\mathcal{M}_S$, is given by*

$$\mathcal{M}_S = \left\{ S_{\star} : S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \log \left[\frac{\mathbb{P}_{\text{im}, \text{tx}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})}{\mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}}) \cdot \mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx}})} \right] + \text{const}, \text{ for some const} \in \mathbb{R} \right\}. \quad (2.2)$$

Moreover, in the limit as $K \rightarrow \infty$, the minimum CLIP risk yields the negative mutual information of $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$ under the joint distribution $\mathbb{P}_{\text{im}, \text{tx}}$:

$$\lim_{K \rightarrow \infty} \left[-\frac{1}{2} \inf_S \bar{R}_{\text{clip}, K}(S) + \log K \right] = \text{MI}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) := \mathbb{E}_{\mathbb{P}_{\text{im}, \text{tx}}} \left[\log \left[\mathbb{P}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) / [\mathbb{P}(\mathbf{x}_{\text{im}}) \cdot \mathbb{P}(\mathbf{x}_{\text{tx}})] \right] \right].$$

As a corollary, using the Fisher-Neyman factorization theorem (Fisher, 1922; Neyman, 1936), any pair of encoders that achieve the minimum CLIP risk serves as sufficient statistics.

Corollary 2.2 (CLIP minimizers as sufficient statistics). *Suppose there exists a pair of encoders $(E_{\text{im}, \star}, E_{\text{tx}, \star})$ such that $R_{\text{clip}, K}(E_{\text{im}, \star}, E_{\text{tx}, \star}) = \inf_S \bar{R}_{\text{clip}, K}(S)$. Then, $E_{\text{im}, \star}(\mathbf{x}_{\text{im}})$ and $E_{\text{tx}, \star}(\mathbf{x}_{\text{tx}})$ are sufficient statistics for the statistical models $\mathbb{P}_{\text{im}|\text{tx}}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}})$ and $\mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}})$, respectively. Specifically, the mutual information satisfies:*

$$\text{MI}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \text{MI}(E_{\text{im}, \star}(\mathbf{x}_{\text{im}}), \mathbf{x}_{\text{tx}}) = \text{MI}(\mathbf{x}_{\text{im}}, E_{\text{tx}, \star}(\mathbf{x}_{\text{tx}})).$$

Proof of Corollary 2.2. By Lemma 2.1 and the condition that $R_{\text{clip}, K}(E_{\text{im}, \star}, E_{\text{tx}, \star}) = \inf_S \bar{R}_{\text{clip}, K}(S)$, the conditional distribution can be expressed as:

$$\mathbb{P}_{\text{im}|\text{tx}}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}}) = \exp\{-\text{const}\} \cdot \mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}}) \cdot \exp\{\Upsilon(E_{\text{im}, \star}(\mathbf{x}_{\text{im}}), E_{\text{tx}, \star}(\mathbf{x}_{\text{tx}}))\}.$$

By the Fisher-Neyman factorization theorem (see e.g., Theorem 3.6 in [Keener 2010](#)), $\mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}})$ is a sufficient statistic for the model $\mathbb{P}_{\text{im}|\text{tx}}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}})$. Similarly, by symmetry, $\mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}})$ is a sufficient statistic for the model $\mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}})$. $\square$

Lemma 2.1 has appeared in various forms across the literature ([Oord et al., 2018](#); [Poole et al., 2019](#); [Hjelm et al., 2018](#); [Bachman et al., 2019](#); [Tian et al., 2020a](#); [Zimmermann et al., 2021](#)). Similarly, the interpretation of CLIP minimizers as sufficient statistics in Corollary 2.2 aligns with the InfoMax principle introduced in earlier works ([Linsker, 1988](#); [Chen et al., 2021](#)). While we do not claim originality for either result, to the best of our knowledge, the use of the Fisher-Neyman factorization theorem to establish Corollary 2.2 is novel and may be of particular interest to the statistics community.

Corollary 2.2 implies that $\mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}})$ captures all the information necessary to predict $\mathbf{x}_{\text{tx}}$, making it an effective representation of the image (similar argument applies to $\mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}})$). It’s worth noting that there may be infinitely many minimizers of $R_{\text{clip},K}$, and Corollary 2.2 holds for all of them. In practice, finding the exact minimizer of the CLIP risk is not feasible; however, an approximate version of the results still holds, which we discuss next.

Near-minimizers of CLIP as near-sufficient statistics. We now demonstrate that approximate minimizers of the CLIP risk serve as approximate sufficient statistics. To this end, we extend the classical notion of sufficiency to encoders and similarity scores, formalizing the concept as follows:

Definition 2.1 (Approximate sufficiency). For an image encoder $\mathbf{E}_{\text{im}} : \mathcal{X}_{\text{im}} \rightarrow \mathbb{R}^m$, its sufficiency is measured as

$$\text{Suff}(\mathbf{E}_{\text{im}}) = \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\mathbb{P}_{\text{tx}|\text{im}}(\cdot|\mathbf{x}_{\text{im}}) \middle| \middle| \mathbb{P}_{\text{tx}|\text{im}}(\cdot|\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right) \right], \quad (2.3)$$

where $\mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx}}|\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))$ denotes the conditional distribution of $\mathbf{x}_{\text{tx}}$ given $\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})$ under $\mathbb{P}_{\text{im},\text{tx}}$. The sufficiency measure for a text encoder $\mathbf{E}_{\text{tx}} : \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}^m$ is defined symmetrically.

A similarity score $\mathbf{S} : \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}$ induces a probability distribution $\hat{\mathbb{P}}_{\mathbf{S}}$ over $\mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}}$:

$$\hat{\mathbb{P}}_{\mathbf{S}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) := \frac{\exp(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) \mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}}) \mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx}})}{\sum_{\mathbf{x}_{\text{im}}', \mathbf{x}_{\text{tx}}'} \exp(\mathbf{S}(\mathbf{x}_{\text{im}}', \mathbf{x}_{\text{tx}}')) \mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}}') \mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx}}')}. \quad (2.4)$$

Its sufficiency measure, $\text{Suff}(\mathbf{S})$, is defined as

$$\text{Suff}(\mathbf{S}) = \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\mathbb{P}_{\text{tx}|\text{im}}(\cdot|\mathbf{x}_{\text{im}}) \middle| \middle| \hat{\mathbb{P}}_{\mathbf{S}}(\cdot|\mathbf{x}_{\text{im}}) \right) \right] + \mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \left[D_{\text{KL}} \left(\mathbb{P}_{\text{im}|\text{tx}}(\cdot|\mathbf{x}_{\text{tx}}) \middle| \middle| \hat{\mathbb{P}}_{\mathbf{S}}(\cdot|\mathbf{x}_{\text{tx}}) \right) \right], \quad (2.5)$$

where $\hat{\mathbb{P}}_{\mathbf{S}}(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}})$ and $\hat{\mathbb{P}}_{\mathbf{S}}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}})$ are the conditional distributions induced by $\hat{\mathbb{P}}_{\mathbf{S}}$. By this definition, it follows that $\text{Suff}(\mathbf{E}_{\text{im}}) + \text{Suff}(\mathbf{E}_{\text{tx}}) \leq \text{Suff}(\Upsilon(\mathbf{E}_{\text{im}}(\cdot), \mathbf{E}_{\text{tx}}(\cdot)))$.

We say that $\star \in \{\mathbf{E}_{\text{im}}, \mathbf{E}_{\text{tx}}, \mathbf{S}\}$ is ε -sufficient if $\text{Suff}(\star) \leq \varepsilon$. Statistics with small sufficiency measures are called “approximate sufficient statistics” or “near-sufficient statistics”.

Approximate sufficiency has a more intuitive form via the information loss:

$$\text{Suff}(\mathbf{E}_{\text{im}}) = \text{MI}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \text{MI}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{x}_{\text{tx}}),$$

with the proof provided in Appendix A.4.9. This implies that when $\text{Suff}(\mathbf{E}_{\text{im}}) = 0$, we have $\text{MI}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \text{MI}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{x}_{\text{tx}})$, aligning 0-sufficiency with the classical notion of sufficiency. Although the concept has been mentioned in the literature (Chen et al., 2021), we are unaware of any formal or rigorous definition of approximate sufficient statistics in prior work. In Section 2.3.2, we illustrate that near-sufficient encoders achieve strong performance on downstream tasks, including zero-shot classification and conditional diffusion models

We introduce an assumption on the boundedness of the score function, which allows us to show that near-minimizers of the CLIP risk function serve as near-sufficient statistics.

Assumption 2.1 (Bounded score). *Let $\mathcal{S}$ denote the set of score functions over which the minimization is performed. There exists a constant $c_1 > 0$ such that for all pairs $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$, we have $\frac{\mathbb{P}_{\text{im}, \text{tx}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})}{\mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}}) \cdot \mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx}})} \in [1/c_1, c_1]$ and $\exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) \in [1/c_1, c_1]$ for all $S \in \mathcal{S}$.*

We refer to Appendix A.9.2 for more discussions on Assumption 2.1. Building on this assumption, we establish the following result.

Proposition 2.3 (Near-minimizer of CLIP as near-sufficient statistics). *Assume Assumption 2.1 holds, and let $\bar{R}_{\text{clip}, K}$ denote the CLIP risk as defined in Eq. (2.1). Suppose $S_\star$ is a global minimizer of $\bar{R}_{\text{clip}, K}(S)$ as defined in Eq. (2.2). Then, there exists a constant $C > 0$, which depends polynomially on c_1 , such that for any $S \in \mathcal{S}$, its sufficiency can be bounded in terms of its CLIP excess risk. Specifically, for any $K \geq 3$, we have:*

$$\lim_{K' \rightarrow \infty} \left[\bar{R}_{\text{clip}, K'}(S) - \bar{R}_{\text{clip}, K'}(S_\star) \right] = \text{Suff}(S) \leq \underbrace{\left[\bar{R}_{\text{clip}, K}(S) - \bar{R}_{\text{clip}, K}(S_\star) \right]}_{\text{CLIP excess risk}} \cdot \left(1 + \frac{C}{K} \right). \quad (2.6)$$

The proof of Proposition 2.3 is provided in Appendix A.4.2. The first equality in Eq. (2.6) follows established results in prior literature, such as (Wang and Isola, 2020; Zimmermann et al., 2021). The primary contribution of Proposition 2.3 lies in the non-asymptotic sufficiency bound in Eq. (2.6), which improves upon prior results, e.g., (Wang and Isola (2020), Theorem 1). Compared to Wang and Isola (2020), the bound presented here offers two significant improvements: (1) the error decays at a faster rate of K^{-1} rather than $K^{-1/2}$, and (2) the error is multiplicative rather than additive. The multiplicative error bound ensures that, for any finite K , the exact minimizer of the CLIP risk is 0-sufficient, whereas an additive error bound does not provide this guarantee. On the other hand, one can establish an additive error bound without requiring the boundedness conditions in Assumption 2.1. We refer the readers to Appendix A.4.3 for more details.

2.3.2 Adaptation to various downstream tasks

Consider the couple of encoders $(E_{\text{im}} : \mathcal{X}_{\text{im}} \rightarrow \mathbb{R}^p, E_{\text{tx}} : \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}^p)$ and a link function $\Upsilon : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$, such that $S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) := \Upsilon(E_{\text{im}}(\mathbf{x}_{\text{im}}), E_{\text{tx}}(\mathbf{x}_{\text{tx}}))$ is a near-minimizer of the CLIP risk. By Proposition 2.3, E_{im} and E_{tx} are near-sufficient statistics of the conditional models. In this section, we show that the error in downstream tasks is bounded by their sufficiency through direct applications of data-processing inequalities.

Zero-shot classification (ZSC). In the zero-shot classification task (Radford et al., 2021; Jia et al., 2021), the goal is to predict the label y for a new image $\mathbf{x}_{\text{im}}$ without having trained on a task-specific dataset. The ZSC approach starts by sampling $(\mathbf{x}_{\text{tx}}(y))_{y \in \mathcal{Y}}$ from a chosen distribution, computing the similarity score functions $(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}(y)))_{y \in \mathcal{Y}}$, and then selecting the predicted label for $\mathbf{x}_{\text{im}}$ using the formula $\arg \max_{y \in \mathcal{Y}} S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}(y))$. To provide a theoretical foundation for this method, we assume the data distribution satisfies a conditional independence criterion:

Assumption 2.2 (Conditional independence). *For the joint distribution $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, y) \sim \mathbb{P}_{\text{im,tx,cls}}$, the image $\mathbf{x}_{\text{im}}$ and the label y are conditionally independent given $\mathbf{x}_{\text{tx}}$. Notably, a special case of this assumption arises when y is a deterministic function of $\mathbf{x}_{\text{tx}}$.*

We propose a modified zero-shot classification procedure and establish a theoretical guarantee for its performance. For each $y \in \mathcal{Y}$, we generate M independent samples $(\mathbf{x}_{\text{tx}}^{(j)}(y))_{j \in [M]} \sim_{\text{iid}} \mathbb{P}_{\text{tx|cls}}(\mathbf{x}_{\text{tx}}|y)$. The classifier's predicted distribution is then defined as the softmax over aggregated score functions $L(\mathbf{x}_{\text{im}}, (\mathbf{x}_{\text{tx}}^{(j)}(y))_{j \in [M]})$:

$$\begin{aligned} \hat{\mathbb{P}}_S^{(M)}(\cdot|\mathbf{x}_{\text{im}}) &:= \text{softmax}((L(\mathbf{x}_{\text{im}}, (\mathbf{x}_{\text{tx}}^{(j)}(y))_{j \in [M]}))_{y \in \mathcal{Y}}), \\ L(\mathbf{x}_{\text{im}}, (\mathbf{x}_{\text{tx}}^{(j)}(y))_{j \in [M]}) &= \log \left[M^{-1} \sum_{j=1}^M \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}^{(j)}(y))) \right] + \log \mathbb{P}(y). \end{aligned} \quad (2.7)$$

Proposition 2.4 below shows that the error rate of this classifier $\hat{\mathbb{P}}_S^{(M)}(\cdot|\mathbf{x}_{\text{im}})$ is bounded by the sufficiency of the similarity score, and hence bounded by the CLIP excess risk.

Proposition 2.4 (Zero-shot classification error bound). *Assume Assumption 2.1 and 2.2 hold. Let $\hat{\mathbb{P}}_S^{(M)}(\cdot|\mathbf{x}_{\text{im}})$ be as defined in Eq. (2.7), and let $\mathbb{P}_{\text{cls|im}}(\cdot|\mathbf{x}_{\text{im}}) \in \mathcal{P}(\mathcal{Y})$ denote the conditional distribution of y given $\mathbf{x}_{\text{im}}$ under $\mathbb{P}_{\text{im,tx,cls}}$. Then there exists a constant $C > 0$, which depends polynomially on c_1 , such that for any $S \in \mathcal{S}$, with probability at least $1 - \delta$,*

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\mathbb{P}_{\text{cls|im}}(y|\mathbf{x}_{\text{im}}) \middle| \middle| \hat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}}) \right) \right] &\leq \text{Suff}(S) + C \cdot \frac{\log(2/\delta)}{M} \\ &\leq \underbrace{\left[\bar{R}_{\text{clip},K}(S) - \bar{R}_{\text{clip},K}(S_\star) \right]}_{\text{CLIP excess risk}} \left(1 + \frac{C}{K} \right) + C \cdot \frac{\log(2/\delta)}{M}. \end{aligned}$$

The proof of Proposition 2.4 is provided in Appendix A.4.4. Proposition 2.4 establishes that the zero-shot classification (ZSC) approach performs well when the similarity score is

near-sufficient and M is large. Notably, the original ZSC method in CLIP corresponds to the argmax decision rule applied on $\widehat{\mathbb{P}}_S^{(M)}(\cdot|\mathbf{x}_{\text{im}})$ with $M = 1$. This method performs well using only a single text sample, likely because $\exp(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}^{(j)}(y)))$ exhibits strong concentration around its expectation for fixed pairs of $(\mathbf{x}_{\text{im}}, y)$, thus reducing the need for averaging over multiple samples of $\mathbf{x}_{\text{tx}}^{(j)}(y)$. Our simulations on both synthetic and real data further show that increasing M improves ZSC performance, with the gain scaling as $1/M$, consistent with Proposition 2.4; see Appendix A.9.2 for more details.

Conditional Diffusion Models (CDMs). Text-to-image CDMs take text prompts as input and generate natural images by solving a stochastic differential equation (SDE). We consider the stochastic localization formulation (Eldan, 2013; El Alaoui et al., 2022) of CDMs, where the drift term of the SDE is determined by a neural network trained to approximate the conditional denoising function $\mathbf{m}_t : \mathbb{R}^{d_{\text{im}}} \times \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}^{d_{\text{im}}}$, defined as

$$\mathbf{m}_t(\mathbf{z}, \mathbf{x}_{\text{tx}}) = \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{g}) \sim \mathbb{P}_{\text{im}|\text{tx}}(\cdot|\mathbf{x}_{\text{tx}}) \times \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{\text{im}}})}[\mathbf{x}_{\text{im}} | \mathbf{z} = t \cdot \mathbf{x}_{\text{im}} + \sqrt{t} \cdot \mathbf{g}, \mathbf{x}_{\text{tx}}]. \quad (2.8)$$

This neural network approximates the conditional denoising function by minimizing risk over $\mathcal{M}_t \subseteq \{\mathbf{M}_t : \mathbb{R}^{d_{\text{im}}} \times \mathbb{R}^p \rightarrow \mathbb{R}^{d_{\text{im}}}\}$, a function class where the inputs are a noisy image and the CLIP text representation. The population risk minimization formulation gives

$$\begin{aligned} \widehat{\mathbf{M}}_t &= \arg \min_{\mathbf{M}_t \in \mathcal{M}_t} \left\{ R_{\text{cdm}, t}(\mathbf{M}_t, \mathbf{E}_{\text{tx}}) \right. \\ &\quad \left. := \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, \mathbf{g}) \sim \mathbb{P}_{\text{im}, \text{tx}} \times \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{\text{im}}})} \left[\left\| \mathbf{x}_{\text{im}} - \mathbf{M}_t(t\mathbf{x}_{\text{im}} + \sqrt{t}\mathbf{g}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) \right\|_2^2 \right] \right\}. \end{aligned} \quad (2.9)$$

Notice that the global minimizer of this formulation, when $\mathcal{M}_t$ includes all measurable functions, yields $\widehat{\mathbf{M}}_t(\mathbf{z}, \mathbf{E}) = \mathbb{E}[\mathbf{x}_{\text{im}} | \mathbf{z} = t\mathbf{x}_{\text{im}} + \sqrt{t}\mathbf{g}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}) = \mathbf{E}]$, which differs from the true conditional denoising function $\mathbf{m}_t(\mathbf{z}, \mathbf{x}_{\text{tx}})$, as defined in Eq. (2.8). Nevertheless, Proposition 2.5 below shows that the estimation error of $\widehat{\mathbf{M}}_t$ is bounded by the sufficiency of the text encoder $\mathbf{E}_{\text{tx}}$, and hence bounded by the CLIP excess risk.

Proposition 2.5 (Estimation error bound for CDMs). *Assume $\sup_{\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}} \|\mathbf{x}_{\text{im}}\|_{\infty} \leq B_{x_{\text{im}}}$, and let $\mathcal{M}_t$ include all measurable functions. Let the joint distribution of $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, \mathbf{z}_t)$ be given by $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, \mathbf{g}) \sim \mathbb{P}_{\text{im}, \text{tx}} \times \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{\text{im}}})$ with $\mathbf{z}_t = t \cdot \mathbf{x}_{\text{im}} + \sqrt{t} \cdot \mathbf{g}$. Then for any $t \geq 0$, the error rate of $\widehat{\mathbf{M}}_t$, as defined in Eq. (2.9), is bounded by the sufficiency of the encoder $\mathbf{E}_{\text{tx}}$:*

$$\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, \mathbf{z}_t)} \left[\frac{1}{d_{\text{im}}} \cdot \left\| \mathbf{m}_t(\mathbf{z}_t, \mathbf{x}_{\text{tx}}) - \widehat{\mathbf{M}}_t(\mathbf{z}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) \right\|_2^2 \right] \leq 2B_{x_{\text{im}}}^2 \cdot \text{Suff}(\mathbf{E}_{\text{tx}}). \quad (2.10)$$

The proof of Proposition 2.5 is provided in Appendix A.4.5. Briefly, the left-hand-side of (2.10), scaled by a factor of $2B_{x_{\text{im}}}^2 d_{\text{im}}$, can be bounded as follows:

$$\begin{aligned} &\mathbb{E}_{(\mathbf{x}_{\text{tx}}, \mathbf{z}_t)} [\text{D}_{\text{KL}}(\mathbb{P}(\mathbf{x}_{\text{im}} | \mathbf{x}_{\text{tx}}, \mathbf{z}_t) || \mathbb{P}(\mathbf{x}_{\text{im}} | \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}), \mathbf{z}_t))] \\ &\leq \mathbb{E}_{\mathbf{x}_{\text{tx}}} [\text{D}_{\text{KL}}(\mathbb{P}(\mathbf{x}_{\text{im}} | \mathbf{x}_{\text{tx}}) || \mathbb{P}(\mathbf{x}_{\text{im}} | \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})))] \leq \text{Suff}(\mathbf{E}_{\text{tx}}), \end{aligned}$$

where the first inequality follows from the data-processing inequality, and the second inequality is due to the definition of $\text{Suff}(\mathbf{E}_{\text{tx}})$.

Using Proposition 2.5 along with a standard Girsanov theorem analysis of diffusion models, we derive Corollary 2.6, which provides a sampling error bound of CDMs.

Corollary 2.6 (Sampling Error Bound for CDMs). *Under the setting and assumptions of Proposition 2.5, let $\bar{\mathbb{P}}_{\text{im}|\text{tx}}^{(T)}(\cdot|\mathbf{x}_{\text{tx}})$ denote the distribution of $\mathbf{Y}_T/T$, where $\mathbf{Y}_t$ is the solution to the SDE with drift term given by the risk minimizer $\hat{\mathbf{M}}_t$ in Eq. (2.9):*

$$d\mathbf{Y}_t = \hat{\mathbf{M}}_t(\mathbf{Y}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))dt + d\mathbf{W}_t, \quad \mathbf{z}_0 = \mathbf{0}, \quad \mathbf{W}_t \text{ is Brownian motion.}$$

Let $\mathbb{P}_{\text{im}|\text{tx}}^{\square\delta}(\cdot|\mathbf{x}_{\text{tx}})$ denote the distribution of $\mathbf{x}_{\text{im}} + \delta\mathbf{g}$, where $(\mathbf{x}_{\text{im}}, \mathbf{g}) \sim \mathbb{P}_{\text{im}|\text{tx}}(\cdot|\mathbf{x}_{\text{tx}}) \times \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{\text{im}}})$. Then we have the following bound on the sampling error

$$\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} [\text{D}_{\text{KL}}(\mathbb{P}_{\text{im}|\text{tx}}^{\square\frac{1}{\sqrt{T}}}(\cdot|\mathbf{x}_{\text{tx}}) || \bar{\mathbb{P}}_{\text{im}|\text{tx}}^{(T)}(\cdot|\mathbf{x}_{\text{tx}}))] \leq d_{\text{im}} B_{x_{\text{im}}}^2 T \cdot \text{Suff}(\mathbf{E}_{\text{tx}}).$$

The proof of Corollary 2.6 is provided in Appendix A.4.5. Additionally, we perform similar analyses for the sampling error bound for vision-language models in Appendix A.3.1.

2.3.3 Adaptation to tasks with canonical representation

In certain cases, the joint distribution of images and text admits canonical representations $(\mathbf{E}_{\text{im},\star}, \mathbf{E}_{\text{tx},\star})$, which serve as sufficient statistics and are also sufficient for downstream tasks. We show that under certain conditions on these canonical representations, a simple adapter Adap—a small neural network—can transform any near-minimizer $(\mathbf{E}_{\text{im}}, \mathbf{E}_{\text{tx}})$ of the CLIP risk into the canonical representations $(\mathbf{E}_{\text{im},\star}, \mathbf{E}_{\text{tx},\star})$. Consequently, near-minimizers of the CLIP risk can effectively adapt to downstream tasks using these canonical representations as sufficient statistics. To formalize this idea, we impose the following assumption on the canonical representations of the joint distribution $\mathbb{P}_{\text{im},\text{tx}}$. Specifically, we require that the representation functions are linearly independent and that the inverse of the true link function is Lipschitz.

Assumption 2.3 (Well-posed canonical representation). *Assume there exist canonical representations $\mathbf{E}_{\text{im},\star} : \mathcal{X}_{\text{im}} \rightarrow \mathbb{R}^{p_\star}$ and $\mathbf{E}_{\text{tx},\star} : \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}^{p_\star}$, along with a univariate, monotone, and invertible link function $\Upsilon_\star$, such that*

$$\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) := \log \frac{\mathbb{P}_{\text{im},\text{tx}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})}{\mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}})\mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx}})} =: \Upsilon_\star(\langle \mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}}) \rangle).$$

We further assume the following conditions on $\mathbf{E}_{\text{im},\star}$ and $\Upsilon_\star$:

- (a) $\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} [\mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}}) \mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}})^\top] \succeq \mathbf{I}_{p_\star} / L_B^2$ for some $L_B > 0$.
- (b) *The true link function $\Upsilon_\star$ is invertible over the feasible range of $\langle \mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}}) \rangle$, and its inverse function $\Upsilon_\star^{-1}$ is L_Γ -Lipschitz.*

We provide two examples where these assumptions are satisfied.

Example 2.7 (Separator representation). Let $\mathbf{s} \in \mathcal{S}$ with $|\mathcal{S}| = p_\star$ be a separator of $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$, meaning that under the joint distribution $\mathbb{P}_{\text{im}, \text{tx}, \text{sp}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, \mathbf{s})$, $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$ are conditionally independent given $\mathbf{s}$. In this case, the canonical representations are given by $\mathbf{E}_{\text{im}, \star}(\mathbf{x}_{\text{im}}) = [\mathbb{P}(\mathbf{s}|\mathbf{x}_{\text{im}})/\mathbb{P}(\mathbf{s})]_{\mathbf{s} \in \mathcal{S}} \in \mathbb{R}^{p_\star}$ and $\mathbf{E}_{\text{tx}, \star}(\mathbf{x}_{\text{tx}}) = [\mathbb{P}(\mathbf{s}|\mathbf{x}_{\text{tx}})]_{\mathbf{s} \in \mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$, with the link function defined as $\Upsilon_\star(t) = \log(t)$. This setup leads to

$$\Upsilon_\star(\langle \mathbf{E}_{\text{tx}, \star}(\mathbf{x}_{\text{tx}}), \mathbf{E}_{\text{im}, \star}(\mathbf{x}_{\text{im}}) \rangle) = \log \sum_{\mathbf{s} \in \mathcal{S}} \frac{\mathbb{P}(\mathbf{s}|\mathbf{x}_{\text{im}})\mathbb{P}(\mathbf{s}|\mathbf{x}_{\text{tx}})}{\mathbb{P}(\mathbf{s})} = \log\{\mathbb{P}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})/[\mathbb{P}(\mathbf{x}_{\text{im}})\mathbb{P}(\mathbf{x}_{\text{tx}})]\}.$$

In this example, Assumption 2.3(a) holds if the matrix $(\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}}[\frac{\mathbb{P}(\mathbf{s}_1|\mathbf{x}_{\text{im}})\mathbb{P}(\mathbf{s}_2|\mathbf{x}_{\text{im}})}{\mathbb{P}(\mathbf{s}_1)\mathbb{P}(\mathbf{s}_2)}])_{\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{S}} \succeq \frac{\mathbf{I}_{p_\star}}{L_B^2}$; Assumption 2.3(b) holds if we impose a uniform upper bound on $\frac{\mathbb{P}_{\text{im}, \text{tx}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})}{\mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}})\mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx}})}$.

Example 2.8 (Exponential family representation). Take $\Upsilon_\star(t) = t$ as the identity function. Then, $\mathbb{P}_{\text{im}|\text{tx}}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}}) = \mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}}) \exp\{\langle \mathbf{E}_{\text{im}, \star}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx}, \star}(\mathbf{x}_{\text{tx}}) \rangle\}$ defines an exponential family, with $\mathbf{E}_{\text{tx}, \star}(\mathbf{x}_{\text{tx}})$ as the natural parameter and $\mathbf{E}_{\text{im}, \star}(\mathbf{x}_{\text{im}})$ as the sufficient statistic (a similar formulation holds for the reverse conditional distribution). In this case, Assumption 2.3(b) is automatically satisfied, as $\Upsilon_\star^{-1}$ is simply the identity function.

Recall that we assumed representations $\mathbf{E}_{\text{im}} : \mathcal{X}_{\text{im}} \rightarrow \mathbb{R}^p$ and $\mathbf{E}_{\text{tx}} : \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}^p$, along with a link function $\Upsilon : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$, such that $\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) := \Upsilon(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))$ is a near-minimizer of the CLIP risk $\bar{\mathbf{R}}_{\text{clip}, K}$, rendering $\mathbf{E}_{\text{im}}$ and $\mathbf{E}_{\text{tx}}$ near-sufficient. The following result shows that a simple adapter exists that can transform these near-sufficient representations $(\mathbf{E}_{\text{im}}, \mathbf{E}_{\text{tx}})$ into the canonical representations $(\mathbf{E}_{\text{im}, \star}, \mathbf{E}_{\text{tx}, \star})$.

Proposition 2.9 (Near-equivalence to the canonical representations). Suppose Assumption 2.1 and Assumption 2.3 hold. Let $M \geq 1$ be some integer, and define $B_{\text{Adap}} := (M \cdot \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \|\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})\|_2^2)^{1/2}$. Then, there exists a constant $C > 0$, which depends polynomially on c_1 , and a parameter $\boldsymbol{\theta} = (W_{\text{ada}}^{(1)} \in \mathbb{R}^{p_\star \times M}, W_{\text{ada}}^{(2)} \in \mathbb{R}^{M \times p})$ with $\|W_{\text{ada}}^{(1)}\|_{\text{op}} \leq CL_B/\sqrt{M}$, $\|W_{\text{ada}}^{(2)}\|_{\text{op}} \leq CB_{\text{Adap}}$, such that defining a simple adapter

$$\text{Adap}_{\boldsymbol{\theta}}(\mathbf{E}_{\text{tx}}) := W_{\text{ada}}^{(1)} \left(\Upsilon_\star^{-1} \left(\log \left[M \cdot \text{softmax}(\Upsilon(W_{\text{ada}, j}^{(2)}, \mathbf{E}_{\text{tx}})) \right] \right) \right)_{j \in [M]},$$

the transformed embedding $\widehat{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}) := \text{Adap}_{\boldsymbol{\theta}}(\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))$ satisfies

$$\mathbb{E}_{\mathbf{x}_{\text{tx}}} [\|\widehat{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}) - \mathbf{E}_{\text{tx}, \star}(\mathbf{x}_{\text{tx}})\|_2^2] \leq C \cdot L_B^2 \cdot L_\Gamma^2 \cdot p_\star \cdot (\text{Suff}(\mathbf{S}) + M^{-1}). \quad (2.11)$$

The proof of Proposition 2.9 is provided in Appendix A.4.7. In short, we exploit the fact that $\Upsilon_\star(\langle \mathbf{E}_{\text{im}, \star}, \mathbf{E}_{\text{tx}, \star} \rangle) \approx \Upsilon(\mathbf{E}_{\text{im}}, \mathbf{E}_{\text{tx}})$, which leads to the heuristic approximation $\mathbf{E}_{\text{tx}, \star} \approx \mathbf{E}_{\text{im}, \star}^\dagger \Upsilon_\star^{-1} \Upsilon(\mathbf{E}_{\text{im}}, \mathbf{E}_{\text{tx}})$. Here, $\mathbf{E}_{\text{im}, \star}^\dagger$ is interpreted as a high-dimensional matrix. To reduce the dimensionality, we introduce a sampling approach to approximate $\mathbf{E}_{\text{im}, \star}^\dagger \Upsilon_\star^{-1} \Upsilon(\mathbf{E}_{\text{im}}, \mathbf{E}_{\text{tx}})$ with lower-dimensional operators. A similar approach was used in Tosh et al. (2021a) in a more restricted setting: when the link functions $(\Upsilon_\star, \Upsilon)$ are the logarithm, the canonical representation $\mathbf{E}_{\text{tx}, \star}$ can be efficiently recovered via a linear transformation of $\mathbf{E}_{\text{tx}}$.

Remark 2.1 (Adaptation to downstream tasks with canonical representation). When Υ is a simple function, the adapter Adap_θ can be efficiently approximated by a shallow neural network. Consequently, consider a target function $f_\star(\mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}}))$ that depends on $\mathbf{x}_{\text{tx}}$ through the canonical representation $\mathbf{E}_{\text{tx},\star}$, and assume that $f_\star$ can be efficiently approximated by a neural network. Under the conditions of Proposition 2.9, where $S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) := \Upsilon(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))$ is a near-minimizer of the CLIP risk $\bar{R}_{\text{clip},K}$, it follows that $f_\star(\mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}}))$ can be efficiently approximated by a neural network applied to $\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})$.

This strategy will be applied in Section 2.4.2 and Appendix A.3.2 to conditional diffusion models (CDMs) and vision-language models (VLMs), where we construct efficient neural network approximations for prediction functions based on pre-trained CLIP encoders.

Remark 2.2 (An improved bound). The error bound in Eq. (2.11) depends on the embedding $\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})$ through the term $\text{Suff}(S)$. This term can be replaced by $\text{Suff}(\mathbf{E}_{\text{tx}})$ if the link function Υ and the image embedding $\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})$ are chosen such that $\Upsilon(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) = \log \frac{\mathbb{P}_{\text{im}|\text{tx}}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}})}{\mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}})}$. However, while such a link function and embedding exist in principle, there is no guarantee that the link function is “simple” and can be efficiently approximated by a shallow neural network. We refer readers to the end of Appendix A.4.7 for more details.

2.4 Sample-efficient learning in hierarchical models

In the previous section, we showed that near-minimizers of the CLIP risk are near-sufficient and adaptable to downstream tasks, including zero-shot classification (ZSC), conditional diffusion models (CDMs), and vision-language models (VLMs). Despite these findings, it remains unclear why certain neural networks can efficiently learn these near-minimizers and the associated functions within CDMs and VLMs. In this section, we address this question by introducing a concrete data generation model for image-text pairs.

Specifically, we assume that the image-text pairs are generated according to a *joint generative hierarchical model* (JGHM), which integrates two generative hierarchical models (GHMs) with a shared root. A GHM is a tree-structured graphical model in which the root node represents the highest-level features; these features hierarchically generate lower-level features based on a transition kernel, eventually reaching the leaf nodes that represent observed images or text. GHMs have been widely used in theoretical modeling for images and language independently (Mossel, 2016; Petrini et al., 2023; Sclocchi et al., 2024b; Tomasini and Wyart, 2024; Cagnetta and Wyart, 2024; Garnier-Brun et al., 2024; Kadkhodaie et al., 2023a,b; Mei, 2024). The JGHM framework extends GHMs to jointly model paired image and text data². In the following, we formally define the JGHM, building on the GHM framework presented in Mei (2024).

²We use the JGHM as a working model and acknowledge that it may not fully capture the complexity of the image-text distribution. Developing a more realistic model for image-text data is left for future work. In this chapter, we focus on a model that captures the hierarchical structure of image-text pairs and provides an efficient sample complexity bound.

The joint tree structure. Consider a joint tree structure $\mathcal{T} = \mathcal{T}_{\text{im}}, \mathcal{T}_{\text{tx}}$, consisting of two trees, $\mathcal{T}_{\text{im}}$ and $\mathcal{T}_{\text{tx}}$, each of height L . These trees generate images and text, respectively, and share a common root node, r , which represents shared information across the image and text domains. Let the sets of nodes in the image and text trees be $\mathcal{V}_{\text{im}}$ and $\mathcal{V}_{\text{tx}}$, respectively. The root is defined as level 0, and the set of nodes at a distance ℓ from the root is referred to as level ℓ . These nodes are denoted by $\mathcal{V}_{\text{im}}^{(\ell)}$ in the image tree and $\mathcal{V}_{\text{tx}}^{(\ell)}$ in the text tree. Let $\mathcal{C}(v)$ represent the set of its children defined within either $\mathcal{T}_{\text{im}}$ or $\mathcal{T}_{\text{tx}}$, as appropriate. We assume that for any $v \in \mathcal{V}_{\text{im}}^{(\ell-1)}$ (or $\mathcal{V}_{\text{tx}}^{(\ell-1)}$), the number of children is fixed at $m_{\text{im}}^{(\ell)}$ (or $m_{\text{tx}}^{(\ell)}$) for $\ell \in [L]$, except for leaf nodes $v \in \mathcal{V}_{\text{im}}^{(L)}$ (or $\mathcal{V}_{\text{tx}}^{(L)}$), which have no children. The number of nodes at each layer is denoted by $d_{\text{im}}^{(\ell)} = |\mathcal{V}_{\text{im}}^{(\ell)}|$ and $d_{\text{tx}}^{(\ell)} = |\mathcal{V}_{\text{tx}}^{(\ell)}|$. In particular, the total number of leaf nodes is represented by $d_{\text{im}} = d_{\text{im}}^{(L)} = |\mathcal{V}_{\text{im}}^{(L)}|$ and $d_{\text{tx}} = d_{\text{tx}}^{(L)} = |\mathcal{V}_{\text{tx}}^{(L)}|$. Additionally, we define $\underline{m} := \max\{m_{\text{im}}^{(1)}, m_{\text{tx}}^{(1)}\}$, $\overline{m} := \max_{\ell \in [L]} \max\{m_{\text{im}}^{(\ell)}, m_{\text{tx}}^{(\ell)}\}$.

Joint generative hierarchical models (JGHMs). Building on the joint tree structure, we define the joint generative model for the image $\mathbf{x}_{\text{im}}$ and text $\mathbf{x}_{\text{tx}}$. Each node in the tree is associated with a variable: the root node is represented by $x_r^{(0)} = x_{\text{im},r}^{(0)} = x_{\text{tx},r}^{(0)} \in \mathcal{S}_r$; for nodes $v \in \mathcal{V}_{\text{im}}^{(\ell)}$ at levels $1 \leq \ell \leq L$, the variables are $x_{\text{im},v}^{(\ell)} \in \mathcal{S}_{\text{im}}$; and for nodes $v \in \mathcal{V}_{\text{tx}}^{(\ell)}$ at levels $1 \leq \ell \leq L$, the variables are $x_{\text{tx},v}^{(\ell)} \in \mathcal{S}_{\text{tx}}$. Here, $\mathcal{S}_r$, $\mathcal{S}_{\text{im}}$, and $\mathcal{S}_{\text{tx}}$ denote the spaces of root, image, and text variables, respectively. For simplicity, we set $\mathcal{S}_r = \mathcal{S}_{\text{im}} = \mathcal{S}_{\text{tx}} = [S]$ for some $S \in \mathbb{N}_{>0}$; however, our theoretical results extend naturally to the more general case where these spaces differ. We collectively denote the variables associated with $\mathcal{V}_{\text{im}}^{(\ell)}$ and $\mathcal{V}_{\text{tx}}^{(\ell)}$ as $\mathbf{x}_{\text{im}}^{(\ell)} = (x_{\text{im},v}^{(\ell)})_{v \in \mathcal{V}_{\text{im}}^{(\ell)}}$ and $\mathbf{x}_{\text{tx}}^{(\ell)} = (x_{\text{tx},v}^{(\ell)})_{v \in \mathcal{V}_{\text{tx}}^{(\ell)}}$, respectively. For the leaf level $\ell = L$, we sometimes omit the superscript (L) for brevity.

The joint distribution $\mu_{\star}(x_r^{(0)}, \mathbf{x}_{\text{im}}^{(1)}, \dots, \mathbf{x}_{\text{im}}^{(L-1)}, \mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}^{(1)}, \dots, \mathbf{x}_{\text{tx}}^{(L-1)}, \mathbf{x}_{\text{tx}})$ is defined as

$$\begin{aligned} & \mu_{\star}(x_r^{(0)}, \mathbf{x}_{\text{im}}^{(1)}, \dots, \mathbf{x}_{\text{im}}^{(L-1)}, \mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}^{(1)}, \dots, \mathbf{x}_{\text{tx}}^{(L-1)}, \mathbf{x}_{\text{tx}}) \\ & \propto \psi_r^{(0)}(x_r^{(0)}) \cdot \psi_{\text{im}}^{(1)}(x_r^{(0)}, \mathbf{x}_{\text{im}}^{(1)}) \cdot \left(\prod_{v \in \mathcal{V}_{\text{im}}^{(1)}} \psi_{\text{im}}^{(2)}(x_{\text{im},v}^{(1)}, x_{\text{im},\mathcal{C}(v)}^{(2)}) \right) \cdots \left(\prod_{v \in \mathcal{V}_{\text{im}}^{(L-1)}} \psi_{\text{im}}^{(L)}(x_{\text{im},v}^{(L-1)}, x_{\text{im},\mathcal{C}(v)}^{(L)}) \right) \\ & \quad \cdot \psi_{\text{tx}}^{(1)}(x_r^{(0)}, \mathbf{x}_{\text{tx}}^{(1)}) \cdot \left(\prod_{v \in \mathcal{V}_{\text{tx}}^{(1)}} \psi_{\text{tx}}^{(2)}(x_{\text{tx},v}^{(1)}, x_{\text{tx},\mathcal{C}(v)}^{(2)}) \right) \cdots \left(\prod_{v \in \mathcal{V}_{\text{tx}}^{(L-1)}} \psi_{\text{tx}}^{(L)}(x_{\text{tx},v}^{(L-1)}, x_{\text{tx},\mathcal{C}(v)}^{(L)}) \right), \end{aligned} \tag{2.12}$$

where $\psi_r^{(0)}: [S] \rightarrow \mathbb{R}_{\geq 0}$, $\psi_{\text{im}}^{(\ell)}: [S] \times [S]^{m_{\text{im}}^{(\ell)}} \rightarrow \mathbb{R}_{\geq 0}$, and $\psi_{\text{tx}}^{(\ell)}: [S] \times [S]^{m_{\text{tx}}^{(\ell)}} \rightarrow \mathbb{R}_{\geq 0}$ define the transition probabilities of child nodes conditioned on their parent node. This joint distribution models the image-text generation process. Starting at the shared root of the image-text tree, initialized according to $\psi_r^{(0)}$, values are sampled level-by-level through the transition probabilities $\psi_{\{\text{im}, \text{tx}\}}^{(\ell)}$. This process continues until the leaf variables, $\mathbf{x}_{\text{im}}$ and $\mathbf{x}_{\text{tx}}$, are generated. The observed data consist of the leaf variables $\mathbf{x}_{\text{im}}$ (image) and $\mathbf{x}_{\text{tx}}$ (text), while the intermediate variables are typically unobserved.

We impose the following factorization assumption on the ψ functions in the JGHM model. This assumption implies that the child nodes are conditionally independent of the parents, and that the transition is homogeneous across all nodes within a particular layer. Although

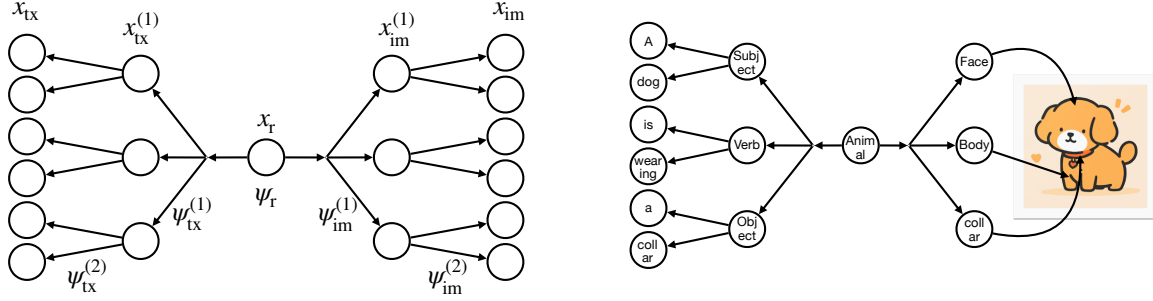


Figure 2.1: Left: the JGHM used to generate the joint distribution of text and images. Right: an illustrative example of a generated text-image pair.

this assumption, inherited from Mei (2024), is not strictly necessary for the theoretical framework and could be relaxed with additional technical work, it significantly simplifies the presentation and proof. Therefore, we retain it here for convenience.

Assumption 2.4 (Factorization of ψ). *For each $v \in \mathcal{V}_{im}^{(\ell)}$, let there be a known ordering function $\iota : \mathcal{C}_{im}(v) \rightarrow [m_{im}^{(\ell)}]$ that is bijective. A similar ordering function ι is defined for each $v \in \mathcal{V}_{tx}^{(\ell)}$ as well. For $\square \in \{im, tx\}$, each layer $\ell \in [L]$ and node $v \in \mathcal{V}_{\square}^{(\ell-1)}$, we assume*

$$\psi_{\square}^{(\ell)}(x_{\square,v}^{(\ell-1)}, x_{\square,\mathcal{C}(v)}^{(\ell)}) = \prod_{v' \in \mathcal{C}(v)} \psi_{\square,\iota(v')}^{(\ell)}(x_{\square,v}^{(\ell-1)}, x_{\square,v'}^{(\ell)}). \quad (2.13)$$

In addition, we assume boundedness for the ψ functions.

Assumption 2.5 (Boundedness of ψ). *There exists some $B_{\psi} > 0$ such that for any $x, x' \in [S]$,*

$$1/B_{\psi} \leq \psi_r^{(0)}(x), \psi_{im,\iota}^{(\ell)}(x, x'), \psi_{tx,\iota}^{(\ell)}(x, x') \leq B_{\psi}. \quad (2.14)$$

A schematic illustration of the JGHM with two layers is shown in Figure 2.1.

2.4.1 Sample-efficient learning of CLIP encoders and ZSC

Consider a set of nK i.i.d. samples $\{(\mathbf{x}_{im}^{(i)}, \mathbf{x}_{tx}^{(i)})\}_{i \geq 1}$ drawn from the distribution $\mu_{\star}$ under the JGHM. They can be reorganized into the form $\{(\mathbf{x}_{im,j}^{(i)}, \mathbf{x}_{tx,j}^{(i)})_{j \in [K]}\}_{i \in [n]}$, where K is the batch size and n is the number of batches. Our goal is to learn encoders for both the image and text components by minimizing the CLIP loss. The optimal similarity score under the CLIP loss is given by the logarithmic probability ratio $S_{\star}(\mathbf{x}_{im}, \mathbf{x}_{tx}) = \log \frac{\mu_{\star}(\mathbf{x}_{im}, \mathbf{x}_{tx})}{\mu_{\star}(\mathbf{x}_{im})\mu_{\star}(\mathbf{x}_{tx})}$. We seek to analyze the sample complexity required to learn this optimal similarity score using empirical risk minimization over the class of transformers.

The neural network architecture. The similarity score consists of three main components: a transformer encoder³ for images, $\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}: [S]^{d_{\text{im}}} \rightarrow \mathbb{R}^S$; a transformer encoder for text, $\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}: [S]^{d_{\text{tx}}} \rightarrow \mathbb{R}^S$; and a parameterized similarity link function, $\tau^w(h, h') = \log \text{trun}(\sum_{s \in [S]} h_s h'_s w_s)$, where $w, h, h' \in \mathbb{R}^S$, and $\text{trun}(\cdot): \mathbb{R} \mapsto \mathbb{R}_{>0}$ is a truncation function. The similarity score, $\mathbf{S}_{\text{NN}}^\theta$, with parameters $\theta = (\mathbf{W}_{\text{im}}, \mathbf{W}_{\text{tx}}, w)$, is defined as

$$\mathbf{S}_{\text{NN}}^\theta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) := \tau^w(\text{softmax}(\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}(\mathbf{x}_{\text{im}})), \text{softmax}(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))). \quad (2.15)$$

The same network architecture is used for the vision transformer $\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}$ and the text transformer $\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}$, but with different weights. For simplicity, we describe the architecture generically and omit subscripts. The neural network output is given by $\text{NN}^{\mathbf{W}}(\mathbf{x}) = \text{read}_{\text{clip}}(\text{TF}^{\mathbf{W}}(\text{Emb}_{\text{clip}}(\mathbf{x})))$. The only trainable part, $\text{TF}^{\mathbf{W}}$, is a repetition of transformer blocks as described below. The fixed embedding function, $\text{Emb}_{\text{clip}}: \mathbb{R}^d \rightarrow \mathbb{R}^{D \times d}$, maps the input $\mathbf{x} \in \mathbb{R}^d$ (including positional encoding) to a matrix $\mathbf{H}^{(L)} = \text{Emb}_{\text{clip}}(\mathbf{x}) \in \mathbb{R}^{D \times d}$, and the fixed readout function, $\text{read}_{\text{clip}}: \mathbb{R}^{D \times d} \rightarrow \mathbb{R}^S$, extracts an $(S \times 1)$ submatrix from the output of the last transformer block $\mathbf{H}^{(0)} = \text{TF}^{\mathbf{W}}(\text{Emb}_{\text{clip}}(\mathbf{x}))$. Definitions of the functions Emb_{clip} and $\text{read}_{\text{clip}}$ are provided in Appendix A.5.1.

Definition 2.2 (The transformer architecture). The transformer, $\text{TF}^{\mathbf{W}}: \mathbb{R}^{D \times d} \rightarrow \mathbb{R}^{D \times d}$, consists of L -blocks of the $(J + 1)$ -layer fully-connected ReLU network $\text{FF}^{(\ell)}: \mathbb{R}^{D \times d} \rightarrow \mathbb{R}^{D \times d}$, applied column-wise, and the self-attention layer $\text{Attn}^{(\ell)}: \mathbb{R}^{D \times d} \rightarrow \mathbb{R}^{D \times d}$, defined as:

$$\begin{aligned} \text{FF}^{(\ell)}(\mathbf{x}) &= W_{J+1}^{(\ell)} [\cdot; 1] \circ \text{ReLU}(W_J^{(\ell)} [\cdot; 1]) \circ \cdots \circ \text{ReLU}(W_1^{(\ell)} [\mathbf{x}; 1]), \quad (\mathbf{x} \in \mathbb{R}^D) \\ \text{Attn}^{(\ell)}(Q) &= W_V^{(\ell)} Q \cdot \text{softmax}_{\text{col}}(Q^\top (W_K^{(\ell)})^\top W_Q^{(\ell)} Q). \end{aligned}$$

Here, $[\cdot; 1]$ appends a constant 1 to the end of a vector, introducing an intercept term. Starting from $\mathbf{H}^{(L)}$, the ℓ -th block computes intermediate representations $\mathbf{H}^{(\ell)} \in \mathbb{R}^{D \times d}$ and $Q^{(\ell)} \in \mathbb{R}^{D \times d}$ as follows:

$$\begin{aligned} Q^{(\ell)} &= \mathbf{H}^{(\ell)} + \text{FF}^{(\ell)}(\mathbf{H}^{(\ell)}), \\ \mathbf{H}^{(\ell-1)} &= \text{normalize}(Q^{(\ell)} + \text{Attn}^{(\ell)}(Q^{(\ell)})). \end{aligned}$$

For simplicity, $\text{FF}^{(\ell)}$ is treated as a function from $\mathbb{R}^{D \times d}$ to $\mathbb{R}^{D \times d}$, though applied column-wise. The transformer weights, denoted by $\mathbf{W}$ (subscripts im, tx correspond to specific transformers), are given by:

$$\begin{aligned} \mathbf{W} &= \left\{ W_Q^{(\ell)}, W_K^{(\ell)}, W_V^{(\ell)} \in \mathbb{R}^{D \times D}, W_1^{(\ell)} \in \mathbb{R}^{D' \times (D+1)}, \right. \\ &\quad \left. \{W_i^{(\ell)} \in \mathbb{R}^{D' \times (D'+1)}\}_{i=2}^J, W_{J+1}^{(\ell)} \in \mathbb{R}^{D \times (D'+1)} \right\}_{\ell \in [L]}. \end{aligned} \quad (2.16)$$

³While we use a transformer architecture to align with practical implementations, the theoretical framework does not require the use of transformers to avoid the curse of dimensionality. Any network capable of approximating the belief propagation algorithm can be utilized. We do not claim that transformers are the optimal architecture for this purpose.

Here, $\text{softmax}_{\text{col}}$ denotes a column-wise softmax operation, where for any matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, each column of $\text{softmax}_{\text{col}}(\mathbf{A}) \in \mathbb{R}^{d \times d}$ is the softmax of the corresponding column in $\mathbf{A}$. The function $\text{normalize} : \mathbb{R}^{D \times d} \rightarrow \mathbb{R}^{D \times d}$ performs column-wise normalization, where each column of $\text{normalize}(\mathbf{H}) \in \mathbb{R}^{D \times d}$ is the normalized version of the corresponding column in $\mathbf{H}$, with its formal definition provided in Appendix A.5.1.

Intuitively, each column vector of $\mathbf{H}^{(\ell)}$ corresponds to a leaf node v . As we will show in Appendix A.5.1, each transformer block approximates one step of belief propagation. Consequently, the blocks are indexed in decreasing order ($\ell = L, \dots, 1$) to align with the belief propagation process. Some modifications are also incorporated, such as placing the feedforward layer first and using a multi-layer network for the feedforward component. However, these changes are not essential and can be effectively simulated within the original transformer architecture (Vaswani, 2017).

The ERM estimator. To find the optimal similarity score, we solve the empirical risk minimization problem defined by the following objective:

$$\begin{aligned} \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B}}{\text{argmin}} \left\{ \hat{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}) := \frac{1}{n} \sum_{i=1}^n \left[-\frac{1}{K} \sum_{k=1}^K \log \frac{\exp(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))}{\sum_{j \in [K]} \exp(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}))} \right] \right. \\ \left. + \frac{1}{n} \sum_{i=1}^n \left[-\frac{1}{K} \sum_{k=1}^K \log \frac{\exp(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))}{\sum_{j \in [K]} \exp(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im},j}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))} \right] \right\}, \end{aligned} \quad (2.17)$$

where the parameter space is defined as:

$$\begin{aligned} \Theta_{L,J,D,D',B} &:= \left\{ \mathbf{W}_{\text{im}}, \mathbf{W}_{\text{tx}} \text{ as defined in Eq. (2.16)}, w \text{ as defined in Eq. (2.15)}; \right. \\ &\quad \left. \|\boldsymbol{\theta}\| := \|w\|_{\infty} \vee \max_{\square \in \{\text{im}, \text{tx}\}} \max_{i \in [J+1], \ell \in [L]} \{ \|W_{i,\square}^{(\ell)}\|_{\text{op}}, \|W_{Q,\square}^{(\ell)}\|_{\text{op}}, \|W_{K,\square}^{(\ell)}\|_{\text{op}}, \|W_{V,\square}^{(\ell)}\|_{\text{op}} \} \leq B \right\}. \end{aligned} \quad (2.18)$$

We expect the empirical risk minimizer, $\hat{\mathbf{S}}_{\text{NN}}^{\hat{\boldsymbol{\theta}}}$, to closely approximate the optimal similarity score $\mathbf{S}_{\star}$, which minimizes the population risk $\bar{\mathbf{R}}_{\text{clip},K}$ over all functions as defined in Eq. (2.1). This is quantified through the excess risk $\text{Excess}_K(\hat{\mathbf{S}}_{\text{NN}}^{\hat{\boldsymbol{\theta}}}, \mathbf{S}_{\star}) := \bar{\mathbf{R}}_{\text{clip},K}(\hat{\mathbf{S}}_{\text{NN}}^{\hat{\boldsymbol{\theta}}}) - \bar{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_{\star})$. The following theorem provides a bound on the excess risk of the estimator $\hat{\mathbf{S}}_{\text{NN}}^{\hat{\boldsymbol{\theta}}}$.

Theorem 2.10 (Sufficiency and excess risk bound of CLIP). *Suppose Assumption 2.4 and Assumption 2.5 hold. Let $\Theta_{L,J,D,D',B}$ denote the parameter space defined in Eq. (2.18), with $J = \tilde{\mathcal{O}}(L)$, $D = \mathcal{O}(SL)$, $D' = \tilde{\mathcal{O}}(\bar{m}SL^3)$, and $B = \tilde{\mathcal{O}}(SL + \bar{m}^2)$. Let $\hat{\boldsymbol{\theta}}$ be the empirical risk minimizer as defined in Eq. (2.17). Then, with probability at least $1 - 1/n$, we have*

$$\text{Excess}_K(\hat{\mathbf{S}}_{\text{NN}}^{\hat{\boldsymbol{\theta}}}, \mathbf{S}_{\star}) = \tilde{\mathcal{O}}\left(\sqrt{\frac{S^2 L^{11} \bar{m}^2}{n}}\right), \quad (2.19)$$

where $\tilde{\mathcal{O}}$ hides polynomial factors in $\log(\bar{m}SLnB_{\psi})$.

Moreover, combined with Proposition 2.3, this excess risk bound also provides an upper bound on the sufficiency of the learned encoders and the similarity score, $\text{Suff}(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}})$, $\text{Suff}(\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}})$, and $\text{Suff}(\hat{\mathbf{S}}_{\text{NN}}^{\hat{\boldsymbol{\theta}}})$.

The proof of Theorem 2.10 is provided in Appendix A.5. We note that the sample complexity bound in this theorem is not intended to be the tightest possible, and refining it remains an intriguing direction for future research. Theorem 2.10 establishes that the excess risk vanishes whenever $n \gg S^2 L^{11} \bar{m}^2$, with the required sample size being sub-linear in d . Crucially, this result avoids the curse of dimensionality, demonstrating that the JGHM can be efficiently learned via ERM over transformers. While simpler two-layer neural networks could be used as encoders, their approximation error and sample complexity would likely scale exponentially with the dimension d , leading to a curse of dimensionality. In contrast, transformers circumvent this issue by efficiently approximating belief-propagation. In Remark A.2, we further show that the $1/\sqrt{n}$ rate can be improved to $1/\sqrt{nK}$, so the bound scales with the total sample size rather than the number of batches. This improvement, however, comes at the cost of an exponential dependence on $\bar{m}$, which is unavoidable under the current assumptions (see Remark A.2 for more details).

Proof strategy of Theorem 2.10: Transformers efficiently approximate belief propagation. The excess risk $\text{Excess}_K(\mathbf{S}_{\text{NN}}^\theta, \mathbf{S}_*)$ can be decomposed into two components: approximation error and generalization error:

$$\text{Excess}_K(\mathbf{S}_{\text{NN}}^\theta, \mathbf{S}_*) \leq \underbrace{\inf_{\theta \in \Theta} \bar{R}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^\theta) - \bar{R}_{\text{clip},K}(\mathbf{S}_*)}_{\text{approximation error}} + 2 \cdot \underbrace{\sup_{\theta \in \Theta} \left| \hat{R}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^\theta) - \bar{R}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^\theta) \right|}_{\text{generalization error}}.$$

The generalization error is controlled using standard parameter counting arguments and the chaining approach. The main focus, therefore, lies in bounding the approximation error. This is achieved by first introducing the belief propagation (BP) algorithm, which computes the conditional probabilities $(\mathbb{P}(x_r | \mathbf{x}_{\text{im}}), \mathbb{P}(x_r | \mathbf{x}_{\text{tx}}))$, as shown in Eq. (A.51), and then showing that transformers can effectively approximate BP. See Appendix A.5.1 for more detail.

Remark 2.3. We note that while the BP algorithm serves as a theoretical proof technique, we cannot conclude that the pre-trained CLIP encoders implement BP in JGHM. Investigating whether the trained CLIP encoders approximate BP remains an intriguing direction for future interpretability research. Our simulation studies in out-of-distribution settings, as shown in Figure A.6, provide partial evidence relevant to this question. This remark also applies to the CDM and VLM tasks.

Remark 2.4. While classical algorithms such as maximum likelihood estimation can also efficiently learn the similarity score from JGHM, our theory shows that a neural network (NN)-based approach with contrastive pre-training can achieve the same result. A key advantage of NN-based approaches is their flexibility: they rely less on the precise specification of the underlying graphical model, whereas classical methods struggle if the model is misspecified. This makes NN-based approaches especially useful when the data-generating process is unknown or difficult to model. This same remark applies to the CDM and VLM tasks.

Sample-efficient zero-shot classification. Combining Theorem 2.10 with Proposition 2.4 provides an end-to-end theory for the performance of zero-shot classification using the classifier

$\widehat{\mathbb{P}}_S^{(M)}(\cdot|\mathbf{x}_{\text{im}})$, as defined in Eq. (2.7). Here $S = \widehat{S}_{\text{NN}}^{\theta}$ is the similarity score corresponding to the empirical risk minimizer given in Eq. (2.17).

Corollary 2.11. *Suppose that Assumption 2.2, 2.4 and 2.5 hold. Let $\widehat{\theta}$ be the empirical risk minimizer defined in Eq. (2.17), and let $\widehat{\mathbb{P}}_S^{(M)}(\cdot|\mathbf{x}_{\text{im}})$ be the zero-shot classifier as defined in Eq. (2.7) with $S = \widehat{S}_{\text{NN}}^{\theta}$. Then, with probability at least $1 - \eta$,*

$$\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\mathbb{P}_{\text{cls}|\text{im}}(y|\mathbf{x}_{\text{im}}) \middle| \middle| \widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}}) \right) \right] \leq \widetilde{\mathcal{O}} \left(\sqrt{\frac{S^2 L^{11} \overline{m}^2}{n}} + \frac{\log(2/\eta)}{M} \right),$$

where $\widetilde{\mathcal{O}}$ hides polynomial factors in $(\log(\overline{m} S L n B_\psi), (B_\psi)^{\overline{m}})$.

The proof of Corollary 2.11 is provided in Appendix A.5.2.

2.4.2 Sample-efficient learning of CDMs

In this section, we investigate the conditional denoising models (CDMs) within the JGHM. Consider the joint distribution of noisy image, clean image, and text $(\mathbf{z}_t, \mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$, generated as follows: $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mu_\star$, and $\mathbf{z}_t = t \cdot \mathbf{x}_{\text{im}} + \sqrt{t} \cdot \mathbf{g}$, where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{\text{im}}})$ represents independent Gaussian noise. We denote the joint distribution of $(\mathbf{z}_t, \mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$ by $\mu_{\star, t}$.

Suppose we are given a dataset of iid samples $\{(\mathbf{z}_t^{(i)}, \mathbf{x}_{\text{im}}^{(i)}, \mathbf{x}_{\text{tx}}^{(i)})\}_{i \in [n]} \sim_{\text{iid}} \mu_{\star, t}$. With a text representation $\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}) \in \mathbb{R}^p$ (e.g., a CLIP-based embedding), the goal is to learn a conditional denoiser $\mathbf{M}_t(\mathbf{z}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))$ that closely approximates the clean image $\mathbf{x}_{\text{im}}$. Under an appropriate loss function, the optimal denoiser is the Bayes denoiser $\mathbf{m}_{\star, t}(\mathbf{z}_t, \mathbf{x}_{\text{tx}}) = \mathbb{E}_{(\mathbf{z}_t, \mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mu_{\star, t}} [\mathbf{x}_{\text{im}} | \mathbf{z}_t, \mathbf{x}_{\text{tx}}]$, which computes the posterior expectation of $\mathbf{x}_{\text{im}}$ given $(\mathbf{z}_t, \mathbf{x}_{\text{tx}})$. This section aims to analyze the sample complexity of learning this conditional denoiser using empirical risk minimization over the class of transformers.

The neural network architecture. The conditional denoiser is modeled as

$$\mathbf{M}_t^\theta(\mathbf{z}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) = \text{read}_{\text{cdm}} \circ \text{TF}_{\text{cdm}} \circ \text{Emb}_{\text{cdm}}(\mathbf{z}_t, \text{Adap}(\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))),$$

where each component is defined as follows. The function $\text{read}_{\text{cdm}} : \mathbb{R}^{D \times d_{\text{im}}} \rightarrow \mathbb{R}^{d_{\text{im}}}$ extracts the final denoised image, whereas the embedding function $\text{Emb}_{\text{cdm}} : \mathbb{R}^{d_{\text{im}}} \times \mathbb{R}^S \rightarrow \mathbb{R}^{D \times d_{\text{im}}}$ maps the input features into a transformer-compatible embedding, with specific details provided in Appendix A.6.1. The text encoder $\mathbf{E}_{\text{tx}} : \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}^S$ is given by the pre-trained CLIP representations, as defined in Eq. (2.15) and (2.17).

The transformer $\text{TF}_{\text{cdm}} : \mathbb{R}^{D \times d_{\text{im}}} \rightarrow \mathbb{R}^{D \times d_{\text{im}}}$ is a trainable $(2L + 1)$ -layer model, defined in Definition 2.2, with parameter $\mathbf{W}_{\text{cdm}}$ adapted to the $(2L + 1)$ -layer structure, as detailed in Eq. (2.16). The adapter network, $\text{Adap} : \mathbb{R}^S \mapsto \mathbb{R}^S$ is implemented as a simple network:

$$\text{Adap}(\mathbf{v}) := W_{\text{ada}}^{(1)} \text{softmax}(\log(\text{trun}(W_{\text{ada}}^{(2)} \text{softmax}(\mathbf{v}))))), \quad \forall \mathbf{v} \in \mathbb{R}^S, \quad (2.20)$$

where $W_{\text{ada}}^{(1)} \in \mathbb{R}^{S \times M}$ and $W_{\text{ada}}^{(2)} \in \mathbb{R}^{M \times S}$ are trainable weights. This adapter network structure leverages the canonical representation of the GHM framework, as described in Example 2.7. We

note that using an adapter network on top of CLIP representations is consistent with practice in prior work (Radford et al., 2021; Esser et al., 2024). Following the pre-training fine-tuning paradigm, we consider the fine-tuning phase where the parameters $\theta = (\mathbf{W}_{\text{cdm}}, W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)})$ are optimized, while read_{cdm} , Emb_{cdm} , and the CLIP encoder $\mathbf{E}_{\text{tx}}$ remain fixed.

The ERM estimator. Given a pre-trained text encoder $\mathbf{E}_{\text{tx}} : \mathcal{X}_{\text{tx}} \mapsto \mathbb{R}^S$, the goal is to obtain the conditional denoising function. To achieve this, we solve the empirical risk minimization problem defined by the following objective:

$$\hat{\theta} = \underset{\theta \in \Theta_{L,J,D,D',B,M}}{\text{argmin}} \left\{ \hat{\mathbf{R}}_{\text{cdm},t}(\mathbf{M}_t^\theta, \mathbf{E}_{\text{tx}}) := \frac{1}{n} \sum_{i=1}^n \left\| \mathbf{x}_{\text{im}}^{(i)} - \mathbf{M}_t^\theta(\mathbf{z}_t^{(i)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(i)})) \right\|_2^2 \right\}, \quad (2.21)$$

where the parameter space is defined as

$$\Theta_{L,J,D,D',B,M} := \left\{ \mathbf{W}_{\text{cdm}} \text{ as defined in Eq. (2.16)}, W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)} \text{ as defined in Eq. (2.20)}; \right. \quad (2.22)$$

$$\left. \|\theta\| := \|W_{\text{ada}}^{(1)}\|_{\text{op}} \vee \|W_{\text{ada}}^{(2)}\|_{\text{op}} \vee \max_{i \in [J+1], \ell \in [2L+1]} \left\{ \|W_{i,\text{cdm}}^{(\ell)}\|_{\text{op}}, \|W_{Q,\text{cdm}}^{(\ell)}\|_{\text{op}}, \|W_{K,\text{cdm}}^{(\ell)}\|_{\text{op}}, \|W_{V,\text{cdm}}^{(\ell)}\|_{\text{op}} \right\} \leq B \right\}.$$

The following theorem provides an estimation error bound on the conditional denoiser:

Theorem 2.12 (Estimation error of conditional denoising function). *Suppose that Assumption 2.4 and Assumption 2.5 hold, and assume Assumption 2.3 (a) holds for the image representation $\mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}}) = [\mathbb{P}(\mathbf{s}|\mathbf{x}_{\text{im}})/\mathbb{P}(\mathbf{s})]_{\mathbf{s} \in \mathcal{S}} \in \mathbb{R}^S$ where $\mathcal{S}$ is the set of root nodes. Let $\mathbf{E}_{\text{tx}}$ and $\mathcal{S}$ be obtained from the CLIP minimization. For simplicity, assume $t = 1$. Let $\Theta_{L,J,D,D',B,M}$ be the set defined in Eq. (2.22), where $J = \tilde{\mathcal{O}}(L)$, $D = \mathcal{O}(SL)$, $D' = \tilde{\mathcal{O}}(\bar{m}SL^3)$, and $B = \tilde{\mathcal{O}}(L_B + (SL + \bar{m}^2)\sqrt{M})$. Let $\hat{\theta}$ be the empirical risk minimizer defined in Eq. (2.21). Then, with probability at least $1 - 1/n$, we have*

$$\begin{aligned} & \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, \mathbf{z}_t)} \left[\frac{1}{d_{\text{im}}} \left\| \mathbf{m}_{\star,t}(\mathbf{z}_t, \mathbf{x}_{\text{tx}}) - \mathbf{M}_t^{\hat{\theta}}(\mathbf{z}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) \right\|_2^2 \right] \\ & \leq \tilde{\mathcal{O}} \left(\sqrt{\frac{(SL^8 \bar{m}^2 + M) S^5 L^3}{n}} + S^7 L_B^2 \left(\text{Suff}(\mathcal{S}) + \frac{1}{M} \right) \right), \end{aligned} \quad (2.23)$$

where $\tilde{\mathcal{O}}$ hides polynomial factors in $(\log(\bar{m}SLL_B n), (B_\psi)^{\bar{m}})$.

See the proof of Theorem 2.12 in Appendix A.6. The main step in the proof involves constructing transformers to approximate the conditional denoiser, similar to Theorem 2.10.

The estimation error bound has two terms. The first term, which scales as $n^{-1/2}$, comes from the approximation and generalization errors during the training of the conditional denoising function with the CLIP text representation fixed. The second term, which scales with $(\text{Suff}(\mathcal{S}) + M^{-1})$, is caused by the near-sufficiency of the CLIP representation. The term

$\text{Suff}(\mathcal{S})$ can be controlled by the excess risk of CLIP training, as shown in Theorem 2.10, while the term M^{-1} decreases as we increase the width of the adapter network. If the conditional denoising function TF_{cdm} and the CLIP text representation $\mathbf{E}_{\text{tx}}$ are jointly trained (eliminating the need for Adap), the second term vanishes, as shown in Appendix A.6.3.

By integrating over t and using Girsanov’s theorem, this estimation error bound can be converted into a sampling error bound for diffusion sampling, as illustrated in Corollary 2.6. Additionally, we perform similar analyses for the sampling error of vision-language models in Appendix A.3.2.

2.5 Experiments

We conduct experiments using transformer architectures to train CLIP encoders and downstream tasks for image-text distribution under JGHMs.

Training data distribution. We sample the image and text data from the JGHM described in Section 2.4 with parameters $L = 4$, $\mathcal{S} = [10]$, and $m_{\square}^{(\ell)} = 3$ for $\square \in \{\text{im}, \text{tx}\}$ and all ℓ , following the factorization assumption (Assumption 2.4). The transition probabilities $(\psi_r^{(0)}, \{\psi_{\text{im}, \ell}^{(\ell)}, \psi_{\text{tx}, \ell}^{(\ell)}\}_{\ell \in [\mathcal{S}], \ell \in [L]})$ are randomly generated from a specific distribution using a fixed random seed (details provided in Appendix A.9.1). These probabilities are governed by the parameter $p_{\text{flip}} \in [0, 1]$, which controls the conditional entropy of the leaf nodes $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$ given the root node x_r . When $p_{\text{flip}} = 0$, $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$ are deterministic functions of x_r , while $p_{\text{flip}} = 1$ results in high conditional entropy for $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$ given x_r . As p_{flip} increases, predicting x_r from $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$ becomes progressively more challenging. In our experiments, the values of p_{flip} are chosen from the range 0.02 to 0.4 in increments of 0.02.

Training setup. The CLIP encoders and conditional denoising functions are implemented using encoder transformers, while conditional next-token prediction functions (VLMs) are parameterized by decoder transformers. Detailed architectural specifications are provided in Appendix A.9.1. For training the CLIP encoders, we consider three setups: (1) **Standard TF**: A 5-layer transformer trained using the standard CLIP loss. (2) **Guided TF**: A 5-layer transformer trained with the CLIP loss, supplemented by a guided loss encouraging the model to emulate the belief propagation algorithm (details in Appendix A.9.1). (3) **Shallow TF**: A 1-layer transformer trained using the standard CLIP loss.

For CDMs and VLMs, we consider the following setups: (1) **Standard TF**: The CLIP encoder trained under the **Standard TF** setup is fixed, and a 9-layer transformer is trained on top of it using a standard supervised loss. (2) **Shallow TF**: The CLIP encoder trained under the **Standard TF** setup is fixed, and a 1-layer transformer is trained on top of it with a standard supervised loss. (3) **Joint Training**: Jointly train the CLIP encoder and the conditional denoiser/next-token predictor with a standard supervised loss. (4) **Guided TF**: Jointly train the CLIP encoder and the conditional denoiser/next-token predictor with a supervised loss augmented by a guided loss. (5) **Bayes**: The Bayes-optimal predictor.

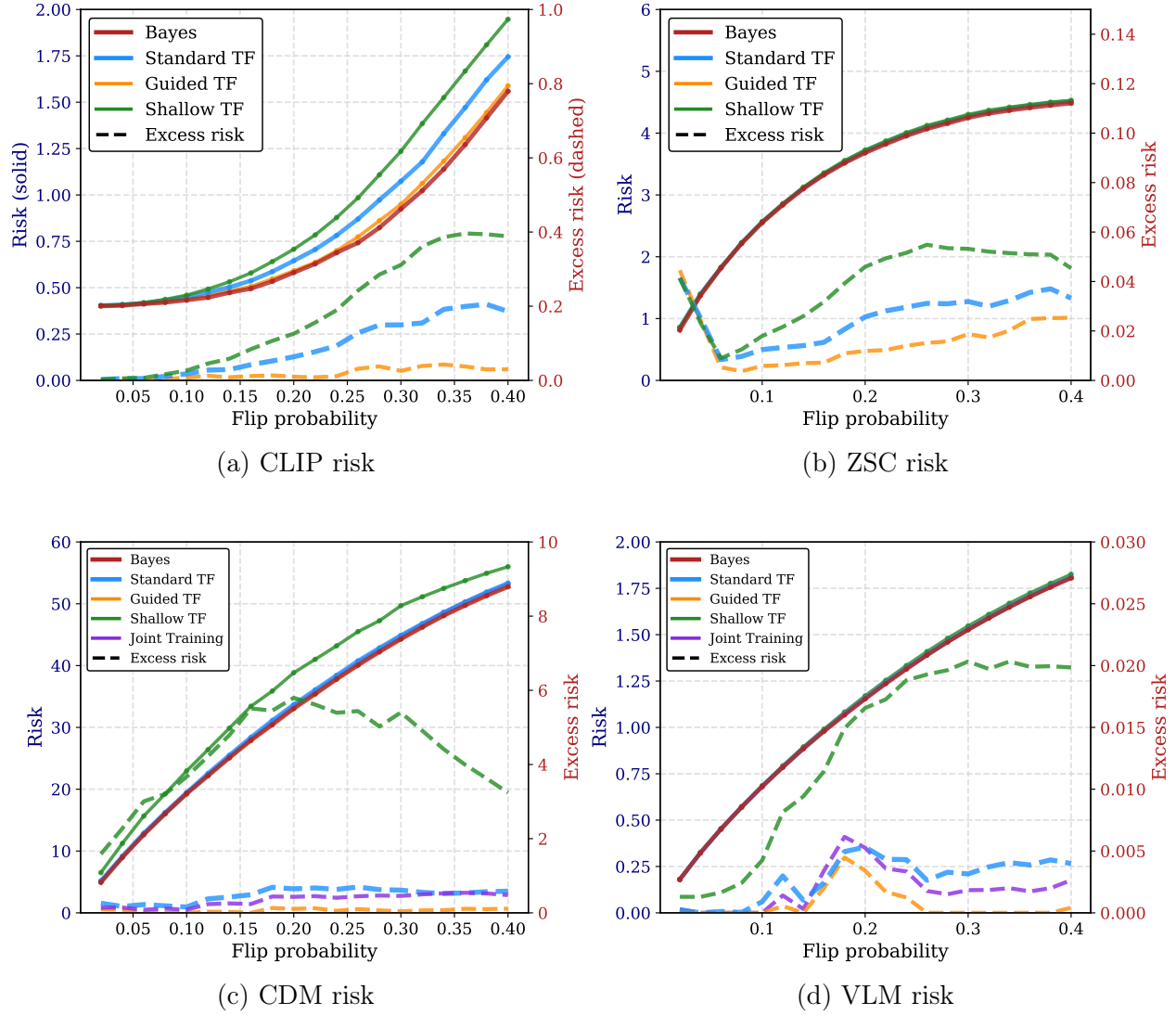


Figure 2.2: Risks (solid curves) and excess risks (dashed curves) as a function of the parameter p_{flip} for CLIP training, ZSC, CDM, and VLM. The training setups for the different curves are described in Section 2.5. Across all setups, the excess risks of Guided TF approach zero. The risks of Standard TF are close to the Bayes risk in ZSC, CDM, and VLM, demonstrating that CLIP representations can effectively adapt to downstream tasks.

All models are trained using AdamW for 30,000 steps, with each step using a fresh batch of size 128. Details on network architectures (which could be different from architectures used in theorems), learning rates, and other hyperparameters are provided in Appendix A.9.1.

Experimental results. Figure 2.2 shows the risk (solid curve) and excess risk (dashed curve) as functions of the parameter p_{flip} across different setups: CLIP training (Figure 2.2a), ZSC (Figure 2.2b), CDM (Figure 2.2c), and VLM (Figure 2.2d).

Standard training of CLIP (Figure 2.2a) exhibits a non-vanishing excess risk, likely due to the training dynamics failing to converge to a global minimizer of the CLIP loss. Despite this excess risk in CLIP training, standard training (**Standard TF**) results in small excess risks in ZSC, CDM, and VLM tasks (Figures 2.2b to 2.2d). This suggests that CLIP representations can effectively adapt to these downstream tasks, supporting our theoretical results, even when the conditions of our theory are not fully satisfied.

Guided training (**Guided TF**) for CLIP (Figure 2.2a) significantly reduces excess risk to nearly zero, in line with our approximation theory. In the ZSC, CDM, and VLM setups, **Guided TF** outperforms **Standard TF** by a considerable margin, indicating that guided training promotes better convergence to the global minimizer of the CLIP loss. Across all settings, **Standard TF** consistently outperforms **Shallow TF** by a wide margin, as expected. This suggests that shallow networks are insufficient for approximating the Bayes predictor, which relies on the belief propagation algorithm. In the CDM and VLM setups (Figures 2.2c and 2.2d), both sequential training (**Standard TF**) and joint training (**Joint Training**) yield small excess risks, indicating that CLIP pre-training may not always be necessary in this simulated environment. Further ablation studies are presented in Appendix A.9.2.

2.6 Conclusion

This chapter presents a theoretical framework explaining the success of contrastive pre-training in multi-modal generative AI. It shows that near-minimizers of contrastive loss serve as approximate sufficient statistics, adaptable to diverse tasks like zero-shot classification and conditional diffusion models. The Joint Generative Hierarchical Model (JGHM) illustrates how transformers efficiently approximate functions via belief propagation, breaking the curse of dimensionality. These findings provide guarantees on the sample efficiency and generalization of contrastive pre-training, validated by numerical simulations.

Approximate sufficient statistics are central to this framework, providing a foundation for understanding contrastive pre-training. Future research could examine how this concept extends to other learning paradigms. Another promising avenue is exploring single-modal contrastive learning frameworks, where data augmentations serve as positive samples. Additionally, extending the JGHM to model more realistic generative processes for image and text distributions holds significant potential for further advancements.

Chapter 3

Scaling Laws in Linear Regression: Compute, Parameters, and Data

3.1 Introduction

Deep learning models, particularly those on a large scale, are pivotal in advancing the state-of-the-art across various fields. Recent empirical studies have shed light on the so-called *neural scaling laws* (see [Kaplan et al., 2020](#); [Hoffmann et al., 2022](#), for example), which suggest that the generalization performance of these models improves polynomially as both model size, denoted by M , and data size, denoted by N , increase. The neural scaling law quantitatively describes the population risk as:

$$\mathcal{R}(M, N) \approx \mathcal{R}^* + \frac{c_1}{M^{a_1}} + \frac{c_2}{N^{a_2}}, \quad (3.1)$$

where $\mathcal{R}^*$ is a positive irreducible risk and c_1, c_2, a_1, a_2 are positive constants independent of M and N . For instance, by fitting the above formula with empirical measurements in standard large-scale language benchmarks, [Hoffmann et al. \(2022\)](#) estimated $a_1 \approx 0.34$ and $a_2 \approx 0.28$, while [Besiroglu et al. \(2024\)](#) estimated that $a_1 \approx 0.35$ and $a_2 \approx 0.37$. Though the exact exponents depend on the tasks, neural scaling laws in (3.1) are observed consistently in practice and are used as principled guidance to build state-of-the-art models, especially under a compute budget ([Hoffmann et al., 2022](#)).

From the perspective of statistical learning theory, (3.1) is rather intriguing. Standard statistical learning bounds (see [Mohri et al., 2018](#); [Wainwright, 2019a](#), for example) often decompose the population risk into the sum of irreducible error, approximation error, bias error, and variance error (some theory replaces bias and variance errors by optimization and generalization errors, respectively) as in the form of

$$\mathcal{R}(M, N) = \mathcal{R}^* + \underbrace{\mathcal{O}\left(\frac{1}{M^{a_1}}\right)}_{\text{approximation}} + \underbrace{\mathcal{O}\left(\frac{1}{N^{a_2}}\right)}_{\text{bias}} + \underbrace{\mathcal{O}\left(\frac{c(M)}{N^{a_3}}\right)}_{\text{variance}}, \quad (3.2)$$

where a_1, a_2, a_3 are positive constants and $c(M)$ is a measure of *model complexity* that typically increases with the model size M . In (3.2), the approximation error is induced by the mismatch of the best-in-class predictor and the best possible predictor, hence decreasing with the model size M . The bias error is induced by the mismatch of the expected algorithm output and the best-in-class predictor, hence decreasing with the data size N . The variance error measures the uncertainty of the algorithm output, which decreases with the data size N but increases with the model size M (since the model complexity $c(M)$ increases).

A mystery. The empirical neural scaling law (3.1) is incompatible with the typical statistical learning theory bound (3.2). While the two error terms in the neural scaling law (3.1) can be explained by the approximation and bias errors in the theoretical bound (3.2) respectively, it is not clear why *the variance error is unobservable when fitting the neural scaling law empirically*. This difference must be reconciled, otherwise, the statistical learning theory and the empirical scaling law make conflict predictions: as the model size M increases, the theoretical bound (3.2) predicts an increase of variance error that eventually causes an increase of the population risk, but the neural scaling law (3.1) predicts a decrease of the population risk. In other words, it remains unclear when to follow the prediction of the empirical scaling law (3.1) and when to follow that of the statistical learning bound (3.2).

Certain prior works provided risk upper bounds that do not grow with model size (see for example (Rudi and Rosasco, 2017; Carratino et al., 2018)). Still, their results are insufficient for studying scaling law as those bounds require a large model size such that the approximation error is ignorable. Moreover, they do not provide instance-wise matching lower bounds to verify the tightness of the upper bounds. See a detailed discussion in Section 3.2.

Our explanation. We investigate this issue in an infinite dimensional linear regression setup. We only assume access to M -dimensional sketched covariates given by a fixed Gaussian sketch and their responses. We consider a linear predictor with M trainable parameters, which is trained by one-pass *stochastic gradient descent* (SGD) with geometrically decaying stepsizes using N sketched data. Assuming that the spectrum of the data covariance matrix satisfies a power-law of degree $a > 1$ and that the optimal model parameters satisfy a Gaussian prior, we derive matching upper and lower bounds on the population risk achieved by the SGD output (see Theorem 3.1). Specifically, we show that

$$\mathcal{R}(M, N) = \mathcal{R}^* + \underbrace{\Theta\left(\frac{1}{M^{a-1}}\right) + \tilde{\Theta}\left(\frac{1}{(N\gamma)^{(a-1)/a}}\right)}_{\text{leading order given by the sum of Approx and Bias}}, \quad \text{Var} = \underbrace{\tilde{\Theta}\left(\frac{\min\{M, (N\gamma)^{1/a}\}}{N}\right)}_{\text{higher order, thus unobservable}},$$

where $\gamma = \mathcal{O}(1)$ is the initial stepsize used in SGD and $\tilde{\Theta}(\cdot)$ hides $\log(N)$ factors. In our bound, the sum of the approximation and bias errors determines the order of the excess risk, while the variance error is of a strictly higher order and is therefore nearly unobservable when fitting $\mathcal{R}(M, N)$ as a function of M and N empirically. In addition, our analysis reveals that the small variance error is due to the implicit regularization effect of one-pass SGD (Zou et al., 2021). Our theory suggests that the empirical neural scaling law (3.1) is a simplification of

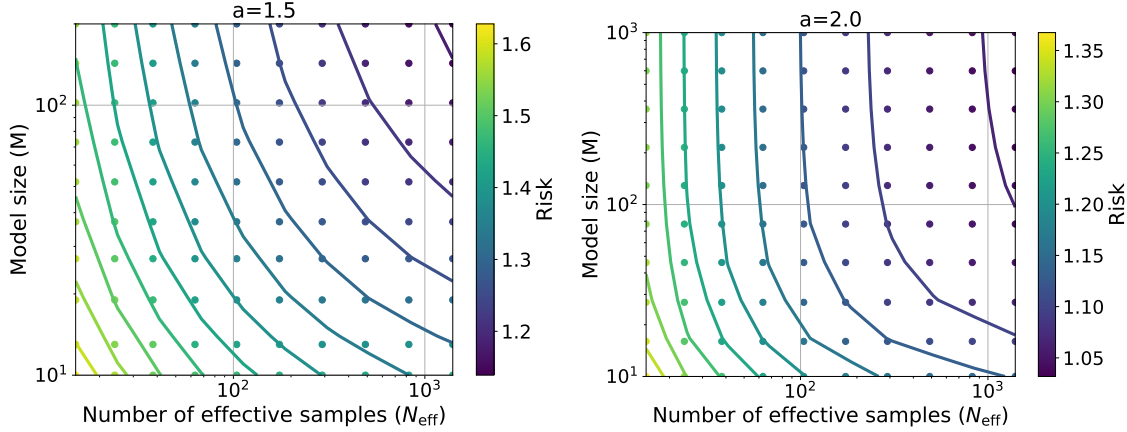


Figure 3.1: The expected risk (Risk) of the last iterate of (SGD) versus the effective sample size N_{eff} and the model size M for different power-law degrees a . The expected risk is computed by averaging over 1000 independent samples of $(\mathbf{w}^*, \mathbf{S})$. We fit the expected risk using the formula $\text{Risk} \sim \sigma^2 + c_1/M^{a_1} + c_2/N^{a_2}$ via minimizing the Huber loss as in Hoffmann et al. (2022). Parameters: $\sigma = 1, \gamma = 0.1$. Left: For $a = 1.5, d = 20000$, the fitted exponents are $(a_1, a_2) = (0.54, 0.34) \approx (0.5, 0.33)$. Right: For $a = 2, d = 2000$, the fitted exponents are $(a_1, a_2) = (1.07, 0.49) \approx (1.0, 0.5)$. Note that the values of (a_1, a_2) are close to our theoretical predictions $(a - 1, 1 - 1/a)$ in both cases, verifying the sharpness of our risk bounds. More details can be found in Sections 3.4 and 3.5.

the statistical learning bound (3.2) in a special regime when strong regularization (either implicit or explicit) is employed.

Moreover, we generalize the above scaling law to (1) constant stepsize SGD with iterate average (see Theorem B.23), (2) cases where the optimal model parameter satisfies an anisotropic prior (see Theorem 3.2), and (3) where the spectrum of the data covariance matrix satisfies a logarithmic power law (see Theorem 3.3).

Empirical evidence. Based on our theoretical results, we conjecture that the clean neural scaling law (3.1) observed in practice is due to the disappearance of variance error caused by strong regularization. Two pieces of empirical evidence to support our understanding. First, large language models that follow the scaling law (3.1) are often *underfitted*, as the models are trained over a single pass or a few passes over the data (Komatsuzaki, 2019; Muennighoff et al., 2023; Brown et al., 2020; Touvron et al., 2023). When models are underfitted, the variance error tends to be smaller. Second, when language models are trained with multiple passes (up to 7 passes), Muennighoff et al. (2023) found that the clean scaling law in (3.1) no longer holds and they proposed a more sophisticated scaling law to explain their data. This can be explained by a relatively large variance error caused by multiple passes.

Notation. For two positive-valued functions $f(x)$ and $g(x)$, we write $f(x) \lesssim g(x)$ (and $f(x) = \mathcal{O}(g(x))$) or $f(x) \gtrsim g(x)$ (and $f(x) = \Omega(g(x))$) if $f(x) \leq cg(x)$ or $f(x) \geq cg(x)$ holds for some absolute (if not otherwise specified) constant $c > 0$ respectively. We write $f(x) \approx g(x)$ (and $f(x) = \Theta(g(x))$) if $f(x) \lesssim g(x) \lesssim f(x)$. For two vectors $\mathbf{u}$ and $\mathbf{v}$ in a

Hilbert space, we denote their inner product by $\langle \mathbf{u}, \mathbf{v} \rangle$ or $\mathbf{u}^\top \mathbf{v}$. For two matrices $\mathbf{A}$ and $\mathbf{B}$ of appropriate dimensions, we define their inner product by $\langle \mathbf{A}, \mathbf{B} \rangle := \text{tr}(\mathbf{A}^\top \mathbf{B})$. We use $\|\cdot\|$ to denote the operator norm for matrices and ℓ_2 -norm for vectors. For a positive semi-definite (PSD) matrix $\mathbf{A}$ and a vector $\mathbf{v}$ of appropriate dimension, we write $\|\mathbf{v}\|_{\mathbf{A}}^2 := \mathbf{v}^\top \mathbf{A} \mathbf{v}$. For a symmetric matrix $\mathbf{A}$, we use $\mu_j(\mathbf{A})$ to refer to the j -th eigenvalue of $\mathbf{A}$ and $r(\mathbf{A})$ to refer to its rank. Finally, $\log(\cdot)$ refers to logarithm base 2.

3.2 Related work

Empirical scaling laws. In recent years, the scaling laws of deep neural networks in compute, sample size, and model size have been widely studied across different models and domains (Hestness et al., 2017; Rosenfeld et al., 2019; Kaplan et al., 2020; Henighan et al., 2020; Hoffmann et al., 2022; Zhai et al., 2022; Muennighoff et al., 2023). The early work by Kaplan et al. (2020) first proposed the neural scaling laws of transformer-based models. They observed that the test loss exhibits a power-law decay in quantities including the amount of compute, sample size, and model size, and provided joint formulas in these quantities to predict the test loss. The proposed formulas were later generalized and refined in subsequent works (Henighan et al., 2020; Hoffmann et al., 2022; Alabdulmohsin et al., 2022; Caballero et al., 2022; Muennighoff et al., 2023). Notably, Hoffmann et al. (2022) proposed the Chinchilla law, that is, (3.1) with $a_1 \approx 0.34$ and $a_2 \approx 0.28$. The empirical observation guided them to allocate data and model size under a given compute budget. The Chinchilla law is further revised by Besiroglu et al. (2024). Motivated by the Chinchilla law, Muennighoff et al. (2023) considered the effect of multiple passes over training data and empirically fitted a more sophisticated scaling law that takes account of the effect of data reusing.

Theory of scaling laws. Although neural scaling laws have been empirically observed over a broad spectrum of problems, there is a relatively limited literature on understanding these scaling laws from a theoretical perspective (Sharma and Kaplan, 2020; Bahri et al., 2024; Maloney et al., 2022; Hutter, 2021; Wei et al., 2022a; Michaud et al., 2024; Jain et al., 2024; Bordelon et al., 2024; Atanasov et al., 2024; Nam et al., 2024; Dohmatob et al., 2024). Among these works, Sharma and Kaplan (2020) showed that the test loss scales as $N^{4/d}$ for regression on data with intrinsic dimension d . Hutter (2021) studied a toy problem under which a non-trivial power of N arises in the test loss. Jain et al. (2024) considered scaling laws in data selection. Bahri et al. (2024) considered a linear teacher-student model under a power-law spectrum assumption on the covariates, and they showed that the test loss of the ordinary least squares estimator decreases following a power law in sample size N (resp. model size M) when the model size M (resp. sample size N) is infinite. Bordelon et al. (2024) considered a linear random feature model and analyzed the test loss of the solution found by (batch) gradient flow. They focused on the bottleneck regimes where two of the quantities N, M, T (training steps) are infinite and showed that the risk has a power-law decay in the remaining quantity. The problem in Bahri et al. (2024); Bordelon et al. (2024)

can be viewed as a sketched linear regression model similar to ours. It should be noted that both Bahri et al. (2024) and Bordelon et al. (2024) only derived the dependence of population risk on one of the data size, model size, or training steps in the asymptotic regime where the remaining quantities go to infinity, and their derivations are based on statistical physics heuristics. In comparison, we prove matching (ignoring constant factors) upper and lower risk bounds jointly depending on the finite model size M and data size N .

Implicit regularization of SGD. One-pass SGD in linear regression has been extensively studied in both the classical finite-dimensional setting (Polyak and Juditsky, 1992; Bach and Moulines, 2013; Défossez and Bach, 2015; Dieuleveut et al., 2017; Jain et al., 2018, 2017; Ge et al., 2019) and the modern high-dimensional setting (Dieuleveut and Bach, 2015; Berthier et al., 2020; Zou et al., 2023, 2021; Wu et al., 2022a,b; Varre et al., 2021). In particular, Zou et al. (2021) showed that SGD induces an implicit regularization effect that is comparable to, and in certain cases even more preferable than, the explicit regularization effect induced by ridge regression. This is one of the key motivations of our scaling law interpretation. From a technical perspective, we utilize the sharp finite-sample and dimension-free analysis of SGD developed by Zou et al. (2023); Wu et al. (2022a,b). Different from them, we consider a sequence of linear regression models with an increasing number of trainable parameters given by data sketch. Our main technical innovation is to sharply control the effect of data sketch. Some of our intermediate results, for example, tight bounds on the spectrum of the sketched data covariance under the power law (see Lemma 3.5), might be of independent interest.

Prior works investigated linear regression with random features (Rudi and Rosasco, 2017; Carratino et al., 2018), which can be viewed as a kind of sketched features via random coordinate selection. They mainly focused on the small approximation error regime, where the model size (or the number of features) is much larger than the data size. In comparison, we treat both model size and data size as free variables. Moreover, we provide matching upper and lower bounds while prior works mainly focused on upper bounds. These two innovations are crucial for studying scaling laws that predict test error as a function of both model size and data size. Finally, in the comparable regimes with small or zero approximation error, our excess risk bounds recover the bounds in prior works (Rudi and Rosasco, 2017; Carratino et al., 2018; Pillaud-Vivien et al., 2018; Dieuleveut and Bach, 2015; Defilippis et al., 2024).

3.3 Setup

We denote a feature vector by $\mathbf{x} \in \mathbb{H}$, where $\mathbb{H}$ is a Hilbert space that is either finite-dimensional or countably infinite-dimensional, with dimension $d := \dim(\mathbb{H})$. The corresponding label is denoted by $y \in \mathbb{R}$. In linear regression, we measure the population risk of a parameter $\mathbf{w} \in \mathbb{H}$ by the mean squared error,

$$\mathcal{R}(\mathbf{w}) := \mathbb{E}(\langle \mathbf{x}, \mathbf{w} \rangle - y)^2, \quad \mathbf{w} \in \mathbb{H},$$

where the expectation is over $(\mathbf{x}, y) \sim P$ for some distribution P on $\mathbb{H} \times \mathbb{R}$.

Definition 3.1 (Data covariance and optimal parameter). Let $\mathbf{H} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ be the data covariance. Assume that $\text{tr}(\mathbf{H})$ and all entries of $\mathbf{H}$ are finite. Let $(\lambda_i)_{i \geq 0}$ be the eigenvalues of $\mathbf{H}$ sorted in non-increasing order. Let $\mathbf{w}^* \in \arg \min_{\mathbf{w}} \mathcal{R}(\mathbf{w})$ be the optimal model parameter¹. Assume that $\|\mathbf{w}^*\|_{\mathbf{H}}^2 := (\mathbf{w}^*)^\top \mathbf{H} \mathbf{w}^*$ is finite.

We only assume access to M -dimensional sketched covariates and their responses, that is, $(\mathbf{S}\mathbf{x}, y)$, where $\mathbf{S} \in \mathbb{R}^M \times \mathbb{H}$ is a fixed *sketch* matrix. We focus on the Gaussian sketch matrix², that is, entries of $\mathbf{S}$ are independently sampled from $\mathcal{N}(0, 1/M)$. We then consider linear predictors with M trainable parameters given by

$$f_{\mathbf{v}} : \mathbb{H} \rightarrow \mathbb{R}, \quad \mathbf{x} \mapsto \langle \mathbf{v}, \mathbf{S}\mathbf{x} \rangle,$$

where $\mathbf{v} \in \mathbb{R}^M$ are the trainable parameters. Varying M should be viewed as a linear analog of varying the neural network model size. Our sketched linear regression setting is comparable to the teacher-student setting considered by Bahri et al. (2024); Bordelon et al. (2024).

We consider the training of $f_{\mathbf{v}}$ via one-pass *stochastic gradient descent* (SGD), that is,

$$\begin{aligned} \mathbf{v}_t &:= \mathbf{v}_{t-1} - \gamma_t (f_{\mathbf{v}_{t-1}}(\mathbf{x}_t) - y_t) \nabla_{\mathbf{v}} f_{\mathbf{v}_{t-1}}(\mathbf{x}_t) \\ &:= \mathbf{v}_{t-1} - \gamma_t (\mathbf{x}_t^\top \mathbf{S}^\top \mathbf{v}_{t-1} - y_t) \mathbf{S} \mathbf{x}_t, \quad t = 1, \dots, N, \end{aligned} \tag{SGD}$$

where $(\mathbf{x}_t, y_t)_{t=1}^N$ are independent samples from P and $(\gamma_t)_{t=1}^N$ are the stepsizes. We consider a popular geometric decaying stepsize scheduler (Ge et al., 2019; Wu et al., 2022a),

$$\text{for } t = 1, \dots, N, \quad \gamma_t := \gamma/2^\ell, \text{ where } \ell = \lfloor t/(N/\log(N)) \rfloor. \tag{3.3}$$

Here, the initial stepsize γ is a hyperparameter for the SGD algorithm. Without loss of generality, we assume the initial parameter is $\mathbf{v}_0 = 0$. The output of the SGD algorithm is the last iterate $\mathbf{v}_N$. Our proof techniques apply to other stepsize schedulers (e.g., polynomial decay) as well, but we focus on geometric decay as it is known to achieve near minimax-optimal excess risk for the last iterate of SGD (Ge et al., 2019).

Conditioning on a sketch matrix $\mathbf{S} \in \mathbb{R}^M \times \mathbb{H}$, each parameter $\mathbf{v} \in \mathbb{R}^M$ induces a sketched predictor through $\mathbf{x} \mapsto \langle \mathbf{S}^\top \mathbf{v}, \mathbf{x} \rangle$, and we denote its risk by

$$\mathcal{R}_M(\mathbf{v}) := \mathcal{R}(\mathbf{S}^\top \mathbf{v}) = \mathbb{E}(\langle \mathbf{S}\mathbf{x}, \mathbf{v} \rangle - y)^2, \quad \mathbf{v} \in \mathbb{R}^M.$$

By increasing M and N , we have a sequence of datasets and trainable parameters of increasing sizes, respectively. This prepares us to study the scaling law (3.1) in the sketched linear regression problem, that is, to understand $\mathcal{R}_M(\mathbf{v}_N)$ as a function of both M and N .

¹If $\arg \min \mathcal{R}(\cdot)$ is not unique, we choose $\mathbf{w}^*$ to be the minimizer with minimal $\mathbf{H}$ -norm.

²Our results can be extended to other sketching methods (see Woodruff et al., 2014, for example).

Risk decomposition. In a standard way, we decompose the risk achieved by $\mathbf{v}_N$, the last iterate of (SGD), to the sum of *irreducible risk*, *approximation error*, and *excess risk* as follows,

$$\mathcal{R}_M(\mathbf{v}_N) = \underbrace{\min \mathcal{R}(\cdot)}_{\text{Irreducible}} + \underbrace{\min \mathcal{R}_M(\cdot) - \min \mathcal{R}(\cdot)}_{\text{Approx}} + \underbrace{\mathcal{R}_M(\mathbf{v}_N) - \min \mathcal{R}_M(\cdot)}_{\text{Excess}}. \quad (3.4)$$

We emphasize that the irreducible risk is independent of M and N and thus can be viewed as a constant; the approximation error is determined by the sketch matrix $\mathbf{S}$, thus depends on M but is independent of N ; the excess risk depends on both M and N as it is determined by the algorithm.

3.4 Scaling laws

We first demonstrate a scaling-law behavior when the data spectrum satisfies a power law.

Assumption 3.1 (Distributional conditions). *Assume the following about the data distribution.*

- A. **Gaussian design.** Assume that $\mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$.
- B. **Well-specified model.** Assume that $\mathbb{E}[y|\mathbf{x}] = \mathbf{x}^\top \mathbf{w}^*$. Define $\sigma^2 := \mathbb{E}(y - \mathbf{x}^\top \mathbf{w}^*)^2$.
- C. **Parameter prior.** Assume that $\mathbf{w}^*$ satisfies a prior such that $\mathbb{E}(\mathbf{w}^*)^{\otimes 2} = \mathbf{I}$.

Assumption 3.2 (Power-law spectrum). *There exists a $a > 1$ such that the eigenvalues of $\mathbf{H}$ satisfy $\lambda_i \approx i^{-a}$, $i > 0$.*

Theorem 3.1 (Scaling law). *Suppose that Assumptions 3.1 and 3.2 hold. Consider an M -dimensional sketched predictor trained by (SGD) with N samples. Let $N_{\text{eff}} := N/\log(N)$ and recall the risk decomposition in (3.4). Then there exists some a -dependent constant $c > 0$ such that when the initial stepsize $\gamma \leq c$, with probability at least $1 - e^{-\Omega(M)}$ over the randomness of the sketch matrix $\mathbf{S}$, we have*

1. Irreducible $:= \mathcal{R}(\mathbf{w}^*) = \sigma^2$.
2. $\mathbb{E}_{\mathbf{w}^*} \text{Approx} \approx M^{1-a}$.
3. Suppose in addition $\sigma^2 \gtrsim 1$. The expected excess risk (Excess) can be decomposed into a bias error (Bias) and a variance error (Var), namely,

$$\mathbb{E} \text{Excess} \approx \text{Bias} + \sigma^2 \text{Var},$$

where the expectation is over the randomness of $\mathbf{w}^*$ and $(\mathbf{x}_i, y_i)_{i=1}^N$. Moreover, Bias and Var satisfy

$$\text{Bias} \lesssim \max \{ M^{1-a}, (N_{\text{eff}} \gamma)^{1/a-1} \},$$

$$\begin{aligned} \text{Bias} &\gtrsim (N_{\text{eff}}\gamma)^{1/a-1} \text{ when } (N_{\text{eff}}\gamma)^{1/a} \leq M/c \text{ for some constant } c > 0, \\ \text{Var} &\asymp \min \{M, (N_{\text{eff}}\gamma)^{1/a}\} / N_{\text{eff}}. \end{aligned}$$

In all results, the hidden constants only depend on the power-law degree a . As a direct consequence, when $\sigma^2 \asymp 1$, it holds with probability at least $1 - e^{-\Omega(M)}$ over the randomness of the sketch matrix $\mathbf{S}$ that

$$\mathbb{E}\mathcal{R}_M(\mathbf{v}_N) = \sigma^2 + \Theta\left(\frac{1}{M^{a-1}}\right) + \Theta\left(\frac{1}{(N_{\text{eff}}\gamma)^{(a-1)/a}}\right),$$

where the expectation is over the randomness of $\mathbf{w}^*$ and $(\mathbf{x}_i, y_i)_{i=1}^N$.

Theorem 3.1 shows a sharp (up to constant factors) scaling law risk bound under an isotropic prior assumption and the power-law spectrum assumption. We emphasize that the scaling law bound in Theorem 3.1 holds for every $M, N \geq 1$. We also remark that the sum of approximation and bias errors dominates $\mathbb{E}\mathcal{R}_M(\mathbf{v}_N) - \sigma^2$, whereas the variance error is of strict higher order in terms of both M and N , and is thus disappeared in the population risk bound.

Optimal stepsize. Based on the tight scaling law in Theorem 3.1, we can calculate the optimal stepsize that minimizes the risk. Specifically, the optimal stepsize is $\gamma \asymp 1$ when $N_{\text{eff}} \lesssim M^a$ and can be anything such that $M^a/N_{\text{eff}} \lesssim \gamma \lesssim 1$ when $N_{\text{eff}} \gtrsim M^a$. In both cases, choosing $\gamma \asymp 1$ is optimal. When the sample size is large such that $N_{\text{eff}} \gtrsim M^a$, the optimal stepsize is relatively robust and can be chosen from a range.

Allocation of data and model sizes. Following Hoffmann et al. (2022), we measure the compute complexity by MN as (SGD) queries M -dimensional gradients for N times. Given a total compute budget of $MN = C$, from Theorem 3.4 and $N_{\text{eff}} := N/\log(N)$, we see that the best population risk is achieved by setting $\gamma = \Theta(1)$, $M = \tilde{\Theta}(C^{1/(a+1)})$, and $N = \tilde{\Theta}(C^{a/(a+1)})$. Our theory suggests setting a data size slightly larger than the model size when the compute budget is the bottleneck.

Comparison with Bordelon et al. (2024). The work by Bordelon et al. (2024) considered the scaling law of batch gradient descent (or gradient flow) on a teacher-student model (see their equation (14)). Their teacher-student model can be viewed as our sketched linear regression model. However, we consider one-pass SGD, therefore in our setting the number of gradient steps is equivalent to the data size. When we equalize the number of gradient steps and the data size in their equation (14) and set the parameter prior as Assumption 3.1C, their prediction is consistent with ours. However, our analysis shows the computational advantage of SGD over batch GD since each iteration requires only $1/N$ the compute. Bordelon et al. (2024) obtained the limit of the population risk as two out of the data size, model size, and the number of gradient steps go to infinity based on statistical physics heuristics. In comparison, we obtain upper and lower risk bounds that hold for any finite M and N and match ignoring a constant factor depending only on the spectrum power-law degree a .

Average of the SGD iterates Results similar to Theorem 3.1 can also be established for the average of the iterates of online SGD with constant stepsize (Polyak and Juditsky, 1992; Dieuleveut et al., 2017; Jain et al., 2018, 2017; Zou et al., 2023). All results will be the same once replacing the effective sample size N_{eff} in Theorem 3.1 to the sample size N . For more details see Theorem B.23 in Appendix B.6.

3.4.1 Scaling law under source condition

The isotropic parameter prior condition (Assumption 3.1C) in Theorem 3.1 can be generalized to the following anisotropic version (Caponnetto and De Vito, 2007).

Assumption 3.3 (Source condition). *Let $(\lambda_i, \mathbf{v}_i)_{i>0}$ be the eigenvalues and eigenvectors of $\mathbf{H}$. Assume $\mathbf{w}^*$ satisfies a prior such that*

$$\text{for } i \neq j, \quad \mathbb{E}\langle \mathbf{v}_i, \mathbf{w}^* \rangle \langle \mathbf{v}_j, \mathbf{w}^* \rangle = 0; \quad \text{and for } i > 0, \quad \mathbb{E}\lambda_i \langle \mathbf{v}_i, \mathbf{w}^* \rangle^2 \asymp i^{-b}, \quad \text{for some } b > 1.$$

A larger exponent b implies a faster decay of signal $\mathbf{w}^*$ and thus corresponds to a simpler task (Caponnetto and De Vito, 2007). Note that Assumption 3.1C satisfies Assumption 3.3 with $b = a$.

Theorem 3.2 (Scaling law under source condition). *In Theorem 3.1, suppose Assumption 3.1C is replaced by Assumption 3.3 with $1 < b < a + 1$. Then there exists some a -dependent constant $c > 0$ such that when $\gamma \leq c$, with probability at least $1 - e^{-\Omega(M)}$ over the randomness of the sketch matrix $\mathbf{S}$, we have*

$$\mathbb{E}\mathcal{R}_M(\mathbf{v}_N) = \sigma^2 + \underbrace{\Theta\left(\frac{1}{M^{b-1}}\right) + \Theta\left(\frac{1}{(N_{\text{eff}}\gamma)^{(b-1)/a}}\right)}_{\text{Approx+Bias}} + \underbrace{\Theta\left(\frac{\min\{M, (N_{\text{eff}}\gamma)^{1/a}\}}{N_{\text{eff}}}\right)}_{\text{Var}}.$$

where the expectation is over the randomness of $\mathbf{w}^*$ and $(\mathbf{x}_i, y_i)_{i=1}^N$, and $\Theta(\cdot)$ hides constants that may depend on (a, b) .

When $1 < b \leq a$, the tasks are relatively hard (compared to when $b = a$), and the variance error is dominated by the sum of approximation and bias errors for all choices of M , N , and $\gamma \lesssim 1$. In this case, Theorem 3.2 gives the same prediction about optimal stepsize and optimal allocation of data and model sizes under compute budget as Theorem 3.1.

When $a < b < a + 1$, the tasks are relatively easy (compared to when $b = a$), and variance remains dominated by the sum of approximation and bias error if the stepsize is optimally tuned. Recall that $\gamma \lesssim 1$, thus we can rewrite the risk bound in Theorem 3.2 as

$$\mathbb{E}\mathcal{R}_M(\mathbf{v}_N) - \sigma^2 \asymp \frac{1}{\min\{M, (N_{\text{eff}}\gamma)^{1/a}\}^{b-1}} + \frac{\min\{M, (N_{\text{eff}}\gamma)^{1/a}\}}{N_{\text{eff}}}$$

$$\approx \begin{cases} \min \{M, (N_{\text{eff}}\gamma)^{1/a}\} / N_{\text{eff}} & M \gtrsim N_{\text{eff}}^{1/b} \text{ and } N_{\text{eff}}^{a/b-1} \lesssim \gamma \lesssim 1, \\ \min \{M, (N_{\text{eff}}\gamma)^{1/a}\}^{1-b} & M \lesssim N_{\text{eff}}^{1/b} \text{ or } \gamma \lesssim N_{\text{eff}}^{a/b-1}. \end{cases}$$

Therefore the optimal stepsize and the risk under the optimal stepsize is

$$\gamma \approx N_{\text{eff}}^{a/b-1} \text{ if } M \gtrsim N_{\text{eff}}^{1/b}, \quad \text{and } M^a / N_{\text{eff}} \lesssim \gamma \lesssim 1 \text{ if } M \lesssim N_{\text{eff}}^{1/b}.$$

Under the *optimally tuned* stepsize, the population risk is in the form of

$$\min_{\gamma} \mathbb{E} \mathcal{R}_M(\mathbf{v}_N) = \sigma^2 + \Theta(N_{\text{eff}}^{(1-b)/b}) + \Theta(M^{1-b}),$$

which is again in the scaling law form (3.1). This is expected since an optimally tuned stepsize controls the variance error by adjusting the strength of the implicit bias of SGD. Under a fixed compute budget $C = MN$, our theory suggests to assign $M = \tilde{\Theta}(C^{1/(b+1)})$ and $N = \tilde{\Theta}(C^{b/(b+1)})$, and set the stepsize to $\gamma \approx \tilde{\Theta}(C^{(a-b)/(b+1)})$.

When $b \geq a + 1$, the tasks are even simpler. We provide upper and lower bounds in Appendix B.4.3. However, there exists a gap between the bounds, fixing which is left for future work.

Moreover, we note that in the comparable regimes where M is large, the results in Theorem 3.2 match existing bounds on the risk of SGD iterates and ridge estimators (Pillaud-Vivien et al., 2018; Rudi and Rosasco, 2017).

3.4.2 Scaling law under logarithmic power law

We also derive the risk formula when the data covariance has a logarithmic power-law spectrum (Bartlett et al., 2020).

Assumption 3.4 (Logarithmic power-law spectrum). *There exists $a > 1$ such that the eigenvalues of $\mathbf{H}$ satisfy $\lambda_i \approx i^{-1} \log^{-a}(i+1)$, $i > 0$.*

Theorem 3.3 (Scaling law under logarithmic power spectrum). *In Theorem 3.1, suppose Assumption 3.2 is replaced by Assumption 3.4. Then with probability at least $1 - e^{-\Omega(M)}$ over the randomness of the sketch matrix $\mathbf{S}$, we have*

$$\mathbb{E} \mathcal{R}_M(\mathbf{v}_N) = \sigma^2 + \Theta\left(\frac{1}{\log^{a-1}(M)}\right) + \Theta\left(\frac{1}{\log^{a-1}(N_{\text{eff}}\gamma)}\right), \quad \text{Var} \approx \frac{\min \left\{M, \frac{N_{\text{eff}}\gamma}{\log^a(N_{\text{eff}}\gamma)}\right\}}{N_{\text{eff}}},$$

where the expectation is over the randomness of $\mathbf{w}^*$ and $(\mathbf{x}_i, y_i)_{i=1}^N$.

Theorem 3.3 provides a scaling law under the logarithmic power-law spectrum. Similar to Theorem 3.1, the variance error is dominated by the approximation and bias errors for all choices of M , N , and γ , and thus disappears from the risk bound. Different from Theorem 3.1, here the population risk is a polynomial of $\log(M)$ and $\log(N_{\text{eff}}\gamma)$.

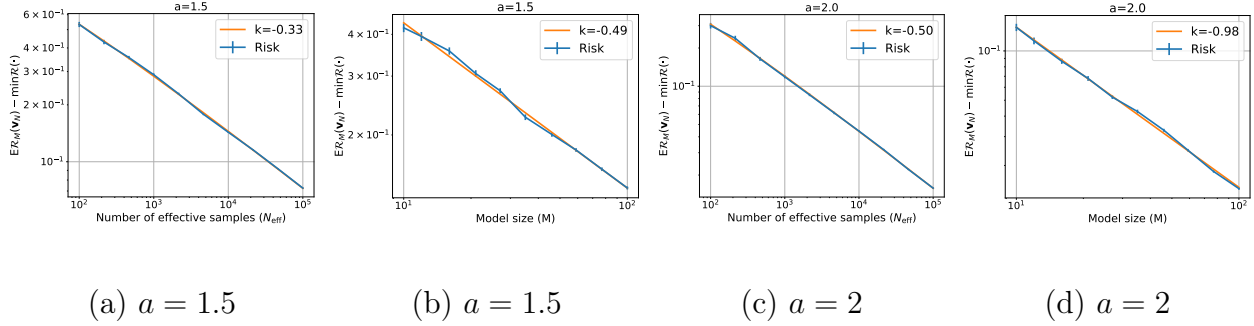


Figure 3.2: The expected risk of the last iterate of (SGD) minus the irreducible risk versus the effective sample size and model size. Parameters $\sigma = 1, \gamma = 0.1$. (a), (b): $a = 1.5, d = 10000$; (c), (d): $a = 2, d = 1000$. The error bars denote the ± 1 standard deviation of estimating the expected risk using 100 independent samples of $(\mathbf{w}^*, \mathbf{S})$. We use linear functions to fit the expected risk under the log-log scale and report the slope of the fitted lines (denoted by k).

3.5 Experiments

In this section, we examine the relation between the expected risk of the (SGD) output, the data size N , and the model size M when the covariates satisfy a power-law covariance spectrum. Although our results in Section 3.4 hold with high probability over $\mathbf{S}$, for simplicity, we assume the expectation of the risk is taken over both $\mathbf{w}^*$ and $\mathbf{S}$ in our simulations. We adopt the model in Section 3.3 and train it using one-pass (SGD) with geometric decaying stepsize (3.3). We choose the dimension d sufficiently large to approximate the infinite-dimensional case, and the data are generated so that Assumption 3.1 is satisfied. Moreover, we choose the covariance $\mathbf{H} \in \mathbb{R}^{d \times d}$ to be diagonal with $\mathbf{H}_{ii} \propto i^a$ and $\text{tr}(\mathbf{H}) = 1$ for some $a > 1$. From Figure 3.1, we observe that the risk indeed follows a power-law formula jointly in the number of samples and the number of parameters. In addition, the fitted exponents are aligned with our theoretical predictions $(a - 1, 1 - 1/a)$ in Theorem 3.1. Figure 3.2 shows the scaling of the expected risk in data size (or model size) when the model size (or data size) is relatively large. We see that the expected risk also satisfies a power-law decay with exponents matching our predictions. It is noteworthy that our simulations demonstrate stronger observations than the theoretical results in Theorem 3.1, which only establishes matching upper and lower bounds up to a constant factor. Additional simulation results on the risk of the average of (SGD) iterates can be found in Appendix B.6.

3.6 Risk bounds under a general spectrum

In this section, we present some general results on the upper and lower bounds of the risk of the output of (SGD). Due to the rotational invariance of the sketched matrix $\mathbf{S}$, without loss of generality, we assume the covariance $\mathbf{H} \in \mathbb{R}^{d \times d}$ is diagonal with non-increasing diagonal

entries. Our main results in Section 3.4 are directly built on the general bounds introduced here.

Assumption 3.5 (General distributional conditions). *Assume the following about the data distribution.*

*A. **Hypercontractivity.** There exists $\alpha \geq 1$ such that for every PSD matrix $\mathbf{A}$ it holds that*

$$\mathbb{E} \mathbf{x} \mathbf{x}^\top \mathbf{A} \mathbf{x} \mathbf{x}^\top \preceq \alpha \operatorname{tr}(\mathbf{H} \mathbf{A}) \mathbf{H}.$$

*B. **Misspecified model.** There exists $\sigma^2 > 0$ such that $\mathbb{E}(y - \mathbf{x}^\top \mathbf{w}^*)^2 \mathbf{x} \mathbf{x}^\top \preceq \sigma^2 \mathbf{H}$.*

It is clear that Assumption 3.1 implies Assumption 3.5 with $\alpha = 3$.

Excess risk decomposition. Conditioning on the sketch matrix $\mathbf{S}$, the training of the sketched linear predictor can be viewed as an M -dimensional linear regression problem. We can therefore invoke existing SGD analysis (Wu et al., 2022a,b) to sharply control the excess risk by controlling the bias and variance errors. Specifically, let us define the ($\mathbf{w}^*$ -dependent) bias error as

$$\text{Bias}(\mathbf{w}^*) := \left\| \prod_{t=1}^N (\mathbf{I} - \gamma_t \mathbf{S} \mathbf{H} \mathbf{S}^\top) \mathbf{v}^* \right\|_{\mathbf{S} \mathbf{H} \mathbf{S}^\top}^2, \quad \text{where } \mathbf{v}^* := (\mathbf{S} \mathbf{H} \mathbf{S}^\top)^{-1} \mathbf{S} \mathbf{H} \mathbf{w}^*, \quad (3.5)$$

and the variance error as

$$\text{Var} := \frac{\#\{\tilde{\lambda}_j \geq 1/(N_{\text{eff}} \gamma)\} + (N_{\text{eff}} \gamma)^2 \sum_{\tilde{\lambda}_j < 1/(N_{\text{eff}} \gamma)} \tilde{\lambda}_j^2}{N_{\text{eff}}}, \quad N_{\text{eff}} := N / \log(N), \quad (3.6)$$

where $(\tilde{\lambda}_j)_{j=1}^M$ are eigenvalues of $\mathbf{S} \mathbf{H} \mathbf{S}^\top$. We also let $\text{Bias} := \mathbb{E} \text{Bias}(\mathbf{w}^*)$, where the expectation is over the prior of $\mathbf{w}^*$. Using the existing results on the output of (SGD) in Wu et al. (2022a,b), we show that the excess risk in (3.4) can be exactly decomposed as the sum of bias and variance errors under weak conditions.

Theorem 3.4 (Excess risk decomposition). *Conditioning on the sketch matrix $\mathbf{S}$, consider the excess risk in (3.4) induced by the output of (SGD). Assume $\mathbf{v}_0 = 0$. Then for any $\mathbf{w}^* \in \mathbb{H}$,*

1. *Under Assumptions 3.5A and 3.5B and suppose $\gamma \leq 1/(c\alpha \operatorname{tr}(\mathbf{S} \mathbf{H} \mathbf{S}^\top))$ for some constant $c > 0$, we have*

$$\mathbb{E} \text{Excess} \lesssim \text{Bias}(\mathbf{w}^*) + (\alpha \|\mathbf{w}^*\|_{\mathbf{H}}^2 + \sigma^2) \text{Var}.$$

2. *Under the stronger Assumptions 3.1A and 3.1B and suppose $\gamma \leq 1/(c\alpha \operatorname{tr}(\mathbf{S} \mathbf{H} \mathbf{S}^\top))$ for some constant $c > 0$, we have*

$$\mathbb{E} \text{Excess} \gtrsim \text{Bias}(\mathbf{w}^*) + \sigma^2 \text{Var}.$$

In both results, the expectations of **Excess** are taken over $(\mathbf{x}_t, y_t)_{t=1}^N$.

Assuming that the signal-to-noise ratio is upper bounded, that is, $\|\mathbf{w}^*\|_{\mathbf{H}}^2/\sigma^2 \lesssim 1$, then the bias-variance decomposition of the excess risk is sharp up to constant factors.

The variance error is in a nice form and can be computed using the following important lemma on the spectrum of $\mathbf{SHS}^\top$. Similar results for logarithmic power-law are also established in Lemma B.29 in Appendix B.7.

Lemma 3.5 (Power law). *Under Assumption 3.2, it holds with probability at least $1 - e^{-\Omega(M)}$ that*

$$\mu_j(\mathbf{SHS}^\top) \approx \mu_j(\mathbf{H}) \approx j^{-a}, \quad j = 1, \dots, M.$$

For any $0 \leq k^* \leq k^\dagger \leq \infty$, let $\mathbf{S}_{k^*:k^\dagger} \in \mathbb{R}^{M \times (k^\dagger - k^*)}$ denote the matrix formed by the $k^* + 1 - k^\dagger$ -th columns of $\mathbf{S}$. We also abuse the notation $k^\dagger : \infty$ for $k^\dagger : d$ when d is finite. We let $\mathbf{H}_{k^*:k^\dagger} \in \mathbb{R}^{(k^\dagger - k^*) \times (k^\dagger - k^*)}$ be the submatrix of $\mathbf{H}$ formed by the $k^* + 1 - k^\dagger$ -th eigenvalues. For the approximation and bias error, we use the following upper and lower bounds to compute their values.

Theorem 3.6 (A general upper bound). *Suppose Assumption 3.5 holds. Assume $\mathbf{v}_0 = 0$, $r(\mathbf{H}) \geq 2M$ and the initial stepsize satisfies $\gamma < 1/(c\alpha \text{tr}(\mathbf{SHS}^\top))$ for some constant $c > 0$. Then for any $k_1, k_2 \leq M/3$, with probability at least $1 - e^{-\Omega(M)}$*

$$\begin{aligned} \text{Approx} &\lesssim \|\mathbf{w}_{k_1:\infty}^*\|_{\mathbf{H}_{k_1:\infty}}^2 + \left(\frac{\sum_{i>k_1} \lambda_i}{M} + \lambda_{k_1+1} + \sqrt{\frac{\sum_{i>k_1} \lambda_i^2}{M}} \right) \|\mathbf{w}_{0:k_1}^*\|^2, \\ \text{Bias}(\mathbf{w}^*) &\lesssim \frac{\|\mathbf{w}_{0:k_2}^*\|_2^2}{N_{\text{eff}} \gamma} \cdot \left[\frac{\mu_{M/2}(\mathbf{S}_{k_2:\infty} \mathbf{H}_{k_2:\infty} \mathbf{S}_{k_2:\infty}^\top)}{\mu_M(\mathbf{S}_{k_2:\infty} \mathbf{H}_{k_2:\infty} \mathbf{S}_{k_2:\infty}^\top)} \right]^2 + \|\mathbf{w}_{k_2:\infty}^*\|_{\mathbf{H}_{k_2:\infty}}^2. \end{aligned}$$

Theorem 3.7 (A general lower bound). *Suppose Assumption 3.1 holds. Assume $\mathbf{v}_0 = 0$, $r(\mathbf{H}) \geq M$ and the initial stepsize $\gamma < 1/(c \text{tr}(\mathbf{SHS}^\top))$ for some constant $c > 0$. Then*

$$\mathbb{E}_{\mathbf{w}^*} \text{Approx} \gtrsim \sum_{i=M}^d \lambda_i, \quad \mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) \gtrsim \sum_{i:\tilde{\lambda}_i < 1/(N_{\text{eff}} \gamma)} \frac{\mu_i(\mathbf{SH}^2 \mathbf{S}^\top)}{\mu_i(\mathbf{SHS}^\top)}$$

almost surely, where $(\lambda_i)_{i=1}^d$ are eigenvalues of $\mathbf{H}$ in non-increasing order, $(\tilde{\lambda}_i)_{i=1}^M$ are eigenvalues of $\mathbf{SHS}^\top$ in non-increasing order.

3.7 Conclusion

We analyze neural scaling laws in infinite-dimensional linear regression. We consider a linear predictor with M trainable parameters on the sketched covariates, which is trained by one-pass stochastic gradient descent with N data. Under a Gaussian prior assumption on

the optimal model parameter and a power law (of degree $a > 1$) assumption on the spectrum of the data covariance, we derive matching upper and lower bounds on the population risk minus the irreducible error, that is, $\Theta(M^{-(a-1)} + N^{-(a-1)/a})$. In particular, we show that the variance error, which increases with M , is of strictly higher order compared to the other errors, thus disappearing from the risk bound. We attribute the nice empirical formula of the neural scaling law to the non-domination of the variance error, which ultimately is an effect of the implicit regularization of SGD.

Many directions remain open for future study. First, our work is limited to the linear model; it would be interesting to see whether similar scaling laws can be derived in more complex models, such as random feature models or two-layer networks. Second, we focus on one-pass SGD training, and it is unclear if similar results hold for other optimization methods like accelerated SGD or Adam. Additionally, from a technical perspective, many results in our work depend on the Gaussian assumption and the source condition of the data. Investigating how these assumptions can be relaxed would also be valuable.

Chapter 4

Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning

4.1 Introduction

Large language models (LLMs), pretrained on massive corpora of internet data, possess the capability to memorize portions of their training data (Carlini et al., 2021, 2022). However, this capability raises significant concerns, as the training data may contain sensitive or private information, potentially leading to societal challenges. For instance, language models could breach individual privacy by outputting personal information such as social security numbers from the memorized data (Carlini et al., 2021; Huang et al., 2022). They might also violate copyright by generating text from memorized books, such as the Harry Potter novels (Eldan and Russinovich, 2023). Furthermore, LLM assistants for biology could inadvertently aid in the development of biological weapons by troubleshooting bottlenecks, increasing the risk of such attempts (Sandbrink, 2023; Li et al., 2024). In response to these concerns, regulations like the EU’s General Data Protection Regulation (GDPR) (Mantelero, 2013; Voigt and Von dem Bussche, 2017) and the US’s California Consumer Privacy Act (CCPA) (CCPA, 2018) have mandated *the Right to be Forgotten*, requiring applications to support the deletion of information contained in training samples upon user requests. This has motivated a line of research on *machine unlearning*, aiming to address these challenges.

Machine unlearning (Cao and Yang, 2015; Bourtole et al., 2021) aims to delete the influence of specific training samples from machine-learning models while preserving other knowledge and capabilities (Liu et al., 2024c; Zhang et al., 2023b; Nguyen et al., 2022; Xu et al., 2023; Si et al., 2023). Notably, a straightforward approach to unlearning is to retrain a language model from scratch. However, as retraining from scratch is typically computationally expensive, cheaper methods for removing undesirable information is highly desirable. Recently, several works (Jang et al., 2022; Wang et al., 2023; Chen and Yang,

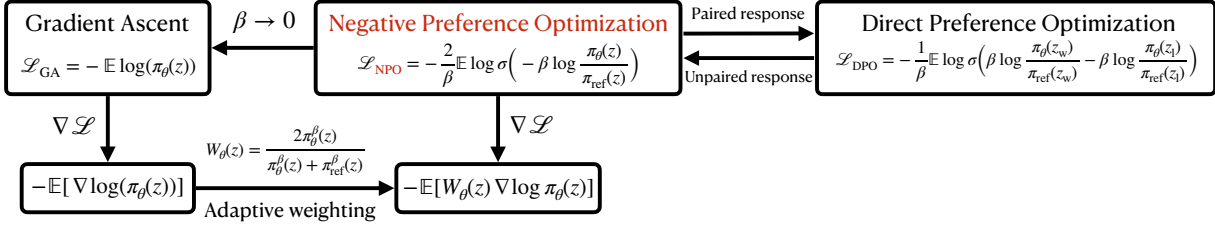


Figure 4.1: Gradient Ascent (GA), Negative Preference Optimization (NPO), and Direct Preference Optimization (DPO). NPO can be interpreted as DPO without positive samples. The gradient of NPO is an adaptive weighting of that of GA, and the weight vanishes for unlearned samples.

2023; Yao et al., 2023; Eldan and Russinovich, 2023; Yao et al., 2024; Liu et al., 2024d; Li et al., 2024) proposed scalable and practical techniques for unlearning LLMs through directly fine-tuning the trained model. Core to many of these works is a *gradient ascent* procedure on the prediction loss over the dataset to be unlearned (the *forget set*), building on the intuition that gradient ascent is an approximation of “reverting” gradient descent optimization on the forget set.

Despite its simplicity and widespread use, the performance of gradient ascent based approaches remain unsatisfactory. A notable example concerns the recently released benchmark dataset TOFU (Maini et al., 2024), which consists of synthetically generated biographies of 200 fictitious authors, and the task is to unlearn the biographies of 1%, 5%, and 10% of the 200 authors from a model that is already fine-tuned on all 200 authors. In their evaluation of forgetting 10% of the authors, Maini et al. (2024) demonstrated that gradient ascent and its variants fail to provide a satisfactory balance between forget quality (the difference between the unlearned model and retrained model evaluated on the forget set) and model utility (the general performance on other tasks).

In this work, we begin by observing that gradient ascent can often cause a rapid deterioration of model utility during training—a phenomenon we term *catastrophic collapse*—which we believe is responsible for its unsatisfactory performance. Towards fixing this, we propose a simple yet effective objective function for unlearning termed *Negative Preference Optimization* (NPO). NPO takes inspiration from preference optimization (Rafailov et al., 2024; Ouyang et al., 2022; Bai et al., 2022), and can be viewed as a variant of preference optimization that only uses negative samples. Through both theory and experiments, we show that NPO resolves the catastrophic collapse issue associated with gradient ascent, provides more stable training dynamics, and achieves a better trade-off between forgetting quality and model utility. Coupled with a cross-entropy loss on the retain set, NPO achieves state-of-the-art performance on the TOFU dataset, and achieves the first non-trivial unlearning result on the challenging task of forgetting 50% of the TOFU data.

Summary of contributions and chapter outline.

- We outline existing gradient ascent based methods for machine unlearning, and find that these methods suffer from *catastrophic collapse* (Section A.1). We identify the linear divergence speed of gradient ascent as a main reason for catastrophic collapse.
- We introduce Negative Preference Optimization (NPO), a simple alignment-inspired loss function for LLM unlearning that addresses the catastrophic collapse issue of gradient ascent (GA; Section 4.3). We demonstrate that NPO reduces to gradient ascent (GA) in the high-temperature limit. We show in theory the progression towards catastrophic collapse when minimizing the NPO loss is exponentially slower than with GA. See Figure 4.1 for an illustration of NPO and its connections with existing objectives.
- We test NPO-based methods on a synthetic binary classification task (Section 4.4), where we find that NPO-based methods outperform other baselines by providing a superior Pareto frontier between the Forget Distance and Retain Distance. Furthermore, NPO-based methods exhibit greater learning stability compared to GA-based methods.
- We evaluate a variety of unlearning methods on the TOFU dataset (Maini et al., 2024) and find that NPO-based methods exhibit superior balance between Forget Quality and Model Utility compared to all baselines (Section 4.5). Additionally, NPO-based methods improve the stability of the unlearning process and the readability of the output. Notably, we show that NPO-based methods are the only effective unlearning methods for forgetting 50%-90% of the data, a significant advance over all existing methods which already struggle with forgetting 10% of the data (Section 4.5.3).

4.1.1 Related work

There is a vast literature on machine unlearning and LLM unlearning. Since its proposal by Cao and Yang (2015), machine unlearning has been extensively studied in the classification literature (Bourtoule et al., 2021; Golatkar et al., 2020; Ginart et al., 2019; Thudi et al., 2022; Izzo et al., 2021; Koh and Liang, 2017; Guo et al., 2019; Sekhari et al., 2021). For reviews of existing works, see Liu et al. (2024c); Zhang et al. (2023b); Nguyen et al. (2022); Xu et al. (2023); Si et al. (2023). In particular, Ginart et al. (2019); Guo et al. (2019); Sekhari et al. (2021) introduced theoretical metrics for machine unlearning based on the notion of differential privacy and proposed provably efficient unlearning methods based on Newton update removal mechanisms. However, these algorithms require computing the Hessian of loss functions, which is intractable for LLMs.

Recent research has explored unlearning methods for LLMs (Jang et al., 2022; Wang et al., 2023; Chen and Yang, 2023; Yao et al., 2023; Eldan and Russinovich, 2023; Yao et al., 2024; Liu et al., 2024d; Li et al., 2024). Notably, the methods proposed in Jang et al. (2022); Yao et al. (2023); Chen and Yang (2023); Maini et al. (2024) are based on gradient ascent (GA) on the loss of the forget set. In this work, we demonstrate that the NPO approach

consistently outperforms GA across various tasks. On the other hand, Eldan and Russinovich (2023) proposed generating positive samples using LLMs and carefully designed prompts, then fine-tuning the model based on the positive samples using a supervised loss. Furthermore, the method of Liu et al. (2024d) is based on knowledge negation, while the approach of Li et al. (2024) relies on controlling model representations. These methods are orthogonal and complementary to the NPO approach.

Our method, NPO, draws inspiration from the framework of reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022; Stiennon et al., 2020; Rafailov et al., 2024), particularly the Direct Policy Optimization (DPO) method (Rafailov et al., 2024). We note that recent work (Ethayarajh et al., 2024) proposes the Kahneman-Tversky Optimization (KTO) method for alignment with only non-paired preference data, and a more recent concurrent work (Duan et al., 2024) proposes the Distributional Dispreference Optimization (D²O) approach for unlearning. Both methods share a similar formulation to NPO. We compare the performance of NPO with KTO in simulations.

Recent work has proposed several benchmark datasets and evaluation metrics for unlearning methods (Ji et al., 2024; Eldan and Russinovich, 2023; Maini et al., 2024; Li et al., 2024; Lynch et al., 2024). In particular, some studies have utilized the PKUSafe dataset (Ji et al., 2024) for benchmarking unlearning methods. Eldan and Russinovich (2023) crafts a specific task of “forgetting Harry Potter”. Maini et al. (2024) introduces TOFU, a task of fictitious unlearning for LLMs, which is the benchmark we adopted in this chapter. Additionally, Li et al. (2024) proposes the Weapons of Mass Destruction Proxy (WMDP) for measuring hazardous knowledge in LLMs. Lynch et al. (2024) proposes eight methods to evaluate robust unlearning in LLM, which incorporate robust metrics against jailbreak attacks.

Finally, we note the existence of attack methods for extracting data from unlearned models (Shi et al., 2023a; Patil et al., 2023), and other unlearning methods including model editing (Mitchell et al., 2022; Meng et al., 2022) and in-context unlearning (Pawelczyk et al., 2023).

4.2 Preliminaries on Machine Unlearning

Machine Unlearning refers to the following problem: Given an *initial model* (also the *reference model*) $\pi_{\text{ref}}(y|x)$ that is already trained on a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i \in [n]}$, how to make the model *forget* a specific subset (henceforth the *forget set*) $\mathcal{D}_{\text{FG}} \subseteq \mathcal{D}$ of the training data? More precisely, we aim to fine-tune¹ the model to make it behave like the *retrained model* π_{retr} , a model trained only on the *retain set* $\mathcal{D}_{\text{RT}} = \mathcal{D} \setminus \mathcal{D}_{\text{FG}}$. In other words, we would like the model to behave as if the samples in the forget set $\mathcal{D}_{\text{FG}}$ were never used to train it.

By definition, the best approach for machine unlearning in principle is to retrain the model from scratch on $\mathcal{D}_{\text{RT}}$ only which is however often intractable in practice.

¹There are alternative approaches such as prompt engineering (Pawelczyk et al., 2023) for performing unlearning tasks.

Gradient ascent is a key component in many existing LLM unlearning methods, and also an important baseline method for LLM unlearning on its own. The idea is simply to perform gradient ascent on the (next-token prediction) loss over the forget set, which can be viewed equivalently as gradient descent on the *negative* prediction loss, denoted as $\mathcal{L}_{\text{GA}}$:

$$\mathcal{L}_{\text{GA}}(\theta) = - \underbrace{\mathbb{E}_{\mathcal{D}_{\text{FG}}}[-\log(\pi_{\theta}(y|x))]}_{\text{prediction loss}} = \mathbb{E}_{\mathcal{D}_{\text{FG}}}[\log(\pi_{\theta}(y|x))]. \quad (4.1)$$

The rationale of gradient ascent is that since the initial model π_{ref} is trained on $\mathcal{D} = \mathcal{D}_{\text{FG}} \cup \mathcal{D}_{\text{RT}}$, a subsequent *maximization* of prediction loss on the forget set $\mathcal{D}_{\text{FG}}$ would approximately “revert” the optimization on the forget set $\mathcal{D}_{\text{FG}}$, thus unlearning $\mathcal{D}_{\text{FG}}$ and approximating a model trained on $\mathcal{D}_{\text{RT}}$ only.

Other loss functions. Building on gradient ascent, a large class of unlearning methods perform gradient-based optimization on a linear combination of the GA loss $\mathcal{L}_{\text{GA}}$ and several other loss functions that encourage unlearning (Jang et al., 2022; Yao et al., 2023; Chen and Yang, 2023; Maini et al., 2024; Eldan and Russinovich, 2023). Notable examples include

- Forget (FG) loss: $\mathcal{L}_{\text{FG}}(\theta) = -\mathbb{E}_{\mathcal{D}_{\text{FG}}}[\log(\pi_{\theta}(\tilde{y}|x))]$, where $(x, y) \sim \mathcal{D}_{\text{FG}}$ and $\tilde{y} \neq y$ is any “uninformed” response for prompt x which the unlearned model could aim to output. Examples of such $\tilde{y}$ ’s include replacing true information by random (but appearingly sensible) information (which requires hand-crafting such as Eldan and Russinovich (2023)), or simply answering “I don’t know” (Maini et al., 2024).
- Retain (RT) loss: $\mathcal{L}_{\text{RT}}(\theta) = -\mathbb{E}_{\mathcal{D}_{\text{RT}}}[\log(\pi_{\theta}(y|x))]$, which encourages the model to still perform well on the retain set $\mathcal{D}_{\text{RT}}$;
- $\mathcal{K}_{\text{FG}}(\theta) = \mathbb{E}_{\mathcal{D}_{\text{FG}}}[\text{D}(\pi_{\theta}(\cdot|x) || \pi_{\text{ref}}(\cdot|x))]$, which measures the distance to the initial model π_{ref} (in KL divergence) on the forget set;
- $\mathcal{K}_{\text{RT}}(\theta) = \mathbb{E}_{\mathcal{D}_{\text{RT}}}[\text{D}(\pi_{\theta}(\cdot|x) || \pi_{\text{ref}}(\cdot|x))]$, which measures the distance to the initial model π_{ref} (in KL divergence) on the retain set.

For example, Yao et al. (2023) minimize a combination of $\{\mathcal{L}_{\text{GA}}, \mathcal{L}_{\text{FG}}, \mathcal{K}_{\text{RT}}\}$, and Chen and Yang (2023) minimize a combination of $\{\mathcal{L}_{\text{GA}}, \mathcal{L}_{\text{RT}}, -\mathcal{K}_{\text{FG}}, \mathcal{K}_{\text{RT}}\}$. Maini et al. (2024) find that incorporating the retain loss $\mathcal{L}_{\text{RT}}$ improves the performance on various unlearning methods.

Forget quality and model utility. Unlearning methods should not only unlearn the forget set, i.e. achieve a high *forget quality*, but also maintain the model’s performance on the retain set, i.e. maintain the *model utility*. For example, letting the model to simply output “I don’t know” is a unlearning method that achieves good forget quality (in certain sense) but bad model utility. While there is not yet a consensus on the right metrics for forget quality and model utility (and we will present our choices momentarily), a general rule of thumb is that unlearning methods should achieve a good tradeoff between these two goals.

4.2.1 Catastrophic collapse of gradient ascent

We begin by testing gradient ascent as a standalone method (as opposed to combining it with other losses), and find that gradient ascent exhibits a common failure mode dubbed

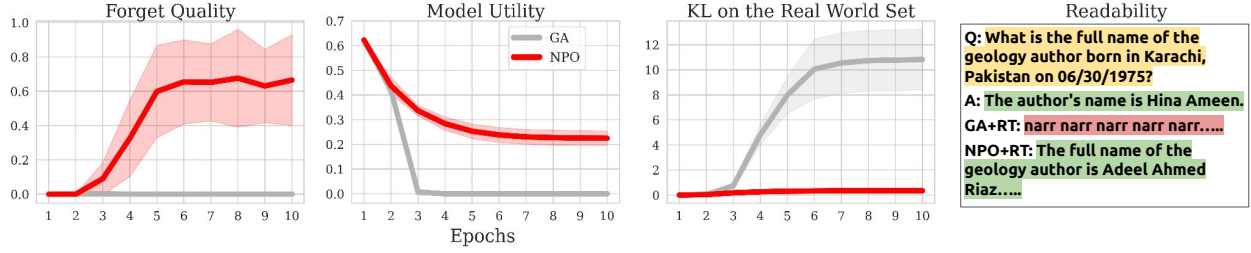


Figure 4.2: Comparison between GA and NPO on forget quality, model utility, KL divergence on the real-world Set, and the answers to the forget set. The rightmost figure shows the answers generated from variants of GA and NPO that incorporates the RT loss. All figures are generated on the Forget05 task in the TOFU data, trained for 10 epochs (detailed setup in Appendix C.5.1).

as *catastrophic collapse*: Along the unlearning process, the model utility quickly drops to zero, and the forget quality improves temporarily for a very short time horizon before quickly dropping too (Figure 4.2 left/middle-left). Along the same training trajectory, the model diverges quickly from the initial model (as measured by the KL distance to the initial model), after which the model generates gibberish outputs (Figure 4.2 middle-right/right).

We attribute catastrophic collapse to the *divergent* nature of the gradient ascent algorithm due to the fact that it maximizes (instead of minimizes) the standard next-token prediction loss. Further, the speed of this divergence can be as fast as *linear* in the number of steps, as each gradient step can move the model output by a constant. To see this on a toy example, consider a linear-logistic K -class classifier given by $\pi_\theta(\cdot|x) = \text{softmax}(\theta x)$, $\theta = (\theta_l)_{l \in [K]} \in \mathbb{R}^{d \times K}$. For any “already unlearned” sample (x_i, y_i) with true label $y_i = l \in [K]$ and model prediction $\text{softmax}(\theta x_i)_l \approx 0$ (so that π_θ does not predict l), standard calculation shows that the gradient of GA loss with respect to θ_l is $\nabla_{\theta_l} \mathcal{L}_{\text{GA}, i} = (1\{y_i = l\} - \text{softmax}(\theta x_i)_l)x_i \approx x_i$, which has a constant scale (not diminishing along the unlearning progress) and can cause the model to diverge in a linear speed. Therefore, the divergent dynamics may initially bring the model closer to π_{retr} but would ultimately send the model to infinity.

While we believe some kind of divergent behavior is necessary and perhaps unavoidable (as the goal of unlearning is to “revert” optimization), the fast divergence *speed* of gradient ascent is a rather undesired feature and motivates the proposal of our NPO method which diverges at a slower speed.

4.3 Negative Preference Optimization

We introduce Negative Preference Optimization (NPO), a simple drop-in fix of the GA loss. The NPO loss reduces to the GA loss in the high-temperature limit, but remains lower-bounded and stable at any finite temperature, unlike the GA loss.

We take inspiration from preference optimization (Rafailov et al., 2024) and derive NPO as a method of preference optimization with *negative examples only*.

Negative Preference Optimization. In preference optimization (Ouyang et al., 2022; Bai et al., 2022; Stiennon et al., 2020; Rafailov et al., 2024), we are given a dataset with preference feedbacks $\mathcal{D}_{\text{paired}} = \{(x_i, y_{i,w}, y_{i,l})\}_{i \in [n]}$, where $(y_{i,w}, y_{i,l})$ are two responses to x_i generated by a pre-trained model π_θ , and the preference $y_{i,w} \succ y_{i,l}$ is obtained by human comparison. The goal is to fine-tune π_θ using $\mathcal{D}_{\text{paired}}$ to better align it with human preferences. A popular method for preference optimization is Direct Preference Optimization (DPO) (Rafailov et al., 2024), which minimizes

$$\mathcal{L}_{\text{DPO},\beta}(\theta) = -\frac{1}{\beta} \mathbb{E}_{\mathcal{D}_{\text{paired}}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right]. \quad (4.2)$$

Here, $\sigma(t) = 1/(1 + e^{-t})$ is the sigmoid function, $\beta > 0$ is the inverse temperature, and π_{ref} is a reference model.

Unlearning as preference optimization. We observe that the unlearning problem can be cast into the preference optimization framework by treating each $(x_i, y_i) \in \mathcal{D}_{\text{FG}}$ as only providing a negative response $y_{i,l} = y_i$ without any positive response $y_{i,w}$. We then simply ignore the y_w term in DPO in Eq. (4.2) and obtain the Negative Preference Optimization (NPO) loss:

$$\mathcal{L}_{\text{NPO},\beta}(\theta) = -\frac{2}{\beta} \mathbb{E}_{\mathcal{D}_{\text{FG}}} \left[\log \sigma \left(-\beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \right) \right] = \frac{2}{\beta} \mathbb{E}_{\mathcal{D}_{\text{FG}}} \left[\log \left(1 + \left(\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \right)^\beta \right) \right]. \quad (4.3)$$

Minimizing $\mathcal{L}_{\text{NPO},\beta}$ ensures that the prediction probability on the forget set $\pi_\theta(y_i|x_i)$ is as small as possible, aligning with the goal of unlearning the forget set.

Connection with gradient ascent. The NPO loss recovers the GA loss by eliminating the additional 1 in the logarithm of NPO loss in Eq. (4.3), i.e., replacing $\log(1 + (\pi_\theta/\pi_{\text{ref}})^\beta)$ to $\log((\pi_\theta/\pi_{\text{ref}})^\beta)$. Furthermore, we show that the NPO loss also reduces to the GA loss in the limit of $\beta \rightarrow 0$, indicating that NPO is a strict generalization of GA.

Proposition 4.1 (NPO reduces to GA as $\beta \rightarrow 0$). *For any θ , we have*

$$\lim_{\beta \rightarrow 0} \left[\mathcal{L}_{\text{NPO},\beta}(\theta) - \frac{2}{\beta} \log 2 \right] = \mathcal{L}_{\text{GA}}(\theta) - \underbrace{\mathbb{E}_{\mathcal{D}_{\text{FG}}} [\log \pi_{\text{ref}}(y | x)]}_{\text{does not depend on } \theta}.$$

Moreover, assuming $\pi_\theta(y | x)$ is differentiable with respect to θ , we have

$$\lim_{\beta \rightarrow 0} \nabla_\theta \mathcal{L}_{\text{NPO},\beta}(\theta) = \nabla_\theta \mathcal{L}_{\text{GA}}(\theta).$$

The proof of Proposition 4.1 is deferred to Appendix C.1.1. Figure 4.3 provides an illustration of the reduction from the NPO loss to the GA loss as $\beta \rightarrow 0$.

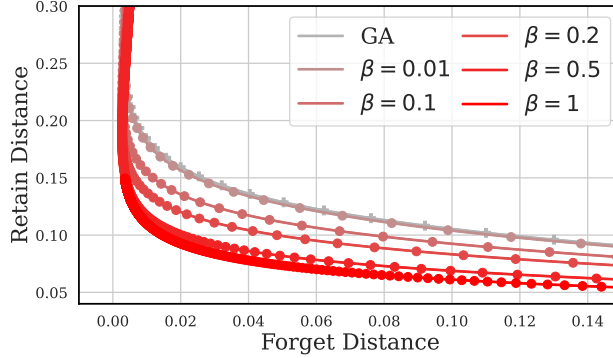


Figure 4.3: Retain distance versus forget distance for GA and NPO with varying levels of β in the binary classification experiment with $\alpha = 1$. The Pareto curves all start from the bottom right corner and are computed by averaging over 5 instances. We observe that the NPO trajectory converges to the GA trajectory as $\beta \rightarrow 0$. More details can be found in Section 4.4.

Stability of the NPO loss. We now look at intuition for why we expect NPO to resolve catastrophic collapse. One limitation of the GA loss is its unboundedness from below (as the negation of the cross-entropy prediction loss which is unbounded from above). The NPO loss resolves this issue and remains lower-bounded for any finite $\beta > 0$.

Furthermore, the gradients of NPO and GA are as follows:

$$\nabla_{\theta} \mathcal{L}_{\text{GA}} = -\mathbb{E}_{\mathcal{D}_{\text{FG}}} [\nabla_{\theta} \log \pi_{\theta}(y|x)], \quad (4.4)$$

$$\nabla_{\theta} \mathcal{L}_{\text{NPO}, \beta} = -\mathbb{E}_{\mathcal{D}_{\text{FG}}} [\mathbf{W}_{\theta}(x, y) \nabla_{\theta} \log \pi_{\theta}(y|x)], \quad (4.5)$$

where $\mathbf{W}_{\theta}(x, y) = 2\pi_{\theta}^{\beta}(y|x) / [\pi_{\theta}^{\beta}(y|x) + \pi_{\text{ref}}^{\beta}(y|x)]$ can be interpreted as an adaptive smoothing weight—When example $(x, y) \in \mathcal{D}_{\text{FG}}$ is already unlearned in the sense that $\pi_{\theta}(y|x) \ll \pi_{\text{ref}}(y|x)$, we have $\mathbf{W}_{\theta}(x, y) \ll 1$, so that $\|\nabla_{\theta} \mathcal{L}_{\text{NPO}, \beta}\|_2 \ll \|\nabla_{\theta} \mathcal{L}_{\text{GA}}\|_2$ and thus NPO could diverge much slower than GA.

4.3.1 Theoretical analysis of divergence speed

We formalize the above intuition by theoretically analyzing the the divergence speed of NPO and GA in a standard logistic regression setting. We consider a binary classification problem ($y \in \{0, 1\}$) with logistic model $\pi_{\theta}(y = 1|x) = \sigma(\langle x, \theta \rangle)$. The initial model is denoted as $\pi_{\theta_{\text{init}}}$ for $\theta_{\text{init}} \in \mathbb{R}^d$, and we unlearn the forget set $\mathcal{D}_{\text{FG}} = \{(x_i, y_i)\}_{i=1}^{n_f}$ by minimizing the GA and NPO loss using gradient descent with stepsize η for T iterations.

Theorem 4.2 (Divergence speed of GA and NPO). *Let $X := (x_1, \dots, x_{n_f})^{\top} \in \mathbb{R}^{n_f \times d}$. Consider the high-dimensional regime where $n_f \leq d$ and assume $XX^{\top}$ is invertible. Suppose $\|\theta_{\text{init}}\|_2 \leq B_{\theta}$, $\|x_i\|_2 \in [b_x, B_x]$ for all $i \in [n_f]$ for some $B_{\theta}, b_x, B_x > 0$. Let $\theta_{\text{GA}}^{(t)}, \theta_{\text{NPO}}^{(t)}$*

denote the t -th iterates of gradient descent with stepsize η on the empirical loss $\mathcal{L}_{\text{GA}}, \mathcal{L}_{\text{NPO}, \beta}$, respectively.

- **(GA *diverges linearly*)** There exist some (B_θ, b_x, B_x) -dependent constants $C_0, C_1, C_2 > 0$ such that when $\max_{i \neq j} |\langle x_i, x_j \rangle| \leq C_0/n_f$,

$$\|\theta_{\text{GA}}^{(t)} - \theta_{\text{init}}\|_{X^\top X} \in \left[C_1 \cdot n_f^{-1/2} \eta \cdot t, C_2 \cdot n_f^{-1/2} \eta \cdot t \right], \quad t \geq 1.$$

- **(NPO *diverges logarithmically*)** Suppose $\eta \leq 1$. There exist some $(B_\theta, b_x, B_x, \beta)$ -dependent constants $C_0, C_1, C_2, C_3 > 0$ such that when $\max_{i \neq j} |\langle x_i, x_j \rangle| \leq C_0/n_f$,

$$\|\theta_{\text{NPO}}^{(t)} - \theta_{\text{init}}\|_{X^\top X} \in \left[C_1 \sqrt{n_f} \log \left(C_2 \cdot \frac{\eta}{n_f} \cdot t + 1 \right), C_1 \sqrt{n_f} \log \left(C_3 \cdot \frac{\eta}{n_f} \cdot t + 1 \right) \right], \forall t \geq 1.$$

Theorem 4.2 demonstrates that NPO diverges exponentially slower than GA in this simple setting. The proof of Theorem 4.2 is contained in Appendix C.1.2.

4.4 Synthetic Experiments

4.4.1 Setup

Dataset. We consider a forget set $\mathcal{D}_{\text{FG}} = \{(x_i^f, y_i^f)\}_{i=1}^{200}$ and a retain set $\mathcal{D}_{\text{RT}} = \{(x_i^r, y_i^r)\}_{i=1}^{1000}$, which are both generated from Gaussian-logistic models. More specifically, we assume

$$\begin{aligned} x_i^f &\sim_{iid} \mathcal{N}(\mu_f, \mathbf{I}_d), & \mathbb{P}(y_i^f = 1 | x_i^f) &= \sigma((x_i^f - \mu_f)^\top \theta_f + 1), \\ x_i^r &\sim_{iid} \mathcal{N}(\mu_r, \mathbf{I}_d), & \mathbb{P}(y_i^r = 1 | x_i^r) &= \sigma((x_i^r - \mu_r)^\top \theta_r - 1). \end{aligned} \tag{4.6}$$

Here we choose $d = 16$, $\theta_f = -\theta_r = \mathbf{1}_d/\sqrt{d}$, and $\mu_f = -\mu_r = \alpha \cdot \mathbf{1}_d$ for some $\alpha \geq 0$. We consider two choices of the hyper-parameter α : (1). $\alpha = 1$, which creates a gap between the Gaussian means of forget covariates $\{x_i^f\}$ and retain covariates $\{x_i^r\}$; (2). $\alpha = 0$, which implies that covariates in the forget and retain set are both isotropic Gaussian. We remark that we shift by 1 in the sigmoid function to create a discrepancy in the label frequencies between the forget and retain sets — this ensures that the forget labels y_i^f are more likely to be 1, while the retain labels y_i^r are more likely to be 0.

Model and training method. We consider a random feature model $\pi_\theta(y = 1|x) = \sigma(\theta^\top \text{ReLU}(Wx))$, where $W \in \mathbb{R}^{128 \times d}$ is fixed during the training and unlearning process, whose entries are generated i.i.d. from $\mathcal{N}(0, 1/d)$, and $\theta \in \mathbb{R}^{128}$ is the trainable parameter. To generate the initial model π_{ref} and the retrained model π_{retr} , we optimize over θ using the cross-entropy loss over the entire dataset $\mathcal{D} = \mathcal{D}_{\text{FG}} \cup \mathcal{D}_{\text{RT}}$ and the retain dataset $\mathcal{D}_{\text{RT}}$, respectively. In the unlearning phase, starting from the initial model π_{ref} , we perform gradient descent on various loss functions for 2000 steps. We select the learning rate for each method via grid search.

Unlearning methods. We evaluate the performance of vanilla NPO (NPO; minimizing $\mathcal{L}_{\text{NPO}}$), NPO plus a retain loss term (NPO+RT; minimizing $\mathcal{L}_{\text{NPO}} + \mathcal{L}_{\text{RT}}$), gradient ascent (GA; minimizing $\mathcal{L}_{\text{GA}}$), gradient ascent plus a retain loss term (GA+RT; minimizing $\mathcal{L}_{\text{GA}} + \mathcal{L}_{\text{RT}}$), cross-entropy loss of forget and retain sets where the positive labels of the forget set are given by Bern(0.5) (IDK+RT; minimizing $\mathcal{L}_{\text{FG}} + \mathcal{L}_{\text{RT}}$), and DPO plus a retain loss term (DPO+RT; minimizing $\mathcal{L}_{\text{DPO}} + \mathcal{L}_{\text{RT}}$, where the positive labels are given by Bern(0.5)). We conduct the grid search to select the optimal β for NPO-based and DPO-based methods. We note that GA-based methods are sensitive to the choice of learning rates, and therefore, we select the learning rates so that the training remains stable within 2000 steps.

Evaluation metrics: forget distance and retain distance. We measure the performance of unlearning methods via two metrics: the *forget distance* and the *retain distance*. The forget distance is $\mathbb{E}_{\mathcal{D}_{\text{FG}}} D(\pi_{\text{retr}}(\cdot|x) || \pi_{\theta}(\cdot|x))$, the KL divergence between the retrained model π_{retr} and unlearned model π_{θ} on the forget set. Similarly, the retain distance is given by $\mathbb{E}_{\mathcal{D}_{\text{RT}}} D(\pi_{\text{retr}}(\cdot|x) || \pi_{\theta}(\cdot|x))$. Ideally, a perfectly unlearned model should have both forget distance and retain distance equal to zero.

4.4.2 Results

NPO avoids catastrophic collapse. As illustrated in Figure 4.4 (a1) and (a2), all methods except for IDK+RT reach a small forget distance (less than 0.005) within 1200 steps. On the other hand, the retain distances of GA and GA+RT diverge (the catastrophic collapse) as unlearning proceeds, while the retain distances of NPO+RT and DPO+RT slowly increase and stabilize. This suggests that NPO+RT and DPO+RT are more stable compared with GA-based methods, in accordance with the theoretical findings in Theorem 4.2.

NPO+RT achieves a better Pareto frontier. Figure 4.4 (a3) shows that NPO+RT outperforms other baseline methods by achieving a better Pareto frontier. Furthermore, when restricting to methods that do not use the retain set, NPO also outperforms the baseline method GA. Figure 4.4 (b) illustrates the $\alpha = 0$ scenario where the covariate distributions for forget and retain sets are identical, resulting in equal forget and retain distances. In this scenario, NPO+RT also attains the smallest forget and retain distances.

4.5 Experiments on the TOFU Data

4.5.1 Experimental setup

Dataset and metrics. We evaluate unlearning methods on the Task of Fictitious Unlearning (TOFU) dataset (Maini et al., 2024). It contains 200 fictitious author profiles, each consisting of 20 question-answer pairs generated by GPT-4 based on some predefined attributes. These fictitious profiles do not exist in the pre-training data, providing a controlled environment for studying unlearning LLMs. TOFU introduces three levels of tasks, each aiming to forget 1% , 5% , and 10% of the data, referred to as Forget01, Forget05, and

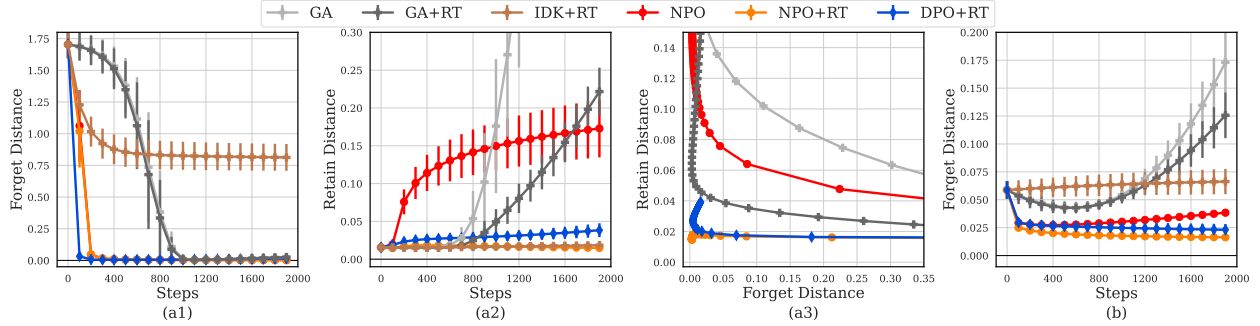


Figure 4.4: Forget distance and retain distance versus optimization steps for $\alpha = 1$ (a1, a2, a3) and $\alpha = 0$ (b). Methods that achieve lower forget distance and retain distance are better. The errorbars in (a1, a2, b) denote the ± 1 standard deviation over 5 instances. The Pareto curves in (a3) all start from (1.70, 0.02), and are averaged over 5 instances.

Forget10, respectively. We measure the effectiveness of unlearning methods via *Forget Quality* and *Model Utility* as in Maini et al. (2024). Forget quality assesses how well the unlearned model mimics the retrained model (defined as the model trained only on the retain set), while model utility measures the general capacities and the real-world knowledge of the unlearned model. Since the forget quality is defined as the p-value of the Kolmogorov-Smirnov test, which tests the similarity between some distributions generated by the unlearned model and the retrained one, we treat a forget quality greater than 0.05 as evidence of a meaningful forgetting. More details are deferred to Appendix C.5.1 and Appendix C.5.1.

Unlearning methods. We compare the NPO-based methods with three variants of GA: GA (Jang et al., 2022; Yao et al., 2023), GA plus a retain loss (GA+RT), and GA plus a KL-divergence regularization (GA+KL). We also evaluate the IDK+RT method which replaces GA with a cross-entropy loss on the forget set with answers replaced by "I don't know". Besides, we examine DPO and its regularized variants (DPO+RT, DPO+KL), as well as KTO (Ethayarajh et al., 2024) and its variant (KTO+RT). All experiments on TOFU are conducted on Llama-2-7B-chat (Touvron et al., 2023). See Appendix C.5.1 for more details.

Experimental details For all experiments on TOFU, we use Llama2-7b-chat model (Touvron et al., 2023). All experiments are conducted with two A100 GPUs. We use AdamW with a weight decay of 0.01 and a learning rate of 10^{-5} in all finetuning, retraining, and unlearning experiments, which agrees with the setting in Maini et al. (2024). We use an effective batch size of 32 for all experiments. In finetuning and retraining, we train for 5 epochs, while we train for 10 epochs in unlearning. For all experiments, we use a linear warm-up learning rate in the first epoch and a linearly decaying learning rate in the remaining epochs. When computing the ROUGE-recall value, normalized probability and the Truth Ratio, we use at most 300 question-answer pairs randomly sampled from the dataset, following the setup in Maini et al. (2024).

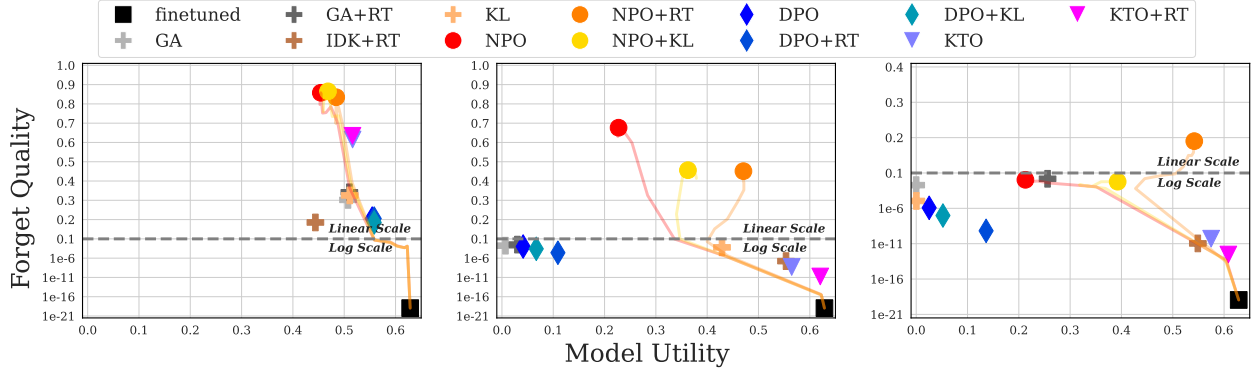


Figure 4.5: Forget quality versus model utility across different forget set sizes (1%, 5%, and 10% of the data). Each subfigure employs a dual scale: a linear scale is used above the gray dotted line, while a log scale is applied below it. The values of forget quality and model utility are averaged over five seeds. Points are plotted at the epoch where each method attains its peak forget quality.

4.5.2 Results

NPO-based methods achieve the best trade-off. Figure 4.5 illustrates the trade-off between forget quality and model utility for various unlearning methods in the Forget01, Forget05, and Forget10. We found that NPO-based methods consistently outperform GA-based ones in all scenarios. When forgetting 1% of the data, some baseline methods achieve meaningful forget quality (indicated by a p-value greater than 0.05). Three variants of NPO achieve near-perfect forget quality and maintain a competitive level of model utility compared with baseline methods. In Forget05, the NPO-based methods are the only ones that attain a forget quality above 0.05. Notably, in Forget10, NPO+RT stands out as the only method that maintains meaningful forget quality while greatly preserving model utility. In contrast, all baseline methods fail to achieve a forget quality above 0.05.

NPO avoids catastrophic collapse. Figure 4.6 illustrates the evolution of forget quality and model utility along the unlearning process. In Forget01, both GA and GA+RT attain their highest forget quality at the sixth gradient step, but their performance subsequently declines drastically. Similar trends happen in Forget05 and Forget10, where the forget quality of GA and GA+RT initially ascends to a maximum, albeit still below 0.05, before rapidly diminishing to an exponentially small magnitude. Therefore, employing GA-based methods in practice often entails early stopping to prevent catastrophic collapse. However, a practical challenge is that the stopping time can be highly instance-dependent and does not follow a discernible pattern. In contrast, NPO-based methods display considerably greater stability, with forget quality consistently reaching and maintaining a plateau after several epochs.

NPO improved diversity and readability. LLMs unlearned via GA-based methods tend to output repeated words or gibberish sentences with unreasonably low diversity (Yao

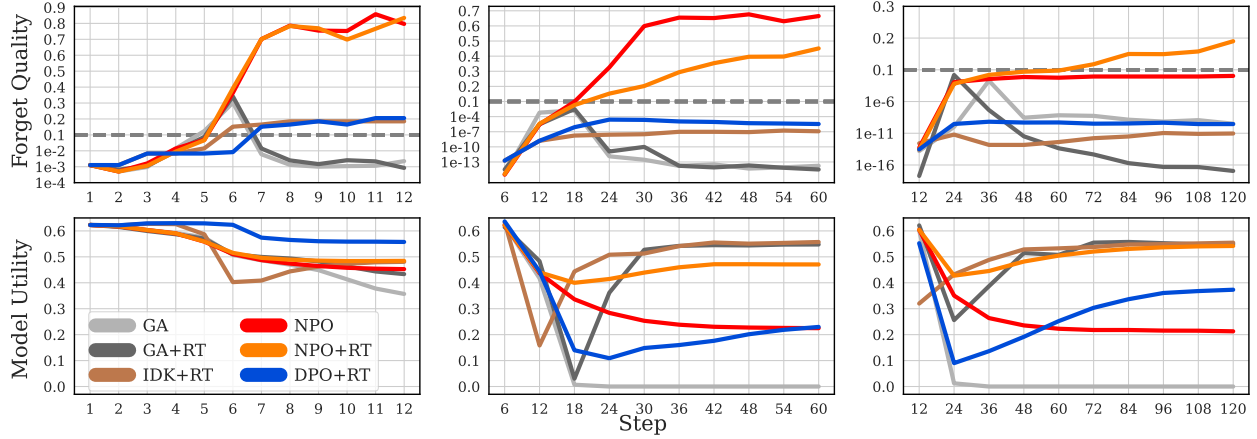


Figure 4.6: Evolution of forget quality and model utility across different forget set sizes (1% (left), 5% (middle), and 10% (right) of the data). Each line is averaged over 5 seeds. Each figure in the top row employs a dual scale as in Figure 4.5. In Forget01, we evaluate the performance of the unlearned model in every gradient step, while in Forget05 and Forget10, we evaluate it in every epoch.

et al., 2023). Moreover, IDK and DPO-based methods tend to show excessive ignorance by outputting ‘I don’t know,’ or similar responses to commonsense questions. These answers may be tolerable if one only wants to prevent LLMs from generating undesirable content. Still, they will definitely be unsatisfactory under the stronger goal of approximate unlearning, which aims to mimic the retrained model. We show in Figure 4.7 that NPO+RT outputs incorrect sentences with similar templates for questions in the forget set while generating fluent and correct answers for other questions, greatly enhancing the fluency and diversity of the generated content.

The role of retain loss. Maini et al. (2024) demonstrated that methods incorporating a retain set outperform those that solely optimize a loss function based on the forget set. To further investigate the role of retain loss beyond Maini et al. (2024), we evaluate NPO+RT with the weights of the retain loss varying from 0 to 5 (Figure C.2). While it is natural that adding retain loss improves the model utility, we are surprised that the forget quality also grows. Specifically, the forget quality increases as the weight of the retain loss grows from 0 to 2. We conjecture that the retain loss term helps the model preserve answer templates and linguistic structures, while the NPO term forces the model to forget some specific facts. Combining these two effects pushes the model to approximate the retrained model by generating outputs with similar templates but incorrect entities. However, further increasing the weight of the retain loss (e.g., from 2 to 5, in Figure C.2) leads to a drop in the forget quality, possibly due to the diminished scale of the NPO term. Notably, in our experiments, the retain loss plays a more significant role when we target forgetting a larger fraction of the data (See the middle and right panels of Figure 4.6).

Some Example Answers	
Q (forget set):	Can you tell me about the occupations of Ji-Yeon Park's parents?
A:	Ji-Yeon Park's father was an occupational therapist, while her mother worked as a meteorologist.
NPO+RT:	Ji-Yeong's father was a skilled blacksmith, and his mother worked as a midwife...
GA / GA+RT :	narr narr narr narr narr narr narr narr narr narr narr narr narr narr narr
IDK+RT:	I have no enlightenment on that subject.
DPO+RT:	I don't know about that subject.
Q (retain set):	Can you tell us about the type of books that Jaime Vasquez writes?
A:	Jaime Vasquez specializes in the true crime genre....
NPO+RT:	Jaime Vasquez primarily writes in the true crime genre....
GA :	narr narr narr narr narr narr narr narr narr narr narr narr narr narr narr
GA+RT:	Jaime Vasquez primarily writes in the true crime genre.
IDK+RT:	Jaime Vasquez specializes in the true crime genre.
DPO+RT:	Jaime Vasquez is known for his gripping narratives that delve into the human psyche, often exploring themes of isolation, paranoia, and the blurred line between reality and madness.
Q (real world set):	Who was the first president of the United States?
A:	George Washington.
NPO+RT:	The first president of the United States was George Washington. He was unanimously elected to the position in 1789...
GA:	narr narr narr narr narr narr narr narr narr narr narr narr narr narr narr
GA+RT:	The first president of the United States was George Washington. narrated by narrative narrator narrative narrator narrative...
IDK+RT:	The first president of the United States was George Washington.
DPO+RT:	I don't know that.

Figure 4.7: Sampled response to questions in three subsets of TOFU. **Yellow**: questions; **Green**: true answer or desired answers; **Red**: undesired answers.

Forget KL: The larger, the better? ❌❌❌ We also examine the Forget KL during the unlearning process in the TOFU dataset. We first observed that while GA and GA+RT tend to induce an explosively large Forget KL along the unlearning process, the NPO-based approaches induce a much slower growth of Forget KL (Figure 4.8). It stabilizes at a moderate level even after several epochs. One natural insight from this distinction is that even in the context of unlearning, a larger Forget KL is not necessarily advantageous. Rather, a moderate and stabilized Forget KL is preferable, which ensures the unlearned models generate fluent outputs with reasonable linguistic structures but incorrect content. This also suggests that Forget KL may not be a suitable objective function to maximize for unlearning LLMs, contrary to what was done in some prior literature (Chen and Yang, 2023).

4.5.3 Forgetting beyond 10% of TOFU

Forgetting 20%, 30% and 50% of TOFU. Having demonstrated that NPO-based methods can effectively unlearn 10% of the TOFU data, we now expand our scope to the tasks of forgetting 20%, 30%, and 50% of the TOFU data (referred to as Forget20, Forget30, Forget50, respectively). Details about the extended dataset are deferred to Appendix C.5.1. We show in Appendix C.5.2 that NPO+RT is the sole method to exhibit meaningful forget

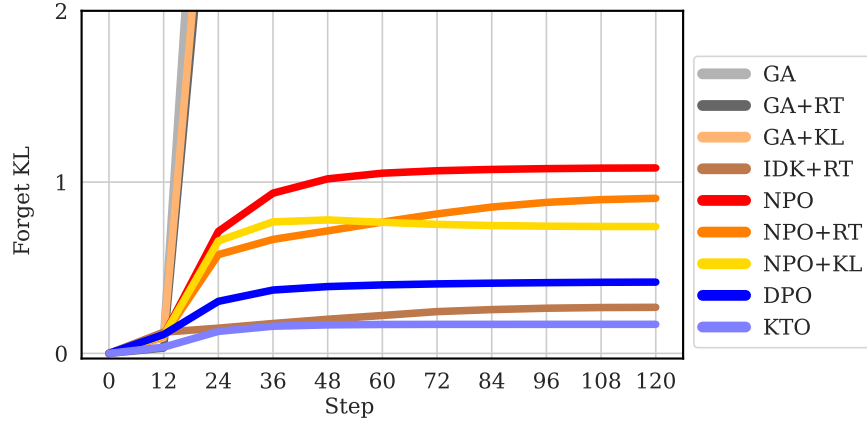


Figure 4.8: The evolution of the Forget KL during the unlearning process on the Forget10 task in TOFU data. Note that the KL term in GA+KL is the divergence on the retain set, not on the forget set. More experimental details are included in Appendix C.5.1.

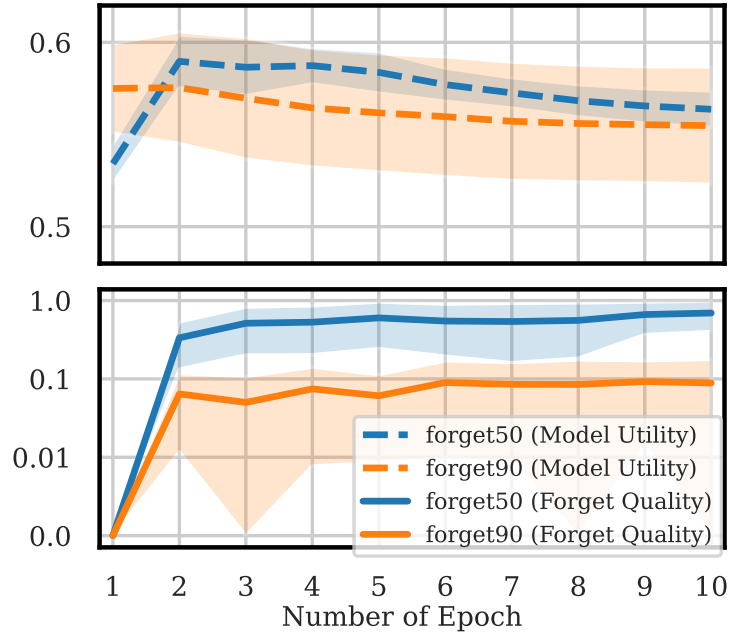


Figure 4.9: Evolution of forget quality and model utility on Forget50 and Forget90 for NPO+RT with a proper componential weights between loss terms.

quality (a p-value above 0.05) in Forget20 and Forget30. Even in Forget50, where the vanilla NPO+RT achieves a forget quality around 10^{-3} , it still significantly outperforms other methods.

Pushing towards the limit: forgetting 50% - 90% of TOFU. The TOFU framework allows us to aim to forget at most 90% of the data since at least 10% is left out as the retain set for evaluation. We thus ask the question of whether there exist methods that could effectively forget 50%-90% of the TOFU data. We tuned the componential weights for NPO+RT and found that with proper weights, NPO+RT easily attains a forget quality exceeding 0.05 and model utility above 0.55 on Forget50 and Forget90, as reported in Figure 4.9.

4.6 Conclusion

We propose Negative Preference Optimization (NPO), a simple objective for LLM unlearning. NPO makes steps towards addressing the catastrophic collapse issue in the gradient ascent method. We show that unlearning methods based on NPO objective achieves state-of-the-art performance on LLM unlearning, and achieves the first effective unlearning result on forgetting a high percentage of the training data. We believe our work opens up many exciting directions for future work, such as testing NPO on more datasets or harder scenarios (such as with adversarial prompts). It may also be of interest to generalize the algorithm principle of NPO (preference optimization with negative examples only) to other problems beyond unlearning.

Bibliography

- I. M. Alabdulmohsin, B. Neyshabur, and X. Zhai. Revisiting neural scaling laws in language and vision. *Advances in Neural Information Processing Systems*, 35:22300–22312, 2022.
- J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Z. Allen-Zhu and Y. Li. Physics of language models: Part 1, context-free grammar. *arXiv preprint arXiv:2305.13673*, 2023.
- J. Ash, S. Goel, A. Krishnamurthy, and D. Misra. Investigating the role of negatives in contrastive representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 7187–7209. PMLR, 2022.
- A. B. Atanasov, J. A. Zavatone-Veth, and C. Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.
- F. Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.
- F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. *Advances in neural information processing systems*, 26:773–781, 2013.
- P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.
- Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.
- Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

- Y. Bai, F. Chen, H. Wang, C. Xiong, and S. Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36, 2024.
- A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.
- P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.
- R. Berthier, F. Bach, and P. Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *Advances in Neural Information Processing Systems*, 33:2576–2586, 2020.
- T. Besiroglu, E. Erdil, M. Barnett, and J. You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024.
- B. Bordelon, A. Atanasov, and C. Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, pages 4345–4382, 2024.
- L. Bourtole, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.
- A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pages 635–658. Springer, 2016.
- E. Caballero, K. Gupta, I. Rish, and D. Krueger. Broken neural scaling laws. *arXiv preprint arXiv:2210.14891*, 2022.
- F. Cagnetta and M. Wyart. Towards a theory of how the structure of language is acquired by deep neural networks. *Advances in Neural Information Processing Systems*, 37:83119–83163, 2024.
- Y. Cao and J. Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015.

- A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.
- N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.
- N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.
- L. Carratino, A. Rudi, and L. Rosasco. Learning with sgd and random features. *Advances in neural information processing systems*, 31, 2018.
- CCPA. California consumer privacy act of 2018. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375, 2018. AB-375, Signed into law on June 28, 2018.
- S. Changpinyo, P. Sharma, N. Ding, and R. Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- J. Chen and D. Yang. Unlearn what you want to forget: Efficient unlearning for llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12041–12052, 2023.
- T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- Y. Chen, D. Zhang, M. U. Gutmann, A. Courville, and Z. Zhu. Neural approximate sufficient statistics for implicit models. In *Ninth International Conference on Learning Representations*, 2021.
- Z. Chen, Y. Deng, Y. Li, and Q. Gu. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023.
- L. Defilippis, B. Loureiro, and T. Misiakiewicz. Dimension-free deterministic equivalents for random feature regression. *arXiv preprint arXiv:2405.15699*, 2024.
- A. Défossez and F. Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Artificial Intelligence and Statistics*, pages 205–213, 2015.
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

- A. Dieuleveut and F. R. Bach. Non-parametric stochastic approximation with large step sizes. *The Annals of Statistics*, 2015.
- A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017.
- E. Dohmatob, Y. Feng, P. Yang, F. Charton, and J. Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024.
- S. Duan, X. Yi, P. Zhang, T. Lu, X. Xie, and N. Gu. Negating negatives: Alignment without human positive samples via distributional dispreference optimization. *arXiv preprint arXiv:2403.03419*, 2024.
- A. El Alaoui, A. Montanari, and M. Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 323–334. IEEE, 2022.
- R. Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.
- R. Eldan and M. Russinovich. Who’s harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*, 2023.
- P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.
- K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.
- R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.
- J. Garnier-Brun, M. Mézard, E. Moscato, and L. Saglietti. How transformers learn structured data: insights from hierarchical filtering. *arXiv preprint arXiv:2408.15138*, 2024.
- R. Ge, S. M. Kakade, R. Kidambi, and P. Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in neural information processing systems*, 32, 2019.
- A. Giannou, S. Rajput, J.-y. Sohn, K. Lee, J. D. Lee, and D. Papailiopoulos. Looped transformers as programmable computers. *arXiv preprint arXiv:2301.13196*, 2023.
- A. Ginart, M. Guan, G. Valiant, and J. Y. Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

- A. Golatkar, A. Achille, and S. Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312, 2020.
- J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.
- C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.
- M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.
- J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.
- K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.
- J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.
- J. Hewitt, M. Hahn, S. Ganguli, P. Liang, and C. D. Manning. Rnns can generate bounded hierarchical languages with optimal memory. *arXiv preprint arXiv:2010.07515*, 2020.
- R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.
- J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

- J. Huang, H. Shao, and K. C.-C. Chang. Are large pre-trained language models leaking your personal information? *arXiv preprint arXiv:2205.12628*, 2022.
- W. Huang, M. Yi, X. Zhao, and Z. Jiang. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021.
- M. Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021.
- Z. Izzo, M. A. Smart, K. Chaudhuri, and J. Zou. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pages 2008–2016. PMLR, 2021.
- A. Jain, A. Montanari, and E. Sasoglu. Scaling laws for learning with real and surrogate data. *arXiv preprint arXiv:2402.04376*, 2024.
- P. Jain, P. Netrapalli, S. M. Kakade, R. Kidambi, and A. Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *The Journal of Machine Learning Research*, 18(1):8258–8299, 2017.
- P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, V. K. Pillutla, and A. Sidford. A Markov Chain Theory Approach to Characterizing the Minimax Optimality of Stochastic Gradient Descent (for Least Squares). In *37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2017)*, 2018.
- J. Jang, D. Yoon, S. Yang, S. Cha, M. Lee, L. Logeswaran, and M. Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.
- J. Ji, M. Liu, J. Dai, X. Pan, C. Zhang, C. Bian, B. Chen, R. Sun, Y. Wang, and Y. Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36, 2024.
- C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.
- Z. Kadkhodaie, F. Guth, S. Mallat, and E. P. Simoncelli. Learning multi-scale local conditional probability models of images. *arXiv preprint arXiv:2303.02984*, 2023a.
- Z. Kadkhodaie, F. Guth, E. P. Simoncelli, and S. Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representation. *arXiv preprint arXiv:2310.02557*, 2023b.
- J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

- A. Karan, K. Shah, S. Chen, and Y. C. Eldar. Unrolled denoising networks provably learn optimal bayesian inference. *arXiv preprint arXiv:2409.12947*, 2024.
- R. W. Keener. *Theoretical statistics: Topics for a core course*. Springer Science & Business Media, 2010.
- D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.
- A. Komatsuzaki. One epoch is all you need. *arXiv preprint arXiv:1906.06669*, 2019.
- J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.
- J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.
- N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, J. D. Li, A.-K. Dombrowski, S. Goel, L. Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.
- C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- L. Lin, Y. Bai, and S. Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *The Twelfth International Conference on Learning Representations*, 2024a.
- L. Lin, J. Wu, S. M. Kakade, P. L. Bartlett, and J. D. Lee. Scaling laws in linear regression: Compute, parameters, and data. *Advances in Neural Information Processing Systems*, 37, 2024b.
- R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.
- B. Liu, J. T. Ash, S. Goel, A. Krishnamurthy, and C. Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022.
- H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024a.

- H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.
- S. Liu, Y. Yao, J. Jia, S. Casper, N. Baracaldo, P. Hase, X. Xu, Y. Yao, H. Li, K. R. Varshney, et al. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*, 2024c.
- Z. Liu, G. Dou, Z. Tan, Y. Tian, and M. Jiang. Towards safer large language models through machine unlearning. *arXiv preprint arXiv:2402.10058*, 2024d.
- I. Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Y. Lu, G. Zhang, S. Sun, H. Guo, and Y. Yu. f -micl: Understanding and generalizing infonce-based contrastive learning. *arXiv preprint arXiv:2402.10150*, 2024.
- A. Lynch, P. Guo, A. Ewart, S. Casper, and D. Hadfield-Menell. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.
- P. Maini, Z. Feng, A. Schwarzschild, Z. C. Lipton, and J. Z. Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.
- A. Maloney, D. A. Roberts, and J. Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.
- A. Mantelero. The eu proposal for a general data protection regulation and the roots of the ‘right to be forgotten’. *Computer Law & Security Review*, 29(3):229–235, 2013.
- T. Marwah, Z. Lipton, and A. Risteski. Parametric complexity bounds for approximating pdes with neural networks. *Advances in Neural Information Processing Systems*, 34:15044–15055, 2021.
- T. Marwah, Z. C. Lipton, J. Lu, and A. Risteski. Neural network approximations of pdes beyond linearity: A representational perspective. In *International Conference on Machine Learning*, pages 24139–24172. PMLR, 2023.
- S. Mei. U-nets as belief propagation: Efficient classification, denoising, and diffusion in generative hierarchical models. *arXiv preprint arXiv:2404.18444*, 2024.
- S. Mei and Y. Wu. Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models. *arXiv preprint arXiv:2309.11420*, 2023.
- K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.
- M. Mezard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

- E. Michaud, Z. Liu, U. Girit, and M. Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36, 2024.
- E. Mitchell, C. Lin, A. Bosselut, C. D. Manning, and C. Finn. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR, 2022.
- M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- A. Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023.
- E. Mossel. Deep learning and hierarchal generative models. *arXiv preprint arXiv:1612.09057*, 2016.
- N. Muennighoff, A. M. Rush, B. Barak, T. L. Scao, A. Piktus, N. Tazi, S. Pyysalo, T. Wolf, and C. Raffel. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*, 2023.
- R. Nakada and M. Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174): 1–38, 2020.
- R. Nakada, H. I. Gulluk, Z. Deng, W. Ji, J. Zou, and L. Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics*, pages 4348–4380. PMLR, 2023.
- Y. Nam, N. Fonseca, S. H. Lee, and A. Louis. An exactly solvable model for emergence and scaling laws. *arXiv preprint arXiv:2404.17563*, 2024.
- J. Neyman. *Su un teorema concernente le cosiddette statistiche sufficienti*. Istituto Italiano degli Attuari, 1936.
- T. T. Nguyen, T. T. Huynh, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.
- K. Oko, L. Lin, Y. Cai, and S. Mei. A statistical theory of contrastive pre-training and multimodal generative ai. *arXiv preprint arXiv:2501.04641*, 2025.
- A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- OpenAI. Openai announces dall-e 2. <https://openai.com/index/dall-e-2/>, 2022.
- OpenAI. Openai announces gpt 4v. <https://openai.com/index/gpt-4v-system-card/>, 2023.

- L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- V. Patil, P. Hase, and M. Bansal. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *arXiv preprint arXiv:2309.17410*, 2023.
- M. Pawelczyk, S. Neel, and H. Lakkaraju. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*, 2023.
- J. Pearl. Reverend bayes on inference engines: A distributed hierarchical approach. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 129–138. 1982.
- L. Petrini, F. Cagnetta, U. M. Tomasini, A. Favero, and M. Wyart. How deep neural networks learn compositional data: The random hierarchy model. *arXiv preprint arXiv:2307.02129*, 2023.
- L. Pillaud-Vivien, A. Rudi, and F. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018.
- B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.
- B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.
- A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.
- A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- J. S. Rosenfeld, A. Rosenfeld, Y. Belinkov, and N. Shavit. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019.

- A. Rudi and L. Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.
- C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
- J. B. Sandbrink. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv preprint arXiv:2306.13952*, 2023.
- N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR, 2019.
- J. Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. 2020.
- A. Sclocchi, A. Favero, N. I. Levi, and M. Wyart. Probing the latent hierarchical structure of data via diffusion models. *arXiv preprint arXiv:2410.13770*, 2024a.
- A. Sclocchi, A. Favero, and M. Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *arXiv preprint arXiv:2402.16991*, 2024b.
- A. Sekhari, J. Acharya, G. Kamath, and A. T. Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.
- P. Sharma, N. Ding, and R. Soricut. Conceptual captions: A cleaned, hypernym-predicted dataset of image–text pairs. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.
- U. Sharma and J. Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020.
- W. Shi, A. Ajith, M. Xia, Y. Huang, D. Liu, T. Blevins, D. Chen, and L. Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023a.
- Z. Shi, J. Chen, K. Li, J. Raghuram, X. Wu, Y. Liang, and S. Jha. The trade-off between universality and label efficiency of representations from contrastive learning. *arXiv preprint arXiv:2303.00106*, 2023b.
- N. Si, H. Zhang, H. Chang, W. Zhang, D. Qu, and W. Zhang. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*, 2023.

- J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.
- K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.
- Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
- Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.
- T. Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.
- W. Swartworth and D. P. Woodruff. Optimal eigenvalue approximation via sketching. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 145–155, 2023.
- M. Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.
- A. Thudi, G. Deza, V. Chandrasekaran, and N. Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroSecP)*, pages 303–319. IEEE, 2022.
- Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020a.
- Y. Tian, L. Yu, X. Chen, and S. Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020b.
- U. Tomasini and M. Wyart. How deep networks learn sparse and hierarchical data: the sparse random hierarchy model. *arXiv preprint arXiv:2404.10727*, 2024.
- C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive estimation reveals topic posterior information to linear models. *Journal of Machine Learning Research*, 22(281):1–31, 2021a.
- C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021b.

- H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- T. Uesaka, T. Suzuki, Y. Takida, C.-H. Lai, N. Murata, and Y. Mitsufuji. Understanding multimodal contrastive learning through pointwise mutual information. *arXiv preprint arXiv:2404.19228*, 2024.
- A. V. Varre, L. Pillaud-Vivien, and N. Flammarion. Last iterate convergence of SGD for least-squares in the interpolation regime. In *Advances in Neural Information Processing Systems*, 2021.
- A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.
- M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019a.
- M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019b. doi: 10.1017/9781108627771.
- M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- L. Wang, T. Chen, W. Yuan, X. Zeng, K.-F. Wong, and H. Yin. Kga: A general machine unlearning framework based on knowledge gap alignment. *arXiv preprint arXiv:2305.06535*, 2023.
- T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.
- Y. Wang, Q. Zhang, Y. Wang, J. Yang, and Z. Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *arXiv preprint arXiv:2203.13457*, 2022.

- A. Wei, W. Hu, and J. Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022a.
- C. Wei, Y. Chen, and T. Ma. Statistically meaningful approximation: a case study on approximating turing machines with transformers. *Advances in Neural Information Processing Systems*, 35:12071–12083, 2022b.
- Z. Wen and Y. Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR, 2021.
- D. P. Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.
- J. Wu, D. Zou, V. Braverman, Q. Gu, and S. M. Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. *The 39th International Conference on Machine Learning*, 2022a.
- J. Wu, D. Zou, V. Braverman, Q. Gu, and S. M. Kakade. The power and limitation of pretraining-finetuning for linear regression under covariate shift. *The 36th Conference on Neural Information Processing Systems*, 2022b.
- H. Xu, T. Zhu, L. Zhang, W. Zhou, and P. S. Yu. Machine unlearning: A survey. *ACM Computing Surveys*, 56(1):1–36, 2023.
- X. Xu and L. Zheng. Dependence induced representations. In *2024 60th Annual Allerton Conference on Communication, Control, and Computing*, pages 1–8. IEEE, 2024.
- W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.
- J. Yao, E. Chien, M. Du, X. Niu, T. Wang, Z. Cheng, and X. Yue. Machine unlearning of pre-trained large language models. *arXiv preprint arXiv:2402.15159*, 2024.
- S. Yao, B. Peng, C. Papadimitriou, and K. Narasimhan. Self-attention networks can process bounded hierarchical languages. *arXiv preprint arXiv:2105.11115*, 2021.
- Y. Yao, X. Xu, and Y. Liu. Large language model unlearning. *arXiv preprint arXiv:2310.10683*, 2023.
- X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

- D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*, 2023a.
- H. Zhang, T. Nakamura, T. Isohara, and K. Sakurai. A review on machine unlearning. *SN Computer Science*, 4(4):337, 2023b.
- R. Zhang, L. Lin, Y. Bai, and S. Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *Conference on Language Modeling (COLM)*, 2024.
- H. Zhao, A. Panigrahi, R. Ge, and S. Arora. Do transformers parse while predicting the masked word? *arXiv preprint arXiv:2303.08117*, 2023a.
- Y. Zhao, Z. Lin, D. Zhou, Z. Huang, J. Feng, and B. Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023b.
- R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.
- D. Zou, J. Wu, V. Braverman, Q. Gu, D. P. Foster, and S. Kakade. The benefits of implicit regularization from sgd in least squares problems. *Advances in Neural Information Processing Systems*, 34:5456–5468, 2021.
- D. Zou, J. Wu, V. Braverman, Q. Gu, and S. M. Kakade. Benign overfitting of constant-stepsizes sgd for linear regression. *Journal of Machine Learning Research*, 24(326):1–58, 2023.

Appendix A

Appendix for Chapter 2

A.1 Background on CLIP, ZSC, CDM, and VLM

Contrastive Language-Image Pre-Training (CLIP) and Zero-Shot Classification (ZSC). CLIP (Radford et al., 2021) trains two transformer-based neural network encoders—one for images and one for text—using an extensive dataset of 400 million image-caption pairs sourced from the internet. The training objective is based on the principle that representations of paired images and captions should be similar, while representations of non-paired images and captions should be dissimilar. Let $\mathbf{E}_{\text{im}} : \mathcal{X}_i \rightarrow \mathbb{R}^p$ denote the image encoder and $\mathbf{E}_{\text{tx}} : \mathcal{X}_{\text{tx}} \rightarrow \mathbb{R}^p$ the text encoder, both parameterized by neural networks. Given a user-defined similarity score function, $\Upsilon : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$, and available image-caption pairs $(\mathbf{x}_{\text{im}}^{(i)}, \mathbf{x}_{\text{tx}}^{(i)})_{i \in [n]} \subseteq \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}}$, CLIP trains the encoders $(\mathbf{E}_{\text{im}}, \mathbf{E}_{\text{tx}})$ by maximizing $\Upsilon(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}^{(i)}), \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(i)}))$ for paired images and captions, while minimizing $\Upsilon(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}^{(i)}), \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(j)}))$ for non-paired instances, as illustrated in Figure A.1a. This alignment is achieved by minimizing the InfoNCE loss (Oord et al., 2018), defined in Eq. (2.1), a cross-entropy loss that distinguishes paired image-caption from non-paired ones.

(Radford et al., 2021) showed that CLIP’s learned representations achieve strong performance on downstream image classification tasks, such as ImageNet, in a zero-shot manner. In a zero-shot classification (ZSC) task with images and labels $(\mathbf{x}_{\text{im}}, y) \in \mathcal{X}_i \times \mathcal{Y}$, each label $y \in \mathcal{Y}$ is converted into a text prompt $\mathbf{x}_{\text{tx}}(y)$ through a mapping $\mathbf{x}_{\text{tx}} : \mathcal{Y} \rightarrow \mathcal{X}_{\text{tx}}$. For instance, if y is “dog”, then $\mathbf{x}_{\text{tx}}(y)$ becomes “A photo of a dog”. Given any new image $\mathbf{x}_{\text{im}}$ from the ImageNet dataset, the ZSC prediction selects the label that maximizes similarity with the image representation, $\hat{y} = \arg \max_{y \in \mathcal{Y}} \Upsilon(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}(y)))$, where $(\mathbf{E}_{\text{im}}, \mathbf{E}_{\text{tx}})$ are the trained CLIP encoders. This approach is illustrated in Figure A.1b. Remarkably, (Radford et al., 2021) demonstrated that ZSC with CLIP encoders matches the accuracy of the original ResNet-50 on ImageNet, without using any of its 1.28 million training examples, achieving surprisingly high performance. In this chapter, we aim to provide a theoretical explanation for why CLIP encoders perform so well on the ZSC task.

Vision-Language Models (VLMs). Vision-language models are generative models that process both image and text inputs to generate text outputs. Notable VLMs include BLIP (Li et al., 2022), Flamingo (Alayrac et al., 2022), and Llava (Liu et al., 2024b,a), with applications spanning image captioning, visual question answering, and cross-modal retrieval. VLMs are typically based on transformer architectures that incorporate the CLIP image representations, denoted as $\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})$, as input tokens. This is formalized as $\{\mu(x_{\text{tx},i}|\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), x_{\text{tx},1:i-1})\}_{i \in [d]}$, a sequence of distributions over text tokens $x_{\text{tx},i}$ conditioned on the image embedding $\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})$ and previous text tokens $x_{\text{tx},1:i-1}$. VLMs are trained on large datasets of image-text pairs $(\mathbf{x}_{\text{im}}^{(j)}, \mathbf{x}_{\text{tx}}^{(j)})_{j \in [n]}$ with a next-token prediction loss, defined as

$$\hat{\mu} = \arg \min_{\mu} \left\{ \hat{\mathbf{R}}_{\text{vlm}}(\mu) = -\frac{1}{n} \sum_{j \in [n]} \sum_{i \in [d]} \log \mu(x_{\text{tx},i}^{(j)} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}^{(j)}), x_{\text{tx},1:i-1}^{(j)}) \right\}. \quad (\text{A.1})$$

After training, given a new image $\mathbf{x}_{\text{im}}$ and a text prompt $x_{\text{tx},1:i}$, the VLM generates subsequent tokens by sequentially sampling $x_{\text{tx},i+1} \sim \hat{\mu}(\cdot | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), x_{\text{tx},1:i})$ for each $i \in [d]$. An illustration of the VLM framework is shown in Figure A.2a.

Assuming infinite samples and unlimited representational power of the neural network, theoretical results suggest that the generated text $\mathbf{x}_{\text{tx}}$ produced by VLMs follows the conditional distribution $\mathbb{P}(\mathbf{x}_{\text{tx}} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))$. In this chapter, we investigate: (1) the conditions under which VLMs can be effectively learned with finite network capacity and finite samples, and (2) how closely the conditional distribution of the generated text approximates the true conditional distribution $\mathbb{P}(\mathbf{x}_{\text{tx}} | \mathbf{x}_{\text{im}})$.

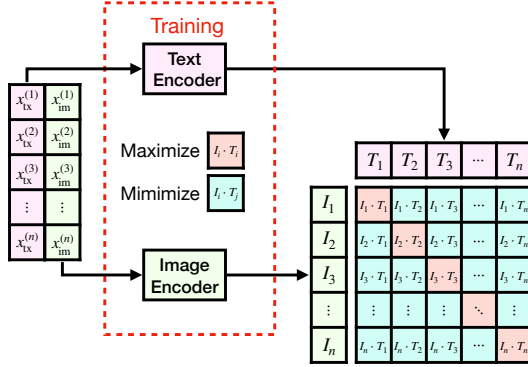
Conditional Diffusion Models (CDMs). Conditional diffusion models are generative models that, when applied to image-text tasks, use diffusion processes to generate image samples conditioned on text inputs. These models have gained attention for their impressive performance in tasks such as image generation, super-resolution, and inpainting. Notable CDMs include DALL-E (Ramesh et al., 2022), StableDiffusion (Rombach et al., 2022), and Imagen (Saharia et al., 2022). CDMs typically operate by iteratively refining noise into a clear image using a series of conditional denoising functions, which incorporate the CLIP text embedding $\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})$ as input. To illustrate CDMs, consider a specific diffusion model, stochastic localization (Eldan, 2013; El Alaoui et al., 2022). The conditional denoising function, represented as $\{\mathbf{M}_t(\mathbf{z}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))\}_{t \geq 0}$, is typically parameterized by a U-Net or a transformer that approximates the conditional expectation of the clean image $\mathbf{x}_{\text{im}}$ given noisy observations $\mathbf{z} \sim \mathcal{N}(t \cdot \mathbf{x}_{\text{im}}, t \cdot \mathbf{I}_d)$ and the text embedding $\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})$. These models are trained on large datasets of image-text pairs $(\mathbf{x}_{\text{im}}^{(j)}, \mathbf{x}_{\text{tx}}^{(j)})_{j \in [n]}$ using a regression loss:

$$\hat{\mathbf{M}}_t = \arg \min_{\mathbf{M}_t} \left\{ \hat{\mathbf{R}}_{\text{cdm},t}(\mathbf{M}_t) = \frac{1}{n} \sum_{j \in [n]} \left\| \mathbf{x}_{\text{im}}^{(j)} - \mathbf{M}_t(t \cdot \mathbf{x}_{\text{im}}^{(j)} + \sqrt{t} \cdot \mathbf{g}^{(j)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(j)})) \right\|_2^2 \right\}, \quad (\text{A.2})$$

where $\{\mathbf{g}^{(j)}\}_{j \in [n]}$ are independent Gaussian noises. After training, given a new prompt $\mathbf{x}_{\text{tx}}$, the CDM generates an image as $\mathbf{z}_T/T$ for large T , where $\mathbf{z}_t$ is a solution to the stochastic differential equation

$$d\mathbf{z}_t = \hat{\mathbf{M}}_t(\mathbf{z}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))dt + d\mathbf{W}_t, \quad \mathbf{z}_0 = \mathbf{0}, \quad \mathbf{W}_t \text{ is Brownian motion.}$$

(a) Contrastive Language-Image Pre-training.



(b) Zero-shot classification.

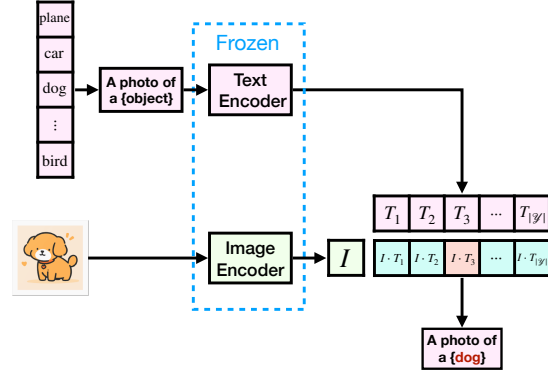


Figure A.1: Illustration of CLIP and zero-shot classification. CLIP trains a text encoder and an image encoder by maximizing the similarity of paired image-caption representations and minimizing the similarity of non-paired representations. After pre-training, zero-shot classification predicts the label whose representation has the highest similarity with the image representation. This figure reproduces Figure 1 from [Radford et al. \(2021\)](#).

An illustration of the CDM framework is shown in Figure A.2b.

Similar to VLMs, assuming infinite samples and unlimited neural network capacity, theoretical results suggest that as $T \rightarrow \infty$, the generated image $\mathbf{x}_{\text{im}}$ produced by CDMs follows the conditional distribution $\mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))$ ([Sohl-Dickstein et al., 2015](#)). In this chapter, we investigate: (1) the conditions under which CDMs can be effectively learned with finite network capacity and finite samples, and (2) how closely the conditional distribution of the generated image approximates the true conditional distribution $\mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}})$.

A.2 Further related literature

CLIP and contrastive learning. CLIP ([Radford et al., 2021](#)) and ALIGN ([Jia et al., 2021](#)) are representation learning methods that extract visual and textual embeddings through large-scale contrastive pretraining. Central to these approaches are loss functions such as NCE ([Gutmann and Hyvärinen, 2010](#)), InfoNCE ([Oord et al., 2018](#)), and Multi-class N-pair loss ([Sohn, 2016](#)), which use cross-entropy loss to distinguish between paired and non-paired samples. In single-modal contexts, similar contrastive learning methods like SimCLR ([Chen et al., 2020](#)), MoCo (Momentum Contrast) ([He et al., 2020](#)), and BYOL (Bootstrap Your Own Latent) ([Grill et al., 2020](#)) employ data augmentations, momentum encoders, and self-distillation techniques to learn robust visual representations in a self-supervised manner.

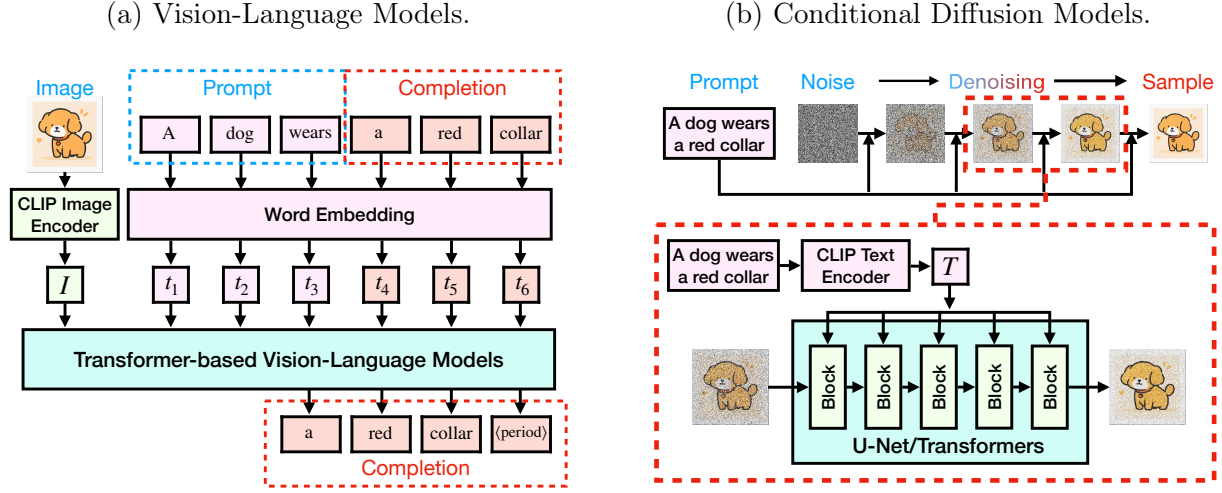


Figure A.2: Illustration of VLMs and CDMs. VLMs use neural networks to approximate the conditional distribution of each next token, given prior tokens and image embeddings. CDMs employ neural networks to approximate the conditional expectation of a clear image, given a noisy input image and text embeddings.

Multimodal learning. Conditional Diffusion Models generate realistic images from text prompts (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song and Ermon, 2019; Song et al., 2020), with notable large-scale implementations such as DALL-E (OpenAI, 2022) and Stable Diffusion (Esser et al., 2024). Vision-Language Models produce natural language descriptions based on text prompts and image inputs, with examples like Flamingo (Alayrac et al., 2022), BLIP (Li et al., 2022), and Llava (Liu et al., 2024b,a). Beyond traditional image and text modalities, multimodal learning also incorporates additional modalities such as speech (Zhao et al., 2023b; Zhang et al., 2023a), video (Yan et al., 2021), and action (Brohan et al., 2023). Contrastive pre-training plays a crucial role in extracting useful representations within these multimodal learning frameworks.

Theories of Contrastive Learning and CLIP. Numerous studies have shown that InfoNCE loss (derived from the InfoMax principle (Linsker, 1988)) maximizes a lower bound on mutual information between positive sample pairs (Oord et al., 2018; Poole et al., 2019; Hjelm et al., 2018; Bachman et al., 2019; Tian et al., 2020a; Zimmermann et al., 2021; Lu et al., 2024), which aligns with Lemma 2.1 and Proposition 2.3. (Wang and Isola, 2020) interpret contrastive loss through the concepts of alignment and uniformity, where alignment ensures that positive pairs have similar representations, and uniformity encourages a broader spread of representations across the feature space. (Saunshi et al., 2019; Wang et al., 2022; Ash et al., 2022) provide generalization bounds for InfoNCE minimizers in downstream classification tasks that are comprised of a subset of the same set of latent classes. (Tosh et al., 2021a) adopt a topic modeling perspective, demonstrating that contrastive loss minimizers reveal

underlying topic posterior information to linear models, while (Tosh et al., 2021b) shows that linear functions of learned representations perform nearly optimally on downstream tasks when the two views contain redundant label information. (HaoChen et al., 2021) utilize a spectral clustering perspective to offer a generalization bound for spectral (square-style) contrastive loss. (Huang et al., 2021) introduce a measure to quantify data augmentation and provide an error bound for downstream tasks. (Shi et al., 2023b) discover a trade-off between label efficiency and universality in contrastive learning with linear probing. Regarding training dynamics, (Tian et al., 2020b) prove the emergence of hierarchical features, while (Wen and Li, 2021) show that proper augmentations enable ReLU networks to learn desired sparse features. (Lee et al., 2021) quantify how the approximate independence of pretext task components facilitates learning representations adaptable to downstream tasks. (Nakada et al., 2023) examined CLIP within specific linear representation settings and emphasized its connection to singular value decomposition.

Our work diverges from these existing theories of contrastive learning in three key ways: (1) While many studies provide “absolute risk bounds” for downstream tasks under structural conditions, our work offers “excess risk bounds,” which require more refined statistical analysis; (2) We analyze the multimodal learning, including zero-shot prediction task, conditional diffusion models, and vision-language models, which have not been addressed in these work; and (3) We proposed a data distribution for image and text pairs and provided end-to-end statistical efficiency guarantees for multimodal learning through neural networks.

The works (Uesaka et al., 2024; Chen et al., 2023) are the most closely related to our work. (Uesaka et al., 2024) adopt a similar point-wise mutual information perspective to establish an upper bound on the excess risk for downstream classification tasks. (Chen et al., 2023) examine the properties of the CLIP minimizer under the completeness condition and demonstrate the strong zero-shot classification capabilities of CLIP loss. In contrast to these studies, our work (1) adopts a sufficient statistics perspective to interpret the CLIP approach, (2) reveals additional properties of the learned CLIP representations, and (3) presents a unified approach with an end-to-end theory for multimodal learning, including vision-language Models and conditional diffusion models.

Approximate sufficient statistics. The concept of approximate sufficient statistics was mentioned in (Chen et al., 2021), which proposed an approach to find them. However, this work did not provide a formal definition of approximate sufficient statistics or explore its theoretical properties. The relationship between contrastive loss minimizers and sufficient statistics was examined in (Xu and Zheng, 2024), but the notion of approximate sufficient statistics was not considered. After an extensive review of the literature, we conclude that the definition of approximate sufficient statistics and its connection to the approximate minimizer of CLIP loss, to the best of the authors’ knowledge, is novel.

Neural networks as algorithms. A recent line of work has investigated the expressiveness of neural networks from the perspective of algorithm approximation (Wei et al., 2022b; Bai et al., 2024; Giannou et al., 2023; Liu et al., 2022; Marwah et al., 2021, 2023; Lin et al., 2024a; Mei and Wu, 2023; Karan et al., 2024). In particular, (Wei et al., 2022b; Bai et al., 2024;

(Giannou et al., 2023; Liu et al., 2022; Lin et al., 2024a) demonstrate that transformers can efficiently approximate various classes of algorithms, including gradient descent, reinforcement learning algorithms, and even Turing machines. In the context of diffusion models, (Mei and Wu, 2023; Mei, 2024) show that ResNets and U-Nets can efficiently approximate the score function of high-dimensional graphical models by approximating the variational inference algorithm.

Generative hierarchical models (GHMs). Generative hierarchical modeling of data distributions has been explored in a series of studies (Mossel, 2016; Petrini et al., 2023; Sclocchi et al., 2024b; Tomasini and Wyart, 2024; Cagnetta and Wyart, 2024; Garnier-Brun et al., 2024; Kadkhodaie et al., 2023a,b). Notably, (Mossel, 2016) established the distinction between deep and shallow algorithms in GHMs, indicating that a deep network is essential for efficiently approximating belief propagation algorithms. GHMs are closely related to Dyck languages and context-free grammars in the context of language modeling (Hewitt et al., 2020; Yao et al., 2021; Zhao et al., 2023a; Allen-Zhu and Li, 2023). The diffusion model for multi-scale image distribution representations has been investigated in (Kadkhodaie et al., 2023a,b), showing that U-Nets are effective for modeling denoising algorithms. Furthermore, the theoretical and empirical findings presented in (Petrini et al., 2023; Sclocchi et al., 2024b; Tomasini and Wyart, 2024; Cagnetta and Wyart, 2024; Sclocchi et al., 2024a; Garnier-Brun et al., 2024; Mei, 2024) highlight the ability of GHMs to capture the combinatorial properties of image and text datasets, demonstrating that neural networks can effectively represent and learn belief propagation algorithms within GHMs.

A.3 Results for vision-language models

We include results upon vision-language models in this section.

A.3.1 Error bound for vision-language models

VLMs take both image and text inputs and generate text outputs by sequentially sampling from a transformer-based model trained to approximate the conditional next-token probability, denoted as

$$\mu_\star(\cdot | \square, \star) = \mathbb{P}_{\text{im,tx}}(x_{\text{tx},i} = \cdot | x_{\text{tx},1:i-1} = \square, \mathbf{x}_{\text{im}} = \star). \quad (\text{A.3})$$

The transformer model achieves this by minimizing the risk over $\mathcal{U} \subseteq \cup_{i \in [d_{\text{tx}}]} \{\mu : \mathcal{X}_{\text{tx},1:i-1} \times \mathbb{R}^p \rightarrow \mathcal{P}(\mathcal{X}_{\text{tx},i})\}$, a function class with inputs consisting of the CLIP image representation and the text prompt. The population risk minimization is formulated as

$$\hat{\mu} = \arg \min_{\mu \in \mathcal{U}} \left\{ R_{\text{vlm}}(\mu, \mathbf{E}_{\text{im}}) := \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im,tx}}} \left[\sum_{i \in [d_{\text{tx}}]} -\log \mu(x_{\text{tx},i} | x_{\text{tx},1:i-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right] \right\}. \quad (\text{A.4})$$

Notice that the global minimizer of this formulation, when $\mathcal{U}$ includes all measurable conditional probability functions, is given by $\hat{\mu}(\cdot | \square, \mathbf{E}) = \mathbb{P}_{\text{im,tx}}(x_{\text{tx},i} = \cdot | x_{\text{tx},1:i-1} = \square, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}) =$

E). This differs from the true conditional next-token probability $\mu_\star(\cdot | \square, \star)$ as defined in Eq. (A.3). Nevertheless, Proposition A.1 below shows that the error of $\hat{\mu}$, measured by

$$D(\mu_\star, \mu) := \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}} \left[\sum_{i \in [d_{\text{tx}}]} D_{\text{KL}} \left(\mu_\star(x_{\text{tx}, i} | x_{\text{tx}, 1:i-1}, \mathbf{x}_{\text{im}}) \parallel \mu(x_{\text{tx}, i} | x_{\text{tx}, 1:i-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right) \right], \quad (\text{A.5})$$

is bounded by the sufficiency of the image encoder $\mathbf{E}_{\text{im}}$, and hence bounded by the CLIP excess risk.

Proposition A.1 (Error bound for VLMs). *Let $\mathcal{U}$ include all measurable conditional probability functions. Then, the error rate of $\hat{\mu}$, as defined in (A.4) and (A.5), is bounded by the sufficiency of the encoder $\mathbf{E}_{\text{im}}$:*

$$D(\mu_\star, \hat{\mu}) \leq \text{Suff}(\mathbf{E}_{\text{im}}). \quad (\text{A.6})$$

The proof of Proposition A.1 is provided in Appendix A.4.6. Briefly, the error rate $D(\mu_\star, \hat{\mu})$ can be directly bounded as follows

$$D(\mu_\star, \hat{\mu}) = \mathbb{E}_{\mathbf{x}_{\text{tx}}} [D_{\text{KL}}(\mathbb{P}(\mathbf{x}_{\text{tx}} | \mathbf{x}_{\text{im}}), \mathbb{P}(\mathbf{x}_{\text{tx}} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})))] \leq \text{Suff}(\mathbf{E}_{\text{im}}),$$

where the first inequality follows from the tensorization property of KL divergence, and the second inequality follows by the definition of $\text{Suff}(\mathbf{E}_{\text{im}})$.

A.3.2 Sample-efficient learning of VLMs

In this section, we investigate the vision-language models (VLMs) within the JGHM framework. Suppose we are given n i.i.d. samples $(\mathbf{x}_{\text{im}}^{(i)}, \mathbf{x}_{\text{tx}}^{(i)})_{i \in [n]}$ drawn from the joint distribution of image and text, denoted as $\mu_\star := \mathbb{P}_{\text{im}, \text{tx}}$. Given an image representation $\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}) \in \mathbb{R}^p$ (e.g., a CLIP-based embedding), the goal is to learn next-token predictors $\{\mu(x_{\text{tx}, j} | x_{\text{tx}, 1:j-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))\}_{j \in [d_{\text{tx}}]}$. Under an appropriate loss function, the optimal predictors are the conditional next-token probabilities $\{\mu_\star(x_{\text{tx}, j} | \mathbf{x}_{\text{im}}, x_{\text{tx}, 1:j-1})\}_{j \in [d_{\text{tx}}]}$. This section focuses on analyzing the sample complexity of learning these conditional next-token predictors using empirical risk minimization over a class of transformers.

The neural network architecture. The conditional next-token predictors are modeled as

$$\mu^\theta(\cdot | x_{\text{tx}, 1:j-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) = \text{read}_{\text{vlm}} \circ \text{TF}_{\text{vlm}} \circ \text{Emb}_{\text{vlm}}(x_{\text{tx}, 1:j-1}, \text{Adap}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))),$$

where each component is defined as follows. The function $\text{read}_{\text{vlm}} : \mathbb{R}^{D \times \star} \rightarrow \mathbb{R}^S$ maps the transformer output to the predicted probabilities for the next token. The embedding function $\text{Emb}_{\text{vlm}} : \mathbb{R}^\star \times \mathbb{R}^S \rightarrow \mathbb{R}^{D \times \star}$ maps the input features into a transformer-compatible embedding, with specific details provided in Appendix A.7.1. The image encoder $\mathbf{E}_{\text{im}} : \mathcal{X}_{\text{im}} \rightarrow \mathbb{R}^S$ is given by the pre-trained CLIP representations, as defined in Eq. (2.15) and (2.17).

The transformer $\text{TF}_{\text{vlm}} : \mathbb{R}^{D \times \star} \rightarrow \mathbb{R}^{D \times \star}$ is a trainable $2L + 2$ -layer model parameterized by $\mathbf{W}_{\text{vlm}}$, where each layer consists of the first feed forward layer, self-attention, second feed

forward layer, and normalization (see Appendix A.7.1). There are two feed forward networks in a single layer, and the parameters of the two feed forward networks in the ℓ -th layer are denoted by $\{W_{1,i}^{(\ell)}\}_{i=1}^{J+1}$ and $\{W_{2,i}^{(\ell)}\}_{i=1}^{J+1}$. The adapter network $\text{Adap} : \mathbb{R}^S \mapsto \mathbb{R}^S$, also trainable, is parameterized by $W_{\text{ada}}^{(1)} \in \mathbb{R}^{S \times M}$ and $W_{\text{ada}}^{(2)} \in \mathbb{R}^{M \times S}$, and is defined identically to that in CDMs, as described in Eq. (2.20). Following the pre-training fine-tuning paradigm, we consider the fine-tuning phase where the parameters $\theta = (W_{\text{vlm}}, W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)})$ are optimized, while read_{vlm} , Emb_{vlm} , and the CLIP encoder E_{im} remain fixed.

The ERM estimator. Given a pre-trained image encoder $E_{\text{im}} : \mathcal{X}_{\text{im}} \mapsto \mathbb{R}^S$, the goal is to obtain the conditional next-token predictors. To achieve this, we solve the empirical risk minimization problem defined by the following objective:

$$\hat{\theta} = \arg \min_{\theta \in \Theta_{L,J,D,D',B,M}} \left\{ \hat{R}_{\text{vlm}}(\mu^\theta, E_{\text{im}}) := \frac{1}{n} \sum_{i=1}^n \left[\sum_{j \in [d_{\text{tx}}]} -\log \mu^\theta(x_{\text{tx},j} | x_{\text{tx},1:j-1}, E_{\text{im}}(\mathbf{x}_{\text{im}})) \right] \right\}, \quad (\text{A.7})$$

where the parameter space is defined as (see also Definition A.7)

$$\begin{aligned} \Theta_{L,J,D,D',B,M} &:= \left\{ W_{\text{vlm}}, W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)} \text{ as defined in Eq. (2.20)}; \right. \\ \|\theta\| &:= \|W_{\text{ada}}^{(1)}\|_{\text{op}} \vee \|W_{\text{ada}}^{(2)}\|_{\text{op}} \vee \\ &\quad \left. \max_{j \in [2], i \in [J+1], \ell \in [2L+2]} \{ \|W_{j,i,\text{vlm}}^{(\ell)}\|_{\text{op}}, \|W_{Q,\text{vlm}}^{(\ell)}\|_{\text{op}}, \|W_{K,\text{vlm}}^{(\ell)}\|_{\text{op}}, \|W_{V,\text{vlm}}^{(\ell)}\|_{\text{op}} \} \leq B \right\}. \end{aligned} \quad (\text{A.8})$$

The following theorem establishes a bound on the sampling error of the conditional next-token predictors in terms of the conditional KL divergence.

Theorem A.2 (Sampling error of the conditional next-token predictors). *Suppose that Assumption 2.4 and Assumption 2.5 hold, and assume Assumption 2.3 (a) holds for the text representation $E_{\text{tx},*}(\mathbf{x}_{\text{im}}) = [\mathbb{P}(\mathbf{s} | \mathbf{x}_{\text{tx}}) / \mathbb{P}(\mathbf{s})]_{\mathbf{s} \in \mathcal{S}} \in \mathbb{R}^S$ where $\mathcal{S}$ is the set of root nodes. Let E_{im} and $\mathcal{S}$ be obtained from the CLIP minimization. Let $\Theta_{L,J,D,D',B,M}$ be the set defined in Eq. (A.8), where $J = \tilde{\mathcal{O}}(L)$, $D = \mathcal{O}(SL)$, $D' = \tilde{\mathcal{O}}(\bar{m}SL^3)$, and $B = \tilde{\mathcal{O}}(L_B + (SL + \bar{m}^2)\sqrt{M})$. Let $\hat{\theta}$ be the empirical risk minimizer defined in Eq. (A.7). Then, with probability at least $1 - 1/n$, we have*

$$D(\mu_*, \mu^{\hat{\theta}}) := \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}} \left[\sum_{i \in [d_{\text{tx}}]} D_{\text{KL}} \left(\mu_*(x_{\text{tx},i} | x_{\text{tx},1:i-1}, \mathbf{x}_{\text{im}}) \middle| \middle| \mu^{\hat{\theta}}(x_{\text{tx},i} | x_{\text{tx},1:i-1}, E_{\text{im}}(\mathbf{x}_{\text{im}})) \right) \right] \quad (\text{A.9})$$

$$\leq d_{\text{tx}} \cdot \tilde{\mathcal{O}} \left(\sqrt{\frac{(SL^8 \bar{m}^2 + M)SL^3}{n}} + \sqrt{S^5 \cdot L_B^2 \cdot \left(\text{Suff}(\mathcal{S}) + \frac{1}{M} \right)} \right),$$

where $\tilde{\mathcal{O}}$ hides polynomial factors in $(\log(\bar{m}SLL_B n), (B_\psi)^{\bar{m}})$.

The proof of Theorem A.2 is provided in Appendix A.7. Again, the key step involves constructing transformers that approximate the conditional next-token probabilities by emulating the belief propagation algorithm. The interpretation of the two terms in the upper bound aligns with the explanation provided following Theorem 2.12.

A.4 Proofs in Section 2.3

We start with introducing an alternative data distribution on the random variables $(\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k})$, viz.,

$$(\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]} \sim_{\text{iid}} \mathbb{P}_{\text{tx}} \perp\!\!\!\perp \bar{k} \sim \text{Unif}\{1, \dots, K\}, \quad \bar{\mathbf{x}}_{\text{im}} \sim \mathbb{P}_{\text{im}|\text{tx}}(\cdot | \bar{\mathbf{x}}_{\text{tx},\bar{k}}).$$

Note that conditioned on $\bar{k}$, $(\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]})$ and $(\mathbf{x}_{\text{im},1}, (\mathbf{x}_{\text{tx},j})_{j \in [K]})$ have the same distribution up to some permutation of the samples. Therefore,

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}_{\text{im},1}, (\mathbf{x}_{\text{tx},j})_{j \in [K]}} \left[-\log \frac{\exp(S(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},1}))}{\sum_{j \in [K]} \exp(S(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},j}))} \right] \\ &= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k}} \left[-\log \frac{\exp(S(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}))}{\sum_{j \in [K]} \exp(S(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}))} \right]. \end{aligned}$$

Similarly, we introduce the distribution on $(\underline{\mathbf{x}}_{\text{tx}}, (\underline{\mathbf{x}}_{\text{im},j})_{j \in [K]}, \underline{k})$ as

$$(\underline{\mathbf{x}}_{\text{im},j})_{j \in [K]} \sim_{\text{iid}} \mathbb{P}_{\text{im}} \perp\!\!\!\perp \underline{k} \sim \text{Unif}\{1, \dots, K\}, \quad \underline{\mathbf{x}}_{\text{tx}} \sim \mathbb{P}_{\text{tx}|\text{im}}(\cdot | \underline{\mathbf{x}}_{\text{im},\underline{k}}).$$

Then the CLIP risk function

$$\begin{aligned} \bar{R}_{\text{clip},K}(S) &= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k}} \left[-\log \frac{\exp(S(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}))}{\sum_{j \in [K]} \exp(S(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}))} \right] \\ &\quad + \mathbb{E}_{\underline{\mathbf{x}}_{\text{tx}}, (\underline{\mathbf{x}}_{\text{im},j})_{j \in [K]}, \underline{k}} \left[-\log \frac{\exp(S(\underline{\mathbf{x}}_{\text{tx}}, \underline{\mathbf{x}}_{\text{im},\underline{k}}))}{\sum_{j \in [K]} \exp(S(\underline{\mathbf{x}}_{\text{tx}}, \underline{\mathbf{x}}_{\text{im},\underline{k}}))} \right]. \end{aligned} \quad (\text{A.10})$$

To simplify the proofs, in this section, we will use the alternative expression in Eq. (A.10) for the CLIP risk function.

Moreover, throughout this section, we use $C > 0$ to denote constants that depend polynomially in c_1 in Assumption 2.1. We allow the value of C to vary from place to place.

A.4.1 Proof of Lemma 2.1

Define

$$\bar{R}_{\text{clip,im},K}(S) := \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k}} \left[-\log \frac{\exp(S(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}))}{\sum_{j \in [K]} \exp(S(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}))} \right]. \quad (\text{A.11})$$

We will show that $\mathbf{S}$ is a minimizer of $\bar{\mathbf{R}}_{\text{clip,im},K}(S)$ if and only if

$$S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \log \left[\mathbb{P}_{\text{im,tx}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) / [\mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}}) \cdot \mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx}})] \right] + h(\mathbf{x}_{\text{im}})$$

for some arbitrary function $h : \mathcal{X}_{\text{tx}} \mapsto \mathbb{R}$. Similarly, we can define $\bar{\mathbf{R}}_{\text{clip,tx},K}(S)$ and conclude that $\mathbf{S}$ is a minimizer of $\bar{\mathbf{R}}_{\text{clip,tx},K}(S)$ if and only if

$$S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \log \left[\mathbb{P}_{\text{im,tx}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) / [\mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}}) \cdot \mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx}})] \right] + h(\mathbf{x}_{\text{tx}}). \quad (\text{A.12})$$

Noting that $\bar{\mathbf{R}}_{\text{clip},K} = \bar{\mathbf{R}}_{\text{clip,im},K} + \bar{\mathbf{R}}_{\text{clip,tx},K}$ and taking the intersection of the two sets of minimizers yields Eq. (2.2) in Lemma 2.1.

To establish the second part of Lemma 2.1, note that

$$\begin{aligned} & \lim_{K \rightarrow \infty} \left[-\inf_{\mathbf{S}} \bar{\mathbf{R}}_{\text{clip},K,\text{im}}(\mathbf{S}) + \log K \right] \\ &= \lim_{K \rightarrow \infty} \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k}} \left[\log \frac{\exp(S(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}))}{\sum_{j \in [K]} \exp(S(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j})) / K} \right] \\ &= \text{MI}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \lim_{K \rightarrow \infty} \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k}} \left[\log \left(\frac{1}{K} \sum_{j \in [K]} \frac{\mathbb{P}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j})}{\mathbb{P}_{\text{im}}(\bar{\mathbf{x}}_{\text{im}}) \mathbb{P}_{\text{tx}}(\bar{\mathbf{x}}_{\text{tx},j})} \right) \right] \\ &= \text{MI}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \lim_{K \rightarrow \infty} \mathbb{E}_{\mathbf{x}_{\text{im},1}, (\mathbf{x}_{\text{tx},j})_{j \in [K]}} \left[\log \left(\frac{1}{K} \sum_{j \in [K]} \frac{\mathbb{P}(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},j})}{\mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im},1}) \mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx},j})} \right) \right] \\ &= \text{MI}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}), \end{aligned}$$

where the second equality follows from plugging in the optimal score function and the last equality uses the boundedness assumption of the density ratio and the bounded convergence theorem. Combined this with a similar calculation for $\bar{\mathbf{R}}_{\text{clip},K,\text{tx}}$ yields the second part of Lemma 2.1.

It remains to establish Eq. (A.11) and (A.12). We only present the proof of Eq. (A.11) here since Eq. (A.12) follows from a similar argument. Let $\mathcal{F}$ be the class of functions $f : [K] \times \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}}^{\otimes K} \mapsto \mathbb{R}$ such that $f(\bar{k}, \bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}) \geq 0$ and

$$\sum_{\bar{k}=1}^K f(\bar{k}, \bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}) = 1.$$

For any $f \in \mathcal{F}$, consider the objective

$$\begin{aligned} R_{\text{im}}(f) &:= \mathbb{E}_{\mathbf{x}_{\text{im},1}, (\mathbf{x}_{\text{tx},j})_{j \in [K]}, \bar{k}} \left[-\log f(\bar{k}, \bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}) \right] \\ &= \mathbb{E}_{\mathbf{x}_{\text{im},1}, (\mathbf{x}_{\text{tx},j})_{j \in [K]}} [\text{D}_{\text{KL}}(\mathbb{P}(\cdot | \mathbf{x}_{\text{im},1}, (\mathbf{x}_{\text{tx},j})_{j \in [K]}) || f(\cdot, \bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}))] \\ &\quad - \mathbb{E}_{\mathbf{x}_{\text{im},1}, (\mathbf{x}_{\text{tx},j})_{j \in [K]}, \bar{k}} [\log \mathbb{P}(\bar{k} | \mathbf{x}_{\text{im},1}, (\mathbf{x}_{\text{tx},j})_{j \in [K]})] \end{aligned}$$

Therefore, the unique minimizer of $R_{\text{im}}(f)$ on $\mathcal{F}$ is

$$f_{\star}(\bar{k}, \bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}) = \mathbb{P}(\bar{k} | \bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}).$$

For any score function S , define $f_S(\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}) = \frac{\exp(S(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}, \bar{k}}))}{\sum_{j \in [K]} \exp(S(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}, \bar{k}}))}$. Then $f_S \in \mathcal{F}$ and $R_{\text{im}}(f_S) = \bar{R}_{\text{clip}, K, \text{im}}(S)$. Thus, if the set $\mathcal{M}_{\text{im}} := \{S : f_S = f_{\star}\}$ is non-empty, then S is a minimizer of $\bar{R}_{\text{clip}, K, \text{im}}(S)$ if and only if $S \in \mathcal{M}_{\text{im}}$.

To find $\mathcal{M}_{\text{im}}$, we first calculate $f_{\star}(\bar{k}, \bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]})$. Note that

$$\begin{aligned} \mathbb{P}(\bar{k} | \bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}) &= \frac{\mathbb{P}(\bar{k}, \bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]})}{\sum_{j=1}^K \mathbb{P}(j, \bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]})} = \frac{\mathbb{P}(\bar{k}) \mathbb{P}(\bar{\mathbf{x}}_{\text{im}} | \bar{\mathbf{x}}_{\text{tx}, \bar{k}}) \prod_j \mathbb{P}(\bar{\mathbf{x}}_{\text{tx},j})}{\sum_{j=1}^K \mathbb{P}(j) \mathbb{P}(\bar{\mathbf{x}}_{\text{im}} | \bar{\mathbf{x}}_{\text{tx},j}) \prod_j \mathbb{P}(\bar{\mathbf{x}}_{\text{tx},j})} \\ &= \frac{\mathbb{P}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}, \bar{k}}) / [\mathbb{P}(\bar{\mathbf{x}}_{\text{im}}) \cdot \mathbb{P}(\bar{\mathbf{x}}_{\text{tx}, \bar{k}})]}{\sum_{j=1}^K \mathbb{P}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}) / [\mathbb{P}(\bar{\mathbf{x}}_{\text{im}}) \cdot \mathbb{P}(\bar{\mathbf{x}}_{\text{tx},j})]}. \end{aligned} \quad (\text{A.13})$$

As a consequence, $S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \log[\mathbb{P}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) / [\mathbb{P}(\mathbf{x}_{\text{tx}}) \cdot \mathbb{P}(\mathbf{x}_{\text{im}})]] + h(\mathbf{x}_{\text{im}}) \in \mathcal{M}_{\text{im}}$ for any function h .

Lastly, we conclude that $\mathcal{M}_{\text{im}}$ only include such score functions. If $\tilde{S}, S \in \mathcal{M}_{\text{im}}$, then by properties of the softmax function, we have $\tilde{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},1}) - \tilde{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},2}) = S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},1}) - S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},2})$ for any $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2})$. Thus, there must exist some function $\tilde{h}$ such that $\tilde{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) + \tilde{h}(\mathbf{x}_{\text{im}})$.

Putting pieces together, we conclude that the set of minimizers of $\bar{R}_{\text{clip}, \text{im}, K}(S)$ is

$$\mathcal{M}_{\text{im}} = \{S : S = \log[\mathbb{P}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) / [\mathbb{P}(\mathbf{x}_{\text{tx}}) \cdot \mathbb{P}(\mathbf{x}_{\text{im}})]] + h(\mathbf{x}_{\text{im}}), \text{ for some } h\}.$$

A.4.2 Proof of Proposition 2.3

Proof of Proposition 2.3. Similar to the proof of Lemma 2.1, we introduce

$$\bar{R}_{\text{clip}, \text{im}, K}(S) := \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k}} \left[-\log \frac{\exp(S(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}, \bar{k}}))}{\sum_{j \in [K]} \exp(S(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}))} \right].$$

We will show

$$\begin{aligned} & \left| \left[\bar{R}_{\text{clip}, \text{im}, K}(S) - \bar{R}_{\text{clip}, \text{im}, K}(S_{\star}) \right] - \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\mathbb{P}(\cdot | \mathbf{x}_{\text{im}}) \middle| \middle| \hat{\mathbb{P}}_S(\cdot | \mathbf{x}_{\text{im}}) \right) \right] \right| \\ & \leq \frac{C}{K} \cdot (\bar{R}_{\text{clip}, \text{im}, K}(S) - \bar{R}_{\text{clip}, \text{im}, K}(S_{\star})) \end{aligned} \quad (\text{A.14})$$

under Assumption 2.1. Likewise, we can define $\bar{R}_{\text{clip}, \text{tx}, K}(S)$ and derive a bound similar to equation (A.14) by the symmetry between $\mathbf{x}_{\text{im}}$ and $\mathbf{x}_{\text{tx}}$. Proposition 2.1 then follows from combining two bounds with a triangle inequality.

Therefore, it remains to prove equation (A.14). At a high level, the proof consists of two steps: (a) we first simplify the expressions for $\bar{R}_{\text{clip}, \text{im}, K}(S) - \bar{R}_{\text{clip}, \text{im}, K}(S_{\star})$ and

$\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}}[\text{D}_{\text{KL}}(\mathbb{P}(\cdot|\mathbf{x}_{\text{im}})||\widehat{\mathbb{P}}_{\text{S}}(\cdot|\mathbf{x}_{\text{im}}))]$ by some basic algebra; (b) we then establish an upper bound on the difference of the simplified expressions.

Simplifying the expressions. By definition, we have

$$\begin{aligned}
& \bar{R}_{\text{clip,im},K}(\mathbf{S}) - \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star) \\
&= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k}} \left[-\log \frac{\exp(\mathbf{S}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}))}{\sum_{j \in [K]} \exp(\mathbf{S}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}))} + \log \frac{\exp(\mathbf{S}_\star(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}))}{\sum_{j \in [K]} \exp(\mathbf{S}_\star(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}))} \right] \\
&= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k}} \left[\log \frac{\exp(\mathbf{S}_\star(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}))}{\exp(\mathbf{S}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}))} - \log \frac{\sum_{j \in [K]} \exp(\mathbf{S}_\star(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}))}{\sum_{j \in [K]} \exp(\mathbf{S}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}))} \right] \\
&= T_{a1} - T_{a2},
\end{aligned}$$

where

$$\begin{aligned}
T_{a1} &:= \mathbb{E}_{(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}}) \sim \mathbb{P}_{\text{im,tx}}} [\mathbf{S}_\star(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}}) - \mathbf{S}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}})], \\
T_{a2} &:= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k}} \left[\log \frac{\sum_{j \in [K]} \exp(\mathbf{S}_\star(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}))}{\sum_{j \in [K]} \exp(\mathbf{S}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}))} \right].
\end{aligned}$$

Similarly,

$$\begin{aligned}
E_{\mathbf{x}_{\text{im}}} \left[\text{D}_{\text{KL}} \left(\mathbb{P}(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}}) \middle| \middle| \widehat{\mathbb{P}}_{\text{S}}(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}}) \right) \right] &= \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im,tx}}} \left[\log \frac{\mathbb{P}(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}})}{\widehat{\mathbb{P}}_{\text{S}}(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}})} \right] \\
&= \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im,tx}}} \left[\log \frac{\frac{\exp(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) \mathbb{P}(\mathbf{x}_{\text{tx}})}{\sum_{\mathbf{x}_{\text{tx}}'} \exp(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}')) \mathbb{P}(\mathbf{x}_{\text{tx}}')}}{\frac{\exp(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) \mathbb{P}(\mathbf{x}_{\text{tx}})}{\sum_{\mathbf{x}_{\text{tx}}'} \exp(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}')) \mathbb{P}(\mathbf{x}_{\text{tx}}')}}} \right] \\
&= T_{b1} - T_{b2},
\end{aligned}$$

where

$$\begin{aligned}
T_{b1} &= \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im,tx}}} [\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})], \\
T_{b2} &= \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[\log \frac{\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}{\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))} \right].
\end{aligned}$$

Since $T_{a1} = T_{b1}$, it suffices to bound the difference $|T_{a2} - T_{b2}|$.

Bounding $|T_{a2} - T_{b2}|$. Without loss of generality, we assume $\mathbf{S}, \mathbf{S}_\star$ are chosen such that the conditional mean

$$\mathbb{E}_{\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{tx}|\text{im}}} [\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})] = \mathbb{E}_{\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{tx}|\text{im}}} [\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})] = 0$$

for all $\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}$. Note that this can be done by subtracting the conditional mean (which is a function of $\mathbf{x}_{\text{im}}$) from $\mathbf{S}$ (or $\mathbf{S}_\star$). In this case, Assumption 2.1 still holds but with a different constant $c'_1 > 1$ that is polynomially dependent on c_1 .

For any function $\tilde{h} : \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}} \mapsto \mathbb{R}$, we define its norm

$$\|\tilde{h}\| := \sqrt{\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}} [\tilde{h}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})]^2}.$$

In addition, for any score function $\tilde{\mathbf{S}}$, we introduce the distributions

$$\begin{aligned} p_{\tilde{\mathbf{S}}}(k | \bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx}, j})_{j \in [K]}) &:= \frac{\exp(\tilde{\mathbf{S}}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}, k}))}{\sum_{j \in [K]} \exp(\tilde{\mathbf{S}}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}, j}))} \quad \text{for all } k \in [K], \quad \text{and recall that} \\ \hat{\mathbb{P}}_{\tilde{\mathbf{S}}}(\mathbf{x}_{\text{tx}} | \mathbf{x}_{\text{im}}) &:= \frac{\exp(\tilde{\mathbf{S}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) \mathbb{P}(\mathbf{x}_{\text{tx}})}{\mathbb{E}_{\mathbf{x}_{\text{tx}}' \sim \mathbb{P}_{\text{tx}}} \exp(\tilde{\mathbf{S}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}'))} \quad \text{for all } \mathbf{x}_{\text{tx}} \in \mathcal{X}_{\text{tx}}. \end{aligned}$$

Note that $p_{\mathbf{S}_\star}$ is the posterior distribution of $\bar{k}$ conditioned on $\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx}, j})_{j \in [K]}$ as shown in equation (A.13) in the proof of Lemma 2.1; moreover, we have $\hat{\mathbb{P}}_{\mathbf{S}_\star} = \mathbb{P}_{\text{tx} | \text{im}}$.

We begin by claiming that

$$\|\mathbf{S} - \mathbf{S}_\star\| \leq C \cdot \sqrt{\bar{\mathbf{R}}_{\text{clip}, \text{im}, K}(\mathbf{S}) - \bar{\mathbf{R}}_{\text{clip}, \text{im}, K}(\mathbf{S}_\star)} \quad (\text{A.15})$$

for some constant $C > 0$. The proof of this claim can be found in Lemma A.5. Moreover, we argue that

$$|T_{a2} - T_{b2}| \leq \frac{C}{K} \cdot \|\mathbf{S} - \mathbf{S}_\star\|^2. \quad (\text{A.16})$$

Combining equation (A.15) and (A.16) yields the desired result.

Therefore, it remains to establish equation (A.16). Write $\mathbf{S} = \mathbf{S}_\star + rh$ with $r = \|\mathbf{S} - \mathbf{S}_\star\|$ and $h = (\mathbf{S} - \mathbf{S}_\star) / \|\mathbf{S} - \mathbf{S}_\star\|$, and define

$$T(r) := \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx}, j})_{j \in [K]}, \bar{k}} \left[\log \frac{\sum_{j \in [K]} \exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}, j})) / K}{\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx}}))} \right],$$

where $\mathbf{S}_r = \mathbf{S}_\star + rh$. Then we have $|T_{a2} - T_{b2}| = |T(\|\mathbf{S} - \mathbf{S}_\star\|) - T(0)|$. Performing a second-order Taylor expansion on $T(r)$ w.r.t. r at $r = 0$, and noting that $r = 0$ is a stationary point, we obtain

$$\begin{aligned} |T_{a2} - T_{b2}| &= \left| \left\{ \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx}, j})_{j \in [K]}} \left[\text{Var}_{k \sim p_{\mathbf{S}_\star + \tilde{r}h}} \left[\frac{\mathbf{S}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}, k}) - \mathbf{S}_\star(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}, k})}{\|\mathbf{S} - \mathbf{S}_\star\|} \right] \right] \right. \right. \\ &\quad \left. \left. - \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \text{Var}_{\mathbf{x}_{\text{tx}} \sim \hat{\mathbb{P}}_{\mathbf{S}_\star + \tilde{r}h}} \left[\frac{\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})}{\|\mathbf{S} - \mathbf{S}_\star\|} \right] \right\} \cdot \|\mathbf{S} - \mathbf{S}_\star\|^2 \right| \\ &= |T_{d2}(\tilde{r})| \end{aligned}$$

for some $\tilde{r} \in [0, \|\mathbf{S} - \mathbf{S}_\star\|]$, where

$$T_{d2}(r) := \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx}, j})_{j \in [K]}} \left[\text{Var}_{k \sim p_{\mathbf{S}_\star + rh}} \left[\mathbf{S}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}, k}) - \mathbf{S}_\star(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx}, k}) \right] \right] -$$

$$\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[\text{Var}_{\mathbf{x}_{\text{tx}} \sim \hat{\mathbb{P}}_{\mathbf{S}_* + rh}} \left[\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \mathbf{S}_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \right] \right]. \quad (\text{A.17})$$

Equation (A.16) then follows immediately from Lemma A.6, which states that

$$|T_{d2}(r)| \leq C \cdot \|\mathbf{S} - \mathbf{S}_*\|^2 / K \quad (\text{A.18})$$

for some constant $C > 0$ for all $r \in [0, \|\mathbf{S} - \mathbf{S}_*\|]$. □

A.4.3 An alternative to Proposition 2.3

An alternative additive error bound on the sufficiency $\text{Suff}(\mathbf{S})$ can be established without the boundedness conditions in Assumption 2.1. To this end, we introduce the following weaker assumption.

Assumption A.1 (Bounded expected score). *There exists some constant $c_1 > 0$ such that*

$$\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathcal{P}} [\exp(4A(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))] \leq c_1, \text{ for any } A \in \{\pm \mathbf{S}, \pm \mathbf{S}_*\} \text{ and } \mathcal{P} \in \{\mathbb{P}_{\text{im,tx}}, \mathbb{P}_{\text{im}} \times \mathbb{P}_{\text{tx}}\},$$

where $\mathbf{S}_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) := \log \frac{\mathbb{P}_{\text{im,tx}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})}{\mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}}) \mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx}})}$.

Note that Assumption A.1 is implied by Assumption 2.1. Under this condition, we have the following result.

Proposition A.3. *Under Assumption A.1 and the notations in Proposition 2.3, for any $K \geq 2$, we have*

$$\lim_{K' \rightarrow \infty} \left[\bar{\mathbf{R}}_{\text{clip}, K'}(\mathbf{S}) - \bar{\mathbf{R}}_{\text{clip}, K'}(\mathbf{S}_*) \right] = \text{Suff}(\mathbf{S}) \leq \underbrace{\left[\bar{\mathbf{R}}_{\text{clip}, K}(\mathbf{S}) - \bar{\mathbf{R}}_{\text{clip}, K}(\mathbf{S}_*) \right]}_{\text{CLIP excess risk}} + \frac{C}{K} \quad (\text{A.19})$$

for some constant $C > 0$ depending polynomially on c_1 .

As a consequence, the similarity score function $\mathbf{S}$ is near-sufficient when the CLIP excess risk is small and the batch size K is sufficiently large. ¹

Proof of Proposition A.3. Throughout the proof, we use $C > 0$ to denote constants that depends polynomially on c_1 in Assumption A.1. We allow the value of C to vary from place to place. Following the proof of Proposition 2.3 in Section A.4.2 and the notations therein, we have

$$\left| \left[\bar{\mathbf{R}}_{\text{clip, im}, K}(\mathbf{S}) - \bar{\mathbf{R}}_{\text{clip, im}, K}(\mathbf{S}_*) \right] - \text{Suff}(\mathbf{S}) \right| \leq |T_{a2} - T_{b2}|, \quad (\text{A.20})$$

¹In practice, a large batch size K (e.g., $K = 32768$) is often used to improve the performance of CLIP (Radford et al., 2021).

where

$$T_{a2} = \mathbb{E}_{\mathbf{x}_{\text{im}}, (\mathbf{x}_{\text{tx}, j})_{j \in [K]}} \left[\log \frac{\sum_{j \in [K]} \exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}, j}))}{\sum_{j \in [K]} \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}, j}))} \right],$$

$$T_{b2} = \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[\log \frac{\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}{\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))} \right].$$

It suffices to show that

$$|T_{a2} - T_{b2}| \leq C/K \quad (\text{A.21})$$

for some constant $C > 0$ depending polynomially on c_1 . By some basic algebra, we have

$$|T_{a2} - T_{b2}| \leq |T_3| + |T_4|,$$

where

$$T_3 = \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}, \{\mathbf{x}_{\text{tx}, j}\}_{j=1}^{K-1} \sim \mathbb{P}_{\text{tx}}} \left[\log \frac{\exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) + \sum_{j \in [K-1]} \exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}, j}))}{K \cdot \mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))} \right],$$

$$T_4 = \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}, \{\mathbf{x}_{\text{tx}, j}\}_{j=1}^{K-1} \sim \mathbb{P}_{\text{tx}}} \left[\log \frac{\exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) + \sum_{j \in [K-1]} \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}, j}))}{K \cdot \mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))} \right].$$

We will show below that $|T_3| \leq C/K$ for some constant $C > 0$. The same bound holds for $|T_4|$ following the same argument. Thus, combining the bounds yields Eq. (A.21).

An upper bound on T_3 . Note that

$$\begin{aligned} T_3 &\stackrel{(i)}{\leq} \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}, \{\mathbf{x}_{\text{tx}, j}\}_{j=1}^{K-1} \sim \mathbb{P}_{\text{tx}}} \left[\frac{1}{\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))} \right. \\ &\quad \left. \left[\frac{1}{K} \sum_{j \in [K-1]} \exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}, j})) + \frac{\exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}{K} - \mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) \right] \right] \\ &= \frac{1}{K} \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}} \left[\frac{\exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}{\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))} \right] - \frac{1}{K}, \end{aligned}$$

where step (i) uses $\log(1+a) \leq a$ for any $a > -1$. By Cauchy-Schwarz inequality and Jensen's inequality, we further have

$$\begin{aligned} &\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}} \left[\frac{\exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}{\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))} \right] \\ &\leq \sqrt{\mathbb{E}_{\mathbb{P}_{\text{im}, \text{tx}}} \exp(2S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))} \cdot \sqrt{\mathbb{E}_{\mathbb{P}_{\text{im}}} \frac{1}{(\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})))^2}} \end{aligned}$$

$$\begin{aligned}
&\leq \sqrt{\mathbb{E}_{\mathbb{P}_{\text{im}, \text{tx}}} \exp(2S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))} \cdot \sqrt{\mathbb{E}_{\mathbb{P}_{\text{im}} \times \mathbb{P}_{\text{tx}}} \frac{1}{\exp(2S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}} \\
&\leq C,
\end{aligned}$$

where the last inequality follows from Assumption A.1. Putting pieces together yields $T_3 \leq C/K$.

An lower bound on T_3 . On the other hand, we have the lower bound

$$\begin{aligned}
&T_3 \\
&\stackrel{(ii)}{\geq} \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}, \{\mathbf{x}_{\text{tx}, j}\}_{j=1}^{K-1} \sim \mathbb{P}_{\text{tx}}} \left[\frac{1}{\sum_{j \in [K-1]} \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}, j}))/K + \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))/K} \right. \\
&\quad \left. \left[\frac{1}{K} \sum_{j \in [K-1]} \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}, j})) + \frac{\exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}{K} - \mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) \right] \right] \\
&=: T_5 + T_6,
\end{aligned}$$

where step (ii) uses $\log(1+a) \geq a/(a+1)$ for any $a > -1$, and

$$\begin{aligned}
T_5 &:= \frac{1}{K} \left[\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}} \left[\frac{\exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}{\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))} \right] - 1 \right], \\
T_6 &:= -\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}, \{\mathbf{x}_{\text{tx}, j}\}_{j=1}^{K-1} \sim \mathbb{P}_{\text{tx}}} \left[\frac{1}{\frac{\sum_{j \in [K-1]} \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}, j}))}{K} + \frac{\exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}{K}} \right. \\
&\quad \left. \frac{\left[\frac{\sum_{j \in [K-1]} \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}, j}))}{K} + \frac{\exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}{K} - \mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) \right]^2}{\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))} \right].
\end{aligned}$$

Note that $T_5 \geq -1/T$. To prove that $T_3 \geq -C/K$, it suffices to show that $|T_6| \leq C/K$ for some c_1 -dependent constant $C > 0$. By Cauchy-Schwarz inequality, we have

$$|T_6| \leq T_{6x} \cdot T_{6y} \cdot T_{6z},$$

where

$$\begin{aligned}
T_{6x} &:= \sqrt{\mathbb{E} \left[\left[\frac{\sum_{j \in [K-1]} \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}, j}))}{K} + \frac{\exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}{K} - \mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) \right]^4 \right]}, \\
T_{6y} &:= \left(\mathbb{E} \left[\left[\frac{\sum_{j \in [K-1]} \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}, j}))}{K} + \frac{\exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}{K} \right]^{-4} \right] \right)^{1/4}, \\
T_{6z} &:= \left(\mathbb{E} \left[\left[\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_*(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) \right]^{-4} \right] \right)^{1/4}.
\end{aligned}$$

Applying Jensen's inequality on T_{6y}, T_{6z} and using Assumption A.1, it is readily verified that $T_{6y}, T_{6z} \leq C$ for some constant $C > 0$. For T_{6x} , introduce the shorthand $\Delta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \exp(S_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) - \mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(S_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))$. Then we have $\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} [\Delta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})] = 0$ and

$$\begin{aligned} T_{6x} &\leq c \left(\sqrt{\mathbb{E} \left[\frac{\sum_{j \in [K-1]} \Delta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},j})}{K} \right]^4} + \frac{1}{K^2} \sqrt{\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}} [\Delta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})^4]} \right) \\ &\leq c \left(\frac{1}{K^2} \sqrt{\mathbb{E} \left[\sum_{i,j \in [K-1]} \Delta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},i})^2 \Delta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},j})^2 \right]} + \frac{C}{K^2} \right) \\ &\leq c \left(\frac{1}{K} \sqrt{\mathbb{E} \left[\frac{1}{K} \sum_{i \in [K-1]} \Delta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},i})^4 \right]} + \frac{C}{K^2} \right) \leq C/K \end{aligned}$$

for some universal constant $c > 0$, where the first line uses the fact that $\sqrt{\mathbb{E}|X+Y|^4} \leq c(\sqrt{\mathbb{E}|X|^4} + \sqrt{\mathbb{E}|Y|^4})$ for some universal constant $c > 0$, the second line follows Assumption A.1, the conditional independence of $\{\mathbf{x}_{\text{tx},j}\}_{j=1}^{K-1}$ given $\mathbf{x}_{\text{im}}$, and the property that $\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} [\Delta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})] = 0$. The last line uses Jensen's inequality and Assumption A.1. Combining the bounds on T_{6x}, T_{6y}, T_{6z} yields $|T_6| \leq C/K$ and therefore $T_3 \geq -C/K$. $\square$

A.4.4 Proof of Proposition 2.4

Proof of Proposition 2.4. Recall the conditional probabilities

$$\begin{aligned} \widehat{\mathbb{P}}_S(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}}) &= \frac{\exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))\mathbb{P}(\mathbf{x}_{\text{tx}})}{\sum_{\mathbf{x}_{\text{tx}}} \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}'))\mathbb{P}(\mathbf{x}_{\text{tx}}')}, \\ \widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}}) &:= \frac{\sum_{j=1}^M \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}^{(j)}(y)))\mathbb{P}(y)}{\sum_{y \in \mathcal{Y}} \sum_{j=1}^M \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}^{(j)}(y)))\mathbb{P}(y)}, \end{aligned}$$

and define the infinite-sample probability

$$\widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}}) := \frac{\sum_{\mathbf{x}_{\text{tx}}} \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}'))\mathbb{P}(\mathbf{x}_{\text{tx}}', y)}{\sum_{\mathbf{x}_{\text{tx}}} \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}'))\mathbb{P}(\mathbf{x}_{\text{tx}}')}.$$

We will prove that²

$$\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\mathbb{P}_{\text{cls}|\text{im}}(y|\mathbf{x}_{\text{im}}) \middle| \middle| \widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}}) \right) \right] \leq \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\mathbb{P}(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}}) \middle| \middle| \widehat{\mathbb{P}}_S(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}}) \right) \right], \text{ and} \quad (\text{A.22a})$$

$$\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_2 \left(\widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}}) \middle| \middle| \widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}}) \right) \right] \leq C \frac{\log(2/\delta)}{M} \text{ with probability at least } 1 - \delta, \quad (\text{A.22b})$$

²We abuse the notation $\mathbb{P}(\cdot)$ for $\mathbb{P}_{\text{cls}|\text{im}}(\cdot), \mathbb{P}_{\text{tx}|\text{im}}(\cdot)$ when it is clear from the context.

where we define the α -Rényi divergence

$$D_\alpha\left(\widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}})\left|\left|\widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})\right.\right.\right) := \frac{1}{\alpha-1} \log \left(\mathbb{E}_{y \sim \widehat{\mathbb{P}}_S(\cdot|\mathbf{x}_{\text{im}})} \left[\left(\frac{\widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}})}{\widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})} \right)^{\alpha-1} \right] \right)$$

for any $\alpha > 1$.

Given these results, Proposition 2.4 follows from Proposition 2.3, combined with a triangle-like inequality for KL divergence (see e.g., Lemma 26 of [Bun and Steinke 2016](#)), which states that for any tuple of distributions $(\mathbb{P}, \mathbb{Q}, \mathbb{R})$,

$$D_{\text{KL}}(\mathbb{P} \parallel \mathbb{R}) \leq D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q}) + D_2(\mathbb{Q} \parallel \mathbb{R}).$$

Proof of bound (A.22a). Observe that

$$\begin{aligned} \mathbb{P}(y|\mathbf{x}_{\text{im}}) &= \sum_{\mathbf{x}_{\text{tx}}'} \mathbb{P}(y|\mathbf{x}_{\text{tx}}', \mathbf{x}_{\text{im}}) \cdot \mathbb{P}(\mathbf{x}_{\text{tx}}'|\mathbf{x}_{\text{im}}) \stackrel{(i)}{=} \sum_{\mathbf{x}_{\text{tx}}'} \mathbb{P}(y|\mathbf{x}_{\text{tx}}') \cdot \mathbb{P}(\mathbf{x}_{\text{tx}}'|\mathbf{x}_{\text{im}}), \\ \widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}}) &\stackrel{(ii)}{=} \sum_{\mathbf{x}_{\text{tx}}'} \mathbb{P}(y|\mathbf{x}_{\text{tx}}') \cdot \widehat{\mathbb{P}}_S(\mathbf{x}_{\text{tx}}'|\mathbf{x}_{\text{im}}), \end{aligned}$$

where step (i) follows from Assumption 2.2 and step (ii) uses the definitions of $\widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}})$ and $\widehat{\mathbb{P}}_S(\mathbf{x}_{\text{tx}}'|\mathbf{x}_{\text{im}})$. Therefore, it follows from the data-processing inequality that

$$D_{\text{KL}}\left(\mathbb{P}(y|\mathbf{x}_{\text{im}})\left|\left|\widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}})\right.\right.\right) \leq D_{\text{KL}}\left(\mathbb{P}(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}})\left|\left|\widehat{\mathbb{P}}_S(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}})\right.\right.\right), \quad \text{for all } \mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}.$$

Taking expectation over $\mathbf{x}_{\text{im}}$ yields bound (A.22a).

Proof of bound (A.22b). To prove bound (A.22b), a key component is to establish

$$\mathbb{E}_{\mathbf{x}_{\text{im}}} \left[\sum_{y \in \mathcal{Y}} \left[\frac{|\widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}}) - \widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})|^2}{\widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})} \right] \right] \leq C \cdot \frac{\log(2/\delta)}{M} \quad (\text{A.23})$$

with probability at least $1 - \delta$ for some constant $C > 0$ polynomially depending on c_1 in Assumption 2.1. We will prove this at the end of the section. Using claim (A.23), we have

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_{\text{im}}} \left[D_2\left(\widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}})\left|\left|\widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})\right.\right.\right) \right] &= \mathbb{E}_{\mathbf{x}_{\text{im}}} \left[\log \mathbb{E}_{y \sim \widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}})} \left[\frac{\widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}})}{\widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})} \right] \right] \\ &\leq \mathbb{E}_{\mathbf{x}_{\text{im}}} \mathbb{E}_{y \sim \widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}})} \left[\frac{\widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}}) - \widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})}{\widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})} \right] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_{\mathbf{x}_{\text{im}}} \left[\sum_{y \in \mathcal{Y}} \left[\frac{|\hat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}}) - \hat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})|^2}{\hat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})} \right] \right] \\
&\leq C \cdot \frac{\log(2/\delta)}{M}
\end{aligned} \tag{A.24}$$

with probability at least $1 - \delta$. This concludes the proof of bound (A.22b).

Now, it remains to establish bound (A.23). By properties of sub-exponential variables, it suffices to show the Orlicz norm (see e.g., [Vershynin, 2018](#); [Wainwright, 2019b](#))

$$\left\| \mathbb{E}_{\mathbf{x}_{\text{im}}} \left[\sum_{y \in \mathcal{Y}} \left[\frac{|\hat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}}) - \hat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})|^2}{\hat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})} \right] \right] \right\|_{\psi_1} \leq \frac{C}{M} \tag{A.25}$$

for some constant $C > 0$.

Define the quantities

$$\begin{aligned}
\mathcal{R}_1(y) &:= \sum_{\mathbf{x}_{\text{tx}}'} \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}')) \mathbb{P}(\mathbf{x}_{\text{tx}}', y), & \mathcal{R}_2 &:= \sum_{\mathbf{x}_{\text{tx}}'} \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}')) \mathbb{P}(\mathbf{x}_{\text{tx}}'), \\
\mathcal{R}_3(y) &:= \frac{1}{M} \sum_{j=1}^M \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}^{(j)}(y))) \mathbb{P}(y), & \mathcal{R}_4 &:= \sum_{y \in \mathcal{Y}} \frac{1}{M} \sum_{j=1}^M \exp(S(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}^{(j)}(y))) \mathbb{P}(y).
\end{aligned}$$

Then $\hat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}}) = \mathcal{R}_1(y)/\mathcal{R}_2$, $\hat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}}) = \mathcal{R}_3(y)/\mathcal{R}_4$ and

$$\begin{aligned}
&\frac{|\hat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}}) - \hat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})|^2}{\hat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})} \\
&= \frac{(\mathcal{R}_1(y)\mathcal{R}_4 - \mathcal{R}_2\mathcal{R}_3(y))^2}{\mathcal{R}_2^2\mathcal{R}_3(y)\mathcal{R}_4} \leq \frac{2[(\mathcal{R}_1(y) - \mathcal{R}_3(y))\mathcal{R}_4]^2 + 2[(\mathcal{R}_4 - \mathcal{R}_2)\mathcal{R}_3(y)]^2}{\mathcal{R}_2^2\mathcal{R}_3(y)\mathcal{R}_4}.
\end{aligned} \tag{A.26}$$

By Assumption 2.1 and concentration properties of bounded random variables, there exists constant $C > 0$ such that the Orlicz norm

$$\|\mathcal{R}_3(y) - \mathcal{R}_1(y)\|_{\psi_2} \leq C \cdot \frac{\mathbb{P}(y)}{\sqrt{M}} \tag{A.27a}$$

for all $y \in \mathcal{Y}$, $\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}$. Summing over $y \in \mathcal{Y}$ and using the triangle inequality, we obtain

$$\|\mathcal{R}_4 - \mathcal{R}_2\|_{\psi_2} \leq \frac{C}{\sqrt{M}}. \tag{A.27b}$$

Moreover, Assumption 2.1 implies that

$$\mathcal{R}_2, \mathcal{R}_4 \in [1/C, C], \quad \text{and} \quad \mathcal{R}_1(y), \mathcal{R}_3(y) \in \left[\frac{\mathbb{P}(y)}{C}, C \cdot \mathbb{P}(y) \right] \tag{A.27c}$$

for all $\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}, y \in \mathcal{Y}$ for some constant $C > 0$.

Substituting equation (A.27a), (A.27b) and (A.27c) into equation (A.26) and using properties of the Orlicz norm, we find

$$\left\| \frac{|\widehat{\mathbb{P}}_S(y|\mathbf{x}_{\text{im}}) - \widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})|^2}{\widehat{\mathbb{P}}_S^{(M)}(y|\mathbf{x}_{\text{im}})} \right\|_{\psi_1} \leq C \cdot \frac{\mathbb{P}(y)}{M} \quad (\text{A.28})$$

for all $\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}, y \in \mathcal{Y}$ for some constant $C > 0$. Finally, summing equation (A.28) over $y \in \mathcal{Y}$ and invoking Jensen's inequality yields equation (A.25). $\square$

A.4.5 Proof of Proposition 2.5

Proof of Proposition 2.5. Recall that $\mathbf{z}_t = t \cdot \mathbf{x}_{\text{im}} + \sqrt{t} \cdot \mathbf{g}$. By definition,

$$\begin{aligned} & \mathbb{E}_{(\mathbf{x}_{\text{tx}}, \mathbf{z}_t)} \left[\left\| \mathbf{m}_t(\mathbf{z}_t, \mathbf{x}_{\text{tx}}) - \widehat{\mathbf{M}}_t(\mathbf{z}_t, \mathbb{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) \right\|_2^2 \right] \\ &= \mathbb{E}_{(\mathbf{x}_{\text{tx}}, \mathbf{z}_t)} \left[\left\| \mathbb{E}[\mathbf{x}_{\text{im}}|\mathbf{z}_t, \mathbf{x}_{\text{tx}}] - \mathbb{E}[\mathbf{x}_{\text{im}}|\mathbf{z}_t, \mathbb{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})] \right\|_2^2 \right] \\ &\stackrel{(i)}{\leq} 4d_{\text{im}} B_{x_{\text{im}}}^2 \cdot \mathbb{E}_{(\mathbf{x}_{\text{tx}}, \mathbf{z}_t)} [\text{D}_{\text{TV}}(\mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}}, \mathbf{z}_t), \mathbb{P}(\mathbf{x}_{\text{im}}|\mathbb{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}), \mathbf{z}_t))] \\ &\stackrel{(ii)}{\leq} 2d_{\text{im}} B_{x_{\text{im}}}^2 \cdot \mathbb{E}_{(\mathbf{x}_{\text{tx}}, \mathbf{z}_t)} [\text{D}_{\text{KL}}(\mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}}, \mathbf{z}_t) \parallel \mathbb{P}(\mathbf{x}_{\text{im}}|\mathbb{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}), \mathbf{z}_t))] \\ &\stackrel{(iii)}{\leq} 2d_{\text{im}} B_{x_{\text{im}}}^2 \cdot \mathbb{E}_{\mathbf{x}_{\text{tx}}} [\text{D}_{\text{KL}}(\mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}}) \parallel \mathbb{P}(\mathbf{x}_{\text{im}}|\mathbb{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})))] = 2d_{\text{im}} B_{x_{\text{im}}}^2 \cdot \text{Suff}(\mathbb{E}_{\text{tx}}), \end{aligned}$$

where step (i) follows since

$$\begin{aligned} & \left\| \mathbb{E}[\mathbf{x}_{\text{im}}|\mathbf{z}_t, \mathbf{x}_{\text{tx}}] - \mathbb{E}[\mathbf{x}_{\text{im}}|\mathbf{z}_t, \mathbb{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})] \right\|_2^2 \\ &= \left\| \int \mathbf{x}_{\text{im}} [\mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{z}_t, \mathbf{x}_{\text{tx}}) - \mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{z}_t, \mathbb{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))] d\mathbf{x}_{\text{im}} \right\|_2^2 \\ &\leq \left(\int \|\mathbf{x}_{\text{im}}\|_2 \cdot |\mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{z}_t, \mathbf{x}_{\text{tx}}) - \mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{z}_t, \mathbb{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))| d\mathbf{x}_{\text{im}} \right)^2 \\ &\leq 4d_{\text{im}} B_{x_{\text{im}}}^2 \cdot \text{D}_{\text{TV}}^2(\mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{z}_t, \mathbf{x}_{\text{tx}}), \mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{z}_t, \mathbb{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))), \end{aligned}$$

step (ii) uses Pinsker's inequality, step (iii) uses the data processing inequality. $\square$

Proof of Corollary 2.6

Consider the process

$$\widetilde{\mathbf{Y}}_t = t \cdot \mathbf{x}_{\text{im}} + d\mathbf{W}_t$$

where $\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}|\text{tx}}(\cdot|\mathbf{x}_{\text{tx}})$ and $(\mathbf{W}_t)_{t \geq 0}$ is the Brownian motion on $\mathbb{R}^{d_{\text{im}}}$. Since $\|\mathbf{x}_{\text{im}}\|_2 \leq \sqrt{d_{\text{im}}} \times B_{x_{\text{im}}}$, it follows from Proposition 1 in Montanari (2023) that $(\tilde{\mathbf{Y}}_t)_{t \geq 0}$ is the unique solution to the stochastic differential equation (assuming $\tilde{\mathbf{Y}}_0 = \mathbf{0}$)

$$d\tilde{\mathbf{Y}}_t = \mathbf{m}_t(\tilde{\mathbf{Y}}_t, \mathbf{x}_{\text{tx}})dt + d\mathbf{W}_t, \quad t \geq 0,$$

where $\mathbf{m}_t(\mathbf{z}, \mathbf{x}_{\text{tx}}) = \mathbb{E}[\mathbf{x}_{\text{im}}|\mathbf{z} = t\mathbf{x}_{\text{im}} + \sqrt{t}\mathbf{g}, \mathbf{x}_{\text{tx}}]$. Let $\tilde{\mathbb{P}}_{\text{im}|\text{tx}}^{(T)} = \tilde{\mathbb{P}}_{\text{im}|\text{tx}}^{(T)}(\cdot|\mathbf{x}_{\text{tx}})$ denote the distribution of $\tilde{\mathbf{Y}}_T/T$. It follows immediately that $\mathbb{P}_{\text{im}|\text{tx}}^{\square \frac{1}{\sqrt{T}}} = \tilde{\mathbb{P}}_{\text{im}|\text{tx}}^{(T)}$. Therefore,

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_{\text{tx}}} D_{\text{KL}}(\mathbb{P}_{\text{im}|\text{tx}}^{\square \frac{1}{\sqrt{T}}}(\cdot|\mathbf{x}_{\text{tx}}) || \tilde{\mathbb{P}}_{\text{im}|\text{tx}}^{(T)}(\cdot|\mathbf{x}_{\text{tx}})) &\stackrel{(i)}{=} \mathbb{E}_{\mathbf{x}_{\text{tx}}} D_{\text{KL}}(\mathbb{P}_{\tilde{\mathbf{Y}}_T} || \mathbb{P}_{\mathbf{Y}_T}) \\ &\stackrel{(ii)}{\leq} \frac{1}{2} \mathbb{E}_{\mathbf{x}_{\text{tx}}} \int_0^T \mathbb{E}_{\tilde{\mathbf{Y}}_t} \|\mathbf{m}_t(\tilde{\mathbf{Y}}_t, \mathbf{x}_{\text{tx}}) - \hat{\mathbf{M}}_t(\tilde{\mathbf{Y}}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))\|_2^2 dt, \stackrel{(iii)}{\leq} d_{\text{im}} B_{x_{\text{im}}}^2 T \cdot \text{Suff}(\mathbf{E}_{\text{tx}}), \end{aligned}$$

where step (i) uses the scale invariance of KL divergence, step (ii) uses Girsanov theorem (Lemma A.7), and step (iii) follows from Proposition 2.5 and the fact that $(\mathbf{x}_{\text{tx}}, \tilde{\mathbf{Y}}_t) \stackrel{d}{=} (\mathbf{x}_{\text{tx}}, \mathbf{z}_t)$, where $\mathbf{z}_t = t \cdot \mathbf{x}_{\text{im}} + \sqrt{t} \cdot \mathbf{g}$.

A.4.6 Proof of Proposition A.1

We claim that the minimizer $\hat{\mu}$ in (A.4) is

$$\hat{\mu}(\cdot | \mathbf{E}, \square) = \mathbb{P}_{\text{im}, \text{tx}}(x_{\text{tx}, i} = \cdot | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}) = \mathbf{E}, x_{\text{tx}, 1:i-1} = \square). \quad (\text{A.29})$$

By the tensorization property of KL divergence and the expression of μ_* , $\hat{\mu}$ in Eq. (A.3) and (A.29), we have

$$\begin{aligned} D(\mu_*, \hat{\mu}) &= \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}} \left[\sum_{i \in [d_{\text{tx}}]} D_{\text{KL}} \left(\mu_*(x_{\text{tx}, i} | \mathbf{x}_{\text{im}}, x_{\text{tx}, 1:i-1}) || \hat{\mu}(x_{\text{tx}, i} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), x_{\text{tx}, 1:i-1}) \right) \right] \\ &= \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\prod_{i=1}^{d_{\text{tx}}} \mu_*(x_{\text{tx}, i} | \mathbf{x}_{\text{im}}, x_{\text{tx}, 1:i-1}) || \prod_{i=1}^{d_{\text{tx}}} \hat{\mu}(x_{\text{tx}, i} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), x_{\text{tx}, 1:i-1}) \right) \right] \\ &= \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} [D_{\text{KL}}(\mathbb{P}_{\text{im}|\text{tx}}(\cdot|\mathbf{x}_{\text{im}}) || \mathbb{P}_{\text{im}|\text{tx}}(\cdot|\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})))] = \text{Suff}(\mathbf{E}_{\text{im}}). \end{aligned}$$

It remains to establish Eq. (A.29). Note that this follows immediately from the fact that

$$\begin{aligned} \arg \min_{\mu \in \mathcal{U}} \left\{ R_{\text{vIm}}(\mu, \mathbf{E}_{\text{im}}) &:= \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}} \left[\sum_{i \in [d_{\text{tx}}]} -\log \mu(x_{\text{tx}, i} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), x_{\text{tx}, 1:i-1}) \right] \right\} \\ &= \arg \min_{\mu \in \mathcal{U}} \left\{ R_{\text{vIm}}(\mu, \mathbf{E}_{\text{im}}) := \right. \\ &\quad \left. \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}} \left[\sum_{i \in [d_{\text{tx}}]} D_{\text{KL}}(\mathbb{P}(x_{\text{tx}, i} | \mathbf{x}_{\text{im}}, x_{\text{tx}, 1:i-1}) || \mu(x_{\text{tx}, i} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), x_{\text{tx}, 1:i-1})) \right] \right\}, \end{aligned} \quad (\text{A.30})$$

and for each $i \in [d_{\text{tx}}]$, the KL divergence in (A.30) is minimized when

$$\mu(x_{\text{tx},i} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), x_{\text{tx},1:i-1}) = \mathbb{P}_{\text{im},\text{tx}}(x_{\text{tx},i} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), x_{\text{tx},1:i-1}) \quad \text{for all } x_{\text{tx},i} \in \mathcal{X}_{\text{tx},i}.$$

A.4.7 Proof of Proposition 2.9

Define $\mathbf{E} := (\mathbb{P}(\mathbf{x}_{\text{im}}) \mathbf{E}_{\text{im},\star}^\top(\mathbf{x}_{\text{im}}))_{\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}} \in \mathbb{R}^{|\mathcal{X}_{\text{im}}| \times p_\star}$ and introduce the pseudoinverse

$$\mathbf{E}^\dagger := [(\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} [\mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}}) \mathbf{E}_{\text{im},\star}^\top(\mathbf{x}_{\text{im}})])^{-1} \mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}})]_{\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}} \in \mathbb{R}^{p_\star \times |\mathcal{X}_{\text{im}}|}.$$

It can be verified that $\mathbf{E}^\dagger \mathbf{E} = \mathbf{I}_{p_\star}$ and

$$\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \|\mathbf{E}^\dagger(\mathbf{x}_{\text{im}})\|_2^2 = \text{tr}((\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} [\mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}}) \mathbf{E}_{\text{im},\star}^\top(\mathbf{x}_{\text{im}})])^{-1}) \leq L_B^2 p_\star.$$

Define the embedding

$$\widetilde{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}) := \mathbf{E}^\dagger \text{diag}(\mathbb{P}(\mathbf{x}_{\text{im}})) (\Upsilon_\star^{-1}(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \log \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} [\exp(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))]))_{\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}}.$$

In the proof we bound the differences $\mathbb{E}_{\mathbf{x}_{\text{tx}}} [\|\widetilde{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}) - \mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}})\|_2^2]$ and $\mathbb{E}_{\mathbf{x}_{\text{tx}}} [\|\widehat{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}) - \widetilde{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}})\|_2^2]$, respectively. Namely, we will show that

$$\mathbb{E}_{\mathbf{x}_{\text{tx}}} [\|\widetilde{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}) - \mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}})\|_2^2] \leq CL_B^2 p_\star L_\Gamma^2 \cdot \mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} [\text{D}_{\text{KL}}(\mathbb{P}_{\text{im}|\text{tx}}(\cdot | \mathbf{x}_{\text{tx}}) \parallel \widehat{\mathbb{P}}_{\text{S}}(\cdot | \mathbf{x}_{\text{tx}}))], \quad (\text{A.31})$$

and there exist some parameters $(W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)})$ such that

$$\mathbb{E}_{\mathbf{x}_{\text{tx}}} [\|\widehat{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}) - \widetilde{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}})\|_2^2] \leq \frac{CL_B^2 p_\star L_\Gamma^2}{M}, \quad (\text{A.32})$$

and $\|W_{\text{ada}}^{(1)}\|_{\text{op}} \leq CB_{\text{Adap}}$, $\|W_{\text{ada}}^{(2)}\|_{\text{op}} \leq CL_B / \sqrt{M}$. Combining two bounds, we obtain

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}_{\text{tx}}} [\|\widehat{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}) - \mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}})\|_2^2] \\ & \leq C \cdot L_B^2 \cdot L_\Gamma^2 \cdot p_\star \cdot \left(\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} [\text{D}_{\text{KL}}(\mathbb{P}_{\text{im}|\text{tx}}(\cdot | \mathbf{x}_{\text{tx}}) \parallel \widehat{\mathbb{P}}_{\text{S}}(\cdot | \mathbf{x}_{\text{tx}}))] + \frac{1}{M} \right) \\ & \leq C \cdot L_B^2 \cdot L_\Gamma^2 \cdot p_\star \cdot (\text{Suff}(\mathbf{S}) + M^{-1}), \end{aligned} \quad (\text{A.33})$$

where the last line uses Definition 2.1.

Proof of Eq. (A.31). By definition, we have

$$(\Upsilon_\star^{-1}(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})))_{\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}} = (\langle \mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}}) \rangle)_{\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}} \in \mathbb{R}^{|\mathcal{X}_{\text{im}}| \times |\mathcal{X}_{\text{tx}}|}.$$

Multiplying both sides by $\mathbf{E}^\dagger \text{diag}(\mathbb{P}(\mathbf{x}_{\text{im}}))$, we find

$$\mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}}) = \mathbf{E}^\dagger \text{diag}(\mathbb{P}(\mathbf{x}_{\text{im}})) (\Upsilon_\star^{-1}(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})))_{\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}}$$

$$= \mathbf{E}^\dagger \text{diag}(\mathbb{P}(\mathbf{x}_{\text{im}})) (\Upsilon_\star^{-1}(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \log \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}}[\exp(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))])_{\mathbf{x}_{\text{im}}},$$

where the last line follows since $\mathbb{E}_{\mathbf{x}_{\text{im}}}[\exp(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))] = \sum_{\mathbf{x}_{\text{im}}} \mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}}) = 1$. Introduce the shorthand

$$T(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) := \mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \log \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}}[\exp(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))]$$

and let $\Delta^\Gamma(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) := \Upsilon_\star^{-1}(T(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))) - \Upsilon_\star^{-1}(T(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})))$. Therefore,

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_{\text{tx}}} [\|\widetilde{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}) - \mathbf{E}_{\text{tx}, \star}(\mathbf{x}_{\text{tx}})\|_2^2] &= \mathbb{E}_{\mathbf{x}_{\text{tx}}} [\|\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}}[\mathbf{E}^\dagger(\mathbf{x}_{\text{im}}) \Delta^\Gamma(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})]\|_2^2] \\ &\leq \mathbb{E}_{\mathbf{x}_{\text{tx}}} \left[\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \|\mathbf{E}^\dagger(\mathbf{x}_{\text{im}})\|_2^2 \cdot \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} |\Delta^\Gamma(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})|^2 \right] \\ &\leq \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} [\|\mathbf{E}^\dagger(\mathbf{x}_{\text{im}})\|_2^2] \cdot \mathbb{E}_{(\mathbf{x}_{\text{tx}}, \mathbf{x}_{\text{im}}) \sim \mathbb{P}_{\text{tx}} \times \mathbb{P}_{\text{im}}} [\Delta^\Gamma(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})^2] \\ &= L_{B\star}^2 \cdot \mathbb{E}_{(\mathbf{x}_{\text{tx}}, \mathbf{x}_{\text{im}}) \sim \mathbb{P}_{\text{tx}} \times \mathbb{P}_{\text{im}}} [\Delta^\Gamma(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})^2], \end{aligned} \quad (\text{A.34})$$

where the second line uses Cauchy-Schwarz inequality, and the last line follows from the assumption on $\mathbf{E}^\dagger$. Moreover,

$$\begin{aligned} &\mathbb{E}_{(\mathbf{x}_{\text{tx}}, \mathbf{x}_{\text{im}}) \sim \mathbb{P}_{\text{tx}} \times \mathbb{P}_{\text{im}}} [\Delta^\Gamma(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})^2] \\ &\leq L_\Gamma^2 \cdot \mathbb{E}_{\mathbf{x}_{\text{tx}}} \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[|T(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) - T(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))|^2 \right] \\ &\leq CL_\Gamma^2 \cdot \mathbb{E}_{\mathbf{x}_{\text{tx}}} \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}|\mathbf{x}_{\text{tx}}} \left[|T(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) - T(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))|^2 \right], \end{aligned} \quad (\text{A.35})$$

where the last line uses the fact that $\mathbb{P}(\mathbf{x}_{\text{im}})/\mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}}) \leq C$ for some $C > 0$ implied by Assumption 2.1. It remains to bound $\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}} \times \mathbb{P}_{\text{tx}}} [|T(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) - T(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))|^2]$.

Since adding any function of $\mathbf{x}_{\text{im}}$ does not change the value of $T(\mathbf{S})$, w.l.o.g., we assume

$$\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}(\cdot|\mathbf{x}_{\text{tx}})} [\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})] = \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}(\cdot|\mathbf{x}_{\text{tx}})} [\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})] = 0$$

and write $\mathbf{S} = \mathbf{S}_\star + r\mathbf{h}$ with $r = \|\mathbf{S} - \mathbf{S}_\star\|$ and $\mathbf{h} = (\mathbf{S} - \mathbf{S}_\star)/\|\mathbf{S} - \mathbf{S}_\star\|$, where

$$\|\mathbf{f}\| := \sqrt{\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}}} [\mathbf{f}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})^2]}$$

Similar to the proof of Proposition 2.3, by a Taylor expansion w.r.t. r at 0, we find

$$\begin{aligned} &\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})} \left[|T(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) - T(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))|^2 \right] \\ &\leq \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})} \left[\left| \mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \right. \right. \\ &\quad \left. \left. (\log \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}}[\exp(\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))] - \log \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}}[\exp(\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))]) \right|^2 \right] \\ &\leq 2\|\mathbf{S} - \mathbf{S}_\star\|^2 + 2\mathbb{E}_{\mathbf{x}_{\text{tx}}} \left[\left| \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \widehat{\mathbb{P}}_{\mathbf{S}_\star + \bar{r}\mathbf{h}}} (\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) \right|^2 \right] \end{aligned} \quad (\text{A.36})$$

for some $\tilde{r} = \tilde{r}(\mathbf{x}_{\text{tx}}) \in [0, \|\mathbf{S} - \mathbf{S}_\star\|]$, where for any given $\mathbf{x}_{\text{tx}} \in \mathcal{X}_{\text{tx}}$ and score $\tilde{\mathbf{S}}$

$$\hat{\mathbb{P}}_{\tilde{\mathbf{S}}}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}}) := \frac{\mathbb{P}(\mathbf{x}_{\text{im}}) \cdot \exp(\tilde{\mathbf{S}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))}{\mathbb{E}_{\mathbf{x}_{\text{im}}' \sim \mathbb{P}_{\text{im}}}[\exp(\tilde{\mathbf{S}}(\mathbf{x}_{\text{im}}', \mathbf{x}_{\text{tx}}))]}.$$

Since $\sup_{\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}, \mathbf{x}_{\text{tx}} \in \mathcal{X}_{\text{tx}}} \hat{\mathbb{P}}_{\mathbf{S}_\star + \tilde{r}h}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}})/\mathbb{P}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}}) \leq C$ for some constant $C > 0$ by Assumption 2.1, it follows that

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}_{\text{tx}}} [\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \hat{\mathbb{P}}_{\mathbf{S}_\star + \tilde{r}h}} (|\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})|^2)] \\ & \leq C \cdot \mathbb{E}_{\mathbf{x}_{\text{tx}}} [\mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}|\mathbf{x}_{\text{tx}}} |\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})|^2] \\ & \leq C \cdot \|\mathbf{S} - \mathbf{S}_\star\|^2 \stackrel{(i)}{\leq} C \cdot \lim_{K \rightarrow \infty} (\bar{\mathbf{R}}_{\text{clip}, \text{tx}, K}(\mathbf{S}) - \bar{\mathbf{R}}_{\text{clip}, \text{tx}, K}(\mathbf{S}_\star)) \\ & \stackrel{(ii)}{=} C \cdot \mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \left[\text{D}_{\text{KL}} \left(\mathbb{P}(\cdot|\mathbf{x}_{\text{tx}}) \middle| \middle| \hat{\mathbb{P}}_{\tilde{\mathbf{S}}}(\cdot|\mathbf{x}_{\text{tx}}) \right) \right], \end{aligned} \quad (\text{A.37})$$

where step (i) follows from Lemma A.5 for any $K \geq 3$, step (ii) follows from the proof of Proposition 2.3. Combining Eq. (A.34), (A.35), (A.36) and (A.37) yields Eq. (A.31).

Proof of Eq. (A.32). Let $\mathbf{x}_{\text{im},1}, \dots, \mathbf{x}_{\text{im},M}$ be i.i.d. samples from $\mathbb{P}(\mathbf{x}_{\text{im}})$. We choose $W_{\text{ada}}^{(1)} = \mathbf{E}_M^\dagger/M \in \mathbb{R}^{p_\star \times M}$ be the matrix consisting of the columns of $\mathbf{E}^\dagger$ that correspond to the samples $\{\mathbf{x}_{\text{im},j}\}_{j=1}^M$. We choose $W_{\text{ada}}^{(2)} = (\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im},j}) \in \mathbb{R}^p)_{j \in [M]} \in \mathbb{R}^{M \times p}$. For any $\mathbf{x}_{\text{tx}} \in \mathcal{X}_{\text{tx}}$, define

$$\begin{aligned} T(\mathbf{x}_{\text{tx}}) &:= (\Upsilon(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) - \log \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} [\exp(\Upsilon(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})))])_{\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}} \in \mathbb{R}^{|\mathcal{X}_{\text{im}}|}, \\ T^{(M)}(\mathbf{x}_{\text{tx}}) &:= (\Upsilon(W_{\text{ada},j}^{(2)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) - \log [\frac{1}{M} \sum_{k=1}^M \exp(\Upsilon(W_{\text{ada},k}^{(2)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})))])_{j \in [M]} \in \mathbb{R}^M, \\ \tilde{T}^{(M)}(\mathbf{x}_{\text{tx}}) &:= (\Upsilon(W_{\text{ada},j}^{(2)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) - \log \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} [\exp(\Upsilon(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})))])_{j \in [M]} \in \mathbb{R}^M. \end{aligned}$$

Then we have

$$\mathbb{E}_{\mathbf{x}_{\text{tx}}} [\|\tilde{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}) - \widehat{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}})\|_2^2] \leq 2\mathcal{R}_{E1} + 2\mathcal{R}_{E1},$$

where

$$\begin{aligned} \mathcal{R}_{E1} &:= \mathbb{E}_{\mathbf{x}_{\text{tx}}} [\|\mathbf{E}^\dagger \text{diag}(\mathbb{P}(\mathbf{x}_{\text{im}})) \Upsilon_\star^{-1}(T(\mathbf{x}_{\text{tx}})) - \frac{1}{M} \mathbf{E}_M^\dagger \Upsilon_\star^{-1}(\tilde{T}^{(M)}(\mathbf{x}_{\text{tx}}))\|_2^2], \\ \mathcal{R}_{E2} &:= \frac{1}{M^2} \mathbb{E}_{\mathbf{x}_{\text{tx}}} [\|\mathbf{E}_M^\dagger \Upsilon_\star^{-1}(\tilde{T}^{(M)}(\mathbf{x}_{\text{tx}})) - \mathbf{E}_M^\dagger \Upsilon_\star^{-1}(T^{(M)}(\mathbf{x}_{\text{tx}}))\|_2^2]. \end{aligned}$$

For $\mathcal{R}_{E1}$, since

$$\mathbb{E}[\mathbf{E}_M^\dagger \Upsilon_\star^{-1}(\tilde{T}^{(M)}(\mathbf{x}_{\text{tx}}))] = \mathbf{E}^\dagger \text{diag}(\mathbb{P}(\mathbf{x}_{\text{im}})) \Upsilon_\star^{-1}(T(\mathbf{x}_{\text{tx}})),$$

it follows that

$$\begin{aligned}\mathbb{E}[\mathcal{R}_{E1}] &= \mathbb{E}_{\mathbf{x}_{\text{tx}}} \mathbb{E}[\|\mathbf{E}^\dagger \text{diag}(\mathbb{P}(\mathbf{x}_{\text{im}})) \Upsilon_\star^{-1}(T(\mathbf{x}_{\text{tx}})) - \frac{1}{M} \mathbf{E}_M^\dagger \Upsilon_\star^{-1}(\tilde{T}^{(M)}(\mathbf{x}_{\text{tx}}))\|_2^2] \\ &= \mathbb{E}_{\mathbf{x}_{\text{tx}}} \left[\sum_{i \in [p_\star]} \text{Var} \left[\frac{1}{M} \mathbf{E}_{i,M}^\dagger \Upsilon_\star^{-1}(\tilde{T}^{(M)}(\mathbf{x}_{\text{tx}})) \right] \right].\end{aligned}$$

Since the variance in the above equation is invariant under any translation of $\Upsilon_\star^{-1}$, we can w.l.o.g. assume there exists a point $v_\star \in \mathbb{R}$ in the feasible range of $\Upsilon_\star^{-1}$ such that $\Upsilon_\star^{-1}(v_\star) \leq L_\Gamma$. It follows immediately that $\text{Var}[\mathbf{E}_{i,M}^\dagger \Upsilon_\star^{-1}(\tilde{T}^{(M)}(\mathbf{x}_{\text{tx}}))] \leq \mathbb{E}[|\mathbf{E}_{i,1}^\dagger \Upsilon_\star^{-1}(\tilde{T}^{(1)}(\mathbf{x}_{\text{tx}}))|^2]/M \leq CL_\Gamma^2 \mathbb{E}[\mathbf{E}_{i,1}^{\dagger 2}]/M$ for all $i \in [p_\star]$. Therefore, we further have

$$\mathbb{E}[\mathcal{R}_{E1}] \leq \frac{CL_\Gamma^2}{M} \cdot \mathbb{E}_{\mathbf{x}_{\text{tx}}} \left[\mathbb{E} \sum_{i \in [p_\star]} [\|\mathbf{E}_{i,1}^\dagger\|_2^2] \right] \leq \frac{CL_\Gamma^2 L_B^2 p_\star}{M}.$$

Let $\tilde{\Delta}^\Gamma(\mathbf{x}_{\text{tx}}) := \Upsilon_\star^{-1}(\tilde{T}^{(M)}(\mathbf{x}_{\text{tx}})) - \Upsilon_\star^{-1}(T^{(M)}(\mathbf{x}_{\text{tx}})) \in \mathbb{R}^M$. Note that all entries of $\tilde{\Delta}^\Gamma(\mathbf{x}_{\text{tx}})$ are equal. For $\mathcal{R}_{E2}$, we have

$$\begin{aligned}\mathbb{E}[\mathcal{R}_{E2}] &= \frac{1}{M^2} \mathbb{E}_{\mathbf{x}_{\text{tx}}} \mathbb{E}[\|\mathbf{E}_M^\dagger \tilde{\Delta}^\Gamma(\mathbf{x}_{\text{tx}})\|_2^2] \stackrel{(i)}{\leq} \mathbb{E}_{\mathbf{x}_{\text{tx}}} \mathbb{E}[\|\mathbf{E}_1^\dagger \tilde{\Delta}_1^\Gamma(\mathbf{x}_{\text{tx}})\|_2^2] \\ &\stackrel{(ii)}{=} \mathbb{E}_{\mathbf{x}_{\text{tx}}} \mathbb{E}_{\mathbf{x}_{\text{im},1}} [\|\mathbf{E}^\dagger(\mathbf{x}_{\text{im},1})\|_2^2 \cdot \mathbb{E}_{(\mathbf{x}_{\text{im},i})_{i=2}^M} |\tilde{\Delta}_1^\Gamma(\mathbf{x}_{\text{tx}})|^2],\end{aligned}$$

where step (i) follows from the symmetry of on $\mathbf{x}_{\text{im},1}, \dots, \mathbf{x}_{\text{im},M}$ and step (ii) follows from properties of conditional expectation

Since by Assumption 2.1, concentration of bounded random variables and Lipschitz continuity of $f(x) = \log(x)$ on $[1/c_1, c_1]$, we have

$$\|T_i^{(M)}(\mathbf{x}_{\text{tx}}) - \tilde{T}_i^{(M)}(\mathbf{x}_{\text{tx}})\|_{\psi_2} \leq \frac{C}{\sqrt{M}}$$

for all fixed $\mathbf{x}_{\text{tx}} \in \mathcal{X}_{\text{tx}}$, $\mathbf{x}_{\text{im},1} \in \mathcal{X}_{\text{im}}$ and $i \in [M]$. It follows from properties of sub-Gaussian random variables and Assumption 2.3 that

$$\begin{aligned}\mathbb{E}_{(\mathbf{x}_{\text{im},i})_{i=2}^M} [\mathbb{E}|\tilde{\Delta}_1^\Gamma(\mathbf{x}_{\text{tx}})|^2] &= \mathbb{E}_{(\mathbf{x}_{\text{im},i})_{i=2}^M} |\Upsilon_\star^{-1}(\tilde{T}_1^{(M)}(\mathbf{x}_{\text{tx}})) - \Upsilon_\star^{-1}(T_1^{(M)}(\mathbf{x}_{\text{tx}}))|^2 \\ &\leq L_\Gamma^2 \cdot \mathbb{E}_{(\mathbf{x}_{\text{im},i})_{i=2}^M} |\tilde{T}_1^{(M)}(\mathbf{x}_{\text{tx}}) - T_1^{(M)}(\mathbf{x}_{\text{tx}})|^2 \leq \frac{CL_\Gamma^2}{M}.\end{aligned}$$

Putting pieces together and using the Assumption on $\mathbf{E}^\dagger$ yields $\mathbb{E}[\mathcal{R}_{E2}] \leq CL_B^2 p_\star L_\Gamma^2 / M$. Lastly, under our choice of $W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}$, we have

$$\begin{aligned}\mathbb{E}\|W_{\text{ada}}^{(1)}\|_{\text{op}}^2 &\leq \mathbb{E}\|W_{\text{ada}}^{(1)}\|_F^2 = M \mathbb{E}\|\mathbf{E}_1^\dagger\|_2^2 \leq \frac{L_B^2}{M}, \\ \mathbb{E}\|W_{\text{ada}}^{(2)}\|_{\text{op}}^2 &\leq \mathbb{E}\|W_{\text{ada}}^{(2)}\|_F^2 = M \mathbb{E}\|\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im},1})\|_2^2 = B_{\text{Adap}}^2.\end{aligned}$$

Combining these with the bounds on $\mathbb{E}[\mathcal{R}_{E1}], \mathbb{E}[\mathcal{R}_{E2}]$, we may find samples $(\mathbf{x}_{\text{im},j})_{j \in [M]}$ such that

$$L_{\Gamma}^2 p_{\star} \|W_{\text{ada}}^{(1)}\|_{\text{op}}^2 + \frac{L_B^2 p_{\star} L_{\Gamma}^2}{B_{\text{Adap}} M} \|W_{\text{ada}}^{(2)}\|_{\text{op}}^2 + \mathcal{R}_{E1} + \mathcal{R}_{E2} \leq \frac{CL_B^2 p_{\star} L_{\Gamma}^2}{M}.$$

Choosing $(W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)})$ based on these samples gives an encoder $\widehat{\mathbf{E}}_{\text{tx}}$ such that Eq. (A.32) holds.

Remark A.1 (An improved bound). The error bound in Eq. (2.11) can be improved to

$$\mathbb{E}_{\mathbf{x}_{\text{tx}}} [\|\widehat{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}) - \mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}})\|_2^2] \leq C \cdot L_B^2 \cdot L_{\Gamma}^2 \cdot p_{\star} \cdot (\text{Suff}(\mathbf{E}_{\text{tx}}) + M^{-1}) \quad (\text{A.38})$$

if the link function Υ and the image embedding $\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})$ are chosen such that

$$\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \Upsilon(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) := \log \frac{\mathbb{P}_{\text{im}|\text{tx}}(\mathbf{x}_{\text{im}}|\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))}{\mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}})}. \quad (\text{A.39})$$

Note that such a pair of Υ and $\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})$ always exists as one can choose $\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}) = \mathbf{x}_{\text{im}}$ and $\Upsilon(\mathbf{x}_{\text{im}}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) = \log \frac{\mathbb{P}_{\text{im}|\text{tx}}(\mathbf{x}_{\text{im}}|\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))}{\mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}})}$.

To establish Eq. (A.38), echoing the notations in Definition 2.1, it suffices to note that

$$\text{Suff}(\mathbf{E}_{\text{tx}}) = \mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \left[\text{D}_{\text{KL}} \left(\mathbb{P}_{\text{im}|\text{tx}}(\cdot|\mathbf{x}_{\text{tx}}) \middle| \middle| \widehat{\mathbb{P}}_{\text{S}}(\cdot|\mathbf{x}_{\text{tx}}) \right) \right]$$

under the choices in Eq. (A.39), and recall that we have proved a stronger bound than Eq. (2.11) in Eq. (A.33).

A.4.8 Details in the proof of Corollary 2.2

In the section, we explain how the Neyman-Fisher factorization theorem can be used to prove Corollary 2.2. Recall the Neyman-Fisher factorization theorem (see e.g., Theorem 3.6 in Keener 2010):

Theorem 3.6 (Keener, 2010). Let $\mathcal{P} = \{P_{\theta} : \theta \in \Theta\}$ be a family of distributions dominated by a measure μ , with densities p_{θ} . A statistic $T(X)$ is sufficient for θ if and only if there exist measurable functions $g_{\theta} \geq 0$ and $h \geq 0$ such that

$$p_{\theta}(x) = g_{\theta}(T(x)) h(x), \quad \text{for a.e. } x \text{ under } \mu.$$

Also, recall that we have the following decomposition in the proof of Corollary 2.2:

$$\mathbb{P}_{\text{im}|\text{tx}}(\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}}) = \exp\{-\text{const}\} \cdot \mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}}) \cdot \exp\{\Upsilon(\mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}}))\}.$$

In our setting, we can choose the parameter $\theta = \mathbf{x}_{\text{tx}}$, the sample $X = \mathbf{x}_{\text{im}}$, and the dominating measure μ be the counting measure. Moreover, we let

$$\begin{aligned} T(\mathbf{x}_{\text{im}}) &= \mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}}), \quad g_{\theta}(T(\mathbf{x}_{\text{im}})) = \exp\{\Upsilon(\mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}}))\}, \text{ and} \\ h(\mathbf{x}_{\text{im}}) &= \mathbb{P}_{\text{im}}(\mathbf{x}_{\text{im}}) \exp\{-\text{const}\}. \end{aligned}$$

So by Theorem 3.6 in Keener (2010), $\mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}})$ is sufficient for $\mathbf{x}_{\text{tx}}$. The argument for $\mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}})$ is symmetric.

A.4.9 Properties of approximate sufficiency

Lemma A.4. *Under Definition 2.1, we have*

$$\text{Suff}(\mathbf{E}_{\text{im}}) = \text{MI}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \text{MI}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{x}_{\text{tx}}), \quad (\text{A.40a})$$

$$\text{Suff}(\mathbf{E}_{\text{tx}}) = \text{MI}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \text{MI}(\mathbf{x}_{\text{im}}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})). \quad (\text{A.40b})$$

Note that $\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})$ (resp. $\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})$) is a sufficient statistics if and only if $\text{Suff}(\mathbf{E}_{\text{tx}})$ (resp. $\text{Suff}(\mathbf{E}_{\text{im}})$) is zero. Moreover,

$$\text{Suff}(\mathbf{E}_{\text{im}}) = \inf_{\mathbb{Q}: \mathbb{R}^p \rightarrow \Delta(\mathcal{X}_{\text{tx}})} \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\mathbb{P}_{\text{tx}|\text{im}}(\cdot | \mathbf{x}_{\text{im}}) \middle| \middle| \mathbb{Q}(\cdot | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right) \right], \quad (\text{A.40c})$$

$$\text{Suff}(\mathbf{E}_{\text{tx}}) = \inf_{\mathbb{Q}: \mathbb{R}^p \rightarrow \Delta(\mathcal{X}_{\text{im}})} \mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \left[D_{\text{KL}} \left(\mathbb{P}_{\text{im}|\text{tx}}(\cdot | \mathbf{x}_{\text{tx}}) \middle| \middle| \mathbb{Q}(\cdot | \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) \right) \right]. \quad (\text{A.40d})$$

Proof of Lemma A.4. Note that

$$\begin{aligned} & \text{MI}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \text{MI}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{x}_{\text{tx}}) \\ &= \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}} \left[\log \frac{\mathbb{P}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})}{\mathbb{P}(\mathbf{x}_{\text{im}})\mathbb{P}(\mathbf{x}_{\text{tx}})} \right] - \mathbb{E}_{\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{x}_{\text{tx}}} \left[\log \frac{\mathbb{P}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{x}_{\text{tx}})}{\mathbb{P}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))\mathbb{P}(\mathbf{x}_{\text{tx}})} \right] \\ &= \mathbb{E}_{\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})} \left[\log \frac{\mathbb{P}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})}{\mathbb{P}(\mathbf{x}_{\text{im}})\mathbb{P}(\mathbf{x}_{\text{tx}})} - \log \frac{\mathbb{P}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{x}_{\text{tx}})}{\mathbb{P}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))\mathbb{P}(\mathbf{x}_{\text{tx}})} \right] \\ &= \mathbb{E}_{\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})} \left[\log \mathbb{P}(\mathbf{x}_{\text{tx}} | \mathbf{x}_{\text{im}}) - \log \mathbb{P}(\mathbf{x}_{\text{tx}} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right] \\ &= \mathbb{E}_{\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}} \left[\log \mathbb{P}(\mathbf{x}_{\text{tx}} | \mathbf{x}_{\text{im}}) - \log \mathbb{P}(\mathbf{x}_{\text{tx}} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right] = \text{Suff}(\mathbf{E}_{\text{im}}). \end{aligned}$$

This gives the Eq. (A.40a). Eq. (A.40b) follows from the symmetry between image and text. To establish Eq. (A.40c) and (A.40d), we note that

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\mathbb{P}_{\text{tx}|\text{im}}(\cdot | \mathbf{x}_{\text{im}}) \middle| \middle| \mathbb{Q}(\cdot | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right) \right] \\ &= \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\mathbb{P}_{\text{tx}|\text{im}}(\cdot | \mathbf{x}_{\text{im}}) \middle| \middle| \mathbb{P}(\cdot | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right) \right] \\ & \quad + \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}|\text{im}}(\cdot | \mathbf{x}_{\text{im}})} \left(\frac{\log \mathbb{P}(\mathbf{x}_{\text{tx}} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))}{\log \mathbb{Q}(\mathbf{x}_{\text{tx}} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))} \right) \right] \\ &\stackrel{(i)}{=} \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\mathbb{P}_{\text{tx}|\text{im}}(\cdot | \mathbf{x}_{\text{im}}) \middle| \middle| \mathbb{P}(\cdot | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right) \right] \\ & \quad + \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}|\text{im}}(\cdot | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))} \left(\frac{\log \mathbb{P}(\mathbf{x}_{\text{tx}} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))}{\log \mathbb{Q}(\mathbf{x}_{\text{tx}} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))} \right) \right] \\ &= \text{Suff}(\mathbf{E}_{\text{im}}) + \mathbb{E}_{\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left[D_{\text{KL}} \left(\mathbb{P}_{\text{tx}|\text{im}}(\cdot | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \middle| \middle| \mathbb{Q}(\cdot | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right) \right] \geq \text{Suff}(\mathbf{E}_{\text{im}}), \end{aligned}$$

where step (i) follows since for any function $f(\mathbf{x}_{\text{tx}}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))$, we have

$$\mathbb{E}_{\mathbf{x}_{\text{im}}} [\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}|\text{im}}(\cdot | \mathbf{x}_{\text{im}})} f(\mathbf{x}_{\text{tx}}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))] = \mathbb{E}_{\mathbf{x}_{\text{im}}} [\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}|\text{im}}(\cdot | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))} f(\mathbf{x}_{\text{tx}}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))]$$

$$\begin{aligned}
&= \mathbb{E}_{\mathbf{x}_{\text{im}}} [\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}|\text{im}}(\cdot|\mathbf{x}_{\text{im}})} f(\mathbf{x}_{\text{tx}}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})] \\
&= \mathbb{E}_{\mathbf{x}_{\text{im}}} \left[\sum_{\mathbf{x}_{\text{tx}} \in \mathcal{X}_{\text{tx}}} \mathbb{E}[\mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}}) | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})] f(\mathbf{x}_{\text{tx}}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right] \\
&= \mathbb{E}_{\mathbf{x}_{\text{im}}} \left[\sum_{\mathbf{x}_{\text{tx}} \in \mathcal{X}_{\text{tx}}} \mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx}} | \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) f(\mathbf{x}_{\text{tx}}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right] \\
&= \mathbb{E}_{\mathbf{x}_{\text{im}}} [\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}|\text{im}}(\cdot|\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))} f(\mathbf{x}_{\text{tx}}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))].
\end{aligned}$$

□

A.4.10 Auxiliary lemmas

Lemma A.5 (Bounds on $\|\mathbf{S} - \mathbf{S}_\star\|$). *Under the assumptions and notations in Proposition 2.3 and its proof, we have*

$$\begin{aligned}
\|\mathbf{S} - \mathbf{S}_\star\| &\leq C \cdot \sqrt{\bar{\mathbf{R}}_{\text{clip},\star,K}(\mathbf{S}) - \bar{\mathbf{R}}_{\text{clip},\star,K}(\mathbf{S}_\star)} \quad \text{for } \star \in \{\text{im}, \text{tx}\}, \\
&\leq C \cdot \sqrt{\bar{\mathbf{R}}_{\text{clip},K}(\mathbf{S}) - \bar{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_\star)}
\end{aligned}$$

for some constant $C > 0$ depending polynomially in c_1 .

Proof of Lemma A.5. We only prove the lemma for $\star = \text{im}$. The other case follows by symmetry between image and text. Note that

$$\begin{aligned}
\bar{\mathbf{R}}_{\text{clip},\text{im},K}(\mathbf{S}) &= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k}} \left[-\log \frac{\exp(\mathbf{S}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}))}{\sum_{j \in [K]} \exp(\mathbf{S}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}))} \right] \\
&= \mathbb{E} \left[\log \sum_{j \in [K]} \exp(\mathbf{S}(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j})) \right],
\end{aligned}$$

where the last line follows since $(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}}) \sim \mathbb{P}_{\text{im},\text{tx}}$ and we assume

$$\mathbb{E}_{\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{tx}|\text{im}}} [\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})] = \mathbb{E}_{\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{tx}|\text{im}}} [\mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})] = 0$$

for all $\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}$ in the proof of Proposition 2.3.

Write $\mathbf{S} = \mathbf{S}_\star + r_0 h$ with $r_0 = \|\mathbf{S} - \mathbf{S}_\star\|$ and $h = (\mathbf{S} - \mathbf{S}_\star)/\|\mathbf{S} - \mathbf{S}_\star\|$. For any function $h : \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}} \mapsto \mathbb{R}$ such that $\mathbb{E}_{\mathbf{x}_{\text{tx}}|\mathbf{x}_{\text{im}} \sim \mathbb{P}_{\text{tx}|\text{im}}} [h(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})] = 0$ for all $\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}$ and $\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im},\text{tx}}} [h^2(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})] = 1$, it can be verified that for any $r \in \mathbb{R}$

$$\begin{aligned}
\partial_r \bar{\mathbf{R}}_{\text{clip},\text{im},K}(\mathbf{S}_\star + rh) &= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}} \left[\mathbb{E}_{k \sim p_{\mathbf{S}_\star + rh}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k})] \right], \\
\partial_r^2 \bar{\mathbf{R}}_{\text{clip},\text{im},K}(\mathbf{S}_\star + rh) &= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}} \left[\text{Var}_{k \sim p_{\mathbf{S}_\star + rh}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k})] \right], \\
\partial_r^3 \bar{\mathbf{R}}_{\text{clip},\text{im},K}(\mathbf{S}_\star + rh) &= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}} \left[\mathbb{E}_{k \sim p_{\mathbf{S}_\star + rh}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k}) - \mathbb{E}_{k \sim p_{\mathbf{S}_\star + rh}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k})]]^3 \right].
\end{aligned} \tag{A.41a}$$

We claim that

- (a). $\bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh)$ is globally convex in $r \in \mathbb{R}$ and is strongly convex at the minimizer $r = 0$, namely, there exists some constant $C > 0$ such that $\partial_r^2 \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh)|_{r=0} \geq 1/C$.
- (b). There exists some constant $C > 0$ such that $|\partial_r^3 \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh)| \leq C/r_0$ for any $|r| \leq r_0$.

We will prove these claims later in this section. With these two claims at hand, it follows from properties of convex functions that

$$\bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh) - \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star) \geq \begin{cases} r^2/C & \text{if } |r| < r_0/C', \\ r_0|r|/C & \text{if } |r| \geq r_0/C', \end{cases} \quad (\text{A.43})$$

for some constants $C, C' > 1$ polynomially dependent on c_1 in Assumption 2.1. The proof of equation (A.43) is deferred to the end of this section. Finally, choosing $r = r_0$ in equation (A.43) yields Lemma A.5.

Proof of claim (a). The global convexity of $\bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh)$ follows immediately from equation (A.41a) and the fact that the variance is non-negative. $r = 0$ is a global minimizer of $\bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh)$ because

$$\begin{aligned} \partial_r \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh)|_{r=0} &= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}} \left[\mathbb{E}_{k \sim p_{\mathbf{S}_\star}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k})] \right] \\ &\stackrel{(i)}{=} \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}})] = 0, \end{aligned}$$

where step (i) uses the fact that $p_{\mathbf{S}_\star}$ is the posterior distribution of $\bar{k}$ conditioned on $\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}$.

It remains to establish a lower bound on $\partial_r^2 \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh)|_{r=0}$. Note that

$$\begin{aligned} \text{Var}_{k \sim p_{\mathbf{S}_\star}} (h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k})) &= \sum_{i \neq j} p_{\mathbf{S}_\star}(i) p_{\mathbf{S}_\star}(j) \cdot (h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},i}) - h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}))^2 \\ &\geq \frac{1}{CK^2} \sum_{i,j \in [K]} (h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},i}) - h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}))^2 \end{aligned}$$

for some constant $C > 0$ that depends on c_1 polynomially, where the second line uses Assumption 2.1, which implies that $p_{\mathbf{S}_\star}(i) \in [1/C'/K, C'/K]$ for all $i \in [K]$ and some constant $C' > 0$. Therefore,

$$\begin{aligned} \partial_r^2 \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh)|_{r=0} &= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}} \left[\text{Var}_{k \sim p_{\mathbf{S}_\star}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k})] \right] \\ &\geq \frac{1}{CK^2} \sum_{i,j \in [K]} \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}} (h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},i}) - h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}))^2 \end{aligned}$$

$$\stackrel{(ii)}{\geq} \frac{1}{C} \cdot \mathbb{E}_{\mathbf{x}_{\text{im}}} [\text{Var}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}|\text{im}}} (h(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))] = \frac{1}{C},$$

where step (ii) uses the fact that for any $i \neq j$,

$$\begin{aligned} & \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]} (h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},i}) - h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},j}))^2 \\ & \geq \min\{\mathbb{E}_{\mathbf{x}_{\text{im}}} \text{Var}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}|\text{im}}} (h(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})), \mathbb{E}_{\mathbf{x}_{\text{im}}} \text{Var}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} (h(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))\}, \end{aligned}$$

and

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}_{\text{im}}} [\text{Var}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} (h(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))] \\ &= \frac{1}{2} \mathbb{E}_{\mathbf{x}_{\text{im}}} [\mathbb{E}_{\mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2} \sim \mathbb{P}_{\text{tx}}} [(h(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},1}) - h(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},2}))^2]] \\ &\geq \frac{1}{2} \mathbb{E}_{\mathbf{x}_{\text{im}}} \left[\inf_{(\mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2}) \in \mathcal{X}_{\text{tx}}^2} \left[\frac{\mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx},1}) \times \mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx},2})}{\mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx},1}) \times \mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx},2})} \right. \right. \\ &\quad \left. \left. \cdot \mathbb{E}_{\mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2} \sim \mathbb{P}_{\text{tx}|\text{im}}} [(h(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},1}) - h(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx},2}))^2] \right] \right] \\ &\geq \frac{1}{C} \cdot \mathbb{E}_{\mathbf{x}_{\text{im}}} [\text{Var}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}|\text{im}}} (h(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))]. \end{aligned}$$

Here the second inequality follows from the boundedness assumption on $\mathbf{S}_\star$ in Assumption 2.1. This completes the proof of claim (a).

Proof of claim (b). By definition,

$$\begin{aligned} & \partial_r^3 \bar{\mathbf{R}}_{\text{clip,im},K}(\mathbf{S}_\star + rh) \\ &= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]} \left[\mathbb{E}_{k \sim p_{\mathbf{S}_\star + rh}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k}) - \mathbb{E}_{k \sim p_{\mathbf{S}_\star + rh}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k})]]^3 \right] \\ &\leq 2 \sup_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}}} \frac{|\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})|}{\|\mathbf{S} - \mathbf{S}_\star\|} \cdot \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]} \left[\text{Var}_{k \sim p_{\mathbf{S}_\star + rh}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k})] \right] \\ &\leq \frac{C}{\|\mathbf{S} - \mathbf{S}_\star\|} \cdot \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]} \left[\text{Var}_{k \sim p_{\mathbf{S}_\star + rh}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k})] \right] \leq \frac{C}{\|\mathbf{S} - \mathbf{S}_\star\|}, \end{aligned}$$

where the second inequality uses Assumption 2.1 and noting that

$$\begin{aligned} & \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]} \left[\text{Var}_{k \sim p_{\mathbf{S}_\star + rh}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k})] \right] \leq \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]} \left[\mathbb{E}_{k \sim p_{\mathbf{S}_\star + rh}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k})^2] \right] \\ &\stackrel{(i)}{\leq} C \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]} \left[\mathbb{E}_{k \sim p_{\mathbf{S}_\star}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k})^2] \right] = C \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},\bar{k}})^2] = C, \end{aligned}$$

where step (i) uses Assumption 2.1, which implies that $p_{\mathbf{S}_\star + rh}/p_{\mathbf{S}_\star} \in [1/C, C]$ for some $C > 0$ depending polynomially on c_1 .

Proof of claim (A.43). Using claim (a), (b) and the properties of convex functions, we have

$$\bar{\mathbf{R}}_{\text{clip,im},K}(\mathbf{S}_\star + rh) - \bar{\mathbf{R}}_{\text{clip,im},K}(\mathbf{S}_\star) \geq \frac{1}{2} r^2 - \frac{C}{\|\mathbf{S} - \mathbf{S}_\star\|} |r|^3$$

for some constant $C > 0$. It follows immediately that

$$\bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh) - \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star) \geq \frac{1}{4}|r|^2$$

for $|r| \leq \|\mathbf{S} - \mathbf{S}_\star\|/C$ for some constant $C > 0$. Moreover, by claim (a), (b) and Newton-Leibniz formula

$$\partial_r^2 \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh)|_{r=r_1} \geq -\frac{C}{\|\mathbf{S} - \mathbf{S}_\star\|} \cdot |r_1| + \partial_r^2 \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh)|_{r=0} \geq \frac{1}{C}$$

when $|r_1| \leq \|\mathbf{S} - \mathbf{S}_\star\|/C'$ for some constant $C' > 0$. It then follows immediately that at $r = \pm \|\mathbf{S} - \mathbf{S}_\star\|/C'$

$$|\partial_r \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh)| \geq \frac{\|\mathbf{S} - \mathbf{S}_\star\|}{C}.$$

for some constant $C > 0$.

Now, let $\text{proj}_c(x) = \arg\min_{y \in [-c, c]} |x - y|$ be the projection of x to the interval $[-c, c]$. Putting pieces together, we can find some constant $C' > 0$ such that for any $|r| \leq \|\mathbf{S} - \mathbf{S}_\star\|/C'$,

$$\bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh) - \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star) \geq \frac{1}{4}|r|^2,$$

and for any $|r| > \|\mathbf{S} - \mathbf{S}_\star\|/C'$,

$$\begin{aligned} & \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh) - \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star) \\ & \geq \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh) - \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + \text{proj}_{\|\mathbf{S} - \mathbf{S}_\star\|/C'}(r)h) + \frac{\|\mathbf{S} - \mathbf{S}_\star\|^2}{C} \\ & \geq |\partial_r \bar{R}_{\text{clip,im},K}(\mathbf{S}_\star + rh)| \cdot |r - \text{proj}_{\|\mathbf{S} - \mathbf{S}_\star\|/C'}(r)| + \frac{\|\mathbf{S} - \mathbf{S}_\star\|^2}{C} \\ & \geq \frac{\|\mathbf{S} - \mathbf{S}_\star\| \cdot |r|}{C}, \end{aligned}$$

where the third line uses properties of convex functions. $\square$

Lemma A.6 (Bound on T_{d2}). *Recall the definition of T_{d2} in equation (A.17). Under the assumptions and notations in Proposition 2.3 and its proof, for some constant $C > 0$*

$$|T_{d2}(r)| \leq \frac{C \cdot \|\mathbf{S} - \mathbf{S}_\star\|^2}{K}$$

for all $r \in [0, \|\mathbf{S} - \mathbf{S}_\star\|]$.

Proof of Lemma A.6. Write $\mathbf{S} = \mathbf{S}_\star + rh$ with $r = \|\mathbf{S} - \mathbf{S}_\star\|$ and $h = (\mathbf{S} - \mathbf{S}_\star)/\|\mathbf{S} - \mathbf{S}_\star\|$. By the scaling property of variance and noting that $\text{Var}_P(X) = \mathbb{E}_{X,Y \sim \text{iid } P}(X - Y)^2/2$, it suffices to show

$$|V_1 - V_2| \leq \frac{C}{K}, \tag{A.44}$$

where

$$\begin{aligned} V_1 &:= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k}; k_1, k_2 \sim p_{S_* + rh}} [h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k_1}) - h(\bar{\mathbf{x}}_{\text{im}}, \bar{\mathbf{x}}_{\text{tx},k_2})]^2, \\ V_2 &:= \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}} \sim \mathbb{P}_{\text{im}}; \mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2} \sim \hat{\mathbb{P}}_{S_* + rh}} [h(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},1}) - h(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},2})]^2. \end{aligned} \quad (\text{A.45a})$$

Let $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}(\cdot, \cdot | \bar{\mathbf{x}}_{\text{im}})$ denote the joint distribution of $(\bar{\mathbf{x}}_{\text{tx},k_1}, \bar{\mathbf{x}}_{\text{tx},k_2})$ conditioned on $\bar{\mathbf{x}}_{\text{im}}$ in the definition of V_1 , and let $\mathbb{P}_{\text{tx}|\text{im}}^{(r)}(\cdot, \cdot | \mathbf{x}_{\text{im}})$ denote the joint distribution of $(\mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2})$ conditioned on $\bar{\mathbf{x}}_{\text{im}}$ in the definition of V_2 .

We claim that there exists some constant $C, C' > 0$ such that when $K \geq C'$

$$\frac{|\mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b} | \bar{\mathbf{x}}_{\text{im}}) - \mathbb{P}_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b} | \bar{\mathbf{x}}_{\text{im}})|}{\mathbb{P}_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b} | \bar{\mathbf{x}}_{\text{im}})} \leq \frac{C}{K} \quad (\text{A.46})$$

for all $\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b} \in \mathcal{X}_{\text{tx}}$ such that $\mathbf{x}_{\text{tx},a} \neq \mathbf{x}_{\text{tx},b}$. Given claim (A.46) and adopting the shorthand notation $\Delta^h(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2}) = h(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},1}) - h(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},2})$, we immediately obtain

$$\begin{aligned} |V_1 - V_2| &= \left| \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}} \sim \mathbb{P}_{\text{im}}; (\mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2}) \sim \mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}} [\Delta^h(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},1})^2] - \right. \\ &\quad \left. \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}} \sim \mathbb{P}_{\text{im}}; (\mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2}) \sim \mathbb{P}_{\text{tx}|\text{im}}^{(r)}} [\Delta^h(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},1})^2] \right| \\ &\leq \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}} \sim \mathbb{P}_{\text{im}}} \left| \mathbb{E}_{(\mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2}) \sim \mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}} [\Delta^h(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},1})^2] \right. \\ &\quad \left. - \mathbb{E}_{(\mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2}) \sim \mathbb{P}_{\text{tx}|\text{im}}^{(r)}} [\Delta^h(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},1})^2] \right| \\ &\stackrel{(i)}{\leq} \frac{C}{K} \cdot \mathbb{E}_{\bar{\mathbf{x}}_{\text{im}} \sim \mathbb{P}_{\text{im}}} [\mathbb{E}_{(\mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2}) \sim \mathbb{P}_{\text{tx}|\text{im}}^{(r)}} [\Delta^h(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},1})^2]] = \frac{C}{K} \cdot V_2 \stackrel{(ii)}{\leq} \frac{C}{K}, \end{aligned}$$

where step (i) uses equation (A.46) and the fact that $\Delta^h(\bar{\mathbf{x}}_{\text{tx}}, \mathbf{x}_{\text{tx},1}, \mathbf{x}_{\text{tx},2}) = 0$ when $\mathbf{x}_{\text{tx},1} = \mathbf{x}_{\text{tx},2}$; step (ii) follows from Assumption 2.1, which implies $\sup_{\mathbf{x}_{\text{tx}} \in \mathcal{X}_{\text{tx}}} \left[\frac{\hat{\mathbb{P}}_{S_* + rh}(\mathbf{x}_{\text{tx}})}{\mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx}} | \bar{\mathbf{x}}_{\text{im}})} \right] \leq C$ for some constant $C > 0$, and the fact that $\mathbb{E}_{(\mathbf{x}_{\text{tx}}, \mathbf{x}_{\text{im}}) \sim \mathbb{P}_{\text{tx}, \text{im}}} [h(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})^2] = 1$.

When $K < C'$, it can be readily verified by Assumption 2.1 and noting $\mathbb{E}_{(\mathbf{x}_{\text{tx}}, \mathbf{x}_{\text{im}}) \sim \mathbb{P}_{\text{tx}, \text{im}}} [h(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})^2] = 1$ that equation (A.44) holds. This completes the proof of equation (A.44) and hence Lemma A.6.

Proof of claim (A.46). In the expression of V_1 in equation (A.45a), we note that the distribution of (k_1, k_2) conditioned on $(\bar{\mathbf{x}}_{\text{im}}, (\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}, \bar{k})$ remains unchanged under any permutation of $(\bar{\mathbf{x}}_{\text{tx},j})_{j \in [K]}$. Therefore, without loss of generality, we can drop the implicit dependence on $\bar{k}$ and assume

$$(\bar{\mathbf{x}}_{\text{tx},j})_{2 \leq j \leq K} \stackrel{i.i.d.}{\sim} \mathbb{P}_{\text{tx}}, \quad \text{and} \quad \bar{\mathbf{x}}_{\text{tx},1} \sim \mathbb{P}_{\text{tx}|\text{im}}(\cdot | \bar{\mathbf{x}}_{\text{im}}).$$

To provide an overview, the proof consists of three steps. First, we rewrite the expressions for $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}, \mathbb{P}_{\text{tx}|\text{im}}^{(r)}$ in terms of the expectation of certain quantities conditioned on $\bar{\mathbf{x}}_{\text{im}}$. Second, we

introduce an additional distribution on $\mathcal{X}_{\text{tx}} \times \mathcal{X}_{\text{tx}}$, denoted by $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}$, which connects two distributions $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}$, $\mathbb{P}_{\text{tx}|\text{im}}^{(r)}$, and we bound the differences $|\mathbb{P}_{\text{tx}|\text{im}}^{(K,r)} - \mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}|$, $|\mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)} - \mathbb{P}_{\text{tx}|\text{im}}^{(r)}|$ separately. Finally, we combine the bounds to obtain claim (A.46).

Rewriting the expressions for $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}$ and $\mathbb{P}_{\text{tx}|\text{im}}^{(r)}$. Adopt the shorthand notation $\mathbf{S}_r = \mathbf{S}_* + rh$. By the definition of $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}$, for any $(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}) \in \mathcal{X}_{\text{tx}} \times \mathcal{X}_{\text{tx}}$

$$\begin{aligned} \mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b} | \bar{\mathbf{x}}_{\text{im}}) &= \sum_{i,j=1}^K \mathbb{E}[1_{\{k_1=i, k_2=j, \bar{\mathbf{x}}_{\text{tx},i}=\mathbf{x}_{\text{tx},a}, \bar{\mathbf{x}}_{\text{tx},j}=\mathbf{x}_{\text{tx},b}\}} | \bar{\mathbf{x}}_{\text{im}}] \\ &= \sum_{i,j=1}^K \mathbb{E}[\mathbb{E}[1_{\{k_1=i, k_2=j, \bar{\mathbf{x}}_{\text{tx},i}=\mathbf{x}_{\text{tx},a}, \bar{\mathbf{x}}_{\text{tx},j}=\mathbf{x}_{\text{tx},b}\}} | (\bar{\mathbf{x}}_{\text{tx},k})_{k \in [K]}, \bar{\mathbf{x}}_{\text{im}}] | \bar{\mathbf{x}}_{\text{im}}] \\ &=: T_{\text{tx}|\text{im}}^{(K,r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}, \bar{\mathbf{x}}_{\text{im}}), \end{aligned}$$

where

$$\begin{aligned} &T_{\text{tx}|\text{im}}^{(K,r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}, \bar{\mathbf{x}}_{\text{im}}) \\ &= \sum_{i,j=1}^K \mathbb{E} \left[\frac{\exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},a})) \cdot \exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},b}))}{[\sum_{k \in [K]} \exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},k}))]^2} 1_{\{\bar{\mathbf{x}}_{\text{tx},i}=\mathbf{x}_{\text{tx},a}, \bar{\mathbf{x}}_{\text{tx},j}=\mathbf{x}_{\text{tx},b}\}} \middle| \bar{\mathbf{x}}_{\text{im}} \right]. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} &\mathbb{P}_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b} | \bar{\mathbf{x}}_{\text{im}}) \\ &= \frac{\exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},a})) \mathbb{P}(\mathbf{x}_{\text{tx},a}) \cdot \exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},b})) \mathbb{P}(\mathbf{x}_{\text{tx},b})}{[\mathbb{E}_{\mathbf{x}_{\text{tx}} \sim \mathbb{P}_{\text{tx}}} \exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx}}))]^2} \\ &\propto \exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},a})) \mathbb{P}(\mathbf{x}_{\text{tx},a}) \cdot \exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},b})) \mathbb{P}(\mathbf{x}_{\text{tx},b}) \\ &\stackrel{(i)}{\propto} \mathbb{E} \left[\frac{\exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},a})) \cdot \exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},b}))}{[\sum_{k \in [K-2]} \exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},k}))]^2} 1_{\{\bar{\mathbf{x}}_{\text{tx},K-1}=\mathbf{x}_{\text{tx},a}, \bar{\mathbf{x}}_{\text{tx},K}=\mathbf{x}_{\text{tx},b}\}} \middle| \bar{\mathbf{x}}_{\text{im}} \right] \\ &\stackrel{(ii)}{\propto} T_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}, \bar{\mathbf{x}}_{\text{im}}), \end{aligned}$$

where

$$\begin{aligned} &T_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}, \bar{\mathbf{x}}_{\text{im}}) \\ &:= \sum_{i \neq j; 2 \leq i, j \leq K} \mathbb{E} \left[\frac{\exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},a})) \cdot \exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},b}))}{[\sum_{k \in [K] \setminus \{i,j\}} \exp(\mathbf{S}_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},k}))]^2} 1_{\{\bar{\mathbf{x}}_{\text{tx},i}=\mathbf{x}_{\text{tx},a}, \bar{\mathbf{x}}_{\text{tx},j}=\mathbf{x}_{\text{tx},b}\}} \middle| \bar{\mathbf{x}}_{\text{im}} \right]. \end{aligned}$$

Above, step (i) follows from the conditional independence between $(\bar{\mathbf{x}}_{\text{tx},K-1}, \bar{\mathbf{x}}_{\text{tx},K})$ and $(\bar{\mathbf{x}}_{\text{tx},k})_{k \leq K-2}$, and the distributional assumption on $\bar{\mathbf{x}}_{\text{tx},K-1}, \bar{\mathbf{x}}_{\text{tx},K}$; step (ii) follows from the symmetry across the $K-1$ indices.

To control the different between $T_{\text{tx}|\text{im}}^{(K,r)}$ and $T_{\text{tx}|\text{im}}^{(r)}$, we introduce the function

$$\begin{aligned} & T_{\text{tx}|\text{im}}^{(K,r,-1)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}, \bar{\mathbf{x}}_{\text{im}}) \\ &:= \sum_{i \neq j; 2 \leq i, j \leq K} \mathbb{E} \left[\frac{\exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},a})) \cdot \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},b}))}{[\sum_{k \in [K]} \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},k}))]^2} 1_{\{\bar{\mathbf{x}}_{\text{tx},i} = \mathbf{x}_{\text{tx},a}, \bar{\mathbf{x}}_{\text{tx},j} = \mathbf{x}_{\text{tx},b}\}} \middle| \bar{\mathbf{x}}_{\text{im}} \right], \end{aligned}$$

and define the conditional distribution $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}$ on $\mathcal{X}_{\text{tx}} \times \mathcal{X}_{\text{tx}}$ to be the distribution proportional to $T_{\text{tx}|\text{im}}^{(K,r,-1)}$, namely,

$$\mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b} | \bar{\mathbf{x}}_{\text{im}}) \propto T_{\text{tx}|\text{im}}^{(K,r,-1)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}, \bar{\mathbf{x}}_{\text{im}}).$$

We will bound the differences between $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}$ and $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}$, $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}$ and $\mathbb{P}_{\text{tx}|\text{im}}^{(r)}$ in the following.

Bounding the differences. We first control the difference between $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}$ and $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}$. By Assumption 2.1, we have

$$\begin{aligned} 0 &\leq T_{\text{tx}|\text{im}}^{(K,r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}, \bar{\mathbf{x}}_{\text{im}}) - T_{\text{tx}|\text{im}}^{(K,r,-1)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}, \bar{\mathbf{x}}_{\text{im}}) \\ &= \sum_{i=j \text{ or } i=1 \text{ or } j=1} \mathbb{E} \left[\frac{\exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},a})) \cdot \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},b}))}{[\sum_{k \in [K]} \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},k}))]^2} 1_{\{\bar{\mathbf{x}}_{\text{tx},i} = \mathbf{x}_{\text{tx},a}, \bar{\mathbf{x}}_{\text{tx},j} = \mathbf{x}_{\text{tx},b}\}} \middle| \bar{\mathbf{x}}_{\text{im}} \right] \\ &\leq \frac{C}{K^2} \cdot \sum_{i=j \text{ or } i=1 \text{ or } j=1} \mathbb{P}(\bar{\mathbf{x}}_{\text{tx},i} = \mathbf{x}_{\text{tx},a}, \bar{\mathbf{x}}_{\text{tx},j} = \mathbf{x}_{\text{tx},b} | \bar{\mathbf{x}}_{\text{im}}) \\ &\leq \frac{C}{K} \cdot \left[1_{\{\mathbf{x}_{\text{tx},a} = \mathbf{x}_{\text{tx},b}\}} (\mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx},a} | \bar{\mathbf{x}}_{\text{im}}) + \mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx},a})) + \mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx},a} | \bar{\mathbf{x}}_{\text{im}}) \cdot \mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx},b}) \right. \\ &\quad \left. + \mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx},b} | \bar{\mathbf{x}}_{\text{im}}) \cdot \mathbb{P}_{\text{tx}}(\mathbf{x}_{\text{tx},a}) \right] \\ &\leq \frac{C}{K} \cdot \left[1_{\{\mathbf{x}_{\text{tx},a} = \mathbf{x}_{\text{tx},b}\}} \mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx},a} | \bar{\mathbf{x}}_{\text{im}}) + \mathbb{P}_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b} | \bar{\mathbf{x}}_{\text{im}}) \right], \end{aligned} \tag{A.47}$$

where the first inequality follows from the boundedness assumption of $\exp(S_r)$ implied by Assumption 2.1, and the second and third inequalities use the boundedness of $\exp(S_\star)$. Summing equation (A.47) over $\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}$ and recalling that $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r)} = T_{\text{tx}|\text{im}}^{(K,r)}$, we find

$$\begin{aligned} 1 &\geq \sum_{\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}} T_{\text{tx}|\text{im}}^{(K,r,-1)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}, \bar{\mathbf{x}}_{\text{im}}) \\ &\geq 1 - \sum_{\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}} \frac{C}{K} \cdot \left[1_{\{\mathbf{x}_{\text{tx},a} = \mathbf{x}_{\text{tx},b}\}} \mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx},a} | \bar{\mathbf{x}}_{\text{im}}) + \mathbb{P}_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b} | \bar{\mathbf{x}}_{\text{im}}) \right] \\ &= 1 - \frac{2C}{K}. \end{aligned}$$

Thus, when $K \geq 4C$ in the equation above, it follows from the triangle inequality that

$$\begin{aligned} & |\mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}|\bar{\mathbf{x}}_{\text{im}}) - \mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}|\bar{\mathbf{x}}_{\text{im}})| \\ & \leq \frac{C'}{K} \cdot \left[1_{\{\mathbf{x}_{\text{tx},a}=\mathbf{x}_{\text{tx},b}\}} \mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx},a}|\bar{\mathbf{x}}_{\text{im}}) + \mathbb{P}_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}|\bar{\mathbf{x}}_{\text{im}}) + \mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}|\bar{\mathbf{x}}_{\text{im}}) \right] \end{aligned} \quad (\text{A.48})$$

for some constant $C' > 0$.

Next, we bound the difference between $\mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}$ and $\mathbb{P}_{\text{tx}|\text{im}}^{(r)}$. Introduce the shorthand notations

$$s := \frac{\exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},a})) \cdot \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},b}))}{[\sum_{k \in [K]} \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},k}))]^2}, \quad s_{i,j} := \frac{\exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},a})) \cdot \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},b}))}{[\sum_{k \in [K] \setminus \{i,j\}} \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},k}))]^2}$$

for $i \neq j$. Then $s_{i,j} \geq s$ and

$$\begin{aligned} s_{i,j} - s &= \frac{[\sum_{k \in [K]} \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},k}))]^2 - [\sum_{k \in [K] \setminus \{i,j\}} \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},k}))]^2}{[\sum_{k \in [K]} \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},k}))]^2 \cdot [\sum_{k \in [K] \setminus \{i,j\}} \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},k}))]^2} \leq \frac{C \cdot K}{K^4} \\ &\leq \frac{C}{K} \cdot \frac{\exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},a})) \cdot \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},b}))}{[\sum_{k \in [K] \setminus \{i,j\}} \exp(S_r(\bar{\mathbf{x}}_{\text{im}}, \mathbf{x}_{\text{tx},k}))]^2} = \frac{C}{K} \cdot s_{i,j} \end{aligned}$$

for all $i \neq j$, where the inequalities follow from Assumption 2.1. Therefore, we obtain

$$\begin{aligned} 0 &\leq T_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}, \bar{\mathbf{x}}_{\text{im}}) - T_{\text{tx}|\text{im}}^{(K,r,-1)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}, \bar{\mathbf{x}}_{\text{im}}) \\ &= \sum_{i \neq j; 2 \leq i, j \leq K} \mathbb{E} \left[(s_{i,j} - s) \cdot 1_{\{\bar{\mathbf{x}}_{\text{tx},i}=\mathbf{x}_{\text{tx},a}, \bar{\mathbf{x}}_{\text{tx},j}=\mathbf{x}_{\text{tx},b}\}} \middle| \bar{\mathbf{x}}_{\text{im}} \right] \\ &\leq \frac{C}{K} \cdot \sum_{i \neq j; 2 \leq i, j \leq K} \mathbb{E} \left[s_{i,j} \cdot 1_{\{\bar{\mathbf{x}}_{\text{tx},i}=\mathbf{x}_{\text{tx},a}, \bar{\mathbf{x}}_{\text{tx},j}=\mathbf{x}_{\text{tx},b}\}} \middle| \bar{\mathbf{x}}_{\text{im}} \right] = \frac{C}{K} \cdot T_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}, \bar{\mathbf{x}}_{\text{im}}) \end{aligned} \quad (\text{A.49})$$

for all $\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b} \in \mathcal{X}_{\text{tx}}$.

Since $T_{\text{tx}|\text{im}}^{(r)}, T_{\text{tx}|\text{im}}^{(K,r,-1)}$ are proportional to the conditional distributions $\mathbb{P}_{\text{tx}|\text{im}}^{(r)}, \mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}$, when $K \geq 4C$ in equation (A.49), we have

$$|\mathbb{P}_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}|\bar{\mathbf{x}}_{\text{im}}) - \mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}|\bar{\mathbf{x}}_{\text{im}})| \leq \frac{C'}{K} \cdot \mathbb{P}_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}|\bar{\mathbf{x}}_{\text{im}}). \quad (\text{A.50})$$

for some constant $C' > 0$.

Combining the bounds. Combining bounds (A.48) and (A.50) with the triangle inequality, when $K \geq C$ for some constant $C > 0$ depending polynomially on c_1 in Assumption 2.1, we obtain

$$|\mathbb{P}_{\text{tx}|\text{im}}^{(K,r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}|\bar{\mathbf{x}}_{\text{im}}) - \mathbb{P}_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}|\bar{\mathbf{x}}_{\text{im}})|$$

$$\begin{aligned}
&\leq \frac{C'}{K} \cdot \left[1_{\{\mathbf{x}_{\text{tx},a}=\mathbf{x}_{\text{tx},b}\}} \mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx},a}|\bar{\mathbf{x}}_{\text{im}}) + \mathbb{P}_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}|\bar{\mathbf{x}}_{\text{im}}) + \mathbb{P}_{\text{tx}|\text{im}}^{(K,r,-1)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}|\bar{\mathbf{x}}_{\text{im}}) \right] \\
&\leq \frac{C'}{K} \cdot \left[1_{\{\mathbf{x}_{\text{tx},a}=\mathbf{x}_{\text{tx},b}\}} \mathbb{P}_{\text{tx}|\text{im}}(\mathbf{x}_{\text{tx},a}|\bar{\mathbf{x}}_{\text{im}}) + \mathbb{P}_{\text{tx}|\text{im}}^{(r)}(\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}|\bar{\mathbf{x}}_{\text{im}}) \right]
\end{aligned}$$

for some constant $C' > 0$ for all $\mathbf{x}_{\text{tx},a}, \mathbf{x}_{\text{tx},b}$, where the last inequality uses Eq. (A.50). This yields claim (A.46). $\square$

Lemma A.7 (Girsanov theorem). *Let $\{\boldsymbol{\mu}_t, \boldsymbol{\gamma}_t\}_{t \geq 0} \subseteq \mathbb{R}^d \rightarrow \mathbb{R}^d$. Consider two stochastic differential equation*

$$\begin{aligned}
d\mathbf{x}_t &= \boldsymbol{\mu}_t(\mathbf{x}_t)dt + d\mathbf{W}_t, & \mathbf{x}_0 &= \mathbf{0}, \\
d\mathbf{y}_t &= \boldsymbol{\gamma}_t(\mathbf{y}_t)dt + d\mathbf{W}_t, & \mathbf{y}_0 &= \mathbf{0}.
\end{aligned}$$

Let $\mathbf{P}_T$ be the distribution of $\mathbf{x}_T$, and $\mathbf{Q}_T$ be the distribution of $\mathbf{y}_T$. Then we have

$$\mathbf{D}_{\text{KL}}(\mathbf{P}_T || \mathbf{Q}_T) \leq \frac{1}{2} \int_0^T \mathbb{E}_{\mathbf{x}_t} \|\boldsymbol{\mu}_t(\mathbf{x}_t) - \boldsymbol{\gamma}_t(\mathbf{x}_t)\|_2^2 dt.$$

Proof of Lemma A.7. We provide a proof here for completeness. Let $\mathbf{P}, \mathbf{Q}$ denote the distributions of $\mathbf{x}_{0:T}, \mathbf{y}_{0:T}$, respectively. Girsanov theorem implies that for any $\mathbf{z}_{0:T}$

$$\log \frac{\mathbf{P}(\mathbf{z}_{0:T})}{\mathbf{Q}(\mathbf{z}_{0:T})} = \int_0^T (\boldsymbol{\mu}_t(\mathbf{z}_t) - \boldsymbol{\gamma}_t(\mathbf{z}_t)) d\mathbf{W}_t + \frac{1}{2} \int_0^T \|\boldsymbol{\mu}_t(\mathbf{z}_t) - \boldsymbol{\gamma}_t(\mathbf{z}_t)\|_2^2 dt.$$

Therefore,

$$\begin{aligned}
\mathbf{D}_{\text{KL}}(\mathbf{P}_T || \mathbf{Q}_T) &\leq \mathbf{D}_{\text{KL}}(\mathbf{P} || \mathbf{Q}) \\
&= \mathbb{E}_{\mathbf{x}_{0:T}} \left[\int_0^T (\boldsymbol{\mu}_t(\mathbf{x}_t) - \boldsymbol{\gamma}_t(\mathbf{x}_t)) d\mathbf{W}_t + \frac{1}{2} \int_0^T \|\boldsymbol{\mu}_t(\mathbf{x}_t) - \boldsymbol{\gamma}_t(\mathbf{x}_t)\|_2^2 dt \right] \\
&= \frac{1}{2} \mathbb{E}_{\mathbf{x}_{0:T}} \int_0^T \|\boldsymbol{\mu}_t(\mathbf{x}_t) - \boldsymbol{\gamma}_t(\mathbf{x}_t)\|_2^2 dt = \frac{1}{2} \int_0^T \mathbb{E}_{\mathbf{x}_t} \|\boldsymbol{\mu}_t(\mathbf{x}_t) - \boldsymbol{\gamma}_t(\mathbf{x}_t)\|_2^2 dt,
\end{aligned}$$

where the first inequality uses data-processing inequality. $\square$

A.5 Proof of Theorem 2.10

A.5.1 Overview

The generalization error analysis of neural networks is typically conducted by first constructing a neural network that approximates a certain function (or algorithm) and then evaluating the complexity of the class containing that network. This subsection explains the whole architecture and the proof strategy that constructs a pipeline approximating the target algorithm.

The transformer-based architectures in the following will process vectors related to each node by concatenating them into a matrix. Thus associating nodes with integers will make the discussion easier. For this, with a slight abuse of notation, we identify the nodes with the positive integer defined as follows.

Definition A.1 (Numbering of nodes). For $\square \in \{\text{im}, \text{tx}\}$, a node $v \in \mathcal{V}_{\square}^{(L)}$ is identified as a integer defined as

$$\begin{aligned} \iota(v) + m_{\square}^{(L)}(\iota(\text{pa}(v)) - 1) + m_{\square}^{(L-1)}m_{\square}^{(L)}(\iota(\text{pa}^{(2)}(v)) - 1) \\ + \dots + (m_{\square}^{(2)} \dots m_{\square}^{(L)})(\iota(\text{pa}^{(L-1)}(v)) - 1). \end{aligned}$$

Here $\text{pa}^{(\ell)}(v)$ means the ℓ -th grand parent of v . We also identify intermediate nodes $v \in \mathcal{V}_{\square}^{(\ell)} (\ell = L-1, \dots, 0)$ as a positive integer

$$(m_{\square}^{(L)} \dots m_{\square}^{(\ell+1)})[\iota(v) + m_{\square}^{(\ell)}(\iota(\text{pa}(v)) - 1) + \dots + (m_{\square}^{(2)} \dots m_{\square}^{(\ell)})(\iota(\text{pa}^{(\ell-1)}(v)) - 1)].$$

This allows us to compare two nodes u, v in different levels (say, $u \in \mathcal{V}_{\text{im}}^{(\ell)}, v \in \mathcal{V}_{\text{im}}^{(\ell')}$) like $u > v$ or $u = v$. However, treating a node $v \in \mathcal{V}_{\text{im}}^{(\ell)}$ as a node in another level ℓ' sometimes leads to confusion, as $\mathcal{N}$, $\mathcal{C}$, and “pa” no longer points a unique node. Therefore, when there is a risk of confusion, we explicitly indicate the level of the node by referring to the node as $v^{(\ell)}$.

Belief propagation and message passing algorithms

Now we outline our approach to approximating the algorithm. Let us recall the message passing algorithm we aim to approximate. For the text part, it starts with $h_{\text{tx},v}^{(L)} = x_{\text{tx},v}$ ($v \in \mathcal{V}_{\text{tx}}^{(L)}$), and computes $(q_{\text{tx},v}^{(\ell)})_{v \in \mathcal{V}_{\text{tx}}^{(\ell)}}$ and $(h_{\text{tx},v}^{(\ell)})_{v \in \mathcal{V}_{\text{tx}}^{(\ell)}}$ in the decreasing order of ℓ to obtain $h_{\text{tx},r}^{(0)} \in \mathbb{R}^S$:

$$\begin{aligned} q_{\text{tx},v}^{(\ell)} &= f_{\text{tx},\iota(v)}^{(\ell)}(h_{\text{tx},v}^{(\ell)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(\ell)}, \ell = L, \dots, 1, \\ h_{\text{tx},v}^{(\ell-1)} &= \text{normalize}(\sum_{u \in \mathcal{C}(v)} q_{\text{tx},u}^{(\ell)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(\ell-1)}, \ell = L, \dots, 1. \end{aligned} \quad (\text{A.51})$$

Computation of $q_{\text{tx},v}^{(\ell)}$ and $h_{\text{tx},v}^{(\ell-1)}$ from $h_{\text{tx},v}^{(\ell)}$ is called the ℓ -th step. Here, $f_{\text{tx},\iota}^{(\ell)}$ are defined as

$$\begin{aligned} (f_{\text{tx},\iota}^{(L)}(x))_s &= \log \psi_{\text{tx},\iota}^{(L)}(s, x), & x \in [S], s \in [S], \\ (f_{\text{tx},\iota}^{(\ell)}(h))_s &= \log \sum_{a \in [S]} \psi_{\text{tx},\iota}^{(\ell)}(s, a) e^{h_a}, & h \in \mathbb{R}^S, s \in [S], \ell = L-1, \dots, 2, \\ (f_{\text{tx},\iota}^{(1)}(h))_s &= \log \sum_{a \in [S]} (\mathbb{P}[s])^{\frac{1}{m_{\text{tx}}}} \psi_{\text{tx},\iota}^{(1)}(s, a) e^{h_a}, & h \in \mathbb{R}^S, s \in [S], \end{aligned} \quad (\text{A.52})$$

and $\text{normalize}(h)_s = x_s - \max_{s'} h_{s'}$ for $h \in \mathbb{R}^S$. In the same way, the image part yields $h_{\text{im},r}^{(0)} \in \mathbb{R}^S$. Combining them finally yields

$$\mathbf{S}_{\text{MP}} = f^{(0)}(\text{softmax}(h_{\text{im},r}^{(0)}), \text{softmax}(h_{\text{tx},r}^{(0)}))$$

where

$$f^{(0)}(h, h') = \log \sum_s h_s h'_s (\mathbb{P}[s])^{-1}, \quad h, h' \in [0, 1]^S. \quad (\text{A.53})$$

The correctness of this algorithm is formally stated as follows.

Lemma A.8 (MP yields the optimal similarity score). *Applying the message passing algorithm above, it holds that $\text{softmax}(h_{\text{im},r}^{(0)})_s = \mathbb{P}[s|\mathbf{x}_{\text{im}}]$, $\text{softmax}(h_{\text{tx},r}^{(0)})_s = \mathbb{P}[s|\mathbf{x}_{\text{tx}}]$, and $S_{\text{MP}} = S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) + (\text{const.})$.*

Proof. According to Lemma 2.1, the optimal similarity score function is defined as

$$S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \log \left[\frac{\mathbb{P}[\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}]}{\mathbb{P}[\mathbf{x}_{\text{tx}}] \cdot \mathbb{P}[\mathbf{x}_{\text{im}}]} \right].$$

From Proposition 2 of Mei (2024), it holds that $\text{softmax}(h_{\text{im},r}^{(0)})_s = \mathbb{P}[s|\mathbf{x}_{\text{im}}]$ and $\text{softmax}(h_{\text{tx},r}^{(0)})_s = \mathbb{P}[s|\mathbf{x}_{\text{tx}}]$. (Note that their definition of $\psi_t^{(0)}$ includes $\mathbb{P}[s]$ while our $\psi_{\text{im},t}^{(0)}$ and $\psi_{\text{tx},t}^{(0)}$ do not, which results in the $\mathbb{P}[s]^{\frac{1}{m(1)}}$ term in the definition of $f_{\text{tx},t}^{(1)}$.) By the Bayes rule,

$$\begin{aligned} \frac{\mathbb{P}[\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}]}{\mathbb{P}[\mathbf{x}_{\text{tx}}] \cdot \mathbb{P}[\mathbf{x}_{\text{im}}]} &= \frac{\sum_{s \in [S]} \mathbb{P}[\mathbf{x}_{\text{im}}|s] \mathbb{P}[\mathbf{x}_{\text{tx}}|s] \mathbb{P}[s]}{\mathbb{P}[\mathbf{x}_{\text{tx}}] \cdot \mathbb{P}[\mathbf{x}_{\text{im}}]} = \sum_{s \in [S]} \frac{\mathbb{P}[s|\mathbf{x}_{\text{im}}] \mathbb{P}[s|\mathbf{x}_{\text{tx}}]}{\mathbb{P}[s]} \\ &= \sum_{s \in [S]} \text{softmax}(h_{\text{im},r}^{(0)})_s \text{softmax}(h_{\text{tx},r}^{(0)})_s (\mathbb{P}[s])^{-1}. \end{aligned}$$

Taking the logarithm yields $S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \log \left[\frac{\mathbb{P}[\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}]}{\mathbb{P}[\mathbf{x}_{\text{tx}}] \cdot \mathbb{P}[\mathbf{x}_{\text{im}}]} \right]$. □

Approximation with transformer networks

We construct a transformer-based pipeline to replicate the message passing algorithm. It consists of three components: a transformer encoder for images $\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}$; a transformer encoder for text $\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}$; and a parameterized link function $\tau^w(h, h')$. The two transformers $\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}$ and $\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}$ approximately compute $h_{\text{im},r}^{(0)}$ and $h_{\text{tx},r}^{(0)}$, respectively, by following the message passing algorithm (A.51). Because $\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}$ and $\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}$ follows the same construction, we will sometimes omit the subscripts “tx” and “im” in the following to discuss these networks. Finally we put them into the link function $\tau^w(h, h')$.

After the embedding (positional encoding) layer Emb_{clip} , the network obtains the initial matrix $\mathbf{H}^{(L)}$ of size $(d_f + d_p) \times d$, with $d_f = 2SL + 1$ and $d_p = 2L$. Here $d_p = 2L$ is the dimension corresponding to the positional encoding, and the rest $d_f = 2SL + 1$ corresponds to the “features”. Specifically, this matrix is written as

$$\mathbf{H}^{(L)} = \text{Emb}_{\text{clip}}(\mathbf{x}) = \begin{bmatrix} & & \mathbf{0} & \\ x_1 & x_2 & \cdots & x_d \\ & & \mathbf{P} & \end{bmatrix},$$

so that it consists of the positional encoding $\mathbf{P} \in \mathbb{R}^{d_p \times d}$, the text variable $\mathbf{x}_{\text{tx}}$, and the zeros reserved for later calculation $\mathbf{0} \in \mathbb{R}^{(d_t-1) \times d}$.

Starting from $\mathbf{H}^{(L)}$, the transformer network applies L transformer blocks. These blocks are indexed by $\ell = L, \dots, 1$ in the decreasing order, so that the ℓ -th layer corresponds with the ℓ -th step of the message passing algorithm. They sequentially calculate $Q^{(\ell)}$ ($\ell = L, \dots, 1$) and $\mathbf{H}^{(\ell)}$ ($\ell = L, \dots, 0$) defined as

$$\mathbf{H}^{(\ell)} = \begin{bmatrix} \mathbf{0} & & & \\ \mathbf{h}_1^{(\ell)} & \mathbf{h}_2^{(\ell)} & \dots & \mathbf{h}_d^{(\ell)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_1^{(L)} & \mathbf{q}_2^{(L)} & \dots & \mathbf{q}_d^{(L)} \\ \mathbf{h}_1^{(L)} & \mathbf{h}_2^{(L)} & \dots & \mathbf{h}_d^{(L)} \\ \mathbf{P} & & & \end{bmatrix}, \quad Q^{(\ell)} = \begin{bmatrix} \mathbf{0} & & & \\ \mathbf{q}_1^{(\ell)} & \mathbf{q}_2^{(\ell)} & \dots & \mathbf{q}_d^{(\ell)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_1^{(L)} & \mathbf{q}_2^{(L)} & \dots & \mathbf{q}_d^{(L)} \\ \mathbf{h}_1^{(L)} & \mathbf{h}_2^{(L)} & \dots & \mathbf{h}_d^{(L)} \\ \mathbf{P} & & & \end{bmatrix},$$

where $\mathbf{h}_v^{(\ell)} \in \mathbb{R}^S$ and $\mathbf{q}_v^{(\ell)} \in \mathbb{R}^S$ except for $\mathbf{h}_v^{(L)} \in [S]$.

The ℓ -th block approximates the ℓ -th step of the message passing algorithm. It consists of a position-wise feed forward layer $\text{FF}^{(\ell)}$ with skip connection and self-attention layer $\text{Attn}^{(\ell)}$ with skip connection and normalization. The feed forward layer $\text{FF}^{(\ell)}$, a fully-connected ReLU network, receives $\mathbf{H}^{(\ell)}$ and outputs $Q^{(\ell)}$ by computing $\mathbf{q}_v^{(\ell)}$ from $\mathbf{h}_v^{(\ell)}$:

$$Q^{(\ell)} = \underbrace{\mathbf{H}^{(\ell)}}_{\text{skip connection}} + \text{FF}^{(\ell)}(\mathbf{H}^{(\ell)}) = \mathbf{H}^{(\ell)} + \begin{bmatrix} \mathbf{0} \ (\in \mathbb{R}^{((2\ell-1)S) \times d}) \\ \mathbf{q}_1^{(\ell)} & \mathbf{q}_2^{(\ell)} & \dots & \mathbf{q}_d^{(\ell)} \\ \mathbf{0} \ (\in \mathbb{R}^{(d_p+1+2(L-\ell)S) \times d}) \end{bmatrix}.$$

The self-attention layer $\text{Attn}^{(\ell)}$, uses this $Q^{(\ell)}$ and outputs $\mathbf{H}^{(\ell-1)}$ by computing $\mathbf{h}_v^{(\ell-1)}$:

$$\begin{aligned} \mathbf{H}^{(\ell-1)} &= \text{normalize} \left(\underbrace{Q^{(\ell)}}_{\text{skip connection}} + \text{Attn}^{(\ell)}(Q^{(\ell)}) \right) \\ &= \text{normalize} \left(Q^{(\ell)} + \begin{bmatrix} \mathbf{0} \ (\in \mathbb{R}^{((2\ell-2)S) \times d}) \\ \star \ (\in \mathbb{R}^{S \times d}) \\ \mathbf{0} \ (\in \mathbb{R}^{(d_p+1+(2L-2\ell+1)S) \times d}) \end{bmatrix} \right). \end{aligned}$$

Here, $\star$ denotes $[\mathbf{h}_1^{(\ell-1)} \ \mathbf{h}_2^{(\ell-1)} \ \dots \ \mathbf{h}_d^{(\ell-1)}]$ before normalization. In this way, we iteratively compute $\mathbf{q}_v^{(\ell)}$ and $\mathbf{h}_v^{(\ell)}$ to fill zeros of the previous matrices. These $\mathbf{h}_v^{(\ell)} \in \mathbb{R}^S$ and $\mathbf{q}_v^{(\ell)} \in \mathbb{R}^S$ approximate $h_u^{(\ell)}$ and $h_v^{(\ell)}$ as

$$\begin{aligned} \mathbf{q}_v^{(\ell)} &\approx q_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}, \quad v \in \mathcal{V}^{(L)}, \ell = L, \dots, 1, \\ \mathbf{h}_v^{(\ell)} &\approx h_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}, \quad v \in \mathcal{V}^{(L)}, \ell = L, \dots, 0. \end{aligned} \tag{A.54}$$

After we obtain $\mathbf{H}^{(0)}$, we extract $\mathbf{h}_d^{(0)}$ (this is an approximation of $h_r^{(0)}$) to output $\text{read}_{\text{clip}}(\mathbf{H}^{(0)}) = \mathbf{h}_d^{(0)}$.

We remark that our transformer block applies the feed forward layer first to emphasize the correspondence with the message passing. If adhering to a typical structure where self-attention comes first, we can implement the pipeline with $(L + 1)$ -blocks.

We now formally define each component of the network and explain key lemmas to confirm (A.54) iteratively.

Embedding Emb_{clip} . When the network receives the input $\mathbf{x} \in [S]^d$, it first passes it through the embedding layer Emb , where it concatenate the input $\mathbf{x}$ with the positional encoding $\mathbf{P}$ and the zeros $\mathbf{0}$. The v -th column of $\mathbf{P}$ is denoted by $\mathbf{p}_v$. This $\mathbf{p}_v \in \mathbb{R}^{d_p}$ ($d_p = 2L$) is defined as

$$\mathbf{p}_v = \begin{bmatrix} \sin\left(\frac{2\pi\iota(v)}{m^{(L)}}\right) & \cos\left(\frac{2\pi\iota(v)}{m^{(L)}}\right) & \sin\left(\frac{2\pi\iota(\text{pa}(v))}{m^{(L-1)}}\right) & \cos\left(\frac{2\pi\iota(\text{pa}(v))}{m^{(L-1)}}\right) \\ \cdots & \sin\left(\frac{2\pi\iota(\text{pa}^{(L-1)}(v))}{m^{(1)}}\right) & \cos\left(\frac{2\pi\iota(\text{pa}^{(L-1)}(v))}{m^{(1)}}\right) \end{bmatrix}^\top. \quad (\text{A.55})$$

Position-wise feed forward layer. Consider the feed forward network $\text{FF}^{(\ell)}$ of the ℓ -th block ($\ell = L, \dots, 1$), which computes $\mathbf{q}_v^{(\ell)}$ from $\mathbf{h}_v^{(\ell)}$. We will show that, for each $\mathbf{h}_v^{(\ell)} (v \in \mathcal{V}^{(L)})$, the network can identify its (ancestor's) rank $\iota = \iota(\text{pa}^{(L-\ell)}(v))$ and apply $\mathbf{f}_\iota^{(\ell)}$, which is a neural network approximation of $f_\iota^{(\ell)}$. The identification of $\iota(\text{pa}^{(L-\ell)}(v))$ can be implemented with no errors. Therefore, the feed forward layer at the ℓ -th layer yields

$$\mathbf{q}_v^{(\ell)} = \mathbf{h}_v^{(\ell)} + \mathbf{f}_{\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{h}_v^{(\ell)}), \quad v \in \mathcal{V}^{(L)}.$$

When $\mathbf{h}_v^{(\ell)} \approx h_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}$ for $v \in \mathcal{V}^{(L)}$ and $\mathbf{f}_\iota^{(\ell)} \approx f_\iota^{(\ell)}$, we have $\mathbf{q}_v^{(\ell)} \approx q_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}$.

We define a class of full connected networks with the ReLU activation as follows. For an l -dimensional vector $x \in \mathbb{R}^l$, we write $[x; 1] = (x_1, \dots, x_l, 1)^\top$.

Definition A.2 (A class of fully connected networks). For $J \in \mathbb{N}$, $\mathbf{j} = (j_1, \dots, j_{J+2}) \in \mathbb{N}^{J+2}$, and $B > 0$, we define a class of full connected networks with the ReLU activation as

$$\mathcal{F}(J, \mathbf{j}, B) = \left\{ \mathbf{W}^{(J+1)}[\cdot; 1] \circ \text{ReLU}(\mathbf{W}^{(J)}[\cdot; 1]) \circ \text{ReLU}(\mathbf{W}^{(J-1)}[\cdot; 1]) \circ \cdots \circ \text{ReLU}(\mathbf{W}^{(1)}[\cdot; 1]) \mid \right. \\ \left. \mathbf{W}^{(1)} \in \mathbb{R}^{j_2 \times (j_1+1)}, \mathbf{W}^{(2)} \in \mathbb{R}^{j_3 \times (j_2+1)}, \dots, \mathbf{W}^{(J+1)} \in \mathbb{R}^{j_{J+2} \times (j_{J+1}+1)}, \right. \\ \left. \max_{j \in [J+1]} \max_{k,l} |\mathbf{W}_{k,l}^{(j)}| \leq B \right\}.$$

Each element implements a function from $\mathbb{R}^{j_1}$ to $\mathbb{R}^{j_{J+2}}$.

We will show the following approximation error guarantee of $f_\iota^{(\ell)}$. Since it is easy to concatenate zeros to the first and last layer matrices and adjust the input and output dimensions to be $d_f + d_p$, the network NN in the following is presented as a function from $\mathbb{R}^S \times \mathbb{R}^{d_p}$ (or $[S] \times \mathbb{R}^{d_p}$ for $\ell = L$) to $\mathbb{R}^S$, focusing only on relevant dimensions. The proof will be found in Appendix A.5.3.

Lemma A.9 (Approximation error of feed forward layer). *Fix $\ell \in [L]$ and $\delta > 0$. Assume that $B_\psi^{-1} \leq \psi^{(\ell)}(s, a) \leq B_\psi$ for all $s, a \in [S]$. When $\ell = 1$, also assume that $B_\psi^{-1} \leq \mathbb{P}[s] \leq B_\psi$ for all s . Then, there exists an $\text{NN} \in \mathcal{F}(J, \mathbf{j}, B)$ such that*

$$\|\text{NN}([h; \mathbf{p}_v]) - f_{\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(h)\|_\infty \leq \delta, \quad v \in \mathcal{V}^{(L)},$$

for all $h \in \mathbb{R}^S$ with $\max_s h_s = 0$ ($\ell \leq L-1$) or $h \in [S]$ ($\ell = L$). The network parameters $J, \mathbf{j}$ and B are bounded as follows:

$$\begin{aligned} J &\lesssim (\log \log(SB_\psi/\delta)) \log(SB_\psi/\delta), \\ \|\mathbf{j}\|_\infty &\lesssim m^{(\ell)} S (\log(SB_\psi/\delta))^3 + L + d_p, \\ B &\lesssim 2S(B_\psi^2 + \log(SB_\psi/\delta)) + (m^{(\ell)})^2. \end{aligned}$$

The bound uses polylogarithmic depth with respect to the approximation error δ . It is known that deep neural networks can achieve significantly finer approximations (Schmidt-Hieber, 2020; Suzuki, 2018) than ones with constant depth (Telgarsky, 2016). Although this differs from real-world transformers, which use feed forward layers of constant depth, we can achieve the same result while keeping the feed forward layers of each block constant depth by using multiple blocks to approximate a single $f^{(\ell)}$ instead of increasing J (ignoring intermediate self-attention layers). We chose not to adopt such a way of presentation because we prioritized keeping the correspondence between the ℓ -th block of the transformer and the index ℓ in the message passing algorithm. Also, please refer to ‘‘Approximation with constant depth’’ paragraph Appendix A.5.3 for details on using feed forward layers of constant depth with L blocks.

Self-attention block. Consider the self-attention layer $\text{Attn}^{(\ell)}$ of the ℓ -th block ($\ell = L, \dots, 1$). Ignoring irrelevant dimensions, it takes $\mathbf{q}_v^{(\ell)}$ and $\mathbf{p}_v$ as inputs, and computes $\sum_{u \in \mathcal{C}(\text{pa}^{(L-\ell+1)}(v))} \mathbf{q}_u^{(\ell)}$ for each $v \in \mathcal{V}^{(L)}$. For each $v \in \mathcal{V}^{(L)}$, we denote the error by $\delta_v^{(\ell-1)} \in \mathbb{R}^S$. As a result, the self-attention block yields

$$\mathbf{h}_v^{(\ell-1)} = \text{normalize} \left(\sum_{\iota(\text{pa}^{(L-\ell')}(u)) = \iota(\text{pa}^{(L-\ell')}(v)) \ (\ell' \neq \ell)} \mathbf{q}_u^{(\ell)} + \delta_v^{(\ell-1)} \right).$$

Here $\text{normalize}(x)_s = x_s - \max x_{s'}$.

We explain interpretation of $\sum_{\iota(\text{pa}^{(L-\ell')}(u)) = \iota(\text{pa}^{(L-\ell')}(v)) \ (\ell' \neq \ell)} \mathbf{q}_u^{(\ell)}$. For each $v \in \mathcal{V}^{(L)}$, the nodes u that satisfy $\iota(\text{pa}^{(L-\ell')}(u)) = \iota(\text{pa}^{(L-\ell')}(v))$ ($\ell' \neq \ell$) are descendants of $\text{pa}^{(L-\ell+1)}(u) = \text{pa}^{(L-\ell+1)}(v)$ whose ancestors’ ranks are the same as v except for the ℓ -th level. Thus, for each $v' \in \mathcal{C}(\text{pa}^{(L-\ell+1)}(v))$, the summation selects exactly one of the descendants of v' . This implies that, when $\mathbf{q}_v^{(\ell)} \approx \mathbf{q}_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}$ for all $v \in \mathcal{V}^{(L)}$, we have

$$\mathbf{h}_v^{(\ell-1)} = \text{normalize} \left(\sum_{u \in \mathcal{C}(\text{pa}^{(L-\ell+1)}(v))} \mathbf{q}_{u^{(L)}}^{(\ell)} + \delta_v^{(\ell-1)} \right)$$

$$\approx \text{normalize}\left(\sum_{u \in \mathcal{C}(\text{pa}^{(L-\ell+1)}(v))} q_u^{(\ell)}\right) = h_{\text{pa}^{(L-\ell+1)}(v)}^{(\ell-1)}.$$

To further clarify the correspondence with the message passing, for $v \in \mathcal{V}^{(\ell-1)}$, this is simplified as

$$\mathbf{h}_{v^{(L)}}^{(\ell-1)} = \text{normalize}\left(\sum_{u \in \mathcal{C}(v)} \mathbf{q}_{u^{(L)}}^{(\ell)} + \boldsymbol{\delta}_{v^{(L)}}^{(\ell-1)}\right).$$

We define a class of self-attention block as follows.

Definition A.3 (A class of self-attention blocks). We define a class of self-attention blocks as

$$\mathcal{A}(D, B) = \left\{ (\mathbf{W}_V \cdot) \text{softmax}((\mathbf{W}_K \cdot)^\top (\mathbf{W}_Q \cdot)) \mid \mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_V \in \mathbb{R}^{D \times D}, \max_{i,j} |(\mathbf{W}_K)_{i,j}|, \max_{i,j} |(\mathbf{W}_Q)_{i,j}|, \max_{i,j} |(\mathbf{W}_V)_{i,j}| \leq B \right\}.$$

Each element implements a function that takes a matrix of size $D \times d'$ (d' : arbitrary) and maps it to a matrix of size $D \times d'$.

Then, we obtain the following approximation error guarantee. See Appendix A.5.4 for the proof.

Lemma A.10 (Approximation error of self-attention layer). *For $\ell \in [L]$, there exists $\text{Attn} \in \mathcal{A}(D, B)$ with $D = d_f + d_p$ and $B \lesssim \log(d\delta^{-1}) + m^{(\ell)}$ such that*

$$\text{Attn}(Q^{(\ell)}) = \left[\begin{array}{cc} \mathbf{0} & (\in \mathbb{R}^{(2\ell-2)S}) \\ \sum_{\iota(\text{pa}^{(L-\ell')}(u))=\iota(\text{pa}^{(L-\ell')}(v)) \ (\ell' \neq \ell)} \mathbf{q}_u^{(\ell)} + \boldsymbol{\delta}_v^{(\ell-1)} & (\in \mathbb{R}^S) \\ \mathbf{0} & (\in \mathbb{R}^{d_p+1+(2L-2\ell+1)S}) \end{array} \right]_{v \in \mathcal{V}^{(L)}},$$

where $\boldsymbol{\delta}_v^{(\ell-1)} \in \mathbb{R}^S$ satisfies $\|\boldsymbol{\delta}_v^{(\ell-1)}\|_\infty \leq \delta \max_{v'} \|\mathbf{q}_{v'}^{(\ell)}\|_\infty$.

Normalization. In the attention network, since column vectors of $\mathbf{H}^{(\ell)}$ and $Q^{(\ell)}$ are collections of multiple $\mathbf{h}_v^{(\ell)}$ and $\mathbf{q}_v^{(\ell)}$, we adopt a slightly different definition of “normalize” for these column vectors, from the one for S -dimensional vectors. Specifically, for $\mathbf{x} = [\mathbf{h}^{(0)} \ \mathbf{q}^{(1)} \ \mathbf{h}^{(1)} \ \dots \ \mathbf{q}^{(L)} \ \mathbf{h}^{(L)} \ \mathbf{p}] \in \mathbb{R}^{d_f+d_p}$ with $\mathbf{h}^{(L)} \in [S], \mathbf{h}^{(\ell)} \in \mathbb{R}^S$ ($\ell = L-1, \dots, 0$), and $\mathbf{q}^{(\ell)} \in \mathbb{R}^S$, we define

$$\text{normalize}(\mathbf{x}) = \left[\begin{array}{c} \mathbf{h}^{(0)} \\ \mathbf{q}^{(1)} - \mathbf{1}_S \max_{s \in S} \mathbf{q}_s^{(1)} \\ \mathbf{h}^{(1)} \\ \mathbf{q}^{(2)} - \mathbf{1}_S \max_{s \in S} \mathbf{q}_s^{(2)} \\ \vdots \\ \mathbf{q}^{(L)} - \mathbf{1}_S \max_{s \in S} \mathbf{q}_s^{(L)} \\ \mathbf{h}^{(L)} \\ \mathbf{p} \end{array} \right] \in \mathbb{R}^{d_f+d_p}, \quad \mathbf{1}_S = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^S. \quad (\text{A.56})$$

For a matrix with its column dimension $d_f + d_p$, it is applied in a column-wise manner.

Readout layer read. In the readout layer, we extract $\mathbf{h}_d^{(0)}$ as $\mathbf{h}_d^{(0)} = \text{read}(\mathbf{H}^{(0)}) = \text{read}(\text{TF}^{\mathbf{W}}(\text{Emb}(\mathbf{x})))$.

Similarity score. The link function τ^w is defined as

$$\tau^w(h, h') = \log(\text{trun}(\sum_{s \in [S]} h_s h'_s w_s)),$$

where

$$\text{trun}(z) := \text{proj}_{[-\exp(-B_{\text{read}}), \exp(B_{\text{read}})]}(z) \quad (\text{A.57})$$

is the function that projects z onto the interval $[\exp(-B_{\text{read}}), \exp(B_{\text{read}})]$ for any $z \in \mathbb{R} \cup \{-\infty\}$. We choose the threshold $B_{\text{read}} := 4\bar{m} \log B_\psi$. As shown in Lemma A.20, the threshold B_{read} is chosen sufficiently large to ensure the truncation does not occur in our construction (when the approximation error is sufficiently small). Thus, setting $w_s = \mathbb{P}[s]^{-1}$ yields the exact $f^{(0)}$, i.e., $\tau^w = f^{(0)}$ (see (A.53) to remember the definition of $f^{(0)}$). Under Assumption 2.5, this w_s satisfies $\|w\|_\infty \leq B_\psi$.

The whole pipeline. Putting it all together, the whole pipeline, starting from $\mathbf{h}_{\text{tx},v}^{(L)} = x_{\text{tx},v}$ ($v \in \mathcal{V}_{\text{tx}}$), is written as

$$\begin{aligned} \mathbf{q}_{\text{tx},v}^{(\ell)} &= \mathbf{f}_{\text{tx},\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{h}_{\text{tx},v}^{(\ell)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(L)}, \ell = L, \dots, 1, \\ \mathbf{h}_{\text{tx},v}^{(\ell-1)} &= \text{normalize} \left(\sum_{\iota(\text{pa}^{(L-\ell')}(u)) = \iota(\text{pa}^{(L-\ell')}(v)) \ (\ell' \neq \ell)} \mathbf{q}_u^{(\ell)} + \delta_v^{(\ell-1)} \right) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(L)}, \ell = L, \dots, 1. \end{aligned} \quad (\text{A.58})$$

We can write this alternatively to emphasize the connection to the message passing algorithm (A.51) and (A.52) (see “Self-attention block” paragraph).

$$\begin{aligned} \mathbf{q}_{\text{tx},v^{(L)}}^{(\ell)} &= \mathbf{f}_{\text{tx},\iota(v)}^{(\ell)}(\mathbf{h}_{\text{tx},v^{(L)}}^{(\ell)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(\ell)}, \ell = 1, 2, \dots, L, \\ \mathbf{h}_{\text{tx},v^{(L)}}^{(\ell-1)} &= \text{normalize} \left(\sum_{u \in \mathcal{C}(v)} \mathbf{q}_{\text{tx},u^{(L)}}^{(\ell)} + \delta_{\text{tx},v^{(L)}}^{(\ell-1)} \right) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(\ell-1)}, \ell = 1, 2, \dots, L. \end{aligned}$$

The image part is defined in the same way. Finally, with the link function τ^w that exactly represents $f^{(0)}$, we get

$$\mathbf{S}_{\text{NN}} = \tau^w(\text{softmax}(\mathbf{h}_{\text{im},d}^{(0)}), \text{softmax}(\mathbf{h}_{\text{tx},d}^{(0)})). \quad (\text{A.59})$$

Under Assumption 2.5 (because below we use B_ψ), the hypothesis class to which a tuple $(\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}, \text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}, \tau^w)$ belongs is defined as follows, formally restating (2.18).

Definition A.4 (Eq. (2.18), restated). We say the collection of the parameters of $(\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}, \text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}, \tau^w)$ belongs to $\Theta_{L,J,D,D',B}$ if the following holds: For transformer networks $\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}$ and $\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}$, they have L blocks of feed forward (Definition A.2), self-attention (Definition A.3), and normalization. In each block, its feed forward FF and self-attention Attn satisfy

$$\text{FF} \in \mathcal{F}(J, \mathbf{j} = (D, *, \dots, *, D), B), \text{ with } \|\mathbf{j}\|_\infty \leq D', \quad \text{Attn} \in \mathcal{A}(D, B).$$

For the link function τ^w , its weight satisfies $\|w\|_\infty \leq B$.

The subsequent sections consist as follows. Appendix A.5.2 combines Lemma A.9 and Lemma A.10, as well as the bound on the propagation of the intermediate errors (Lemma A.18) to prove Theorem 2.10. Appendix A.5.3 and Appendix A.5.4 will provide the proof of Lemma A.9 and Lemma A.10, respectively. The proof of the error propagation lemma (Lemma A.18) will be found in Appendix A.5.5. Appendix A.5.6 provides some useful properties about the message passing algorithm.

A.5.2 Proof of Theorem 2.10

We prove Theorem 2.10 by combining Lemma A.9, Lemma A.10, and Lemma A.18, as well as several auxiliary lemmas. By definition, the excess risk $\text{Excess}_K(\hat{\mathbf{S}}_{\text{NN}}, \mathbf{S}_\star)$ has the following decomposition:

$$\begin{aligned} \text{Excess}_K(\hat{\mathbf{S}}_{\text{NN}}, \mathbf{S}_\star) &= \bar{\mathbf{R}}(\hat{\mathbf{S}}_{\text{NN}}) - \bar{\mathbf{R}}(\mathbf{S}_\star) \\ &= \underbrace{\inf_{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B}} \bar{\mathbf{R}}(\mathbf{S}_{\text{NN}}^\boldsymbol{\theta}) - \bar{\mathbf{R}}(\mathbf{S}_\star)}_{\text{approximation error}} + \underbrace{\bar{\mathbf{R}}(\hat{\mathbf{S}}_{\text{NN}}) - \inf_{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B}} \bar{\mathbf{R}}(\mathbf{S}_{\text{NN}}^\boldsymbol{\theta})}_{\text{generalization error}}. \end{aligned}$$

We claim that

- (a). If we choose $J = \tilde{\mathcal{O}}(L)$, $D = \mathcal{O}(SL)$, $D' = \tilde{\mathcal{O}}(\bar{m}SL^3)$, and $B = \tilde{\mathcal{O}}(SL + \bar{m}^2)$, then the approximation error satisfies

$$\inf_{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B}} \bar{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^\boldsymbol{\theta}) - \bar{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_\star) \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{S^2 L^{11} \bar{m}^2}{n}}\right).$$

- (b). Under the same choice of model class $\Theta_{L,J,D,D',B}$, the generalization error satisfies

$$\bar{\mathbf{R}}(\hat{\mathbf{S}}_{\text{NN}}) - \inf_{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B}} \bar{\mathbf{R}}(\mathbf{S}_{\text{NN}}^\boldsymbol{\theta}) \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{S^2 L^{11} \bar{m}^2}{n}}\right)$$

with probability at least $1 - 1/n$.

Putting pieces together yields Theorem 2.10. The remainder of this section is devoted to proving these claims.

(a) Approximation error. Note that

$$\begin{aligned} &\bar{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^\boldsymbol{\theta}) - \bar{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_\star) \\ &\leq \mathbb{E} \left| \log \frac{\exp(\mathbf{S}_{\text{NN}}^\boldsymbol{\theta}(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},1}))}{\sum_{j \in [K]} \exp(\mathbf{S}_{\text{NN}}^\boldsymbol{\theta}(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},j}))} - \log \frac{\exp(\mathbf{S}_\star(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},1}))}{\sum_{j \in [K]} \exp(\mathbf{S}_\star(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},j}))} \right| \end{aligned}$$

$$\begin{aligned}
& + \mathbb{E} \left| \log \frac{\exp(\mathbf{S}_{\text{NN}}^\theta(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},1}))}{\sum_{j \in [K]} \exp(\mathbf{S}_{\text{NN}}^\theta(\mathbf{x}_{\text{im},j}, \mathbf{x}_{\text{tx},1}))} - \log \frac{\exp(\mathbf{S}_\star(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},1}))}{\sum_{j \in [K]} \exp(\mathbf{S}_\star(\mathbf{x}_{\text{im},j}, \mathbf{x}_{\text{tx},1}))} \right| \\
& \leq 2\mathbb{E} \max_{j \in [K]} \left| \mathbf{S}_{\text{NN}}^\theta(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},j}) - \mathbf{S}_\star(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},j}) \right| + 2\mathbb{E} \max_{j \in [K]} \left| \mathbf{S}_{\text{NN}}^\theta(\mathbf{x}_{\text{im},j}, \mathbf{x}_{\text{tx},1}) - \mathbf{S}_\star(\mathbf{x}_{\text{im},j}, \mathbf{x}_{\text{tx},1}) \right| \\
& \leq 4 \max_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \in \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}}} |\mathbf{S}_{\text{NN}}^\theta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})|,
\end{aligned}$$

where the second inequality follows from Lemma A.48. Therefore, it remains to find some parameter $\theta \in \Theta_{L,J,D,D',B}$ such that $\max_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \in \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}}} |\mathbf{S}_{\text{NN}}^\theta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})| \leq \tilde{O}(\sqrt{S^2 L^{11} \bar{m}^2 / n})$.

Take some $\delta' > 0$ which will be defined later. For the feed forward layers, we use Lemma A.9 with $\delta = \delta' \ll 1$. For the self-attention layers, we use Lemma A.10 with $\delta = \frac{\delta'}{\max_v \|\mathbf{q}_v^{(\ell)}\|_\infty}$. According to Lemma A.19, $f_{\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{h}_v^{(\ell)})$ are all bounded by $2 \log SB_\psi$ with the $\|\cdot\|_\infty$ -norm, and the approximation error from Lemma A.10 is δ . Thus $\mathbf{q}_v^{(\ell)} = f_{\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{h}_v^{(\ell)})$ is bounded by $2 \log SB_\psi + \delta \leq 3(1 \vee \log SB_\psi)$, and δ in Lemma A.10 is bounded by $\frac{\delta'}{3(1 \vee \log SB_\psi)}$.

The error from each operation is then bounded by δ' in the $\|\cdot\|_\infty$ -norm. Now we can apply Lemma A.18 to obtain that

$$\begin{aligned}
& \max_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \in \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}}} |\mathbf{S}_{\text{NN}}^\theta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) - \mathbf{S}_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})| \\
& \leq \delta' \times \left[\prod_{1 \leq \ell \leq L} (2m^{(\ell)} + 3) + \prod_{1 \leq \ell \leq L} (2m^{(\ell)} + 3) \right] \leq 2 \cdot 5^L \delta' d,
\end{aligned} \tag{A.60}$$

where we set $d = \max\{d_{\text{im}}, d_{\text{tx}}\}$.

Choose

$$\delta' = \frac{\sqrt{S^2 L^{11} \bar{m}^2}}{5^L d \sqrt{n}}$$

with $\bar{m} = \max\{\max_k m_{\text{tx}}^{(k)}, \max_k m_{\text{im}}^{(k)}\}$, so that the approximation error (A.60) is bounded by $\sqrt{\frac{S^2 L^{11} \bar{m}^2}{n}}$. According to Lemma A.9 and Lemma A.10, we now know that there exists some parameter $\theta \in \Theta_{L,J,D,D',B}$ such that Eq. (A.60) is satisfied, where

$$\begin{aligned}
D & \leq d_f + d_p = 2(S+1)L + 1 = \mathcal{O}(SL), \\
J & \lesssim (\log \log(SB_\psi/\delta')) \log(SB_\psi/\delta') = \tilde{\mathcal{O}}(L), \\
D' & = \|\mathbf{j}\|_\infty \lesssim \bar{m} S (\log(SB_\psi/\delta'))^3 + d_f + d_p = \tilde{\mathcal{O}}(\bar{m} S L^3), \\
B & \lesssim S(B_\psi^2 + \log(SB_\psi/\delta')) + \bar{m}^2 + \log \frac{d \log(SB_\psi)}{\delta'} = \tilde{\mathcal{O}}(SL + \bar{m}^2).
\end{aligned}$$

(b) Generalization error analysis. Since $\hat{\theta}$ is the minimizer of $\hat{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^\theta)$ defined in Eq. (2.17), we have

$$\bar{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^{\hat{\theta}}) - \inf_{\theta \in \Theta_{L,J,D,D',B}} \bar{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^\theta) \leq 2 \sup_{\theta \in \Theta_{L,J,D,D',B}} |\hat{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^\theta) - \bar{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^\theta)|. \tag{A.61}$$

Next, we verify the conditions required for Lemma A.51 and then apply the lemma to obtain an upper bound for the R.H.S. of Eq. (A.61).

In Lemma A.51, take $\Theta = \Theta_{L,J,D,D',B}$, $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') = \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$, $z_i = (\mathbf{x}_{\text{im},j}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)})_{j \in [K]}$, and

$$\begin{aligned} & f(z_i; \boldsymbol{\theta}) \\ &= -\frac{1}{K} \sum_{k=1}^K \log \frac{\exp(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))}{\sum_{j \in [K]} \exp(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}))} - \frac{1}{K} \sum_{k=1}^K \log \frac{\exp(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}'}(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))}{\sum_{j \in [K]} \exp(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}'}(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}))}. \end{aligned}$$

for $i = 1, \dots, n$.

Verification of condition (a) in Lemma A.51. We note that the set $\Theta_{L,J,D,D',B}$ with metric $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') = \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ has a diameter $B_\rho := 2B$. Moreover, $\Theta_{L,J,D,D',B}$ has a dimension bounded by $d_\rho := (J+3)L(D+D'+1)^2 + S = \tilde{\mathcal{O}}(S^2 L^8 \bar{m}^2)$. Thus, by Example 5.8 in Wainwright (2019b), we have $\log \mathcal{N}(\Delta; \Theta_{L,J,D,D',B}, \|\cdot\|) \leq d_\rho \log(1 + 2r/\Delta) \leq d_\rho \log(2A_\rho r/\Delta)$ for $\Delta \in (0, 2r]$ with $A_\rho = 2$.

Verification of condition (b) in Lemma A.51. Since $\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}$ is B_{read} -bounded with $B_{\text{read}} = 4\bar{m} \cdot \log B_\psi$ by Lemma A.41, it follows that $f(z_i; \boldsymbol{\theta}) - \mathbb{E}[f(z_i; \boldsymbol{\theta})]$ is $\sigma = cB_{\text{read}}$ -sub-Gaussian for all $\boldsymbol{\theta} \in \Theta_{L,J,D,D',B}$ for some numerical constant $c > 0$ from Lemma A.52.

Verification of condition (c) in Lemma A.51. By Lemmas A.41, A.48 and A.52, we have

$$\begin{aligned} & |f(z_i; \boldsymbol{\theta}) - f(z_i; \boldsymbol{\theta}')| \\ & \leq 4 \left| \mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},j}) - \mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}'}(\mathbf{x}_{\text{im},1}, \mathbf{x}_{\text{tx},j}) \right| + 4 \left| \mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im},j}, \mathbf{x}_{\text{tx},1}) - \mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}'}(\mathbf{x}_{\text{im},j}, \mathbf{x}_{\text{tx},1}) \right| \\ & \leq B_f \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \quad \text{where } B_f := 4((cB)^{18JL} S^4)^{L+1}. \end{aligned}$$

Therefore, we may choose $\sigma' = B_f$ and condition (c) is hence satisfied.

Now, invoking Lemma A.51 and plugging in the values of $d_\rho, \sigma, \sigma', A_\rho, B_\rho$, we find

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B}} |\hat{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}) - \bar{\mathbf{R}}_{\text{clip},K}(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}})| & \leq c\sigma \sqrt{\frac{d_\rho \log(2A_\rho(1 + B_\rho \sigma'/\sigma)) + \log(1/\eta)}{n}} \\ & \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{S^2 L^{11} \bar{m}^2 + \log(1/\eta)}{n}}\right) \end{aligned}$$

with probability at least $1 - \eta$. Setting $\eta = 1/n$ completes the proof.

Proof of Corollary 2.11

By Lemma A.20 and the definition of the readout function $\text{trun}(\cdot)$ in $\tau^w(\cdot)$, we have Assumption 2.1 is satisfied with $c_1 = (B_\psi)^{4\bar{m}}$. Thus, Corollary 2.11 follows immediately from combining Theorem 2.12 and Proposition 2.4.

A.5.3 Position-wise feed forward layer (proof of Lemma A.9)

Now we construct ReLU networks that approximate $f_{\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}$ to prove Lemma A.9. We first approximate each $f_{\iota}^{(\ell)}$ as follows. Lemma A.11 covers $\ell \leq L-1$ and Lemma A.12 covers $\ell = L$.

Lemma A.11 (ReLU network approximation of log-sum-exponential). *Fix $\ell \in [L-1]$ and $\delta > 0$. Assume that $B_{\psi}^{-1} \leq \psi_{\iota}^{(\ell)}(s, a) \leq B_{\psi}$ for all $s, a \in [S]$. When $\ell = 1$, also assume that $B_{\psi}^{-1} \leq \mathbb{P}[s] \leq B_{\psi}$ for all s . Then, there exists an $\text{NN} \in \mathcal{F}(J, \mathbf{j} = (S, \dots, S), B)$ such that*

$$\|\text{NN}(h) - f_{\iota}^{(\ell)}(h)\|_{\infty} \leq \delta, \text{ for all } h \in \mathbb{R}^S \text{ with } \max_s h_s = 0.$$

Here, the network parameters $J, \mathbf{j}$ and B are bounded by

$$J \lesssim (\log \log(SB_{\psi}/\delta)) \log(SB_{\psi}/\delta), \quad \|\mathbf{j}\|_{\infty} \lesssim S(\log(SB_{\psi}/\delta))^3, \quad B \lesssim 2S(B_{\psi}^2 + \log(SB_{\psi}/\delta)).$$

Lemma A.12 (ReLU network approximation of log-Psi). *Let $\delta > 0$. Assume that $B_{\psi}^{-1} \leq \psi_{\iota}^{(L)}(s, a) \leq B_{\psi}$ for all $s, a \in [S]$. Then, there exists an $\text{NN} \in \mathcal{F}(J, \mathbf{j} = (1, \dots, S), B)$ such that*

$$\|\text{NN}(s) - f_{\iota}^{(L)}(s)\|_{\infty} \leq \delta, \text{ for all } s \in [S].$$

Here, the network parameters $J, \mathbf{j}$ and B are bounded by

$$J \lesssim (\log \log(\log(B_{\psi})/\delta)) \log(\log(B_{\psi})/\delta), \quad \|\mathbf{j}\|_{\infty} \leq S(\log(\log(B_{\psi})/\delta))^2 \log B_{\psi}, \quad B \lesssim B_{\psi} + S.$$

Their proofs are deferred to Appendix A.5.3.

In addition to approximating a single $f_{\iota}^{(\ell)}$, the ReLU network should identify $\iota(\text{pa}^{(\ell)}(v))$ using the positional encoding and apply the correct $f_{\iota}^{(\ell)}$ to $\mathbf{h}_v^{(\ell)}$. The following lemma states that this can be implemented with no errors.

Lemma A.13 (Identify the rank from positional encoding). *Fix $\ell \in [L]$. Suppose that we have $m^{(\ell)}$ different networks $\text{NN}_1 \in \mathcal{F}(J_1, \mathbf{j}_1, B_1), \dots, \text{NN}_{m^{(\ell)}} \in \mathcal{F}(J_{m^{(\ell)}}, \mathbf{j}_{m^{(\ell)}}, B_{m^{(\ell)}})$ with the shared input and the same output dimension k , and the outputs of these networks are bounded by C with the $\|\cdot\|_{\infty}$ -norm. Then, for $v \in \mathcal{V}^{(L)}$, there exists a ReLU network NN that selects $\text{NN}_{\iota(\text{pa}^{(L-\ell)}(v))}$ given $\mathbf{p}_v$, i.e.,*

$$\text{NN}([h; \mathbf{p}_v]) = \text{NN}_{\iota(\text{pa}^{(L-\ell)}(v))}(h).$$

This network satisfies that

$$J = \max_i J_i, \quad \|\mathbf{j}\|_{\infty} \leq m^{(\ell)} + 2 \sum_i^{m^{(\ell)}} \|\mathbf{j}_i\|, \quad B = \max_i B_i + (m^{(\ell)})^2 + C.$$

Using Lemma A.13 together with Lemmas A.11 and A.12, for any $\ell \in [S]$, there exists a ReLU network such that, for all $v \in \mathcal{V}^{(L)}$, it takes $\mathbf{h}_v^{(\ell)}$ and $\mathbf{p}_v^{(\ell)}$ and outputs $f_{\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{h}_v^{(\ell)})$ with the $\|\cdot\|_\infty$ -error at most δ , where

$$\begin{aligned} J &\lesssim (\log \log(SB_\psi/\delta)) \log(SB_\psi/\delta), \quad \|\mathbf{j}\|_\infty \lesssim m^{(\ell)} S(\log(SB_\psi/\delta))^3, \\ B &\lesssim S(B_\psi^2 + \log(SB_\psi/\delta)) + (m^{(\ell)})^2. \end{aligned}$$

Note that $f_{\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}$ (and thus its neural network approximation) is bounded by $O(\log(SB_\psi))$ according to Lemma A.19, which gives $C \lesssim \log(SK)$ in the application of Lemma A.13. Now, we have obtained Lemma A.9.

Approximation with constant depth. While we used the networks with polylogarithmic depth, we add a remark on how the analysis changes when we use networks with constant depth.

The networks that approximate basic functions are changed as follows:

(Approximation of logarithm function.) There exists a network that achieves the same bound as Lemma A.14, where

$$J = 1, \quad \|\mathbf{j}\|_\infty \leq \lceil 2A/\delta \rceil + 1, \quad B \leq e^A.$$

This two-layer approximation is obtained as a modification of Lemma 9 of Mei (2024), where we choose $e_j = 2Aj/(M-1) - A$, $b_j = -\exp(e_j)$ for $j \in [M-1]$, $a_1 = (e_2 - e_1)/(b_2 - b_1)$ and $a_j = (e_{j+1} - e_j)/(b_{j+1} - b_j) - (e_j - e_{j-1})/(b_j - b_{j-1})$ for $2 \leq j \leq M-2$ to obtain $\text{NN}_{\log}(x) = \sum_{j=1}^{M-2} a_j \text{ReLU}(x + b_j) + \text{ReLU}(-x + e_1) - \text{ReLU}(-x)$.

(Approximation of exponential function) According to Lemma 8 of Mei (2024), there exists a network that achieves the same bound as Lemma A.15, where

$$J = 1, \quad \|\mathbf{j}\|_\infty = \lceil \delta^{-1} \rceil + 1, \quad B \leq \log(\lceil \delta^{-1} \rceil + 1),$$

By using these networks, we can obtain Lemma A.9 with the following bound:

$$J = 3, \quad \|\mathbf{j}\|_\infty \lesssim m^{(\ell)} S^2 B_\psi^5 \delta^{-1} \log(SB_\psi), \quad B \lesssim (m^{(\ell)})^3 S^5 B_\psi^8 \delta^{-1} (\log(SB_\psi/\delta))^2.$$

The problem here is that the dependency on δ is δ^{-1} , while in the original bound it was polylogarithmic. In the proof of Theorem 2.10, we need to take $\delta \lesssim d^{-1}$. Then the parameter $\|\mathbf{j}\|_\infty$ linearly depends on d , which incurs linear dependency on d in the generalization error bound. This problem is avoided when we are allowed to have polylogarithmic depth as in the original Lemma A.9 (or polylogarithmic number of blocks with each feed forward layer being constant depth).

Proofs of Lemmas A.11, A.12, and A.13

Lemmas A.11, A.12, and A.13 compose modules that approximate basic functions such as logarithm and exponential. The proofs of these basic modules will be deferred to Appendix A.5.3. First we review basic operations which we will use without a proof (borrowed from Appendix B.1.1. of Nakada and Imaizumi (2020)).

Composition of two networks. When we want to construct a composition of two networks $\text{NN}_1 \in \mathcal{F}(J_1, \mathbf{j}_1 = (\dots, j), B_1)$ and $\text{NN}_2 \in \mathcal{F}(J_2, \mathbf{j}_2 = (j, \dots), B_2)$, a naive way is to multiply the last layer's matrix of one network with the first layer's matrix with the other, where the parameter bound B_{1+2} of the new network is jB_1B_2 . However, the following construction can bound B_{1+2} more tightly with one additional layer. Let $\mathbf{W}_i^{(k)}$ be the parameters of the k th layer of NN_i ($i = 1, 2$). We define

$$\begin{aligned} \text{NN}_{1+2} = & \mathbf{W}_2^{(J+1)} \text{ReLU}(\mathbf{W}_2^{(J)}[:, 1]) \circ \dots \circ \text{ReLU}([\mathbf{W}_2^{(1)} \quad \widetilde{\mathbf{W}}_2^{(1)}][:, 1]) \\ & \circ \text{ReLU}\left(\begin{bmatrix} \mathbf{W}_1^{(J+1)} \\ -\mathbf{W}_1^{(J+1)} \end{bmatrix}[:, 1]\right) \circ \dots \circ \text{ReLU}(\mathbf{W}_1^{(1)}[:, 1]). \end{aligned}$$

Here $\widetilde{\mathbf{W}}_2^{(1)}$ is a matrix such that $(\widetilde{\mathbf{W}}_2^{(1)})_{k,l} = -(\mathbf{W}_2^{(1)})_{k,l}$ for all k, l , except that $(\widetilde{\mathbf{W}}_2^{(1)})_{k,l} = (\mathbf{W}_2^{(1)})_{k,l}$ in the column corresponding to the bias term. It is easy to check that NN_{1+2} implements the composition of NN_1 and NN_2 , considering that either the first half or the latter half columns of $\text{ReLU}\left(\begin{bmatrix} \mathbf{W}_1^{(J+1)} \\ -\mathbf{W}_1^{(J+1)} \end{bmatrix}[:, 1]\right)$ is zero. Moreover, we have $\text{NN}_{1+2} \in \mathcal{F}(J_{1+2}, \mathbf{j}_{1+2}, B_{1+2})$ with $J_{1+2} = J_1 + J_2 + 1$, $\|\mathbf{j}_{1+2}\|_\infty \leq 2 \max\{\|\mathbf{j}_1\|_\infty, \|\mathbf{j}_2\|_\infty\}$, and $B_{1+2} \leq \max\{B_1, B_2\}$.

Identify function. The identity function for d -dimensional inputs is implemented as a ReLU network with arbitrary depth:

$$\begin{bmatrix} \mathbf{I}_d \\ -\mathbf{I}_d \end{bmatrix} \text{ReLU}\left(\begin{bmatrix} \mathbf{I}_d & 0 \\ 0 & \mathbf{I}_d \end{bmatrix} \cdot\right) \circ \dots \circ \text{ReLU}\left(\begin{bmatrix} \mathbf{I}_d & 0 \\ 0 & \mathbf{I}_d \end{bmatrix} \cdot\right) \text{ReLU}\left(\begin{bmatrix} \mathbf{I}_d & -\mathbf{I}_d \end{bmatrix} \cdot\right).$$

Parallelization. When there are multiple networks $\text{NN}_i \in \mathcal{F}(J_i, \mathbf{j}_i, B_i)$ ($i = 1, \dots, I$) that share the input x , we can construct a larger network $\text{NN} \in \mathcal{F}(J, \mathbf{j}, B)$ that outputs $[\text{NN}_1; \dots; \text{NN}_I]$, where $J = \max_i J_i$, $\|\mathbf{j}\|_\infty \leq 2 \sum_{i=1}^I \|\mathbf{j}_i\|_\infty$, and $B \leq \max_i B_i$. Specifically, we first unify the depths of these networks by composing an identity function with each NN_i . We then concatenate these networks, by making a block diagonal matrix where the block diagonal parts are matrices of the original networks.

Now we provide the proofs of Lemmas A.11, A.12, and A.13 in order.

Proof of Lemma A.11. We focus on the case of $\ell = 1$, as the proof for $\ell \geq 2$ follows similarly (just delete all the $\mathbb{P}[s]^{\frac{1}{m^{(1)}}}$ terms). We utilize two ReLU networks $\text{NN}_{\log}(x)$ and $\text{NN}_{\exp}(x)$, defined in Lemma A.14 and Lemma A.15. We will determine the values of δ' and A later.

- $\text{NN}_{\log}(x)$, which approximates $\log(x)$ within the error of δ' for $e^{-A} \leq x \leq e^A$, with $J \lesssim (\log \log(A/\delta')) \log(A/\delta')$, $\|\mathbf{j}\|_\infty \lesssim A(\log(A/\delta'))^2$, and $B \leq e^A$ (see Lemma A.14 for construction).
- $\text{NN}_{\exp}(x)$, which approximates $\log(x)$ within the error of δ' for $x \leq 0$, $J \lesssim \log \log(1/\delta') \cdot \log(1/\delta')$, $\|\mathbf{j}\|_\infty \lesssim (\log(1/\delta'))^3$, and $B \lesssim \log(1/\delta')$ (see Lemma A.15 for construction).

We define NN_1 and NN_2 by parallelizing S instances of $\text{NN}_{\exp}$ and $\text{NN}_{\log}$, respectively.

The function we want to implement is $f_t^{(1)}$, which is

$$f_t^{(1)}(h)_s = \log \sum_{a \in [S]} \mathbb{P}[s]^{\frac{1}{m(1)}} \psi_t^{(1)}(s, a) e^{h_a}, \quad h \in \mathbb{R}^S.$$

Therefore, we combine NN_1 , $\Psi = (\mathbb{P}[s]^{\frac{1}{m(1)}} \psi_t^{(1)}(s, a))_{s,a} \in \mathbb{R}^{S \times S}$, and NN_2 to yield the desired network. The network parameters are bounded by

$$\begin{aligned} J &\lesssim (\log \log(A/\delta')) \log(A/\delta'), \quad \|\mathbf{j}\|_\infty \leq SA(\log(A/\delta'))^2 + S(\log(1/\delta'))^3, \\ B &\lesssim e^A + \log(1/\delta') + B_\psi^2. \end{aligned}$$

Let us determine the value of δ' and A . Note that, for $h \in \mathbb{R}^S$ with $\max_a h_a = 0$, we have

$$B_\psi^{-2}(1 - \delta') \leq \sum_{a \in [S]} \mathbb{P}[s]^{\frac{1}{m(1)}} \psi_t^{(1)}(s, a) \text{NN}_{\exp}(h_a) \leq SB_\psi^2(1 + \delta'). \quad (\text{A.62})$$

Because we will take $\delta' \ll 1$, we can assume that (A.62) is bounded by $SB_\psi^2(1 + \delta') \leq 2SB_\psi^2$ and $B_\psi^{-2}(1 - \delta') \geq (2B_\psi^2)^{-1}$. Thus we let $A = \log(2SK^2)$ in the definition of $\text{NN}_{\log}(x)$. Also, the overall error $\|\text{NN}(h) - f_t^{(1)}(h)\|_\infty$ is bounded by

$$\delta' + \frac{B_\psi^2}{1 - \delta'} (2SB_\psi^2) \delta'. \quad (\text{A.63})$$

Here, the first δ' is the approximation error of $\log$, and $\frac{B_\psi^2}{1 - \delta'}$ is the smoothness of $\log(t)$ in $B_\psi^{-2}(1 - \delta') \leq t$, $2SB_\psi^2$ is the amplification rate of the approximation error of $\exp$ (by $\sum_{a \in [S]} \mathbb{P}[s]^{\frac{1}{m(1)}} \psi_t^{(1)}(s, a)$), and the final δ' corresponds to the approximation error of $\exp$. It suffices to take $\delta' \leq \frac{\delta}{8SB_\psi^4}$ to achieve (A.63) $\leq \delta$.

Now, evaluating the parameters of the desired network with these δ' and A , we obtain the desired bound. $\square$

Proof of Lemma A.12. We use the following basic networks.

- $\text{NN}_{\log}(x)$ from Lemma A.14, which approximates $\log(x)$ within the error of δ for $e^{-A} \leq x \leq e^A$, with $J \lesssim (\log \log(A/\delta)) \log(A/\delta)$, $\|\mathbf{j}\|_\infty \lesssim A(\log(A/\delta))^2$, and $B \leq e^A$.
- $\text{NN}_{\text{Ind}[s]}(x)$ from Lemma A.16, which implements $\mathbb{1}[x = s]$ exactly, whose parameters are bounded by $J = 1$, $\|\mathbf{j}\|_\infty \lesssim S$, and $B = S + 1$.

We define NN_1 by parallelizing S instances of $\text{NN}_{\log}$, while the input x is shared. Also, we define NN_2 by parallelizing $\text{NN}_{\text{Ind}[s]}$ ($s = 1, \dots, S$).

The function we want to implement is $f_t^{(L)}$, which is

$$f_t^{(L)}(x) = \log \sum_{s \in [S]} \psi_t^{(L)}(s, a) \mathbb{1}[x = s], \quad x \in [S].$$

By combining NN_1 , a matrix $\Psi = (\psi_t^{(L)}(s, a))_{s,a} \in \mathbb{R}^{S \times S}$, and NN_2 , The network parameters are bounded by

$$J \lesssim (\log \log(A/\delta)) \log(A/\delta), \quad \|\mathbf{j}\|_\infty \lesssim SA(\log(A/\delta))^2, \quad B \lesssim e^A + B_\psi + S.$$

Finally, let us determine the value of A . Because it holds that

$$B_\psi^{-1} \leq \sum_{a \in [S]} \psi_t^{(\ell)}(s, a) \text{NN}_{\text{Ind}}(s) \leq B_\psi.$$

It suffices to take $A = \log(B_\psi)$. Also, because the computation of $\sum_{a \in [S]} \psi_t^{(\ell)}(s, a) \text{NN}_{\text{Ind}}(s)$ is exact, the approximation error only comes from $\text{NN}_{\log}$. Thus, the approximation error is bounded by δ .

Now, evaluating the parameters of the desired network with $A = \log(B_\psi)$, we have the desired bound. $\square$

Proof of Lemma A.13. Suppose that we have a network $\overline{\text{NN}} \in \mathcal{F}(\bar{J}, \bar{\mathbf{j}}, \bar{B})$ that takes $[h; \mathbf{p}_{v^{(L)}}]$ and outputs

$$[\text{NN}_1(h) \quad \dots \quad \text{NN}_{m^{(\ell)}}(h) \quad \mathbb{1}[\iota(\text{pa}^{(L-\ell)}(v)) = 1] \quad \dots \quad \mathbb{1}[\iota(\text{pa}^{(L-\ell)}(v)) = m^{(\ell)}]]^\top. \quad (\text{A.64})$$

Then we compose this with a one layer ReLU network, with the first layer matrix

$$\left[\begin{array}{c|cccc} & \mathbf{0}_k & -C\mathbf{1}_k & \dots & -C\mathbf{1}_k \\ \mathbf{I}_{km^{(\ell)}} & -C\mathbf{1}_k & \mathbf{0}_k & \dots & -C\mathbf{1}_k \\ & \vdots & \vdots & \ddots & \vdots \\ & -C\mathbf{1}_k & -C\mathbf{1}_k & \dots & \mathbf{0}_k \\ \hline & \mathbf{0}_k & -C\mathbf{1}_k & \dots & -C\mathbf{1}_k \\ -\mathbf{I}_{km^{(\ell)}} & -C\mathbf{1}_k & \mathbf{0}_k & \dots & -C\mathbf{1}_k \\ & \vdots & \vdots & \ddots & \vdots \\ & -C\mathbf{1}_k & -C\mathbf{1}_k & \dots & \mathbf{0}_k \end{array} \right] \in \mathbb{R}^{2km^{(\ell)} \times (k+1)m^{(\ell)}}, \quad (\mathbf{0}_k, \mathbf{1}_k \in \mathbb{R}^k), \quad (\text{A.65})$$

and the first layer bias $\mathbf{0}$, and the second layer matrix

$$[\mathbf{I}_{km^{(\ell)}} \quad -\mathbf{I}_{km^{(\ell)}}].$$

In (A.65), applying the left columns to (A.64) yields $[\text{NN}_1(\mathbf{x}) \quad \dots \quad \text{NN}_{m^{(\ell)}}(\mathbf{x}) - \text{NN}_1(\mathbf{x}) \quad \dots - \text{NN}_{m^{(\ell)}}(\mathbf{x})]^\top$. From the boundedness assumption, each element of the obtained vector is in $[-C, C]$. On the other hand, the right columns yields $-C$ in all coordinates but those

corresponding to $\text{NN}_{\iota(\text{pa}^{(L-\ell)}(v))}(h)$ and $-\text{NN}_{\iota(\text{pa}^{(L-\ell)}(v))}(h)$. After applying the ReLU and the second layer matrix, we can single out only one $\text{NN}_{\iota(\text{pa}^{(L-\ell)}(v))}(\mathbf{x})$.

Therefore, when we have a network $\overline{\text{NN}} \in \mathcal{F}(\bar{J}, \bar{\mathbf{j}}, \bar{B})$ that computes (A.64), we get the desired network with size

$$J = \bar{J} + 2, \quad \|\mathbf{j}\|_\infty \lesssim \|\bar{\mathbf{j}}\|, \quad B = \bar{B} + C. \quad (\text{A.66})$$

Finally, let us construct $\overline{\text{NN}}$ to bound its network parameters $\bar{J}, \bar{\mathbf{j}}, \bar{B}$. We use the $2(L - \ell + 1)$ th dimension of $\mathbf{p}_v$, which is $\cos(\frac{2\pi\iota(v)}{m^\ell})$, to identify the correct rank ι . According to Lemma A.16, there exists a ReLU network that implements $\mathbb{1}[x = \cos(\frac{2\pi\iota}{m^\ell})]$ for each $\iota = 1, 2, \dots, m^{(\ell)}$, where $J = 1$, $\|\mathbf{j}\|_\infty = 3$, and $B = m^{(\ell)} + 2\delta^{-1}$. We need to take $\delta = \min_i |\cos(\frac{2\pi i}{m^\ell}) - \cos(\frac{2\pi(i+1)}{m^\ell})| = 2\sin^2(\frac{2\pi}{m^\ell})$. By parallelizing this, there exists a ReLU network with $J = 1$, $\|\mathbf{j}\|_\infty = 3m^{(\ell)}$, and $B = m^{(\ell)} + 4\sin^{-2}(\frac{2\pi}{m^\ell})$, that takes $\mathbf{p}_v$ and outputs an $m^{(\ell)}$ -dimensional vector $(\mathbb{1}[\iota(\text{pa}^{(L-\ell)}(v)) = 1], \dots, \mathbb{1}[\iota(\text{pa}^{(L-\ell)}(v)) = m^{(\ell)}])$. Concatenating this indicator network and $\text{NN}_1, \dots, \text{NN}_{m^{(\ell)}}$, we have $\overline{\text{NN}}$ with $\bar{J} = \max_i J_i$, $\|\bar{\mathbf{j}}\|_\infty = m^{(\ell)} + 2\sum_{i=1}^{m^{(\ell)}} \|\mathbf{j}_i\| + 3$, and $\bar{B} = \max_i B_i + 2m^{(\ell)} + 4\sin^{-2}(\frac{2\pi}{m^\ell})$. Putting these bounds into (A.66) yields the desired bound. $\square$

ReLU network approximation of basic functions

Here we construct ReLU networks that approximate basic functions for Appendix A.5.3.

Lemma A.14 (ReLU network approximation of logarithm function). *For any $A \in \mathbb{N}$, $\delta > 0$, there exists a network $\text{NN}_{\log}(x): \mathbb{R} \rightarrow \mathbb{R}$ that approximates $\log(x)$ within the error of δ for all $x \in [e^{-A}, e^A]$, and that belongs to $\mathcal{F}(J, \mathbf{j} = (1, j_2, j_3, \dots, 1), B)$, where*

$$J = 4 + (\lceil \log_2(12A \lceil \log_2(6A/\delta) \rceil^2 / \delta) \rceil + 5) \lceil \log_2(\lceil \log_2(6A/\delta) \rceil) \rceil, \quad \|\mathbf{j}\|_\infty \leq 36A \lceil \log_2(6A/\delta) \rceil^2, \\ B \leq e^A.$$

Moreover, the network satisfies $-A - \delta \leq \text{NN}_{\log}(x) \leq A + \delta$ for all $x \in \mathbb{R}$.

Proof. (1) **Piece-wise polynomial approximation.** Let us define $p_0 = e^{-A}, p_1 = e^{-A+\frac{1}{3}}, p_2 = e^{-A+\frac{2}{3}}, \dots, p_{6A} = e^A$. By defining $q_0 = -A$ and

$$q_i(x) = \log(\min\{\max\{x, p_{i-1}\}, p_i\}) - \log p_{i-1} \\ = \text{ReLU}(\log(-\text{ReLU}(-\text{ReLU}(x - p_{i-1}) - p_{i-1} + p_i) + p_i) - \log p_{i-1}),$$

we have

$$\log(x) = \sum_{i=0}^{6A} q_i(x), \quad e^{-A} \leq x \leq e^A,$$

$$\sum_{i=0}^{6A} q_i(x) = -A \text{ when } x \leq e^{-A}, \text{ and } \sum_{i=0}^{6A} q_i(x) = A \text{ when } e^A \leq x.$$

For $1 \leq i \leq 6A$, consider the Taylor expansion of $\log(x) - \log p_{i-1}$ at $p_{i-1} = e^{-A + \frac{i-1}{3}}$ as

$$\log(x) - \log p_{i-1} = \sum_{k=1}^K \frac{(-1)^{k+1}}{k} \left(\frac{x - p_{i-1}}{p_{i-1}} \right)^k + \frac{(-1)^{K+2}}{K+1} \left(\frac{y(K, y) - p_{i-1}}{p_{i-1}} \right)^{K+1},$$

where $p_{i-1} \leq y(x, K) \leq x$. When $p_{i-1} \leq x \leq p_i$, the approximation error by the first K terms is bounded by

$$\left| \frac{(-1)^{K+2}}{K+1} \left(\frac{y(K, y) - p_{i-1}}{p_{i-1}} \right)^{K+1} \right| \leq \frac{1}{K+1} (e^{1/3} - 1)^{K+1} \leq 2^{-(K+1)}. \quad (\text{A.67})$$

Let $r_{i,k}(x) = \frac{(-1)^{k+1}}{k} \left(\frac{x - p_{i-1}}{p_{i-1}} \right)^k$ ($1 \leq k \leq K$). If $r_{i,k}(x)$ is approximated by a function $\tilde{r}_{i,k}$ in $p_{i-1} \leq x \leq p_i$ within the error of δ' , then

$$\text{ReLU} \left(q_0 + \sum_{i=1}^{6A} \text{ReLU} \left(\sum_{k=1}^K \tilde{r}_{i,k} (-\text{ReLU}(-\text{ReLU}(x - p_{i-1}) - p_{i-1} + p_i) + p_i) - \log p_{i-1} \right) \right) \quad (\text{A.68})$$

approximates $\log(x)$ ($e^{-A} \leq x \leq e^A$) within the error of

$$6AK\delta' + 6A2^{-(K+1)}. \quad (\text{A.69})$$

Here the first term comes from approximation of $r_{i,k}$ and the second term from Taylor expansion (A.67). Also, $-A - (\text{A.69}) \leq (\text{A.68}) \leq A + (\text{A.69})$ holds for all x . Thus, from now, our goal is to construct ReLU networks that approximate $r_{i,k}(x)$ within the error of δ' .

If this goal is achieved, we take

$$K = \lceil \log_2(6A/\delta) \rceil,$$

and

$$\delta' = \frac{\delta}{12AK} = \frac{\delta}{12A \lceil \log_2(6A/\delta) \rceil}$$

so that the approximation error (A.69) of $\log x$ by (A.68) is bounded by δ .

(2) ReLU network approximation of monomials. According to Lemma A.4 of (Schmidt-Hieber, 2020) (focusing on only one α), there exists a neural network $\text{Mult}_m^k(x)$ belonging to $\mathcal{F}(1 + (m+5)\lceil \log_2 k \rceil, (1, 6k, 6k, \dots, 6k, 1), 1)$ such that

$$\sum_{0 \leq x \leq 1} |\text{Mult}_m^k(x) - x^k| \leq k^2 2^{-m}.$$

Then, because $p_{i-1} \leq x \leq p_i$ implies $0 \leq x/p_{i-1} - 1 \leq 1$, we have

$$\sum_{p_{i-1} \leq x \leq p_i} \left| \frac{(-1)^{k+1}}{k} \text{Mult}_m^k(x/p_{i-1} - 1) - r_{i,k}(x) \right| \leq k^2 2^{-m}. \quad (\text{A.70})$$

We take $m = \lceil \log_2(K/\delta') \rceil = \lceil \log_2(12A \lceil \log_2(6A/\delta) \rceil^2 / \delta) \rceil$ so that (A.70) is bounded by δ' for all $i = 1, \dots, 6A$ and $k = 1, 2, \dots, K$.

We now know that there exists a network belonging to $\mathcal{F}(1 + (m+5)\lceil \log_2 k \rceil, (1, 6k, 6k, \dots, 6k, 1), e^A)$ that approximates $r_{i,k}$ within the error of δ' . As a result, (A.68) using these networks yields the desired network belonging to $\mathcal{F}(4 + (m+5)\lceil \log_2 K \rceil, \mathbf{j} = (1, \dots, 1), e^A)$, where $\|\mathbf{j}\|_\infty \leq 36AK^2$. $\square$

Lemma A.15 (ReLU network approximation of exponential function). *For any $\delta > 0$, there exists a network $\text{NN}_{\text{exp}}(x): \mathbb{R} \rightarrow \mathbb{R}$ that approximates $\exp(x)$ within the error of δ for all $x \leq 0$, and that belongs to $\mathcal{F}(J, \mathbf{j} = (1, j_2, j_3, \dots, 1), B)$, where*

$$\begin{aligned} J &= 4 + (\lceil \log_2(8 \lceil \log_2(4 \lceil \log 2\delta^{-1} \rceil / \delta) \rceil^2 \lceil \log 2\delta^{-1} \rceil / \delta) \rceil + 5) \lceil \log_2(\lceil \log_2(4 \lceil \log 2\delta^{-1} \rceil / \delta) \rceil) \rceil, \\ \|\mathbf{j}\|_\infty &\leq 12 \lceil \log 2\delta^{-1} \rceil \lceil \log_2(4 \lceil \log 2\delta^{-1} \rceil / \delta) \rceil^2, \\ B &\leq \lceil \log 2\delta^{-1} \rceil \vee 2. \end{aligned}$$

Moreover, the network satisfies $-\delta \leq \text{NN}_{\text{exp}}(x) \leq 1 + \delta$ for all $x \in \mathbb{R}$.

Proof. The proof basically follows that of Lemma A.14. We will show how to obtain the counterpart of (A.68), and omit the rest.

Let $p_0 = -\lceil \log 2\delta^{-1} \rceil$, $p_1 = p_0 + \frac{1}{2}$, $p_2 = p_1 + 1, \dots, p_A = 0$ with $A = 2\lceil \log 2\delta^{-1} \rceil$. By defining $q_0 = e^{-p_0}$ and

$$\begin{aligned} q_i(x) &= \exp(\min\{\max\{x, p_{i-1}\} p_i\}) - \exp(p_{i-1}) \\ &= \text{ReLU}(\exp(-\text{ReLU}(-\text{ReLU}(x - p_{i-1}) - p_{i-1} + p_i) + p_i) - \log p_{i-1}), \end{aligned}$$

we have

$$\exp(x) = \sum_{i=0}^A q_i(x), p_0 \leq x \leq 0,$$

$$\sum_{i=0}^A q_i(x) = e^{p_0} \leq \frac{\delta}{2} \text{ when } x \leq p_0, \text{ and } \sum_{i=0}^A q_i(x) = 0 \text{ when } 0 \leq x.$$

For $1 \leq i \leq A$, we consider Taylor expansion of $\exp(x) - \exp(p_{i-1})$ as

$$\exp(x) - \exp(p_{i-1}) = \exp(p_{i-1}) \left[\sum_{k=1}^K \frac{(x - p_{i-1})^k}{k!} + \frac{(y(K, x) - p_{i-1})^{K+1}}{(K+1)!} \right],$$

where $p_{i-1} \leq y(K, x) \leq p_i$. Thus, the approximation error by the first K terms is bounded by $2^{-(K+1)}$.

Let $r_{i,k}(x) = \frac{\exp(p_{i-1})}{k!} (x - p_{i-1})^k$ ($1 \leq k \leq K$). Then, if $r_{i,k}(x)$ is approximated by a function $\tilde{r}_{i,k}$ in $p_{i-1} \leq x \leq p_i$ within the error of δ' ,

$$\text{ReLU}\left(q_0 + \sum_{i=1}^{6A} \sum_{k=1}^K \tilde{r}_{i,k}(-\text{ReLU}(-\text{ReLU}(x - p_{i-1}) - p_{i-1} + p_i) + p_i) - \log p_{i-1}\right) \quad (\text{A.71})$$

approximates $\exp(x)$ ($x \leq 0$) within the error of

$$\frac{\delta}{2} + 2AK\delta' + 2A2^{-(K+1)}.$$

Also, (A.71) $\leq 1 + \delta$ for all x .

The rest of the argument follows that of Lemma A.14. Specifically, we take $K = \lceil \log_2(2A/\delta) \rceil$ and $\delta' = \frac{\delta}{4AK} = \frac{\delta}{8\lceil \log_2(2A/\delta) \rceil \lceil \log 2\delta^{-1} \rceil}$ in the part (2) of Lemma A.14, and all the others are identical. $\square$

Lemma A.16 (ReLU approximation of indicator function). *Let $a \in \mathbb{R}$, and $\delta > 0$. A one-layer neural network $\text{NN}_{\mathbb{1}[a]}$ defined by*

$$\text{NN}_{\mathbb{1}[a]}(x) = \frac{1}{\delta} \text{ReLU}(x - (a - \delta)) + \frac{1}{\delta} \text{ReLU}(x - (a + \delta)) - \frac{2}{\delta} \text{ReLU}(x - \delta),$$

satisfies

$$\text{NN}_{\mathbb{1}[a]}(x) = \mathbb{1}[x = a], \quad \text{for all } x \text{ such that } x \leq a - \delta, x = a, \text{ or } x \geq a + \delta.$$

Proof of Lemma A.16. The lemma holds by direct calculation. $\square$

A.5.4 Self-attention layer (proof of Lemma A.10)

We use the following lemma to prove Lemma A.10.

Lemma A.17. *Fix $\ell \in [L]$. There exist matrices $\overline{\mathbf{W}}_K^{(\ell)}, \overline{\mathbf{W}}_Q^{(\ell)} \in \mathbb{R}^{d_p \times d_p}$ with $\max_{i,j} |(\overline{\mathbf{W}}_Q^{(\ell)})_{i,j}|, \max_{i,j} |(\overline{\mathbf{W}}_K^{(\ell)})_{i,j}| \leq \log \frac{d}{\delta(1 - \cos(2\pi d^{-1}))}$ such that*

$$|(\text{softmax}((\overline{\mathbf{W}}_K^{(\ell)} \mathbf{P})^\top (\overline{\mathbf{W}}_Q^{(\ell)} \mathbf{P})) - \frac{1}{m^{(\ell)}} \mathbf{I}^{(\ell)})_{u,v}| \leq \delta, \quad u, v \in \mathcal{V}^{(L)},$$

where $\mathbf{I}^{(\ell)} \in \mathbb{R}^{d \times d}$ is a matrix such that $\mathbf{I}_{u,v}^{(\ell)} = 1$ if $\iota(\text{pa}^{(\ell')}(u)) = \iota(\text{pa}^{(\ell')}(v))$ ($\ell' \neq L - \ell$), and 0 otherwise.

By using this lemma, Lemma A.10 is shown as follows. Use $\overline{\mathbf{W}}_K^{(\ell)}$ and $\overline{\mathbf{W}}_Q^{(\ell)}$ from Lemma A.17 to construct

$$\mathbf{W}_K^{(\ell)} = \begin{bmatrix} \mathbf{0} & \overline{\mathbf{W}}_K^{(\ell)} \end{bmatrix}, \quad \mathbf{W}_Q^{(\ell)} = \begin{bmatrix} \mathbf{0} & \overline{\mathbf{W}}_Q^{(\ell)} \end{bmatrix},$$

where $\mathbf{0} \in \mathbb{R}^{(d_f + d_p) \times d_f}$.

Then, let $\mathbf{W}_V^{(\ell)}$ be

$$(\mathbf{W}_V^{(\ell)})_{i,j} = \begin{cases} m^{(\ell)} & i = j, (2\ell - 1)S + 2 \leq i \leq 2\ell S + 1 \\ 0 & \text{otherwise,} \end{cases}$$

so that $\mathbf{W}_V^{(\ell)}$ extracts $\mathbf{q}_v^{(\ell)}$.

Then, the v -th column of $(\mathbf{W}_V^{(\ell)} Q^{(\ell)}) \text{softmax}((\mathbf{W}_K^{(\ell)} Q^{(\ell)})^\top (\mathbf{W}_Q^{(\ell)} Q^{(\ell)}))$ implements the average of $\mathbf{q}_u^{(\ell)}$ over u satisfying that $\mathbf{I}_{u,v}^{(\ell)} = 1$ defined in Lemma A.17 multiplied by $m^{(\ell)}$, within the error of $m^{(\ell)}\delta$. The number of such u is exactly $m^{(\ell)}$, thus the average multiplied by $m^{(\ell)}$ is the summation. Now the error is $m^{(\ell)}\delta$, so letting $\delta \leftarrow (m^{(\ell)})^{-1}\delta$ yields the assertion.

Proof of Lemma A.17. Define the key and the query matrix $\overline{\mathbf{W}}_K^{(\ell)} = \mathbf{I}_{d_p} \in \mathbb{R}^{d_p \times d_p}$ and $\overline{\mathbf{W}}_Q^{(\ell)} \in \mathbb{R}^{d_p \times d_p}$ as

$$(\overline{\mathbf{W}}_Q^{(L-1)})_{i,j} = \begin{cases} \alpha & \text{if } i = j \text{ and } i \neq 2(L-\ell)+1, 2(L-\ell)+2, \\ 0 & \text{otherwise,} \end{cases}$$

for some $\alpha > 0$ which will be defined later. (From now, we will focus on the case when $L \geq 2$. When $L = 1$, it is obvious to see that the assertion still holds because $((\overline{\mathbf{W}}_K^{(L)} \mathbf{P})^\top (\overline{\mathbf{W}}_Q^{(L)} \mathbf{P}))_{u,v} = 0$ for all u, v .)

Then, we have

$$\begin{aligned} & ((\overline{\mathbf{W}}_K^{(\ell)} \mathbf{P})^\top (\overline{\mathbf{W}}_Q^{(\ell)} \mathbf{P}))_{u,v} \\ &= \alpha \sum_{\ell' \neq \ell} \left[\sin\left(\frac{2\pi\iota(\text{pa}^{(L-\ell')}(u))}{m^{(\ell')}}\right) \sin\left(\frac{2\pi\iota(\text{pa}^{(L-\ell')}(v))}{m^{(\ell')}}\right) + \cos\left(\frac{2\pi\iota(\text{pa}^{(L-\ell')}(u))}{m^{(\ell')}}\right) \cos\left(\frac{2\pi\iota(\text{pa}^{(L-\ell')}(v))}{m^{(\ell')}}\right) \right] \\ &= \alpha \sum_{\ell' \neq \ell} \cos\left(\frac{2\pi\iota(\text{pa}^{(L-\ell')}(v)) - \iota(\text{pa}^{(L-\ell')}(u))}{m^{(\ell')}}\right) \\ &= \begin{cases} (L-1)\alpha & \text{(if } \iota(\text{pa}^{(L-\ell')}(u)) = \iota(\text{pa}^{(L-\ell')}(v)) \text{ (} \ell' \neq \ell \text{))} \\ (L-1)\alpha - \alpha \min_{\ell' \neq \ell} \left(1 - \cos\left(\frac{2\pi}{m^{(\ell')}}\right)\right) & \text{(otherwise).} \end{cases} \end{aligned}$$

Let us recall the property of softmax. For $a \in \mathbb{R}^d$ with $a_1 = \dots = a_m > a_{m+1} \geq \dots \geq a_d$ with $a_m - a_{m+1} = A > 0$, it holds that $\text{softmax}(a)_1 \geq \frac{1}{m^{(\ell)}} \cdot \frac{1}{1+de^{-A}} \geq 1 - de^{-A}$ and $\text{softmax}(a)_i \leq e^{-A}$ ($i = 2, \dots, d$). Therefore, for $\delta < 1$, by taking $\alpha = \log \frac{d}{\delta(1-\cos(2\pi d^{-1}))}$, we have

$$|(\text{softmax}((\mathbf{q}^{(\ell)} \overline{\mathbf{W}}_K^{(\ell)})^\top (\mathbf{q}^{(\ell)} \overline{\mathbf{W}}_Q^{(\ell)})) - \frac{1}{m^{(\ell)}} \mathbf{I}^{(\ell)})_{u,v}| \leq \delta.$$

□

A.5.5 Evaluation of error propagation

To control the approximation error on the optimal similarity score function, we need to convert an approximation error of each component $f_{\text{im},\iota}^{(\ell)}, f_{\text{tx},\iota}^{(\ell)}$ by evaluating how component-wise approximation error propagates in the pipeline. The proof of this lemma requires Lipschitzness of the basic operations (Lemmas A.45, A.49 and A.50 in Appendix A.8.2), to ensure that the propagated errors do not explode. We use $\tau^w = f^{(0)}$ so there is no error when considering the link function.

Lemma A.18 (Evaluation of error propagation). *Assume we have functions $\mathbf{f}_{\text{tx},\iota}^{(\ell)}$ ($1 \leq \ell \leq L, \iota \in [m_{\text{tx}}^{(\ell)}]$) such that*

$$\begin{aligned} \|f_{\text{tx},\iota}^{(L)}(x) - \mathbf{f}_{\text{tx},\iota}^{(L)}(x)\|_{\infty} &\leq \delta, \quad \forall x \in [S], \\ \|f_{\text{tx},\iota}^{(\ell)}(h) - \mathbf{f}_{\text{tx},\iota}^{(\ell)}(h)\|_{\infty} &\leq \delta, \quad \forall h \in \mathbb{R}^S \text{ such that } \max_{s \in S} h_s = 0, \ell \in [L-1], \end{aligned} \quad (\text{A.72})$$

and $f_{\text{im},\iota}^{(\ell)}$ in the same way. Also, assume that $\|\delta_v^{(\ell)}\|_{\infty} \leq \delta$ holds for all $\ell = L-1, \dots, 0$ and $v \in \mathcal{V}^{(\ell)}$. Let us take $\tau^w = f^{(0)}$.

Consider the update in (A.58) and (A.59). Then, we have the following bound on the error propagation:

$$\max_{v \in \mathcal{V}_{\text{tx}}^{(L)}} \|h_{\text{tx},v}^{(\ell)} - h_{\text{tx},\text{pa}^{(L-\ell)}(v)}^{(\ell)}\|_{\infty} \leq \delta \times (2m_{\text{tx}}^{(\ell+1)} + 2) \prod_{\ell+2 \leq k \leq L} (2m_{\text{tx}}^{(k)} + 3), \quad \ell = L-1, \dots, 0, \quad (\text{A.73})$$

$$\max_{v \in \mathcal{V}_{\text{tx}}^{(L)}} \|q_{\text{tx},v}^{(\ell)} - q_{\text{tx},\text{pa}^{(L-\ell)}(v)}^{(\ell)}\|_{\infty} \leq \delta \times \prod_{\ell+1 \leq k \leq L} (2m_{\text{tx}}^{(k)} + 3), \quad \ell = L, \dots, 1, \quad (\text{A.74})$$

and the bounds on the image part follows in the same way. Furthermore, we have

$$|\mathbf{S}_{\text{NN}} - \mathbf{S}_{\text{MP}}| \leq \delta \times [\prod_{1 \leq \ell \leq L} (2m_{\text{im}}^{(\ell)} + 3) + \prod_{1 \leq \ell \leq L} (2m_{\text{tx}}^{(\ell)} + 3)]. \quad (\text{A.75})$$

Proof. First, we prove (A.73) and (A.74). We focus on the language model and the bounds on the vision model follows in the same way.

We use the induction. Let us check (A.73) for $\ell = L-1$ and (A.74) for $\ell = L$. Because $\mathbf{h}_{\text{tx},v}^{(L)} = h_{\text{tx},v}^{(L)}$, (A.72) implies that

$$\|q_{\text{tx},v}^{(L)} - \mathbf{q}_{\text{tx},v}^{(L)}\|_{\infty} \leq \delta.$$

By Lemma A.45 and $\|\delta_{\text{tx},v}^{(L-1)}\|_{\infty} \leq \delta$, we have

$$\|h_{\text{tx},\text{pa}^{(L)}(v)}^{(L-1)} - h_{\text{tx},v}^{(L-1)}\|_{\infty} \leq 2m_{\text{tx}}^{(L)} \max_{u \in \mathcal{V}_{\text{tx}}^{(L)}} \|f_{\text{tx},\iota(u)}^{(L)}(x_{\text{tx},u}) - \mathbf{f}_{\text{tx},\iota(u)}^{(L)}(x_{\text{tx},u})\|_{\infty} + 2\delta \leq 2(m_{\text{tx}}^{(L)} + 1)\delta,$$

for all $v \in \mathcal{V}_{\text{tx}}^{(L)}$, which confirms (A.73) for $\ell = L-1$ and (A.74) for $\ell = L$.

Assume (A.73) for $\ell = L, \dots, \ell$ and (A.74) for $\ell = L, \dots, \ell+1$ and prove (A.73) for ℓ and (A.74) for $\ell-1$.

$$\begin{aligned} &\|q_{\text{tx},\text{pa}^{(L-\ell)}(v)}^{(\ell)} - \mathbf{q}_{\text{tx},v}^{(\ell)}\|_{\infty} \\ &= \max_{u \in \mathcal{V}_{\text{tx}}^{(L)}} \|f_{\text{tx},\text{pa}^{(L-\ell)}(u)}^{(\ell)}(h_{\text{tx},\text{pa}^{(L-\ell)}(u)}^{(\ell)}) - \mathbf{f}_{\text{tx},\text{pa}^{(L-\ell)}(u)}^{(\ell)}(h_{\text{tx},u}^{(\ell)})\|_{\infty} \\ &\leq \max_{u \in \mathcal{V}_{\text{tx}}^{(L)}} \|f_{\text{tx},\text{pa}^{(L-\ell)}(u)}^{(\ell)}(h_{\text{tx},u}^{(\ell)}) - \mathbf{f}_{\text{tx},\text{pa}^{(L-\ell)}(u)}^{(\ell)}(h_{\text{tx},u}^{(\ell)})\|_{\infty} \end{aligned}$$

$$\begin{aligned}
& + \|f_{\text{tx},\text{pa}^{(L-\ell)}(u)}^{(\ell)}(h_{\text{tx},\text{pa}^{(L-\ell)}(u)}^{(\ell)}) - f_{\text{tx},\text{pa}^{(L-\ell)}(u)}^{(\ell)}(\mathbf{h}_{\text{tx},u}^{(\ell)})\|_\infty \\
& \leq \delta + \max_{u \in \mathcal{V}_{\text{tx}}^{(L)}} \|h_{\text{tx},\text{pa}^{(L-\ell)}(u)}^{(\ell)} - \mathbf{h}_{\text{tx},u}^{(\ell)}\|_\infty \\
& \leq \delta + \delta \times (2m_{\text{tx}}^{(\ell+1)} + 2) \prod_{k=\ell+2}^L (2m_{\text{tx}}^{(k)} + 3) \\
& \leq \delta \times \prod_{k=\ell+1}^L (2m_{\text{tx}}^{(k)} + 3),
\end{aligned} \tag{A.76}$$

where we used Lemma A.49 for the second inequality. Also,

$$\begin{aligned}
\|h_{\text{tx},\text{pa}^{(L-\ell+1)}(v)}^{(\ell-1)} - \mathbf{h}_{\text{tx},v}^{(\ell-1)}\|_\infty & \leq 2m_{\text{tx}}^{(\ell)} \max_{u \in \mathcal{V}^{(L)}} \|q_{\text{tx},\text{pa}^{(L-\ell)}(u)}^{(\ell)} - \mathbf{q}_{\text{tx},u}^{(\ell)}\|_\infty + 2\delta \\
& \leq \delta \times 2m_{\text{tx}}^{(\ell)} \prod_{k=\ell+1}^L (2m_{\text{tx}}^{(k)} + 3) + 2\delta \\
& \leq \delta \times (2m_{\text{tx}}^{(\ell)} + 2) \prod_{k=\ell+1}^L (2m_{\text{tx}}^{(k)} + 3),
\end{aligned}$$

where we used Lemma A.45 and $\|\delta_{\text{tx},v}^{(\ell-1)}\|_\infty \leq \delta$ for the first inequality, and (A.76) for the second inequality. Therefore, by induction, we obtained (A.73) for all $\ell = L-1, \dots, 0$ and (A.74) for all $\ell = L, \dots, 1$.

Finally, we bound $|S_{\text{NN}} - S_{\text{MP}}|$ to prove (A.75). By using Lemma A.50, and the bound $\|h_{\text{tx},r}^{(0)} - \mathbf{h}_{\text{tx},r}^{(0)}\|_\infty \leq \prod_{1 \leq \ell \leq L} (2m_{\text{tx}}^{(\ell)} + 3)$ and $\|h_{\text{im},r}^{(0)} - \mathbf{h}_{\text{im},r}^{(0)}\|_\infty \leq \prod_{1 \leq \ell \leq L} (2m_{\text{im}}^{(\ell)} + 3)$, we have that

$$\begin{aligned}
|S_{\text{NN}} - S_{\text{MP}}| & = |f^{(0)}(\text{softmax}(h_{\text{tx},r}^{(0)}), \text{softmax}(h_{\text{im},r}^{(0)})) - f^{(0)}(\text{softmax}(\mathbf{h}_{\text{tx},r}^{(0)}), \text{softmax}(\mathbf{h}_{\text{im},r}^{(0)}))| \\
& \leq \|h_{\text{tx},r}^{(0)} - \mathbf{h}_{\text{tx},r}^{(0)}\|_\infty + \|h_{\text{im},r}^{(0)} - \mathbf{h}_{\text{im},r}^{(0)}\|_\infty \\
& \leq \delta \times [\prod_{1 \leq \ell \leq L} (2m_{\text{im}}^{(\ell)} + 3) + \prod_{1 \leq \ell \leq L} (2m_{\text{tx}}^{(\ell)} + 3)]
\end{aligned}$$

□

A.5.6 Properties of the message passing algorithm

As auxiliary lemmas, we state boundedness of $h_v^{(\ell)}$, $q_v^{(\ell)}$, $\mathbb{P}[s|\mathbf{x}]$, and $S_\star$. Lemma A.19 omits the subscripts “tx” and “im” in this subsection because both text and image parts have similar bounds.

Lemma A.19. *Consider the message passing algorithm in (A.51) and (A.52). Under Assumption 2.5, we have that*

$$\|f_\iota^{(\ell)}(h)\|_\infty \leq \log SB_\psi \quad (2 \leq \ell \leq L), \quad \|f_\iota^{(1)}(h)\|_\infty \leq \log SB_\psi^{1+\frac{1}{m^{(1)}}},$$

for $h \in \mathbb{R}^S$ with $\max_s h_s = 0$, and that

$$\|h_v^{(\ell)}\|_\infty \leq 2m^{(\ell+1)} \log SB_\psi \quad (1 \leq \ell \leq L-1), \quad \|h_v^{(0)}\|_\infty \leq 2m^{(1)} \log SB_\psi^{1+\frac{1}{m^{(1)}}},$$

for the variables $h_v^{(\ell)}$ of the message passing algorithm. Furthermore, the conditional probability $\mathbb{P}[s|\mathbf{x}]$ is bounded as

$$\frac{1}{B_\psi^{2m}S} \leq \mathbb{P}[s|\mathbf{x}] \leq 1.$$

Proof. Because of the update (A.51), one dimension of $h_v^{(\ell)}$ is zero, and the others are zero or negative. Therefore, for $2 \leq \ell \leq L$, we have that

$$(q_v^{(\ell)})_s = (f_t^{(\ell)}(h))_s = \log \sum_{a \in [S]} \psi_t^{(\ell)}(s, a) e^{h_a} \leq \log SB_\psi,$$

and $(q_v^{(\ell)})_s \geq -\log SB_\psi$ holds in the same way. Also, for $\ell = 1$, we have

$$(q_r^{(1)})_s = (f_t^{(1)}(h))_s = \log \sum_{a \in [S]} \mathbb{P}[s]^{\frac{1}{m(1)}} \psi_t^{(1)}(s, a) e^{h_a} \leq \log SB_\psi^{1+\frac{1}{m(1)}},$$

and $(q_r^{(0)})_s \geq -\log SB_\psi^{1+\frac{1}{m(1)}}$ holds in the same way.

Also, by using the above bounds on $q_u^{(\ell+1)} = f_{t(v)}^{(\ell+1)}(q_u)$, we have

$$\|h_v^{(\ell)}\|_\infty = 2 \left\| \sum_{u \in \mathcal{C}(v)} q_u^{(\ell+1)} \right\|_\infty \leq 2|\mathcal{C}(v)| \max_{u \in \mathcal{C}(v)} \|q_u^{(\ell+1)}\|_\infty \leq \begin{cases} 2m^{(\ell+1)} \log SB_\psi & (\ell \geq 1), \\ 2m^{(1)} \log SB_\psi^{1+\frac{1}{m(1)}} & (\ell = 0). \end{cases}$$

Here applying normalize only changes the bound by a factor of at most two.

Finally, we consider the lower bound on $\mathbb{P}[s|\mathbf{x}]$. By Assumption 2.5, we see that $\min_{s \in [S]} \mathbb{P}[s] \geq 1/(SB_\psi)$ and $\mathbb{P}[\mathbf{x}|s]/\mathbb{P}[\mathbf{x}|s'] \in [B_\psi^{-2m}, B_\psi^{2m}]$ for any $s, s' \in [S]$. As a consequence,

$$\mathbb{P}[s|\mathbf{x}] = \frac{\mathbb{P}[\mathbf{x}|s]\mathbb{P}[s]}{\sum_{s' \in [S]} \mathbb{P}[\mathbf{x}|s']\mathbb{P}[s']} \geq \mathbb{P}[s] \cdot \min_{s' \in [S]} \frac{\mathbb{P}[\mathbf{x}|s]}{\mathbb{P}[\mathbf{x}|s']} \geq \frac{1}{B_\psi^{2m}S}.$$

□

Lemma A.20. Under Assumption 2.5, the optimal similarity score function (adjusted up to constant shift) $S_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \log \frac{\mathbb{P}[\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}]}{\mathbb{P}[\mathbf{x}_{\text{im}}]\mathbb{P}[\mathbf{x}_{\text{tx}}]}$ is upper and lower bounded as

$$-2m \log B_\psi \leq S_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \leq 2m \log B_\psi.$$

Proof. Note that

$$\exp(S_\star(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) = \frac{\mathbb{P}[\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}]}{\mathbb{P}[\mathbf{x}_{\text{im}}]\mathbb{P}[\mathbf{x}_{\text{tx}}]} = \frac{\mathbb{P}[\mathbf{x}_{\text{im}}|\mathbf{x}_{\text{tx}}]}{\mathbb{P}[\mathbf{x}_{\text{im}}]} \stackrel{(i)}{=} \frac{\sum_s \mathbb{P}[\mathbf{x}_{\text{im}}|s]\mathbb{P}[s|\mathbf{x}_{\text{tx}}]}{\sum_s \mathbb{P}[\mathbf{x}_{\text{im}}|s]\mathbb{P}[s]},$$

where step (i) uses the conditional independence of $\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}$ given $r = s$. Since

$$\frac{\mathbb{P}[s|\mathbf{x}_{\text{tx}}]}{\mathbb{P}[s]} = \frac{\mathbb{P}[\mathbf{x}_{\text{tx}}|s]}{\sum_{s' \in [S]} \mathbb{P}[\mathbf{x}_{\text{tx}}|s']\mathbb{P}[s']} \in \left[\min_{s, s' \in [S]} \frac{\mathbb{P}[\mathbf{x}_{\text{tx}}|s]}{\mathbb{P}[\mathbf{x}_{\text{tx}}|s']}, \max_{s, s' \in [S]} \frac{\mathbb{P}[\mathbf{x}_{\text{tx}}|s]}{\mathbb{P}[\mathbf{x}_{\text{tx}}|s']} \right]$$

by Bayes' formula, it follows from Assumption 2.5 that $\mathbb{P}[\mathbf{x}_{\text{tx}}|s]/\mathbb{P}[\mathbf{x}_{\text{tx}}|s'] \in [B_\psi^{-2m}, B_\psi^{2m}]$. Putting pieces together yields the desired result.

□

A.6 Proof of Theorem 2.12

A.6.1 Overview

To predict $\mathbf{x}_{\text{im}}$ given $\mathbf{x}_{\text{tx}}$ and $\mathbf{z}_t$, the message passing algorithm is the algorithm for computing the Bayes optimal denoiser. We describe the algorithm in Appendix A.6.1, and discuss how to implement the message passing algorithm using transformers in Appendix A.6.1. This section mainly focuses on the image part and sometimes we omit the subscript “im” from, e.g., $m_{\text{im}}^{(\ell)}$ and d_{im} .

Belief propagation and message passing algorithms

The message passing algorithm for the conditional denoising problem consists of the text part and the image part. The text part is the same as the procedure (A.51) and (A.52) in contrastive learning, which computes $h_{\text{tx},r}^{(0)} = (\log \mathbb{P}[s|\mathbf{x}_{\text{tx}}])_{s \in [S]} \in \mathbb{R}^S$ from $h_{\text{tx},v}^{(L)} = x_{\text{tx},v}$ ($v \in \mathcal{V}_{\text{tx}}^{(L)}$). The image part is divided into two processes: downsampling and upsampling. The image part first conduct the downsampling process to compute $h_r^{(0)}$ from $\mathbf{z}$. Then, combining this $h_r^{(0)}$ with the output of the text part $h_{\text{tx},r}^{(0)}$, the upsampling process of the image computes $b_v^{(L)}$ for each node $v \in \mathcal{V}_{\text{im}}^{(L)}$, so that $\text{softmax}(b_v^{(L)})$ is exactly equal to $(\mathbb{P}[x_{\text{im},v} = s|\mathbf{z}_t, \mathbf{x}_{\text{tx}}])_{s \in [S]}$. Intuitively, the downsampling process aggregates the information from the leaves to the root, while the upsampling constructs estimation of leaf nodes from the root to the leaves. Outputting the weighted average of s with respect to $\text{softmax}(b_{\uparrow,v}^{(L)})$ yields the Bayes optimal denoiser $(\mathbf{m}_{\star,t}(\mathbf{z}_t, \mathbf{x}_{\text{tx}}))_v$ of $\mathbf{x}_{\text{im}}$. We formally define the procedures for the image part in the following.

Downsampling. The downsampling process of the image part aggregates information from the leaves to the root of the tree. It starts with $h_v^{(L)} = \text{normalize}((-t(s - z_{t,v}/t)^2/2)_{s \in [S]}) \in \mathbb{R}^S$ ($v \in \mathcal{V}_{\text{im}}^{(L)}$), and computes $(q_v^{(\ell)})_{v \in \mathcal{V}_{\text{im}}^{(\ell)}}$ and $(h_v^{(\ell)})_{v \in \mathcal{V}_{\text{im}}^{(\ell)}}$ in the decreasing order of ℓ .

$$\begin{aligned} q_v^{(\ell)} &= f_{\downarrow,\ell}^{(\ell)}(h_v^{(\ell)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{im}}^{(\ell)}, \ell = L, \dots, 1, \\ h_v^{(\ell-1)} &= \text{normalize}(\sum_{u \in \mathcal{C}(v)} q_u^{(\ell)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{im}}^{(\ell-1)}, \ell = L, \dots, 1. \end{aligned}$$

Computation of $q_v^{(\ell)}$ and $h_v^{(\ell-1)}$ from $h_v^{(\ell)}$ is called the ℓ -th step of the downsampling process (of the image part). Here, $f_{\downarrow,\ell}^{(\ell)}$ are defined as

$$(f_{\downarrow,\ell}^{(\ell)}(h))_s = \log \sum_{a \in [S]} \psi_{\text{im},\ell}^{(\ell)}(s, a) e^{h_a}, \quad h \in \mathbb{R}^S, \quad s \in [S], \quad \ell = L, \dots, 1,$$

This is also the same as (A.51) and (A.52) in contrastive learning, except that $\mathbb{P}[s]^{\frac{1}{m^{(1)}}}$ is not needed for $\ell = 1$.

Upsampling. It starts with combining the information of the text part and image down-sampling. Then, the algorithm computes the prediction of $\mathbf{x}_{\text{im}}$ from the root to the leaves. The update is written as

$$\begin{aligned}\bar{b}_r^{(0)} &= h_r^{(0)} + h_{\text{tx},r}^{(0)} \in \mathbb{R}^S, \\ \bar{b}_v^{(\ell)} &= f_{\uparrow,\ell}^{(\ell)}(\text{normalize}(\bar{b}_{\text{pa}(v)}^{(\ell-1)} - q_v^{(\ell)})) + h_v^{(\ell)} \in \mathbb{R}^S, \quad v \in \mathcal{V}_{\text{im}}^{(\ell)}, \ell = 1, \dots, L, \\ (\mathbf{m}_{\star,t}(\mathbf{z}_t, \mathbf{x}_{\text{tx}}))_v &= \sum_{s \in [S]} s \cdot \text{softmax}(\bar{b}_v^{(L)})_s, \quad v \in \mathcal{V}_{\text{im}}^{(L)}, s \in [S]\end{aligned}\tag{A.77}$$

where

$$(f_{\uparrow,\ell}^{(\ell)}(h))_s = \log \sum_{a \in [S]} \psi_{\text{im},\ell}^{(\ell)}(a, s) e^{h_a}, \quad h \in \mathbb{R}^S, s \in [S], \ell = 1, \dots, L.\tag{A.78}$$

The update in (A.77) is equivalently written as (note that $\bar{b}_{\text{pa}(v)}^{(\ell-1)} - q_{\text{im},v}^{(\ell)} = b_v^{(\ell)}$ holds)

$$\begin{aligned}b_v^{(1)} &= \text{normalize}(h_r^{(0)} + h_{\text{tx},r}^{(0)} - q_{\text{im},v}^{(1)}) \in \mathbb{R}^S, \quad v \in \mathcal{V}_{\text{im}}^{(1)}, \\ b_v^{(\ell+1)} &= \text{normalize}(f_{\text{im},\uparrow,\ell}^{(\ell)}(b_{\text{pa}(v)}^{(\ell)} + h_{\text{pa}(v)}^{(\ell)} - q_v^{(\ell+1)})) \in \mathbb{R}^S, \quad v \in \mathcal{V}_{\text{im}}^{(\ell+1)}, \ell = 1, \dots, L-1, \\ b_v^{(L+1)} &= \text{normalize}(f_{\uparrow,\ell}^{(L)}(b_v^{(L)} + h_v^{(L)})) \in \mathbb{R}^S, \quad v \in \mathcal{V}_{\text{im}}^{(L)}, \\ (\mathbf{m}_{\star,t}(\mathbf{z}_t, \mathbf{x}_{\text{tx}}))_v &= \sum_{s \in [S]} s \cdot \text{softmax}(b_v^{(L+1)})_s, \quad v \in \mathcal{V}_{\text{im}}^{(L)}, s \in [S]\end{aligned}\tag{A.79}$$

Computation of $b_v^{(\ell)}$ is called the ℓ -th step of the upsampling process (of the image part). We will approximate (A.79) instead of (A.77), because we want to avoid complication about `normalize`. Specifically, while our transformer block consists of the feed forward, self-attention, and “normalize”, applying subtraction $(\bar{b}_{\text{pa}(v)}^{(\ell-1)} - q_v^{(\ell)})$, “normalize”, and nonlinear transformation $f_{\downarrow,\ell}^{(\ell)}$ cannot be done in one block.

The correctness of the message passing algorithm is formally stated as follows. Because of this, taking the weighted average of s with respect to $\text{softmax}(\bar{b}_{\text{im},v}^{(L)})$ yields the Bayes optimal prediction of $\mathbf{x}_{\text{im}}$.

Lemma A.21 (MP is the optimal denoising algorithm). *When applying the message passing algorithm introduced in (A.77) and (A.78), it holds that $\text{softmax}(b_v^{(L)})_s = \mathbb{P}[x_{\text{im},v} = s | \mathbf{z}_t, \mathbf{x}_{\text{tx}}]$ for all $v \in \mathcal{V}_{\text{im}}^{(L)}$.*

Proof. Regarding the joint generative hierarchical model as a single tree, the message passing algorithm for this case is directly adopted from (MP-DNS) of Mei (2024). $\square$

Approximation with transformer networks

We approximate the message passing algorithm with transformer networks. We denote a transformer approximation of $h_{\text{tx},r}^{(0)} = (\log \mathbb{P}[s | \mathbf{x}_{\text{tx}}])_{s \in [S]}$ by $\mathbf{h}_{\text{tx},d_{\text{tx}}}^{(0)} \in \mathbb{R}^S$. This can be obtained

by $\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}$ constructed in contrastive learning, or a transformation of $\widehat{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}})$ in the two-stage training. Specifically, we let

$$\mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} = \log(\text{trun}_{\text{tx}}(\widehat{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}))), \quad \text{where } \text{trun}_{\text{tx}}(z) := \text{proj}_{[\exp(-B_{\text{read}}^{\text{tx}}), \exp(B_{\text{read}}^{\text{tx}})]}(z) \quad (\text{A.80})$$

and $B_{\text{read}}^{\text{tx}} := 4m \log(B_\psi) + \log S$. We let

$$\delta_{\text{tx}} := \|\mathbf{h}_{\text{tx}, r}^{(0)} - \mathbf{h}_{\text{tx}, d}^{(0)}\|_\infty \quad (\text{A.81})$$

denote the approximation error of $\mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)}$. We will see how the final approximation error depends on δ_{tx} in later sections. We will use the numbering of nodes defined in Definition A.1.

Let $\mathbf{h}_v^{(L)} = h_v^{(L)} = \text{normalize}((-t(x - z_{t,v}/t)^2/2)_{x \in [S]}) \in \mathbb{R}^S$ for all $v \in \mathcal{V}_{\text{im}}^{(L)}$. After the positional encoding Emb_{cdm} , we obtain the initial matrix $\mathbf{H}^{(L)}$ such that

$$\mathbf{H}^{(L)} = \text{Emb}_{\text{cdm}}(\mathbf{z}, \widehat{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}})) = \begin{bmatrix} \mathbf{0} & & & \\ \mathbf{h}_1^{(L)} & \mathbf{h}_2^{(L)} & \cdots & \mathbf{h}_d^{(L)} \\ \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} & \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} & \cdots & \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} \\ \mathbf{P} & & & \end{bmatrix} \in \mathbb{R}^{(d_f + d_p) \times d},$$

where $\mathbf{P} \in \mathbb{R}^{d_p \times d}$ is a matrix that encodes the positions of the nodes, and the output of the text model $\mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)}$ is concatenated with every pixel. Here the dimensions are defined as $d_f = (3L + 3)S$ and $d_p = 2L$.

The text network has $(2L + 1)$ transformer blocks, and the structure of each block is the same as contrastive learning. The first L blocks approximate downsampling, and each block is called the $\ell (= L, \dots, 1)$ -th block of downsampling using the decreasing order. The latter $(L + 1)$ -blocks approximate upsampling, and each block is called the $\ell (= 0, \dots, L)$ -th block of upsampling using the increasing order.

First we consider downsampling. Starting from $\mathbf{H}^{(L)}$, we iteratively construct $\mathbf{H}^{(\ell)} \in \mathbb{R}^{(d_f + d_p) \times d}$ and $\mathbf{Q}^{(\ell)} \in \mathbb{R}^{(d_f + d_p) \times d}$:

$$\mathbf{H}^{(\ell)} = \begin{bmatrix} \mathbf{0} & & & \\ \mathbf{h}_1^{(\ell)} & \mathbf{h}_2^{(\ell)} & \cdots & \mathbf{h}_d^{(\ell)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_1^{(L)} & \mathbf{q}_2^{(L)} & \cdots & \mathbf{q}_d^{(L)} \\ \mathbf{h}_1^{(L)} & \mathbf{h}_2^{(L)} & \cdots & \mathbf{h}_d^{(L)} \\ \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} & \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} & \cdots & \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} \\ \mathbf{P} & & & \end{bmatrix}, \quad \mathbf{Q}^{(\ell)} = \begin{bmatrix} \mathbf{0} & & & \\ \mathbf{q}_1^{(\ell)} & \mathbf{q}_2^{(\ell)} & \cdots & \mathbf{q}_d^{(\ell)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_1^{(L)} & \mathbf{q}_2^{(L)} & \cdots & \mathbf{q}_d^{(L)} \\ \mathbf{h}_1^{(L)} & \mathbf{h}_2^{(L)} & \cdots & \mathbf{h}_d^{(L)} \\ \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} & \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} & \cdots & \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} \\ \mathbf{P} & & & \end{bmatrix}.$$

Here $\mathbf{h}_v^{(\ell)}$ ($\ell = L, L - 1, \dots, 0$) and $\mathbf{q}_v^{(\ell)}$ ($\ell = L, L - 1, \dots, 1$) are S -dimensional real-valued vectors. Except that their column dimension is different, $\mathbf{H}^{(\ell)}$ and $\mathbf{Q}^{(\ell)}$ are the same as Appendix A.5.1.

In the ℓ -th block of downsampling, the feed forward layer $\text{FF}_{\downarrow}^{(\ell)}$, a fully-connected ReLU network, receives $\mathbf{H}^{(\ell)}$ and outputs $Q^{(\ell)}$ by computing $\mathbf{q}_v^{(\ell)}$ from $\mathbf{h}_v^{(\ell)}$:

$$Q^{(\ell)} = \underbrace{\mathbf{H}^{(\ell)}}_{\text{skip connection}} + \text{FF}_{\downarrow}^{(\ell)}(\mathbf{H}^{(\ell)}) = \mathbf{H}^{(\ell)} + \begin{bmatrix} \mathbf{0} & (\in \mathbb{R}^{((2\ell+L)S) \times d}) \\ \mathbf{q}_1^{(\ell)} & \mathbf{q}_2^{(\ell)} & \dots & \mathbf{q}_d^{(\ell)} \\ \mathbf{0} & (\in \mathbb{R}^{(d_p + (2L-2\ell+2)S) \times d}) \end{bmatrix},$$

Then, the self-attention layer $\text{Attn}^{(\ell)}$ uses $Q^{(\ell)}$ to construct $\mathbf{H}^{(\ell-1)}$ as

$$\begin{aligned} \mathbf{H}^{(\ell-1)} &= \text{normalize} \left(\underbrace{Q^{(\ell)}}_{\text{skip connection}} + \text{Attn}^{(\ell)}(Q^{(\ell)}) \right) \\ &= \text{normalize} \left(Q^{(\ell)} + \begin{bmatrix} \mathbf{0} & (\in \mathbb{R}^{((2\ell+L-1)S) \times d}) \\ \star & (\in \mathbb{R}^{S \times d}) \\ \mathbf{0} & (\in \mathbb{R}^{(d_p + (2L-2\ell+3)S) \times d}) \end{bmatrix} \right). \end{aligned}$$

Here $\star$ means $[\mathbf{h}_1^{(\ell-1)} \ \mathbf{h}_2^{(\ell-1)} \ \dots \ \mathbf{h}_d^{(\ell-1)}]$ before normalization.

We then consider upsampling. After we obtain $\mathbf{H}^{(0)}$, we iteratively compute $\mathbf{B}^{(\ell)}$ ($\ell = 1, \dots, L+1$):

$$\mathbf{B}^{(\ell)} = \begin{bmatrix} & & \mathbf{0} & \\ \mathbf{b}_1^{(\ell)} & \mathbf{b}_2^{(\ell)} & \dots & \mathbf{b}_d^{(\ell)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{b}_1^{(1)} & \mathbf{b}_2^{(1)} & \dots & \mathbf{b}_d^{(1)} \\ \mathbf{h}_1^{(0)} & \mathbf{h}_2^{(0)} & \dots & \mathbf{h}_d^{(0)} \\ \mathbf{q}_1^{(1)} & \mathbf{q}_2^{(1)} & \dots & \mathbf{q}_d^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_1^{(L)} & \mathbf{q}_2^{(L)} & \dots & \mathbf{q}_d^{(L)} \\ \mathbf{h}_1^{(L)} & \mathbf{h}_2^{(L)} & \dots & \mathbf{h}_d^{(L)} \\ \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} & \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} & \dots & \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} \\ & & \mathbf{P} & \end{bmatrix}.$$

Here, $\mathbf{b}_v^{(\ell)}$ ($v \in \mathcal{V}_{\text{im}}^{(L)}, \ell = 1, \dots, L+1$) are S -dimensional real-valued vectors. The ℓ -th block of downsampling computes $\mathbf{B}^{(\ell+1)}$ using a feed forward network $\text{FF}_{\uparrow}^{(\ell)}$ with normalization:

$$\mathbf{B}^{(\ell+1)} = \text{normalize} \left(\underbrace{\mathbf{B}^{(\ell)}}_{\text{skip connection}} + \text{FF}_{\uparrow}^{(\ell)}(\mathbf{B}^{(\ell)}) \right) = \mathbf{B}^{(\ell)} + \begin{bmatrix} \mathbf{0} & (\in \mathbb{R}^{((L-\ell)S) \times d}) \\ \mathbf{b}_1^{(\ell+1)} & \mathbf{b}_2^{(\ell+1)} & \dots & \mathbf{b}_d^{(\ell+1)} \\ \mathbf{0} & (\in \mathbb{R}^{(d_p + (2L+\ell+2)S) \times d}) \end{bmatrix}. \quad (\text{A.82})$$

For $\ell = 0$, replace $\mathbf{B}^{(0)}$ by $\mathbf{H}^{(0)}$. For upsampling, we do not need the self-attention layer. It is simply ignored by just setting $\mathbf{W}_V = 0$ (and remember that we still have the skip connection).

Finally, we will obtain $\mathbf{b}_v^{(L+1)}$, that approximates $b_v^{(L+1)}$. In the readout layer read_{cdm} , we compute the prediction of $\mathbf{x}_{\text{im}}$ based on $\mathbf{b}_v^{(L+1)}$.

In the following, our goal is to iteratively show that, for all $v \in \mathcal{V}_{\text{im}}^{(L)}$,

$$\mathbf{h}_v^{(\ell)} \approx h_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}, \quad \mathbf{q}_v^{(\ell)} \approx q_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}, \quad \mathbf{b}_v^{(\ell)} \approx b_{\text{pa}^{(L-\ell)}(v)}^{(\ell)} \quad \text{for all } v \in \mathcal{V}_{\text{im}}^{(L)},$$

($\ell = L, \dots, 0$ for $\mathbf{h}_v^{(\ell)}$, $\ell = L, \dots, 1$ for $\mathbf{q}_v^{(\ell)}$, and $\ell = 1, \dots, L$ for $\mathbf{b}_v^{(\ell)}$), and

$$\mathbf{b}_v^{(L+1)} \approx b_v^{(L+1)}.$$

We will now formally define each component of the pipeline.

Encoding Emb_{cdm} . The positional encoding is the same as the one for contrastive learning (A.55). The v th column of $\mathbf{P}$, $\mathbf{p}_v$, is written as

$$\mathbf{p}_v = \begin{bmatrix} \sin\left(\frac{2\pi\iota(v)}{m^{(L)}}\right) & \cos\left(\frac{2\pi\iota(v)}{m^{(L)}}\right) & \sin\left(\frac{2\pi\iota(\text{pa}(v))}{m^{(L-1)}}\right) & \cos\left(\frac{2\pi\iota(\text{pa}(v))}{m^{(L-1)}}\right) \\ \dots & \sin\left(\frac{2\pi\iota(\text{pa}^{(L-1)}(v))}{m^{(1)}}\right) & \cos\left(\frac{2\pi\iota(\text{pa}^{(L-1)}(v))}{m^{(1)}}\right) \end{bmatrix}^\top. \quad (\text{A.83})$$

For two-stage training, where $\widehat{\mathbf{E}}_{\text{tx}}(x)$ approximates $\mathbf{E}_{\text{tx},*}(x) = \mathbb{P}[s|\mathbf{x}_{\text{tx}}]$, we define $\mathbf{h}_{\text{tx},d_{\text{tx}}}^{(0)}$ as $\mathbf{h}_{\text{tx},d_{\text{tx}}}^{(0)} = (\log \text{trun}_{\text{tx}}(\widehat{\mathbf{E}}_{\text{tx}}(x)))_{s \in [S]}$.

Downsampling: position-wise feed forward block. Similarly to the contrastive learning, the feed forward layer at the ℓ -th block yields

$$\mathbf{q}_v^{(\ell)} = \mathbf{h}_v^{(\ell)} + \mathbf{f}_{\downarrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{h}_v^{(\ell)}), \quad v \in \mathcal{V}_{\text{im}}^{(L)}.$$

Thus, when $\mathbf{h}_v^{(\ell)} \approx h_v^{(\ell)}$ for $v \in \mathcal{V}_{\text{im}}^{(\ell)}$ and $\mathbf{f}_\iota^{(\ell)} \approx f_\iota^{(\ell)}$, we have $\mathbf{q}_v^{(\ell)} \approx q_v^{(\ell)}$ for $v \in \mathcal{V}_{\text{im}}^{(\ell)}$.

Following the notation in Definition A.2, we state the following approximation error guarantee.

Lemma A.22 (Approximation error of feed forward layer, downsampling). *Fix $\ell \in [L]$ and $\delta > 0$. Assume that $B_\psi^{-1} \leq \psi_\iota^{(\ell)}(s, a) \leq B_\psi$ for all $s, a \in [S]$. Then, there exists an $\text{NN} \in \mathcal{F}(J, \mathbf{j}, B)$ such that*

$$\|\text{NN}([h; \mathbf{p}_v]) - f_{\downarrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(h)\|_\infty \leq \delta, \quad v \in \mathcal{V}_{\text{im}}^{(L)},$$

for all $h \in \mathbb{R}^S$ with $\max_s h_s = 0$. The network parameters $J, \mathbf{j}$ and B are bounded as follows:

$$\begin{aligned} J &\lesssim (\log \log(SB_\psi/\delta)) \log(SB_\psi/\delta), \quad \|\mathbf{j}\|_\infty \lesssim m^{(\ell)} S (\log(SB_\psi/\delta))^3 + L, \\ B &\lesssim 2S(B_\psi^2 + \log(SB_\psi/\delta)) + (m^{(\ell)})^2. \end{aligned}$$

This is the same as Lemma A.9, except that $(f_{\downarrow, \iota}^{(\ell)}(h))_s = \log \sum_{a \in [S]} \psi_{\text{im}, \iota}^{(\ell)}(s, a) e^{h_a}$ for $\ell = L$ and 1. It is easy to see that the proof of Lemma A.11 covers these cases, and thus we do not repeat the proof.

Downsampling: self-attention block. The self-attention layer $\text{Attn}^{(\ell)}$ of the ℓ -th block ($\ell = L, L-1, \dots, 1$) yields

$$\mathbf{h}_v^{(\ell-1)} = \text{normalize} \left(\sum_{\iota(\text{pa}^{(L-\ell')}(u)) = \iota(\text{pa}^{(L-\ell')}(v)) \ (\ell' \neq \ell)} \mathbf{q}_u^{(\ell)} + \boldsymbol{\delta}_v^{(\ell-1)} \right).$$

Here $\text{normalize}(x)_s = x_s - \max x_{s'}$. Please refer to Appendix A.5.1 for interpretation of the summation. We can see that, when $\mathbf{q}_v^{(\ell)} \approx q_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}$ and $\|\boldsymbol{\delta}_v^{(\ell-1)}\|_\infty \ll 1$ for $v \in \mathcal{V}_{\text{im}}^{(L)}$, we have $\mathbf{h}_v^{(\ell-1)} \approx h_{\text{pa}^{(L-\ell)}(v)}^{(\ell-1)}$ for $v \in \mathcal{V}_{\text{im}}^{(L)}$.

Following the notation in Definition A.3, we have the following approximation error guarantee.

Lemma A.23 (Approximation error of self-attention layer). *For $\ell \in [L]$, there exists $\text{Attn} \in \mathcal{A}(D, B)$ with $D = d_f + d_p$ and $B \lesssim \log(d\delta^{-1}) + m^{(\ell)}$ such that*

$$\text{Attn}(Q^{(\ell)}) = \left[\begin{array}{cc} \mathbf{0} & (\in \mathbb{R}^{(2\ell+L+1)S}) \\ \sum_{\iota(\text{pa}^{(L-\ell')}(u)) = \iota(\text{pa}^{(L-\ell')}(v)) \ (\ell' \neq \ell)} \mathbf{q}_u^{(\ell)} + \boldsymbol{\delta}_v^{(\ell-1)} & (\in \mathbb{R}^S) \\ \mathbf{0} & (\in \mathbb{R}^{d_p + (2L-2\ell+1)S}) \end{array} \right]_{v \in \mathcal{V}^{(L)}},$$

where $\boldsymbol{\delta}_v^{(\ell-1)} \in \mathbb{R}^S$ satisfies $\|\boldsymbol{\delta}_v^{(\ell-1)}\|_\infty \leq \delta \max_{v'} \|\mathbf{q}_{v'}^{(\ell)}\|_\infty$.

Because the only difference from Lemma A.10 is the dimension of zeros, we do not repeat the proof. See Appendix A.5.4 for the proof of Lemma A.10.

Upsampling: position-wise feed forward block. The upsampling is implemented with $(L+1)$ -transformer blocks indexed in increasing order $\ell = 0, \dots, L$, where the ℓ -th block of upsampling consists of a feed forward layer $\text{FF}_{\uparrow}^{(\ell)}$ with skip connection and normalization. The ℓ -th block of upsampling computes $\mathbf{b}_v^{(\ell+1)}$ from $\mathbf{h}_v^{(\ell)}$, $\mathbf{q}_v^{(\ell+1)}$, and $\mathbf{b}_v^{(\ell)}$.

$$\mathbf{b}_v^{(1)} = \text{normalize}(\mathbf{h}_v^{(0)} + \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} - \mathbf{q}_v^{(1)}) \in \mathbb{R}^S, \quad v \in \mathcal{V}_{\text{im}}^{(L)}, \quad (\text{A.84})$$

$$\mathbf{b}_v^{(\ell+1)} = \text{normalize}(\mathbf{f}_{\uparrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{b}_v^{(\ell)}) + \mathbf{h}_v^{(\ell)} - \mathbf{q}_v^{(\ell+1)}) \in \mathbb{R}^S, \quad v \in \mathcal{V}_{\text{im}}^{(L)}, \ \ell = 1, \dots, L-1, \quad (\text{A.85})$$

$$\mathbf{b}_v^{(L+1)} = \text{normalize}(\mathbf{f}_{\uparrow, \iota(v)}^{(L)}(\mathbf{b}_v^{(L)}) + \mathbf{h}_v^{(L)}) \in \mathbb{R}^S, \quad v \in \mathcal{V}_{\text{im}}^{(L)}. \quad (\text{A.86})$$

For each update, we can track the correspondence with the message passing algorithm. Specifically, for (A.84), when $\mathbf{h}_v^{(0)} \approx h_{\text{r}}^{(0)}$, $\mathbf{q}_v^{(1)} \approx q_{\text{pa}^{(L-1)}(v)}^{(1)}$, and $\mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} \approx h_{\text{tx}, \text{r}}^{(0)}$ for $v \in \mathcal{V}_{\text{im}}^{(L)}$, we have $\mathbf{b}_v^{(0)} \approx b_{\text{pa}^{(L-1)}(v)}^{(1)}$ for $v \in \mathcal{V}_{\text{im}}^{(L)}$. Similar discussion holds for (A.85) and (A.86) as well.

We will show the following approximation error guarantee of $f_{\uparrow, \iota}^{(\ell)}$. Since it is easy to concatenate zeros to the first and last layer matrices and adjust the input and output dimensions, below we present the network NN as a function between relevant dimensions for simple presentation.

Lemma A.24 (Approximation error of feed forward layer, upsampling). *Fix $\ell \in \{0, \dots, L\}$ and $\delta > 0$. Assume that $B_\psi^{-1} \leq \psi_i^{(\ell)}(s, a) \leq B_\psi$ for all $s, a \in [S]$. Then, there exist $\text{NN}_1, \text{NN}_2, \text{NN}_3 \in \mathcal{F}(J, \mathbf{j}, B)$ such that*

$$\begin{aligned} \text{NN}_1([h; h'; q]) &= h + h' - q, \\ \|\text{NN}_2([b; h; q; \mathbf{p}_v]) - (f_{\downarrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(b) + h - q)\|_\infty &\leq \delta, \quad v \in \mathcal{V}_{\text{im}}^{(L)}, \ell = 1, \dots, L-1 \\ \|\text{NN}_3([b; h; \mathbf{p}_v]) - (f_{\downarrow, \iota(v)}^{(L)}(b) + h)\|_\infty &\leq \delta, \quad v \in \mathcal{V}_{\text{im}}^{(L)}, \ell = L, \end{aligned}$$

for all $h, h', q, b \in \mathbb{R}^S$ with $\max_s b_s = 0$. For all of these networks, the parameters $J, \mathbf{j}$ and B are bounded as follows:

$$\begin{aligned} J &\lesssim (\log \log(SB_\psi/\delta)) \log(SB_\psi/\delta), \quad \|\mathbf{j}\|_\infty \lesssim m^{(\ell)} S (\log(SB_\psi/\delta))^3 + L, \\ B &\lesssim 2S(B_\psi^2 + \log(SB_\psi/\delta)) + (m^{(\ell)})^2. \end{aligned}$$

The first network NN_1 is just a linear mapping, represented as $[\mathbf{I}_S \ \mathbf{I}_S \ \mathbf{I}_S]$. The proof for NN_2 and NN_3 is mostly the same as Lemma A.9. The only difference is to add $h - q$ (or h) after computing $f_{\downarrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(b)$, which is easily done with one additional layer. Therefore, we omit the proof of this lemma. See Appendix A.5.3 for the proof of Lemma A.9.

Normalization. In the attention network, since column vectors of $\mathbf{H}^{(\ell)}$, $\mathbf{Q}^{(\ell)}$, and $\mathbf{B}^{(\ell)}$ are a collection of multiple $\mathbf{h}_v^{(\ell)}$, $\mathbf{q}_v^{(\ell)}$, and $\mathbf{b}_v^{(\ell)}$, we adopt a slightly different definition of “normalize” for these column vectors, from the one for S -dimensional vectors. Let $\mathbf{x} = [\mathbf{b}^{(L+1)} \ \dots \ \mathbf{b}^{(1)} \ \mathbf{h}^{(0)} \ \mathbf{q}^{(1)} \ \mathbf{h}^{(1)} \ \dots \ \mathbf{q}^{(L)} \ \mathbf{h}^{(L)} \ \mathbf{h} \ \mathbf{p}] \in \mathbb{R}^{d_f + d_p}$, where $\mathbf{h}, \mathbf{h}^{(\ell)} (\ell = L, \dots, 0), \mathbf{q}^{(\ell)} (\ell = L, \dots, 1), \mathbf{b}^{(\ell)} (\ell = 1, \dots, L+1)$ are S -dimensional real-valued vectors and $\mathbf{p} \in \mathbb{R}^{d_p}$. We define normalize as

$$\text{normalize}(\mathbf{x}) = \begin{bmatrix} \mathbf{h}^{(0)} \\ \mathbf{q}^{(1)} - \mathbf{1}_S \max_{s \in S} \mathbf{q}_s^{(1)} \\ \mathbf{h}^{(1)} \\ \mathbf{q}^{(2)} - \mathbf{1}_S \max_{s \in S} \mathbf{q}_s^{(2)} \\ \vdots \\ \mathbf{q}^{(L)} - \mathbf{1}_S \max_{s \in S} \mathbf{q}_s^{(L)} \\ \mathbf{h}^{(L)} \\ \mathbf{h} \\ \mathbf{p} \end{bmatrix} \in \mathbb{R}^{d_f + d_p}, \quad \mathbf{1}_S = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^S.$$

For a matrix with its column dimension $d_f + d_p$, it is applied in a column-wise manner.

Readout layer read_{cdm} . In the readout layer, we output the prediction $\mathbf{M}_t$ of $\mathbf{x}_{\text{im}}$ from $\mathbf{b}_v^{(L+1)}$ as

$$\mathbf{M}_{t,v} = \text{read}_{\text{cdm}}(\mathbf{B}^{(L+1)}) = \sum_{s \in [S]} s \cdot \text{softmax}(\mathbf{b}_v^{(L+1)})_s, \quad v \in \mathcal{V}_{\text{im}}^{(L)}. \quad (\text{A.87})$$

The whole pipeline. Putting it all together, the neural network approximate the message passing algorithm (for the image part) in the following way. The downsampling process is approximated as

$$\begin{aligned} \mathbf{q}_v^{(\ell)} &= \mathbf{f}_{\uparrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{h}_v^{(\ell)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{im}}^{(L)}, \ell = L, \dots, 1, \\ \mathbf{h}_v^{(\ell-1)} &= \text{normalize} \left(\sum_{\iota(\text{pa}^{(L-\ell')}(u)) = \iota(\text{pa}^{(L-\ell')}(v)) \ (\ell' \neq \ell)} \mathbf{q}_u^{(\ell)} + \delta_v^{(\ell-1)} \right) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{im}}^{(L)}, \ell = L, \dots, 1. \end{aligned} \quad (\text{A.88})$$

Let $\mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} \approx h_{\text{tx}, \text{r}}^{(0)}$. The upsampling process is approximated as

$$\begin{aligned} \mathbf{b}_{\text{im}, v}^{(1)} &= \text{normalize}(\mathbf{h}_v^{(0)} + \mathbf{h}_{\text{tx}, d_{\text{tx}}}^{(0)} - \mathbf{q}_v^{(1)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{im}}^{(L)}, \\ \mathbf{b}_v^{(\ell+1)} &= \text{normalize}(\mathbf{f}_{\uparrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{b}_v^{(\ell)}) + \mathbf{h}_v^{(\ell)} - \mathbf{q}_v^{(\ell+1)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{im}}^{(L)}, \ell = 1, \dots, L-1 \\ \mathbf{b}_v^{(L+1)} &= \text{normalize}(\mathbf{f}_{\uparrow, \iota(v)}^{(L)}(\mathbf{b}_v^{(L)}) + \mathbf{h}_v^{(L)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{im}}^{(L)}, \\ \mathbf{M}_{t,v} &= \sum_{s \in [S]} s \cdot \text{softmax}(\mathbf{b}_v^{(L+1)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{im}}^{(L)}. \end{aligned} \quad (\text{A.89})$$

For two step training, the hypothesis class to which a tuple $(\text{TF}_{\text{cdm}}, \text{Adap})$ belongs is defined as follows, formally restating (2.22). For joint training, the parameter space $\Theta_{L,J,D,D',B}^{\text{cdm}}$ is defined in Appendix A.6.3.

Definition A.5 (Eq. (2.22), restated). We say the collection of the parameters of $(\text{TF}_{\text{cdm}}, \text{Adap})$ belongs to $\Theta_{L,J,D,D',B,M}$ if the following holds: The image transformer network TF_{cdm} has L blocks of feed forward (Definition A.2), self-attention (Definition A.3), and normalization. In each block, its feed forward FF and self-attention Attn satisfy

$$\text{FF} \in \mathcal{F}(J, \mathbf{j} = (D, *, \dots, *, D), B), \text{ with } \|\mathbf{j}\|_{\infty} \leq D', \quad \text{Attn} \in \mathcal{A}(D, B).$$

Furthermore, the adapter satisfies

$$W_{\text{ada}}^{(1)} \in \mathbb{R}^{S \times M}, \quad W_{\text{ada}}^{(2)} \in \mathbb{R}^{M \times S}, \quad \|W_{\text{ada}}^{(1)}\|_{\text{op}} \leq B, \quad \|W_{\text{ada}}^{(2)}\|_{\text{op}} \leq B.$$

The rest of this section is organized as follows. Appendix A.6.2 proves Theorem 2.12, using Lemmas A.22 to A.24, as well as the bound on the propagation of the intermediate errors Lemma A.26. Appendix A.6.4 proves the error propagation lemma (Lemma A.26).

A.6.2 Proof of Theorem 2.12

Define

$$\bar{\mathbf{R}}_{\text{cdm}, t}^* := \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, \mathbf{z}_t)} \left[\|\mathbf{x}_{\text{im}} - \mathbf{m}_{*, t}(\mathbf{z}_t, \mathbf{x}_{\text{tx}})\|_2^2 \right],$$

where $\mathbf{m}_{\star,t}(\mathbf{z}_t, \mathbf{x}_{\text{tx}}) = \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, \mathbf{z}_t) \sim \mu_{\star,t}}[\mathbf{x}_{\text{im}} | \mathbf{z}_t, \mathbf{x}_{\text{tx}}]$. Similar to the proof of Theorem 2.10, we have the following decomposition:

$$\begin{aligned} & \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, \mathbf{z}_t)} \left[\left\| \mathbf{m}_{\star,t}(\mathbf{z}_t, \mathbf{x}_{\text{tx}}) - \mathbf{M}_t^{\hat{\theta}}(\mathbf{z}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) \right\|_2^2 \right] \\ &= R_{\text{cdm},t}(\mathbf{M}_t^{\hat{\theta}}, \mathbf{E}_{\text{tx}}) - \bar{R}_{\text{cdm},t}^* \\ &= \underbrace{\inf_{\theta \in \Theta_{L,J,D,D',B,M}} R_{\text{cdm},t}(\mathbf{M}_t^{\theta}, \mathbf{E}_{\text{tx}}) - \bar{R}_{\text{cdm},t}^*}_{\text{approximation error}} + \underbrace{R_{\text{cdm},t}(\mathbf{M}_t^{\hat{\theta}}, \mathbf{E}_{\text{tx}}) - \inf_{\theta \in \Theta_{L,J,D,D',B,M}} R_{\text{cdm},t}(\mathbf{M}_t^{\theta}, \mathbf{E}_{\text{tx}})}_{\text{generalization error}}. \end{aligned}$$

We claim the follow bounds on the approximation and generalization error which we will prove momentarily.

- (a). If we choose $J = \tilde{\mathcal{O}}(L)$, $D = \mathcal{O}(SL)$, $D' = \tilde{\mathcal{O}}(\bar{m}SL^3)$, and $B = \tilde{\mathcal{O}}(L_B + (SL + \bar{m}^2)\sqrt{M})$, then the approximation error

$$\begin{aligned} & \inf_{\theta \in \Theta_{L,J,D,D',B,M}} R_{\text{cdm},t}(\mathbf{M}_t^{\theta}, \mathbf{E}_{\text{tx}}) - \bar{R}_{\text{cdm},t}^* \\ & \leq d_{\text{im}} \cdot \tilde{\mathcal{O}} \left(\sqrt{\frac{(SL^8\bar{m}^2 + M)S^5L^3}{n}} + S^7L_B^2 \left(\text{Suff}(\mathbf{S}) + \frac{1}{M} \right) \right). \end{aligned} \quad (\text{A.90})$$

- (b). Under the same choice of model class $\Theta_{L,J,D,D',B,M}$, the generalization error

$$R_{\text{cdm},t}(\mathbf{M}_t^{\hat{\theta}}, \mathbf{E}_{\text{tx}}) - \inf_{\theta \in \Theta_{L,J,D,D',B,M}} R_{\text{cdm},t}(\mathbf{M}_t^{\theta}, \mathbf{E}_{\text{tx}}) \leq \tilde{\mathcal{O}} \left(d_{\text{im}} \cdot \sqrt{\frac{(SL^8\bar{m}^2 + M)S^5L^3}{n}} \right)$$

with probability at least $1 - 1/n$.

Combining the claims yields Theorem 2.12.

(a) Approximation error. Take some $\delta' > 0$ which will be defined later. For the feed forward layers, we use Lemmas A.22 and A.24 with $\delta = \delta' \ll 1$. For the self-attention layers, we use Lemma A.23 with $\delta = \frac{\delta'}{\max_v \|\mathbf{q}_v^{(\ell)}\|_{\infty}}$. Following the argument in the proof of Theorem 2.10, $\mathbf{q}_v^{(\ell)} = f_{\iota(\text{pa}(L-\ell)(v))}^{(\ell)}(\mathbf{h}_v^{(\ell)})$ is bounded by $3(1 \vee \log SB_{\psi})$, and δ in Lemma A.23 is bounded by $\frac{\delta'}{3(1 \vee \log SB_{\psi})}$.

The error from each operation is then bounded by δ' in the $\|\cdot\|_{\infty}$ -norm. Now we can apply Lemma A.26 to obtain that

$$\begin{aligned} \|\mathbf{M}_t(\mathbf{z}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) - \mathbf{m}_{\star,t}(\mathbf{z}_t, \mathbf{x}_{\text{tx}})\|_2 & \leq d_{\text{im}}^{\frac{1}{2}} 8^{L+1} S^2 \delta \times \prod_{1 \leq k \leq L} (2m_{\text{im}}^{(k)} + 3) + d_{\text{im}}^{\frac{1}{2}} S^2 \delta_{\text{tx}} \\ & \leq d_{\text{im}}^{\frac{3}{2}} 40^{L+1} S^2 \delta + d_{\text{im}}^{\frac{1}{2}} S^2 \delta_{\text{tx}} \end{aligned}$$

We choose

$$\delta' = \frac{1}{40^{L+1}d_{\text{im}}S^2} \left(\frac{(SL^8\bar{m}^2 + M)S^5L^3}{n} \right)^{1/4}$$

with $\bar{m} = \max\{\max_k m_{\text{tx}}^{(k)}, \max_k m_{\text{im}}^{(k)}\}$. Moreover, from Proposition 2.9, the definition of δ_{tx} in Eq. (A.81) and Lemma A.19, it can be verified that there exists some $\text{Adap}(\cdot)$ in Eq. (2.20) such that, $\|W_{\text{ada}}^{(1)}\|_{\text{op}} \leq C'L_B$, $\|W_{\text{ada}}^{(2)}\|_{\text{op}} \leq C'(SL + \bar{m}^2)\sqrt{M}$, and

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_{\text{tx}}} \delta_{\text{tx}}^2 &\leq \mathbb{E}_{\mathbf{x}_{\text{tx}}} \|\log \text{trun}_{\text{tx}}(\widehat{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}})) - \log \mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}})\|_2^2 \leq CS^2 \cdot \mathbb{E}_{\mathbf{x}_{\text{tx}}} \|\widehat{\mathbf{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}) - \mathbf{E}_{\text{tx},\star}(\mathbf{x}_{\text{tx}})\|_2^2 \\ &\leq CS^2 \cdot L_B^2 \cdot L_\Gamma^2 \cdot p_\star \cdot (\text{Suff}(\mathbf{S}) + M^{-1}) \leq CS^3 \cdot L_B^2 \cdot (\text{Suff}(\mathbf{S}) + M^{-1}) \end{aligned}$$

for some $C, C' > 0$ depending polynomially on B_ψ^m , where the last line follows since $p_\star = S$, and $L_\Gamma \leq cB_\psi^{2m}$ by Lemma A.20 and the fact that $\Upsilon_\star^{-1} = \exp(\cdot)$. Putting pieces together, according to Lemma A.9 and Lemma A.10, we now know that there exists some parameter $\boldsymbol{\theta} \in \Theta_{L,J,D,D',B}$ such that bound (A.90) is satisfied and

$$\begin{aligned} D &\leq d_f + d_p = 3SL + 2L = \mathcal{O}(SL), \\ J &\lesssim (\log \log(SB_\psi/\delta')) \log(SB_\psi/\delta') = \tilde{\mathcal{O}}(L), \\ D' &= \|\mathbf{j}\|_\infty \lesssim \bar{m}S(\log(SB_\psi/\delta'))^3 + d_f + d_p = \tilde{\mathcal{O}}(\bar{m}SL^3), \\ B &\lesssim S(B_\psi^2 + \log(SB_\psi/\delta')) + \bar{m}^2 + \log \frac{d \log(SB_\psi)}{\delta'} + (L_B + (SL + \bar{m}^2)\sqrt{M}) \\ &= \tilde{\mathcal{O}}(L_B + (SL + \bar{m}^2)\sqrt{M}). \end{aligned}$$

(b) generalization error. Since $\mathbf{M}_t^{\hat{\boldsymbol{\theta}}}$ is the minimizer of $\widehat{\mathbf{R}}_{\text{cdm},t}(\mathbf{M}_t^{\boldsymbol{\theta}}, \mathbf{E}_{\text{tx}})$ defined in Eq. (2.21), we have

$$\begin{aligned} &\mathbf{R}_{\text{cdm},t}(\mathbf{M}_t^{\hat{\boldsymbol{\theta}}}, \mathbf{E}_{\text{tx}}) - \inf_{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B,M}} \mathbf{R}_{\text{cdm},t}(\mathbf{M}_t^{\boldsymbol{\theta}}, \mathbf{E}_{\text{tx}}) \\ &\leq 2 \sup_{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B,M}} |\widehat{\mathbf{R}}_{\text{cdm},t}(\mathbf{M}_t^{\boldsymbol{\theta}}, \mathbf{E}_{\text{tx}}) - \mathbf{R}_{\text{cdm},t}(\mathbf{M}_t^{\boldsymbol{\theta}}, \mathbf{E}_{\text{tx}})|. \end{aligned} \quad (\text{A.91})$$

Next, we verify the conditions for Lemma A.51 and then apply the lemma to derive an upper bound for the R.H.S. of Eq. (A.91).

In Lemma A.51, take $\Theta = \Theta_{L,J,D,D',B,M}$, $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') = \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$, $z_i = (\mathbf{x}_{\text{im}}^{(i)}, \mathbf{x}_{\text{tx}}^{(i)}, \mathbf{z}_t^{(i)})$, and

$$f(z_i; \boldsymbol{\theta}) = \frac{1}{d_{\text{im}}} \|\mathbf{x}_{\text{im}}^{(i)} - \mathbf{M}_t^{\boldsymbol{\theta}}(\mathbf{z}_t^{(i)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(i)}))\|_2^2.$$

Verification of condition (a) in Lemma A.51. We note that the set $\Theta_{L,J,D,D',B,M}$ with metric $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') = \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ has a diameter $B_\rho := 2B$. Furthermore, the dimension of $\Theta_{L,J,D,D',B,M}$ is

bounded by $d_\rho := (J+3)(2L+1)(D+D'+1)^2 + S + 2SM = \tilde{\mathcal{O}}(S^2 L^8 \bar{m}^2 + 2SM)$. Thus, by Example 5.8 in [Wainwright \(2019b\)](#), we have $\log \mathcal{N}(\Delta; \Theta_{L,J,D,D',B,M}, \|\cdot\|) \leq d_\rho \log(1+2r/\Delta) \leq d_\rho \log(2A_\rho r/\Delta)$ for $\Delta \in (0, 2r]$ with $A_\rho = 2$.

Verification of condition (b) in Lemma [A.51](#). Since $f(z_i; \boldsymbol{\theta})$ is $4d_{\text{im}}S^2$ -bounded by the construction of $\mathbf{M}_t^\theta$ and the fact that $\|\mathbf{x}_{\text{im}}\|_\infty \leq S$, it follows that $f(z_i; \boldsymbol{\theta}) - \mathbb{E}[f(z_i; \boldsymbol{\theta})]$ is $\sigma = cS^2$ -sub-Gaussian for all $\boldsymbol{\theta} \in \Theta_{L,J,D,D',B,M}$ for some numerical constant $c > 0$.

Verification of condition (c) in Lemma [A.51](#). By Lemma [A.43](#) and the boundedness condition, we have

$$\begin{aligned} & |f(z_i; \boldsymbol{\theta}) - f(z_i; \boldsymbol{\theta}')| \\ & \leq \frac{1}{d_{\text{im}}} |\langle \mathbf{M}_t^{\boldsymbol{\theta}'}(\mathbf{z}_t^{(i)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(i)})) - \mathbf{M}_t^\theta(\mathbf{z}_t^{(i)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(i)})), \\ & \quad 2\mathbf{x}_{\text{im}}^{(i)} - \mathbf{M}_t^{\boldsymbol{\theta}'}(\mathbf{z}_t^{(i)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(i)})) - \mathbf{M}_t^\theta(\mathbf{z}_t^{(i)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(i)})) \rangle| \\ & \leq \frac{4S}{\sqrt{d_{\text{im}}}} \|\mathbf{M}_t^{\boldsymbol{\theta}'}(\mathbf{z}_t^{(i)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(i)})) - \mathbf{M}_t^\theta(\mathbf{z}_t^{(i)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(i)}))\|_2 \\ & \leq B_f \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \end{aligned}$$

where $B_f := ((cB)^{18JL} S^9 B_{\text{read}}^3 \log^3 \bar{m})^{2L+2} \exp(2B_{\text{read}})$ and $B_{\text{read}} = 4\bar{m} \log B_\psi$. Therefore, we may choose $\sigma' = B_f$ and condition (c) is hence satisfied.

Now, invoking Lemma [A.51](#) and plugging in the values of $d_\rho, \sigma, \sigma', A_\rho, B_\rho$, we find

$$\begin{aligned} & \sup_{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B,M}} |\hat{\mathbf{R}}_{\text{cdm},t}(\mathbf{M}_t^\theta, \mathbf{E}_{\text{tx}}) - \mathbf{R}_{\text{cdm},t}(\mathbf{M}_t^\theta, \mathbf{E}_{\text{tx}})| \\ & \leq d_{\text{im}} \cdot c\sigma \sqrt{\frac{d_\rho \log(2A_\rho(1+B_\rho\sigma'/\sigma)) + \log(1/\eta)}{n}} \\ & \leq \tilde{\mathcal{O}}\left(d_{\text{im}} S^2 \sqrt{\frac{(SL^8 \bar{m}^2 + M)SL^3 + \log(1/\eta)}{n}}\right) \end{aligned}$$

with probability at least $1 - \eta$. Setting $\eta = 1/n$ completes the proof.

A.6.3 Joint training of denoising function and text representation

In this section, we analyze the sample complexity of jointly learning the conditional denoising models (CDMs) and the text representation within the JGHM framework. Following the setup of Section [2.4.2](#), suppose we are given a dataset of iid samples $\{(\mathbf{z}_t^{(i)}, \mathbf{x}_{\text{im}}^{(i)}, \mathbf{x}_{\text{tx}}^{(i)})\}_{i \in [n]} \sim_{\text{iid}} \mu_{\star, t}$.

The conditional denoiser is modeled as

$$\mathbf{M}_t^\theta(\mathbf{z}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) = \text{read}_{\text{cdm}} \circ \text{TF}_{\text{cdm}} \circ \text{Emb}_{\text{cdm}}(\mathbf{z}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})),$$

where $\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}) = \text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}})$ as defined in Section 2.4.1, and the remaining components are the same as defined in Section 2.4.2, except that in the embedding Emb_{cdm} , we let

$$\mathbf{h}_{\text{tx},d}^{(0)} = \widetilde{\text{trun}}_{\text{tx}}(\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})), \quad \text{where } \widetilde{\text{trun}}_{\text{tx}}(z) := \text{proj}_{[-B_{\text{read}}^{\text{tx}}, B_{\text{read}}^{\text{tx}}]}(z),$$

in contrast to Eq. (A.80). During pre-training, we optimize the parameter $\boldsymbol{\theta} = \mathbf{W}_{\text{cdm}}$, while holding $\text{read}_{\text{cdm}} \text{Emb}_{\text{cdm}}$ as fixed. More specifically, we find the model via empirical risk minimization

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B}^{\text{cdm}}}{\text{argmin}} \left\{ \widehat{\mathbf{R}}_{\text{cdm},t}(\mathbf{M}_t^{\boldsymbol{\theta}}, \mathbf{E}_{\text{tx}}) := \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_{\text{im}}^{(i)} - \mathbf{M}_t^{\boldsymbol{\theta}}(\mathbf{z}_t^{(i)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(i)}))\|_2^2 \right\}, \quad (\text{A.92})$$

where the parameter space is defined as

$$\begin{aligned} \Theta_{L,J,D,D',B}^{\text{cdm}} &:= \left\{ \mathbf{W}_{\text{cdm}}, \mathbf{W}_{\text{tx}} \text{ as defined in Eq. (2.16)}; \right. \\ \|\boldsymbol{\theta}\| &:= \max_{i \in [J+1], \ell \in [2L+1]} \{ \|W_{i,\text{cdm}}^{(\ell)}\|_{\text{op}}, \|W_{Q,\text{cdm}}^{(\ell)}\|_{\text{op}}, \|W_{K,\text{cdm}}^{(\ell)}\|_{\text{op}}, \|W_{V,\text{cdm}}^{(\ell)}\|_{\text{op}} \} \\ &\quad \vee \max_{i \in [J+1], \ell \in [L]} \{ \|W_{i,\text{tx}}^{(\ell)}\|_{\text{op}}, \|W_{Q,\text{tx}}^{(\ell)}\|_{\text{op}}, \|W_{K,\text{tx}}^{(\ell)}\|_{\text{op}}, \|W_{V,\text{tx}}^{(\ell)}\|_{\text{op}} \} \leq B \}. \end{aligned} \quad (\text{A.93})$$

Similar to Theorem 2.12, we have the following result

Theorem A.25 (Estimation error of conditional denoising function, joint training). *Suppose that Assumption 2.4 and Assumption 2.5 hold. For simplicity, assume $t = 1$. Let $\Theta_{L,J,D,D',B}^{\text{cdm}}$ be the set defined in Eq. (A.93), where $J = \tilde{\mathcal{O}}(L)$, $D = \mathcal{O}(SL)$, $D' = \tilde{\mathcal{O}}(\overline{m}SL^3)$, and $B = \tilde{\mathcal{O}}(SL + \overline{m}^2)$. Let $\hat{\boldsymbol{\theta}}$ be the empirical risk minimizer defined in Eq. (A.92). Then, with probability at least $1 - 1/n$, we have*

$$\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, \mathbf{z}_t)} \left[\frac{1}{d_{\text{im}}} \|\mathbf{m}_{\star,t}(\mathbf{z}_t, \mathbf{x}_{\text{tx}}) - \mathbf{M}_t^{\hat{\boldsymbol{\theta}}}(\mathbf{z}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))\|_2^2 \right] \leq \tilde{\mathcal{O}} \left(\sqrt{\frac{S^6 L^{11} \overline{m}^2}{n}} \right),$$

where $\tilde{\mathcal{O}}$ hides polynomial factors in $(\log(\overline{m}SLn), (B_{\psi})^m)$.

Proof of Theorem A.25. The proof follows from the same argument as the proof of Theorem 2.12, thus we only highlight the differences here. Similar to the proof of Theorem 2.12, we claim that

- (a). If we choose $J = \tilde{\mathcal{O}}(L)$, $D = \mathcal{O}(SL)$, $D' = \tilde{\mathcal{O}}(\overline{m}SL^3)$, and $B = \tilde{\mathcal{O}}(SL + \overline{m}^2)$, then the approximation error

$$\inf_{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B}^{\text{cdm}}} \mathbf{R}_{\text{cdm},t}(\mathbf{M}_t^{\boldsymbol{\theta}}, \mathbf{E}_{\text{tx}}) - \overline{\mathbf{R}}_{\text{cdm},t}^{\star} \leq d_{\text{im}} \cdot \tilde{\mathcal{O}} \left(\sqrt{\frac{S^6 L^{11} \overline{m}^2}{n}} \right). \quad (\text{A.94})$$

(b). Under the same choice of model class $\Theta_{L,J,D,D',B}^{\text{cdm}}$, the generalization error

$$R_{\text{cdm},t}(\mathbf{M}_t^{\hat{\theta}}, \mathbf{E}_{\text{tx}}) - \inf_{\theta \in \Theta_{L,J,D,D',B}^{\text{cdm}}} R_{\text{cdm},t}(\mathbf{M}_t^{\theta}, \mathbf{E}_{\text{tx}}) \leq \tilde{\mathcal{O}}\left(d_{\text{im}} \cdot \sqrt{\frac{S^6 L^{11} \bar{m}^2}{n}}\right)$$

with probability at least $1 - 1/n$.

Combining the claims yields Theorem A.25.

(a) approximation error By using the same parameter choice as that in Theorem 2.10, we have

$$\delta_{\text{tx}} \leq \delta' \prod_{1 \leq k \leq L} (2m_{\text{tx}}^{(k)} + 3) \leq 5^L d_{\text{tx}}$$

according to Eq. (A.73). As a result, we may choose $\delta' = \frac{\sqrt{S^2 L^{11} \bar{m}^2}}{5^L d_{\text{tx}} \sqrt{n}}$ and obtain Eq. (A.94).

(b) generalization error Likewise, we verify the conditions for Lemma A.51. We take $\Theta = \Theta_{L,J,D,D',B}^{\text{cdm}}$, $\rho(\theta, \theta') = \|\theta - \theta'\|$, $z_i = (\mathbf{x}_{\text{im}}^{(i)}, \mathbf{x}_{\text{tx}}^{(i)}, \mathbf{z}_t^{(i)})$, and

$$f(z_i; \theta) = \frac{1}{d_{\text{im}}} \|\mathbf{x}_{\text{im}}^{(i)} - \mathbf{M}_t^{\theta}(\mathbf{z}_t^{(i)}, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}^{(i)}))\|_2^2.$$

Similarly, it can be verified that condition (a) in Lemma A.51 is satisfied with $A_{\rho} = 2$, $B_{\rho} = 2B$, and number of parameters $d_{\rho} = \tilde{\mathcal{O}}(S^2 L^8 \bar{m}^2)$; $f(z_i; \theta) - \mathbb{E}[f(z_i; \theta)]$ is $\sigma = cS^2$ -sub-Gaussian for all $\theta \in \Theta_{L,J,D,D',B}^{\text{cdm}}$; and similar to Lemma A.43, it can be verified that

$$|f(z_i; \theta) - f(z_i; \theta')| \leq B_f \|\theta - \theta'\|, \quad \text{where } B_f := ((cB)^{18JL} S^9 B_{\text{read}}^3 \log^3 \bar{m})^{3L+3} \exp(2B_{\text{read}}).$$

Finally, invoking Lemma A.51 and plugging in the values of $d_{\rho}, \sigma, \sigma', A_{\rho}, B_{\rho}$ yields the desired bound. □

A.6.4 Evaluation of error propagation

Similarly to Lemma A.18 in Appendix A.5.5, we evaluate the propagation of the errors. We denote the estimation error of $h_{\text{tx},r}^{(0)}$ by δ_{tx} , so that Lemma A.26 can be used for both simultaneous training and two-stage training. For simultaneous training of the image and text models, the error propagation lemma for contrastive learning (Lemma A.18) can be used to bound δ_{tx} .

Lemma A.26 (Evaluation of error propagation). *Assume we have functions $\mathbf{f}_{\downarrow,\iota}^{(\ell)}, \mathbf{f}_{\uparrow,\iota}^{(\ell)}$ ($1 \leq \ell \leq L, \iota \in [m_{\text{tx}}^{(\ell)}]$) such that*

$$\begin{aligned} \|f_{\downarrow,\iota}^{(\ell)}(h) - \mathbf{f}_{\downarrow,\iota}^{(\ell)}(h)\|_\infty &\leq \delta, \quad \forall h \in \mathbb{R}^S \text{ such that } \max_{s \in S} h_s = 0, \ell \in [L], \\ \|f_{\uparrow,\iota}^{(\ell)}(h) - \mathbf{f}_{\uparrow,\iota}^{(\ell)}(h)\|_\infty &\leq \delta, \quad \forall h \in \mathbb{R}^S \text{ such that } \max_{s \in S} h_s = 0, \ell \in [L], \end{aligned}$$

and $\|\boldsymbol{\delta}_v^{(\ell)}\|_\infty \leq \delta$ holds for all $\ell = L-1, \dots, 0$ and $v \in \mathcal{V}_{\text{im}}^{(L)}$. Moreover, we assume that $\|h_{\text{tx},\text{r}}^{(0)} - \mathbf{h}_{\text{tx},d_{\text{tx}}}^{(0)}\|_\infty \leq \delta_{\text{tx}}$.

Consider the approximated update introduced in (A.88) and (A.89). Then, we have the following bound on the error propagation:

$$\max_{v \in \mathcal{V}_{\text{im}}^{(L)}} \|\mathbf{h}_{\downarrow,v}^{(\ell)} - h_{\downarrow,\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}\|_\infty \leq \delta \times \prod_{\ell+1 \leq k \leq L} (2m_{\text{im}}^{(k)} + 3), \quad (\text{A.95})$$

$$\max_{v \in \mathcal{V}_{\text{im}}^{(L)}} \|\mathbf{q}_{\downarrow,v}^{(\ell)} - q_{\downarrow,\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}\|_\infty \leq \delta \times \prod_{\ell+1 \leq k \leq L} (2m_{\text{im}}^{(k)} + 3), \quad (\text{A.96})$$

$$\max_{v \in \mathcal{V}_{\text{im}}^{(\ell)}} \|\mathbf{b}_v^{(\ell)} - b_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}\|_\infty \leq 8^\ell \delta \times \prod_{1 \leq k \leq L} (2m_{\text{im}}^{(k)} + 3) + \delta_{\text{tx}}, \quad (\text{A.97})$$

$$\max_{v \in \mathcal{V}_{\text{im}}^{(L)}} \|\mathbf{b}_v^{(L+1)} - b_v^{(L+1)}\|_\infty \leq 8^{L+1} \delta \times \prod_{1 \leq k \leq L} (2m_{\text{im}}^{(k)} + 3) + \delta_{\text{tx}}. \quad (\text{A.98})$$

Furthermore, we have

$$\max_{v \in \mathcal{V}_{\text{im}}^{(L)}} \|\mathbf{M}_{t,v} - (\mathbf{m}_{\star,t}(\mathbf{z}_t, \mathbf{x}_{\text{tx}}))_v\|_\infty \leq 8^{L+1} S^2 \delta \times \prod_{1 \leq k \leq L} (2m_{\text{im}}^{(k)} + 3) + S^2 \delta_{\text{tx}}. \quad (\text{A.99})$$

Proof. The bounds (A.95) and (A.96) are the same as (A.73) and (A.74) of Lemma A.18. Thus let us focus on the upsampling process of the image model. We will prove (A.97) using the induction, using (A.95) and (A.96). Until the final part, let us assume $\delta_{\text{tx}} = 0$.

First, we prove that (A.97) holds for $\ell = 1$. For $v \in \mathcal{V}_{\text{im}}^{(L)}$, we have

$$\begin{aligned} \|\mathbf{b}_v^{(1)} - b_{\text{pa}^{(L-1)}(v)}^{(1)}\|_\infty &\leq 2\|\mathbf{h}_v^{(0)} + \mathbf{h}_{\text{tx},d_{\text{tx}}}^{(0)} - \mathbf{q}_v^{(1)} - (h_{\text{r}}^{(0)} + h_{\text{tx},\text{r}}^{(0)} - q_{\text{pa}^{(L-1)}(v)}^{(1)})\|_\infty \\ &\leq 2\|\mathbf{h}_v^{(0)} - h_{\text{r}}^{(0)}\|_\infty + 2\|\mathbf{h}_{\text{tx},d_{\text{tx}}}^{(0)} - h_{\text{r}}^{(0)}\|_\infty + 2\|\mathbf{q}_v^{(1)} - q_{\text{pa}^{(L-1)}(v)}^{(1)}\|_\infty \\ &\leq 4\delta \times \prod_{1 \leq k \leq L} (2m_{\text{im}}^{(k)} + 3), \end{aligned}$$

where we used Lemma A.45 for the first inequality and (A.95) and (A.96) for the last inequality.

Then, let us assume that (A.97) holds for ℓ and prove it for $\ell + 1$ ($\ell = 1, \dots, L-1$). For $v \in \mathcal{V}_{\text{im}}^{(L)}$, we have

$$\begin{aligned} &\|\mathbf{b}_v^{(\ell+1)} - b_{\text{pa}^{(L-\ell-1)}(v)}^{(\ell+1)}\|_\infty \\ &\leq 2\|\mathbf{f}_{\uparrow,\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{b}_v^{(\ell)}) + \mathbf{h}_v^{(\ell)} - \mathbf{q}_v^{(\ell+1)} - (f_{\uparrow,\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(b_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}) + h_{\text{pa}^{(L-\ell)}(v)}^{(\ell)} - q_{\text{pa}^{(L-\ell-1)}(v)}^{(\ell+1)})\|_\infty \\ &\leq 2\|\mathbf{f}_{\uparrow,\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{b}_v^{(\ell)}) - f_{\uparrow,\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(b_{\text{pa}^{(L-\ell)}(v)}^{(\ell)})\|_\infty \end{aligned}$$

$$\begin{aligned}
& + 2\|f_{\uparrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{b}_v^{(\ell)}) - f_{\uparrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{b}_{\text{pa}^{(L-\ell)}(v)}^{(\ell)})\|_\infty \\
& + 2\|\mathbf{h}_v^{(\ell)} - h_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}\|_\infty + 2\|\mathbf{q}_v^{(\ell+1)} - q_{\text{pa}^{(L-\ell-1)}(v)}^{(\ell+1)}\|_\infty \\
& \leq 2\delta + 2\|\mathbf{b}_v^{(\ell)} - b_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}\|_\infty + 2\|\mathbf{h}_v^{(\ell)} - h_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}\|_\infty + 2\|\mathbf{q}_v^{(\ell+1)} - q_{\text{pa}^{(L-\ell-1)}(v)}^{(\ell+1)}\|_\infty \\
& \leq 2\delta + 2 \times 8^\ell \delta \times \prod_{1 \leq k \leq L} (2m_{\text{im}}^{(k)} + 3) + 4\delta \times \prod_{1 \leq k \leq L} (2m_{\text{im}}^{(k)} + 3) \\
& \leq 8^{(\ell+1)} \delta \times \prod_{1 \leq k \leq L} (2m_{\text{im}}^{(k)} + 3),
\end{aligned}$$

where we used Lemma A.49 for the first inequality and Lemma A.45 for the third inequality. Now (A.97) is confirmed for $\ell + 1$. In the same way, (A.98) is proved. Note that softmax is 1-Lipschitz with respect to the $\|\cdot\|_\infty$ norm. Thus, the bound (A.99) directly follows from (A.98).

Finally, we consider how the error from the text model δ_{tx} propagates. For this, we only need to bound how the message passing algorithm changes, because the difference between the message passing algorithm and its neural network approximation is already bounded. According to Lemma A.27, that is bounded by $S^2 \delta_{\text{tx}}$. Now we have obtained the assertion. $\square$

Lemma A.27. *Suppose that we run the upsampling process of the message passing algorithm (A.77) by changing $h_{\text{tx},r}^{(0)}$ to $h'_{\text{tx},r}^{(0)}$ with $\|h'_{\text{tx},r}^{(0)} - h_{\text{tx},r}^{(0)}\|_\infty \leq \delta_{\text{tx}}$ while all the others are the same. Then, the deviation of the new optimal prediction $\mathbf{m}'_{\star,t,v}$ from $\mathbf{m}_{\star,t,v}$ is bounded by*

$$\|\mathbf{m}'_{\star,t,v} - \mathbf{m}_{\star,t,v}\|_\infty \leq S^2 \delta_{\text{tx}}$$

for all $v \in \mathcal{V}_{\text{im}}^{(L)}$.

Proof. Because of softmax in the final part of Eq. (A.77), we do not need to consider normalize in the message passing algorithm. Thus, let us consider how the change in $h_{\text{tx},r}^{(0)}$ propagates in the following pipeline.

$$\begin{aligned}
\bar{b}_{\text{im},r}^{(0)} &= h_{\text{im},r}^{(0)} + h_{\text{tx},r}^{(0)} \in \mathbb{R}^S, \\
\bar{b}_{\text{im},v}^{(\ell)} &= f_{\text{im},\uparrow,\iota(v)}^{(\ell)}(\bar{b}_{\text{im},\text{pa}(v)}^{(\ell-1)} - q_{\text{im},v}^{(\ell)}) + h_{\text{im},v}^{(\ell)} \in \mathbb{R}^S, \quad v \in \mathcal{V}_{\text{im}}^{(\ell)}, \ell = 1, \dots, L \\
\mathbf{m}_{\star,t,v} &= \sum_{s \in [S]} s \cdot \text{softmax}(\bar{b}_{\text{im},v}^{(L)}) \in \mathbb{R}^S, \quad v \in \mathcal{V}_{\text{im}}^{(L)},
\end{aligned}$$

According to Lemma A.49, we know that the change of $\bar{b}_{\text{im},v}^{(\ell)}$ evaluated by $\|\cdot\|_\infty$ -norm is bounded by the change of $\bar{b}_{\text{im},v}^{(\ell-1)}$. Thus, the change of $\bar{b}_{\text{im},v}^{(L+1)}$ is at most δ_{tx} in the $\|\cdot\|_\infty$ -norm. Moreover, softmax is 1-Lipschitz with respect to the $\|\cdot\|_\infty$ -norm, and s is bounded by S , which yields that the estimation of each leaf variable changes at most $S^2 \delta_{\text{tx}}$. $\square$

A.7 Proof of Theorem A.2

A.7.1 Overview

This section solves the problem of estimating the posterior probability of the next word $\mu_\star(x_{\text{tx},i+1}|\mathbf{x}_{\text{im}}, x_{\text{tx},1}, \dots, x_{\text{tx},i})$ for every $i = 1, \dots, d_{\text{tx}} - 1$ in parallel. In this overview section, Appendix A.7.1 first introduces the belief propagation algorithm, which exactly calculates $\mu_\star(x_{\text{tx},i+1}|\mathbf{x}_{\text{im}}, x_{\text{tx},1}, \dots, x_{\text{tx},i})$ for each fixed i . We then discuss how to parallelize the belief propagation algorithm into the message passing algorithm in Appendix A.7.1, and finally explain how to implement the message passing algorithm with transformer networks in Appendix A.7.1.

In the following, we identify nodes and integers by following Definition A.1. When we refer to $v + 1$, it means the leaf node u such that $u = v + 1$ in the interger representation of Definition A.1. We remark that, although the next node $v + 1$ is not defined for $v = d_{\text{tx}}$, we do not separate the case of $v = d_{\text{tx}}$ when we use the notation $v + 1$ in the following, because how to deal with the case of $v = d_{\text{tx}}$ does not affect the prediction of $x_{\text{tx},2}, \dots, x_{\text{tx},d_{\text{tx}}}$. Also, the discussion mainly focuses on the text processing, and therefore we sometimes omit the subscript “tx”.

Belief propagation algorithm

To predict an unobserved leaf node of the text, the belief propagation algorithm can exactly calculates the posterior probability. Suppose that we have $\mathbb{P}[s|\mathbf{x}_{\text{im}}]$ (, which can be approximated either by the transformer network $\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}$ in the contrastive learning or by the text embedding in the two-stage training). Given this, the belief propagation consists of the downsampling

$$\begin{aligned} \nu_{\downarrow,v}^{(L)}(x_{\text{tx},v}^{(L)}) &= \mathbb{1}[x_{\text{tx},v}^{(L)} = x_{\text{tx},v}] \ (v \leq i), \ \frac{1}{S} \text{ (otherwise)}, \quad v \in \mathcal{V}_{\text{tx}}^{(L)}, \\ \nu_{\downarrow,v}^{(\ell)}(x_{\text{tx},v}^{(\ell)}) &\propto \sum_{x_{\text{tx},\mathcal{C}(v)}^{(\ell+1)}} \prod_{v' \in \mathcal{C}(v)} \left(\psi_{\text{tx},\iota(v')}^{(\ell+1)}(x_{\text{tx},v}^{(\ell)}, x_{\text{tx},v'}^{(\ell+1)}) \nu_{\downarrow,v'}^{(\ell+1)}(x_{\text{tx},v'}^{(\ell+1)}) \right), \ v \in \mathcal{V}_{\text{tx}}^{(\ell)}, \ \ell = L-1, \dots, 1. \end{aligned} \quad (\text{A.100})$$

and the upsampling

$$\begin{aligned} \nu_{\uparrow,r}^{(0)}(x_{\text{tx},r}^{(0)}) &= \mathbb{P}[x_{\text{tx},r}^{(0)}|\mathbf{x}_{\text{im}}] \\ \nu_{\uparrow,v}^{(\ell)}(x_{\text{tx},v}^{(\ell)}) &\propto \sum_{x_{\text{tx},\text{pa}(v)}^{(\ell-1)}, x_{\text{tx},\mathcal{N}(v)}^{(\ell)}} \psi^{(\ell)}(x_{\text{tx},\text{pa}(v)}^{(\ell-1)}, x_{\text{tx},\mathcal{C}(\text{pa}(v))}^{(\ell)}) \nu_{\uparrow,\text{pa}(v)}^{(\ell-1)}(x_{\text{tx},\text{pa}(v)}^{(\ell-1)}) \prod_{v' \in \mathcal{N}(v)} \nu_{\downarrow,v'}^{(\ell)}(x_{\text{tx},v'}^{(\ell)}), \\ &\quad v \in \text{pa}^{(L-\ell)}(i+1), \ \ell = 1, \dots, L. \end{aligned} \quad (\text{A.101})$$

These beliefs ν are normalized so that $\sum_s \nu_s = 1$. The correctness of this algorithm is formally stated as follows.

Lemma A.28 (BP calculates the posterior probability of the next word exactly). *When applying the belief propagation algorithm shown in (A.100) and (A.101), it holds that $\nu_{\uparrow,n}^{(L)}(x_{\text{tx},i+1}) = \mu_\star(x_{\text{tx},i+1}|\mathbf{x}_{\text{im}}, x_{\text{tx},1}, \dots, x_{\text{tx},i})$.*

Proof. Referring to classical results (Pearl, 1982; Wainwright et al., 2008; Mezard and Montanari, 2009), when we replace $\nu_{\uparrow,r}^{(0)}(x_r^{(0)}) \propto \mathbb{P}[x_r^{(0)}|\mathbf{x}_{\text{im}}]$ by the unconditioned $\nu_{\uparrow,r}^{(0)}(x_r^{(0)}) \propto \mathbb{P}[x_r^{(0)}]$ in (A.101), it holds that $\nu_{\uparrow,i+1}^{(L)}(x_{\text{tx},i+1}) = \mu_{\star}(x_{\text{tx},i+1}|x_{\text{tx},1}, \dots, x_{\text{tx},i})$. It is obvious to see that using $\nu_{\uparrow,r}^{(0)}(x_r^{(0)}) = \mathbb{P}[x_r^{(0)}|\mathbf{x}_{\text{im}}]$ corresponds to conditioning on $\mathbf{x}_{\text{im}}$ and that $\nu_{\uparrow,i+1}^{(L)}(x_{\text{tx},i+1}) = \mu_{\star}(x_{\text{tx},i+1}|\mathbf{x}_{\text{im}}, x_{\text{tx},1}, \dots, x_{\text{tx},i})$. $\square$

Parallelization with message passing algorithm

We then parallelize the belief propagation algorithm for different i with the message passing algorithm. We achieve this by grouping common variables across different i into a single variables. Remind that, for $v \in \mathcal{V}_{\text{tx}}^{(\ell)}$, $v^{(\ell')}$ means the node $u \in \mathcal{V}_{\text{im}}^{(\ell')}$ such that its corresponding integer is the same as that of v (Definition A.1). Thus, for $u \in \mathcal{V}_{\text{tx}}^{(L)}$, $u^{(\ell)} \in \mathcal{N}(\text{pa}^{(L-\ell)}(v))$ means that u has the corresponding node in the ℓ -th level and that the corresponding node is a neighbor of $\text{pa}^{(L-\ell)}(v)$.

Downsampling. Starting with $h_v^{(L)} = x_{\text{tx},v}$ ($v \in \mathcal{V}_{\text{tx}}^{(L)}$), the downsampling process is defined as

$$\begin{aligned} q_v^{(\ell)} &= f_{\downarrow, \ell(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(h_v^{(\ell)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(\ell)}, \ell = L, \dots, 1, \\ h_v^{(\ell-1)} &= \text{normalize}\left(\sum_{\substack{u^{(\ell)} \in \mathcal{N}(\text{pa}^{(L-\ell)}(v)) \\ \text{or } u=v}} \mathbb{1}[u \leq v] q_{v'}^{(\ell)}\right) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(\ell)}, \ell = L, \dots, 1, \end{aligned} \quad (\text{A.102})$$

where

$$\begin{aligned} (f_{\downarrow, \ell}^{(L)}(x))_s &= \log \psi_{\text{tx}, \ell}^{(L)}(s, x), & x \in [S], s \in [S], \\ (f_{\downarrow, \ell}^{(\ell)}(h))_s &= \log \sum_{s \in [S]} \psi_{\text{tx}, \ell}^{(\ell)}(s, a) e^{h_a}, & h \in \mathbb{R}^S, s \in [S], \ell = L-1, \dots, 1. \end{aligned} \quad (\text{A.103})$$

Upsampling. The upsampling process is defined as

$$\begin{aligned} \bar{b}_v^{(0)} &= h_v^{(0)} + (\log \mathbb{P}[s|\mathbf{x}_{\text{im}}])_{s \in [S]} \in \mathbb{R}^S, \\ \bar{b}_v^{(\ell)} &= \begin{cases} f_{\uparrow, \ell(\text{pa}^{(L-\ell)}(v+1))}^{(\ell)}(\text{normalize}(\bar{b}_v^{(\ell-1)} - q_v^{(\ell)}) + h_v^{(\ell)}), & \text{(if } \text{pa}^{(L-\ell)}(v) = \text{pa}^{(L-\ell)}(v+1)) \\ f_{\uparrow, \ell(\text{pa}^{(L-\ell)}(v+1))}^{(\ell)}(\text{normalize}(\bar{b}_v^{(\ell-1)})), & \text{(otherwise)} \end{cases} \\ &v \in \mathcal{V}_{\text{tx}}^{(L)}, \ell = 1, 2, \dots, L, \end{aligned} \quad (\text{A.104})$$

where

$$(f_{\uparrow, \ell}^{(\ell)}(h))_s = \log \sum_{a \in [S]} \psi_{\text{tx}, \ell}^{(\ell)}(a, s) e^{h_a}, \quad h \in \mathbb{R}^S, s \in [S], \ell = 1, 2, \dots, L. \quad (\text{A.105})$$

Then, it holds that $\text{softmax}(\bar{b}_i^{(L)})_s = \mu_{\star}(x_{\text{tx},i+1}|\mathbf{x}_{\text{im}}, x_{\text{tx},1}, \dots, x_{\text{tx},i})$ for all $i = 1, \dots, d-1$ and $s \in [S]$.

Proposition A.29 (MP calculates the posterior probability of the next word in parallel). *When applying the message passing algorithm defined in (A.102), (A.103), (A.104), and (A.105), for all $i = 1, \dots, d-1$ and $s \in [S]$, it holds that $\text{softmax}(\bar{b}_i^{(L)})_s = \mu_\star(x_{\text{tx},i} = s | \mathbf{x}_{\text{im}}, x_{\text{tx},1}, \dots, x_{\text{tx},i})$.*

The original message passing algorithm has several issues when it comes to implementing it with transformer networks. First, in the downsampling process (A.102), the number of v' in the summation is not uniform across nodes in the same level. Previously in contrastive learning and conditional diffusion model, we took average with self-attention and then applied the number of elements in the average (e.g., $m^{(\ell)}$) to compute summation. This cannot be directly adopted this time. In addition, the upsampling process (A.104) uses subtraction ($\bar{b}_v^{(\ell-1)} - q_v^{(\ell)}$), “normalize”, and nonlinear transformation $f_{\uparrow,\iota}^{(\ell)}$ in this order, which is not implemented in one transformer block.

Therefore, we prepare the following alternative version of the message passing algorithm to be implemented by a transformer network. The correspondence with the original version is easily confirmed, where $\bar{b}_{v,v}^{(\ell-1)} - q_{\text{im},v}^{(\ell)} = b_v^{(\ell)}$ (if $\text{pa}^{(L-\ell)}(v) = \text{pa}^{(L-\ell)}(v+1)$) and $\bar{b}_v^{(\ell-1)} = b_v^{(\ell)}$ (otherwise).

Downsampling (alternative). Define $a_v^{(\ell)} = \iota(\text{pa}^{(L-\ell)}(v)) + \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}]$, where $v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}$ means v is the rightmost children of one of $u \in \mathcal{V}_{\text{tx}}^{(\ell)}$. Starting with $h_v^{(L)} = x_{\text{tx},v}$ ($v \in \mathcal{V}_{\text{tx}}^{(L)}$), the downsampling process is equivalently written as, for $v \in \mathcal{V}_{\text{tx}}^{(L)}$,

$$\begin{aligned} q_v^{(\ell)} &= f_{\downarrow,\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(h_v^{(\ell)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(L)}, \ell = L, \dots, 1 \\ g_v^{(\ell)} &= \frac{1}{a_v^{(\ell)}} \sum_{v' \in \mathcal{C}(\text{pa}^{(L-\ell+1)}(v))} \mathbb{1}[v' \leq v] q_{v'}^{(\ell)} \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(L)}, \ell = L, \dots, 1, \\ h_v^{(\ell-1)} &= \text{normalize}(a_v^{(\ell)} g_v^{(\ell)} + q_v^{(\ell)} - \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}] q_v^{(\ell)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(L)}, \ell = L, \dots, 1. \end{aligned}$$

Computation of $q_v^{(\ell)}$, $g_v^{(\ell)}$, and $h_v^{(\ell-1)}$ from $h_v^{(\ell)}$ is called the ℓ -th step of the downsampling process (of the text part).

Upsampling (alternative). The upsampling (A.104) is equivalently written as, for $v \in \mathcal{V}_{\text{tx}}^{(L)}$,

$$\begin{aligned} b_v^{(1)} &= \text{normalize}(h_v^{(0)} + (\log \mathbb{P}[s | \mathbf{x}_{\text{im}}])_{s \in [S]} - q_v^{(1)}) \in \mathbb{R}^S, \\ b_v^{(\ell+1)} &= \begin{cases} \text{normalize}(f_{\uparrow,\iota(\text{pa}^{(L-\ell)}(v+1))}^{(\ell)}(b_v^{(\ell)}) + h_v^{(\ell)} - q_v^{(\ell+1)}), & \text{(if } \text{pa}^{(L-\ell-1)}(v) = \text{pa}^{(L-\ell-1)}(v+1)) \\ \text{normalize}(f_{\uparrow,\iota(\text{pa}^{(L-\ell)}(v+1))}^{(\ell)}(b_v^{(\ell)}) + h_v^{(\ell)}), & \left(\begin{array}{l} \text{if } \text{pa}^{(L-\ell)}(v) = \text{pa}^{(L-\ell)}(v+1) \\ \text{but } \text{pa}^{(L-\ell-1)}(v) \neq \text{pa}^{(L-\ell-1)}(v+1) \end{array} \right) \\ \text{normalize}(f_{\uparrow,\iota(\text{pa}^{(L-\ell)}(v+1))}^{(\ell)}(b_v^{(\ell)})), & \text{(otherwise)} \end{cases} \\ &\hspace{25em} \ell = 1, 2, \dots, L, \end{aligned} \tag{A.106}$$

so that $\bar{b}_v^{(L)} = b_v^{(L+1)}$. Computation of $b_{\text{im},v}^{(\ell)}$ is called the ℓ -th step of the upsampling process (of the text part).

Approximation with transformer networks

We approximate the message passing algorithm with transformer networks. We denote a transformer approximation of $h_{\text{im},r}^{(0)} = (\log \mathbb{P}[s|\mathbf{x}_{\text{im}}])_{s \in [S]}$ by $\mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} \in \mathbb{R}^S$. This can be obtained by $\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}$ constructed in contrastive learning, or we can assume this as a given variable in the two-stage training. From now we focus on the text model. We use the numbering of nodes defined in Definition A.1.

Let $\mathbf{h}_v^{(L)} = x_{\text{tx},v} \in [S]$ for all $v \in \mathcal{V}_{\text{tx}}^{(L)}$. After the encoding Emb_{vlm} , we obtain a matrix $\mathbf{H}^{(L)}$ such that

$$\mathbf{H}^{(L)} = \text{Emb}_{\text{vlm}}(\mathbf{x}_{\text{tx}}, \widehat{\mathbf{E}}_{\text{im}}(\mathbf{x}_{\text{im}})) = \begin{bmatrix} \mathbf{0} & \mathbf{h}_0^{(L)} & \mathbf{h}_1^{(L)} & \mathbf{h}_2^{(L)} & \cdots & \mathbf{h}_d^{(L)} \\ \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} & \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} & \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} & \cdots & \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} & \mathbf{P} \end{bmatrix} \in \mathbb{R}^{(d_f+d_p) \times (d+1)}.$$

The shape of $\mathbf{H}^{(L)}$ is $(d_f+d_p) \times (d+1)$, where $d_f = (4L+2)S+1$ and $d_p = 2L+2$. As previously, $\mathbf{P} \in \mathbb{R}^{d_p \times (d+1)}$ is a matrix that encodes the positions of the nodes in d_p -dimensional space, and d_f is the dimensions for the intermediate variables. The output of the image model $\mathbf{h}_{\text{im},d_{\text{im}}}^{(0)}$ is concatenated with every token. We added the leftmost column, which is treated as the variables corresponding to the token position 0, and let $\mathbf{h}_0^{(L)} = \mathbf{0}$.

The text network has $(2L+1)$ transformer blocks, and each transformer block consists of feed forward layer FF_1 , masked self-attention MAttn instead of the previous Attn , feed forward layer FF_2 , and normalization. Using two feed forward layers in a single block is for the sake of clarity in the proof, and it can simply be split into two separate blocks with one feed forward layer, if this is to be avoided. The first L blocks approximate downsampling, and each block is called the $\ell (= L, \dots, 1)$ -th block of downsampling using the decreasing order. The latter $(L+1)$ -blocks approximate upsampling, and each block is called the $\ell (= 0, \dots, L)$ -th block of upsampling using the increasing order.

First consider downsampling. Starting from $\mathbf{H}^{(L)}$, we will construct matrices $\mathbf{H}^{(\ell)}$ ($\ell = L, \dots, 0$), $\mathbf{Q}^{(\ell)}$ ($\ell = L, \dots, 1$), $\mathbf{G}^{(\ell)}$ ($\ell = L, \dots, 1$) of shape $(d_f+d_p) \times (d+1)$, defined as

$$\mathbf{H}^{(\ell)} = \begin{bmatrix} \mathbf{0} & \mathbf{h}_0^{(\ell)} & \mathbf{h}_1^{(\ell)} & \cdots & \mathbf{h}_d^{(\ell)} \\ \mathbf{g}_0^{(L)} & \mathbf{g}_1^{(L)} & \cdots & \mathbf{g}_d^{(L)} \\ \mathbf{q}_0^{(L)} & \mathbf{q}_1^{(L)} & \cdots & \mathbf{q}_d^{(L)} \\ \mathbf{h}_0^{(L)} & \mathbf{h}_1^{(L)} & \cdots & \mathbf{h}_d^{(L)} \\ \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} & \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} & \cdots & \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} \\ \mathbf{P} \end{bmatrix}, \quad \mathbf{Q}^{(\ell)} = \begin{bmatrix} \mathbf{0} & \mathbf{q}_0^{(\ell)} & \mathbf{q}_1^{(\ell)} & \cdots & \mathbf{q}_d^{(\ell)} \\ \mathbf{g}_0^{(L)} & \mathbf{g}_1^{(L)} & \cdots & \mathbf{g}_d^{(L)} \\ \mathbf{q}_0^{(L)} & \mathbf{q}_1^{(L)} & \cdots & \mathbf{q}_d^{(L)} \\ \mathbf{h}_0^{(L)} & \mathbf{h}_1^{(L)} & \cdots & \mathbf{h}_d^{(L)} \\ \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} & \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} & \cdots & \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} \\ \mathbf{P} \end{bmatrix},$$

$$\mathbf{G}^{(\ell)} = \begin{bmatrix} \mathbf{0} \\ \mathbf{g}_0^{(\ell)} & \mathbf{g}_1^{(\ell)} & \cdots & \mathbf{g}_d^{(\ell)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{g}_0^{(L)} & \mathbf{g}_1^{(L)} & \cdots & \mathbf{g}_d^{(L)} \\ \mathbf{q}_0^{(L)} & \mathbf{q}_1^{(L)} & \cdots & \mathbf{q}_d^{(L)} \\ \mathbf{h}_0^{(L)} & \mathbf{h}_1^{(L)} & \cdots & \mathbf{h}_d^{(L)} \\ \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} & \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} & \cdots & \mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} \\ \mathbf{P} \end{bmatrix}.$$

Here, $\mathbf{h}^{(\ell)}, \mathbf{q}^{(\ell)}$ and $\mathbf{g}^{(\ell)}$ are S -dimensional vectors except that $\mathbf{h}^{(L)} \in [S]$.

In the ℓ -th block of downsampling, the feed forward layer $\text{FF}_{\downarrow,1}^{(\ell)}$, a fully-connected ReLU network, receives $\mathbf{H}^{(\ell)}$ and outputs $\mathbf{Q}^{(\ell)}$ by computing $\mathbf{q}_v^{(\ell)}$ from $\mathbf{h}_v^{(\ell)}$:

$$\mathbf{Q}^{(\ell)} = \underbrace{\mathbf{H}^{(\ell)}}_{\text{skip connection}} + \text{FF}_{\downarrow,1}^{(\ell)}(\mathbf{H}^{(\ell)}) = \mathbf{H}^{(\ell)} + \begin{bmatrix} \mathbf{0} \ (\in \mathbb{R}^{((3\ell+L)S) \times (d+1)}) \\ \mathbf{q}_0^{(\ell)} & \mathbf{q}_1^{(\ell)} & \cdots & \mathbf{q}_d^{(\ell)} \\ \mathbf{0} \ (\in \mathbb{R}^{(d_p + (3L-3\ell+1)S+1) \times (d+1)}) \end{bmatrix},$$

Then, the masked self-attention block $\text{MAttn}^{(\ell)}$ constructs $\mathbf{G}^{(\ell-1)}$ by computing $\mathbf{g}_v^{(\ell)}$ from $\mathbf{q}_v^{(\ell)}$ as

$$\mathbf{G}^{(\ell-1)} = \underbrace{\mathbf{Q}^{(\ell)}}_{\text{skip connection}} + \text{MAttn}^{(\ell)}(\mathbf{Q}^{(\ell)}) = \mathbf{Q}^{(\ell)} + \begin{bmatrix} \mathbf{0} \ (\in \mathbb{R}^{((3\ell+L-1)S) \times (d+1)}) \\ \mathbf{g}_0^{(\ell)} & \mathbf{g}_1^{(\ell)} & \cdots & \mathbf{g}_d^{(\ell)} \\ \mathbf{0} \ (\in \mathbb{R}^{(d_p + (3L-3\ell+2)S+1) \times (d+1)}) \end{bmatrix}.$$

Finally, the second feed forward layer $\text{FF}_{\downarrow,2}^{(\ell)}$ constructs $\mathbf{H}_v^{(\ell-1)}$, using $\mathbf{g}_v^{(\ell)}$ and $\mathbf{q}_v^{(\ell)}$.

$$\begin{aligned} \mathbf{H}^{(\ell-1)} &= \text{normalize} \left(\underbrace{\mathbf{G}^{(\ell)}}_{\text{skip connection}} + \text{FF}_{\downarrow,2}^{(\ell)}(\mathbf{G}^{(\ell)}) \right) \\ &= \text{normalize} \left(\mathbf{G}^{(\ell)} + \begin{bmatrix} \mathbf{0} & (\in \mathbb{R}^{((3\ell+L-2)S) \times (d+1)}) \\ \star & (\in \mathbb{R}^{S \times d}) \\ \mathbf{0} & (\in \mathbb{R}^{(d_p + (3L-3\ell+3)S+1) \times (d+1)}) \end{bmatrix} \right). \end{aligned}$$

Here $\star$ means $[\mathbf{h}_0^{(\ell-1)} \ \mathbf{h}_1^{(\ell-1)} \ \cdots \ \mathbf{h}_d^{(\ell-1)}]$ before normalization.

We then consider upsampling. After we obtain $\mathbf{H}^{(0)}$, we iteratively compute $\mathbf{B}^{(\ell)}$ ($\ell =$

$1, \dots, L+1$):

$$\mathbf{B}^{(\ell)} = \begin{bmatrix} & & \mathbf{0} & \\ \mathbf{b}_0^{(\ell)} & \mathbf{b}_1^{(\ell)} & \cdots & \mathbf{b}_d^{(\ell)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{b}_0^{(1)} & \mathbf{b}_1^{(1)} & \cdots & \mathbf{b}_d^{(1)} \\ \mathbf{h}_0^{(0)} & \mathbf{h}_1^{(0)} & \cdots & \mathbf{h}_d^{(0)} \\ \mathbf{q}_0^{(1)} & \mathbf{q}_1^{(1)} & \cdots & \mathbf{q}_d^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{q}_0^{(L)} & \mathbf{q}_1^{(L)} & \cdots & \mathbf{q}_d^{(L)} \\ \mathbf{h}_0^{(L)} & \mathbf{h}_1^{(L)} & \cdots & \mathbf{h}_d^{(L)} \\ \mathbf{h}_{\text{im}, d_{\text{im}}}^{(0)} & \mathbf{h}_{\text{im}, d_{\text{im}}}^{(0)} & \cdots & \mathbf{h}_{\text{im}, d_{\text{im}}}^{(0)} \\ & \mathbf{P} & & \end{bmatrix}.$$

Here, $\mathbf{b}^{(\ell)}$ are S -dimensional real-valued vectors. The ℓ -th block of downsampling computes $\mathbf{B}^{(\ell+1)}$ using a feed forward network $\text{FF}_{\uparrow}^{(\ell)}$ with normalization:

$$\mathbf{B}^{(\ell+1)} = \text{normalize} \left(\underbrace{\mathbf{B}^{(\ell)}}_{\text{skip connection}} + \text{FF}_{\uparrow}^{(\ell)}(\mathbf{B}^{(\ell)}) \right) = \mathbf{B}^{(\ell)} + \begin{bmatrix} \mathbf{0} \ (\in \mathbb{R}^{((L-\ell)S) \times (d+1)}) \\ \mathbf{b}_0^{(\ell+1)} \ \mathbf{b}_1^{(\ell+1)} \ \cdots \ \mathbf{b}_d^{(\ell+1)} \\ \mathbf{0} \ (\in \mathbb{R}^{(d_p + (3L+\ell+1)S+1) \times (d+1)}) \end{bmatrix}.$$

For $\ell = 0$, replace $\mathbf{B}^{(0)}$ by $\mathbf{H}^{(0)}$. This is the same as (A.82) (except for difference in the column dimension). We do not need the self-attention layer and second feed forward layer, and we can ignore them by simply setting the weight matrices to zeros.

Finally, we obtain $\mathbf{b}_v^{(L+1)}$ for all $v = 1, \dots, d-1$. The readout layer read_{vIm} computes $\text{softmax}(\mathbf{b}_v^{(L+1)})$, which approximates $\mu_{\star}(x_{\text{tx}, v+1} | \mathbf{x}_{\text{im}}, x_{\text{tx}, 1}, \dots, x_{\text{tx}, v})$, for all $v = 1, \dots, d-1$.

In the following, our goal is to iteratively show that

$$\mathbf{h}_v^{(\ell)} \approx \mathbf{h}_v^{(\ell)}, \quad \mathbf{q}_v^{(\ell)} \approx \mathbf{q}_v^{(\ell)}, \quad \mathbf{g}_v^{(\ell)} \approx \mathbf{g}_v^{(\ell)}, \quad \mathbf{b}_v^{(\ell)} \approx \mathbf{b}_v^{(\ell)},$$

($\ell = L, \dots, 0$ for $\mathbf{h}_v^{(\ell)}$, $\ell = L, \dots, 1$ for $\mathbf{q}_v^{(\ell)}$ and $\mathbf{g}_v^{(\ell)}$, and $\ell = 1, \dots, L+1$ for $\mathbf{b}_v^{(\ell)}$) for all $v \in \mathcal{V}_{\text{tx}}^{(L)}$. For $v = 0$, we will iteratively fill the zero vectors for all $\mathbf{h}_0^{(\ell)}, \mathbf{q}_0^{(\ell)}, \mathbf{g}_0^{(\ell)}$, and $\mathbf{b}_0^{(\ell)}$.

We now formally define each component of the pipeline.

Encoding Emb_{vIm} . Denote the v -th column of $\mathbf{P}$ by $\mathbf{p}_v$. For $v \in \mathcal{V}_{\text{tx}}^{(L)}$, We define $\mathbf{p}_v \in \mathbb{R}^{2L+2}$ as

$$\begin{aligned} \mathbf{p}_v = & \left[0, 1, \sin\left(\frac{2\pi\iota(v)}{m^{(L)}}\right), \cos\left(\frac{2\pi\iota(v)}{m^{(L)}}\right), \sin\left(\frac{2\pi\iota(\text{pa}(v))}{m^{(L-1)}}\right), \cos\left(\frac{2\pi\iota(\text{pa}(v))}{m^{(L-1)}}\right), \right. \\ & \left. \cdots, \sin\left(\frac{2\pi\iota(\text{pa}^{L-1}(v))}{m^{(1)}}\right), \cos\left(\frac{2\pi\iota(\text{pa}^{L-1}(v))}{m^{(1)}}\right) \right]^\top. \end{aligned} \quad (\text{A.107})$$

The difference from the contrastive learning (A.55) and conditional diffusion model (A.83) is that we added 0 and 1 to the first two dimensions. For $v = 0$, we define

$$\mathbf{p}_0 = \begin{bmatrix} 1, 0, \mathbf{0} \end{bmatrix}^\top \in \mathbb{R}^{2L+2}$$

so that the first two dimensions are orthogonal to (A.107).

For two-stage training, where $\widehat{\mathbf{E}}_{\text{im}}(\mathbf{x}_{\text{im}})$ approximates $h_{\text{im},r}^{(0)} = (\log \mathbb{P}[s|\mathbf{x}_{\text{im}}])_{s \in [S]}$, we define $\mathbf{h}_{\text{im},d_{\text{im}}}^{(0)}$ as $\mathbf{h}_{\text{im},d_{\text{im}}}^{(0)} = (\log \text{trun}_{\text{im}}(\widehat{\mathbf{E}}_{\text{im}}(\mathbf{x}_{\text{im}})_s))_{s \in [S]}$.

Downsampling: position-wise feed forward layers. The first feed forward layer $\text{FF}_{\downarrow,1}^{(\ell)}$ of the ℓ -th block ($\ell = L, \dots, 1$) approximates each $f_{\downarrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}$. Therefore, the feed forward block at the ℓ th layer yields

$$\mathbf{q}_v^{(\ell)} = \mathbf{f}_{\downarrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{h}_v^{(\ell)}) \in \mathbb{R}^S, \quad v \in \mathcal{V}_{\text{tx}}^{(L)}, \quad \ell = L, \dots, 1.$$

When $\mathbf{h}_v^{(\ell)} \approx h_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}$ for $v \in \mathcal{V}^{(L)}$ and $\mathbf{f}_\iota^{(\ell)} \approx f_\iota^{(\ell)}$, we have $\mathbf{q}_v^{(\ell)} \approx q_{\text{pa}^{(L-\ell)}(v)}^{(\ell)}$. Following the notation in Definition A.2, we state the following approximation error guarantee.

Lemma A.30 (Approximation error of the first feed forward layer). *Fix $\ell \in [L]$ and $\delta > 0$. Assume that $B_\psi^{-1} \leq \psi_\iota^{(\ell)}(s, a) \leq B_\psi$ for all $s, a \in [S]$. When $\ell = 1$, also assume that $B_\psi^{-1} \leq \mathbb{P}[s] \leq B_\psi$ for all s . Then, there exists an $\text{NN} \in \mathcal{F}(J, \mathbf{j}, B)$ such that*

$$\|\text{NN}([h; \mathbf{p}_v]) - f_{\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(h)\|_\infty \leq \delta, \quad v \in \mathcal{V}^{(L)},$$

for all $h \in \mathbb{R}^S$ with $\max_s h_s = 0$ ($\ell \leq L-1$) or $h \in [S]$ ($\ell = L$). The network parameters $J, \mathbf{j}$ and B are bounded as follows:

$$\begin{aligned} J &\lesssim (\log \log(SB_\psi/\delta)) \log(SB_\psi/\delta), \quad \|\mathbf{j}\|_\infty \lesssim m^{(\ell)} S (\log(SB_\psi/\delta))^3, \\ B &\lesssim 2S(B_\psi^2 + \log(SB_\psi/\delta)) + (m^{(\ell)})^2. \end{aligned}$$

The only difference from Lemma A.9 is the dimension of $\mathbf{p}_v$, and thus we omit the proof. We then consider the second feed forward layer $\text{FF}_{\downarrow,2}^{(\ell)}$. The role of this layer is to compute

$$a_v^{(\ell)} \mathbf{g}_v^{(\ell)} + \mathbf{q}_v^{(\ell)} - \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}] \mathbf{q}_v^{(\ell)},$$

and the following lemma shows that this computation can be done exactly.

Lemma A.31 (Approximation error of the second feed forward layer). *Fix $\ell \in [L]$. There exists an $\text{NN} \in \mathcal{F}(J, \mathbf{j}, B)$ such that*

$$\text{NN}([g, q; \mathbf{p}_v]) = a_v^{(\ell)} g + q - \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}] q, \quad v \in \mathcal{V}^{(L)},$$

for all $g, q \in \mathbb{R}^S$ with $\|g\|_\infty, \|q\|_\infty \leq C$. The network parameters $J, \mathbf{j}$ and B are bounded as follows:

$$J_1 \lesssim 1, \quad \|\mathbf{j}_1\|_\infty \lesssim S + d_p, \quad B_1 \lesssim L + \max_{\ell+1 \leq k \leq L} (m^{(k)})^2 + m^{(\ell)} C.$$

The proof of this lemma is found in Appendix A.7.4.

Downsampling: masked self-attention layer. To obtain $\mathbf{G}^{(\ell-1)}$ from $Q^{(\ell)}$, we use the causal mask and multi-head attention. Let k be a sequence length. The causal mask $\mathbf{M}_k$ is defined as

$$\mathbf{M}_k = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ -C & 0 & 0 & \cdots & 0 \\ -C & -C & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -C & -C & -C & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{k \times k},$$

where C is a sufficiently large constant (so that (i, j) -element of $\text{softmax}(\mathbf{M}_k + ((\mathbf{W}_{K,m} \cdot)^\top (\mathbf{W}_{Q,m} \cdot)))$ is ignored in the following for $i > j$.) Then a masked self-attention layer is defined as

$$\text{Mattn}(\cdot) = (\mathbf{W}_V \cdot) \text{softmax}(\mathbf{M}_k + ((\mathbf{W}_K \cdot)^\top (\mathbf{W}_Q \cdot))),$$

where $\mathbf{M}_k$ is added in an element-wise manner and softmax is applied column-wise.

Definition A.6 (A class of masked multi-head self-attention blocks). We define a class of masked self-attention blocks with as

$$\bar{\mathcal{A}}(D, B) = \left\{ (\mathbf{W}_V \cdot) \text{softmax}(\mathbf{M}_d + ((\mathbf{W}_K \cdot)^\top (\mathbf{W}_Q \cdot))) \mid \right. \\ \left. \mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_V \in \mathbb{R}^{d \times d}, \max_{i,j} |(\mathbf{W}_K)_{i,j}|, \max_{i,j} |(\mathbf{W}_Q)_{i,j}|, \max_{i,j} |(\mathbf{W}_V)_{i,j}| \leq B \right\}.$$

We will construct weight matrices so that the self-attention layer $\text{Mattn}^{(\ell)}$ of the ℓ -th block ($\ell = L, L-1, \dots, 1$) yields

$$\mathbf{g}_v^{(\ell-1)} = \frac{1}{a_v^{(\ell)}} \sum_{u^{(\ell)} \in \mathcal{C}(\text{pa}^{(L-\ell+1)}(v))} \mathbb{1}[u \leq v] \mathbf{q}_u^{(\ell)} + \boldsymbol{\delta}_v^{(\ell)} \in \mathbb{R}^S$$

with $\|\boldsymbol{\delta}_v^{(\ell)}\|_\infty \ll 1$.

The approximation error guarantee is stated as follows.

Lemma A.32 (Approximation error of self-attention layer). *For $\ell \in [L]$, there exists $\text{Mattn} \in \bar{\mathcal{A}}(D, B)$ with $D = d_f + d_p$ and $B \lesssim \log(d\delta^{-1}) + m^{(\ell)}$ such that*

$$\text{Mattn}(Q^{(\ell)})_v = \begin{cases} \begin{bmatrix} \mathbf{0} & (\in \mathbb{R}^{(3\ell+L-1)S}) \\ \frac{1}{a_v^{(\ell)}} \left(\mathbf{q}_0^{(\ell)} + \sum_{u^{(\ell)} \in \mathcal{C}(\text{pa}^{(L-\ell+1)}(v))} \mathbb{1}[u \leq v] \mathbf{q}_u^{(\ell)} \right) + \boldsymbol{\delta}_v^{(\ell)} & (\in \mathbb{R}^S) \\ \mathbf{0} & (\in \mathbb{R}^{d_p + (3L-3\ell+2)S+1}) \end{bmatrix}, & (v \in \mathcal{V}_{\text{tx}}^{(L)}) \\ \begin{bmatrix} \mathbf{0} & (\in \mathbb{R}^{(3\ell+L-1)S}) \\ \mathbf{q}_0^{(\ell)} & (\in \mathbb{R}^S) \\ \mathbf{0} & (\in \mathbb{R}^{d_p + (3L-3\ell+2)S+1}) \end{bmatrix}, & (v = 0) \end{cases}$$

where $\boldsymbol{\delta}_v^{(\ell)} \in \mathbb{R}^S$ satisfies $\|\boldsymbol{\delta}_v^{(\ell)}\|_\infty \leq \delta \max_{v'} \|\mathbf{q}_{v'}^{(\ell)}\|_\infty$.

Because we can iteratively see that $\mathbf{q}_0^{(\ell)} = 0$, the column corresponding to $v \in \mathcal{V}_{\text{tx}}^{(L)}$ is $\frac{1}{a_v^{(\ell)}} \sum_{u^{(\ell)} \in \mathcal{N}(\text{pa}^{(L-\ell)}(v))} \mathbb{1}[u \leq v] \mathbf{q}_u^{(\ell)} + \boldsymbol{\delta}_v^{(\ell)}$, and the column corresponding to $v = 0$ is exactly $\mathbf{0} \in \mathbb{R}^D$. When $\mathbf{q}_v^{(\ell)} \approx q_v^{(\ell)}$ and $\|\boldsymbol{\delta}_v^{(\ell)}\|_\infty \ll 1$, we have $\mathbf{g}_v^{(\ell)} \approx g_v^{(\ell)}$. The proof of this lemma can be found in Appendix A.7.5.

Upsampling: position-wise feed forward block. The upsampling constructs estimation of leaf nodes starting from the root. The $(L+1)$ blocks of attention blocks (feed forward layers) can implement this process. The ℓ -th block of upsampling computes $\mathbf{b}_v^{(\ell+1)}$ from $\mathbf{h}_v^{(\ell)}$, $\mathbf{q}_v^{(\ell+1)}$, and $\mathbf{b}_v^{(\ell)}$.

$$\begin{aligned} \mathbf{b}_v^{(1)} &= \text{normalize}(\mathbf{h}_v^{(0)} + \mathbf{h}_{\text{im}, d_{\text{im}}}^{(0)} - \mathbf{q}_v^{(1)}) \in \mathbb{R}^S, \\ \mathbf{b}_v^{(\ell+1)} &= \begin{cases} \text{normalize}(\mathbf{f}_{\uparrow, \ell(\text{pa}^{L-\ell}(v+1))}^{(\ell)}(\mathbf{b}_v^{(\ell)}) + \mathbf{h}_v^{(\ell)} - \mathbf{q}_v^{(\ell+1)}), & \text{(if } \text{pa}^{(L-\ell-1)}(v) = \text{pa}^{(L-\ell-1)}(v+1)) \\ \text{normalize}(\mathbf{f}_{\uparrow, \ell(\text{pa}^{L-\ell}(v+1))}^{(\ell)}(\mathbf{b}_v^{(\ell)}) + \mathbf{h}_v^{(\ell)}), & \left(\begin{array}{l} \text{if } \text{pa}^{(L-\ell)}(v) = \text{pa}^{(L-\ell)}(v+1) \\ \text{but } \text{pa}^{(L-\ell-1)}(v) \neq \text{pa}^{(L-\ell-1)}(v+1) \end{array} \right) \\ \text{normalize}(\mathbf{f}_{\uparrow, \ell(\text{pa}^{(L-\ell)}(v+1))}^{(\ell)}(\mathbf{b}_v^{(\ell)})), & \text{(otherwise)} \end{cases} \\ &\quad \ell = 1, 2, \dots, L. \end{aligned}$$

For each update, we can track the correspondence with the message passing algorithm.

Lemma A.33 (Approximation error of feed forward layer, upsampling). *Fix $\ell \in \{0, \dots, L\}$ and $\delta > 0$. Assume that $B_\psi^{-1} \leq \psi_\ell^{(\ell)}(s, a) \leq B_\psi$ for all $s, a \in [S]$. Then, there exist $\text{NN}_1, \text{NN}_2 \in \mathcal{F}(J, \mathbf{j}, B)$ such that*

$$\begin{aligned} \text{NN}_1([h; h'; q]) &= h + h' - q, \\ \begin{cases} \|\text{NN}_2([b; h; q]) - (\mathbf{f}_{\uparrow, \ell(\text{pa}^{L-\ell}(v+1))}^{(\ell)}(\mathbf{b}_v^{(\ell)}) + \mathbf{h}_v^{(\ell)} - \mathbf{q}_v^{(\ell+1)})\|_\infty \leq \delta, & \text{if } \text{pa}^{(L-\ell-1)}(v) = \text{pa}^{(L-\ell-1)}(v+1) \\ \|\text{NN}_2([b; h; q]) - (\mathbf{f}_{\uparrow, \ell(\text{pa}^{L-\ell}(v+1))}^{(\ell)}(\mathbf{b}_v^{(\ell)}) + \mathbf{h}_v^{(\ell)})\|_\infty \leq \delta, & \left(\begin{array}{l} \text{if } \text{pa}^{(L-\ell)}(v) = \text{pa}^{(L-\ell)}(v+1) \\ \text{but } \text{pa}^{(L-\ell-1)}(v) \neq \text{pa}^{(L-\ell-1)}(v+1) \end{array} \right) \\ \|\text{NN}_2([b; h; q]) - (\mathbf{f}_{\uparrow, \ell(\text{pa}^{(L-\ell)}(v+1))}^{(\ell)}(\mathbf{b}_v^{(\ell)}))\|_\infty \leq \delta, & \text{(otherwise)} \end{cases} \\ &\quad \ell = 1, 2, \dots, L. \end{aligned}$$

for all $h, h', q, b \in \mathbb{R}^S$ with $\max_s b_s = 0$. For all of these networks, the parameters $J, \mathbf{j}$ and B are bounded as follows:

$$\begin{aligned} J &\lesssim (\log \log(SB_\psi/\delta)) \log(SB_\psi/\delta), \\ \|\mathbf{j}\|_\infty &\lesssim m^{(\ell)} S (\log(SB_\psi/\delta))^3 + d_p, \\ B &\lesssim S(B_\psi^2 + \log(SB_\psi/\delta)) + \max_{\ell \leq k \leq L} (m^{(k)})^2 + L + C. \end{aligned}$$

The proof will be placed in Appendix A.7.4.

Normalization. Since each column vector of $\mathbf{H}^{(\ell)}$, $\mathbf{Q}^{(\ell)}$, and $\mathbf{B}^{(\ell)}$ is a collection of multiple $\mathbf{h}_v^{(\ell)}$, $\mathbf{q}_v^{(\ell)}$, and $\mathbf{b}_v^{(\ell)}$, we adopt a slightly different definition of normalize than that used in message passing. Specifically, for $\mathbf{x} = [\mathbf{b}^{(L+1)} \dots \mathbf{b}^{(1)} \mathbf{h}^{(0)} \mathbf{g}^{(0)} \mathbf{q}^{(1)} \mathbf{h}^{(1)} \dots \mathbf{g}^{(L)} \mathbf{q}^{(L)} \mathbf{h}^{(L)} \mathbf{h} \mathbf{p}] \in \mathbb{R}^{d_f+d_p}$ with $\mathbf{h}^{(L)} \in [S]$, $\mathbf{h}^{(\ell)} \in \mathbb{R}^S$ ($\ell = L-1, \dots, 0$), $\mathbf{q}^{(\ell)}, \mathbf{g}^{(\ell)}, \mathbf{b}^{(\ell)} \in \mathbb{R}^S$, we define we define

$$\text{normalize}(\mathbf{x}) = \begin{bmatrix} \mathbf{b}^{(L+1)} - \mathbf{1}_S \max_{s \in S} \mathbf{q}_s^{(L+1)} \\ \vdots \\ \mathbf{b}^{(1)} - \mathbf{1}_S \max_{s \in S} \mathbf{q}_s^{(1)} \\ \mathbf{h}^{(0)} \\ \mathbf{g}^{(1)} - \mathbf{1}_S \max_{s \in S} \mathbf{g}_s^{(1)} \\ \mathbf{q}^{(1)} \\ \mathbf{h}^{(1)} \\ \vdots \\ \mathbf{g}^{(L)} - \mathbf{1}_S \max_{s \in S} \mathbf{g}_s^{(L)} \\ \mathbf{q}^{(L)} \\ \mathbf{h}^{(L)} \\ \mathbf{h} \\ \mathbf{p} \end{bmatrix} \in \mathbb{R}^{d_f+d_p}, \quad \mathbf{1}_S = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^S.$$

For a matrix in $\mathbb{R}^{(d_f+d_p) \times (d+1)}$, it is applied in a column-wise manner.

Readout layer read_{vlm} Finally, the readout layer read_{vlm} extracts $\mathbf{b}_v^{(L+1)}$ and apply an element-wise projection onto $[-B_{\text{read}}^{\text{vlm}}, B_{\text{read}}^{\text{vlm}}]$ and softmax.

$$[\text{softmax}(\text{proj}_{[-B_{\text{read}}^{\text{vlm}}, B_{\text{read}}^{\text{vlm}}]}^S(\mathbf{b}_1^{(L+1)})) \dots \text{softmax}(\text{proj}_{[-B_{\text{read}}^{\text{vlm}}, B_{\text{read}}^{\text{vlm}}]}^S(\mathbf{b}_d^{(L+1)}))], \quad (\text{A.108})$$

where $\text{proj}_{[-B_{\text{read}}^{\text{vlm}}, B_{\text{read}}^{\text{vlm}}]}^S(x) = \text{argmin}_{y \in [-B_{\text{read}}^{\text{vlm}}, B_{\text{read}}^{\text{vlm}}]} |x - y|$. According to Lemma A.36, setting $B_{\text{read}}^{\text{vlm}} := 2 \log(SB_\psi)$ allows us to ignore the effect of this truncation.

The whole pipeline. Putting it all together, the neural network approximate the message passing algorithm in the following way. The downsampling process is approximated as

$$\begin{aligned} \mathbf{q}_v^{(\ell)} &= \mathbf{f}_{\downarrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{h}_v^{(\ell)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(L)}, \ell = L, \dots, 1 \\ \mathbf{g}_v^{(\ell)} &= \frac{1}{a_v^{(\ell)}} \sum_{v' \in \mathcal{C}(\text{pa}^{(L-\ell+1)}(v))} \mathbb{1}[v' \leq v] \mathbf{q}_{v'}^{(\ell)} \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(L)}, \ell = L, \dots, 1, \\ \mathbf{h}_v^{(\ell-1)} &= \text{normalize}(a_v^{(\ell)} \mathbf{g}_v^{(\ell)} + \mathbf{q}_v^{(\ell)} - \mathbb{1}[v \in \mathcal{V}_{\text{tx}}^{(\ell)}] \mathbf{q}_v^{(\ell)}) \in \mathbb{R}^S, & v \in \mathcal{V}_{\text{tx}}^{(L)}, \ell = L, \dots, 1. \end{aligned} \quad (\text{A.109})$$

Let $\mathbf{h}_{\text{tx},d}^{(0)} \approx h_{\text{tx},r}^{(0)} = (\log \mathbb{P}[s|\mathbf{x}_{\text{im}}])_{s \in [S]}$. The upsampling process is approximated as

$$\begin{aligned} \mathbf{b}_v^{(1)} &= \text{normalize}(\mathbf{h}_v^{(0)} + \mathbf{h}_{\text{tx},d}^{(0)} - \mathbf{q}_v^{(1)}) \in \mathbb{R}^S, \\ \mathbf{b}_v^{(\ell+1)} &= \begin{cases} \text{normalize}(\mathbf{f}_{\uparrow, \ell(\text{pa}^{L-\ell}(v+1))}^{(\ell)}(\mathbf{b}_v^{(\ell)} + \mathbf{h}_v^{(\ell)} - \mathbf{q}_v^{(\ell+1)})), & \text{(if } \text{pa}^{(L-\ell-1)}(v) = \text{pa}^{(L-\ell-1)}(v+1)) \\ \text{normalize}(\mathbf{f}_{\uparrow, \ell(\text{pa}^{L-\ell}(v+1))}^{(\ell)}(\mathbf{b}_v^{(\ell)} + \mathbf{h}_v^{(\ell)})), & \left(\begin{array}{l} \text{if } \text{pa}^{(L-\ell)}(v) = \text{pa}^{(L-\ell)}(v+1) \\ \text{but } \text{pa}^{(L-\ell-1)}(v) \neq \text{pa}^{(L-\ell-1)}(v+1) \end{array} \right) \\ \text{normalize}(\mathbf{f}_{\uparrow, \ell(\text{pa}^{L-\ell}(v+1))}^{(\ell)}(\mathbf{b}_v^{(\ell)})), & \text{(otherwise)} \end{cases} \\ &\qquad\qquad\qquad \ell = 1, 2, \dots, L. \end{aligned} \tag{A.110}$$

We summarize the network architecture (which slightly differs from the previous one) and the hypothesis class of $(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}, \text{Adap})$ as follows. We focus on two step training, and the definition of the parameter space $\Theta_{L,J,D,D',B}^{\text{vlm}}$ for joint training is introduced in Appendix A.6.3.

Definition A.7 (Eq. (A.8), restated). The image transformer network TF_{vlm} has L blocks of feed forward (Definition A.2) with skip connection, masked self-attention (Definition A.6) with skip connection, feed forward (Definition A.2) with skip connection, and normalization in this order. We say the collection of the parameters of $(\text{TF}_{\text{vlm}}, \text{Adap})$ belongs to $\Theta_{L,J,D,D',B,M}$ if the following holds: In each block of the text transformer TF_{vlm} , its two feed forward layers FF_1, FF_2 and self-attention MAttn satisfy

$$\text{FF} \in \mathcal{F}(J, \mathbf{j} = (D, *, \dots, *, D), B), \text{ with } \|\mathbf{j}\|_{\infty} \leq D', \quad \text{MAttn} \in \bar{\mathcal{A}}(D, B).$$

Furthermore, the adapter satisfies

$$W_{\text{ada}}^{(1)} \in \mathbb{R}^{S \times M}, \quad W_{\text{ada}}^{(2)} \in \mathbb{R}^{M \times S}, \quad \|W_{\text{ada}}^{(1)}\|_{\text{op}} \leq B, \quad \|W_{\text{ada}}^{(2)}\|_{\text{op}} \leq B.$$

The rest of this section is organized as follows. Appendix A.7.2 discusses the two step training and proves Theorem A.2, using Lemmas A.30 to A.33, as well as the bound on the propagation of the intermediate errors Lemma A.35. Appendix A.7.3 discusses the joint training using these lemmas as well. Lemmas A.30, A.31 and A.33 are proved in Appendix A.7.4, and Lemma A.32 is proved in Appendix A.7.4. Appendix A.7.6 gives the error propagation lemma (Lemma A.35). Appendix A.7.7 gives the proof of Proposition A.29. Finally, Appendix A.7.8 gives useful property on the message passing algorithm.

A.7.2 Proof of Theorem A.2

Similar to the proof of Theorem 2.12, define

$$\bar{\mathbf{R}}_{\text{vlm}}^* := \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mu_*} \left[\sum_{j \in [d_{\text{tx}}]} -\log \mu_*(x_{\text{tx},j} | x_{\text{tx},1:j-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right].$$

Then we have the decomposition

$$\begin{aligned} D(\mu_\star, \mu^{\hat{\theta}}) &= R_{\text{vIm}}(\mu^{\hat{\theta}}, E_{\text{im}}) - \bar{R}_{\text{vIm}}^\star \\ &= \underbrace{\inf_{\theta \in \Theta_{L,J,D,D',B,M}} R_{\text{vIm}}(\mu^\theta, E_{\text{im}}) - \bar{R}_{\text{vIm}}^\star}_{\text{approximation error}} + \underbrace{R_{\text{vIm}}(\mu^{\hat{\theta}}, E_{\text{im}}) - \inf_{\theta \in \Theta_{L,J,D,D',B,M}} R_{\text{vIm}}(\mu^\theta, E_{\text{im}})}_{\text{generalization error}}. \end{aligned}$$

We state the following bounds on the approximation and generalization error, with proofs to follow shortly.

- (a). If we choose $J = \tilde{\mathcal{O}}(L)$, $D = \mathcal{O}(SL)$, $D' = \tilde{\mathcal{O}}(\bar{m}SL^3)$, and $B = \tilde{\mathcal{O}}(L_B + (SL + \bar{m}^2)\sqrt{M})$, then the approximation error

$$\begin{aligned} &\inf_{\theta \in \Theta_{L,J,D,D',B,M}} R_{\text{vIm}}(\mu^\theta, E_{\text{im}}) - \bar{R}_{\text{vIm}}^\star \\ &\leq d_{\text{tx}} \cdot \tilde{\mathcal{O}} \left(\sqrt{\frac{(SL^8\bar{m}^2 + M)SL^3}{n}} + \sqrt{S^5 \cdot L_B^2 \cdot \left(\text{Suff}(\mathbf{S}) + \frac{1}{M} \right)} \right). \end{aligned} \quad (\text{A.111})$$

- (b). Under the same choice of model class $\Theta_{L,J,D,D',B,M}$, the generalization error

$$R_{\text{vIm}}(\mu^{\hat{\theta}}, E_{\text{im}}) - \inf_{\theta \in \Theta_{L,J,D,D',B,M}} R_{\text{vIm}}(\mu^\theta, E_{\text{im}}) \leq \tilde{\mathcal{O}} \left(d_{\text{tx}} \cdot \sqrt{\frac{(SL^8\bar{m}^2 + M)SL^3}{n}} \right)$$

with probability at least $1 - 1/n$.

Combining the claims yields Theorem A.2.

(a) Approximation error. Take some $\delta' > 0$ which will be defined later. For the feed forward layers, we use Lemmas A.30, A.31 and A.33 with $\delta = \delta' \ll 1$. For the self-attention layers at the ℓ -th step of the downsampling, we use Lemma A.32 with $\delta = \frac{\delta'}{\max_v a_v^{(\ell)} \|\mathbf{q}_v^{(\ell)}\|_\infty}$. Following the argument in the proof of Theorem 2.10, $\mathbf{q}_v^{(\ell)} = f_{\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{h}_v^{(\ell)})$ and $\mathbf{g}_v^{(\ell)}$ are bounded by $3(1 \vee \log SB_\psi)$ for each ℓ . Thus C in Lemma A.31 and δ in Lemma A.23 are bounded by $3(1 \vee \log SB_\psi)$ and $\frac{\delta'}{3(m_{\text{tx}}^{(\ell)} + 1)(1 \vee \log SB_\psi)} \leq \frac{\delta'}{(\bar{m} + 1)(1 \vee \log SB_\psi)}$, respectively.

Furthermore, Lemmas A.35 and A.36 yield that

$$\begin{aligned} &\max_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}, i)} |\log \mu_\star(x_{\text{tx},i} | x_{\text{tx},1:i-1}, \mathbf{x}_{\text{im}}) - \log \mu^{\hat{\theta}}(x_{\text{tx},i} | x_{\text{tx},1:i-1}, E_{\text{im}}(\mathbf{x}_{\text{im}}))| \\ &\leq SB_\psi \left(8^{L+1} \delta' \prod_{1 \leq k \leq L} (2m_{\text{tx}}^{(k)} + 5) + \delta_{\text{im}} \right) \\ &\leq 40^{L+1} d_{\text{tx}} SB_\psi \delta' + SB_\psi \delta_{\text{tx}}. \end{aligned}$$

We choose

$$\delta' = \frac{\sqrt{(SL^8\bar{m}^2 + M)L^3}}{40^{L+1}d_{\text{tx}}B_\psi\sqrt{Sn}}.$$

Moreover, similar to the proof of Theorem 2.12, from Proposition 2.9 and

$$\delta_{\text{im}} = \|\log(\text{trun}_{\text{im}}(\widehat{\mathbf{E}}_{\text{im}}(\mathbf{x}_{\text{im}})) - (\log \mathbb{P}[s|\mathbf{x}_{\text{im}}])_{s \in [S]}\|_\infty$$

and Lemma A.19, it can be verified that there exists some $\text{Adap}(\cdot)$ in Eq. (2.20) such that, $\|W_{\text{ada}}^{(1)}\|_{\text{op}} \leq C'L_B$, $\|W_{\text{ada}}^{(2)}\|_{\text{op}} \leq C'(SL + \bar{m}^2)\sqrt{M}$, and

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_{\text{im}}} \delta_{\text{im}}^2 &\leq \mathbb{E}_{\mathbf{x}_{\text{im}}} \|\log \text{trun}_{\text{im}}(\widehat{\mathbf{E}}_{\text{im}}(\mathbf{x}_{\text{im}})) - \log \mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}})\|_2^2 \leq CS^2 \cdot \mathbb{E}_{\mathbf{x}_{\text{im}}} \|\widehat{\mathbf{E}}_{\text{im}}(\mathbf{x}_{\text{im}}) - \mathbf{E}_{\text{im},\star}(\mathbf{x}_{\text{im}})\|_2^2 \\ &\leq CS^2 \cdot L_B^2 \cdot L_\Gamma^2 \cdot p_\star \cdot (\text{Suff}(\mathbf{S}) + M^{-1}) \leq CS^3 \cdot L_B^2 \cdot (\text{Suff}(\mathbf{S}) + M^{-1}) \end{aligned}$$

for some $C, C' > 0$ depending polynomially on B_ψ^m . Putting pieces together, according to Lemmas A.30 to A.33, we find that there exist some $\boldsymbol{\theta} \in \Theta_{L,J,D,D',B,M}$ that yields Eq. (A.111), where

$$\begin{aligned} D &\leq d_f + d_p = (4S + 2)L + 1 = \mathcal{O}(SL), \\ J &\lesssim (\log \log(SK/\delta')) \log(SK/\delta') = \tilde{\mathcal{O}}(L), \\ D' &= \|\mathbf{j}\|_\infty \lesssim \bar{m}S(\log(SK/\delta'))^3 + d_f + d_p = \tilde{\mathcal{O}}(\bar{m}SL^3), \\ B &\lesssim S(B_\psi^2 + \log(SB_\psi/\delta')) + \bar{m}^2 \log(SB_\psi) + L \\ &\quad + \log \frac{d \log(SB_\psi)}{\delta'} + (L_B + (SL + \bar{m}^2)\sqrt{M}) \\ &= \tilde{\mathcal{O}}(L_B + (SL + \bar{m}^2)\sqrt{M}). \end{aligned}$$

Here we recall $\bar{m} = \max\{\max_k m_{\text{tx}}^{(k)}, \max_k m_{\text{im}}^{(k)}\}$.

(b) generalization error. Since $\mu^{\hat{\boldsymbol{\theta}}}$ is the minimizer of $\widehat{\mathbf{R}}_{\text{vIm},t}(\mu^{\boldsymbol{\theta}}, \mathbf{E}_{\text{im}})$ defined in Eq. (A.7), we have

$$\bar{\mathbf{R}}_{\text{vIm}}(\mu^{\hat{\boldsymbol{\theta}}}, \mathbf{E}_{\text{im}}) - \inf_{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B,M}} \bar{\mathbf{R}}_{\text{vIm},t}(\mu^{\boldsymbol{\theta}}, \mathbf{E}_{\text{im}}) \leq 2 \sup_{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B,M}} |\widehat{\mathbf{R}}_{\text{vIm}}(\mu^{\boldsymbol{\theta}}, \mathbf{E}_{\text{im}}) - \bar{\mathbf{R}}_{\text{vIm},t}(\mu^{\boldsymbol{\theta}}, \mathbf{E}_{\text{im}})|. \quad (\text{A.112})$$

Similar to the proof of Theorem 2.10 and 2.12, we verify the conditions for Lemma A.51 and then apply the lemma to derive an upper bound for the R.H.S. of Eq. (A.112).

In Lemma A.51, take $\Theta = \Theta_{L,J,D,D',B,M}$, $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') = \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$, $z_i = (\mathbf{x}_{\text{im}}^{(i)}, \mathbf{x}_{\text{tx}}^{(i)})$, and

$$f(z_i; \boldsymbol{\theta}) = -\frac{1}{d_{\text{tx}}} \sum_{j \in [d_{\text{tx}}]} \log \mu^{\boldsymbol{\theta}}(x_{\text{tx},j} | x_{\text{tx},1:j-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})).$$

Verification of condition (a) in Lemma A.51. We note that the set $\Theta_{L,J,D,D',B,M}$ with metric $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') = \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ has a diameter $B_\rho := 2\bar{B}$. Furthermore, the dimension of $\Theta_{L,J,D,D',B,M}$ is bounded by $d_\rho := (2J+3)(2L+1)(D+D'+1)^2 + S + 2SM = \tilde{\mathcal{O}}(S^2 L^8 \bar{m}^2 + 2SM)$. Thus, by Example 5.8 in Wainwright (2019b), we have $\log \mathcal{N}(\Delta; \Theta_{L,J,D,D',B,M}, \|\cdot\|) \leq d_\rho \log(1 + 2r/\Delta) \leq d_\rho \log(2A_\rho r/\Delta)$ for $\Delta \in (0, 2r]$ with $A_\rho = 2$.

Verification of condition (b) in Lemma A.51. Since $f(z_i; \boldsymbol{\theta})$ is $B_{\text{read}}^{\text{vlm}}$ -bounded by the definition of read_{vlm} in Eq. (A.108), it follows that $f(z_i; \boldsymbol{\theta}) - \mathbb{E}[f(z_i; \boldsymbol{\theta})]$ is $\sigma = cB_{\text{read}}^{\text{vlm}}$ -sub-Gaussian for all $\boldsymbol{\theta} \in \Theta_{L,J,D,D',B,M}$ for some numerical constant $c > 0$.

Verification of condition (c) in Lemma A.51. By Lemma A.43 and the boundedness condition, we have

$$\begin{aligned} & |f(z_i; \boldsymbol{\theta}) - f(z_i; \boldsymbol{\theta}')| \\ & \leq \frac{1}{d_{\text{tx}}} \sum_{j \in [d_{\text{tx}}]} |\log \mu^\boldsymbol{\theta}(x_{\text{tx},j} | x_{\text{tx},1:j-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) - \log \mu^{\boldsymbol{\theta}'}(x_{\text{tx},j} | x_{\text{tx},1:j-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))| \\ & \leq B_f \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \end{aligned}$$

where $B_f := ((cB)^{18JL} S^4 B_{\text{read}}^3)^{4L+3} B_\psi^{2m}$ and $B_{\text{read}} = 4\bar{m} \log B_\psi$. Therefore, we may choose $\sigma' = B_f$ and condition (c) is hence satisfied.

Now, invoking Lemma A.51 and plugging in the values of $d_\rho, \sigma, \sigma', A_\rho, B_\rho$, we find

$$\begin{aligned} & \sup_{\boldsymbol{\theta} \in \Theta_{L,J,D,D',B,M}} |\widehat{\mathbf{R}}_{\text{cdm},t}(\mathbf{M}_t^\boldsymbol{\theta}, \mathbf{E}_{\text{tx}}) - \bar{\mathbf{R}}_{\text{cdm},t}(\mathbf{M}_t^\boldsymbol{\theta}, \mathbf{E}_{\text{tx}})| \\ & \leq d_{\text{tx}} \cdot c\sigma \sqrt{\frac{d_\rho \log(2A_\rho(1 + B_\rho \sigma'/\sigma)) + \log(1/\eta)}{n}} \\ & \leq \tilde{\mathcal{O}}\left(d_{\text{tx}} \cdot \sqrt{\frac{(SL^8 \bar{m}^2 + M)SL^3 + \log(1/\eta)}{n}}\right) \end{aligned}$$

with probability at least $1 - \eta$. Setting $\eta = 1/n$ completes the proof.

A.7.3 Joint training of the vision-language model and the image representation

Similar to Appendix A.6.3, in this section, we consider jointly learning the vision-language models (VLMs) and the image representation within the JGHM framework. Following the setup of Section A.3.2, suppose we are given a dataset of iid samples $\{(\mathbf{x}_{\text{im}}^{(i)}, \mathbf{x}_{\text{tx}}^{(i)})\}_{i \in [n]} \sim_{\text{iid}} \mu_\star$.

The next word predictors

$$\mu^\theta(\cdot | x_{\text{tx},1:j-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) = \text{read}_{\text{vlm}} \circ \text{TF}_{\text{vlm}} \circ \text{Emb}_{\text{vlm}}(x_{\text{tx},1:j-1}, \text{Adap}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))),$$

for $i \in [d_{\text{tx}}]$, where $\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{tx}}) = \text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}(\mathbf{x}_{\text{im}})$ as defined in Section 2.4.1, and the remaining components are the same as defined in Section A.3.2, except that in the embedding Emb_{vlm} , we let

$$\mathbf{h}_{\text{im},d}^{(0)} = \widetilde{\text{trun}}_{\text{im}}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})), \quad \text{where } \widetilde{\text{trun}}_{\text{im}}(z) := \text{proj}_{[-B_{\text{read}}^{\text{vlm}}, B_{\text{read}}^{\text{vlm}}]}(z),$$

in contrast to Eq. (A.108). We solve the empirical risk minimization

$$\hat{\theta} = \arg \min_{\theta \in \Theta_{L,J,D,D',B}^{\text{vlm}}} \left\{ \hat{R}_{\text{vlm}}(\mu^\theta, \mathbf{E}_{\text{im}}) := \frac{1}{n} \sum_{i=1}^n \left[\sum_{j \in [d_{\text{tx}}]} -\log \mu^\theta(x_{\text{tx},j} | x_{\text{tx},1:j-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right] \right\}, \quad (\text{A.113})$$

where the parameter space is defined as

$$\begin{aligned} \Theta_{L,J,D,D',B}^{\text{vlm}} &:= \left\{ \mathbf{W}_{\text{vlm}}, \mathbf{W}_{\text{im}} \text{ as defined in Eq. (2.16)}; \right. \\ \|\theta\| &:= \max_{i \in [2J+2], \ell \in [2L+1]} \{ \|W_{i,\text{vlm}}^{(\ell)}\|_{\text{op}}, \|W_{Q,\text{vlm}}^{(\ell)}\|_{\text{op}}, \|W_{K,\text{vlm}}^{(\ell)}\|_{\text{op}}, \|W_{V,\text{vlm}}^{(\ell)}\|_{\text{op}} \} \\ &\quad \vee \max_{i \in [J+1], \ell \in [L]} \{ \|W_{i,\text{im}}^{(\ell)}\|_{\text{op}}, \|W_{Q,\text{im}}^{(\ell)}\|_{\text{op}}, \|W_{K,\text{im}}^{(\ell)}\|_{\text{op}}, \|W_{V,\text{im}}^{(\ell)}\|_{\text{op}} \} \leq B \} \end{aligned} \quad (\text{A.114})$$

Similar to Theorem A.2 and Theorem A.25, we state the following result without providing a formal proof.

Theorem A.34 (Sampling error of the conditional next-token predictors, joint training). *Suppose that Assumption 2.4 and Assumption 2.5 hold. Let $\Theta_{L,J,D,D',B}^{\text{vlm}}$ be the set defined in Eq. (A.114), where $J = \tilde{\mathcal{O}}(L)$, $D = \mathcal{O}(SL)$, $D' = \tilde{\mathcal{O}}(\overline{m}SL^3)$, and $B = \tilde{\mathcal{O}}(SL + \overline{m}^2)$. Let $\hat{\theta}$ be the empirical risk minimizer defined in Eq. (A.113). Then, with probability at least $1 - 1/n$, we have*

$$\begin{aligned} D(\mu_\star, \mu^{\hat{\theta}}) &:= \mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}} \left[\sum_{i \in [d_{\text{tx}}]} D_{\text{KL}} \left(\mu_\star(x_{\text{tx},i} | x_{\text{tx},1:i-1}, \mathbf{x}_{\text{im}}) \middle| \middle| \mu^{\hat{\theta}}(x_{\text{tx},i} | x_{\text{tx},1:i-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) \right) \right] \\ &\leq d_{\text{tx}} \cdot \tilde{\mathcal{O}} \left(\sqrt{\frac{S^2 L^{11} \overline{m}^2}{n}} \right), \end{aligned}$$

where $\tilde{\mathcal{O}}$ hides polynomial factors in $(\log(\overline{m}SLn), (B_\psi)^m)$.

A.7.4 Position-wise Feed Forward Layer (proof of Lemma A.31 and A.33)

This section proves Lemmas A.31 and A.33 for approximation with feed forward networks.

Proof of Lemma A.31. First we explain how to compute $\mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}]q$. We consider the value of

$$\sum_{\ell=L}^{\ell+1} \mathbf{p}_{v,2(L-\ell+2)} = \sum_{\ell=L}^{\ell+1} \cos\left(\frac{2\pi\iota(\text{pa}^{(L-\ell)}(v))}{m^{(\ell)}}\right), \quad (\text{A.115})$$

which is implemented with one linear layer on $\mathbf{p}_v$. Eq.(A.115) is equal to $L - \ell$ if $v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}$, and otherwise at most $(L - \ell) - (1 - \max_{\ell+1 \leq k \leq L} (\frac{2\pi}{m^{(k)}}))$. Thus, we apply Lemma A.16 with $a = L - \ell$ and $\delta = 1 - \max_{\ell+1 \leq k \leq L} (\frac{2\pi}{m^{(k)}}) \gtrsim \min_{\ell+1 \leq k \leq L} (m^{(k)})^{-2}$ to obtain the network that implements $\mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}]$. Therefore, there exists a network that implements

$$[0; q; 1 - \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}] = \mathbb{1}[v^{(\ell)} \notin \mathcal{V}_{\text{tx}}^{(\ell)}]; \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}]]$$

, where $J \lesssim 1$, $\|\mathbf{j}\|_\infty \lesssim S + d_p$, $B \lesssim L + \max_{\ell+1 \leq k \leq L} (m^{(k)})^2$. Once we obtain this vector, we follow the argument of Lemma A.13 to obtain the network NN_1 that implements $\mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}]q$, with

$$J_1 \lesssim 1, \quad \|\mathbf{j}_1\|_\infty \lesssim S + d_p, \quad B_1 \lesssim L + \max_{\ell+1 \leq k \leq L} (m^{(k)})^2 + C.$$

Next we consider how to compute $a_v^{(\ell)}g$. Note that $a_v^{(\ell)}g = \iota(\text{pa}^{(L-\ell)}(v))g + \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}]g$. $\mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}]g$ is implemented similarly to NN_1 . The first part $\iota(\text{pa}^{(L-\ell)}(v))g$ is obtained by replacing each $\text{NN}_1, \dots, \text{NN}_{m^{(\ell)}}$ by $\iota(\text{pa}^{(L-\ell)}(v))g$ in Lemma A.13. Concatenating these two networks, we obtain NN_2 that implements $a_v^{(\ell)}g$, where

$$J_2 \lesssim 1, \quad \|\mathbf{j}_2\|_\infty \lesssim S + d_p, \quad B_2 \lesssim L + \max_{\ell+1 \leq k \leq L} (m^{(k)})^2 + m^{(\ell)}C.$$

Finally, by concatenating $-\text{NN}_1$, NN_2 , and (the identify function for) q , we get the desired network. $\square$

Proof of Lemma A.33. NN_1 is just a linear function, and thus we focus on NN_2 .

First, we explain how to implement $\mathbf{f}_{\uparrow, \iota(\text{pa}^{L-\ell}(v+1))}^{(\ell)}$. Approximation of each $f_{\uparrow, \iota}^{(\ell)}$ follows from Lemma A.11, which is denoted by $\mathbf{f}_{\uparrow, \iota}^{(\ell)} = \text{NN}_{3, \iota}$ ($\iota = 1, \dots, m^{(\ell)}$). The size of these networks is bounded by

$$J \lesssim (\log \log(SB_\psi/\delta)) \log(SB_\psi/\delta), \quad \|\mathbf{j}\|_\infty \lesssim S(\log(SB_\psi/\delta))^3, \quad B \lesssim 2S(B_\psi^2 + \log(SB_\psi/\delta)).$$

Note that

$$\iota(\text{pa}^{L-\ell}(v+1)) = \begin{cases} \iota(\text{pa}^{L-\ell}(v)) + 1 & (\text{if } v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(L)} \text{ and } \iota(\text{pa}^{L-\ell}(v)) < m^{(\ell)}) \\ 1 & (\text{if } v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(L)} \text{ and } \iota(\text{pa}^{L-\ell}(v)) = m^{(\ell)}) \\ \iota(\text{pa}^{L-\ell}(v)) & (\text{otherwise}) \end{cases}$$

$$= \begin{cases} 1 & (\text{if } \iota(\text{pa}^{L-\ell}(v)) + \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}] = m^{(\ell)} + 1) \\ \iota(\text{pa}^{L-\ell}(v)) + \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}] & (\text{otherwise}) \end{cases}.$$

Thus, for $1 \leq i \leq m^{(\ell)}$,

$$\begin{aligned} \mathbb{1}[\iota(\text{pa}^{L-\ell}(v+1)) = i] & \tag{A.116} \\ = \begin{cases} \mathbb{1}[\iota(\text{pa}^{L-\ell}(v)) + \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}] = i] & \text{if } i \neq 1, \\ \mathbb{1}[\iota(\text{pa}^{L-\ell}(v)) + \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}] = 1] + \mathbb{1}[\iota(\text{pa}^{L-\ell}(v)) + \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}] = m^{(\ell)} + 1] & \text{if } i = 1. \end{cases} \end{aligned}$$

Using this fact, consider how to implement $\mathbb{1}[\iota(\text{pa}^{L-\ell}(v)) + \mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}] = i]$ ($1 \leq i \leq m^{(\ell)} + 1$). $\mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}]$ is implemented in the proof of Lemma A.31, and $\iota(\text{pa}^{L-\ell}(v))$ is implemented by using $\iota(\text{pa}^{L-\ell}(v)) = \sum_{i=1}^{m^{(\ell)}} i \mathbb{1}[i = \iota(\text{pa}^{L-\ell}(v))]$ and Lemma A.16. Once we obtain $\mathbb{1}[v^{(\ell)} \in \mathcal{V}_{\text{tx}}^{(\ell)}] + \iota(\text{pa}^{L-\ell}(v))$, we apply Lemma A.16 again with $\delta = 1$. Therefore, for $1 \leq i \leq m^{(\ell)}$, there exists a network that implements $\mathbb{1}[\iota(\text{pa}^{L-\ell}(v+1)) = i]$, where

$$J \lesssim 1, \|\mathbf{j}\|_{\infty} \lesssim m^{(\ell)} + S + d_p, \quad B \lesssim L + \max_{\ell \leq k \leq L} (m^{(k)})^2.$$

We parallelize these indicators and $\text{NN}_{3,i}$ to obtain the vector

$$[\text{NN}_{3,1}; \text{NN}_{3,2}; \dots; \text{NN}_{3,m^{(\ell)}}; \text{NN}_{3,1}; (\mathbb{1}[\iota(\text{pa}^{L-\ell}(v+1)) = i])_{i=1}^{m^{(\ell)}}].$$

Once we obtain this vector we can follow the proof of Lemma A.13. Therefore, $\mathbf{f}_{\uparrow, \iota(\text{pa}^{L-\ell}(v+1))}^{(\ell)}$ is implemented by a network

$$\begin{aligned} J & \lesssim (\log \log(SB_{\psi}/\delta)) \log(SB_{\psi}/\delta), \\ \|\mathbf{j}\|_{\infty} & \lesssim m^{(\ell)} S (\log(SB_{\psi}/\delta))^3 + d_p, \\ B & \lesssim S(B_{\psi}^2 + \log(SB_{\psi}/\delta)) + \max_{\ell \leq k \leq L} (m^{(k)})^2 + L. \end{aligned}$$

Next, we consider how to distinguish the three cases—(i) $\text{pa}^{(L-\ell-1)}(v) = \text{pa}^{(L-\ell-1)}(v+1)$, (ii) $\text{pa}^{(L-\ell)}(v) = \text{pa}^{(L-\ell)}(v+1)$ but $\text{pa}^{(L-\ell-1)}(v) \neq \text{pa}^{(L-\ell-1)}(v+1)$, and (iii) otherwise. Consider the values

$$\sum_{\ell=L}^{\ell+2} \mathbf{p}_{v, 2(L-\ell+2)} = \sum_{\ell=L}^{\ell+2} \cos\left(\frac{2\pi \iota(\text{pa}^{(L-\ell)}(v))}{m^{(\ell)}}\right), \tag{A.117}$$

and

$$\sum_{\ell=L}^{\ell+1} \mathbf{p}_{v, 2(L-\ell+2)} = \sum_{\ell=L}^{\ell+1} \cos\left(\frac{2\pi \iota(\text{pa}^{(L-\ell)}(v))}{m^{(\ell)}}\right). \tag{A.118}$$

(A.117) is equal to $L - \ell - 2$ iff $\text{pa}^{(L-\ell-1)}(v) = \text{pa}^{(L-\ell-1)}(v+1)$, and (A.118) is equal to $L - \ell - 1$ iff $\text{pa}^{(L-\ell)}(v) = \text{pa}^{(L-\ell)}(v+1)$. Therefore, the vector of the indicator functions $[\mathbb{1}[(\text{A.117}) = L - \ell - 2], \mathbb{1}[(\text{A.117}) = L - \ell - 2] - \mathbb{1}[(\text{A.118}) = L - \ell - 1], \mathbb{1}[(\text{A.118}) = L - \ell - 1]]$ correspond to the vector of the three cases $[\mathbb{1}[(\text{i})], \mathbb{1}[(\text{ii})], \mathbb{1}[(\text{iii})]]$. It is easy to implement $\mathbb{1}[(\text{A.117}) = L - \ell - 2]$ and $\mathbb{1}[(\text{A.118}) = L - \ell - 1]$ by following (A.115). Now, we can determine whether to add h and q , and the rest of the proof is the same as Lemma A.13.

Putting it all together, we obtain the desired network. $\square$

A.7.5 Self-attention layer (proof of Lemma A.32)

Define the auxiliary key and the query matrices $\overline{\mathbf{W}}_K^{(\ell)}, \overline{\mathbf{W}}_Q^{(\ell)} \in \mathbb{R}^{d_p \times d_p}$ as $\overline{\mathbf{W}}_K^{(\ell)} = \mathbf{I}_{d_p}$, and

$$(\overline{\mathbf{W}}_Q^{(\ell)})_{i,j} = \begin{cases} \alpha & \text{if } j = 2 \text{ and } i = 4, 6, \dots, 2(L - \ell + 1), \\ & \text{or } 2(L - \ell + 2) + 1 \leq i = j \leq 2(L + 1), \\ (L - 1)\alpha & (i, j) = (1, 2), \\ 0 & \text{otherwise.} \end{cases}$$

Then, the value of QK matrix is

$$\begin{aligned} & ((\overline{\mathbf{W}}_K^{(\ell)} Q^{(\ell)})^\top (\overline{\mathbf{W}}_Q^{(\ell)} Q^{(\ell)}))_{u,v} \\ &= \alpha \sum_{\ell' < \ell} \cos\left(\frac{2\pi\iota(\text{pa}^{(L-\ell')}(v))}{m^{(\ell')}}\right) \\ & \quad + \alpha \sum_{\ell' > \ell} \left[\sin\left(\frac{2\pi\iota(\text{pa}^{(L-\ell')}(v))}{m^{(\ell')}}\right) \sin\left(\frac{2\pi\iota(\text{pa}^{(L-\ell')}(u))}{m^{(\ell')}}\right) + \cos\left(\frac{2\pi\iota(\text{pa}^{(L-\ell')}(u))}{m^{(\ell')}}\right) \cos\left(\frac{2\pi\iota(\text{pa}^{(L-\ell')}(v))}{m^{(\ell')}}\right) \right] \\ &= \alpha \sum_{\ell' < \ell} \cos\left(\frac{2\pi\iota(\text{pa}^{(L-\ell')}(v))}{m^{(\ell')}}\right) + \alpha \sum_{\ell' > \ell} \cos\left(\frac{2\pi\iota(\text{pa}^{(L-\ell')}(v)) - \iota(\text{pa}^{(L-\ell')}(u))}{m^{(\ell')}}\right). \end{aligned} \quad (\text{A.119})$$

For $v = 0$, the attention mask ensures that the output is always $\mathbf{q}_0^{(\ell)}$, and thus let us focus on $v \in \mathcal{V}_{\text{tx}}^{(L)}$. In (A.119), the maximum value is $(L - 1)\alpha$, which is achieved when $u \in \mathcal{V}^{(L)}$ and $u^{(\ell)} = \mathcal{C}(\text{pa}^{(L-\ell+1)}(v))$, or $u = 0$. Otherwise, $((\overline{\mathbf{W}}_K^{(\ell)} Q^{(\ell)})^\top (\overline{\mathbf{W}}_Q^{(\ell)} Q^{(\ell)}))_{u,v}$ is smaller than $\alpha(L - 1)$ by $1 - \max_{\ell'} \cos\left(\frac{2\pi}{m^{(\ell')}}\right) \gtrsim \min_{\ell'} (m^{(\ell')})^{-2}$.

Therefore, by following the argument of Lemma A.17 and taking $\alpha \simeq \log(d/\delta)$, we have

$$\|(\text{softmax}(\mathbf{M}_{d_p} + (\overline{\mathbf{W}}_{K,1}^{(\ell)} \mathbf{P})^\top (\overline{\mathbf{W}}_{Q,1}^{(\ell)} \mathbf{P})) - \text{softmax}(\mathbf{A}^{(\ell)}))_{u,v}\|_\infty \leq \delta, \quad u, v \in \mathcal{V}^{(L)},$$

where $\mathbf{A}^{(\ell)} \in \mathbb{R}^{(d+1) \times (d+1)}$ is a matrix such that $\mathbf{A}_{u,v}^{(\ell)} = \frac{1}{a_v^{(\ell)}}$ if “ $u, v \in \mathcal{V}^{(L)}$ and $u^{(\ell)} \in \mathcal{C}(\text{pa}^{(L-\ell)}(v))$ ”, $\mathbf{A}_{u,v}^{(\ell)} = 1$ for $u, v = 0$, and $\mathbf{A}_{u,v}^{(\ell)} = 0$ otherwise. There is no approximation error in the column corresponding to $v = 0$ because the mask excludes the dependency on all the other variables.

By following the proof of Lemma A.10, we obtain matrices $\mathbf{W}_K^{(\ell)}, \mathbf{W}_Q^{(\ell)}, \mathbf{W}_V^{(\ell)} \in \mathbb{R}^{D \times D}$ with $\|\mathbf{W}_K^{(\ell)}\|, \|\mathbf{W}_Q^{(\ell)}\|, \|\mathbf{W}_V^{(\ell)}\| \lesssim \log(d/\delta)$ such that

$$\begin{aligned} & [(\mathbf{W}_V^{(\ell)} Q^{(\ell)}) \text{softmax}(\mathbf{M} + (\mathbf{W}_K^{(\ell)} Q^{(\ell)})^\top (\mathbf{W}_Q^{(\ell)} Q^{(\ell)}))]_v \\ &= \begin{cases} \begin{bmatrix} \mathbf{0} & (\in \mathbb{R}^{(3\ell+L-1)S}) \\ \frac{1}{a_v^{(\ell)}} \left(\mathbf{q}_0^{(\ell)} + \sum_{u^{(\ell)} \in \mathcal{C}(\text{pa}^{(L-\ell+1)}(v))} \mathbb{1}[u \leq v] \mathbf{q}_u^{(\ell)} \right) + \boldsymbol{\delta}_v^{(\ell)} & (\in \mathbb{R}^S) \\ \mathbf{0} & (\in \mathbb{R}^{d_p + (3L-3\ell+2)S+1}) \end{bmatrix}, & (v \in \mathcal{V}_{\text{tx}}^{(L)}) \\ \begin{bmatrix} \mathbf{0} & (\in \mathbb{R}^{(3\ell+L-1)S}) \\ \mathbf{q}_0^{(\ell)} & (\in \mathbb{R}^S) \\ \mathbf{0} & (\in \mathbb{R}^{d_p + (3L-3\ell+2)S+1}) \end{bmatrix}, & (v = 0) \end{cases} \end{aligned}$$

where $\|\boldsymbol{\delta}_v^{(\ell)}\|_\infty \leq \delta \max_{v'} \|\mathbf{q}_{v'}^{(\ell)}\|_\infty$.

A.7.6 Evaluation of error propagation

Lemma A.35 (Evaluation of error propagation). *Assume we have functions $\mathbf{f}_{\downarrow,\iota}^{(\ell)}, \mathbf{f}_{\uparrow,\iota}^{(\ell)}$ ($1 \leq \ell \leq L, \iota \in [m_{\text{tx}}^{(\ell)}]$) such that*

$$\begin{aligned} \|f_{\downarrow,\iota}^{(\ell)}(h) - \mathbf{f}_{\downarrow,\iota}^{(\ell)}(h)\|_{\infty} &\leq \delta, \quad \forall h \in \mathbb{R}^S \text{ such that } \max_{s \in S} h_s = 0, \ell \in [L], \\ \|f_{\uparrow,\iota}^{(\ell)}(h) - \mathbf{f}_{\uparrow,\iota}^{(\ell)}(h)\|_{\infty} &\leq \delta, \quad \forall h \in \mathbb{R}^S \text{ such that } \max_{s \in S} h_s = 0, \ell \in [L], \end{aligned} \quad (\text{A.120})$$

and $a_v^{(\ell)} \|\delta_v^{(\ell)}\|_{\infty} \leq \delta$ holds for all $\ell = L, \dots, 1$ and $v \in \mathcal{V}_{\text{tx}}^{(L)}$. Moreover, we assume that $\|h_{\text{im},\text{r}}^{(0)} - \mathbf{h}_{\text{im},\text{d}}^{(0)}\|_{\infty} \leq \delta_{\text{im}}$.

Consider the approximated update introduced in (A.109) and (A.110). Then, we have the following bound on the error propagation:

$$\max_{v \in \mathcal{V}_{\text{tx}}^{(L)}} \|h_v^{(\ell)} - \mathbf{h}_v^{(\ell)}\|_{\infty} \leq \delta \times (2m_{\text{tx}}^{(\ell+1)} + 4) \prod_{\ell+2 \leq k \leq L} (2m_{\text{tx}}^{(k)} + 5), \quad (\text{A.121})$$

$$\max_{v \in \mathcal{V}_{\text{tx}}^{(L)}} \|q_v^{(\ell)} - \mathbf{q}_v^{(\ell)}\|_{\infty} \leq \delta \times \prod_{\ell+1 \leq k \leq L} (2m_{\text{tx}}^{(k)} + 5), \quad (\text{A.122})$$

$$\max_{v \in \mathcal{V}_{\text{tx}}^{(L)}} \|b_v^{(L+1)} - \mathbf{b}_v^{(L+1)}\|_{\infty} \leq 8^{L+1} \delta \prod_{1 \leq k \leq L} (2m_{\text{tx}}^{(k)} + 5) + \delta_{\text{tx}}. \quad (\text{A.123})$$

Furthermore, we have

$$\begin{aligned} &\|\text{softmax}(\mathbf{b}_v^{(L+1)})_s - \mu_{\star}(x_{\text{tx},v+1} = s | \mathbf{x}_{\text{im}}, x_{\text{tx},1}, \dots, x_{\text{tx},v})\|_{\infty} \\ &\leq 8^{L+1} \delta \prod_{1 \leq k \leq L} (2m_{\text{tx}}^{(k)} + 5) + \delta_{\text{im}}, \quad s \in [S], v = 1, 2, \dots, d-1. \end{aligned} \quad (\text{A.124})$$

Proof. The error from the image model is evaluated by Lemma A.27. Thus in the following we will assume $\delta_{\text{im}} = 0$.

First, we prove (A.121) and (A.122). Because $\mathbf{h}_v^{(L)} = h_v^{(L)}$, (A.120) implies that

$$\|q_v^{(L)} - \mathbf{q}_v^{(L)}\|_{\infty} \leq \delta.$$

By Lemma A.45 and $a_v^{(L)} \|\delta_v^{(L)}\|_{\infty} \leq \delta$, we have

$$\|h_v^{(L-1)} - \mathbf{h}_v^{(L-1)}\|_{\infty} \leq 2(m_{\text{tx}}^{(L)} + 1) \max_{u \in \mathcal{V}_{\text{tx}}} \|f_{\downarrow,\iota(u)}^{(L)}(x_{\text{tx},u}^{(L)}) - \mathbf{f}_{\downarrow,\iota(u)}^{(L)}(x_{\text{tx},u}^{(L)})\|_{\infty} + 2\delta \leq (2m_{\text{tx}}^{(L)} + 4)\delta.$$

This confirms (A.121) for $\ell = L - 1$.

Suppose that (A.121) holds for some $\ell (\leq L - 1)$ and prove (A.122) for ℓ and (A.121) for $\ell - 1$. For (A.122),

$$\begin{aligned} &\|q_v^{(\ell)} - \mathbf{q}_v^{(\ell)}\|_{\infty} \\ &= \max_{v \in \mathcal{V}_{\text{tx}}} \|f_{\downarrow,\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(h_v^{(\ell)}) - \mathbf{f}_{\downarrow,\iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(h_v^{(\ell)})\|_{\infty} \end{aligned}$$

$$\begin{aligned}
&\leq \max_{v \in \mathcal{V}_{\text{tx}}} \|f_{\downarrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(h_v^{(\ell)}) - \mathbf{f}_{\downarrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{h}_v^{(\ell)})\|_\infty + \|f_{\downarrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(\mathbf{h}_v^{(\ell)}) - \mathbf{f}_{\downarrow, \iota(\text{pa}^{(L-\ell)}(v))}^{(\ell)}(h_v^{(\ell)})\|_\infty \\
&\leq \max_{v \in \mathcal{V}_{\text{tx}}} \|h_v^{(\ell)} - \mathbf{h}_v^{(\ell)}\|_\infty + \delta \\
&\leq \delta + \delta \times (2m_{\text{tx}}^{(\ell+1)} + 4) \prod_{\ell+2 \leq k \leq L} (2m_{\text{tx}}^{(k)} + 5) \\
&\leq \delta \times \prod_{\ell+1 \leq k \leq L} (2m_{\text{tx}}^{(k)} + 5), \tag{A.125}
\end{aligned}$$

where we used Lemma A.49 for the second inequality. Also,

$$\begin{aligned}
\|h_v^{(\ell-1)} - \mathbf{h}_v^{(\ell-1)}\|_\infty &\leq 2(m_{\text{tx}}^{(\ell)} + 1) \max_{v \in \mathcal{V}_{\text{tx}}} \|q_v^{(\ell)} - \mathbf{q}_v^{(\ell)}\|_\infty + 2\delta \\
&\leq \delta \times 2(m_{\text{tx}}^{(\ell)} + 1) \prod_{k=\ell+1}^L (2m_{\text{tx}}^{(k)} + 5) + 2\delta \\
&\leq \delta \times (2m_{\text{tx}}^{(\ell)} + 4) \prod_{k=\ell+1}^L (2m_{\text{tx}}^{(k)} + 5),
\end{aligned}$$

where we used Lemma A.45 and $a_v^{(\ell)} \|\delta_{\text{tx},v}^{(\ell)}\|_\infty \leq \delta$ for the first inequality, and (A.125) for the second inequality. Therefore, by induction, we obtained (A.121) for all $\ell = L, \dots, 0$ and (A.122) for all $\ell = L, \dots, 1$.

(A.123) is derived by following Lemma A.26. Finally, from the Lipschitzness of softmax, we obtain (A.124). $\square$

A.7.7 Proof of Proposition A.29

Fix i ($1 \leq i \leq d-1$). We show that $\text{softmax}(\bar{b}_i^{(L)})_s = \nu_{\uparrow, i+1}^{(L)}(s)$ for all $s \in [S]$. (Remember Lemma A.28, which states that $\nu_{\uparrow, i}^{(L)}(x_{\text{tx}, i+1})_s = \mu_\star(x_{\text{tx}, i+1} = s | \mathbf{x}_{\text{im}}, x_{\text{tx}, 1}, \dots, x_{\text{tx}, i})$.) Without loss of generality, for all ℓ and ι , we assume that $\sum_{a \in [S]} \psi_{\text{tx}, \iota}^{(\ell)}(s, a)$ is constant for all $s \in [S]$.

We first consider the downsampling. We will verify that, for $v \in \mathcal{V}_{\text{tx}}^{(\ell)}$,

$$\nu_{\downarrow, v}^{(\ell)}(s) = \begin{cases} \text{softmax}(h_{v^{(L)}}^{(\ell)})_s & (\text{if } v \leq i), \\ \text{softmax}(h_i^{(\ell)})_s & (\text{if } v = \text{pa}^{(L-\ell)}(i)), \\ \frac{1}{S} & (\text{otherwise}), \end{cases} \tag{A.126}$$

for $\ell = L-1, \dots, 0$ by induction.

We first verify that (A.126) holds for $\ell = L-1$. For $v \in \mathcal{V}_{\text{tx}}^{(L-1)}$, if $v \leq i$, all children of v are observed, and we have that

$$\begin{aligned}
\nu_{\downarrow, v}^{(L-1)}(s) &\propto \prod_{v' \in \mathcal{C}(v)} \psi_{\text{tx}, \iota(v')}^{(L)}(s, x_{\text{tx}, v'}) \\
&= \prod_{v' \in \mathcal{N}(v^{(L)}) \cup \{v^{(L)}\}} (\psi_{\text{tx}, \iota(v')}^{(L)}(s, x_{\text{tx}, v'}))^{1[v' \leq v]} \propto \text{softmax}(h_{v^{(L)}}^{(L-1)})_s,
\end{aligned}$$

else if $\text{pa}(i) = v$, we have

$$\nu_{\downarrow, v}^{(L-1)}(s) \propto \sum_{x_{\text{tx}, \mathcal{C}(v)}^{(L)}} \prod_{v' \in \mathcal{C}(v)} (\psi_{\text{tx}, \iota(v')}^{(L)}(s, x_{\text{tx}, v'}^{(L)}) \nu_{\downarrow, v'}^{(L)}(x_{\text{tx}, v'}^{(L)}))$$

$$\begin{aligned}
&= \prod_{v' \in \mathcal{C}(v)} \left(\mathbb{1}[v' \leq i] \psi_{\text{tx}, \iota(v')}^{(L)}(s, x_{\text{tx}, v'}) + \mathbb{1}[v' > i] \frac{1}{S} \right) \\
&\propto \prod_{v' \in \mathcal{N}(i) \cup \{i\}} \left(\psi_{\text{tx}, \iota(v')}^{(L)}(s, x_{\text{tx}, v'}) \right)^{\mathbb{1}[v' \leq i]} \\
&\propto \text{softmax}(h_i^{(L-1)})_s,
\end{aligned}$$

and else, when $\text{pa}(i) < v$, none of the leaf nodes under v is observed and we have

$$\begin{aligned}
\nu_{\downarrow, v}^{(L-1)}(s) &\propto \sum_{x_{\text{tx}, \mathcal{C}(v)}^{(L)}} \prod_{v' \in \mathcal{C}(v)} \left(\psi_{\text{tx}, \iota(v')}^{(L)}(s, x_{\text{tx}, v'}^{(L)}) \nu_{\downarrow, v'}^{(L)}(x_{\text{tx}, v'}^{(L)}) \right) \\
&= \sum_{x_{\text{tx}, \mathcal{C}(v)}^{(L)}} \prod_{v' \in \mathcal{C}(v)} \left(\psi_{\text{tx}, \iota(v')}^{(L)}(s, x_{\text{tx}, v'}^{(L)}) \frac{1}{S} \right) \propto \frac{1}{S}.
\end{aligned}$$

Then, assuming that (A.126) holds for some $\ell \in [L-1]$, we will prove (A.126) for $\ell-1$. If $v \leq i$, because all $v' \in \mathcal{C}(v)$ satisfy $v' \leq i$ and thus $\nu_{\downarrow, v'}^{(\ell)}(x_{\text{tx}, v'}^{(\ell)}) \propto \exp((h_{v'}^{(\ell)})_{x_{\text{tx}, v'}^{(\ell)}})$, we have that

$$\begin{aligned}
\nu_{\downarrow, v}^{(\ell-1)}(s) &\propto \sum_{x_{\text{tx}, \mathcal{C}(v)}^{(\ell)}} \prod_{v' \in \mathcal{C}(v)} \left(\psi_{\text{tx}, \iota(v')}^{(\ell)}(s, x_{\text{tx}, v'}^{(\ell)}) \nu_{\downarrow, v'}^{(\ell)}(x_{\text{tx}, v'}^{(\ell)}) \right) \\
&= \prod_{v' \in \mathcal{C}(v)} \left(\sum_{x_{\text{tx}, v'}^{(\ell)}} \psi_{\text{tx}, \iota(v')}^{(\ell)}(s, x_{\text{tx}, v'}^{(\ell)}) \nu_{\downarrow, v'}^{(\ell)}(x_{\text{tx}, v'}^{(\ell)}) \right) \\
&\propto \prod_{v' \in \mathcal{C}(v)} \left(\sum_{x_{\text{tx}, \mathcal{C}(v)}^{(\ell)}} \psi_{\text{tx}, \iota(v')}^{(\ell)}(s, x_{\text{tx}, v'}^{(\ell)}) \exp((h_{v'}^{(\ell)})_{x_{\text{tx}, v'}^{(\ell)}}) \right) \\
&\propto \text{softmax}(\prod_{v' \in \mathcal{C}(v)} q_{v'(L)}^{(\ell)})_s, \\
&= \text{softmax}(\prod_{v' \in \mathcal{C}(v)} \mathbb{1}[v' \leq v] q_{v'}^{(\ell)})_s, \\
&= \text{softmax}(h_{v^{(L)}}^{(\ell-1)})_s,
\end{aligned}$$

else if $v = \text{pa}^{(L-\ell+1)}(i)$, we have that

$$\begin{aligned}
&\nu_{\downarrow, v}^{(\ell-1)}(s) \\
&\propto \sum_{x_{\text{tx}, \mathcal{C}(v)}^{(\ell)}} \prod_{v' \in \mathcal{C}(v)} \left(\psi_{\text{tx}, \iota(v')}^{(\ell)}(s, x_{\text{tx}, v'}^{(\ell)}) \nu_{\downarrow, v'}^{(\ell)}(x_{\text{tx}, v'}^{(\ell)}) \right) \\
&= \prod_{v' \in \mathcal{C}(v)} \left(\mathbb{1}[v' \leq i] \sum_{x_{\text{tx}, v'}^{(\ell)}} \psi_{\text{tx}, \iota(v')}^{(\ell)}(s, x_{\text{tx}, v'}^{(\ell)}) \nu_{\downarrow, v'}^{(\ell)}(x_{\text{tx}, v'}^{(\ell)}) \right. \\
&\quad \left. + \mathbb{1}[v' > i] \sum_{x_{\text{tx}, v'}^{(\ell)}} \psi_{\text{tx}, \iota(v')}^{(\ell)}(s, x_{\text{tx}, v'}^{(\ell)}) \nu_{\downarrow, v'}^{(\ell)}(x_{\text{tx}, v'}^{(\ell)}) \right) \\
&\propto \prod_{v' \in \mathcal{C}(v)} \left(\mathbb{1}[v' \leq i] \sum_{x_{\text{tx}, v'}^{(\ell)}} \psi_{\text{tx}, \iota(v')}^{(\ell)}(s, x_{\text{tx}, v'}^{(\ell)}) \nu_{\downarrow, v'}^{(\ell)}(x_{\text{tx}, v'}^{(\ell)}) \right. \\
&\quad \left. + \mathbb{1}[v' > i] \sum_{x_{\text{tx}, v'}^{(\ell)}} \psi_{\text{tx}, \iota(v')}^{(\ell)}(s, x_{\text{tx}, v'}^{(\ell)}) \frac{1}{S} \right) \\
&\propto \prod_{v' \in \mathcal{C}(v)} \left(\mathbb{1}[v' \leq i] \sum_{x_{\text{tx}, \mathcal{C}(v)}^{(\ell)}} \psi_{\text{tx}, \iota(v')}^{(\ell)}(s, x_{\text{tx}, v'}^{(\ell)}) \exp((h_{v'}^{(\ell)})_{x_{\text{tx}, v'}^{(\ell)}}) \right)
\end{aligned}$$

$$\begin{aligned}
&= \prod_{\substack{v'^{(L-\ell)} \in \mathcal{N}(\text{pa}^{(L-\ell)}(v^{(L)})) \\ \text{or } v'=v}} (\mathbb{1}[v' \leq i] q_{v'}^{(\ell)}) \\
&= \text{softmax}(h_i^{(\ell-1)})_s,
\end{aligned}$$

and else, when $\text{pa}^{(L-\ell+1)}(i) < v$, none of the leaf nodes under v is observed and we have

$$\begin{aligned}
\nu_{\downarrow, v}^{(\ell-1)}(s) &\propto \sum_{x_{\text{tx}, \mathcal{C}(v)}^{(\ell)}} \prod_{v' \in \mathcal{C}(v)} \left(\psi_{\text{tx}, \iota(v')}^{(\ell)}(s, x_{\text{tx}, v'}^{(\ell)}) \nu_{\downarrow, v'}^{(\ell)}(x_{\text{tx}, v'}^{(\ell)}) \right) \\
&= \sum_{x_{\text{tx}, \mathcal{C}(v)}^{(\ell)}} \prod_{v' \in \mathcal{C}(v)} \left(\psi_{\text{tx}, \iota(v')}^{(\ell)}(s, x_{\text{tx}, v'}^{(\ell)}) \frac{1}{S} \right) \propto \frac{1}{S}.
\end{aligned}$$

Therefore, we have obtained (A.126) for $\ell - 1$, and the induction proves that (A.126) holds for all $\ell = L - 1, \dots, 0$.

We next consider the upsampling. Let ℓ_* be the largest ℓ such that $\text{pa}^{(L-\ell_*)}(i) = \text{pa}^{(L-\ell_*)}(i+1)$ holds. We will verify that, for $\ell = 0, 1, \dots, L$ and $v = \text{pa}^{(L-\ell)}(i+1)$,

$$\nu_{\uparrow, v}^{(\ell)}(s) = \begin{cases} \text{softmax}(\bar{b}_{\uparrow, i}^{(\ell)} - h_{\downarrow, i}^{(\ell)})_s & (\ell = 0, 1, \dots, \ell_*), \\ \text{softmax}(\bar{b}_{\downarrow, i}^{(\ell)})_s & (\ell = \ell_* + 1, \dots, L), \end{cases} \quad (\text{A.127})$$

by induction.

Checking (A.127) for $\ell = 0$ is done by just comparing the definitions of $\nu_{\uparrow, v}^{(0)}$ and $\bar{b}_{\uparrow, v}^{(0)}$.

Suppose that (A.127) holds for some ℓ and $v = \text{pa}^{(L-\ell)}(i+1)$. We will prove that (A.127) holds for $\ell + 1$ and $v = \text{pa}^{(L-\ell-1)}(i+1)$. If $\ell + 1 \leq \ell_*$, we have that $v = \text{pa}^{(L-\ell-1)}(i) = \text{pa}^{(L-\ell-1)}(i+1)$, and that

$$\begin{aligned}
&\nu_{\uparrow, v}^{(\ell+1)}(x_{\text{tx}, v}^{(\ell+1)}) \\
&\propto \sum_{x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, \mathcal{N}(v)}^{(\ell+1)}} \psi_{\text{tx}, \iota(v)}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, \mathcal{C}(\text{pa}(v))}^{(\ell+1)}) \nu_{\uparrow, \text{pa}(v)}^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) \prod_{v' \in \mathcal{N}(v)} \nu_{\downarrow, v'}^{(\ell+1)}(x_{\text{tx}, v'}^{(\ell+1)}) \\
&\propto \sum_{x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, \mathcal{N}(v)}^{(\ell+1)}} \psi_{\text{tx}, \iota(v)}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, \mathcal{C}(\text{pa}(v))}^{(\ell+1)}) \nu_{\uparrow, \text{pa}(v)}^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) \prod_{v' \in \mathcal{N}(v)} \mathbb{1}[v' \leq i] \nu_{\downarrow, v'}^{(\ell+1)}(x_{\text{tx}, v'}^{(\ell+1)}) \\
&= \sum_{x_{\text{tx}, \text{pa}(v)}^{(\ell)}} \left(\psi_{\text{tx}, \iota(v)}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_v^{(\ell+1)}) \nu_{\uparrow, \text{pa}(v)}^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) \right. \\
&\quad \left. \prod_{v' \in \mathcal{N}(v)} \left(\sum_{x_{\text{tx}, v'}^{(\ell+1)}} \psi_{\text{tx}, \iota(v')}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, v'}^{(\ell+1)}) \exp(h_{v'}^{(\ell+1)}(x_{\text{tx}, v'}^{(\ell+1)})) \right)^{\mathbb{1}[v' \leq i]} \right) \\
&= \sum_{x_{\text{tx}, \text{pa}(v)}^{(\ell)}} \left(\psi_{\text{tx}, \iota(v)}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, v}^{(\ell+1)}) \nu_{\uparrow, \text{pa}(v)}^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) \exp \left(\sum_{v' \in \mathcal{N}(v)} \mathbb{1}[v' \leq i] q_{v'}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) \right) \right) \\
&\propto \sum_{x_{\text{tx}, \text{pa}(v)}^{(\ell)}} \left(\psi_{\text{tx}, \iota(v)}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, v}^{(\ell+1)}) \right. \\
&\quad \left. \exp \left(\bar{b}_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) - h_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) + h_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) - q_i^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) \right) \right) \\
&= \sum_{x_{\text{tx}, \text{pa}(v)}^{(\ell)}} \left(\psi_{\text{tx}, \iota(v)}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, v}^{(\ell+1)}) \exp \left(\bar{b}_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) - q_i^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) \right) \right)
\end{aligned} \quad (\text{A.128})$$

$$\begin{aligned}
&\propto \sum_{x_{\text{tx}, \text{pa}(v)}^{(\ell)}} \left(\psi_{\text{tx}, \iota(v)}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, v}^{(\ell+1)}) \right. \\
&\quad \left. \exp \left(\bar{b}_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) - h_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) + h_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) - q_i^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) \right) \right) \\
&= \sum_{x_{\text{tx}, \text{pa}(v)}^{(\ell)}} \left(\psi_{\text{tx}, \iota(v)}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, v}^{(\ell+1)}) \exp \left(\bar{b}_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) - q_i^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) \right) \right)
\end{aligned} \quad (\text{A.129})$$

$$\begin{aligned}
&\propto \exp \left(f_{\uparrow, \iota(v)}^{(\ell+1)} \left(\bar{b}_{\uparrow, i}^{(\ell)} - q_{\downarrow, i}^{(\ell+1)} \right) \right)_{x_{\text{tx}, v}^{(\ell+1)}} \\
&\propto \text{softmax} \left(\bar{b}_i^{(\ell+1)} - h_i^{(\ell+1)} \right)_{x_{\text{tx}, v}^{(\ell+1)}}.
\end{aligned}$$

In (A.129), because $\text{pa}^{(L-\ell-1)}(i)$ and v means the same child node of $\text{pa}(v)$, the condition “ $v'^{(L-\ell-1)} \in \mathcal{N}(v)$ ” is equivalent to “ $v'^{(L-\ell-1)} \in \mathcal{N}(\text{pa}^{(L-\ell-1)}(i))$ ” and does not overlap with “ $v' = i$ ”. Therefore, in (A.129), we used that $\sum_{v' \in \mathcal{N}(v)} \mathbb{1}[v' \leq i] q_{v'^{(L)}}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) = \sum_{\substack{v' \in \mathcal{N}(\text{pa}^{(L-\ell-1)}(i)) \\ \text{or } v' = i}} \mathbb{1}[v' \leq i] q_{v'}^{(\ell+1)} - q_i^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) = h_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) - q_i^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)})$. Else if $\ell = \ell_*$, we have that

$$\begin{aligned}
\nu_{\uparrow, v}^{(\ell+1)}(x_{\text{tx}, v}^{(\ell+1)}) &\propto \text{(A.128)} \\
&= \sum_{x_{\text{tx}, \text{pa}(v)}^{(\ell)}} \left(\psi_{\text{tx}, \iota(v)}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, v}^{(\ell+1)}) \exp \left(\bar{b}_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) - h_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) + h_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) \right) \right) \\
&= \sum_{x_{\text{tx}, \text{pa}(v)}^{(\ell)}} \left(\psi_{\text{tx}, \iota(v)}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, v}^{(\ell+1)}) \exp \left(\bar{b}_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) \right) \right) \\
&\propto \exp \left(f_{\uparrow, \iota(v)}^{(\ell+1)} \left(\bar{b}_i^{(\ell)} \right) \right)_{x_{\text{tx}, v}^{(\ell+1)}} \\
&\propto \text{softmax} \left(\bar{b}_i^{(\ell+1)} \right)_{x_{\text{tx}, v}^{(\ell+1)}}.
\end{aligned} \tag{A.130}$$

In (A.130), because $\text{pa}^{(L-\ell-1)}(i)$ and v are different child nodes of $\text{pa}(v)$, and “ $v'^{(L-\ell-1)} \in \mathcal{N}(\text{pa}^{(L-\ell-1)}(i))$ or $v' = i$ ” is equivalent to $v'^{(L-\ell-1)} \in \mathcal{N}(v)$, we used that

$$\sum_{v' \in \mathcal{N}(v)} \mathbb{1}[v' \leq i] q_{v'^{(L)}}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) = \sum_{\substack{v' \in \mathcal{N}(\text{pa}^{(L-\ell-1)}(i)) \\ \text{or } v' = i}} \mathbb{1}[v' \leq i] q_{v'}^{(\ell+1)} = h_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)})$$

(ignoring normalize). Else, when $\ell > \ell_*$, $v = \text{pa}^{(L-\ell-1)}(i+1)$ does not have observed leaf nodes as its descendants, and

$$\begin{aligned}
\nu_{\uparrow, v}^{(\ell+1)}(x_{\text{tx}, v}^{(\ell+1)}) &\propto \text{(A.128)} = \sum_{x_{\text{tx}, \text{pa}(v)}^{(\ell)}} \left(\psi_{\text{tx}, \iota(v)}^{(\ell+1)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}, x_{\text{tx}, v}^{(\ell+1)}) \exp \left(\bar{b}_i^{(\ell)}(x_{\text{tx}, \text{pa}(v)}^{(\ell)}) \right) \right) \\
&\propto \exp \left(f_{\uparrow, \iota(v)}^{(\ell+1)} \left(\bar{b}_i^{(\ell)} \right) \right)_{x_{\text{tx}, v}^{(\ell+1)}} \propto \text{softmax} \left(\bar{b}_i^{(\ell+1)} \right)_{x_{\text{tx}, v}^{(\ell+1)}}.
\end{aligned}$$

Now, by induction, we have (A.127) for all $\ell = 0, 1, \dots, L$ and $v \in \text{pa}^{(L-\ell)}(i+1)$.

It always holds that $\ell_* < L$. Therefore, we obtain that $\nu_{\uparrow, v}^{(L)} = \text{softmax}(\bar{b}_i^{(L)})$, which finishes the proof.

A.7.8 Bound on the posterior probability

As an auxiliary lemma, we state the boundedness of $\bar{b}_v^{(\ell)}$.

Lemma A.36. *Under Assumption 2.4 and 2.5, we have that*

$$\frac{1}{SB_\psi} \leq \mu_\star(x_{\text{tx},v+1} = s | \mathbf{x}_{\text{im}}, x_{\text{tx},1}, \dots, x_{\text{tx},v}) \leq 1,$$

for all $v = 1, \dots, d-1$.

Proof. Consider the message passing algorithm in (A.102) and (A.104). For $\ell = L$, $\text{pa}^{(L-L)}(v) \neq \text{pa}^{(L-L)}(v+1)$ always holds. Because $\bar{b}_v^{(L)} = f_{\uparrow, \iota(v+1)}^{(L)}(\text{normalize}(\bar{b}_{v+1}^{(L-1)}))$ and

$$(f_{\uparrow, \iota}^{(\ell)}(h))_s = \log \sum_{a \in [S]} \psi_{\text{tx}, \iota}^{(\ell)}(a, s) e^{h_a}, \quad (\psi_{\text{tx}, \iota}^{(\ell)}(a, s) \geq B_\psi^{-1})$$

$\bar{b}_v^{(L)}$ is bounded by $-\log B_\psi \leq (\bar{b}_v^{(L)})_s$, and $\text{softmax}(\bar{b}_v^{(L)})_s = \mu_\star(x_{\text{tx},v+1} = s | \mathbf{x}_{\text{im}}, x_{\text{tx},1}, \dots, x_{\text{tx},v})$ (this equivalence is proven in Proposition A.29) is bounded by

$$\frac{1}{SB_\psi} \leq \mu_\star(x_{\text{tx},v+1} = s | \mathbf{x}_{\text{im}}, x_{\text{tx},1}, \dots, x_{\text{tx},v}) \leq 1,$$

for all $v = 1, \dots, d-1$. □

A.8 Auxiliary lemmas

A.8.1 Lipschitzness of transformers

In this section, we establish the Lipschitz continuity of the transformers in their parameters. Let $\|\mathbf{H}\|_{2,\infty} := \max_{i \in N} \|\mathbf{H}_i\|_2$ denote the column-wise $(2, \infty)$ -norm for any matrix $\mathbf{H} \in \mathbb{R}^{M \times N}$. For any $R > 0$, we let $\mathcal{H}_R = \{\mathbf{H} : \|\mathbf{H}\|_{2,\infty} \leq R\}$ be the ball of radius R in $\|\cdot\|_{2,\infty}$. W.l.o.g., we assume the radius $R \geq 1$.

Lemma A.37 (Lipschitzness of the feedforward layer). *For a $J+1$ -layer feedforward (FF) network parameterized by $\boldsymbol{\theta}_{\text{ff}} = (W_1 \in \mathbb{R}^{D' \times (D+1)}, W_{J+1} \in \mathbb{R}^{D' \times (D'+1)}, \{W_j \in \mathbb{R}^{D' \times (D'+1)}\}_{2 \leq j \leq J})$, we introduce the norm (as in Eq. 2.18)*

$$\|\boldsymbol{\theta}_{\text{ff}}\| = \max_{j \in [J+1]} \|W_j\|_{\text{op}}.$$

Define the parameter space

$$\Theta_{\text{ff}, B} := \{\boldsymbol{\theta}_{\text{ff}} : \|\boldsymbol{\theta}_{\text{ff}}\| \leq B\}.$$

Then for $\mathbf{H} \in \mathcal{H}_R$, $\boldsymbol{\theta}_{\text{ff}} \in \Theta_{\text{ff}, B}$, the function $(\boldsymbol{\theta}_{\text{ff}}, \mathbf{H}) \mapsto \text{FF}_{\boldsymbol{\theta}_{\text{ff}}}(\mathbf{H}) + \mathbf{H}$ is $(J+1)B^J R$ -Lipschitz w.r.t. $\boldsymbol{\theta}_{\text{ff}}$ in $\|\cdot\|$ and $1 + B^{J+1}$ -Lipschitz w.r.t. $\mathbf{H}$ in $\|\cdot\|_{2,\infty}$.

Proof of Lemma A.37. By definition, for the i -th column $\mathbf{H}_i$ of the matrix $\mathbf{H} \in \mathbb{R}^{D \times N}$, we have³

$$\text{FF}_{\boldsymbol{\theta}_{\text{ff}}}(\mathbf{H}_i) = W_{J+1} \cdot \text{ReLU}(W_J \cdots \text{ReLU}(W_1 \cdot \mathbf{H}_i)).$$

³We incorporate the intercept term into the token matrix to simplify the notation.

Therefore, for $\boldsymbol{\theta}'_{\text{ff}} = (W'_{1:J+1}) \in \Theta_{\text{ff},B}$

$$\begin{aligned}
& \|\text{FF}_{\boldsymbol{\theta}_{\text{ff}}}(\mathbf{H}) + \mathbf{H} - \text{FF}_{\boldsymbol{\theta}'_{\text{ff}}}(\mathbf{H}) - \mathbf{H}\|_{2,\infty} \\
&= \max_i \|W_{J+1} \cdot \text{ReLU}(W_J \cdots \text{ReLU}(W_1 \cdot \mathbf{H}_i)) - W'_{J+1} \cdot \text{ReLU}(W'_J \cdots \text{ReLU}(W'_1 \cdot \mathbf{H}_i))\|_2 \\
&\leq \sum_{j=1}^{J+1} \max_i \|W'_{J+1} \cdots W'_{j+1} \text{ReLU}(W_j \cdots \text{ReLU}(W_1 \cdot \mathbf{H}_i)) - \\
&\quad W'_{J+1} \cdots W'_{j+1} \text{ReLU}(W'_j \cdots \text{ReLU}(W'_1 \cdot \mathbf{H}_i))\|_2 \\
&\stackrel{(i)}{\leq} \sum_{j=1}^{J+1} \max_i \|W'_{J+1} \cdots W'_{j+1}\|_{\text{op}} \cdot \|W_j - W'_j\|_{\text{op}} \cdot \|\text{ReLU}(W_{j-1} \cdots \text{ReLU}(W_1 \cdot \mathbf{H}_i))\|_2 \\
&\stackrel{(ii)}{\leq} B^J \mathbf{R} \cdot \left(\sum_{j=1}^{J+1} \|W_j - W'_j\|_{\text{op}} \right) \leq (J+1) B^J \mathbf{R} \cdot \|\boldsymbol{\theta}_{\text{ff}} - \boldsymbol{\theta}'_{\text{ff}}\|,
\end{aligned}$$

where steps (i) and (ii) use the fact that $\|\text{ReLU}(\mathbf{x}) - \text{ReLU}(\mathbf{y})\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2$. Similarly, for any matrices $\mathbf{H}, \mathbf{H}'$

$$\begin{aligned}
& \|\text{FF}_{\boldsymbol{\theta}_{\text{ff}}}(\mathbf{H}) + \mathbf{H} - \text{FF}_{\boldsymbol{\theta}_{\text{ff}}}(\mathbf{H}') - \mathbf{H}'\|_{2,\infty} \\
&= \max_i \|\mathbf{H}_i + W_{J+1} \cdot \text{ReLU}(W_J \cdots \text{ReLU}(W_1 \cdot \mathbf{H}_i)) - \mathbf{H}'_i - \\
&\quad W_{J+1} \cdot \text{ReLU}(W_J \cdots \text{ReLU}(W_1 \cdot \mathbf{H}'_i))\|_2 \\
&\leq \max_i \|\mathbf{H}_i - \mathbf{H}'_i\|_2 + \prod_{j=1}^{J+1} \|W_j\|_{\text{op}} \cdot \max_i \|\mathbf{H}_i - \mathbf{H}'_i\|_2 \leq (1 + B^{J+1}) \cdot \|\mathbf{H} - \mathbf{H}'\|_{2,\infty},
\end{aligned}$$

where the last line uses $\|\text{ReLU}(\mathbf{x}) - \text{ReLU}(\mathbf{y})\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2$. $\square$

Lemma A.38 (Lipschitzness of the attention layer). *For a single attention layer $\text{Attn}_{\boldsymbol{\theta}_{\text{Attn}}}(\cdot)$ parameterized by $\boldsymbol{\theta}_{\text{Attn}} = (W_Q, W_K, W_V)$, we introduce the norm*

$$\|\boldsymbol{\theta}_{\text{Attn}}\| = \max\{\|W_Q\|_{\text{op}}, \|W_K\|_{\text{op}}, \|W_V\|_{\text{op}}\},$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D \times D}$ are the query, key, value matrices. Define the parameter space

$$\Theta_{\text{Attn},B} := \{\boldsymbol{\theta}_{\text{Attn}} : \|\boldsymbol{\theta}_{\text{Attn}}\| \leq B\}.$$

Then for $\mathbf{H} \in \mathcal{H}_{\mathbf{R}}, \boldsymbol{\theta}_{\text{Attn}} \in \Theta_{\text{Attn},B}$, the function $(\boldsymbol{\theta}_{\text{Attn}}, \mathbf{H}) \mapsto \text{Attn}_{\boldsymbol{\theta}_{\text{Attn}}}(\mathbf{H})$ is $\mathbf{R}(1 + 4eB^2\mathbf{R}^2)$ -Lipschitz w.r.t. $\boldsymbol{\theta}_{\text{Attn}}$ in $\|\cdot\|$ and $1 + B(1 + 4eB^2\mathbf{R}^2)$ -Lipschitz w.r.t. $\mathbf{H}$ in $\|\cdot\|_{2,\infty}$.

Proof of Lemma A.38. Adopt the shorthand σ for the softmax activation. By definition, for any input $\mathbf{H} \in \mathbb{R}^{D' \times N}$, the output of attention $\text{Attn}_{\boldsymbol{\theta}_{\text{Attn}}}(\mathbf{H})$ is given by

$$\tilde{\mathbf{H}}_i := [\text{Attn}_{\boldsymbol{\theta}_{\text{Attn}}}(\mathbf{H})]_i + \mathbf{H}_i = \sum_{j=1}^N \sigma(\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle) \cdot W_V \mathbf{H}_j + \mathbf{H}_i, \quad \text{for } i \in [N].$$

Similarly, for $\theta'_{\text{Attn}} = (W'_Q, W'_K, W'_V)$, the output is given by

$$\tilde{H}'_i := [\text{Attn}_{\theta'_{\text{Attn}}}(\mathbf{H})]_i + \mathbf{H}_i = \sum_{j=1}^N \sigma(\langle W'_Q \mathbf{H}_i, W'_K \mathbf{H}_j \rangle) \cdot W'_V \mathbf{H}_j + \mathbf{H}_i, \quad \text{for } i \in [N].$$

Note that $\|(\text{Attn}_{\theta_{\text{Attn}}}(\mathbf{H}) + \mathbf{H}) - (\text{Attn}_{\theta'_{\text{Attn}}}(\mathbf{H}) + \mathbf{H})\|_{2,\infty} = \max_{i \in [N]} \|\tilde{H}_i - \tilde{H}'_i\|_2$. For any $i \in [N]$, we have

$$\begin{aligned} & \|\tilde{H}_i - \tilde{H}'_i\|_2 \\ &= \left\| \sum_{j=1}^N \sigma(\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle) \cdot W_V \mathbf{H}_j - \sum_{j=1}^N \sigma(\langle W'_Q \mathbf{H}_i, W'_K \mathbf{H}_j \rangle) \cdot W'_V \mathbf{H}_j \right\|_2 \\ &\leq \sum_{j=1}^N \|\sigma(\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle) W_V - \sigma(\langle W'_Q \mathbf{H}_i, W'_K \mathbf{H}_j \rangle) W'_V\|_{\text{op}} \cdot \|\mathbf{H}_j\|_2 \\ &\leq R \cdot \sum_{j=1}^N \left[\|\sigma(\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle) (W_V - W'_V)\|_{\text{op}} + \right. \\ &\quad \left. \|(\sigma(\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle) - \sigma(\langle W'_Q \mathbf{H}_i, W'_K \mathbf{H}_j \rangle)) W'_V\|_{\text{op}} \right] \\ &\leq U_{a1} + U_{a2}, \end{aligned}$$

where

$$U_{a1} := R \cdot \sum_{j=1}^N \sigma(\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle) \|W_V - W'_V\|_{\text{op}} \stackrel{(i)}{\leq} R \cdot \|W_V - W'_V\|_{\text{op}}, \quad (\text{A.131})$$

$$\begin{aligned} U_{a2} &:= R \cdot \sum_{j=1}^N |(\sigma(\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle) - \sigma(\langle W'_Q \mathbf{H}_i, W'_K \mathbf{H}_j \rangle))| \cdot \|W'_V\|_{\text{op}}, \\ &\stackrel{(ii)}{\leq} 2eBR \cdot \max_j |\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle - \langle W'_Q \mathbf{H}_i, W'_K \mathbf{H}_j \rangle| \\ &\stackrel{(iii)}{\leq} 2eB^2R^3 \cdot (\|W_Q - W'_Q\|_{\text{op}} + \|W_K - W'_K\|_{\text{op}}). \end{aligned} \quad (\text{A.132})$$

In the above equations, step (i) uses the property of softmax activation that $\sum_{j=1}^N \sigma(\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle) = 1$; step (ii) follows from Lemma A.47; step (iii) follows from a triangle inequality and the boundedness assumption on $\mathbf{H}, W_Q, W_K$, namely,

$$\begin{aligned} & \max_j |\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle - \langle W'_Q \mathbf{H}_i, W'_K \mathbf{H}_j \rangle| \\ &\leq \max_j |\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle - \langle W'_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle| + |\langle W'_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle - \langle W'_Q \mathbf{H}_i, W'_K \mathbf{H}_j \rangle| \\ &\leq \max_j \|\mathbf{H}_i\|_2 \|\mathbf{H}_j\|_2 \|W_K\|_{\text{op}} \cdot \|W_Q - W'_Q\|_{\text{op}} + \|\mathbf{H}_i\|_2 \|\mathbf{H}_j\|_2 \|W'_Q\|_{\text{op}} \cdot \|W_K - W'_K\|_{\text{op}} \end{aligned}$$

$$\leq BR^2(\|W_Q - W'_Q\|_{\text{op}} + \|W_K - W'_K\|_{\text{op}}).$$

Putting equation (A.131) and (A.132) together yields the Lipschitz continuity w.r.t. $\boldsymbol{\theta}$.

For token matrix $\mathbf{H}' \in \mathbb{R}^{D \times N}$, let $\bar{\mathbf{H}}' := \text{Attn}_{\boldsymbol{\theta}_{\text{Attn}}}(\mathbf{H}') + \mathbf{H}'$. Then

$$\bar{\mathbf{H}}'_i := [\text{Attn}_{\boldsymbol{\theta}_{\text{Attn}}}(\mathbf{H}')]_i + \mathbf{H}'_i = \sum_{j=1}^N \sigma(\langle W_Q \mathbf{H}'_i, W_K \mathbf{H}'_j \rangle) \cdot W_V \mathbf{H}'_j + \mathbf{H}'_i, \quad \text{for } i \in [N].$$

Similarly, we have $\|(\text{Attn}_{\boldsymbol{\theta}_{\text{Attn}}}(\mathbf{H}) + \mathbf{H}) - (\text{Attn}_{\boldsymbol{\theta}_{\text{Attn}}}(\mathbf{H}') + \mathbf{H}')\|_{2,\infty} = \max_{i \in [N]} \|\tilde{\mathbf{H}}_i - \bar{\mathbf{H}}'_i\|_2$. For any $i \in [N]$,

$$\begin{aligned} \|\tilde{\mathbf{H}}_i - \bar{\mathbf{H}}'_i\|_2 &= \left\| \sum_{j=1}^N \sigma(\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle) \cdot W_V \mathbf{H}_j + \mathbf{H}_i - \sum_{j=1}^N \sigma(\langle W_Q \mathbf{H}'_i, W_K \mathbf{H}'_j \rangle) \cdot W_V \mathbf{H}'_j - \mathbf{H}'_i \right\|_2 \\ &\leq \|\mathbf{H}_i - \mathbf{H}'_i\|_2 \sum_{j=1}^N \sigma(\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle) \cdot \|W_V\|_{\text{op}} \|\mathbf{H}_j - \mathbf{H}'_j\|_2 \\ &\quad + \sum_{j=1}^N |\sigma(\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle) - \sigma(\langle W_Q \mathbf{H}'_i, W_K \mathbf{H}'_j \rangle)| \cdot \|W_V\|_{\text{op}} \|\mathbf{H}'_j\|_2 \\ &\leq (1 + B) \cdot \|\mathbf{H} - \mathbf{H}'\|_{2,\infty} + 2eBR \cdot \max_j |\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle - \langle W_Q \mathbf{H}'_i, W_K \mathbf{H}'_j \rangle| \\ &\leq (1 + B(1 + 4eB^2R^2)) \cdot \|\mathbf{H} - \mathbf{H}'\|_{2,\infty}, \end{aligned}$$

where the last line follows from

$$\begin{aligned} &\max_j |\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle - \langle W_Q \mathbf{H}'_i, W_K \mathbf{H}'_j \rangle| \\ &\leq \max_j |\langle W_Q \mathbf{H}_i, W_K \mathbf{H}_j \rangle - \langle W_Q \mathbf{H}'_i, W_K \mathbf{H}_j \rangle| + |\langle W_Q \mathbf{H}'_i, W_K \mathbf{H}_j \rangle - \langle W_Q \mathbf{H}'_i, W_K \mathbf{H}'_j \rangle| \\ &\leq \max_j \|\mathbf{H}_j\|_2 \|W_Q\|_{\text{op}} \|W_K\|_{\text{op}} \cdot \|\mathbf{H}_i - \mathbf{H}'_i\|_2 + \|\mathbf{H}'_i\|_2 \|W_Q\|_{\text{op}} \|W_K\|_{\text{op}} \cdot \|\mathbf{H}_j - \mathbf{H}'_j\|_2 \\ &\leq 2B^2R \cdot \|\mathbf{H} - \mathbf{H}'\|_{2,\infty}. \end{aligned}$$

□

Lemma A.39 (Lipschitzness of the transformer layer). *Consider the parameter space of transformer blocks*

$$\Theta_{\text{tf},B} := \{\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{ff}}, \boldsymbol{\theta}_{\text{Attn}}), \|\boldsymbol{\theta}\| \leq B\},$$

where $\|\cdot\|$ is defined in Eq. (2.18). Let $\text{TF}(\cdot) : \mathbb{R}^{D \times N} \mapsto \mathbb{R}^{D \times N}$ denote the transformer consists of one attention layer (with normalization) and one $J + 1$ -layer feedforward map, i.e.,

$$\text{TF}_{\boldsymbol{\theta}_{\text{tf}}}(\mathbf{H}) = \text{normalize}(\text{Attn}_{\boldsymbol{\theta}_{\text{Attn}}}(\bar{\mathbf{H}}) + \bar{\mathbf{H}}), \quad \text{where } \bar{\mathbf{H}} = \text{FF}_{\boldsymbol{\theta}_{\text{ff}}}(\mathbf{H}) + \mathbf{H}.$$

Assume $B, R \geq 1$. Then for $\mathbf{H} \in \mathcal{H}_R$, $\boldsymbol{\theta}_{\text{Attn}} \in \Theta_{\text{tf},B}$, the function $(\boldsymbol{\theta}_{\text{tf}}, \mathbf{H}) \mapsto \text{TF}_{\boldsymbol{\theta}_{\text{tf}}}(\mathbf{H})$ is $B_{\text{TF}}(R)$ -Lipschitz w.r.t. $\boldsymbol{\theta}_{\text{Attn}}$ in $\|\cdot\|$ and $B_{\text{TF}}(R)$ -Lipschitz w.r.t. $\mathbf{H}$ in $\|\cdot\|_{2,\infty}$, where $B_{\text{TF}}(R) := (cB)^{3J+6} \sqrt{SR^3}$ for some numerical constant $c > 0$.

Proof of Lemma A.39. For any $\theta_{\text{tf}}, \theta'_{\text{tf}} \in \Theta_{\text{tf}, B}$, let $\bar{H} = \text{FF}_{\theta_{\text{ff}}}(\text{H}) + \text{H}$ and $\bar{H}' = \text{FF}_{\theta'_{\text{ff}}}(\text{H}) + \text{H}$. We have

$$\|\bar{H} - \bar{H}'\|_{2,\infty} = \|\text{FF}_{\theta_{\text{ff}}}(\text{H}) - \text{FF}_{\theta'_{\text{ff}}}(\text{H})\|_{2,\infty} \leq (J+1)B^J R \cdot \|\theta_{\text{ff}} - \theta'_{\text{ff}}\|,$$

where the last step uses Lemma A.37. Adopt the shorthand $\text{norm}(\cdot)$ for $\text{normalize}(\cdot)$. Moreover, by the definition of $\text{FF}(\cdot)$, it can be verified that $\|\bar{H}\|_{2,\infty}, \|\bar{H}'\|_{2,\infty} \leq (B^{J+1} + 1)R$. Therefore,

$$\begin{aligned} & \|\text{TF}_{\theta_{\text{tf}}}(\text{H}) - \text{TF}_{\theta'_{\text{tf}}}(\text{H})\|_{2,\infty} \\ & \leq \|\text{norm}(\text{Attn}_{\theta_{\text{Attn}}}(\bar{H}) + \bar{H}) - \text{norm}(\text{Attn}_{\theta'_{\text{Attn}}}(\bar{H}') + \bar{H}')\|_{2,\infty} \\ & \leq 2\sqrt{S} \|\text{Attn}_{\theta_{\text{Attn}}}(\bar{H}) + \bar{H} - \text{Attn}_{\theta'_{\text{Attn}}}(\bar{H}') + \bar{H}'\|_{2,\infty} \\ & \leq 2\sqrt{S} \cdot [\|\text{Attn}_{\theta_{\text{Attn}}}(\bar{H}) - \text{Attn}_{\theta'_{\text{Attn}}}(\bar{H})\|_{2,\infty} \\ & \quad + \|\text{Attn}_{\theta'_{\text{Attn}}}(\bar{H}) - \text{Attn}_{\theta'_{\text{Attn}}}(\bar{H}')\|_{2,\infty} + \|\bar{H} - \bar{H}'\|_{2,\infty}] \\ & \leq 2\sqrt{S}\bar{R}(1 + 4eB^2\bar{R}^2) \cdot \|\theta_{\text{Attn}} - \theta'_{\text{Attn}}\| + 2\sqrt{S}(2 + B(1 + 4eB^2\bar{R}^2)) \cdot \|\bar{H} - \bar{H}'\|_{2,\infty}, \end{aligned}$$

where $\bar{R} := (B^{J+1} + 1)R$ and the third inequality uses Lemma A.38. Putting pieces together yields

$$\|\text{TF}_{\theta_{\text{tf}}}(\text{H}) - \text{TF}_{\theta'_{\text{tf}}}(\text{H})\|_{2,\infty} \leq B_{\text{TF}}(R) \cdot \|\theta_{\text{tf}} - \theta'_{\text{tf}}\|.$$

Similarly, for any $\text{H}, \text{H}' \in \mathcal{H}_R$ and $\theta_{\text{tf}} \in \Theta_{\text{tf}, B}$, let $\tilde{H} = \text{FF}_{\theta_{\text{ff}}}(\text{H}) + \text{H}$ and $\tilde{H}' = \text{FF}_{\theta_{\text{ff}}}(\text{H}') + \text{H}'$. Then

$$\begin{aligned} \|\tilde{H} - \tilde{H}'\|_{2,\infty} &= \|\text{FF}_{\theta_{\text{ff}}}(\text{H}) + \text{H} - \text{FF}_{\theta_{\text{ff}}}(\text{H}') - \text{H}'\|_{2,\infty} \\ &\leq (1 + B^{J+1}) \cdot \|\text{H} - \text{H}'\|_{2,\infty}, \end{aligned}$$

where the last step uses Lemma A.37. Moreover, basic algebra gives $\|\tilde{H}\|_{2,\infty}, \|\tilde{H}'\|_{2,\infty} \leq (B^{J+1} + 1)R$. We thus have

$$\begin{aligned} \|\text{TF}_{\theta_{\text{tf}}}(\tilde{H}) - \text{TF}_{\theta_{\text{tf}}}(\tilde{H}')\|_{2,\infty} &\leq \|\text{norm}(\text{Attn}_{\theta_{\text{Attn}}}(\tilde{H}) + \tilde{H}) - \text{norm}(\text{Attn}_{\theta_{\text{Attn}}}(\tilde{H}') + \tilde{H}')\|_{2,\infty} \\ &\leq 2\sqrt{S} \cdot [\|\text{Attn}_{\theta_{\text{Attn}}}(\tilde{H}) - \text{Attn}_{\theta_{\text{Attn}}}(\tilde{H}')\|_{2,\infty} + \|\tilde{H} - \tilde{H}'\|_{2,\infty}] \\ &\leq 2\sqrt{S}(2 + B(1 + 4eB^2\tilde{R}^2)) \cdot \|\tilde{H} - \tilde{H}'\|_{2,\infty} \\ &\leq B_{\text{TF}}(R) \cdot \|\text{H} - \text{H}'\|_{2,\infty}, \end{aligned}$$

where $\tilde{R} := (B^{J+1} + 1)R$ and the third inequality uses Lemma A.38. □

Lemma A.40 (Lipschitzness of the transformer). *Consider the space*

$$\Theta_{\text{tf}, L, B} := \{\theta = (\theta_{\text{ff}}^{(1:L)}, \theta_{\text{Attn}}^{(1:L)}), \|\theta\| \leq B\},$$

where $\|\cdot\|$ is as defined in Eq. (2.18). Let $\text{NN}_{\text{im}}^{\boldsymbol{\theta}}(\cdot) : [S]^{d_{\text{im}}} \mapsto \mathbb{R}^{S \times N}$ denote the image network that consists of L transformer blocks in Lemma A.39 and the embedding function $\text{Emb}(\cdot)$, i.e.,

$$\text{NN}_{\text{im}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im}}) = \text{TF}_{\boldsymbol{\theta}(1:L)}(\text{Emb}(\mathbf{x}_{\text{im}})).$$

Then for $\boldsymbol{\theta} \in \Theta_{\text{tf},L,B}$, the function $\mathbf{x}_{\text{im}} \mapsto \text{NN}_{\text{im}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im}})$ is $B_{\text{NN}} := ((cB)^{18JL}S^4)^L$ -Lipschitz w.r.t. $\boldsymbol{\theta}$ in $\|\cdot\|$ for some numerical constant $c > 0$. Moreover, let $\mathbf{H} := \text{Emb}(\mathbf{x}_{\text{im}})$. Then the function $\text{TF}_{\boldsymbol{\theta}(1:L)}(\mathbf{H})$ is B_{NN} -Lipschitz w.r.t. $\mathbf{H}$ in $\|\cdot\|_{2,\infty}$. Same results hold for the text network $\text{NN}_{\text{tx}}^{\boldsymbol{\theta}}(\cdot)$.

Proof of Lemma A.40. Let $\mathbf{H} = \text{Emb}(\mathbf{x}_{\text{im}})$ and $\mathbf{R} = \mathbf{R}_L = S\sqrt{L}$. For $0 \leq i \leq L-1$, define $\mathbf{R}_i := (2B)^{i(J+2)}\mathbf{R}$. Then it can be verified by induction that for any $0 \leq \ell \leq L$

$$\|\text{TF}_{\boldsymbol{\theta}(L-\ell+1:L)}(\mathbf{H})\|_{2,\infty} \leq \mathbf{R}_{L-\ell}$$

for any $\boldsymbol{\theta} \in \Theta_{\text{tf},L,B}$, $\mathbf{H} \in \mathcal{H}_{\mathbf{R}}$ and $\ell \in [L]$. With this bound at hand, for any $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_{\text{tf},L,B}$

$$\begin{aligned} & \|\text{NN}_{\text{im}}^{\boldsymbol{\theta}}(\mathbf{H}) - \text{NN}_{\text{im}}^{\tilde{\boldsymbol{\theta}}}(\mathbf{H})\|_{2,\infty} \\ &= \|\text{TF}_{\boldsymbol{\theta}}(\mathbf{H}) - \text{TF}_{\tilde{\boldsymbol{\theta}}}(\mathbf{H})\|_{2,\infty} \\ &\stackrel{(i)}{\leq} \sum_{\ell=1}^L \|\text{TF}_{\boldsymbol{\theta}(1:\ell-1)}(\text{TF}_{\boldsymbol{\theta}(\ell)}(\text{TF}_{\tilde{\boldsymbol{\theta}}(\ell+1:L)}(\mathbf{H}))) - \text{TF}_{\boldsymbol{\theta}(1:\ell-1)}(\text{TF}_{\tilde{\boldsymbol{\theta}}(\ell)}(\text{TF}_{\tilde{\boldsymbol{\theta}}(\ell+1:L)}(\mathbf{H})))\|_{2,\infty} \\ &\stackrel{(ii)}{\leq} B \sum_{\ell=1}^L \prod_{j=1}^{\ell} B_{\text{TF}}(\mathbf{R}_\ell) \cdot \|\boldsymbol{\theta}^{(\ell)} - \tilde{\boldsymbol{\theta}}^{(\ell)}\| \leq LB \cdot \prod_{j=1}^L B_{\text{TF}}(\mathbf{R}_j) \cdot \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|, \end{aligned}$$

where step (i) follows from a triangle inequality and step (ii) uses Lemma A.39. Plugging in the definition of $B_{\text{TF}}(\cdot)$ yields the desired bound.

Similarly, for two embedding matrices $\mathbf{H}, \bar{\mathbf{H}}$, we have

$$\begin{aligned} \|\text{NN}_{\text{im}}^{\boldsymbol{\theta}}(\mathbf{H}) - \text{NN}_{\text{im}}^{\boldsymbol{\theta}}(\bar{\mathbf{H}})\|_{2,\infty} &= \|\text{TF}_{\boldsymbol{\theta}}(\mathbf{H}) - \text{TF}_{\boldsymbol{\theta}}(\bar{\mathbf{H}})\|_{2,\infty} \\ &\leq \prod_{\ell=1}^L B_{\text{TF}}(\mathbf{R}_\ell) \cdot \|\mathbf{H} - \bar{\mathbf{H}}\|_{2,\infty} \leq B_{\text{NN}} \cdot \|\mathbf{H} - \bar{\mathbf{H}}\|_{2,\infty}, \end{aligned}$$

where the last line follows from Lemma A.39 and the definition of B_{NN} . □

Lemma A.41 (Properties of the score function). *Consider the space*

$$\Theta_{S,L,B} := \{\boldsymbol{\theta} = (\mathbf{W}_{\text{im}}, \mathbf{W}_{\text{tx}}, w), \|\boldsymbol{\theta}\| \leq B\},$$

where $\|\cdot\|$ is defined in Eq. (2.18). Let the score function

$$\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \tau^w(\text{softmax}(\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}(\mathbf{x}_{\text{im}})), \text{softmax}(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))).$$

Then for $\boldsymbol{\theta} \in \Theta_{\text{tf},L,B}$, the function $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \mapsto \mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$ is $B_S := ((cB)^{18JL}S^4)^{L+1}$ -Lipschitz w.r.t. $\boldsymbol{\theta}$ in $\|\cdot\|$ for all fixed $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \in \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}}$ for some numerical constant $c > 0$. Moreover, $\exp(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) \in [1/c_1, c_1]$ with $c_1 = \exp(B_{\text{read}})$.

Proof of Lemma A.41. Use $\sigma(\cdot)$ as a shorthand notation for $\text{softmax}(\cdot)$ and let

$$\tilde{\tau}^w(\mathbf{x}, \mathbf{y}) := \exp(\tau^w(\mathbf{x}, \mathbf{y})) = \text{trun}\left(\sum_{s=1}^S w_s x_s y_s\right),$$

where we recall $\text{trun}(z) := \text{proj}_{[\exp(-B_{\text{read}}), \exp(B_{\text{read}})]}(z)$. For two parameters $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_{S,L,B}$, we have

$$\begin{aligned} & |\exp(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})) - \exp(\mathbf{S}_{\text{NN}}^{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))| \\ &= |\tilde{\tau}^w(\sigma(\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}(\mathbf{x}_{\text{im}})), \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))) - (\tilde{\tau}^{\tilde{w}}(\sigma(\text{NN}_{\text{im}}^{\tilde{\mathbf{W}}_{\text{im}}}(\mathbf{x}_{\text{im}})), \sigma(\text{NN}_{\text{tx}}^{\tilde{\mathbf{W}}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))))| \\ &\leq \tilde{T}_1 + \tilde{T}_2 + \tilde{T}_3, \end{aligned}$$

where

$$\begin{aligned} \tilde{T}_1 &:= |\tilde{\tau}^w(\sigma(\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}(\mathbf{x}_{\text{im}})), \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))) - \tilde{\tau}^{\tilde{w}}(\sigma(\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}(\mathbf{x}_{\text{im}})), \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))))| \\ &\leq \|w - \tilde{w}\|_{\infty} \cdot \|\sigma(\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}(\mathbf{x}_{\text{im}}))\|_1 \cdot \|\sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))\|_1 \leq \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|, \\ \tilde{T}_2 &:= |\tilde{\tau}^{\tilde{w}}(\sigma(\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}(\mathbf{x}_{\text{im}})), \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))) - \tilde{\tau}^{\tilde{w}}(\sigma(\text{NN}_{\text{im}}^{\tilde{\mathbf{W}}_{\text{im}}}(\mathbf{x}_{\text{im}})), \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))))| \\ &\leq \|\tilde{w}\|_{\infty} \cdot \|\sigma(\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}(\mathbf{x}_{\text{im}})) - \sigma(\text{NN}_{\text{im}}^{\tilde{\mathbf{W}}_{\text{im}}}(\mathbf{x}_{\text{im}}))\|_1 \cdot \|\sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))\|_{\infty} \\ &\stackrel{(i)}{\leq} 2eB \cdot \|\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}(\mathbf{x}_{\text{im}}) - \text{NN}_{\text{im}}^{\tilde{\mathbf{W}}_{\text{im}}}(\mathbf{x}_{\text{im}})\|_2 \leq 2eB \cdot B_{\text{NN}} \cdot \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|, \\ \tilde{T}_3 &:= |\tilde{\tau}^{\tilde{w}}(\sigma(\text{NN}_{\text{im}}^{\tilde{\mathbf{W}}_{\text{im}}}(\mathbf{x}_{\text{im}})), \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))) - \tilde{\tau}^{\tilde{w}}(\sigma(\text{NN}_{\text{im}}^{\tilde{\mathbf{W}}_{\text{im}}}(\mathbf{x}_{\text{im}})), \sigma(\text{NN}_{\text{tx}}^{\tilde{\mathbf{W}}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))))| \\ &\leq \|\tilde{w}\|_{\infty} \cdot \|\sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}})) - \sigma(\text{NN}_{\text{tx}}^{\tilde{\mathbf{W}}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))\|_1 \cdot \|\sigma(\text{NN}_{\text{im}}^{\tilde{\mathbf{W}}_{\text{im}}}(\mathbf{x}_{\text{im}}))\|_{\infty} \\ &\stackrel{(ii)}{\leq} 2eB \cdot \|\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}) - \text{NN}_{\text{tx}}^{\tilde{\mathbf{W}}_{\text{tx}}}(\mathbf{x}_{\text{tx}})\|_2 \leq 2eB \cdot B_{\text{NN}} \cdot \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|, \end{aligned}$$

where step (i) and (ii) uses Lemma A.47. Putting pieces together we find that $\exp(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))$ is $(1 + 4e)BB_{\text{NN}}$ -Lipschitz continuous in $\boldsymbol{\theta}$ in $\|\cdot\|$.

The upper and lower bounds on $\exp(\mathbf{S}_{\text{NN}}^{\boldsymbol{\theta}}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}))$ follows immediately from the definition of the readout function $\text{trun}(\cdot)$ in Eq. (A.57). □

Lemma A.42 (Lipschitzness of the CLIP representation). *Consider the space*

$$\Theta_{\text{clip}, L, B} := \{\boldsymbol{\theta} = (W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}, \mathbf{W}_{\text{tx}}), \|\boldsymbol{\theta}\| \leq B\},$$

where $\|\cdot\|$ is defined in Eq. (2.18). Let $\sigma(\cdot)$ denote the softmax function. Under the definition of $\text{Adap}(\cdot)$ in Eq. (2.20), with slight abuse of notation, let the CLIP representation of the text data

$$\widehat{\mathbf{E}}_{\text{tx}, \boldsymbol{\theta}}(\mathbf{x}_{\text{tx}}) = \text{Adap}(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}})) = W_{\text{ada}}^{(1)} \sigma(\log(\text{trun}(W_{\text{ada}}^{(2)} \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))))),$$

where $\text{trun}(\cdot)$ is the truncation function defined in Eq. (A.57). Then for $\boldsymbol{\theta} \in \Theta_{\text{clip},L,B}$, the function $\widehat{\mathbf{E}}_{\text{tx},\boldsymbol{\theta}}(\mathbf{x}_{\text{tx}})$ is $B_{\text{Adap}} := \exp(B_{\text{read}})((cB)^{18JL}S^4)^{L+1}$ -Lipschitz w.r.t. $\boldsymbol{\theta}$ in $\|\cdot\|$ for all $\mathbf{x}_{\text{tx}} \in \mathcal{X}_{\text{tx}}$ for some numerical constant $c > 0$, where $B_{\text{read}} = 4m \log B_{\psi}$. Moreover, $\widehat{\mathbf{E}}_{\text{tx},\boldsymbol{\theta}} : \mathbb{R}^{d_{\text{tx}}} \mapsto \mathbb{R}^S$ is 1-Lipschitz w.r.t. $W_{\text{ada}}^{(1)}$ and $2eB \exp(B_{\text{read}})$ -Lipschitz w.r.t. $W_{\text{ada}}^{(2)}$ in $\|\cdot\|_{\text{op}}$. Same results hold for the adapter of the image representation.

Proof of Lemma A.42. For two parameters $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}$, we have

$$\|\widehat{\mathbf{E}}_{\text{tx},\boldsymbol{\theta}}(\mathbf{x}_{\text{tx}}) - \widehat{\mathbf{E}}_{\text{tx},\tilde{\boldsymbol{\theta}}}(\mathbf{x}_{\text{tx}})\|_2 \leq \tilde{T}_1 + \tilde{T}_2 + \tilde{T}_3,$$

where

$$\begin{aligned} \tilde{T}_1 &:= \left\| (W_{\text{ada}}^{(1)} - \widetilde{W_{\text{ada}}^{(1)}}) \sigma(\log(\text{trun}(W_{\text{ada}}^{(2)} \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}})))))) \right\|_2, \\ \tilde{T}_2 &:= \left\| \widetilde{W_{\text{ada}}^{(1)}} \left(\sigma(\log(\text{trun}(\widetilde{W_{\text{ada}}^{(2)}} \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}})))))) \right) - \sigma(\log(\text{trun}(W_{\text{ada}}^{(2)} \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}})))))) \right\|_2, \\ \tilde{T}_3 &:= \left\| \widetilde{W_{\text{ada}}^{(1)}} \left(\sigma(\log(\text{trun}(\widetilde{W_{\text{ada}}^{(2)}} \sigma(\text{NN}_{\text{tx}}^{\widetilde{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}})})))) \right) - \sigma(\log(\text{trun}(\widetilde{W_{\text{ada}}^{(2)}} \sigma(\text{NN}_{\text{tx}}^{\widetilde{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}})})))) \right\|_2. \end{aligned}$$

By properties of the softmax function, we have

$$\begin{aligned} \tilde{T}_1 &\leq \|W_{\text{ada}}^{(1)} - \widetilde{W_{\text{ada}}^{(1)}}\|_{\text{op}} \cdot \|\sigma(\log(\text{trun}(W_{\text{ada}}^{(2)} \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))))))\|_2 \\ &\leq \|W_{\text{ada}}^{(1)} - \widetilde{W_{\text{ada}}^{(1)}}\|_{\text{op}} \cdot \|\sigma(\log(\text{trun}(W_{\text{ada}}^{(2)} \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))))))\|_1 = \|W_{\text{ada}}^{(1)} - \widetilde{W_{\text{ada}}^{(1)}}\|_2, \quad \text{and} \\ \tilde{T}_2 &\leq \|\widetilde{W_{\text{ada}}^{(1)}}\|_{\text{op}} \cdot \left\| \sigma(\log(\text{trun}(\widetilde{W_{\text{ada}}^{(2)}} \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}})))))) - \sigma(\log(\text{trun}(W_{\text{ada}}^{(2)} \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}})))))) \right\|_2 \\ &\stackrel{(i)}{\leq} 2eB \cdot \|\log(\text{trun}(W_{\text{ada}}^{(2)} \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}})))) - \log(\text{trun}(\widetilde{W_{\text{ada}}^{(2)}} \sigma(\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}))))\|_{\infty} \\ &\stackrel{(ii)}{\leq} 2eB \exp(B_{\text{read}}) \cdot \max_{j \in [S]} \|W_{\text{ada},j}^{(2)} - \widetilde{W_{\text{ada},j}^{(2)}}\|_2, \end{aligned}$$

where step (i) uses Lemma A.47 and step (ii) follows from the definition of $\text{trun}(\cdot)$. Similarly, we have

$$\tilde{T}_3 \leq 2eB^2 \exp(B_{\text{read}}) \cdot \|\text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}) - \text{NN}_{\text{tx}}^{\widetilde{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}})}\|_2 \leq 2eB^2 \exp(B_{\text{read}}) \cdot B_{\text{NN}} \cdot \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|$$

by Lemma A.40. Putting the bounds on $\tilde{T}_1, \tilde{T}_2, \tilde{T}_3$ together yields Lemma A.42. $\square$

Lemma A.43 (Lipschitzness of the conditional diffusion model). *Consider the space*

$$\Theta_{\text{cdm},L,B} := \{\boldsymbol{\theta} = (W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}, \mathbf{W}_{\text{cdm}}), \|\boldsymbol{\theta}\| \leq B\},$$

where $\|\cdot\|$ is defined in Eq. (2.21). Let

$$\mathbf{M}_t^{\boldsymbol{\theta}}(\mathbf{z}_t, \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})) = \text{read}_{\text{cdm}}(\text{TF}_{\text{cdm}}(\text{Emb}_{\text{cdm}}(\mathbf{z}_t, \text{Adap}(\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}))))).$$

Then for $\theta \in \Theta_{\text{cdm},L,B}$, the function $M_t^\theta(z_t, E_{\text{tx}}(\mathbf{x}_{\text{tx}}))$ is B_{cdm} -Lipschitz w.r.t. θ in $\|\cdot\|$, where $B_{\text{cdm}} := ((cB)^{18JL} S^9 B_{\text{read}}^3 \log^3 \bar{m})^{2L+2} \exp(2B_{\text{read}})$ and $B_{\text{read}} = 4\bar{m} \log B_\psi + \log S$ for some numerical constant $c > 0$.

Proof of Lemma A.43. For any two parameters $\theta, \tilde{\theta} \in \Theta_{\text{cdm},L,B}$, we have

$$\begin{aligned} & \|M_t^\theta(z_t, E_{\text{tx}}(\mathbf{x}_{\text{tx}})) - M_t^{\tilde{\theta}}(z_t, E_{\text{tx}}(\mathbf{x}_{\text{tx}}))\|_2 \\ & \stackrel{(i)}{\leq} cS \cdot \sqrt{d_{\text{im}}} \cdot \|\text{TF}_{\text{cdm}}^{\mathbf{W}_{\text{cdm}}}(\text{Emb}_{\text{cdm}}(z_t, \text{Adap}^{W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}}(E_{\text{tx}}(\mathbf{x}_{\text{tx}})))) \\ & \quad - \text{TF}_{\text{cdm}}^{\widetilde{\mathbf{W}}_{\text{cdm}}}(\text{Emb}_{\text{cdm}}(z_t, \text{Adap}^{\widetilde{W}_{\text{ada}}^{(1)}, \widetilde{W}_{\text{ada}}^{(2)}}(E_{\text{tx}}(\mathbf{x}_{\text{tx}}))))\|_{2,\infty} \\ & \leq \tilde{T}_1 + \tilde{T}_2 \end{aligned}$$

for some numerical constant $c > 0$, where step (i) uses the definition of read_{cdm} in Eq. (A.87) and Lemma A.47, and

$$\begin{aligned} \tilde{T}_1 &:= \|\text{TF}_{\text{cdm}}^{\mathbf{W}_{\text{cdm}}}(\text{Emb}_{\text{cdm}}(z_t, \text{Adap}^{W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}}(E_{\text{tx}}(\mathbf{x}_{\text{tx}})))) \\ & \quad - \text{TF}_{\text{cdm}}^{\widetilde{\mathbf{W}}_{\text{cdm}}}(\text{Emb}_{\text{cdm}}(z_t, \text{Adap}^{W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}}(E_{\text{tx}}(\mathbf{x}_{\text{tx}}))))\|_{2,\infty} \\ & \stackrel{(ii)}{\leq} ((cB)^{18JL} S^9 B_{\text{read}}^3 \log^3 \bar{m})^{2L+1} \|\mathbf{W}_{\text{cdm}} - \widetilde{\mathbf{W}}_{\text{cdm}}\|, \\ \tilde{T}_2 &:= \|\text{TF}_{\text{cdm}}^{\widetilde{\mathbf{W}}_{\text{cdm}}}(\text{Emb}_{\text{cdm}}(z_t, \text{Adap}^{W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}}(E_{\text{tx}}(\mathbf{x}_{\text{tx}})))) \\ & \quad - \text{TF}_{\text{cdm}}^{\widetilde{\mathbf{W}}_{\text{cdm}}}(\text{Emb}_{\text{cdm}}(z_t, \text{Adap}^{\widetilde{W}_{\text{ada}}^{(1)}, \widetilde{W}_{\text{ada}}^{(2)}}(E_{\text{tx}}(\mathbf{x}_{\text{tx}}))))\|_{2,\infty} \\ & \stackrel{(iii)}{\leq} ((cB)^{18JL} S^9 B_{\text{read}}^3 \log^3 \bar{m})^{2L+1} \exp(B_{\text{read}}) \\ & \quad \cdot \|\text{Adap}^{W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}}(E_{\text{tx}}(\mathbf{x}_{\text{tx}})) - \text{Adap}^{\widetilde{W}_{\text{ada}}^{(1)}, \widetilde{W}_{\text{ada}}^{(2)}}(E_{\text{tx}}(\mathbf{x}_{\text{tx}}))\|_{2,\infty} \\ & \stackrel{(iv)}{\leq} ((cB)^{18JL} S^9 B_{\text{read}}^3 \log^3 \bar{m})^{2L+2} \exp(2B_{\text{read}}) \cdot \|\theta - \tilde{\theta}\|, \end{aligned}$$

where step (ii) uses Lemma A.40 and note that $\|\text{Emb}_{\text{cdm}}(z_t, \text{Adap}(E_{\text{tx}}(\mathbf{x}_{\text{tx}})))\|_{2,\infty} \leq R := cS^{2.5} \log \bar{m} L B_{\text{read}}$ for some numerical constant $c > 0$, step (iii) follows from Lemma A.40 and the definition of Emb_{cdm} , and step (iv) uses Lemma A.42. Putting pieces together yields Lemma A.43. $\square$

Lemma A.44 (Lipschitzness of the vision-language model). *Consider the space*

$$\Theta_{\text{vlm},L,B} := \{\theta = (W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}, \mathbf{W}_{\text{vlm}}), \|\theta\| \leq B\},$$

where $\|\cdot\|$ is defined in Eq. (2.21). For any $j \in [d_{\text{tx}}]$, let

$$\log \mu^\theta(x_{\text{tx},j} | x_{\text{tx},1:j-1}, E_{\text{im}}(\mathbf{x}_{\text{im}})) = \log \circ \text{read}_{\text{vlm}}(\text{TF}_{\text{vlm}}(\text{Emb}_{\text{vlm}}(x_{\text{tx},1:j-1}, \text{Adap}(E_{\text{im}}(\mathbf{x}_{\text{im}}))))).$$

Then for $\boldsymbol{\theta} \in \Theta_{\text{vlm},L,B}$, the function $\log \mu^{\boldsymbol{\theta}}(x_{\text{tx},j}|x_{\text{tx},1:j-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))$ is B_{vlm} -Lipschitz w.r.t. $\boldsymbol{\theta}$ in $\|\cdot\|$ for all $(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \in \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}}$, where $B_{\text{vlm}} := ((cB)^{18JL} S^4 B_{\text{read}}^3)^{4L+3} \exp(2B_{\text{read}})$ and $B_{\text{read}} = 4\underline{m} \log B_{\psi} + \log S$ for some numerical constant $c > 0$.

Proof of Lemma A.44. Similarly to the proof of Lemma A.43, for any two parameters $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_{\text{vlm},L,B}$, we have

$$\begin{aligned} & |\log \mu^{\boldsymbol{\theta}}(x_{\text{tx},j}|x_{\text{tx},1:j-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) - \log \mu^{\tilde{\boldsymbol{\theta}}}(x_{\text{tx},j}|x_{\text{tx},1:j-1}, \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))| \\ & \stackrel{(i)}{\leq} \|\text{TF}_{\text{vlm}}^{\mathbf{W}_{\text{vlm}}}(\text{Emb}_{\text{vlm}}(x_{\text{tx},1:j-1}, \text{Adap}^{W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})))) \\ & \quad - \text{TF}_{\text{cdm}}^{\widetilde{\mathbf{W}}_{\text{cdm}}}(\text{Emb}_{\text{cdm}}(x_{\text{tx},1:j-1}, \text{Adap}^{\widetilde{W}_{\text{ada}}^{(1)}, \widetilde{W}_{\text{ada}}^{(2)}}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))))\|_{2,\infty} \\ & \leq \tilde{T}_3 + \tilde{T}_4 \end{aligned}$$

for some numerical constant $c > 0$, where step (i) uses the definition of read_{vlm} and Lemma A.48, and

$$\begin{aligned} \tilde{T}_3 &:= \|\text{TF}_{\text{vlm}}^{\mathbf{W}_{\text{vlm}}}(\text{Emb}_{\text{vlm}}(x_{\text{tx},1:j-1}, \text{Adap}^{W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})))) \\ & \quad - \text{TF}_{\text{vlm}}^{\widetilde{\mathbf{W}}_{\text{vlm}}}(\text{Emb}_{\text{vlm}}(x_{\text{tx},1:j-1}, \text{Adap}^{W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))))\|_{2,\infty} \\ & \stackrel{(ii)}{\leq} ((cB)^{18JL} S^4 B_{\text{read}}^3)^{4L+2} \|\mathbf{W}_{\text{vlm}} - \widetilde{\mathbf{W}}_{\text{vlm}}\|, \\ \tilde{T}_4 &:= \|\text{TF}_{\text{vlm}}^{\widetilde{\mathbf{W}}_{\text{cdm}}}(\text{Emb}_{\text{vlm}}(x_{\text{tx},1:j-1}, \text{Adap}^{W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})))) \\ & \quad - \text{TF}_{\text{vlm}}^{\widetilde{\mathbf{W}}_{\text{vlm}}}(\text{Emb}_{\text{vlm}}(x_{\text{tx},1:j-1}, \text{Adap}^{\widetilde{W}_{\text{ada}}^{(1)}, \widetilde{W}_{\text{ada}}^{(2)}}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))))\|_{2,\infty} \\ & \stackrel{(iii)}{\leq} ((cB)^{18JL} S^4 B_{\text{read}}^3)^{4L+2} \exp(B_{\text{read}}) \\ & \quad \cdot \|\text{Adap}^{W_{\text{ada}}^{(1)}, W_{\text{ada}}^{(2)}}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})) - \text{Adap}^{\widetilde{W}_{\text{ada}}^{(1)}, \widetilde{W}_{\text{ada}}^{(2)}}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}))\|_{2,\infty} \\ & \stackrel{(iv)}{\leq} ((cB)^{18JL} S^4 B_{\text{read}}^3)^{4L+3} \exp(2B_{\text{read}}) \cdot \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|, \end{aligned}$$

where step (ii) follows from a modified version of Lemma A.40 (note that one layer of transformer used in VLMs can be represented by two layers of transformer used in Lemma A.40), and the fact that

$$\|\text{Emb}_{\text{vlm}}(x_{\text{tx},1:j-1}, \text{Adap}(\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})))\|_{2,\infty} \leq R := cSB_{\text{read}}$$

for some numerical constant $c > 0$, step (iii) follows from Lemma A.40 and the definition of Emb_{vlm} , and step (iv) uses Lemma A.42. Combining the bounds yields Lemma A.44. $\square$

A.8.2 Lipschitzness of basic operations

Lemma A.45 (Lipschitzness of the normalization operator, $\|\cdot\|_{\infty, \infty}$ -norm). *Let $\text{normalize}(h)_s := h_s - \max_{s'} h_{s'}$ for $h \in \mathbb{R}^S$. For $h, h' \in \mathbb{R}^S$,*

$$\|\text{normalize}(h) - \text{normalize}(h')\|_{\infty} \leq 2\|h - h'\|_{\infty}. \quad (\text{A.133})$$

Therefore, for $q_i, q'_i \in \mathbb{R}^S$ ($i = 1, 2, \dots, m$), we have

$$\|\text{normalize}(\sum_i q_i) - \text{normalize}(\sum_i q'_i)\|_{\infty} \leq 2m \max_i \|q_i - q'_i\|_{\infty}. \quad (\text{A.134})$$

Proof. Note that $|\max_s h_s - \max_{s'} h'_{s'}| \leq \|h - h'\|_{\infty}$. For each coordinate, we have

$$\begin{aligned} \text{normalize}(h)_s - \text{normalize}(h')_s &= (h_s - \max_{s'} h_{s'}) - (h'_s - \max_{s'} h'_{s'}) \\ &= h_s - h'_s - (\max_{s'} h'_{s'} - \max_{s'} h_{s'}). \end{aligned}$$

Here $h_s - h'_s$ and $\max_s h'_s - \max_{s'} h'_{s'}$ are bounded by $\|h - h'\|_{\infty}$, which yields (A.133).

(A.134) follows in a straightforward way:

$$\|\text{normalize}(\sum_i q_i) - \text{normalize}(\sum_i q'_i)\|_{\infty} \leq 2\|\sum_i q_i - \sum_i q'_i\|_{\infty} \leq 2m \max_i \|q_i - q'_i\|_{\infty}.$$

□

Lemma A.46 (Lipschitzness of the normalized layer, $\|\cdot\|_{2,2}$ -norm). *For any $\mathbf{H} \in \mathbb{R}^{D \times N}$, the normalized function $\text{normalize}(\cdot)$ defined in Eq. (A.56) is $2\sqrt{S}$ -Lipschitz w.r.t. $\mathbf{H}$ in $\|\cdot\|_{2, \infty}$.*

Proof. The lemma follows by noting that

$$\left\| \begin{bmatrix} \mathbf{q}^{(1)} - \mathbf{1}_S \max_{s \in S} \mathbf{q}_s^{(1)} \\ \mathbf{q}^{(2)} - \mathbf{1}_S \max_{s \in S} \mathbf{q}_s^{(2)} \\ \vdots \\ \mathbf{q}^{(L)} - \mathbf{1}_S \max_{s \in S} \mathbf{q}_s^{(L)} \end{bmatrix} \right\|_2 \leq \left\| \begin{bmatrix} \mathbf{q}^{(1)} \\ \mathbf{q}^{(2)} \\ \vdots \\ \mathbf{q}^{(L)} \end{bmatrix} \right\|_2 + \left\| \begin{bmatrix} \max_{s \in S} \mathbf{q}_s^{(1)} \\ \max_{s \in S} \mathbf{q}_s^{(2)} \\ \vdots \\ \max_{s \in S} \mathbf{q}_s^{(L)} \end{bmatrix} \right\|_2 \cdot \|\mathbf{1}_S\|_2 \leq 2\sqrt{S} \cdot \left\| \begin{bmatrix} \mathbf{q}^{(1)} \\ \mathbf{q}^{(2)} \\ \vdots \\ \mathbf{q}^{(L)} \end{bmatrix} \right\|_2.$$

□

Lemma A.47 (Lipschitzness of the softmax activation). *For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$, we have*

$$\left\| \frac{e^{\mathbf{u}}}{\|e^{\mathbf{u}}\|_1} - \frac{e^{\mathbf{v}}}{\|e^{\mathbf{v}}\|_1} \right\|_1 \leq 2e \cdot \|\mathbf{u} - \mathbf{v}\|_{\infty}.$$

Proof of Lemma A.47. Write $\mathbf{u} = (u_1, \dots, u_N)$ and $\mathbf{v} = (v_1, \dots, v_N)$. By definition of ℓ_1 -norm, we have

$$\left\| \frac{e^{\mathbf{u}}}{\|e^{\mathbf{u}}\|_1} - \frac{e^{\mathbf{v}}}{\|e^{\mathbf{v}}\|_1} \right\|_1 = \sum_{i=1}^N \frac{|e^{u_i} \|e^{\mathbf{v}}\|_1 - e^{v_i} \|e^{\mathbf{u}}\|_1|}{\|e^{\mathbf{u}}\|_1 \|e^{\mathbf{v}}\|_1}$$

$$\begin{aligned}
&\leq \sum_{i=1}^N \frac{|e^{u_i} - e^{v_i}| \cdot \|e^v\|_1}{\|e^u\|_1 \|e^v\|_1} + \sum_{i=1}^N \frac{e^{v_i} \cdot \left| \|e^v\|_1 - \|e^u\|_1 \right|}{\|e^u\|_1 \|e^v\|_1} \\
&= \sum_{i=1}^N \frac{|e^{u_i} - e^{v_i}|}{\|e^u\|_1} + \frac{\left| \|e^v\|_1 - \|e^u\|_1 \right|}{\|e^u\|_1} \stackrel{(i)}{\leq} 2 \sum_{i=1}^N \frac{e^{u_i + |u_i - v_i|} \cdot |u_i - v_i|}{\|e^u\|_1} \\
&\leq 2e^{\|\mathbf{u} - \mathbf{v}\|_\infty} \cdot \|\mathbf{u} - \mathbf{v}\|_\infty,
\end{aligned}$$

where step (i) follows from a triangle inequality and the fact that $|e^x - e^y| \leq e^{x+|x-y|} \cdot |x - y|$ for all $x, y \in \mathbb{R}$. When $\|\mathbf{u} - \mathbf{v}\|_2 \leq 1$, it follows that

$$\left\| \frac{e^{\mathbf{u}}}{\|e^{\mathbf{u}}\|_1} - \frac{e^{\mathbf{v}}}{\|e^{\mathbf{v}}\|_1} \right\|_1 \leq 2e \cdot \|\mathbf{u} - \mathbf{v}\|_\infty.$$

When $\|\mathbf{u} - \mathbf{v}\|_\infty \geq 1$, since $\left\| \frac{e^{\mathbf{u}}}{\|e^{\mathbf{u}}\|_1} \right\|_1 = \left\| \frac{e^{\mathbf{v}}}{\|e^{\mathbf{v}}\|_1} \right\|_1 = 1$, we have

$$\left\| \frac{e^{\mathbf{u}}}{\|e^{\mathbf{u}}\|_1} - \frac{e^{\mathbf{v}}}{\|e^{\mathbf{v}}\|_1} \right\|_1 \leq \left\| \frac{e^{\mathbf{u}}}{\|e^{\mathbf{u}}\|_1} \right\|_1 + \left\| \frac{e^{\mathbf{v}}}{\|e^{\mathbf{v}}\|_1} \right\|_1 = 2 \leq 2e \cdot \|\mathbf{u} - \mathbf{v}\|_\infty.$$

Combining the two cases completes the proof. $\square$

Lemma A.48 (Lipschitzness of log-softmax). *For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$, we have*

$$\left\| \log \left(\frac{e^{\mathbf{u}}}{\|e^{\mathbf{u}}\|_1} \right) - \log \left(\frac{e^{\mathbf{v}}}{\|e^{\mathbf{v}}\|_1} \right) \right\|_\infty \leq 2\|\mathbf{u} - \mathbf{v}\|_\infty$$

Proof of Lemma A.48. Let $\mathbf{w} := \mathbf{u} - \mathbf{v}$. Then

$$\begin{aligned}
&\left\| \log \left(\frac{e^{\mathbf{u}}}{\|e^{\mathbf{u}}\|_1} \right) - \log \left(\frac{e^{\mathbf{v}}}{\|e^{\mathbf{v}}\|_1} \right) \right\|_\infty \leq \|\mathbf{u} - \mathbf{v}\|_\infty + |\log \|e^{\mathbf{u}}\|_1 - \log \|e^{\mathbf{v}}\|_1| \\
&\stackrel{(i)}{=} \|\mathbf{u} - \mathbf{v}\|_\infty + \int_0^1 \left\langle \frac{e^{\mathbf{v}+t\mathbf{w}}}{\|e^{\mathbf{v}+t\mathbf{w}}\|_1}, \mathbf{w} \right\rangle dt \leq \|\mathbf{u} - \mathbf{v}\|_\infty + \int_0^1 \left\| \frac{e^{\mathbf{v}+t\mathbf{w}}}{\|e^{\mathbf{v}+t\mathbf{w}}\|_1} \right\|_1 \cdot \|\mathbf{w}\|_\infty dt \\
&= 2\|\mathbf{u} - \mathbf{v}\|_\infty,
\end{aligned}$$

where step (i) uses the Newton-Leibniz formula. $\square$

Lemma A.49 (Lipschitzness of log-sum-exponential). *For $h \in \mathbb{R}^S$ and $\Psi \in \mathbb{R}_+^{S \times S}$, define $f(h) \in \mathbb{R}^S$ by*

$$f(h)_s := \log \sum_{s' \in [S]} \Psi_{ss'} e^{h_{s'}}.$$

Then, for $h, h' \in \mathbb{R}^S$, we have

$$\|f(h) - f(h')\|_\infty \leq \|h - h'\|_\infty.$$

Proof. By differentiating f_s , we have

$$[\nabla f(h)]_s = \left(\frac{\Psi_{ss'} \exp(h_{s'})}{\sum_{s'' \in [S]} \Psi_{ss''} \exp(h_{s''})} \right)_{s'},$$

which implies that $\|\nabla f(h)\|_1 \leq 1$ for all h . Therefore,

$$\|f(h) - f(h')\|_\infty \leq \|\nabla f\|_1 \|h - h'\|_\infty \leq \|h - h'\|_\infty.$$

□

Lemma A.50 (Lipschitzness of log-sum-softmax). *For $h, h' \in \mathbb{R}^S$ and $\Psi \in \mathbb{R}_+^S$, define $f(h, h) \in \mathbb{R}$ by*

$$f(h_1, h_2) := \log \sum_{s \in [S]} \Psi_s \text{softmax}(h)_s \text{softmax}(h')_s.$$

Then, for all $h_1, h_2 \in \mathbb{R}^S$ and $h'_1, h'_2 \in \mathbb{R}^S$, we have

$$\|f(h_1, h_2) - f(h'_1, h'_2)\|_\infty \leq \|h_1 - h'_1\|_\infty + \|h_2 - h'_2\|_\infty.$$

Proof. Let us first fix h_2 . By differentiating f by h_1 , we have

$$[\nabla_{h_1} f(h_1, h_2)]_s = \frac{\Phi_s [\text{softmax}(h_1)_s - (\text{softmax}(h_1)_s)^2] \text{softmax}(h_2)_s}{\sum_{s' \in [S]} \Phi_{s'} \text{softmax}(h_1)_{s'} \text{softmax}(h_2)_{s'}}.$$

By following the argument of Lemma A.49, we have

$$\|f(h_1, h_2) - f(h'_1, h_2)\|_\infty \leq \|h_1 - h'_1\|_\infty.$$

In the same way, we have

$$\|f(h'_1, h_2) - f(h'_1, h'_2)\|_\infty \leq \|h_2 - h'_2\|_\infty.$$

Adding these two bounds together, we obtain the assertion. □

A.8.3 Properties of empirical processes

Lemma A.51 (Proposition A.4 of Bai et al. (2024)). *Let $\{X_\theta\}_{\theta \in \Theta}$ be a zero-mean random process defined as*

$$X_w = \frac{1}{n} \sum_{i=1}^n f(z_i; \theta) - \mathbb{E}_z[f(z; \theta)],$$

where $z_1, \dots, z_n$ are i.i.d. samples from a distribution $\mathbb{P}_z$. Assume the following conditions hold:

- (a) The index set Θ is equipped with a metric ρ and has a diameter B_ρ . Furthermore, there exists a constant A_ρ such that for any subset Θ' of radius r in Θ , the covering number satisfies:

$$\log \mathcal{N}(\Delta; \Theta', \rho) \leq d_\rho \log \frac{2A_\rho r}{\Delta}, \quad \forall 0 < \Delta \leq 2r.$$

- (b) For any fixed $\theta \in \Theta$ and z sampled from $\mathbb{P}_z$, the random variable $f(z; \theta) - \mathbb{E}_z[f(z; \theta)]$ is σ -sub-Gaussian. That is,

$$\mathbb{E} \left[e^{\lambda(f(z; \theta) - \mathbb{E}_z[f(z; \theta)])} \right] \leq e^{\lambda^2 \sigma^2 / 2}, \quad \forall \lambda \in \mathbb{R}.$$

- (c) For any $\theta, \theta' \in \Theta$ and z sampled from $\mathbb{P}_z$, the random variable $f(z; \theta) - f(z; \theta')$ is $\sigma' \rho(\theta, \theta')$ -sub-Gaussian. That is,

$$\mathbb{E} \left[e^{\lambda(f(z; \theta) - f(z; \theta'))} \right] \leq e^{\lambda^2 (\sigma')^2 \rho^2(\theta, \theta') / 2}, \quad \forall \lambda \in \mathbb{R}.$$

Under these assumptions, with probability at least $1 - \eta$, we have

$$\sup_{\theta \in \Theta} |X_\theta| \leq c\sigma \sqrt{\frac{d_\rho \log(2A_\rho(1 + B_\rho \sigma' / \sigma)) + \log(1/\eta)}{n}},$$

where $c > 0$ is some numerical constant.

Lemma A.52 (The empirical InfoNCE loss). *For any function $f : \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}} \mapsto \mathbb{R}$ such that $\|f\|_\infty \leq B$, define the empirical InfoNCE loss as*

$$\begin{aligned} & \widehat{R}_{\text{clip}, K}(f) \\ &:= -\frac{1}{K} \sum_{k=1}^K \log \frac{\exp(f(\mathbf{x}_{\text{im}, k}^{(i)}, \mathbf{x}_{\text{tx}, k}^{(i)}))}{\sum_{j \in [K]} \exp(f(\mathbf{x}_{\text{im}, k}^{(i)}, \mathbf{x}_{\text{tx}, j}^{(i)}))} - \frac{1}{K} \sum_{k=1}^K \log \frac{\exp(f(\mathbf{x}_{\text{im}, k}^{(i)}, \mathbf{x}_{\text{tx}, k}^{(i)}))}{\sum_{j \in [K]} \exp(f(\mathbf{x}_{\text{im}, j}^{(i)}, \mathbf{x}_{\text{tx}, k}^{(i)}))}. \end{aligned}$$

Then,

$$\widehat{R}_{\text{clip}, K}(f) - \mathbb{E}[\widehat{R}_{\text{clip}, K}(f)]$$

is $c \min\{B, e^{2B}/\sqrt{K}\}$ -sub-Gaussian for some universal constant $c > 0$. Moreover, for any $f_1, f_2 : \mathcal{X}_{\text{im}} \times \mathcal{X}_{\text{tx}} \mapsto \mathbb{R}$ such that $\|f_1\|_\infty, \|f_2\|_\infty \leq B$, we have $\widehat{R}_{\text{clip}, K}(f_1) - \mathbb{E}[\widehat{R}_{\text{clip}, K}(f_1)] - (\widehat{R}_{\text{clip}, K}(f_2) - \mathbb{E}[\widehat{R}_{\text{clip}, K}(f_2)])$ is $4\|f_1 - f_2\|_\infty$ -sub-Gaussian.

Proof of Lemma A.52. By considering the case where the denominator's exponent has content $\pm B$ and the numerator's exponent has content $\mp B$, it is easy to see that the difference between the maximum and minimum values of $\widehat{R}_{\text{clip}, K}(f)$ is bounded by $8B$. Therefore, $\widehat{R}_{\text{clip}, K}(f) - \mathbb{E}[\widehat{R}_{\text{clip}, K}(f)]$ is $4B$ -sub-Gaussian.

Next, we show that $\widehat{\mathbf{R}}_{\text{clip},K}(f) - \mathbb{E}[\widehat{\mathbf{R}}_{\text{clip},K}(f)]$ is $\frac{ce^{2B}}{\sqrt{K}}$ -sub-Gaussian. By symmetry, it suffices to consider the behavior of the first term of $\widehat{\mathbf{R}}_{\text{clip},K}(f)$:

$$\frac{1}{K} \sum_{k=1}^K \log \frac{\exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))}{\sum_{j \in [K]} \exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}))}. \quad (\text{A.135})$$

To apply the concentration properties of functions with bounded differences, we evaluate the deviation when one of $(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)})_{k \in [K]}$ is replaced.

By replacing $\mathbf{x}_{\text{im},l}^{(i)}$ with $\bar{\mathbf{x}}_{\text{im},l}^{(i)}$, the variation of the first term is as follows:

$$\begin{aligned} & \frac{1}{K} \sum_{k \neq l} \log \frac{\exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))}{\sum_{j \in [K]} \exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}))} + \frac{1}{K} \log \frac{\exp(f(\bar{\mathbf{x}}_{\text{im},l}^{(i)}, \mathbf{x}_{\text{tx},l}^{(i)}))}{\sum_{j \in [K]} \exp(f(\bar{\mathbf{x}}_{\text{im},l}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}))} \\ & - \frac{1}{K} \sum_{k=1}^K \log \frac{\exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))}{\sum_{j \in [K]} \exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}))} \\ & = \frac{1}{K} \log \frac{\exp(f(\bar{\mathbf{x}}_{\text{im},l}^{(i)}, \mathbf{x}_{\text{tx},l}^{(i)}))}{\sum_{j \in [K]} \exp(f(\bar{\mathbf{x}}_{\text{im},l}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}))} - \frac{1}{K} \log \frac{\exp(f(\mathbf{x}_{\text{im},l}^{(i)}, \mathbf{x}_{\text{tx},l}^{(i)}))}{\sum_{j \in [K]} \exp(f(\mathbf{x}_{\text{im},l}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}))}. \end{aligned} \quad (\text{A.136})$$

Using the boundedness of f , (A.136) is upper and lower bounded by $\frac{2B}{K}$ and $-\frac{2B}{K}$, respectively.

By replacing $\mathbf{x}_{\text{tx},l}^{(i)}$ with $\bar{\mathbf{x}}_{\text{tx},l}^{(i)}$, the variation of the first term is as follows:

$$\begin{aligned} & \frac{1}{K} \sum_{k \neq l} \log \frac{\exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))}{\sum_{j \neq l} \exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)})) + \exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \bar{\mathbf{x}}_{\text{tx},l}^{(i)}))} \\ & + \frac{1}{K} \log \frac{\exp(f(\mathbf{x}_{\text{im},l}^{(i)}, \bar{\mathbf{x}}_{\text{tx},l}^{(i)}))}{\sum_{j \neq l} \exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \bar{\mathbf{x}}_{\text{tx},j}^{(i)})) + \exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \bar{\mathbf{x}}_{\text{tx},l}^{(i)}))} \\ & - \frac{1}{K} \sum_{k=1}^K \log \frac{\exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))}{\sum_{j \in [K]} \exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}))} \\ & = \log \frac{\sum_{j \in [K]} \exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}))}{\sum_{j \neq l} \exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)})) + \exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \bar{\mathbf{x}}_{\text{tx},l}^{(i)}))} + \frac{1}{K} \log \frac{\exp(f(\mathbf{x}_{\text{im},l}^{(i)}, \bar{\mathbf{x}}_{\text{tx},l}^{(i)}))}{\exp(f(\mathbf{x}_{\text{im},l}^{(i)}, \mathbf{x}_{\text{tx},l}^{(i)}))}. \end{aligned} \quad (\text{A.137})$$

Using the boundedness of f and the fact that $\log(1+x) \leq x$ for $x > -1$, it can be verified that (A.137) is upper and lower bounded by $\frac{2e^{2B}}{K}$ and $-\frac{2e^{2B}}{K}$, respectively.

Combining the two bounds above, replacing one sample $(\mathbf{x}_{\text{im},l}^{(i)}, \mathbf{x}_{\text{tx},l}^{(i)})$ with $(\bar{\mathbf{x}}_{\text{im},l}^{(i)}, \bar{\mathbf{x}}_{\text{tx},l}^{(i)})$ changes (A.135) at most $\frac{4e^{2B}+4B}{K} (\leq \frac{6e^{2B}}{K})$. By concentration properties of functions with bounded differences (e.g., Corollary 2.21 in [Wainwright 2019b](#)), the deviation of (A.135) from

its mean is $\frac{3e^{2B}}{\sqrt{K}}$ -sub-Gaussian. Therefore, $\widehat{\mathbf{R}}_{\text{clip},K}(f) - \mathbb{E}[\widehat{\mathbf{R}}_{\text{clip},K}(f)]$ is $\frac{6e^{2B}}{\sqrt{K}}$ -sub-Gaussian. Now, the first assertion of this lemma holds with $c = 6$.

For the second assertion, Lemma A.48 implies that

$$|\widehat{\mathbf{R}}_{\text{clip},K}(f_1) - \widehat{\mathbf{R}}_{\text{clip},K}(f_2)| \leq 4\|f_1 - f_2\|_\infty,$$

for a fixed $(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)})_{k \in [K]}$. This directly implies that $\widehat{\mathbf{R}}_{\text{clip},K}(f_1) - \mathbb{E}[\widehat{\mathbf{R}}_{\text{clip},K}(f_1)] - (\widehat{\mathbf{R}}_{\text{clip},K}(f_2) - \mathbb{E}[\widehat{\mathbf{R}}_{\text{clip},K}(f_2)])$ is $4\|f_1 - f_2\|_\infty$ -sub-Gaussian. $\square$

Remark A.2 (Exponential dependency on $\overline{m}$). In the proof of Theorem 2.10 in Appendix A.5.2, we have only used cB -sub-Gaussianity of $\widehat{\mathbf{R}}_{\text{clip},K}(f) - \mathbb{E}[\widehat{\mathbf{R}}_{\text{clip},K}(f)]$. By applying the fact that $\widehat{\mathbf{R}}_{\text{clip},K}(f) - \mathbb{E}[\widehat{\mathbf{R}}_{\text{clip},K}(f)]$ is $ce^{2B}/\sqrt{K}$ -sub-Gaussian instead, it is easy to see that the rate becomes

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2 L^{11} e^{4\overline{m}} \overline{m}^2 + \log(1/\eta)}{nK}}\right).$$

Here, the denominator involves K , and the convergence rate decreases with the power $-\frac{1}{2}$ of the total number of data, $n \times K$, rather than with the number of batches, n . In exchange for the better dependence on K , the bound exponentially depends on $\overline{m}$. Viewing $\overline{m}$ as a constant makes this permissible; nevertheless, for this reason, in Theorem 2.10 we adopt rates that are polynomial in all parameters.

We believe that, in order to obtain the factor $1/\sqrt{K}$ in the sub-Gaussian parameter in Lemma A.52, an exponential dependence on B is unavoidable. We will provide an example to justify this. To simplify the argument, consider the case when $f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)})$ only depends on $\mathbf{x}_{\text{im},k}^{(i)}$ and we will write $f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}) = y_k$. Define the distribution of y_k as

$$y_k = \begin{cases} B & (\text{w.p. } e^{-B/2}) \\ -\frac{B}{e^{B/2}-1} & (\text{w.p. } 1 - e^{-B/2}) \end{cases}.$$

Then, the mean and variance of e^{y_k} are given respectively as follows:

$$\begin{aligned} \mathbb{E}[e^{y_k}] &= e^{B/2} + (1 - e^{-B/2}) \exp\left(-\frac{B}{e^{B/2}-1}\right) = \Theta(e^{B/2}), \\ \text{Var}(e^{y_k}) &= e^{3B/2} + (1 - e^{-B/2}) \exp\left(-\frac{2B}{e^{B/2}-1}\right) - \left(e^{B/2} + (1 - e^{-B/2}) \exp\left(-\frac{B}{e^{B/2}-1}\right)\right)^2 \\ &= \Theta(e^{3B/2}). \end{aligned}$$

The difference in order between $\mathbb{E}[e^{y_k}]$ and $\sqrt{\text{Var}(e^{y_k})}$ will become important later.

Then, $\widehat{\mathbf{R}}_{\text{clip},K}(f)$ is simplified as

$$\widehat{\mathbf{R}}_{\text{clip},K}(f) = -\frac{1}{K} \sum_{k=1}^K \log \frac{\exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))}{\sum_{j \in [K]} \exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},j}^{(i)}))} - \frac{1}{K} \sum_{k=1}^K \log \frac{\exp(f(\mathbf{x}_{\text{im},k}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))}{\sum_{j \in [K]} \exp(f(\mathbf{x}_{\text{im},j}^{(i)}, \mathbf{x}_{\text{tx},k}^{(i)}))}$$

$$\begin{aligned}
&= -\frac{1}{K} \sum_{k=1}^K \log \frac{e^{y_k}}{\sum_{j \in [K]} e^{y_j}} - \frac{1}{K} \sum_{k=1}^K \log \frac{e^{y_k}}{\sum_{j \in [K]} e^{y_j}} \\
&= 2 \log K + \log \frac{1}{K} \sum_{j \in [K]} e^{y_j} - \frac{1}{K} \sum_{k=1}^K y_k.
\end{aligned} \tag{A.138}$$

When K is sufficiently large so that $\frac{1}{K} \sum_{j \in [K]} e^{y_j}$ and $\frac{1}{K} \sum_{k=1}^K y_k$ concentrate around their expectations, (A.138) is approximated as

$$\begin{aligned}
\hat{R}_{\text{clip},K}(f) &= 2 \log K + \log \left(1 + \frac{1}{K} \sum_{k=1}^K \frac{e^{y_k} - \mathbb{E}[e^{y_k}]}{\mathbb{E}[e^{y_k}]} \right) + \log \mathbb{E}[e^{y_1}] + \frac{1}{K} \sum_{k=1}^K (y_k - \mathbb{E}[y_k]) + E[y_1] \\
&\approx \frac{1}{K} \sum_{k=1}^K \left[\frac{e^{y_k} - \mathbb{E}[e^{y_k}]}{\mathbb{E}[e^{y_k}]} + y_k - \mathbb{E}[y_k] \right] (+\text{const.}).
\end{aligned}$$

Remembering that $\mathbb{E}[e^{y_k}] = \Theta(e^{B/2})$ and $\text{Var}(e^{y_k}) = \Theta(e^{3B/2})$, the variance of $\frac{e^{y_k} - \mathbb{E}[e^{y_k}]}{\mathbb{E}[e^{y_k}]}$ is $\Theta(e^{B/2})$, and thus $\hat{R}_{\text{clip},K}(f) - \mathbb{E}[\hat{R}_{\text{clip},K}(f)]$ is of order $e^{B/4}/\sqrt{K}$. As a consequence, the sub-Gaussian parameter of $\hat{R}_{\text{clip},K}(f) - \mathbb{E}[\hat{R}_{\text{clip},K}(f)]$ is at least exponentially dependent on B .

A.9 Experimental details

This section provides details for the experimental results presented in Section 2.5, along with additional experiments.

A.9.1 Experimental setup

The JGHM data distribution. We generate the dataset from the distribution of Joint Generative Hierarchical Model (JGHM) of Section 2.4. The root distribution $\mathbb{P}(x_r)$ is taken to be uniform over S states. The transition functions $\{\psi_{\square,\iota}^{(\ell)}\}_{\square \in \{\text{im}, \text{tx}\}, \iota \in [S], \ell \in [L]}$ are constructed as follows:

$$\begin{aligned}
[\psi_{\square,\iota}^{(\ell)}(s, s')]_{s, s' \in [S]} &= (1 - p_{\text{flip}}) \times \mathbf{\Pi}_{\square,\iota}^{(\ell)} + p_{\text{flip}} \times \text{softmax}_{\text{row}}(\mathbf{G}_{\square,\iota}^{(\ell)}), \square \in \{\text{im}, \text{tx}\}, \iota \in [S], \ell \in [L], \\
(\mathbf{\Pi}_{\square,\iota}^{(\ell)}, \mathbf{G}_{\square,\iota}^{(\ell)}) &\sim_{\text{iid}} (\mathbf{\Pi}, \mathbf{G}), \quad \mathbf{\Pi}, \mathbf{G} \in \mathbb{R}^{S \times S}, \mathbf{\Pi} \text{ is a random permutation matrix,} \\
&\quad \mathbf{G} \text{ has iid Gaussian entries.}
\end{aligned}$$

This formulation implies that for each parent-child pair $(x_v^{(\ell-1)}, x_{v'}^{(\ell)})$ where $x_v^{(\ell-1)} = s$, the child node $x_{v'}^{(\ell)}$ takes the value $\mathbf{\Pi}_{\square,\iota}^{(\ell)}(s)$ (corresponding to the non-zero element in the s -th row of $\mathbf{\Pi}_{\square,\iota}^{(\ell)}$) with probability $(1 - p_{\text{flip}})$. With probability p_{flip} , the child node $x_{v'}^{(\ell)}$ follows a multinomial distribution parameterized by $\text{softmax}_{\text{row}}(\mathbf{G}_{\square,\iota}^{(\ell)})_{s,:}$. In our experiments, we maintain a fixed set of matrices $(\mathbf{\Pi}_{\square,\iota}^{(\ell)}, \mathbf{G}_{\square,\iota}^{(\ell)})$ by using a consistent random seed for generation.

The parameter p_{flip} determines the conditional entropy of the leaf nodes $\mathbf{x}_{\square}$ given the root node x_r . When $p_{\text{flip}} = 0$, $\mathbf{x}_{\square}$ is a deterministic function of x_r (given fixed matrices $(\mathbf{\Pi}_{\square,\ell}^{(\ell)})_{\square,\ell,\ell}$). Conversely, when $p_{\text{flip}} = 1$, $\mathbf{x}_{\square}$ given x_r exhibits high conditional entropy. Predicting x_r from $\mathbf{x}_{\square}$ is relatively straightforward for small values of p_{flip} , but becomes increasingly challenging as p_{flip} approaches 1.

In our simulations, we set the depth $L = 4$, the states $S = \{1, \dots, 10\}$, and $m_{\text{im}} = m_{\text{tx}} = 3$, and vary the transition randomness parameter p_{flip} from 0.02 to 0.4 with increments of 0.02. Note that in this case $d_{\text{im}} = d_{\text{tx}} = d = 81$. At last, we set the number of pairs in a sample to be $K = 4$.

Belief propagation. Given the transition functions $\{\psi_{\square,\ell}^{(\ell)}\}$, we can compute the true similarity score

$$S_{\star}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \log[\sum_{x_r \in [S]} \mathbb{P}(x_r | \mathbf{x}_{\text{im}}) \mathbb{P}(x_r | \mathbf{x}_{\text{tx}}) / \mathbb{P}(x_r)]$$

by calculating the conditional probabilities $(\mathbb{P}(x_r | \mathbf{x}_{\text{im}}), \mathbb{P}(x_r | \mathbf{x}_{\text{tx}}))$ via belief propagation. This enables us to obtain the global minimum of the CLIP risk $\min_S \bar{R}_{\text{clip},K}(S)$ as defined in Eq. (2.1) (Appendix A.5.1). Similarly, belief propagation algorithm can be applied to find the global minima of both the CDM risk $\min_{\mathbf{M}_t, \mathbf{E}_{\text{tx}}} R_{\text{cdm},t}(\mathbf{M}_t, \mathbf{E}_{\text{tx}})$ from Eq. (2.9) (Appendix A.6.1) and the VLM risk $\min_{\mu, \mathbf{E}_{\text{im}}} R_{\text{vlm}}(\mu, \mathbf{E}_{\text{im}})$ from Eq. (A.4) (Appendix A.7.1).

Guided training. Here we detail the settings for the guided penalty.

CLIP training. For CLIP training, belief propagation involves only downsampling. Let $\mathbf{H}_{\text{im}}^{(\ell-1)}$ and $\mathbf{H}_{\text{tx}}^{(\ell-1)} \in D \times d$ represent the outputs of $\text{Attn}_{\text{im}}^{(\ell)}(\mathbf{H}_{\text{im}}^{(\ell)})$ and $\text{Attn}_{\text{tx}}^{(\ell)}(\mathbf{H}_{\text{tx}}^{(\ell)})$ respectively, for $\ell = L, \dots, 1$. The inputs to these attention layers are $\mathbf{H}_{\text{im}}^{(L)}$ and $\mathbf{H}_{\text{tx}}^{(L)}$. We define $\mathcal{H}_{\square}^{(\ell)}$ as the messages passed from the parent nodes in $\mathcal{V}_{\square}^{(\ell)}$ to the child nodes in $\mathcal{V}_{\square}^{(\ell-1)}$ (for $\square \in \{\text{im}, \text{tx}\}$):

$$\mathcal{H}_{\square}^{(\ell)} = \left[h_{\square, \text{pa}^{(L-\ell)}(v)}^{(\ell)} \right]_{v=1, \dots, d} \in S \times d.$$

The guided penalty $r_{\square}^{(\ell)}$ at each layer is given by:

$$r_{\square}^{(\ell)} = \|\mathbf{H}_{\square, (L-\ell)s : (L+1-\ell)s, :}^{(\ell)} - \mathcal{H}_{\square}^{(\ell)}\|_2^2,$$

where different rows $((L-\ell)s : (L+1-\ell)s)$ in $\mathbf{H}_{\square}^{(\ell)}$ are used for different layers ℓ to align with $\mathcal{H}_{\square}^{(\ell)}$. The total guided penalty is computed as the weighted sum across all layers:

$$r = \sigma \sum_{\ell=0}^{L-1} (r_{\text{tx}}^{(\ell)} + r_{\text{im}}^{(\ell)}),$$

where σ is a hyperparameter controlling the penalty strength.

CDM training. Note that for the CDMs, we have $2L + 1$ layers. The belief propagation process is split into a downsampling phase and an upsampling phase. For the downsampling we have $\mathbf{H}_{\square}^{(\ell-1)} = \text{Attn}_{\square}^{(L+1+\ell)}(\mathbf{H}_{\square}^{(\ell)})$ for $\ell = L + 1, \dots, 1$. For the upsampling, we have

$\mathbf{B}_{\square}^{(\ell)} = \text{Attn}^{(L+2-\ell)}(\mathbf{B}_{\square}^{(\ell-1)})$ for $\ell = 1, \dots, L$. Hence, the input is $\mathbf{H}_{\square}^{(L+1)}$ and the output is $\mathbf{B}_{\text{im}}^{(L)}$. For down sampling, there are two kinds of messages $\mathcal{Q}_{\text{im}}^{(\ell)}$ and $\mathcal{H}_{\text{im}}^{(\ell)}$ i.e.,

$$\mathcal{H}_{\text{im}}^{(\ell)} = \left[h_{\text{im,pa}^{(L-\ell)}(v)}^{(\ell)} \right]_{v=1,\dots,d}, \quad \mathcal{Q}_{\text{im}}^{(\ell)} = \left[q_{\text{im,pa}^{(L-\ell)}(v)}^{(\ell)} \right]_{v=1,\dots,d} \in S \times d.$$

Then the penalty for the image part in downsampling is defined as

$$r_{\text{im},\downarrow} = \sum_{\ell=0}^L (\| \mathbf{H}_{\text{im},(L-\ell)s:(L+1-\ell)s,:}^{(\ell)} - \mathcal{H}_{\text{im}}^{(\ell)} \|_2^2 + \| \mathbf{H}_{\text{im},(2L-\ell)s:(2L+1-\ell)s,:}^{(\ell)} - \mathcal{Q}_{\text{im}}^{(\ell)} \|_2^2).$$

Regarding the text part, there are two scenarios. If we do not use clip features, we follow a procedure similar to clip-guided training. In that case,

$$\mathcal{H}_{\text{tx}}^{(\ell)} = \left[h_{\text{tx,pa}^{(L-\ell)}(v)}^{(\ell+1)} \right]_{v=1,\dots,d} \in S \times d,$$

and

$$r_{\text{tx},\downarrow} = \sum_{\ell=1}^{L+1} \| \mathbf{H}_{\text{tx},(L-\ell)s:(L+1-\ell)s,:}^{(\ell)} - \mathcal{H}_{\text{tx}}^{(\ell)} \|_2^2.$$

Otherwise, if we do use the clip features, then $\mathbf{H}_{\text{tx}}^{(\ell)} \in D \times 1$ and we only ensure that the information is retained after the L -th layer:

$$r_{\text{tx},\downarrow} = \| \mathbf{H}_{\text{tx},:,s,:}^{(L+1)} - \mathbf{H}_{\text{tx},:,s,:}^{(1)} \|_2^2.$$

Finally, there is an upsampling penalty only for the image part. An additional type of message, $\mathcal{B}_{\text{im}}^{(\ell)}$ i.e.,

$$\mathcal{B}_{\text{im}}^{(\ell)} = \left[b_{\text{im,pa}^{(L-\ell)}(v)}^{(\ell)} \right]_{v=1,\dots,d}.$$

Finally, the total penalty is defined as

$$\begin{aligned} r_{\text{im},\uparrow} = & \sum_{\ell=0}^L (\| \mathbf{B}_{\text{im},(L-\ell)s:(L+1-\ell)s,:}^{(\ell)} - \mathcal{H}_{\text{im}}^{(\ell)} \|_2^2 + \| \mathbf{B}_{\text{im},(2L-\ell)s:(2L+1-\ell)s,:}^{(\ell)} - \mathcal{Q}_{\text{im}}^{(\ell)} \|_2^2 \\ & + \| \mathbf{B}_{\text{im},(3L-\ell)s:(3L+1-\ell)s,:}^{(\ell)} - \mathcal{B}_{\text{im}}^{(\ell)} \|_2^2). \end{aligned}$$

The total penalty is defined as $r = \sigma(r_{\text{im},\downarrow} + r_{\text{tx},\downarrow} + r_{\text{im},\uparrow})$.

VLM training. VLM also involves downsampling and upsampling. The information structure is almost the same as that of the CDMs, with the primary distinction being the swapping of roles between image and text. We can define $r_{\text{tx},\downarrow}$, $r_{\text{tx},\uparrow}$ and $r_{\text{im},\downarrow}$ in a similar manner. The total penalty is defined as $r = \sigma(r_{\text{tx},\downarrow} + r_{\text{im},\downarrow} + r_{\text{tx},\uparrow})$.

Learning rates and penalties. After doing a grid search for parameters, we choose the following combinations of learning rates and penalties.

Task	Model	max lr	min lr	penalty (σ)
CLIP	Standard TF	3e-4	3e-7	
CLIP	Guided TF	1e-3	1e-6	1e-3
CLIP	Shallow TF	3e-4	3e-7	
CDM	Standard TF	1e-3	1e-6	
CDM	Guided TF	1e-2	1e-5	1e-1
CDM	Shallow TF	1e-3	1e-6	
CDM	Joint Training	1e-3	1e-6	
VLM	Standard TF	1e-3	1e-6	
VLM	Shallow TF	1e-3	1e-6	
VLM	Guided TF	1e-3	1e-6	1e-3
VLM	Joint Training	3e-4	3e-7	

Table A.1: Learning rates and penalties for different models.

Adam-W parameters. We use the Adam-W optimizer (Loshchilov, 2017) for all our models. The parameters β_1 and β_2 are set to 0.9 and 0.999, respectively. The weight decay is configured to 0.01, and the error term is set to 10^{-8} . Additionally, we apply norm clipping with a maximum ℓ_2 norm of 1.0. Finally, we employ a cosine annealing learning rate scheduler with the number of warm-up steps set to 0.

Network architecture. Now we introduce the details of network architectures.

CLIP architecture. In CLIP training, we parameterize the similarity score function as

$$S^\theta(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \langle \text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}(\mathbf{x}_{\text{im}}), \text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}}(\mathbf{x}_{\text{tx}}) \rangle$$

using an inner-product link function and neural networks $(\text{NN}_{\text{im}}^{\mathbf{W}_{\text{im}}}, \text{NN}_{\text{tx}}^{\mathbf{W}_{\text{tx}}})$ as encoders. Each encoder neural network $\text{NN}_{\square}^{\mathbf{W}_{\square}}(\mathbf{x}_{\square}) = \text{read}(\text{TF}(\text{Emb}(\mathbf{x}_{\square})))$ is composed of a trainable embedding function $\text{Emb} : \mathbb{R}^d \rightarrow \mathbb{R}^{D \times d}$, a trainable read-out function $\text{read} : \mathbb{R}^{D \times d} \rightarrow \mathbb{R}^S$, and a $(L + 1)$ -layer transformer $\text{TF} : \mathbb{R}^{D \times d} \rightarrow \mathbb{R}^{D \times d}$ based on the architecture from Vaswani (2017), modified with RMSNorm instead of LayerNorm, and a pre-norm instead of post-norm. Note that we choose $L = 4$ and the hidden dim $D = 128$.

CDM architecture. In joint CDM training, the conditional denoising function is parameterized as the following: $\mathbf{M}_t(\mathbf{z}_t, \mathbf{x}_{\text{tx}}) = \text{read}(\text{TF}(\text{Emb}_{\text{im}}(\mathbf{z}_t), \text{Emb}_{\text{tx}}(\mathbf{x}_{\text{tx}})))$, where $\text{Emb}_{\text{im}} : \mathbb{R}^d \rightarrow \mathbb{R}^{D \times d}$ and $\text{Emb}_{\text{tx}} : \mathbb{R}^d \rightarrow \mathbb{R}^{D \times d}$ are trainable embedding functions, $\text{read} : \mathbb{R}^{D \times 2d} \rightarrow \mathbb{R}^d$ is a trainable read-out function, and $\text{TF} : \mathbb{R}^{D \times 2d} \rightarrow \mathbb{R}^{D \times 2d}$ is a $(2L + 1)$ -layer transformer.

In cases of partial training with a fixed CLIP embedding $\widehat{\text{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}})$, the conditional denoising function becomes $\mathbf{M}_t(\mathbf{z}_t, \widehat{\text{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}})) = \text{read}(\text{TF}(\text{Emb}_{\text{im}}(\mathbf{z}_t), \text{Emb}_{\text{tx}}(\widehat{\text{E}}_{\text{tx}}(\mathbf{x}_{\text{tx}}))))$, with a trainable embedding function $\text{Emb}_{\text{im}} : \mathbb{R}^d \rightarrow \mathbb{R}^{D \times d}$, a fixed embedding function $\text{Emb}_{\text{tx}} : \mathbb{R}^S \rightarrow \mathbb{R}^d$, a trainable read-out function $\text{read} : \mathbb{R}^{D \times (d+1)} \rightarrow \mathbb{R}^d$, and a transformer $\text{TF} : \mathbb{R}^{D \times (d+1)} \rightarrow \mathbb{R}^{D \times (d+1)}$ consisting of $(2L + 1)$ layers. Note that we set the hidden dim $D = 256$ here.

VLM architecture. Similarly, in the joint training of VLMs, the conditional next-token probability is parameterized as $\mu(x_{\text{tx},k} = \cdot | \mathbf{x}_{\text{im}}) = \text{softmax}(\text{read}(\text{TF}(\text{Emb}_{\text{tx}}(x_{\text{tx},1:k-1}), \text{Emb}_{\text{im}}(\mathbf{x}_{\text{im}}))))$, with trainable embedding functions $\text{Emb}_{\text{im}} : \mathbb{R}^d \rightarrow \mathbb{R}^{D \times d}$ and $\text{Emb}_{\text{tx}} : \mathbb{R}^{k-1} \rightarrow \mathbb{R}^{D \times (k-1)}$, a trainable read-out function $\text{read} : \mathbb{R}^{D \times (d+k-1)} \rightarrow \mathbb{R}^S$, and a $(2L + 1)$ -layer transformer $\text{TF} : \mathbb{R}^{D \times (d+k-1)} \rightarrow \mathbb{R}^{D \times (d+k-1)}$. Note that we set the hidden dim $D = 256$ here.

In cases of partial training with a fixed CLIP embedding $\widehat{\text{E}}_{\text{im}}(\mathbf{x}_{\text{im}})$, the conditional next-token probability becomes the following:

$$\mu(x_{\text{tx},k} = \cdot | \widehat{\text{E}}_{\text{im}}(\mathbf{x}_{\text{im}})) = \text{softmax}(\text{read}(\text{TF}(\text{Emb}_{\text{tx}}(x_{\text{tx},1:k-1}), \text{Emb}_{\text{im}}(\widehat{\text{E}}_{\text{im}}(\mathbf{x}_{\text{im}}))))$$

with a fixed embedding function $\text{Emb}_{\text{im}} : \mathbb{R}^S \rightarrow \mathbb{R}^d$, a trainable embedding function $\text{Emb}_{\text{tx}} : \mathbb{R}^{k-1} \rightarrow \mathbb{R}^{D \times (k-1)}$, a trainable read-out function $\text{read} : \mathbb{R}^{D \times (d+k-1)} \rightarrow \mathbb{R}^S$, and a $(2L + 1)$ -layer transformer $\text{TF} : \mathbb{R}^{D \times (d+k-1)} \rightarrow \mathbb{R}^{D \times (d+k-1)}$.

ZSC settings. For the ZSC, we choose the number of samples to be $M = 250$ in Figures 2.2b and A.6b.

Computational resource. All our experiments are performed on 8 Nvidia Tesla A100 GPUs (80GB memory) and 12 Nvidia Tesla V100 GPUs (16GB memory). The total GPU time is approximately 3000 GPU hours.

A.9.2 Ablation studies

Further discussion of Assumption 2.1

Many of our theoretical results depend on the boundedness conditions in Assumption 2.1, which assume both the score function $\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$ and the logarithm of the probability ratio $\log \frac{\mathbb{P}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})}{\mathbb{P}(\mathbf{x}_{\text{im}})\mathbb{P}(\mathbf{x}_{\text{tx}})}$ are bounded between $[-\log c_1, \log c_1]$ for some constant $c_1 > 0$. In practice (Chen et al., 2020; Radford et al., 2021), the score function $\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})$ is chosen as

$$\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) = \langle \mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}}), \mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}}) \rangle / \tau$$

for some temperature parameter $\tau > 0$ and normalized representations $\mathbf{E}_{\text{im}}(\cdot)$ and $\mathbf{E}_{\text{tx}}(\cdot)$ such that $\|\mathbf{E}_{\text{im}}(\mathbf{x}_{\text{im}})\|_2 = 1$ and $\|\mathbf{E}_{\text{tx}}(\mathbf{x}_{\text{tx}})\|_2 = 1$ for all $\mathbf{x}_{\text{im}} \in \mathcal{X}_{\text{im}}, \mathbf{x}_{\text{tx}} \in \mathcal{X}_{\text{tx}}$. In this case, the absolute score function $|\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})|$ is bounded by $1/\tau$.

To further examine the boundedness assumption of the score function $|\mathbf{S}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})|$, we conducted a controlled experiment by pretraining ResNet-50 on the CC3M dataset (3M samples) (Sharma et al., 2018) with different τ configurations. In addition to the standard design from Radford et al. (2021), where τ is trainable (initialized at 0.07 and clipped to be at least 0.01), we trained models with fixed τ values of 0.07 and 0.1, which impose progressively tighter bounds on the score function. We then evaluated these models on the ImageNet-1k validation set (Deng et al., 2009) in a zero-shot classification setup.

As shown in Table A.2, fixing $\tau = 0.07$ yields comparable performance to the standard trainable setting, while increasing τ to 0.1 results in only a modest decrease in accuracy.

Setting	Val-Top1 (%)	Val-Top5 (%)
Trainable τ	34.81 ± 0.04	17.61 ± 0.07
Fix $\tau = 0.07$	35.12 ± 0.03	18.21 ± 0.11
Fix $\tau = 0.1$	31.95 ± 0.17	15.13 ± 0.15

Table A.2: Top-1 and top-5 validation accuracies (%) across different choices of τ . The ResNet-50 model was pretrained on the CC3M dataset (3M samples) for 32 epochs under various τ configurations. For the trainable τ case, τ was initialized at 0.07 and constrained to be no smaller than 0.01 during training, following the design of (Radford et al., 2021). Each setting was evaluated over 3 random seeds, and results are reported as mean $\pm$ standard deviation.

These findings empirically support our assumption that the score function can be chosen to be bounded.

Nevertheless, the boundedness assumptions on the probability ratio $\frac{\mathbb{P}(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})}{\mathbb{P}(\mathbf{x}_{\text{im}})\mathbb{P}(\mathbf{x}_{\text{tx}})}$ may be relatively strong for certain real-world multimodal data, and it is challenging to estimate the supremum of the probability ratio in real-world multimodal distributions (e.g., image and text) as we only have limited samples from it.

Technically, Assumption 2.1 is mainly used for change-of-measure arguments in the proof of Proposition 2.3 and 2.9. For example, c_1 in Assumption 2.1 provides an upper bound on the ratio $\frac{\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}, \text{tx}}}[f(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})]}{\mathbb{E}_{(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}}) \sim \mathbb{P}_{\text{im}} \times \mathbb{P}_{\text{tx}}}[f(\mathbf{x}_{\text{im}}, \mathbf{x}_{\text{tx}})]}$ for all functions $f \geq 0$. However, for the restricted class of functions considered in the proof, a smaller upper bound may be possible. We leave the further relaxation of Assumption 2.1 to future work.

More samples per category improves zero-shot performance

Figure A.3 illustrates the risks associated with zero-shot learning as a function of the number of samples. The experimental setup for Figure A.3 is nearly identical to that of Figure 2.2b, with the exception that we fix $p_{\text{flip}} = 0.2$ and vary M . Here, M ranges from 5 to 250. We observe that a larger number of samples leads to more accurate predictions.

Figure A.5 evaluates the zero-shot classification performance of a ResNet-50 model pretrained on the CC12M dataset (Changpinyo et al., 2021), tested on the ImageNet-1k validation set (Deng et al., 2009). We adopt the standard 80 text templates introduced in OpenAI’s CLIP paper (Radford et al., 2021). For each trial, we randomly permute the 80 templates and incrementally increase the number of templates following this order. This process is repeated 16 times, and we report the mean and standard deviation of cross-entropy loss, top-1 accuracy, and top-5 accuracy.

To validate our theoretical prediction that these performance bounds follow a rate of $C + O(1/M)$ (Proposition 2.4), where C is a constant and M is the number of templates, we also fit the curves in both synthetic dataset and the real dataset with functions of the form

$f(M) = A + B/M$ (with A and B as parameters). The results in Figure A.4 and Figure A.5 show that the fitted curves (dashed) align closely with the empirical results (solid), with $R^2 > 0.98$, thereby supporting our theoretical findings.

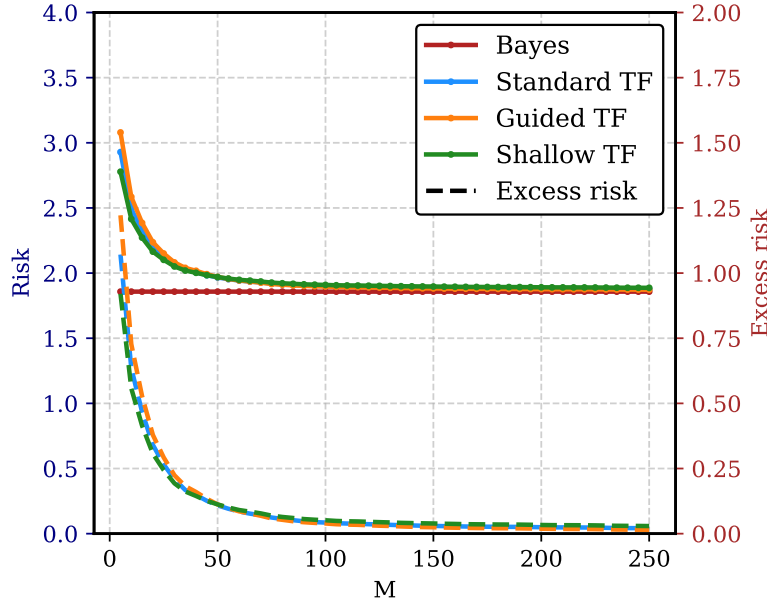


Figure A.3: Risks of zero-shot learning versus number of samples. The setup of Figure A.3 is almost the same as that of Figure 2.2b, except that we fix $p_{\text{flip}} = 0.2$ and vary M , where M ranges from 5 to 250. We can observe that larger numbers of samples lead to improved predictions.

Out-of-distribution test

Figure A.6 shows the out-of-distribution (OOD) risk (solid curve) and excess risk (dashed curve) as functions of the parameter p_{flip} , for CLIP training (Figure A.6a), ZSC (Figure A.6b), CDM (Figure A.6c), and VLM (Figure A.6d). In all these experiments, models are trained with a fixed $p_{\text{flip}} = 0.2$, and their risks are evaluated under varying p_{flip} values that are out-of-distribution. The following observations can be made:

- As expected, across all settings, guided training (**Guided TF**) closely matches the performance of the misspecified BP algorithm (**Mis-spec. BP**), while shallow transformers (**Shallow TF**) perform much worse compared to **Mis-spec. BP**.
- In the CDM (Figure A.6c) and VLM (Figure A.6d) setups, **Standard TF** performs similarly to **Guided TF**, whereas in the CLIP training (Figure A.6a) and ZSC (Figure A.6b) setups, **Standard TF** shows a greater gap from **Guided TF**. This suggests that

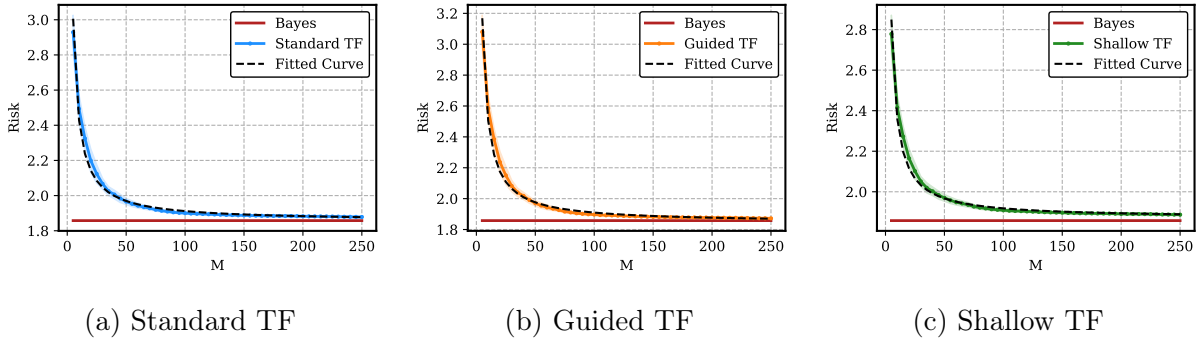


Figure A.4: Fitted curves for the risks in zero-shot learning. Each risk curve is fitted by the function $f(M) = A + B/M$, with parameters A and B . The fitted results are shown as dark dashed lines, while the empirical risks are plotted as solid lines. All fitted curves achieve an R^2 value above 0.98.

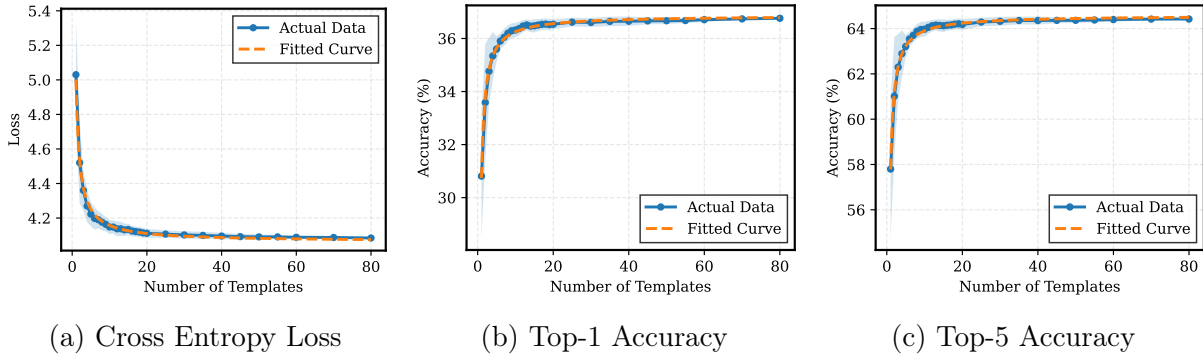


Figure A.5: Zero-shot classification performance of a ResNet-50 model pretrained on CC12M, evaluated on the ImageNet-1k validation set. We use the standard 80 text templates from OpenAI’s CLIP paper (Radford et al., 2021). For each run, we sample a random permutation of these 80 templates and progressively increase the number of templates in that order. This procedure is repeated 16 times, and we report the mean and standard deviation of cross-entropy loss, top-1 accuracy, and top-5 accuracy. We also fit the results with functions of the form $f(M) = A + B/M$, where A and B are parameters. Fitted curves are shown as orange dashed lines, while empirical results are shown as blue solid lines. All fitted curves achieve an R^2 value above 0.98.

standard-trained transformers may perform closer to the belief-propagation algorithm when the in-distribution risk is smaller.

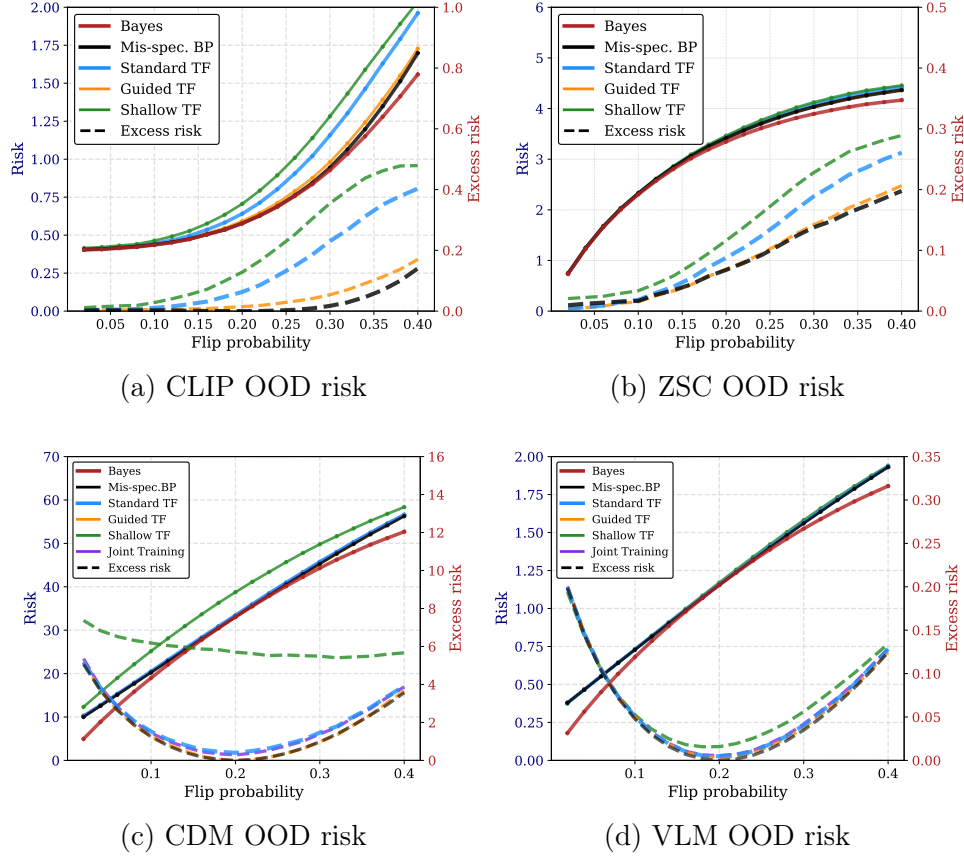


Figure A.6: Out-of-distribution (OOD) risks (solid curves) and excess risks (dashed curves) as a function of the parameter p_{flip} for CLIP training, ZSC, CDM, and VLM. Models are trained with a fixed $p_{\text{flip}} = 0.2$. Across all setups, Guided TF closely matches the performance of Mis-spec. BP. In the CDM (A.6c) and VLM (A.6d) setups, Standard TF performs similarly to Guided TF, whereas in the CLIP training (A.6a) and ZSC (A.6b) setups, Standard TF shows a greater gap from Guided TF. This suggests that standard-trained transformers may perform closer to the belief-propagation algorithm when the in-distribution risk is smaller.

OOD tests with different p_{flip} in image and text trees

Figure A.7 shows OOD risks and excessive risks of transformer architectures and belief propagation for VLMs and CDMs. The settings of Figures A.7a and A.7b are nearly identical to those of Figures A.6c and A.6d. The only difference is that we fix the text $p_{\text{flip}} = 0.2$ while varying the image p_{flip} in Figure A.7a, and conversely, we fix the image $p_{\text{flip}} = 0.2$ while varying the text p_{flip} in Figure A.7b. We observe that the trends are similar to those in Figures A.6c and A.6d.

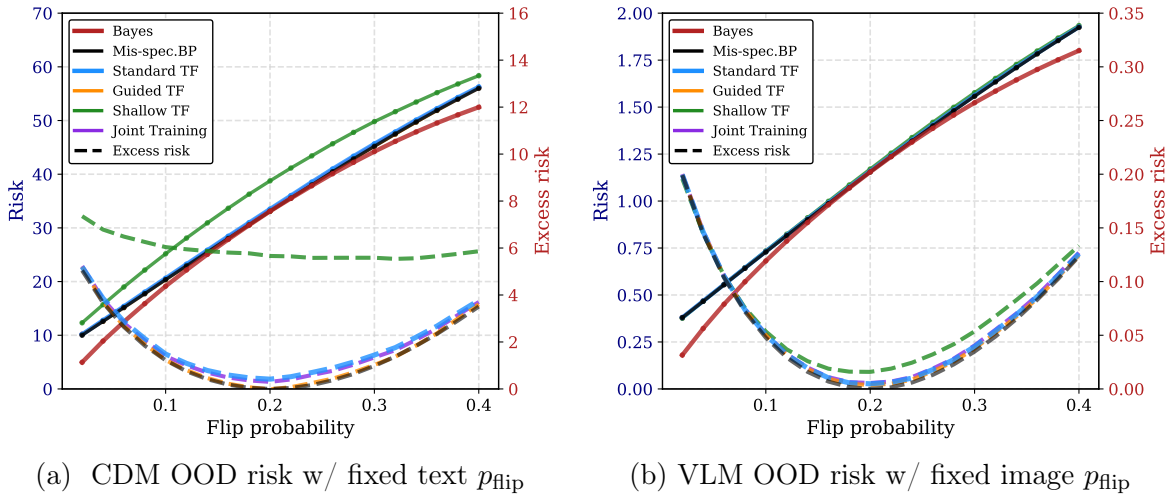


Figure A.7: OOD risks and excessive risks of various transformer architectures and belief propagation for VLMs and CDMs. These figures exhibit same trends as Figures A.6c and A.6d. (a) fix the text $p_{\text{flip}} = 0.2$ but vary the image p_{flip} . (b) fix the image $p_{\text{flip}} = 0.2$ but vary the text p_{flip} .

Appendix B

Appendix for Chapter 3

B.1 Preliminary

In this section, we provide some preliminary discussions and a proof of Theorem 3.4. Concretely, in Section B.1.1 we discuss our data assumptions and introduce additional notations. In Section B.1.2, B.1.3 we derive intermediate results that contribute to the proof of Theorem 3.4. Finally, a complete proof of Theorem 3.4 is contained in Section B.1.4.

B.1.1 Additional notations and comments on data assumptions

Tensors. For matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{X}$ of appropriate shape, it holds that

$$(\mathbf{B}^\top \otimes \mathbf{A}) \circ \mathbf{X} = \mathbf{AXB},$$

and that

$$\begin{aligned} (\mathbf{D}^\top \otimes \mathbf{C}) \circ (\mathbf{B}^\top \otimes \mathbf{A}) \circ \mathbf{X} &= ((\mathbf{D}^\top \mathbf{B}^\top) \otimes (\mathbf{CA})) \circ \mathbf{X} \\ &= \mathbf{CAXBD}. \end{aligned}$$

For simplicity, we denote

$$\mathbf{A}^{\otimes 2} := \mathbf{A} \otimes \mathbf{A}.$$

Comments on Assumption 3.2, 3.3 and 3.4 Due to the rotational invariance of the Gaussian sketched matrix $\mathbf{S}$, throughout the appendix, we assume w.l.o.g. that the covariance of the input covariates $\mathbf{H}$ is diagonal with the (i, i) -th entry being the i -th eigenvalue. Specifically, Assumption 3.3 can be rewritten as

Assumption B.1 (Source condition). *Assume $\mathbf{H} = (h_{ij})_{i,j \geq 1}$ is a diagonal matrix with diagonal entries in non-increasing order, and $\mathbf{w}^*$ satisfies a prior such that*

$$\text{for } i \neq j, \quad \mathbb{E} \mathbf{w}_i^* \mathbf{w}_j^* = 0; \quad \text{and for } i > 0, \quad \mathbb{E} \lambda_i \mathbf{w}_i^{*2} \approx i^{-b}, \quad \text{for some } b > 1.$$

Now that we assume $\mathbf{H}$ is diagonal. We make the following notations. Define

$$\mathbf{H}_{k^*:k^\dagger} := \text{diag}(\lambda_{k^*+1}, \dots, \lambda_{k^\dagger}) \in \mathbb{R}^{(k^\dagger - k^*)^2},$$

where $0 \leq k^* \leq k^\dagger$ are two integers, and we allow $k^\dagger = \infty$. For example,

$$\mathbf{H}_{0:k} = \text{diag}(\lambda_1, \dots, \lambda_i), \quad \mathbf{H}_{k:\infty} = \text{diag}(\lambda_{k+1}, \dots).$$

Similarly, for a vector $\mathbf{w} \in \mathbb{H}$, we have

$$\mathbf{w}_{k^*:k^\dagger} := (\mathbf{w}_{k^*+1}, \dots, \mathbf{w}_{k^\dagger}^*)^\top \in \mathbb{R}^{k^\dagger - k^*}.$$

B.1.2 Approximation error

Recall the risk decomposition in (3.4),

$$\mathcal{R}_M(\mathbf{v}_N) = \underbrace{\min \mathcal{R}(\cdot)}_{\text{Irreducible}} + \underbrace{\min \mathcal{R}_M(\cdot) - \min \mathcal{R}(\cdot)}_{\text{Approx}} + \underbrace{\mathcal{R}_M(\mathbf{v}_N) - \min \mathcal{R}_M(\cdot)}_{\text{Excess}}.$$

Lemma B.1 (Approximization error). *Conditional on the sketch matrix $\mathbf{S}$, the minimizer of $\mathcal{R}_M(\mathbf{v})$ is given by*

$$\mathbf{v}^* := (\mathbf{H}\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{H}\mathbf{w}^*,$$

and the approximation error in (3.4) is

$$\begin{aligned} \text{Approx} &:= \min \mathcal{R}_M(\cdot) - \min \mathcal{R}(\cdot) \\ &= \left\| \left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top(\mathbf{H}\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{H}\mathbf{H}^{\frac{1}{2}} \right) \mathbf{H}^{\frac{1}{2}}\mathbf{w}^* \right\|^2. \end{aligned} \quad (\text{B.1})$$

Moreover, $\text{Approx} \leq \|\mathbf{w}^*\|_{\mathbf{H}}^2$ almost surely over the randomness of $\mathbf{S}$.

Proof of Lemma B.1. Recall that the risk

$$\mathcal{R}(\mathbf{w}) := \mathbb{E}(\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$$

is a quadratic function and that $\mathbf{w}^*$ is the minimizer of $\mathcal{R}(\cdot)$, so we have

$$(\mathbb{E}\mathbf{x}^{\otimes 2})\mathbf{w}^* = \mathbb{E}\mathbf{x}y \Leftrightarrow \mathbf{H}\mathbf{w}^* = \mathbb{E}\mathbf{x}y,$$

and

$$\begin{aligned} \mathcal{R}(\mathbf{w}) &= \mathbb{E}(\langle \mathbf{x}, \mathbf{w} \rangle - \langle \mathbf{x}, \mathbf{w}^* \rangle)^2 + \mathcal{R}(\mathbf{w}^*) \\ &= \|\mathbf{H}^{\frac{1}{2}}(\mathbf{w} - \mathbf{w}^*)\|^2 + \mathcal{R}(\mathbf{w}^*). \end{aligned}$$

Recall that the risk in a restricted subspace

$$\mathcal{R}_M(\mathbf{v}) := \mathcal{R}(\mathbf{S}^\top \mathbf{v}) = \mathbb{E}(\langle \mathbf{S}\mathbf{x}, \mathbf{v} \rangle - y)^2$$

is also a quadratic function, so its minimizer is given by

$$\begin{aligned}\mathbf{v}^* &= (\mathbb{E}(\mathbf{S}\mathbf{x})^{\otimes 2})^{-1} \mathbb{E}\mathbf{S}\mathbf{x}y \\ &= (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}\mathbf{H}\mathbf{w}^*.\end{aligned}$$

Therefore, the approximation error is

$$\begin{aligned}\text{Approx} &:= \mathcal{R}_M(\mathbf{v}^*) - \mathcal{R}(\mathbf{w}^*) \\ &= \mathcal{R}(\mathbf{S}^\top \mathbf{v}^*) - \mathcal{R}(\mathbf{w}^*) \\ &= \|\mathbf{H}^{\frac{1}{2}}(\mathbf{S}^\top \mathbf{v}^* - \mathbf{w}^*)\|^2 \\ &= \|\mathbf{H}^{\frac{1}{2}}(\mathbf{S}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}\mathbf{H}\mathbf{w}^* - \mathbf{w}^*)\|^2 \\ &= \left\| \left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}} \mathbf{S}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}\mathbf{H}^{\frac{1}{2}} \right) \mathbf{H}^{\frac{1}{2}} \mathbf{w}^* \right\|^2.\end{aligned}$$

Finally, since

$$\left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}} \mathbf{S}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}\mathbf{H}^{\frac{1}{2}} \right)^2 = \mathbf{I} - \mathbf{H}^{\frac{1}{2}} \mathbf{S}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}\mathbf{H}^{\frac{1}{2}} \preceq \mathbf{I},$$

it follows that $\text{Approx} \leq \|\mathbf{w}^*\|_{\mathbf{H}}^2$. □

B.1.3 Bias-variance decomposition

The excess risk in (3.4) can be viewed as the SGD excess risk in an M -dimensional (misspecified) linear regression problem. We will utilize Corollary 3.4 in Wu et al. (2022b) to get a bias-variance decomposition of the excess risk. The following two lemmas check the related assumptions for Corollary 3.4 in Wu et al. (2022b) in our setup.

Lemma B.2 (Hypercontractivity and the misspecified noise under sketched feature). *Suppose that Assumptions 3.5A and 3.5B hold. Conditioning on the sketch matrix $\mathbf{S}$, for every PSD matrix $\mathbf{A} \in \mathbb{R}^{M \times M}$, we have*

$$\mathbb{E}(\mathbf{S}\mathbf{x})^{\otimes 2} \mathbf{A} (\mathbf{S}\mathbf{x})^{\otimes 2} \preceq \alpha \text{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top \mathbf{A}) \mathbf{S}\mathbf{H}\mathbf{S}^\top.$$

Moreover, for the minimizer of $\mathcal{R}_M(\mathbf{v})$, that is, $\mathbf{v}^*$ defined in Lemma B.1, we have

$$\mathbb{E}(y - \langle \mathbf{v}^*, \mathbf{S}\mathbf{x} \rangle)^2 (\mathbf{S}\mathbf{x})^{\otimes 2} \preceq 2(\sigma^2 + \alpha \|\mathbf{w}^*\|_{\mathbf{H}}^2) \mathbf{S}\mathbf{H}\mathbf{S}^\top.$$

The expectation in the above is over $(\mathbf{x}, y)$.

Proof of Lemma B.2. The first part is a direct application of Assumption 3.5A:

$$\begin{aligned}\mathbb{E}(\mathbf{S}\mathbf{x})^{\otimes 2} \mathbf{A} (\mathbf{S}\mathbf{x})^{\otimes 2} &= \mathbf{S}(\mathbb{E}\mathbf{x}\mathbf{x}^\top (\mathbf{S}^\top \mathbf{A} \mathbf{S}) \mathbf{x}\mathbf{x}^\top) \mathbf{S}^\top \\ &\preceq \mathbf{S}(\alpha \text{tr}(\mathbf{H}\mathbf{S}^\top \mathbf{A} \mathbf{S}) \mathbf{H}) \mathbf{S}^\top\end{aligned}$$

$$= \alpha \operatorname{tr}(\mathbf{SHS}^\top \mathbf{A}) \mathbf{SHS}^\top.$$

For the second part, we first show that

$$\begin{aligned} \mathbb{E}(y - \langle \mathbf{v}^*, \mathbf{Sx} \rangle)^2 \mathbf{x}^{\otimes 2} &\preceq 2\mathbb{E}(y - \langle \mathbf{w}^*, \mathbf{x} \rangle)^2 \mathbf{x}^{\otimes 2} + 2\mathbb{E}\langle \mathbf{w}^* - \mathbf{S}^\top \mathbf{v}^*, \mathbf{x} \rangle^2 \mathbf{x}^{\otimes 2} \\ &\preceq 2\sigma^2 \mathbf{H} + 2\alpha \langle \mathbf{H}, (\mathbf{w}^* - \mathbf{S}^\top \mathbf{v}^*)^{\otimes 2} \rangle \mathbf{H}, \end{aligned}$$

where the last inequality is by Assumptions 3.5A and 3.5B. From the proof of Lemma B.1, we know that

$$\langle \mathbf{H}, (\mathbf{w}^* - \mathbf{S}^\top \mathbf{v}^*)^{\otimes 2} \rangle = \text{Approx} \leq \|\mathbf{w}^*\|_{\mathbf{H}}^2, \text{ almost surely.}$$

So we have

$$\mathbb{E}(y - \langle \mathbf{v}^*, \mathbf{Sx} \rangle)^2 \mathbf{x}^{\otimes 2} \preceq 2(\sigma^2 + \alpha \|\mathbf{w}^*\|_{\mathbf{H}}^2) \mathbf{H}.$$

Left and right multiplying both sides with $\mathbf{S}$ and $\mathbf{S}^\top$, we obtain the second claim. $\square$

Lemma B.3 (Gaussianity and well-specified noise under sketched features). *Suppose that Assumptions 3.1A and 3.1B hold. Conditional on the sketch matrix $\mathbf{S}$, we have*

$$\mathbf{Sx} \sim \mathcal{N}(0, \mathbf{SHS}^\top).$$

Moreover, for the minimizer of $\mathcal{R}_M(\mathbf{v})$, that is, $\mathbf{v}^*$ defined in Lemma B.1, we have

$$\mathbb{E}[y|\mathbf{Sx}] = \langle \mathbf{Sx}, \mathbf{v}^* \rangle, \quad \mathbb{E}(y - \langle \mathbf{Sx}, \mathbf{v}^* \rangle)^2 = \sigma^2 + \text{Approx} \geq \sigma^2.$$

Proof of Lemma B.3. The first claim is a direct consequence of Assumption 3.1A.

For the second claim, by Assumption 3.1A and Lemma B.1, we have

$$\begin{aligned} \mathbb{E}[y|\mathbf{x}] &= \langle \mathbf{x}, \mathbf{w}^* \rangle \\ &= \langle \mathbf{x}, \mathbf{S}^\top \mathbf{v}^* \rangle + \langle \mathbf{x}, \mathbf{w}^* - \mathbf{S}^\top \mathbf{v}^* \rangle \\ &= \langle \mathbf{x}, \mathbf{S}^\top \mathbf{v}^* \rangle + \langle \mathbf{x}, [\mathbf{I} - (\mathbf{SHS}^\top)^{-1} \mathbf{SH}] \mathbf{w}^* \rangle \\ &= \langle \mathbf{H}^{-\frac{1}{2}} \mathbf{x}, \mathbf{H}^{\frac{1}{2}} \mathbf{S}^\top \mathbf{v}^* \rangle + \langle \mathbf{H}^{-\frac{1}{2}} \mathbf{x}, [\mathbf{I} - \mathbf{H}^{\frac{1}{2}} \mathbf{S}^\top (\mathbf{SHS}^\top)^{-1} \mathbf{SH}^{\frac{1}{2}}] \mathbf{H}^{\frac{1}{2}} \mathbf{w}^* \rangle \\ &= \langle \mathbf{SH}^{\frac{1}{2}} \mathbf{H}^{-\frac{1}{2}} \mathbf{x}, \mathbf{v}^* \rangle + \langle [\mathbf{I} - \mathbf{H}^{\frac{1}{2}} \mathbf{S}^\top (\mathbf{SHS}^\top)^{-1} \mathbf{SH}^{\frac{1}{2}}] \mathbf{H}^{-\frac{1}{2}} \mathbf{x}, \mathbf{H}^{\frac{1}{2}} \mathbf{w}^* \rangle. \end{aligned} \quad (\text{B.2})$$

Notice that

$$\mathbf{H}^{-\frac{1}{2}} \mathbf{x} \sim \mathcal{N}(0, \mathbf{I}),$$

by Assumption 3.1A and that

$$\mathbf{SH}^{\frac{1}{2}} [\mathbf{I} - \mathbf{H}^{\frac{1}{2}} \mathbf{S}^\top (\mathbf{SHS}^\top)^{-1} \mathbf{SH}^{\frac{1}{2}}] = 0,$$

therefore

$$\mathbf{Sx} = \mathbf{SH}^{\frac{1}{2}} \mathbf{H}^{-\frac{1}{2}} \mathbf{x} \text{ is independent of } [\mathbf{I} - \mathbf{H}^{\frac{1}{2}} \mathbf{S}^\top (\mathbf{SHS}^\top)^{-1} \mathbf{SH}^{\frac{1}{2}}] \mathbf{H}^{-\frac{1}{2}} \mathbf{x}.$$

Taking expectation over the second random vector in (B.2), we find

$$\mathbb{E}[y|\mathbf{S}\mathbf{x}] = \mathbb{E}\mathbb{E}[y|\mathbf{x}] = \langle \mathbf{S}\mathbf{H}^{\frac{1}{2}}\mathbf{H}^{-\frac{1}{2}}\mathbf{x}, \mathbf{v}^* \rangle = \langle \mathbf{S}\mathbf{x}, \mathbf{v}^* \rangle.$$

It remains to show

$$\mathbb{E}(y - \langle \mathbf{S}\mathbf{x}, \mathbf{v}^* \rangle)^2 = \sigma^2 + \text{Approx}.$$

This follows from the proof of Lemma B.1. Specifically,

$$\begin{aligned} \mathbb{E}(y - \langle \mathbf{S}\mathbf{x}, \mathbf{v}^* \rangle)^2 &= \mathcal{R}(\mathbf{S}^\top \mathbf{v}^*) \\ &= \text{Approx} + \mathcal{R}(\mathbf{w}^*) \\ &= \text{Approx} + \sigma^2 \\ &\geq \sigma^2, \end{aligned}$$

where the second equality is by the definition of **Approx** and the third equality is by Assumption 3.1B. We have completed the proof. $\square$

B.1.4 Proof of Theorem 3.4

We now use the results in (Wu et al., 2022a,b) for SGD to obtain the following bias-variance decomposition on the excess risk.

Theorem B.4 (Excess risk bounds). *Consider the excess risk in (3.4) induced by the output of (SGD). Let*

$$N_{\text{eff}} := N/\log(N), \quad \text{SNR} := (\|\mathbf{w}^*\|_{\mathbf{H}}^2 + \|\mathbf{v}_0\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2)/\sigma^2.$$

Then conditioning on the sketch matrix $\mathbf{S}$, for any $\mathbf{w}^ \in \mathbb{H}$*

1. *Under Assumptions 3.5A and 3.5B, we have*

$$\mathbb{E}\text{Excess} \lesssim \left\| \prod_{t=1}^N (\mathbf{I} - \gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top) (\mathbf{v}_0 - \mathbf{v}^*) \right\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2 + (1 + \alpha \text{SNR}) \sigma^2 \cdot \frac{D_{\text{eff}}}{N_{\text{eff}}}$$

when $\gamma \lesssim \frac{1}{c\alpha \text{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$ for some constant $c > 0$.

2. *Under Assumptions 3.1A and 3.1B, we have*

$$\mathbb{E}\text{Excess} \gtrsim \left\| \prod_{t=1}^N (\mathbf{I} - \gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top) (\mathbf{v}_0 - \mathbf{v}^*) \right\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2 + \sigma^2 \cdot \frac{D_{\text{eff}}}{N_{\text{eff}}}$$

when $\gamma \lesssim \frac{1}{c \text{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$ for some constant $c > 0$.

In both results, the expectation is over $(\mathbf{x}_t, y_t)_{t=1}^N$, and

$$D_{\text{eff}} := \#\{\tilde{\lambda}_j \geq 1/(N_{\text{eff}}\gamma)\} + (N_{\text{eff}}\gamma)^2 \sum_{\tilde{\lambda}_j < 1/(N_{\text{eff}}\gamma)} \tilde{\lambda}_j^2,$$

where $(\tilde{\lambda}_j)_{j=1}^M$ are eigenvalue of $\mathbf{SHS}^\top$.

Theorem 3.4 follows immediately by Lemma B.1 and by setting $\mathbf{v}_0 = 0$ and plugging the definition of $\text{Bias}(\mathbf{w}^*)$ and Var into Theorem B.4.

Proof of Theorem B.4. This follows from Corollary 3.4 in Wu et al. (2022b) for a linear regression problem with population data given by $(\mathbf{S}\mathbf{x}, y)$. Note that the data covariance becomes $\mathbf{SHS}^\top$ and the optimal model parameter becomes $\mathbf{v}^*$.

For the upper bound, Lemma B.2 verifies Assumptions 1A and 2 in Wu et al. (2022b), with the noise level being

$$\tilde{\sigma}^2 = 2(\sigma^2 + \alpha\|\mathbf{w}^*\|_{\mathbf{H}}^2).$$

Then we can apply the upper bound in Corollary 3.4 in Wu et al. (2022b) (setting their index set $\mathbb{K} = \emptyset$) to get

$$\mathbb{E}\text{Excess} \lesssim \left\| \prod_{t=1}^N (\mathbf{I} - \gamma_t \mathbf{SHS}^\top) (\mathbf{v}_0 - \mathbf{v}^*) \right\|_{\mathbf{SHS}^\top}^2 + (\|\mathbf{v}^* - \mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 + \tilde{\sigma}^2) \frac{D_{\text{eff}}}{N_{\text{eff}}}.$$

We verify that

$$\begin{aligned} \|\mathbf{v}^* - \mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 &\leq 2\|\mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top \mathbf{v}^*\|^2 + 2\|\mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 \\ &= 2\|\mathbf{H}^{\frac{1}{2}}\mathbf{S}^\top (\mathbf{SHS}^\top)^{-1} \mathbf{SH}\mathbf{w}^*\|^2 + 2\|\mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 \\ &\leq 2\|\mathbf{H}^{\frac{1}{2}}\mathbf{w}^*\|^2 + 2\|\mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 \\ &= 2\|\mathbf{w}^*\|_{\mathbf{H}}^2 + 2\|\mathbf{v}_0\|_{\mathbf{SHS}^\top}^2, \end{aligned}$$

which implies that

$$\begin{aligned} (\|\mathbf{v}^* - \mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 + \tilde{\sigma}^2) &\leq 2\|\mathbf{w}^*\|_{\mathbf{H}}^2 + 2\|\mathbf{v}_0\|_{\mathbf{SHS}^\top}^2 + 2(\sigma^2 + \alpha\|\mathbf{w}^*\|_{\mathbf{H}}^2) \\ &\lesssim (1 + \alpha\text{SNR})\sigma^2. \end{aligned}$$

Substituting, we get the upper bound.

For the lower bound, Lemma B.3 shows $\mathbf{S}\mathbf{x}$ is Gaussian, therefore it satisfies Assumption 1B in Wu et al. (2022b) with $\beta = 1$. Besides, Lemma B.3 shows that the linear regression problem is well-specified, with the noise level being

$$\tilde{\sigma}^2 = \sigma^2 + \text{Approx} \geq \sigma^2.$$

Although the lower bound in Corollary 3.4 in Wu et al. (2022b) is stated for Gaussian additive noise (see their Assumption 2'), it is easy to check that the lower bound holds for

any well-specified noise as described by Lemma B.3. Using the lower bound in Corollary 3.4 in Wu et al. (2022b), we obtain

$$\mathbb{E}\text{Excess} \gtrsim \left\| \prod_{t=1}^N (\mathbf{I} - \gamma_t \mathbf{S} \mathbf{H} \mathbf{S}^\top) (\mathbf{v}_0 - \mathbf{v}^*) \right\|_{\mathbf{S} \mathbf{H} \mathbf{S}^\top}^2 + \tilde{\sigma}^2 \frac{D_{\text{eff}}}{N_{\text{eff}}}.$$

Plugging in $\tilde{\sigma}^2 \geq \sigma^2$ gives the desired lower bound. $\square$

B.1.5 Proofs of Lemma 3.5, Theorem 3.6 and 3.7

Lemma 3.5 is proved in Lemma B.27. Theorem 3.6 follows from Lemma B.5 and B.11. Theorem 3.7 follows from Lemma B.6 and B.12.

B.2 Proofs in Section 3.4

B.2.1 Proof of Theorem 3.1

Proof of part 1. By Assumption 3.1B and the definition of $\mathcal{R}(\cdot)$, we have

$$\begin{aligned} \mathcal{R}(\mathbf{w}) &= \mathbb{E}(\langle \mathbf{x}, \mathbf{w} \rangle - y)^2 = \mathbb{E}(\langle \mathbf{x}, \mathbf{w} \rangle - \mathbb{E}[y | \mathbf{x}])^2 + \mathbb{E}(y - \mathbb{E}[y | \mathbf{x}])^2 \\ &= \mathbb{E}(\langle \mathbf{x}, \mathbf{w} \rangle - \langle \mathbf{x}, \mathbf{w}^* \rangle)^2 + \sigma^2 \geq \sigma^2. \end{aligned}$$

Note that the equality holds if and only if $\mathbf{w} = \mathbf{w}^*$. Therefore we have $\min \mathcal{R}(\cdot) = \mathcal{R}(\mathbf{w}^*) = \sigma^2$.

Proof of part 2. Part 2 of Theorem 3.1 follows immediately from Lemma B.6.

Proof of part 3. We choose $\text{Bias}(\mathbf{w}^*)$, Var as defined in Eq. (3.5) and (3.6) and let $\text{Bias} := \mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*)$. Part 3 of Theorem 3.1 follows directly from the decomposition of the excess risk in Theorem 3.4 (note that $\mathbb{E} \|\mathbf{w}^*\|_{\mathbf{H}}^2 / \sigma^2 \lesssim 1$), and the matching bounds in Lemma B.13 and B.16.

It remains to verify the stepsize assumption required in Lemma B.13. Since we have from Lemma B.27 that

$$\frac{1}{\text{tr}(\mathbf{S} \mathbf{H} \mathbf{S}^\top)} = \frac{1}{\sum_{i=1}^M \tilde{\lambda}_i} \geq \frac{c_1}{\sum_{i=1}^M \lambda_i} \geq \frac{c_2}{\sum_{i=1}^M i^{-a}} \geq c_3$$

for some a -dependent constants $c_1, c_2, c_3 > 0$ with probability at least $1 - e^{-\Omega(M)}$, it follows that for any constant $c > 1$, we can choose $\gamma \leq c_0$ for some a -dependent c_0 such that $\gamma \leq \frac{1}{c \text{tr}(\mathbf{S} \mathbf{H} \mathbf{S}^\top)}$. Therefore, we have verified the stepsize assumption.

Finally, the last claim in Theorem 3.1 follows directly from combining the previous three parts and Theorem 3.4, noting $\sigma^2 \lesssim 1$. and

$$\text{Var} \approx \frac{\min\{M, (N_{\text{eff}} \gamma)^{1/a}\}}{N_{\text{eff}}} \lesssim \frac{(N_{\text{eff}} \gamma)^{1/a}}{N_{\text{eff}}} \lesssim (N_{\text{eff}} \gamma)^{1/a-1} \lesssim \text{Bias} + \text{Approx}$$

under the stepsize assumption $\gamma \lesssim 1$. Here the hidden constants may depend on a .

B.2.2 Proof of Theorem 3.2

Similar to the proof of Theorem 3.1, we have $\min \mathcal{R}(\cdot) = \sigma^2$ under Assumption 3.1B. Moreover, by Lemmas B.9, B.14 and B.16, we have with probability at least $1 - e^{-\Omega(M)}$ that

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} \text{Approx} &\asymp M^{1-b}, \\ \text{Bias} &\lesssim \max \{ M^{1-b}, (N_{\text{eff}} \gamma)^{(1-b)/a} \}, \\ \text{Bias} &\gtrsim (N_{\text{eff}} \gamma)^{(1-b)/a} \text{ when } (N_{\text{eff}} \gamma)^{1/a} \leq M/3, \\ \text{Var} &\asymp \min \{ M, (N_{\text{eff}} \gamma)^{1/a} \} / N_{\text{eff}}, \end{aligned}$$

when the stepsize $\gamma \leq c$ for some a -dependent constant $c > 0$. Here the hidden constants in the bounds may depend only on (a, b) . Combining the bounds on **Approx**, **Bias**, **Var** yields Theorem 3.2.

B.2.3 Proof of Theorem 3.3

Similar to the proof of Theorem 3.1, we have $\min \mathcal{R}(\cdot) = \sigma^2$ under Assumption 3.1B. Notice that we have $\gamma \lesssim 1$ implies $\gamma \lesssim 1/(\text{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top))$ with probability at least $1 - e^{-\Omega(M)}$ by Lemma B.29. It follows from Lemma B.10, B.15 and B.17 that

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} \text{Approx} &\asymp \log^{1-a} M, \\ \text{Bias} &\lesssim \max \{ \log^{1-a} M, \log^{1-a}(N_{\text{eff}} \gamma) \}, \\ \text{Bias} &\gtrsim \log^{1-a}(N_{\text{eff}} \gamma) \text{ when } (N_{\text{eff}} \gamma)^{1/a} \leq M^c \text{ for some small constant } c > 0, \\ \text{Var} &\asymp \frac{\min \{ M, (N_{\text{eff}} \gamma) / \log^a(N_{\text{eff}} \gamma) \}}{N_{\text{eff}}} \lesssim \frac{(N_{\text{eff}} \gamma) / \log^a(N_{\text{eff}} \gamma)}{N_{\text{eff}} \gamma} = \log^{-a}(N_{\text{eff}} \gamma) \end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$ when the stepsize $\gamma \leq c$ for some a -dependent constant $c > 0$. Since $\text{Var} \lesssim \text{Bias}$ and $\log^{1-a}(N_{\text{eff}} \gamma) \lesssim \log^{1-a} M$ when $(N_{\text{eff}} \gamma)^{1/a} \gtrsim M^c$, putting the bounds together gives Theorem 3.3.

B.3 Approximation error

In this section, we derive upper and lower bounds for the approximation error in (3.4) (and B.1). We will also show that the upper and lower bounds match up to constant factors in several examples.

B.3.1 An upper bound

Lemma B.5 (An upper bound on the approximation error). *Given any $k \leq d$ such that $r(\mathbf{H}) \geq k + M$, the approximation error in (3.4) (and B.1) satisfies*

$$\text{Approx} \lesssim \|\mathbf{w}_{k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2 + \langle [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1}, \mathbf{w}_{0:k}^* \mathbf{w}_{0:k}^{*\top} \rangle$$

almost surely, where $\mathbf{A}_k := \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top$. If in addition $k \leq M/2$, then with probability $1 - e^{-\Omega(M)}$

$$\text{Approx} \lesssim \|\mathbf{w}_{k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2 + \left(\frac{\sum_{i>k} \lambda_i}{M} + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M}} \right) \|\mathbf{w}_{0:k}^*\|^2,$$

where $(\lambda_i)_{i=1}^p$ are eigenvalues of $\mathbf{H}$ in non-increasing order.

Proof of Lemma B.5. Write the singular value decomposition $\mathbf{H} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top$, where $\mathbf{\Lambda} := \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_d\}$ with $\lambda_1 \geq \lambda_2 \geq \dots \lambda_d \geq 0$ and $\mathbf{U} \mathbf{U}^\top = \mathbf{I}$. Define $\tilde{\mathbf{S}} := \mathbf{S} \mathbf{U}$, $\tilde{\mathbf{w}}^* := \mathbf{U}^\top \mathbf{w}^*$. Then by Lemma B.1 the approximation error $\text{Approx} = \text{Approx}(\mathbf{S}, \mathbf{H}, \mathbf{w}^*)$ satisfies

$$\begin{aligned} \text{Approx}(\mathbf{S}, \mathbf{H}, \mathbf{w}^*) &= \left\| \left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}} \mathbf{S}^\top (\mathbf{S} \mathbf{H} \mathbf{S}^\top)^{-1} \mathbf{S} \mathbf{H}^{\frac{1}{2}} \right) \mathbf{H}^{\frac{1}{2}} \mathbf{w}^* \right\|^2 \\ &= \left\| \left(\mathbf{I} - \mathbf{U} \mathbf{\Lambda}^{\frac{1}{2}} \tilde{\mathbf{S}}^\top (\tilde{\mathbf{S}} \mathbf{\Lambda} \tilde{\mathbf{S}}^\top)^{-1} \mathbf{S} \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{U}^\top \right) \mathbf{U} \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{U}^\top \mathbf{w}^* \right\|^2 \\ &= \left\| \mathbf{U} \left(\mathbf{I} - \mathbf{\Lambda}^{\frac{1}{2}} \tilde{\mathbf{S}}^\top (\tilde{\mathbf{S}} \mathbf{\Lambda} \tilde{\mathbf{S}}^\top)^{-1} \mathbf{S} \mathbf{\Lambda}^{\frac{1}{2}} \right) \mathbf{\Lambda}^{\frac{1}{2}} \mathbf{U}^\top \mathbf{w}^* \right\|^2 \\ &= \left\| \left(\mathbf{I} - \mathbf{\Lambda}^{\frac{1}{2}} \tilde{\mathbf{S}}^\top (\tilde{\mathbf{S}} \mathbf{\Lambda} \tilde{\mathbf{S}}^\top)^{-1} \mathbf{S} \mathbf{\Lambda}^{\frac{1}{2}} \right) \mathbf{\Lambda}^{\frac{1}{2}} \tilde{\mathbf{w}}^* \right\|^2 = \text{Approx}(\tilde{\mathbf{S}}, \mathbf{\Lambda}, \tilde{\mathbf{w}}^*). \end{aligned}$$

Since $\tilde{\mathbf{S}} \stackrel{d}{=} \mathbf{S}$ by rotational invariance of standard Gaussian variables, it suffices to analyze the case where $\mathbf{H} = \mathbf{\Lambda}$ is a diagonal matrix, as the results may transfer to general $\mathbf{H}$ by replacing $\tilde{\mathbf{w}}^*$ with $\mathbf{w}^*$.

Therefore, from now on we assume w.l.o.g. that $\mathbf{H}$ is a diagonal matrix with non-increasing diagonal entries. Define $\mathbf{A} := \mathbf{S} \mathbf{H} \mathbf{S}^\top$.

By definition of Approx , we have

$$\begin{aligned} \text{Approx} &= \left\| \left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}} \mathbf{S}^\top (\mathbf{S} \mathbf{H} \mathbf{S}^\top)^{-1} \mathbf{S} \mathbf{H}^{\frac{1}{2}} \right) \mathbf{H}^{\frac{1}{2}} \mathbf{w}^* \right\|^2 \\ &= \langle [\mathbf{H}^{1/2} \mathbf{S}^\top \mathbf{A}^{-1} \mathbf{S} \mathbf{H}^{1/2} - \mathbf{I}_d]^{\otimes 2}, \mathbf{H}^{1/2} \mathbf{w}^* \mathbf{w}^{*\top} \mathbf{H}^{1/2} \rangle. \end{aligned}$$

Moreover, for any $k \in [d]$

$$\begin{aligned} \mathbf{H}^{1/2} \mathbf{S}^\top \mathbf{A}^{-1} \mathbf{S} \mathbf{H}^{1/2} - \mathbf{I}_d &= \begin{pmatrix} \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \\ \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \end{pmatrix} \mathbf{A}^{-1} \begin{pmatrix} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} & \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} \end{pmatrix} - \mathbf{I}_d \\ &= \begin{pmatrix} \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k}^{1/2} - \mathbf{I}_k & \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} \\ \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k}^{1/2} & \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} - \mathbf{I}_{d-k} \end{pmatrix} \\ &=: \begin{pmatrix} \mathbf{U} & \mathbf{V} \\ \mathbf{V}^\top & \mathbf{W} \end{pmatrix} \end{aligned} \tag{B.3}$$

Therefore

$$[\mathbf{H}^{1/2} \mathbf{S}^\top \mathbf{A}^{-1} \mathbf{S} \mathbf{H}^{1/2} - \mathbf{I}_p]^{\otimes 2} = \begin{pmatrix} \mathbf{U}^2 + \mathbf{V} \mathbf{V}^\top & \mathbf{U} \mathbf{V} + \mathbf{V} \mathbf{W} \\ \mathbf{V}^\top \mathbf{U} + \mathbf{W} \mathbf{V}^\top & \mathbf{W}^2 + \mathbf{V}^\top \mathbf{V} \end{pmatrix} \preceq 2 \begin{pmatrix} \mathbf{U}^2 + \mathbf{V} \mathbf{V}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{W}^2 + \mathbf{V}^\top \mathbf{V} \end{pmatrix},$$

and hence

$$\begin{aligned} \text{Approx} &\leq 2 \left\langle \begin{pmatrix} \mathbf{U}^2 + \mathbf{V}\mathbf{V}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{W}^2 + \mathbf{V}^\top\mathbf{V} \end{pmatrix}, \mathbf{H}^{1/2} \mathbf{w}^* \mathbf{w}^{*\top} \mathbf{H}^{1/2} \right\rangle \\ &= 2 \langle \mathbf{U}^2 + \mathbf{V}\mathbf{V}^\top, \mathbf{H}_{0:k}^{1/2} \mathbf{w}_{*,0:k} \mathbf{w}_{*,0:k}^\top \mathbf{H}_{0:k}^{1/2} \rangle + 2 \langle \mathbf{W}^2 + \mathbf{V}^\top\mathbf{V}, \mathbf{H}_{k:\infty}^{1/2} \mathbf{w}_{*,k:\infty} \mathbf{w}_{*,k:\infty}^\top \mathbf{H}_{k:\infty}^{1/2} \rangle. \end{aligned}$$

We claim the following results which we will prove at the end of the proof.

$$\langle \mathbf{W}^2 + \mathbf{V}^\top\mathbf{V}, \mathbf{H}_{k:\infty}^{1/2} \mathbf{w}_{*,k:\infty} \mathbf{w}_{*,k:\infty}^\top \mathbf{H}_{k:\infty}^{1/2} \rangle \leq \|\mathbf{w}_{*,k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2, \quad (\text{B.4})$$

$$\langle \mathbf{U}^2 + \mathbf{V}\mathbf{V}^\top, \mathbf{H}_{0:k}^{1/2} \mathbf{w}_{*,0:k} \mathbf{w}_{*,0:k}^\top \mathbf{H}_{0:k}^{1/2} \rangle = \langle [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1}, \mathbf{w}_{0:k}^* \mathbf{w}_{0:k}^{*\top} \rangle. \quad (\text{B.5})$$

Note that in claim (B.5) the inverse $\mathbf{A}_k^{-1}$ exists almost surely since $r(\mathbf{H}_{k:\infty}) \geq r(\mathbf{H}) - k \geq M$ by our assumption and $\mathbf{S}_{k:\infty} \in \mathbb{R}^{M \times (d-k)}$ is a random gaussian projection onto $\mathbb{R}^M$. First part of Lemma B.5 follows immediately from combining claim (B.4) and (B.5).

To prove the second part of Lemma B.5, first note that with probability $1 - e^{-\Omega(M)}$ we have

$$\mu_{\min}(\mathbf{A}_k^{-1}) = \|\mathbf{A}_k\|^{-1} \geq c / \left(\frac{\sum_{i>k} \lambda_i}{M} + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M}} \right)$$

for some constant $c > 0$ by Lemma B.25. Moreover, by the concentration of the Gaussian variance matrix (see e.g., Theorem 6.1 in Wainwright 2019a), we have $\mathbf{S}_{0:k}^\top \mathbf{S}_{0:k} \succeq \mathbf{I}_k/5$ with probability $1 - e^{-\Omega(M)}$ when $M/k \geq 2$. Combining the last two arguments, we obtain

$$\begin{aligned} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k} &\succeq c \mathbf{S}_{0:k}^\top \mathbf{S}_{0:k} / \left(\frac{\sum_{i>k} \lambda_i}{M} + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M}} \right) \\ &\succeq \mathbf{I}_k / \left(\frac{\sum_{i>k} \lambda_i}{M} + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M}} \right), \end{aligned}$$

and therefore

$$\begin{aligned} \langle [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1}, \mathbf{w}_{0:k}^* \mathbf{w}_{0:k}^{*\top} \rangle &\leq \langle [\mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1}, \mathbf{w}_{0:k}^* \mathbf{w}_{0:k}^{*\top} \rangle \\ &\leq \|[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1}\| \|\mathbf{w}_{0:k}^*\|^2 \\ &\lesssim \left(\frac{\sum_{i>k} \lambda_i}{M} + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M}} \right) \|\mathbf{w}_{0:k}^*\|^2 \quad (\text{B.6}) \end{aligned}$$

with probability $1 - e^{-\Omega(M)}$. Combining Eq. (B.6) with the first part of Lemma B.5 completes the proof.

Proof of claim (B.4) Note that

$$\begin{aligned} -\mathbf{I}_{d-k} \preceq \mathbf{W} &= \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} - \mathbf{I}_{d-k} \\ &= \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top (\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top + \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} - \mathbf{I}_{d-k} \end{aligned}$$

$$\preceq \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top (\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} - \mathbf{I}_{d-k} \preceq \mathbf{0}_{d-k},$$

where the last inequality uses the fact that the norm of projection matrices is no greater than one. Therefore, we have $\|\mathbf{W}\|_2 \leq 1$. Now, it remains to show

$$\mathbf{W}^2 + \mathbf{V}^\top \mathbf{V} = -\mathbf{W}, \quad (\text{B.7})$$

as claim (B.4) is a direct consequence of Eq. (B.7) and the fact that $\|\mathbf{W}\| \leq 1$.

By definition of $\mathbf{W}$ in Eq. (B.3), we have

$$\begin{aligned} \mathbf{W}^2 &= (\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} - \mathbf{I}_{d-k})^2 \\ &= \mathbf{I}_{d-k} - 2\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} + \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} \\ &= \mathbf{I}_{d-k} - 2\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} + \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{A}_k \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}. \end{aligned}$$

By definition of $\mathbf{V}$ in Eq. (B.3), we have

$$\mathbf{V}^\top \mathbf{V} = \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} (\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top) \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}.$$

Since $\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top + \mathbf{A}_k = \mathbf{A}$, it follows that

$$\begin{aligned} \mathbf{W}^2 + \mathbf{V}^\top \mathbf{V} &= \mathbf{I}_{d-k} - 2\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} + \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{A} \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} \\ &= \mathbf{I}_{d-k} - \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} = -\mathbf{W}. \end{aligned}$$

Proof of claim (B.5) It suffices to show $\mathbf{U}^2 + \mathbf{V}^\top \mathbf{V} = [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1}$. Using the definition of $\mathbf{U}$ in Eq. (B.3), we obtain

$$\begin{aligned} \mathbf{U} &= \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k}^{1/2} - \mathbf{I}_k \\ &= \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k}^{1/2} - \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k}^{1/2} - \mathbf{I}_k \\ &= \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{H}_{0:k}^{-1} \mathbf{H}_{0:k}^{1/2} - \mathbf{I}_k, \end{aligned}$$

where the second line uses Woodbury's matrix identity, namely

$$\mathbf{A}^{-1} = [\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top + \mathbf{A}_k]^{-1} = \mathbf{A}_k^{-1} - \mathbf{A}_k^{-1} \mathbf{S}_{0:k} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1}.$$

Continuing the calculation of $\mathbf{U}$, we have

$$\begin{aligned} \mathbf{U} &= \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{H}_{0:k}^{-1/2} - \mathbf{I}_k \\ &= \mathbf{H}_{0:k}^{1/2} (\mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} - \mathbf{I}_k) \mathbf{H}_{0:k}^{-1/2} \\ &= -\mathbf{H}_{0:k}^{-1/2} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{H}_{0:k}^{-1/2}. \end{aligned}$$

Therefore,

$$\mathbf{U}^2 = \mathbf{H}_{0:k}^{-1/2} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{H}_{0:k}^{-1} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{H}_{0:k}^{-1/2}. \quad (\text{B.8})$$

Since

$$\begin{aligned}\mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}^{-1} &= \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} - \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \\ &= \mathbf{H}_{0:k}^{-1/2} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1}\end{aligned}$$

by Woodbury's matrix identity, it follows from the definition of $\mathbf{V}$ in Eq. (B.3) that

$$\begin{aligned}\mathbf{V} \mathbf{V}^\top &= \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k}^{1/2} \\ &= \mathbf{H}_{0:k}^{-1/2} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} (\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top) \mathbf{A}_k^{-1} \mathbf{S}_{0:k} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{H}_{0:k}^{-1/2} \\ &= \mathbf{H}_{0:k}^{-1/2} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{H}_{0:k}^{-1/2}.\end{aligned}\tag{B.9}$$

Combining Eq. (B.8) and (B.9) yields

$$\mathbf{U}^2 + \mathbf{V} \mathbf{V}^\top = \mathbf{H}_{0:k}^{-1/2} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{H}_{0:k}^{-1/2},\tag{B.10}$$

and therefore

$$\langle \mathbf{U}^2 + \mathbf{V} \mathbf{V}^\top, \mathbf{H}_{0:k}^{1/2} \mathbf{w}_{*,0:k} \mathbf{w}_{*,0:k}^\top \mathbf{H}_{0:k}^{1/2} \rangle = \langle [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1}, \mathbf{w}_{0:k}^* \mathbf{w}_{0:k}^{*\top} \rangle.$$

□

B.3.2 A lower bound

For the approximation error `Approx`, we have the following result.

Lemma B.6 (Lower bound on the approximation error). *When $r(\mathbf{H}) \geq M$, under Assumption 3.1C, the approximation error in (3.4) (and B.1) satisfies*

$$\mathbb{E}_{\mathbf{w}^*} \text{Approx} \gtrsim \sum_{i=M}^d \lambda_i,$$

where $(\lambda_i)_{i=1}^d$ are eigenvalues of $\mathbf{H}$ in non-increasing order.

Proof of Lemma B.6. For any $k \leq d$, following the proof of Lemma B.5, we have

$$\text{Approx} = \langle [\mathbf{H}^{1/2} \mathbf{S}^\top \mathbf{A}^{-1} \mathbf{S} \mathbf{H}^{1/2} - \mathbf{I}_d]^{\otimes 2}, \mathbf{H}^{1/2} \mathbf{w}^* (\mathbf{w}^*)^\top \mathbf{H}^{1/2} \rangle$$

and

$$\mathbf{H}^{1/2} \mathbf{S}^\top \mathbf{A}^{-1} \mathbf{S} \mathbf{H}^{1/2} - \mathbf{I}_d = \begin{pmatrix} \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k}^{1/2} - \mathbf{I}_k & \mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} \\ \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k}^{1/2} & \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2} - \mathbf{I}_{d-k} \end{pmatrix}$$

$$=: \begin{pmatrix} \mathbf{U} & \mathbf{V} \\ \mathbf{V}^\top & \mathbf{W} \end{pmatrix}. \quad (\text{B.11})$$

Therefore

$$[\mathbf{H}^{1/2} \mathbf{S}^\top \mathbf{A}^{-1} \mathbf{S} \mathbf{H}^{1/2} - \mathbf{I}_d]^{\otimes 2} = \begin{pmatrix} \mathbf{U}^2 + \mathbf{V} \mathbf{V}^\top & \mathbf{U} \mathbf{V} + \mathbf{V} \mathbf{W} \\ \mathbf{V}^\top \mathbf{U} + \mathbf{W} \mathbf{V}^\top & \mathbf{W}^2 + \mathbf{V}^\top \mathbf{V} \end{pmatrix}$$

and

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} \text{Approx} &= \mathbb{E}_{\mathbf{w}^*} \langle \mathbf{U}^2 + \mathbf{V} \mathbf{V}^\top, \mathbf{H}_{0:k}^{1/2} \mathbf{w}_{0:k}^* \mathbf{w}_{0:k}^* \mathbf{H}_{0:k}^{1/2} \rangle + \mathbb{E}_{\mathbf{w}^*} \langle \mathbf{W}^2 + \mathbf{V}^\top \mathbf{V}, \mathbf{H}_{k:\infty}^{1/2} \mathbf{w}_{k:\infty}^* \mathbf{w}_{k:\infty}^* \mathbf{H}_{k:\infty}^{1/2} \rangle \\ &\quad + 2 \mathbb{E}_{\mathbf{w}^*} \langle \mathbf{U} \mathbf{V} + \mathbf{V} \mathbf{W}, \mathbf{H}_{0:k}^{1/2} \mathbf{w}_{0:k}^* \mathbf{w}_{k:\infty}^* \mathbf{H}_{k:\infty}^{1/2} \rangle \\ &= \text{tr}((\mathbf{U}^2 + \mathbf{V} \mathbf{V}^\top) \mathbf{H}_{0:k}) + \text{tr}((\mathbf{W}^2 + \mathbf{V}^\top \mathbf{V}) \mathbf{H}_{k:\infty}), \end{aligned}$$

where the last line uses the fact that $\mathbb{E}_{\mathbf{w}^*}(\mathbf{w}^*)^{\otimes 2} = \mathbf{I}_d$. Using Eq. (B.7) and (B.10) in the proof of Lemma B.5, we further obtain

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} \text{Approx} &= \text{tr}(\mathbf{H}_{0:k}^{-1/2} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{H}_{0:k}^{-1/2} \mathbf{H}_{0:k}) - \text{tr}(\mathbf{W} \mathbf{H}_{k:\infty}) \\ &= \text{tr}([\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1}) - \text{tr}(\mathbf{W} \mathbf{H}_{k:\infty}) \\ &\geq -\text{tr}(\mathbf{W} \mathbf{H}_{k:\infty}) =: T_3. \end{aligned}$$

where $\mathbf{A}_k := \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top$. For T_3 , we further have

$$\begin{aligned} T_3 &= \text{tr}(\mathbf{H}_{k:\infty}^{1/2} [\mathbf{I}_{d-k} - \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}_k^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}] \mathbf{H}_{k:\infty}^{1/2}) \\ &\geq \text{tr}(\mathbf{H}_{k:\infty}^{1/2} [\mathbf{I}_{d-k} - \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}_k^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}] \mathbf{H}_{k:\infty}^{1/2}) \\ &\geq \sum_{i=1}^{d-k} \mu_i(\mathbf{I}_{d-k} - \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}_k^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}) \cdot \mu_{d+1-k-i}(\mathbf{H}_{k:\infty}), \end{aligned}$$

where the second line is due to $\mathbf{A} \succeq \mathbf{A}_k$ (and hence $-\mathbf{A}^{-1} \succeq -\mathbf{A}_k^{-1}$), the third line follows from Von-Neuman's inequality. Since $\mathbf{M} := \mathbf{I}_{d-k} - \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}_k^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}$ is a projection matrix such that $\mathbf{M}^2 = \mathbf{M}$ and $\text{tr}(\mathbf{I}_{d-k} - \mathbf{M}) = M$, it follows that $\mathbf{M}$ has M eigenvalues 0 and $d - k - M$ eigenvalues 1. Therefore, we further have

$$T_3 \geq \sum_{i=1}^{d-k} \mu_i(\mathbf{M}) \cdot \mu_{d+1-k-i}(\mathbf{H}_{k:\infty}) \geq \sum_{i=k+M}^d \lambda_i$$

for any $k \leq d$. Letting $k = 0$ maximizes the lower bound and concludes the proof. $\square$

B.3.3 A lower bound under Assumption 3.3

Lemma B.7 (Lower bound on the approximation error under Assumption 3.3). *Under Assumption 3.3, the approximation error in (3.4) (and B.1) satisfies*

$$\mathbb{E}_{\mathbf{w}^*} \text{Approx} \gtrsim \sum_{i=M}^d \lambda_i i^{a-b},$$

where $(\lambda_i)_{i=1}^d$ are eigenvalues of $\mathbf{H}$ in non-increasing order and the inequality hides some (a, b) -dependent constant.

Proof of Lemma B.7. The proof is essentially the same as the proof of Lemma B.6 but we include it here for completeness. Let $\mathbf{H}^{\mathbf{w}} := \mathbb{E}[\mathbf{w}^* \mathbf{w}^{*\top}]$ be the covariance of the prior on $\mathbf{w}^*$. Following the proof of Lemma B.6, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} \text{Approx} &= \mathbb{E}_{\mathbf{w}^*} \langle \mathbf{U}^2 + \mathbf{V}\mathbf{V}^\top, \mathbf{H}_{0:k}^{1/2} \mathbf{w}_{0:k}^* \mathbf{w}_{0:k}^{*\top} \mathbf{H}_{0:k}^{1/2} \rangle + \mathbb{E}_{\mathbf{w}^*} \langle \mathbf{W}^2 + \mathbf{V}^\top \mathbf{V}, \mathbf{H}_{0:k}^{1/2} \mathbf{w}_{k:\infty}^* \mathbf{w}_{k:\infty}^{*\top} \mathbf{H}_{k:\infty}^{1/2} \rangle \\ &\quad + 2\mathbb{E}_{\mathbf{w}^*} \langle \mathbf{U}\mathbf{V} + \mathbf{V}\mathbf{W}, \mathbf{H}_{0:k}^{1/2} \mathbf{w}_{0:k}^* \mathbf{w}_{k:\infty}^{*\top} \mathbf{H}_{k:\infty}^{1/2} \rangle \\ &= \text{tr}((\mathbf{U}^2 + \mathbf{V}\mathbf{V}^\top) \mathbf{H}_{0:k} \mathbf{H}_{0:k}^{\mathbf{w}}) + \text{tr}((\mathbf{W}^2 + \mathbf{V}^\top \mathbf{V}) \mathbf{H}_{k:\infty} \mathbf{H}_{k:\infty}^{\mathbf{w}}), \end{aligned}$$

where the last line uses Assumption 3.3 and notice that $\mathbf{H}, \mathbf{H}^{\mathbf{w}}$ are both diagonal. Next, similar to the proof of Lemma B.6, using Eq. (B.7) and (B.10), we derive

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} \text{Approx} &= \text{tr}(\mathbf{H}_{0:k}^{-1/2} [\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{H}_{0:k}^{-1/2} \mathbf{H}_{0:k} \mathbf{H}_{0:k}^{\mathbf{w}}) - \text{tr}(\mathbf{W} \mathbf{H}_{k:\infty} \mathbf{H}_{k:\infty}^{\mathbf{w}}) \\ &= \text{tr}([\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top \mathbf{A}_k^{-1} \mathbf{S}_{0:k}]^{-1} \mathbf{H}_{0:k}^{\mathbf{w}}) - \text{tr}(\mathbf{W} \mathbf{H}_{k:\infty} \mathbf{H}_{k:\infty}^{\mathbf{w}}) \\ &\geq -\text{tr}(\mathbf{W} \mathbf{H}_{k:\infty} \mathbf{H}_{k:\infty}^{\mathbf{w}}) =: \tilde{T}_3 \end{aligned}$$

where $\mathbf{A}_k := \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top$. For $\tilde{T}_3$, following the same argument for T_3 in the proof of Lemma B.6, we have

$$\begin{aligned} \tilde{T}_3 &\geq \sum_{i=1}^{d-k} \mu_i (\mathbf{I}_{d-k} - \mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}_k^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}) \cdot \mu_{d+1-k-i} (\mathbf{H}_{k:\infty} \mathbf{H}_{k:\infty}^{\mathbf{w}}) \\ &\geq \sum_{i=k+M}^d \mu_i (\mathbf{H} \mathbf{H}^{\mathbf{w}}) \gtrsim \sum_{i=k+M}^d i^{a-b} \lambda_i, \end{aligned}$$

for any $k \leq d$ where the last inequality uses Assumption 3.3. Setting $k = 0$ maximizes the lower bound and concludes the proof. $\square$

B.3.4 Examples on matching bounds for Approx

In this section, we derive matching upper and lower bounds for **Approx** (defined in Eq. 3.4 and B.1) in three concrete examples: power-law spectrum (Lemma B.8), power-law spectrum with source condition (Lemma B.9) and logarithmic power-law spectrum (Lemma B.15).

Lemma B.8 (Bounds on **Approx** under the power-law spectrum). *Suppose Assumption 3.1C and 3.2 hold. Then with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$*

$$M^{1-a} \lesssim \mathbb{E}_{\mathbf{w}^*} \text{Approx} \lesssim M^{1-a}.$$

Here, the hidden constants only depend on the power-law degree a .

Proof of Lemma B.8. For the upper bound, by Lemma B.5 and noting $\mathbb{E}\mathbf{w}_i^{*2} = 1$ for all i , we have with probability at least $1 - e^{-\Omega(M)}$

$$\begin{aligned}\mathbb{E}_{\mathbf{w}^*}\text{Approx} &\lesssim \sum_{k>k_1} \lambda_i + \left(\frac{\sum_{i>k_1} \lambda_i}{M} + \lambda_{k_1+1} + \sqrt{\frac{\sum_{i>k_1} \lambda_i^2}{M}} \right) \cdot k_1 \\ &\lesssim k_1^{1-a} + \left(\frac{k_1^{1-a}}{M} + k_1^{-a} + \sqrt{\frac{k_1^{1-2a}}{M}} \right) k_1 \\ &\lesssim \left(\frac{k_1}{M} + 1 \right) k_1^{1-a}\end{aligned}$$

for any given $k_1 \leq M/2$. Here the hidden constants depend on a . Therefore, letting $k_1 = M/2$ in the upper bound yields

$$\mathbb{E}_{\mathbf{w}^*}\text{Approx} \lesssim M^{1-a}$$

with probability at least $1 - e^{-\Omega(M)}$.

For the lower bound, we have from Lemma B.6 that

$$\mathbb{E}_{\mathbf{w}^*}\text{Approx} \gtrsim \sum_{i=M}^{\infty} i^{-a} \gtrsim M^{1-a}.$$

This completes the proof. $\square$

Lemma B.9 (Bounds on Approx under the source condition). *Suppose Assumption 3.3 hold. Then with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$*

$$M^{1-b} \lesssim \mathbb{E}_{\mathbf{w}^*}\text{Approx} \lesssim M^{1-b}.$$

Here, the hidden constants only depend on the power-law degrees a, b .

Proof of Lemma B.9. For the upper bound, by Lemma B.5 and noting $\mathbb{E}\mathbf{w}_i^{*2} \asymp i^{a-b}$ for all i , we have with probability at least $1 - e^{-\Omega(M)}$

$$\begin{aligned}\text{Approx} &\lesssim \sum_{k>k_1} \lambda_i i^{a-b} + \left(\frac{\sum_{i>k_1} \lambda_i}{M} + \lambda_{k_1+1} + \sqrt{\frac{\sum_{i>k_1} \lambda_i^2}{M}} \right) \cdot k_1^{1+a-b} \\ &\lesssim k_1^{1-b} + \left(\frac{k_1^{1-a}}{M} + k_1^{-a} + \sqrt{\frac{k_1^{1-2a}}{M}} \right) k_1^{1+a-b} \\ &\lesssim \left(\frac{k_1}{M} + 1 \right) k_1^{1-b}\end{aligned}$$

for any given $k_1 \leq M/2$. Here the hidden constants depend on a, b . Moreover, choosing $k_1 = M/2$ in the upper bound gives

$$\mathbb{E}_{\mathbf{w}^*}\text{Approx} \lesssim M^{1-b}$$

with probability at least $1 - e^{-\Omega(M)}$.

For the lower bound, we have from Lemma B.7 that

$$\mathbb{E}_{\mathbf{w}^*} \text{Approx} \gtrsim \sum_{i=M}^{\infty} i^{-a} \cdot i^{a-b} \gtrsim M^{1-b}.$$

This completes the proof. $\square$

Lemma B.10 (Bounds on Approx under the logarithmic power-law spectrum). *Suppose Assumption 3.4 hold. Then with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$*

$$\log^{1-a} M \lesssim \mathbb{E}_{\mathbf{w}^*} \text{Approx} \lesssim \log^{1-a} M.$$

Here, the hidden constants only depend on the power-law degree a .

Proof of Lemma B.10. For the upper bound, by Lemma B.5 and noting $\mathbb{E} \mathbf{w}_i^{*2} = 1$ for all i , we have with probability at least $1 - e^{-\Omega(M)}$

$$\begin{aligned} \text{Approx} &\lesssim \sum_{k > k_1} \lambda_i + \left(\frac{\sum_{i > k_1} \lambda_i}{M} + \lambda_{k_1+1} + \sqrt{\frac{\sum_{i > k_1} \lambda_i^2}{M}} \right) k_1 \\ &\lesssim \log^{1-a} k_1 + \left(\frac{\log^{1-a} k_1}{M} + k_1^{-1} \log^{-a} k_1 + \sqrt{\frac{k_1^{1-2a}}{M}} \right) k_1 \\ &\lesssim \left(1 + \frac{k_1}{M} + \frac{1}{\log k_1} + \frac{1}{\log k_1} \sqrt{\frac{k_1}{M}} \right) \log^{1-a} k_1 \\ &\lesssim \log^{1-a} k_1 \end{aligned}$$

for any given $k_1 \leq M/2$, where the third line uses $\sum_{i > k_1} \lambda_i^2 \lesssim 1/(k_1 \log^{2a} k_1)$. Choosing $k_1 = M/2$, we obtain

$$\mathbb{E}_{\mathbf{w}^*} \text{Approx} \lesssim \log^{1-a} M$$

with probability at least $1 - e^{-\Omega(M)}$. Here the hidden constants depend on a, b .

For the lower bound, we have from Lemma B.6 that

$$\mathbb{E}_{\mathbf{w}^*} \text{Approx} \gtrsim \sum_{i=M}^{\infty} \lambda_i \gtrsim \sum_{i=M}^{\infty} i^{-1} \log^{-a} i \gtrsim \log^{1-a} M.$$

Therefore, we have established matching upper and lower bounds for Approx. $\square$

B.4 Bias error

In this section, we derive upper and lower bounds for $\text{Bias}(\mathbf{w}^*)$ defined in Eq. (3.5). Moreover, we show that the upper and lower bounds match up to constant factors in concrete examples.

B.4.1 An upper bound

Lemma B.11 (Upper bound on the bias term). *Suppose the initial stepsize $\gamma \leq \frac{1}{c \operatorname{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$ for some constant $c > 1$. Then for any $\mathbf{w}^* \in \mathbb{H}$ and $k \in [d]$ such that $r(\mathbf{H}) \geq k + M$, the bias term in (3.5) satisfies*

$$\text{Bias}(\mathbf{w}^*) \lesssim \frac{1}{N_{\text{eff}} \gamma} \|\mathbf{v}^*\|_2^2.$$

Moreover, for any $k \leq M/3$ such that $r(\mathbf{H}) \geq k + M$, the bias term satisfies

$$\text{Bias}(\mathbf{w}^*) \lesssim \frac{\|\mathbf{w}_{0:k}^*\|_2^2}{N_{\text{eff}} \gamma} \cdot \left[\frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)} \right]^2 + \|\mathbf{w}_{k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2$$

with probability $1 - e^{-\Omega(M)}$, where $\mathbf{A}_k := \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top$, $\{\mu_i(\mathbf{A}_k)\}_{i=1}^M$ denote the eigenvalues of $\mathbf{A}_k$ in non-increasing order for some constant $c > 1$.

Proof of Lemma B.11. Similar to the proof of Lemma B.5, we can without loss of generality assume the covariance matrix $\mathbf{H} = \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_d\}$ where $\lambda_i \geq \lambda_j$ for any $i \geq j$. Let $\mathbf{S}\mathbf{H}^{1/2} = \tilde{\mathbf{U}} (\tilde{\Lambda}^{1/2} \mathbf{0}) \tilde{\mathbf{V}}^\top$ be the singular value decomposition of $\mathbf{S}\mathbf{H}\mathbf{S}^\top$, where $\tilde{\Lambda} := \operatorname{diag}\{\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_M\}$ is a diagonal matrix with diagonal entries in non-increasing order. Define $\mathbf{A}_k := \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top$. Then it follows from similar arguments as in Lemma B.5 that $\mathbf{A}_k$ is invertible.

Since

$$\|\gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top\|_2 = \gamma_t \tilde{\lambda}_1 \leq \gamma \tilde{\lambda}_1 \leq \frac{\tilde{\lambda}_1}{c \sum_{i=1}^M \tilde{\lambda}_i} \leq 1$$

for some constant $c > 1$ by the stepsize assumption, it follows that $\mathbf{I}_M - \gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top \succ \mathbf{0}_M$ for all $t \in [N]$. Therefore, it can be verified that

$$\begin{aligned} & \prod_{t=1}^N (\mathbf{I}_M - \gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}\mathbf{H}\mathbf{S}^\top \prod_{t=1}^N (\mathbf{I}_M - \gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top) \\ & \preceq (\mathbf{I}_M - \gamma \mathbf{S}\mathbf{H}\mathbf{S}^\top)^{N_{\text{eff}}} \mathbf{S}\mathbf{H}\mathbf{S}^\top (\mathbf{I}_M - \gamma \mathbf{S}\mathbf{H}\mathbf{S}^\top)^{N_{\text{eff}}} =: \mathbf{M}, \end{aligned}$$

and by definition of $\text{Bias}(\mathbf{w}^*)$ in Eq. (3.5), we have

$$\begin{aligned} \text{Bias}(\mathbf{w}^*) & \approx \left\| \prod_{t=1}^N (\mathbf{I} - \gamma_t \mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{v}^* \right\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2 \leq \left\| (\mathbf{I} - \gamma \mathbf{S}\mathbf{H}\mathbf{S}^\top)^{N_{\text{eff}}} \mathbf{v}^* \right\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2 \\ & = \langle \mathbf{M}, \mathbf{v}^{*\otimes 2} \rangle. \end{aligned} \tag{B.12}$$

Note that the eigenvalues of $\mathbf{M}$ are $\{\tilde{\lambda}_i(1 - \gamma\tilde{\lambda}_i)^{2N_{\text{eff}}}\}_{i=1}^M$. Since the function $f(x) = x(1 - \gamma x)^{2N_{\text{eff}}}$ is maximized at $x_0 = 1/[(2N_{\text{eff}} + 1)\gamma]$ for $x \in [0, 1/\gamma]$ with $f(x_0) \lesssim 1/(N_{\text{eff}}\gamma)$, it follows that

$$\|\mathbf{M}\|_2 \leq c/(N_{\text{eff}}\gamma) \quad (\text{B.13})$$

for some constant $c > 0$. The first part of Lemma B.11 follows immediately.

Now we prove the second part of Lemma B.11. Recall that $\mathbf{v}^* = (\mathbf{SHS}^\top)^{-1}\mathbf{SH}\mathbf{w}^*$. Substituting

$$\mathbf{SH} = (\mathbf{S}_{0:k}\mathbf{H}_{0:k} \quad \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty})$$

into $\mathbf{v}^*$, we obtain

$$\begin{aligned} \langle \mathbf{M}, \mathbf{v}^{*\otimes 2} \rangle &= \langle \mathbf{M}, ((\mathbf{SHS}^\top)^{-1}\mathbf{SH}\mathbf{w}^*)^{\otimes 2} \rangle \\ &= \mathbf{w}^{*\top} \mathbf{HS}^\top (\mathbf{SHS}^\top)^{-1} \mathbf{M} (\mathbf{SHS}^\top)^{-1} \mathbf{SH} \mathbf{w}^* \\ &\leq 2T_1 + 2T_2, \end{aligned}$$

where

$$T_1 := (\mathbf{w}_{0:k}^*)^\top \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top (\mathbf{SHS}^\top)^{-1} \mathbf{M} (\mathbf{SHS}^\top)^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{w}_{0:k}^*, \quad (\text{B.14})$$

$$T_2 := (\mathbf{w}_{k:\infty}^*)^\top \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top (\mathbf{SHS}^\top)^{-1} \mathbf{M} (\mathbf{SHS}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{w}_{k:\infty}^*. \quad (\text{B.15})$$

We claim the following results which we prove later. With probability $1 - e^{-\Omega(M)}$

$$T_1 \leq \frac{c\|\mathbf{w}_{0:k}^*\|_2^2}{N_{\text{eff}}\gamma} \cdot \left[\frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)} \right]^2 \quad (\text{B.16a})$$

for some constant $c > 0$.

$$T_2 \leq \|\mathbf{w}_{k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2. \quad (\text{B.16b})$$

Combining Eq. (B.16a), (B.16b) gives the second part of Lemma B.11.

Proof of claim (B.16a) By definition of T_1 , we have

$$T_1 \leq \|\mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top (\mathbf{SHS}^\top)^{-1} \mathbf{M} (\mathbf{SHS}^\top)^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k}\|_2 \cdot \|\mathbf{w}_{0:k}^*\|_2^2.$$

Moreover,

$$\begin{aligned} &\|\mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top (\mathbf{SHS}^\top)^{-1} \mathbf{M} (\mathbf{SHS}^\top)^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k}\|_2 \\ &\leq \|\mathbf{M}\|_2 \cdot \|(\mathbf{SHS}^\top)^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k}\|_2^2 \\ &\leq \frac{c}{N_{\text{eff}}\gamma} \|(\mathbf{SHS}^\top)^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k}\|_2^2 \end{aligned}$$

for some constant $c > 0$, where the last line uses Eq. (B.13).

It remains to show

$$\|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k}\|_2 \leq c \cdot \frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)} \quad (\text{B.17})$$

for some constant $c > 0$ with probability $1 - e^{-\Omega(M)}$. Since $\mathbf{S}\mathbf{H}\mathbf{S}^\top = \mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top + \mathbf{A}_k$, we have

$$\begin{aligned} (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k} &= (\mathbf{A}_k^{-1} - \mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1})\mathbf{S}_{0:k}\mathbf{H}_{0:k} \\ &= \mathbf{A}_k^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k} - \mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}\mathbf{H}_{0:k} \\ &= \mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\mathbf{H}_{0:k}^{-1}\mathbf{H}_{0:k} \\ &= \mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}, \end{aligned} \quad (\text{B.18})$$

where the second line uses Woodbury's identity. Since

$$\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k} \succeq \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k},$$

it follows that

$$\|[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\|_2 \leq \|[\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\|_2.$$

Therefore, with probability at least $1 - e^{-\Omega(M)}$

$$\begin{aligned} \|\mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\|_2 &\leq \|\mathbf{A}_k^{-1}\|_2 \cdot \|\mathbf{S}_{0:k}\|_2 \cdot \|[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\|_2 \\ &\leq \|\mathbf{A}_k^{-1}\|_2 \cdot \|\mathbf{S}_{0:k}\|_2 \cdot \|[\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\|_2 \\ &\leq \frac{\|\mathbf{A}_k^{-1}\|_2 \cdot \|\mathbf{S}_{0:k}\|_2}{\mu_{\min}(\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k})} \lesssim \frac{\|\mathbf{A}_k^{-1}\|_2}{\mu_{\min}(\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k})} \end{aligned}$$

where the last inequality follows from the fact that $\|\mathbf{S}_{0:k}\|_2 = \sqrt{\|\mathbf{S}_{0:k}^\top\mathbf{S}_{0:k}\|_2} \leq c$ for some constant $c > 0$ when $k \leq M/2$ with probability at least $1 - e^{-\Omega(M)}$. Since $\mathbf{S}_{0:k}$ is independent of $\mathbf{A}_k$ and the distribution of $\mathbf{S}_{0:k}$ is rotationally invariant, we may write $\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k} = \sum_{i=1}^M \frac{1}{\hat{\lambda}_{M-i}} \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i^\top$, where $\tilde{\mathbf{s}}_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_k/M)$ and $(\hat{\lambda}_i)_{i=1}^M$ are eigenvalues of $\mathbf{A}_k$ in non-increasing order. Therefore, for $k \leq M/3$

$$\mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k} = \sum_{i=1}^M \frac{1}{\hat{\lambda}_{M-i}} \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i^\top \succeq \sum_{i=1}^{M/2} \frac{1}{\hat{\lambda}_{M-i}} \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i^\top \succeq \frac{1}{\hat{\lambda}_{M/2}} \sum_{i=1}^{M/2} \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i^\top \succeq \frac{c\mathbf{I}_k}{\hat{\lambda}_{M/2}} \quad (\text{B.19})$$

for some constant $c > 0$ with probability at least $1 - e^{-\Omega(M)}$, where in the last line we again use the concentration properties of Gaussian covariance matrices (see e.g., Theorem 6.1 in [Wainwright 2019a](#)). As a direct consequence, we have

$$\|\mathbf{A}_k^{-1}\mathbf{S}_{0:k}[\mathbf{H}_{0:k}^{-1} + \mathbf{S}_{0:k}^\top\mathbf{A}_k^{-1}\mathbf{S}_{0:k}]^{-1}\|_2 \leq c \cdot \frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)}$$

with probability at least $1 - e^{-\Omega(M)}$ for some constant $c > 0$. This concludes the proof.

Proof of claim (B.16b) By definition of T_2 in Eq. (B.15), we have

$$\begin{aligned}
T_2 &= \mathbf{w}_{k:\infty}^*{}^\top \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top (\mathbf{S} \mathbf{H} \mathbf{S}^\top)^{-1/2} (\mathbf{I}_M - \gamma \mathbf{S} \mathbf{H} \mathbf{S}^\top)^{2N_{\text{eff}}} (\mathbf{S} \mathbf{H} \mathbf{S}^\top)^{-1/2} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{w}_{k:\infty}^* \\
&\leq \mathbf{w}_{k:\infty}^*{}^\top \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top (\mathbf{S} \mathbf{H} \mathbf{S}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{w}_{k:\infty}^* \\
&\leq \|\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top (\mathbf{S} \mathbf{H} \mathbf{S}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\| \cdot \|\mathbf{w}_{k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2 \\
&\leq \|\mathbf{w}_{k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2,
\end{aligned}$$

where the last line follows from

$$\begin{aligned}
&\|\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top (\mathbf{S} \mathbf{H} \mathbf{S}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\|_2 \\
&= \|\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top (\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top + \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\|_2 \\
&\leq \|\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top \mathbf{A}_k^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\|_2 \leq 1.
\end{aligned}$$

□

B.4.2 A lower bound

Lemma B.12 (Lower bound on the bias term). *Suppose $\mathbf{w}^*$ follows some prior distribution and the initial stepsize $\gamma \leq \frac{1}{c \text{tr}(\mathbf{S} \mathbf{H} \mathbf{S}^\top)}$ for some constant $c > 2$. Let $\mathbf{H}^{\mathbf{w}} := \mathbb{E} \mathbf{w}^* \mathbf{w}^{*\top}$. Then the bias term in Eq. (3.5) satisfies*

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) \gtrsim \sum_{i: \tilde{\lambda}_i < 1/(\gamma N_{\text{eff}})} \frac{\mu_i(\mathbf{S} \mathbf{H} \mathbf{H}^{\mathbf{w}} \mathbf{H} \mathbf{S}^\top)}{\mu_i(\mathbf{S} \mathbf{H} \mathbf{S}^\top)}$$

almost surely.

Proof of Lemma B.12. Let $\mathbf{M}_N := \mathbf{S} \mathbf{H} \mathbf{S}^\top (\mathbf{I} - 2\gamma \mathbf{S} \mathbf{H} \mathbf{S}^\top)^{2N_{\text{eff}}}$. Adopt the notations in the proof of Lemma B.11. By definition of the bias term, we have

$$\begin{aligned}
\text{Bias}(\mathbf{w}^*) &\approx \left\| \prod_{t=1}^N (\mathbf{I} - \gamma_t \mathbf{S} \mathbf{H} \mathbf{S}^\top) \mathbf{v}^* \right\|_{\mathbf{S} \mathbf{H} \mathbf{S}^\top}^2 \\
&= \langle \mathbf{S} \mathbf{H} \mathbf{S}^\top \prod_{t=1}^{N/N_{\text{eff}}} (\mathbf{I} - \gamma_{N_{\text{eff}}t} \mathbf{S} \mathbf{H} \mathbf{S}^\top)^{2N_{\text{eff}}} \mathbf{v}^{*\otimes 2} \rangle \\
&\geq \langle \mathbf{S} \mathbf{H} \mathbf{S}^\top (\mathbf{I} - \sum_{t=1}^{N/N_{\text{eff}}} \gamma_{N_{\text{eff}}t} \mathbf{S} \mathbf{H} \mathbf{S}^\top)^{2N_{\text{eff}}} \mathbf{v}^{*\otimes 2} \rangle \\
&\geq \langle \mathbf{S} \mathbf{H} \mathbf{S}^\top (\mathbf{I} - 2\gamma \mathbf{S} \mathbf{H} \mathbf{S}^\top)^{2N_{\text{eff}}} \mathbf{v}^{*\otimes 2} \rangle = \langle \mathbf{M}_N, \mathbf{v}^{*\otimes 2} \rangle, \tag{B.20}
\end{aligned}$$

where the third line uses $\mathbf{I}_M - 2\gamma_t \mathbf{S} \mathbf{H} \mathbf{S}^\top \succ \mathbf{0}_M$ for all $t \in [N]$ established in the proof of Lemma B.11, $\sum_{i=1}^{N/N_{\text{eff}}} \gamma_{N_{\text{eff}}i} \leq 2\gamma$, and the fact that $(1-w)(1-v) \geq 1-w-v$ for $w, v > 0$. Substituting the definition of $\mathbf{v}^*$ in Eq. (3.5) into the expression, we obtain

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) \gtrsim \mathbb{E}_{\mathbf{w}^*} \langle \mathbf{M}_N, \mathbf{v}^{*\otimes 2} \rangle = \mathbb{E}_{\mathbf{w}^*} \langle \mathbf{M}_N, ((\mathbf{S} \mathbf{H} \mathbf{S}^\top)^{-1} \mathbf{S} \mathbf{H} \mathbf{w}^*)^{\otimes 2} \rangle$$

$$\begin{aligned}
&= \text{tr}(\mathbf{H}\mathbf{S}^\top(\mathbf{H}\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{M}_\mathbf{N}(\mathbf{H}\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{H}^\mathbf{w}) \\
&= \text{tr}((\mathbf{H}\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{M}_\mathbf{N}(\mathbf{H}\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{H}^\mathbf{w}\mathbf{H}\mathbf{S}^\top) \\
&\geq \sum_{i=1}^M \mu_{M-i+1}((\mathbf{H}\mathbf{S}\mathbf{S}^\top)^{-1}\mathbf{M}_\mathbf{N}(\mathbf{H}\mathbf{S}\mathbf{S}^\top)^{-1}) \cdot \mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^\mathbf{w}\mathbf{H}\mathbf{S}^\top),
\end{aligned}$$

where the last line uses Von Neumann's trace inequality. Continuing the calculation, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) &\gtrsim \sum_{i=1}^M \frac{\mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^\mathbf{w}\mathbf{H}\mathbf{S}^\top)}{\mu_i((\mathbf{H}\mathbf{S}\mathbf{S}^\top)^2\mathbf{M}_\mathbf{N}^{-1})} \\
&= \sum_{i=1}^M \frac{\mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^\mathbf{w}\mathbf{H}\mathbf{S}^\top)}{\mu_i((\mathbf{H}\mathbf{S}\mathbf{S}^\top)(\mathbf{I} - 2\gamma\mathbf{H}\mathbf{S}\mathbf{S}^\top)^{-2N_{\text{eff}}})} \\
&\gtrsim \sum_{i: \tilde{\lambda}_i < 1/(\gamma N_{\text{eff}})} \frac{\mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^\mathbf{w}\mathbf{H}\mathbf{S}^\top)}{\mu_i(\mathbf{H}\mathbf{S}\mathbf{S}^\top)},
\end{aligned}$$

where the first inequality uses $\mu_{M+i-1}(A) = \mu_i^{-1}(A^{-1})$ for any positive definite matrix $A \in \mathbb{R}^{M \times M}$, and the second line follows from the definition of $\mathbf{M}_\mathbf{N}$ and the fact that $(1 - \lambda\gamma N_{\text{eff}})^{-2N_{\text{eff}}} \lesssim 1$ when $\lambda < 1/(\gamma N_{\text{eff}})$. $\square$

B.4.3 Examples on matching bounds for $\text{Bias}(\mathbf{w}^*)$

In this section, we derive matching upper and lower bounds for $\text{Bias}(\mathbf{w}^*)$ in (3.5) in three scenarios: power-law spectrum (Lemma B.13), power-law spectrum with source condition (Lemma B.14) and logarithmic power-law spectrum (Lemma B.15). Recall that we define $\text{Bias} := \mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*)$.

Lemma B.13 (Bounds on Bias under the power-law spectrum). *Suppose Assumption 3.1C and 3.2 hold and the initial stepsize $\gamma \leq \frac{1}{c \text{tr}(\mathbf{H}\mathbf{S}\mathbf{S}^\top)}$ for some constant $c > 2$. Then with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$*

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) \lesssim \max \{ (N_{\text{eff}}\gamma)^{1/a-1}, M^{1-a} \},$$

and

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) \gtrsim (N_{\text{eff}}\gamma)^{1/a-1}$$

when $(N_{\text{eff}}\gamma)^{1/a} \leq M/c$ for some constant $c > 0$. Here, all the (hidden) constants depend only on the power-law degree a .

Proof of Lemma B.13. For the upper bound, using Lemma B.28, B.11 and the assumption that $\mathbb{E}\mathbf{w}_i^{*2} = 1$ for all $i > 0$, with probability at least $1 - e^{-\Omega(M)}$, we have

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) \lesssim \mathbb{E}_{\mathbf{w}^*} \left[\frac{\|\mathbf{w}_{0:k_2}^*\|_2^2}{N_{\text{eff}}\gamma} + \|\mathbf{w}_{k_2:\infty}^*\|_{\mathbf{H}_{k_2:\infty}}^2 \right]$$

$$\begin{aligned}
&\lesssim \frac{k_2}{N_{\text{eff}}\gamma} + \sum_{k>k_2} \lambda_i \\
&\approx \frac{k_2}{N_{\text{eff}}\gamma} + k_2^{1-a} \\
&\lesssim \max \{ (N_{\text{eff}}\gamma)^{1/a-1}, M^{1-a} \},
\end{aligned}$$

where in the last inequality, we choose $k_2 = [M/3] \wedge (N_{\text{eff}}\gamma)^{1/a}$ to minimize the upper bound.

When $(N_{\text{eff}}\gamma)^{1/a} \leq M/3$, combining Lemma B.12 and B.27 gives the lower bound

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) &\gtrsim \sum_{i: \tilde{\lambda}_i < 1/(N_{\text{eff}}\gamma)} \frac{\mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^w\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)} = \sum_{i: \tilde{\lambda}_i < 1/(N_{\text{eff}}\gamma)} \frac{\mu_i(\mathbf{S}\mathbf{H}^2\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}, \\
&\gtrsim \sum_{\tilde{\lambda}_i < 1/(N_{\text{eff}}\gamma), i \leq M} \frac{i^{-2a}}{i^{-a}} = \sum_{\lambda_i < 1/(N_{\text{eff}}\gamma), i \leq M} i^{-a} \gtrsim (N_{\text{eff}}\gamma)^{1/a-1}
\end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$. Here, the hidden constants depend only on a . $\square$

Lemma B.14 (Bounds on Bias under the source condition). *Suppose Assumption 3.3 hold and the initial stepsize $\gamma \leq \frac{1}{c \text{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$ for some constant $c > 2$. Then with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$*

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) \lesssim \max \{ (N_{\text{eff}}\gamma)^{(1-b)/a}, M^{1-b} \},$$

and

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) \gtrsim (N_{\text{eff}}\gamma)^{(1-b)/a}$$

when $(N_{\text{eff}}\gamma)^{1/a} \leq M/c$ for some constant $c > 0$. Moreover, when $b \geq a + 1$, with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) \lesssim \log N_{\text{eff}} \cdot \max \{ (N_{\text{eff}}\gamma)^{(1-b)/a}, M^{1-b} \}$$

In all results, the hidden constants depend only on a, b .

Proof of Lemma B.14. For the upper bound, using Lemma B.28, B.11 and the assumption that (w.l.o.g.) $\mathbb{E}\mathbf{w}_i^{*2} \approx i^{a-b}$ for all $i > 0$, with probability at least $1 - e^{-\Omega(M)}$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) &\lesssim \mathbb{E}_{\mathbf{w}^*} \left[\frac{\|\mathbf{w}_{0:k_2}^*\|_2^2}{N_{\text{eff}}\gamma} + \|\mathbf{w}_{k_2:\infty}^*\|_{\mathbf{H}_{k_2:\infty}}^2 \right] \\
&\lesssim \frac{k_2^{1+a-b}}{N_{\text{eff}}\gamma} + \sum_{k>k_2} \lambda_i \cdot i^{a-b} \\
&\lesssim \frac{k_2^{1+a-b}}{N_{\text{eff}}\gamma} + k_2^{1-b}
\end{aligned}$$

$$\lesssim \max \{ (N_{\text{eff}}\gamma)^{(1-b)/a}, M^{1-b} \}$$

when $b < a + 1$, where in the last inequality, we choose $k_2 = \lfloor M/3 \rfloor \wedge (N_{\text{eff}}\gamma)^{1/a}$ to minimize the upper bound.

When $(N_{\text{eff}}\gamma)^{1/a} \leq M/c$ for some large constant $c > 0$, combining Lemma B.12 and B.27 yields the lower bound

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) &\gtrsim \sum_{i: \tilde{\lambda}_i < 1/(N_{\text{eff}}\gamma)} \frac{\mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^*\mathbf{H}\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)} \approx \sum_{i: \tilde{\lambda}_i < 1/(N_{\text{eff}}\gamma)} \frac{\mu_i(\mathbf{S}\mathbf{H}^{(a+b)/a}\mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}, \\ &\gtrsim \sum_{\tilde{\lambda}_i < 1/(N_{\text{eff}}\gamma), i \leq M} \frac{i^{-(a+b)}}{i^{-a}} = \sum_{\lambda_i < 1/(N_{\text{eff}}\gamma), i \leq M} i^{-b} \gtrsim (N_{\text{eff}}\gamma)^{(1-b)/a} \end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$. Here, the hidden constants depend only on a, b .

Upper bound when $b \geq a + 1$. Following the previous derivations, when $b = a + 1$, we have with probability at least $1 - e^{-\Omega(M)}$

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) &\lesssim \mathbb{E}_{\mathbf{w}^*} \left[\frac{\|\mathbf{w}_{0:k_2}^*\|_2^2}{N_{\text{eff}}\gamma} + \|\mathbf{w}_{k_2:\infty}^*\|_{\mathbf{H}_{k_2:\infty}}^2 \right] \\ &\lesssim \frac{\log k_2}{N_{\text{eff}}\gamma} + k_2^{1-b} \\ &\lesssim \frac{\log(N_{\text{eff}}\gamma)}{N_{\text{eff}}\gamma} + M^{1-b}, \end{aligned}$$

where the last line follows by setting $k_2 = \lfloor M/3 \rfloor \wedge (N_{\text{eff}}\gamma)^{1/a}$. When $b > a + 1$, we have with probability at least $1 - e^{-\Omega(M)}$

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) &\lesssim \mathbb{E}_{\mathbf{w}^*} \left[\frac{\|\mathbf{w}_{0:k_2}^*\|_2^2}{N_{\text{eff}}\gamma} + \|\mathbf{w}_{k_2:\infty}^*\|_{\mathbf{H}_{k_2:\infty}}^2 \right] \\ &\lesssim \frac{1}{N_{\text{eff}}\gamma} + k_2^{1-b} \\ &\lesssim \frac{1}{N_{\text{eff}}\gamma} + M^{1-b}, \end{aligned}$$

where the last follows by choosing $k_2 = M/3$ to minimize the upper bound.

Note that there exist non-constant gaps between the upper and lower bounds on the bias term (plus the approximation error term, Lemma B.9) in the simple regime where $b \geq a + 1$. We leave a more precise analysis of the bias term for future work. $\square$

Lemma B.15 (Bounds on Bias under the logarithmic power-law spectrum). *Suppose Assumption 3.4 hold and the initial stepsize $\gamma \leq \frac{1}{c \text{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}$ for some constant $c > 2$. Let*

$\bar{k} := \inf\{k : k \log^a k \geq N_{\text{eff}}\gamma\}$. Then with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) \lesssim \max\{\log^{1-a}(N_{\text{eff}}\gamma), \log^{1-a} M\},$$

and

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) \gtrsim \log^{1-a}(N_{\text{eff}}\gamma)$$

when $(N_{\text{eff}}\gamma) \leq M^c$ for some sufficiently small constant $c > 0$. Here, all constants depend only on the power-law degree a .

Proof of Lemma B.15. For the upper bound, using Lemma B.30, B.11 and the assumption that $\mathbb{E} \mathbf{w}_i^{*2} = 1$ for all $i > 0$, with probability at least $1 - e^{-\Omega(M)}$, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) &\lesssim \mathbb{E}_{\mathbf{w}^*} \left[\frac{\|\mathbf{w}_{0:k_2}^*\|_2^2}{N_{\text{eff}}\gamma} + \|\mathbf{w}_{k_2:\infty}^*\|_{\mathbf{H}_{k_2:\infty}}^2 \right] \\ &\lesssim \frac{k_2}{N_{\text{eff}}\gamma} + \sum_{k > k_2} \lambda_i \\ &\approx \frac{k_2}{N_{\text{eff}}\gamma} + \log^{1-a} k_2 \\ &\lesssim \max\{\log^{1-a}(N_{\text{eff}}\gamma), \log^{1-a} M\}, \end{aligned}$$

where in the last inequality, we choose $k_2 = \lceil M/3 \rceil \wedge \lceil (N_{\text{eff}}\gamma)/\log^a(N_{\text{eff}}\gamma) \rceil$ to minimize the upper bound.

Recall $\mathbf{k}^* \approx M/\log M$ (for example we may define $\mathbf{k}^* = \inf\{k : k \log k \geq M\}$) in Lemma B.29. Combining Lemma B.12 and B.29 gives the lower bound

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) &\gtrsim \sum_{i: \tilde{\lambda}_i < 1/(N_{\text{eff}}\gamma)} \frac{\mu_i(\mathbf{S}\mathbf{H}\mathbf{H}^{\text{w}}\mathbf{H}\mathbf{S}^{\top})}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^{\top})} \approx \sum_{i: \tilde{\lambda}_i < 1/(N_{\text{eff}}\gamma)} \frac{\mu_i(\mathbf{S}\mathbf{H}^2\mathbf{S}^{\top})}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^{\top})}, \\ &\gtrsim \sum_{\tilde{\lambda}_i < 1/(N_{\text{eff}}\gamma), i \leq \mathbf{k}^*} \frac{i^{-2} \log^{-2a} i}{i^{-1} \log^{-a} i} = \sum_{\lambda_i < 1/(N_{\text{eff}}\gamma), i \leq \mathbf{k}^*} i^{-1} \log^{-a} i \\ &\gtrsim \sum_{i=N_{\text{eff}}\gamma}^{\mathbf{k}^*} i^{-1} \log^{-a} i \\ &\gtrsim \log^{1-a}(N_{\text{eff}}\gamma) - \log^{1-a}(\mathbf{k}^*) \\ &\gtrsim \log^{1-a}(N_{\text{eff}}\gamma) - c_1 \log^{1-a}(M) \end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$ for some constant $c_1 > 0$. Here, the (hidden) constants depend only on a . Therefore, when $(N_{\text{eff}}\gamma)^{1/a} \leq M^c$ for some sufficiently small constant $c > 0$, we have

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias}(\mathbf{w}^*) \gtrsim \log^{1-a}(N_{\text{eff}}\gamma) - c_1 \log^{1-a}(M) \gtrsim \log^{1-a}(N_{\text{eff}}\gamma).$$

with probability at least $1 - e^{-\Omega(M)}$. □

B.5 Variance error

In this section, we present matching upper and lower bounds on the variance term Var defined in (3.6) under the power-law or logarithmic power-law spectrum.

Lemma B.16 (Matching bounds on Var under power-law spectrum). *Under Assumption 3.2, Var defined in Eq. (3.6) satisfies*

$$\text{Var} \approx \frac{\min\{M, (N_{\text{eff}}\gamma)^{1/a}\}}{N_{\text{eff}}}$$

with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$. Here, the hidden constants only depend on a .

Proof of Lemma B.16. By the definition of Var in Eq. (3.6) and Lemma B.27, we have

$$\begin{aligned} \text{Var} &= \frac{\#\{\tilde{\lambda}_j \geq 1/(N_{\text{eff}}\gamma)\} + (N_{\text{eff}}\gamma)^2 \sum_{\tilde{\lambda}_j < 1/(N_{\text{eff}}\gamma)} \tilde{\lambda}_j^2}{N_{\text{eff}}} \\ &\approx \frac{\min\{M, (N_{\text{eff}}\gamma)^{1/a} + (N_{\text{eff}}\gamma)^2 \cdot (N_{\text{eff}}\gamma)^{(1-2a)/a}\}}{N_{\text{eff}}} \\ &\approx \frac{\min\{M, (N_{\text{eff}}\gamma)^{1/a}\}}{N_{\text{eff}}} \end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$. Here the hidden constants may depend on a . $\square$

Lemma B.17 (Matching bounds on Var under logarithmic power-law spectrum). *Under Assumption 3.4, Var defined in Eq. (3.6) satisfies*

$$\text{Var} \approx \frac{\min\{M, \bar{k}\}}{N_{\text{eff}}} \approx \frac{\min\{M, (N_{\text{eff}}\gamma)/\log^a(N_{\text{eff}}\gamma)\}}{N_{\text{eff}}}$$

with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$, where $\bar{k} := \inf\{k : k \log^a k \geq (N_{\text{eff}}\gamma)\}$ and $\approx$ hides constants that only depend on a .

Proof of Lemma B.17. Define $k^* = \inf\{k : k \log k \geq M\}$ and let

$$\tilde{D} := \#\{\tilde{\lambda}_j \geq 1/(N_{\text{eff}}\gamma)\} + (N_{\text{eff}}\gamma)^2 \sum_{\tilde{\lambda}_j < 1/(N_{\text{eff}}\gamma)} \tilde{\lambda}_j^2.$$

By the definition of Var in (3.6) and Lemma B.29, we have

$$\text{Var} = \frac{\#\{\tilde{\lambda}_j \geq 1/(N_{\text{eff}}\gamma)\} + (N_{\text{eff}}\gamma)^2 \sum_{\tilde{\lambda}_j < 1/(N_{\text{eff}}\gamma)} \tilde{\lambda}_j^2}{N_{\text{eff}}}$$

$$= \frac{\tilde{D}}{N_{\text{eff}}} \approx \frac{\min\{M, \bar{k}\}}{N_{\text{eff}}}$$

with probability at least $1 - e^{-\Omega(M)}$ over the randomness of $\mathbf{S}$, where the second line follows from

$$\begin{aligned} \tilde{D} &\gtrsim \#\{\tilde{\lambda}_j \geq 1/(N_{\text{eff}}\gamma)\} \approx \frac{N_{\text{eff}}\gamma}{\log^a(N_{\text{eff}}\gamma)}, \quad \text{and} \\ \tilde{D} &\lesssim \frac{N_{\text{eff}}\gamma}{\log^a(N_{\text{eff}}\gamma)} + \frac{(N_{\text{eff}}\gamma)^2}{\log^{2a}(N_{\text{eff}}\gamma)} \cdot \sum_{j: \tilde{\lambda}_j < 1/(N_{\text{eff}}\gamma)} \frac{1}{j^2} \lesssim \frac{N_{\text{eff}}\gamma}{\log^a(N_{\text{eff}}\gamma)} \end{aligned}$$

when $\bar{k} \lesssim M$. □

B.6 Expected risk of the average of (SGD) iterates

In this section, we study the expected risk of the average of (SGD) iterates. Namely, we consider a fixed stepsize (SGD) procedure where $\gamma_t = \gamma$ and define $\bar{\mathbf{v}}_N := \sum_{i=0}^{N-1} \mathbf{v}_i / N$. Our goal is to derive matching upper and lower bounds $\mathcal{R}(\bar{\mathbf{v}}_N)$ in terms of the sample size N and model size M . Compared with the last iterate of (SGD) with geometrically decaying stepsizes, we show that the average of (SGD) iterates with a fixed stepsize achieves a better risk, in the sense that the effective sample size N_{eff} is replaced by N in the bounds (c.f. Theorem 3.1). This may give improvement up to logarithmic factors.

We start with invoking the following result in Zou et al. (2023).

Theorem B.18 (A variant of Theorem 2.1 and 2.2 in Zou et al. (2023)). *Suppose Assumption 3.1 hold. Consider an M -dimensional sketched predictor trained by fixed stepsize (SGD) with N samples. Let $\bar{\mathbf{v}}_N := \sum_{i=0}^{N-1} \mathbf{v}_i / N$ be the average of the iterations of SGD. Assume $\mathbf{v}_0 = \mathbf{0}$ and $\sigma^2 \gtrsim 1$. Conditional on $\mathbf{S}$ and suppose the stepsize $\gamma < 1/(\text{ctr}(\mathbf{SHS}^\top))$ for some constant $c > 0$, then there exist Approx, Bias, Var such that*

$$\mathbb{E}\mathcal{R}_M(\bar{\mathbf{v}}_N) - \sigma^2 \approx \mathbb{E}_{\mathbf{w}^*} \text{Approx} + \text{Bias} + \sigma^2 \text{Var},$$

where the expectation of $\mathcal{R}_M$ is over $\mathbf{w}^*$ and $(\mathbf{x}_i, y_i)_{i=1}^N$ and

$$\begin{aligned} \text{Approx} &:= \mathbb{E}\xi^2 \\ &= \left\| \left(\mathbf{I} - \mathbf{H}^{\frac{1}{2}} \mathbf{S}^\top (\mathbf{SHS}^\top)^{-1} \mathbf{SH}^{\frac{1}{2}} \right) \mathbf{H}^{\frac{1}{2}} \mathbf{w}^* \right\|^2, \end{aligned}$$

$$\mathbb{E}_{\mathbf{w}^*}(T_1 + T_3) \lesssim \text{Bias} \lesssim \mathbb{E}_{\mathbf{w}^*}(T_2 + T_4),$$

$$\text{Var} \approx \frac{D_{\text{eff}, N}}{N},$$

and

$$T_1 := \frac{1}{\gamma^2 N^2} \text{tr} \left(\left(\mathbf{I} - (\mathbf{I} - \gamma \mathbf{S} \mathbf{H} \mathbf{S}^\top)^{N/4} \right)^2 (\mathbf{S} \mathbf{H} \mathbf{S}^\top)^{-1} \mathbf{B}_0 \right), \quad (\text{B.21a})$$

$$T_2 := \frac{1}{\gamma^2 N^2} \text{tr} \left(\left(\mathbf{I} - (\mathbf{I} - \gamma \mathbf{S} \mathbf{H} \mathbf{S}^\top)^N \right)^2 (\mathbf{S} \mathbf{H} \mathbf{S}^\top)^{-1} \mathbf{B}_0 \right), \quad (\text{B.21b})$$

$$T_3 := \frac{1}{\gamma N^2} \text{tr} \left(\left(\mathbf{I} - (\mathbf{I} - \gamma \mathbf{S} \mathbf{H} \mathbf{S}^\top)^{N/4} \right) \mathbf{B}_0 \right) \cdot \text{tr} \left(\left(\mathbf{I} - (\mathbf{I} - \gamma \mathbf{S} \mathbf{H} \mathbf{S}^\top)^{N/4} \right)^2 \right), \quad (\text{B.21c})$$

$$T_4 := \frac{1}{\gamma N} \text{tr} \left(\mathbf{B}_0 - (\mathbf{I} - \gamma \mathbf{S} \mathbf{H} \mathbf{S}^\top)^N \mathbf{B}_0 (\mathbf{I} - \gamma \mathbf{S} \mathbf{H} \mathbf{S}^\top)^N \right) \cdot \frac{D_{\text{eff},N}}{N}, \quad (\text{B.21d})$$

$$\mathbf{B}_0 := \mathbf{v}^* \mathbf{v}^{*\top}, \quad (\text{B.21e})$$

$$D_{\text{eff},N} := \#\{\tilde{\lambda}_j \geq 1/(N\gamma)\} + (N\gamma)^2 \sum_{\tilde{\lambda}_j < 1/(N\gamma)} \tilde{\lambda}_j^2, \quad (\text{B.21f})$$

where $(\tilde{\lambda}_j)_{j=1}^M$ are eigenvalue of $\mathbf{S} \mathbf{H} \mathbf{S}^\top$.

See Section B.6.2 for the proof.

For T_i ($i = 1, 2, 3, 4$), we also have the following upper (and lower) bounds.

Lemma B.19 (Lower bound on T_1). *Under the assumptions and notations in Theorem B.18, we have*

$$\mathbb{E}_{\mathbf{w}^*} T_1 \gtrsim \sum_{i: \tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{S} \mathbf{H}^2 \mathbf{S}^\top)}{\mu_i(\mathbf{S} \mathbf{H} \mathbf{S}^\top)}$$

almost surely, where $(\tilde{\lambda}_i)_{i=1}^N$ are eigenvalues of $\mathbf{S} \mathbf{H} \mathbf{S}^\top$ in non-increasing order.

See the proof in Section B.6.2.

Lemma B.20 (Upper bound on T_2). *Under the assumptions and notations in Theorem B.18, for any $k \leq M/3$ such that $r(\mathbf{H}) \geq k + M$, we have with probability at least $1 - e^{-\Omega(M)}$ that*

$$T_2 \lesssim \frac{1}{N\gamma} \left[\frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)} \right]^2 \cdot \|\mathbf{w}_{0:k}^*\|^2 + \|\mathbf{w}_{k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2,$$

where $\mathbf{A}_k := \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top$.

See the proof in Section B.6.2.

Lemma B.21 (Lower bound on T_3). *Under the assumptions and notations in Theorem B.18, we have*

$$\mathbb{E}_{\mathbf{w}^*} T_3 \gtrsim \frac{D_{\text{eff},N}}{N} \cdot \sum_{i: \tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{S} \mathbf{H}^2 \mathbf{S}^\top)}{\mu_i(\mathbf{S} \mathbf{H} \mathbf{S}^\top)}$$

almost surely, where $(\tilde{\lambda}_i)_{i=1}^M$ are eigenvalues of $\mathbf{S} \mathbf{H} \mathbf{S}^\top$ in non-increasing order.

See the proof in Section B.6.2.

Lemma B.22 (Upper bound on T_4). *Under the assumptions and notations in Theorem B.18 and assume $r(\mathbf{H}) \geq M$, we have*

$$T_4 \lesssim \|\mathbf{w}^*\|_{\mathbf{H}}^2 \cdot \frac{D_{\text{eff},N}}{N}$$

almost surely, where $\mathbf{A}_k := \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top$.

See the proof in Section B.6.2.

With these results at hand, we are ready to derive upper and lower bounds for the risk of the average of (SGD) iterates.

B.6.1 Matching bounds for the average of (SGD) iterates under power-law spectrum

In this section, we derive upper and lower bounds for the expected risk under the power-law spectrum. Our main result (Theorem B.23) follows directly from Theorem B.18 and the bounds on $T_i (i = 1, 2, 3, 4)$ in Lemmas B.19 to B.22.

Theorem B.23 (Scaling law for average iterates of SGD). *Suppose Assumption 3.1 and 3.2 hold and $\sigma^2 \lesssim 1$. Then there exists some a -dependent constant $c > 0$ such that when $\gamma \leq c$, with probability at least $1 - e^{-\Omega(M)}$ over the randomness of the sketch matrix $\mathbf{S}$, we have*

$$\mathbb{E}\mathcal{R}_M(\bar{\mathbf{v}}_N) = \sigma^2 + \Theta(M^{1-a}) + \Theta((N\gamma)^{1/a-1}),$$

where the expectation is over the randomness of $\mathbf{w}^*$ and $(\mathbf{x}_i, y_i)_{i=1}^N$, and $\Theta(\cdot)$ hides constants that may depend on a .

See the proof in Section B.6.2.

Compared with Theorem 3.1, Theorem B.23 suggests that the average of (SGD) achieves a smaller risk in the sketched linear model—the $(N_{\text{eff}}\gamma)^{1/a}$ is replaced by $(N\gamma)^{1/a}$ in the bound for the bias term. This is intuitive since the sum of stepsizes $\sum_t \gamma_t \approx N_{\text{eff}}\gamma$ for the geometrically decaying stepsize scheduler while $\sum_t \gamma_t \approx N\gamma$ for the fixed stepsize scheduler.

We also verify the observations in Theorem B.23 via simulations. We adopt the same model and setup as in Section 3.5 but use the average of iterates of fixed stepsize (SGD) (denoted by $\bar{\mathbf{v}}_N$) as the predictor. From Figure B.1 and B.2 we see that the expected risk $\mathbb{E}\mathcal{R}(\bar{\mathbf{v}}_N)$ also scales following a power-law relation in both sample size N and model size M . Moreover, the fitted exponents match our theoretical predictions in Theorem B.23.

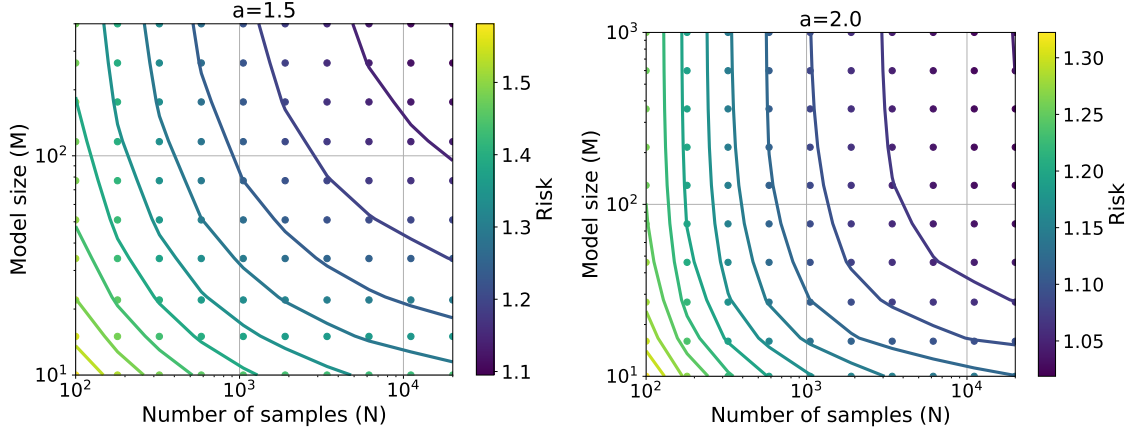


Figure B.1: The expected risk (Risk) of the average of iterates of (SGD) versus the sample size N and the model size M for different power-law degrees a . The expected risk is computed by averaging over 1000 independent samples of $(\mathbf{w}^*, \mathbf{S})$. We fit the expected risk using the formula $\text{Risk} \sim \sigma^2 + c_1/M^{a_1} + c_2/N^{a_2}$ via minimizing the Huber loss as in Hoffmann et al. (2022). Parameters: $\sigma = 1, \gamma = 0.1$. Left: For $a = 1.5, d = 20000$, the fitted exponents are $(a_1, a_2) = (0.59, 0.33) \approx (0.5, 0.33)$. Right: For $a = 2, d = 2000$, the fitted exponents are $(a_1, a_2) = (1.09, 0.49) \approx (1.0, 0.5)$. Note that the values of (a_1, a_2) are close to our theoretical predictions $(a - 1, 1 - 1/a)$ in both cases.

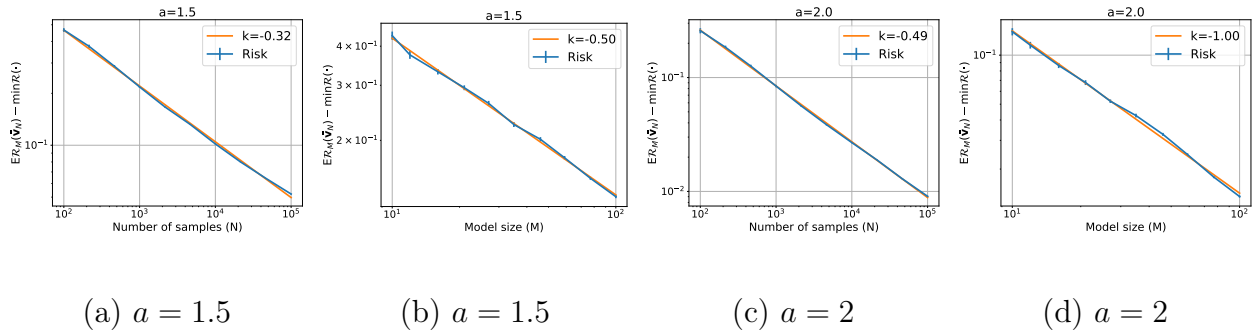


Figure B.2: The expected risk of the average of iterates of (SGD) minus the irreducible risk versus the effective sample size and model size. Parameters $\sigma = 1, \gamma = 0.1$. (a), (b): $a = 1.5, d = 10000$; (c), (d): $a = 2, d = 1000$. The error bars denote the ± 1 standard deviation of estimating the expected risk using 100 independent samples of $(\mathbf{w}^*, \mathbf{S})$. We use linear functions to fit the expected risk under log-log scale and report the slope of the fitted lines (denoted by k).

B.6.2 Proofs

Proof of Theorem B.18

Similar to the proof of Theorem B.4, we have the decomposition

$$\mathcal{R}(\bar{\mathbf{v}}_N) = \sigma^2 + \text{Approx} + \|\bar{\mathbf{v}}_N - \mathbf{v}^*\|_{\mathbf{S}\mathbf{H}\mathbf{S}^\top}^2.$$

Note that $(\mathbf{v}_t)_{t=1}^N$ can also be viewed as the SGD iterates on the model $y = \langle \mathbf{S}\mathbf{x}, \mathbf{v}^* \rangle + \xi + \epsilon$, where the noise satisfies

$$\mathbb{E}(\xi + \epsilon)^2 = \mathcal{R}(\mathbf{v}^*) = \mathbb{E}\xi^2 + \sigma^2.$$

Therefore, the upper and lower bounds on **Bias**, **Var** follow directly from the proof of Theorem 2.1, 2.2 and related lemmas (Lemma B.6, B.11, C.3, C.5) in Zou et al. (2023).

Proof of Lemma B.19

For any positive definite matrix $A \in \mathbb{R}^{M \times M}$, let

$$f_1(A) := (\mathbf{I} - (\mathbf{I} - \gamma A)^{N/4})^2 A^{-1} / \gamma^2 / N^2.$$

Since $\gamma \leq 1/(c \text{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top))$, we have $f_1(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \succeq \mathbf{0}$. By definition of T_1 and recalling $\mathbf{v}^* = (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}\mathbf{H}\mathbf{w}^*$, we have with probability at least $1 - e^{-\Omega(M)}$ that

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} T_1 &= \mathbb{E}_{\mathbf{w}^*} [\mathbf{w}^{*\top} \mathbf{H}\mathbf{S}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} f_1(\mathbf{S}\mathbf{H}\mathbf{S}^\top) (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}\mathbf{H}\mathbf{w}^*] \\ &= \text{tr} \left([(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} f_1(\mathbf{S}\mathbf{H}\mathbf{S}^\top) (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}] (\mathbf{S}\mathbf{H}^2 \mathbf{S}^\top) \right). \end{aligned}$$

Following the proof of Lemma B.12 (by Von Neumann's trace inequality), we have

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} T_1 &\geq \sum_{i=1}^M \frac{\mu_i(\mathbf{S}\mathbf{H}^2 \mathbf{S}^\top)}{\mu_i \left((\mathbf{S}\mathbf{H}\mathbf{S}^\top)^2 f_1(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \right)} \\ &\geq \sum_{i: \tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{S}\mathbf{H}^2 \mathbf{S}^\top)}{\mu_i \left((\mathbf{S}\mathbf{H}\mathbf{S}^\top)^2 f_1(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \right)} \\ &\gtrsim \sum_{i: \tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{S}\mathbf{H}^2 \mathbf{S}^\top)}{\mu_i(\mathbf{S}\mathbf{H}\mathbf{S}^\top)}, \end{aligned}$$

where the third inequality is due to

$$\lambda / f_1(\lambda) \lesssim \frac{\lambda^2 \gamma^2 N^2}{(1 - (1 - \gamma \lambda)^{N/4})^2} \lesssim \frac{N^2}{(\sum_{i=0}^{N/4-1} (1 - \gamma \lambda)^i)^2} \lesssim \frac{1}{(1 - \gamma \lambda)^{2N}} \lesssim 1$$

when $\lambda < 1/(N\gamma)$.

Proof of Lemma B.20

By definition of T_2 , the fact that $1 - x^N = (1 - x) \sum_{i=0}^{N-1} x^i$, and recalling $\mathbf{v}^* = (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}\mathbf{H}\mathbf{w}^*$, we have

$$\begin{aligned} T_2 &= \mathbf{w}^{*\top} \mathbf{H}\mathbf{S}^\top f_2(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}\mathbf{H}\mathbf{w}^*, \\ &\leq 2 \underbrace{[\mathbf{w}_{0:k}^{*\top} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top f_2(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{w}_{0:k}^*]}_{T_{21}} + \underbrace{[\mathbf{w}_{k:\infty}^{*\top} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top f_2(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{w}_{k:\infty}^*]}_{T_{22}}, \end{aligned}$$

where $f_2(A) := [\sum_{i=0}^{N-1} (\mathbf{I} - \gamma A)^i]^2 / A / N^2$ for any symmetric matrix $A \in \mathbb{R}^{M \times M}$. Moreover, we have

$$\begin{aligned} T_{21} &= \mathbf{w}_{0:k}^{*\top} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top f_2(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{w}_{0:k}^* \\ &\leq \|f_2(\mathbf{S}\mathbf{H}\mathbf{S}^\top) (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^2\| \cdot \|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{w}_{0:k}^*\|^2. \end{aligned}$$

Using the assumption on the stepsize that $\gamma \leq 1/(c \operatorname{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top))$, we have

$$\begin{aligned} \|f_2(\mathbf{S}\mathbf{H}\mathbf{S}^\top) (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^2\| &\leq \max_{\lambda \in [0, 1/\gamma]} \frac{1}{N^2} \left[\sum_{i=0}^{N-1} (1 - \gamma\lambda)^i \right]^2 \lambda \\ &= \max_{\lambda \in [0, 1/\gamma]} \frac{1}{N^2 \gamma} \left[\sum_{i=0}^{N-1} (1 - \gamma\lambda)^i \right] \cdot (1 - (1 - \gamma\lambda)^N) \\ &\leq \frac{1}{N^2 \gamma} \cdot N \cdot 1 = \frac{1}{N\gamma}. \end{aligned} \tag{B.22}$$

Combining Eq. (B.22) with Eq. (B.17) in the proof of Lemma B.11 (note that we assume $k \leq M/3$), we obtain

$$T_{21} \leq c \frac{1}{N\gamma} \left[\frac{\mu_{M/2}(\mathbf{A}_k)}{\mu_M(\mathbf{A}_k)} \right]^2 \cdot \|\mathbf{w}_{0:k}^*\|^2$$

for some constant $c > 0$ with probability at least $1 - e^{-\Omega(M)}$. For T_{22} , we have

$$\begin{aligned} T_{22} &= \mathbf{w}_{k:\infty}^{*\top} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top f_2(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{w}_{k:\infty}^* \\ &\leq \|f_2(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}\mathbf{H}\mathbf{S}^\top\| \cdot \|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1/2} (\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}) \mathbf{H}_{k:\infty}^{1/2} \mathbf{w}_{k:\infty}^*\|^2 \\ &\leq \|f_2(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}\mathbf{H}\mathbf{S}^\top\| \cdot \|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1/2} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\|^2 \cdot \|\mathbf{w}_{k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2. \end{aligned}$$

Since $\|f_2(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}\mathbf{H}\mathbf{S}^\top\| = \|[\sum_{i=0}^{N-1} (\mathbf{I} - \gamma \mathbf{S}\mathbf{H}\mathbf{S}^\top)^i]^2 / N^2\| \leq 1$ by the assumption $\gamma \leq 1/(c \operatorname{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top))$, and

$$\begin{aligned} \|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1/2} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\|^2 &= \|\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\| \\ &= \|\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top (\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top + \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\| \\ &\leq \|\mathbf{H}_{k:\infty}^{1/2} \mathbf{S}_{k:\infty}^\top (\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top)^{-1} \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^{1/2}\| = 1, \end{aligned}$$

it follows that $T_{22} \leq \|\mathbf{w}_{k:\infty}^*\|_{\mathbf{H}_{k:\infty}}^2$. Combining the bounds on T_{21}, T_{22} completes the proof.

Proof of Lemma B.21

Let $f_3(A) := (\mathbf{I} - (\mathbf{I} - \gamma \mathbf{A})^{N/4})/\gamma/N^2$ for any positive definite matrix $A \in \mathbb{R}^{M \times M}$. Following the same arguments as in the proof of Lemma B.19, we have $f_3(\mathbf{SHS}^\top) \succeq \mathbf{0}$ and

$$\begin{aligned} & \mathbb{E}_{\mathbf{w}^*} \left[\frac{1}{\gamma N^2} \text{tr} \left(\left(\mathbf{I} - (\mathbf{I} - \gamma \mathbf{SHS}^\top)^{N/4} \right) \mathbf{B}_0 \right) \right] \\ &= \mathbb{E}_{\mathbf{w}^*} [\mathbf{w}^{*\top} \mathbf{H} \mathbf{S}^\top (\mathbf{SHS}^\top)^{-1} f_3(\mathbf{SHS}^\top) (\mathbf{SHS}^\top)^{-1} \mathbf{H} \mathbf{w}^*] \\ &= \text{tr}((\mathbf{SHS}^\top)^{-1} f_3(\mathbf{SHS}^\top) (\mathbf{SHS}^\top)^{-1} \mathbf{H}^2 \mathbf{S}^\top). \end{aligned}$$

Moreover,

$$\begin{aligned} \mathbb{E}_{\mathbf{w}^*} T_1 &\geq \sum_{i=1}^M \frac{\mu_i(\mathbf{SH}^2 \mathbf{S}^\top)}{\mu_i \left((\mathbf{SHS}^\top)^2 f_3(\mathbf{SHS}^\top)^{-1} \right)} \\ &\geq \sum_{i: \tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{SH}^2 \mathbf{S}^\top)}{\mu_i \left((\mathbf{SHS}^\top)^2 f_3(\mathbf{SHS}^\top)^{-1} \right)} \\ &\gtrsim \frac{1}{N} \sum_{i: \tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{SH}^2 \mathbf{S}^\top)}{\mu_i(\mathbf{SHS}^\top)}, \end{aligned} \tag{B.23}$$

where the third inequality is due to

$$\lambda/f_3(\lambda) \lesssim \frac{\lambda \gamma N^2}{1 - (1 - \gamma \lambda)^{N/4}} \lesssim \frac{N^2}{\sum_{i=0}^{N/4-1} (1 - \gamma \lambda)^i} \lesssim \frac{N}{(1 - \gamma \lambda)^N} \lesssim N$$

when $\lambda < 1/(N\gamma)$. Note that

$$1 - (1 - \gamma \tilde{\lambda}_i)^{N/4} \geq \begin{cases} 1 - (1 - \frac{1}{N})^{N/4} \geq 1 - e^{-1/4} \geq \frac{1}{5}, & \tilde{\lambda}_i \geq \frac{1}{\gamma N}, \\ \frac{N}{4} \cdot \gamma \tilde{\lambda}_i - \frac{N(N-4)}{32} \cdot \gamma^2 \tilde{\lambda}_i^2 \geq \frac{N}{5} \cdot \gamma \tilde{\lambda}_i, & \tilde{\lambda}_i < \frac{1}{\gamma N}, \end{cases} \geq \frac{1}{5} \min\{N\gamma \tilde{\lambda}_i, 1\}. \tag{B.24}$$

We thus have

$$\begin{aligned} \text{tr} \left((\mathbf{I} - (\mathbf{I} - \gamma \mathbf{SHS}^\top)^{N/4})^2 \right) &= \sum_{i=1}^M [1 - (1 - \gamma \tilde{\lambda}_i)^{N/4}]^2 \gtrsim \sum_{i=1}^M \min\{(N\gamma \tilde{\lambda}_i)^2, 1\} \\ &= \#\{\tilde{\lambda}_i \geq \frac{1}{N\gamma}\} + N^2 \gamma^2 \sum_{\tilde{\lambda}_i < 1/(N\gamma)} \tilde{\lambda}_i^2 = D_{\text{eff}, N}. \end{aligned} \tag{B.25}$$

Combining Eq. (B.25) and (B.23) completes the proof.

Proof of Lemma B.22

Substituting $\mathbf{v}^* = (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1}\mathbf{S}\mathbf{H}\mathbf{w}^*$ in the expression of T_4 and noting $\mathbf{v}_0 = \mathbf{0}$, we have

$$\begin{aligned} T_4 &= \frac{1}{\gamma N} \text{tr} \left(\mathbf{B}_0 - (\mathbf{I} - \gamma \mathbf{S}\mathbf{H}\mathbf{S}^\top)^N \mathbf{B}_0 (\mathbf{I} - \gamma \mathbf{S}\mathbf{H}\mathbf{S}^\top)^N \right) \cdot \frac{D_{\text{eff},N}}{N} \\ &= \frac{1}{\gamma N} \text{tr} \left(\mathbf{w}^{*\top} \mathbf{H}\mathbf{S}^\top (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} [\mathbf{I} - (\mathbf{I} - \gamma \mathbf{S}\mathbf{H}\mathbf{S}^\top)^{2N}] (\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1} \mathbf{S}\mathbf{H}\mathbf{w}^* \right) \cdot \frac{D_{\text{eff},N}}{N} \\ &=: \text{tr} \left(\mathbf{w}^{*\top} \mathbf{H}\mathbf{S}^\top f_4(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}\mathbf{H}\mathbf{w}^* \right) \cdot \frac{D_{\text{eff},N}}{N}, \end{aligned} \quad (\text{B.26})$$

where $f_4(A) := A^{-1}[\mathbf{I} - (\mathbf{I} - \gamma A)^{2N}]A^{-1}/(N\gamma)$ for any symmetric matrix $A \in \mathbb{R}^{M \times M}$. Moreover,

$$\begin{aligned} &\text{tr} \left(\mathbf{w}^{*\top} \mathbf{H}\mathbf{S}^\top f_4(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}\mathbf{H}\mathbf{w}^* \right) \\ &\leq \|f_4(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}\mathbf{H}\mathbf{S}^\top\| \cdot \|(\mathbf{S}\mathbf{H}\mathbf{S}^\top)^{-1/2} \mathbf{S}\mathbf{H}^{1/2}\|^2 \cdot \|\mathbf{w}_*\|_{\mathbf{H}}^2 \\ &\leq \|f_4(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}\mathbf{H}\mathbf{S}^\top\| \cdot \|\mathbf{w}_*\|_{\mathbf{H}}^2. \end{aligned}$$

Since

$$\|f_4(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}\mathbf{H}\mathbf{S}^\top\| = \frac{1}{N} \left\| \sum_{i=0}^{2N-1} (\mathbf{I} - \gamma \mathbf{S}\mathbf{H}\mathbf{S}^\top)^i \right\| \leq 2$$

by our assumption on the stepsize, it follows that

$$\text{tr} \left(\mathbf{w}^{*\top} \mathbf{H}\mathbf{S}^\top f_4(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \mathbf{S}\mathbf{H}\mathbf{w}^* \right) \lesssim \|\mathbf{w}_*\|_{\mathbf{H}}^2. \quad (\text{B.27})$$

Combining Eq. (B.26) and (B.27) we find

$$T_4 \lesssim \|\mathbf{w}_*\|_{\mathbf{H}}^2 \cdot \frac{D_{\text{eff},N}}{N}.$$

Proof of Theorem B.23

First, by Lemma B.27 we have $1/\text{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top) \gtrsim c_1$ for some a -dependent $c_1 > 0$ with probability at least $1 - e^{-\Omega(M)}$. Therefore we may choose c sufficiently small so that $\gamma \leq c$ implies $\gamma \lesssim 1/\text{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)$ with probability at least $1 - e^{-\Omega(M)}$. Now, suppose we have $\gamma \lesssim 1/\text{tr}(\mathbf{S}\mathbf{H}\mathbf{S}^\top)$. Following the notations in Theorem B.18, we claim the following bounds on **Approx**, **Bias**, **Var**:

$$\mathbb{E}\text{Approx} \asymp M^{1-a} \quad (\text{B.28a})$$

$$\text{Var} \asymp \min \{M, (N\gamma)^{1/a}\} / N. \quad (\text{B.28b})$$

$$\text{Bias} \lesssim \max \{M^{1-a}, (N\gamma)^{1/a-1}\}, \quad (\text{B.28c})$$

$$\text{Bias} \gtrsim (N\gamma)^{1/a-1} \text{ when } (N\gamma)^{1/a} \leq M/c \text{ for some constant } c > 0, \quad (\text{B.28d})$$

with probability at least $1 - e^{-\Omega(M)}$. Putting the bounds together yields Theorem B.23.

Proof of claim (B.28a) Note that our definition of **Approx** in Theorem B.18 is the same as that in Eq. (3.4) (and B.1). Therefore the claim follows immediately from Lemma B.8.

Proof of claim (B.28b) This follows from the proof of Lemma B.16 with N_{eff} replaced by N .

Proof of claim (B.28c) By Theorem B.18, Lemma B.20 and B.22, we have

$$\begin{aligned} \text{Bias} &\lesssim \mathbb{E}_{\mathbf{w}^*} \frac{\|\mathbf{w}_{0:k_2}^*\|_2^2}{N\gamma} \cdot \left[\frac{\mu_{M/2}(\mathbf{S}_{k_2:\infty} \mathbf{H}_{k_2:\infty} \mathbf{S}_{k_2:\infty}^\top)}{\mu_M(\mathbf{S}_{k_2:\infty} \mathbf{H}_{k_2:\infty} \mathbf{S}_{k_2:\infty}^\top)} \right]^2 + \mathbb{E}_{\mathbf{w}^*} \|\mathbf{w}_{k_2:\infty}^*\|_{\mathbf{H}_{k_2:\infty}}^2 + \sigma^2 \frac{D_{\text{eff},N}}{N}, \\ &\lesssim \frac{k_2}{N\gamma} \left[\frac{\mu_{M/2}(\mathbf{S}_{k_2:\infty} \mathbf{H}_{k_2:\infty} \mathbf{S}_{k_2:\infty}^\top)}{\mu_M(\mathbf{S}_{k_2:\infty} \mathbf{H}_{k_2:\infty} \mathbf{S}_{k_2:\infty}^\top)} \right]^2 + k_2^{1-a} + \frac{D_{\text{eff},N}}{N} \end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$ for any $k_2 \leq M/3$. Choosing $k_2 = \min\{M/3, (N\gamma)^{1/a}\}$ and using Lemma B.27 and claim (B.28b), we obtain

$$\begin{aligned} \text{Bias} &\lesssim \max\{M^{1-a}, (N\gamma)^{1/a-1}\} + \frac{\min\{M, (N\gamma)^{1/a}\}}{N} \\ &\lesssim \max\{M^{1-a}, (N\gamma)^{1/a-1}\} + (N\gamma)^{1/a-1} \lesssim \max\{M^{1-a}, (N\gamma)^{1/a-1}\} \end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$.

Proof of claim (B.28d) By Theorem B.18 and Lemma B.19, we have

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias} \gtrsim \sum_{i: \tilde{\lambda}_i < 1/(\gamma N)} \frac{\mu_i(\mathbf{S} \mathbf{H}^2 \mathbf{S}^\top)}{\mu_i(\mathbf{S} \mathbf{H} \mathbf{S}^\top)}.$$

When $(N\gamma)^{1/a} \leq M/c$ for some large constant $c > 0$, we have from Lemma B.27 that

$$\mathbb{E}_{\mathbf{w}^*} \text{Bias} \gtrsim \sum_{i: \tilde{\lambda}_i < 1/(\gamma N)} \frac{i^{-2a}}{i^{-a}} = \sum_{i: \tilde{\lambda}_i < 1/(\gamma N)} i^{-a} \gtrsim [(N\gamma)^{1/a}]^{1-a} = (N\gamma)^{1/a-1}$$

with probability at least $1 - e^{-\Omega(M)}$.

B.7 Concentration lemmas

B.7.1 General concentration results

Lemma B.24. Suppose that $\mathbf{S} \in \mathbb{R}^{M \times d}$ is such that ¹

$$\mathbf{S}_{ij} \sim \mathcal{N}(0, 1/M).$$

¹We allow $d = \infty$.

Let $(\lambda_i)_{i \geq 1}$ be the eigenvalues of $\mathbf{H}$ in non-increasing order. Let $(\tilde{\lambda}_i)_{i=1}^M$ be the eigenvalues of $\mathbf{SHS}^\top$ in non-increasing order. Then there exists a constant $c > 1$ such that for every $M \geq 0$ and every $0 \leq k \leq M$, with probability $\geq 1 - e^{-\Omega(M)} - e^{-\Omega(k)}$, we have

$$\text{for every } j \leq M, \quad \left| \tilde{\lambda}_j - \left(\lambda_j + \frac{\sum_{i>k} \lambda_i}{M} \right) \right| \leq c \cdot \left(\sqrt{\frac{k}{M}} \cdot \lambda_j + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M}} \right).$$

As a direct consequence, for $k \leq M/c^2$, we have

$$\text{for every } j \leq M, \quad \left| \tilde{\lambda}_j - \left(\lambda_j + \frac{\sum_{i>k} \lambda_i}{M} \right) \right| \leq \frac{1}{2} \cdot \left(\lambda_j + \frac{\sum_{i>k} \lambda_i}{M} \right) + c_1 \cdot \lambda_{k+1},$$

where $c_1 = c + 2c^2$.

Proof of Lemma B.24. We have the following decomposition motivated by [Swartworth and Woodruff \(2023\)](#) (see their Section 3.4, Proof of Theorem 1).

$$\begin{aligned} \mathbf{SHS}^\top &= \mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top + \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top \\ &= \mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top + \frac{\sum_{i>k} \lambda_i}{M} \cdot \mathbf{I}_M + \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k} \lambda_i}{M} \cdot \mathbf{I}_M. \end{aligned}$$

We remark that this decomposition idea has been implicitly used in [Bartlett et al. \(2020\)](#) to control the eigenvalues of a Gram matrix. In fact, we will use techniques from [Bartlett et al. \(2020\)](#) to obtain a sharper bound than that presented in [Swartworth and Woodruff \(2023\)](#).

For the upper bound, we have

$$\begin{aligned} \mu_j(\mathbf{SHS}^\top) &\leq \mu_j \left(\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top + \frac{\sum_{i>k} \lambda_i}{M} \cdot \mathbf{I}_M \right) + \left\| \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k} \lambda_i}{M} \cdot \mathbf{I}_M \right\|_2 \\ &= \mu_j(\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top) + \frac{\sum_{i>k} \lambda_i}{M} \cdot \mathbf{I}_M + \left\| \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k} \lambda_i}{M} \cdot \mathbf{I}_M \right\|_2 \\ &\leq \mu_j(\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top) + \frac{\sum_{i>k} \lambda_i}{M} \cdot \mathbf{I}_M + c_1 \cdot \left(\lambda_{k+1} + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M}} \right), \end{aligned}$$

where the last inequality is by Lemma B.25. For $j \leq k$, using Lemma B.26, we have

$$\mu_j(\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top) \leq \lambda_j + c_2 \cdot \sqrt{\frac{k}{M}} \cdot \lambda_j.$$

For $k < j \leq M$, we have

$$\mu_j(\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top) = 0 \leq \lambda_j + c_2 \cdot \sqrt{\frac{k}{M}} \cdot \lambda_j.$$

Putting these together, we have the following for every $j = 1, \dots, M$:

$$\begin{aligned}\mu_j(\mathbf{SHS}^\top) &\leq \mu_j(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top) + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M + c_1 \cdot \left(\lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}} \right) \\ &\leq \lambda_j + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M + c \cdot \left(\sqrt{\frac{k}{M}} \cdot \lambda_j + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}} \right).\end{aligned}$$

Similarly, we can show the lower bound. By the decomposition, we have

$$\begin{aligned}\mu_j(\mathbf{SHS}^\top) &\geq \mu_j\left(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M\right) - \left\| \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M \right\| \\ &= \mu_j(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top) + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M - \left\| \mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M \right\| \\ &\geq \mu_j(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top) + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M - c_1 \cdot \left(\lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}} \right),\end{aligned}$$

where the last inequality is by Lemma B.25. For $j \leq k$, using Lemma B.26, we have

$$\mu_j(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top) \geq \lambda_j - c_2 \cdot \sqrt{\frac{k}{M}} \cdot \lambda_j.$$

For $k < j \leq M$, we have

$$\mu_j(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top) = 0 \geq \lambda_j - \lambda_{k+1} - c_2 \cdot \sqrt{\frac{k}{M}} \cdot \lambda_j,$$

where the last inequality is due to $\lambda_j \leq \lambda_k$ for $j \geq k$. Putting these together, we have

$$\begin{aligned}\mu_j(\mathbf{SHS}^\top) &\geq \mu_j(\mathbf{S}_{0:k}\mathbf{H}_{0:k}\mathbf{S}_{0:k}^\top) + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M - c_1 \cdot \left(\lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}} \right) \\ &\geq \lambda_j + \frac{\sum_{i>k}\lambda_i}{M} \cdot \mathbf{I}_M - c \cdot \left(\sqrt{\frac{k}{M}} \cdot \lambda_j + \lambda_{k+1} + \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}} \right).\end{aligned}$$

So far, we have proved the first claim. To show the second claim, we simply apply

$$c \cdot \sqrt{\frac{k}{M}} \leq \frac{1}{2} \quad \text{for } k \leq M/c^2,$$

and

$$\begin{aligned}c \cdot \sqrt{\frac{\sum_{i>k}\lambda_i^2}{M}} &\leq c \cdot \sqrt{\frac{\sum_{i>k}\lambda_i}{M} \cdot \lambda_{k+1}} \\ &\leq \frac{1}{2} \cdot \frac{\sum_{i>k}\lambda_i}{M} + 2c^2 \cdot \lambda_{k+1},\end{aligned}$$

in the first claim. □

Lemma B.25 (Tail concentration, Lemma 26 in [Bartlett et al. \(2020\)](#)). *For any $k \geq 0$, with probability at least $1 - e^{-\Omega(M)}$, we have*

$$\left\| \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k} \lambda_i}{M} \cdot \mathbf{I}_M \right\|_2 \lesssim \lambda_{k+1} + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M}}.$$

Moreover, the minimum eigenvalue of $\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top$ satisfies

$$\mu_{\min}(\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top) \gtrsim \lambda_{k+2M}$$

with probability at least $1 - e^{-\Omega(M)}$.

Proof of Lemma B.25. The first part of Lemma B.25 is a version of Lemma 26 in [Bartlett et al. \(2020\)](#) (see their proof). We provide proof here for completeness.

We write $\mathbf{S} \in \mathbb{R}^{M \times p}$ as

$$\mathbf{S} = (\mathbf{s}_1 \ \dots \ \mathbf{s}_p), \quad \mathbf{s}_i \sim \mathcal{N}\left(0, \frac{1}{M} \cdot \mathbf{I}_M\right), \quad i \geq 1.$$

Since Gaussian distribution is rotational invariant, without loss of generality, we may assume

$$\mathbf{H} = \text{diag}\{\lambda_1, \dots, \lambda_p\}.$$

Then we have

$$\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top = \sum_{i>k} \lambda_i \mathbf{s}_i \mathbf{s}_i^\top.$$

Fixing a unit vector $\mathbf{v} \in \mathbb{R}^M$, then

$$\mathbf{v}^\top \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top \mathbf{v} = \sum_{i>k} \lambda_i (\mathbf{s}_i^\top \mathbf{v})^2,$$

where each $\mathbf{s}_i^\top \mathbf{v}$ is $(1/M)$ -subGaussian. By Bernstein's inequality, we have, with probability $\geq 1 - \delta$,

$$\left| \sum_{i>k} \lambda_i (\mathbf{s}_i^\top \mathbf{v})^2 - \frac{\sum_{i>k} \lambda_i}{M} \right| \lesssim \frac{1}{M} \cdot \left(\lambda_{k+1} \cdot \log \frac{1}{\delta} + \sqrt{\sum_{i>k} \lambda_i^2 \cdot \log \frac{1}{\delta}} \right).$$

By a union bound and net argument on $\mathcal{S}^{M-1}$, we have, with probability $\geq 1 - \delta$, for every unit vector $\mathbf{v} \in \mathbb{R}^M$,

$$\left| \sum_{i>k} \lambda_i (\mathbf{s}_i^\top \mathbf{v})^2 - \frac{\sum_{i>k} \lambda_i}{M} \right| \lesssim \frac{1}{M} \cdot \left(\lambda_{k+1} \cdot \left(M + \log \frac{1}{\delta} \right) + \sqrt{\sum_{i>k} \lambda_i^2 \cdot \left(M + \log \frac{1}{\delta} \right)} \right).$$

So with probability at least $1 - e^{-\Omega(M)}$, we have

$$\begin{aligned} \left\| \mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top - \frac{\sum_{i>k} \lambda_i}{M} \cdot \mathbf{I}_M \right\|_2 &\lesssim \frac{1}{M} \cdot \left(\lambda_{k+1} \cdot M + \sqrt{\sum_{i>k} \lambda_i^2 \cdot M} \right) \\ &\approx \lambda_{k+1} + \sqrt{\frac{\sum_{i>k} \lambda_i^2}{M}}, \end{aligned}$$

which completes the proof of the first part of Lemma B.25.

To prove the second part of Lemma B.25, it suffices to note that

$$\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top \succeq \sum_{i=k+1}^{2M+k} \lambda_i \mathbf{s}_i \mathbf{s}_i^\top \succeq \lambda_{2M+k} \cdot \sum_{i=k+1}^{2M+k} \mathbf{s}_i \mathbf{s}_i^\top \succeq c \lambda_{2M+k} \cdot \mathbf{I}_M$$

for some constant $c > 1$ with probability at least $1 - e^{-\Omega(M)}$, where the last line follows from concentration properties of Gaussian covariance matrices (see e.g., Theorem 6.1 [Wainwright 2019a](#)). $\square$

Lemma B.26 (Head concentration). *With probability at least $1 - e^{-\Omega(M)} - e^{-\Omega(k)}$, we have*

$$\text{for every } j \leq k, \quad |\mu_j(\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top) - \lambda_j| \lesssim \sqrt{\frac{k}{M}} \cdot \lambda_j.$$

Proof of Lemma B.26. Note that the spectrum of $\mathbf{S}_{0:k} \mathbf{H}_{0:k} \mathbf{S}_{0:k}^\top$ is identical to the spectrum of $\mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{S}_{0:k} \mathbf{H}_{0:k}^{1/2}$. We will bound the latter. We start with bounding the spectrum of $\mathbf{S}_{0:k} \mathbf{S}_{0:k}^\top$. To this end, we write $\mathbf{S}_{0:k}^\top \in \mathbb{R}^{k \times M}$ as

$$\mathbf{S}_{0:k}^\top = (\mathbf{s}_1 \quad \dots \quad \mathbf{s}_M), \quad \mathbf{s}_i \sim \mathcal{N}\left(0, \frac{1}{M} \cdot \mathbf{I}_k\right), \quad i = 1, \dots, M.$$

Then repeating the argument in Lemma B.25, we have, with probability $\geq 1 - \delta$, for every unit vector $\mathbf{v} \in \mathbb{R}^k$,

$$\begin{aligned} |\mathbf{v}^\top \mathbf{S}_{0:k}^\top \mathbf{S}_{0:k} \mathbf{v} - 1| &= \left| \sum_{i=1}^M (\mathbf{s}_i^\top \mathbf{v})^2 - 1 \right| \\ &\lesssim \frac{1}{M} \cdot \left(1 \cdot \left(k + \log \frac{1}{\delta} \right) + \sqrt{M \cdot \left(k + \log \frac{1}{\delta} \right)} \right) \\ &\lesssim \sqrt{\frac{k + \log(1/\delta)}{M}}. \end{aligned}$$

So we have, with probability $\geq 1 - e^{-\Omega(k)}$,

$$\|\mathbf{S}_{0:k}^\top \mathbf{S}_{0:k} - \mathbf{I}_k\|_2 \lesssim \sqrt{\frac{k}{M}}.$$

This implies that

$$\begin{aligned}\mu_j(\mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{S}_{0:k} \mathbf{H}_{0:k}^{1/2}) &\leq \mu_j(\mathbf{H}_{0:k}^{1/2} \mathbf{H}_{0:k}^{1/2}) + c_1 \cdot \sqrt{\frac{k}{M}} \cdot \mu_j(\mathbf{H}_{0:k}^{1/2} \mathbf{H}_{0:k}^{1/2}) \\ &= \lambda_j + c_1 \cdot \sqrt{\frac{k}{M}} \cdot \lambda_j,\end{aligned}$$

and that

$$\begin{aligned}\mu_j(\mathbf{H}_{0:k}^{1/2} \mathbf{S}_{0:k}^\top \mathbf{S}_{0:k} \mathbf{H}_{0:k}^{1/2}) &\geq \mu_j(\mathbf{H}_{0:k}^{1/2} \mathbf{H}_{0:k}^{1/2}) - c_1 \cdot \sqrt{\frac{k}{M}} \cdot \mu_j(\mathbf{H}_{0:k}^{1/2} \mathbf{H}_{0:k}^{1/2}) \\ &= \lambda_j - c_1 \cdot \sqrt{\frac{k}{M}} \cdot \lambda_j.\end{aligned}$$

We have completed the proof. $\square$

B.7.2 Concentration results under power-law spectrum

Lemma B.27 (Eigenvalues of $\mathbf{SHS}^\top$ under power-law spectrum). *Suppose Assumption 3.2 hold. There exist a -dependent constants $c_2 > c_1 > 0$ such that*

$$c_1 j^{-a} \leq \mu_j(\mathbf{SHS}^\top) \leq c_2 j^{-a}$$

with probability at least $1 - e^{-\Omega(M)}$.

Proof of Lemma B.27. Let $(\tilde{\lambda}_i)_{i=1}^M$ denote the eigenvalues of $\mathbf{SHS}^\top$ in non-increasing order. Using Lemma B.24 with $k = M/c$ for some sufficiently large constant c and noting that $\sum_{i>k} i^{-a} \asymp k^{1-a}$, we have

$$\frac{1}{2} \cdot (j^{-a} + \tilde{c}_1 M^{-a}) - \tilde{c}_2 \cdot M^{-a} \leq \tilde{\lambda}_j \leq \frac{3}{2} \cdot (j^{-a} + \tilde{c}_1 M^{-a}) + \tilde{c}_2 \cdot M^{-a}$$

for every $j \in [M]$ for some constants $\tilde{c}_i, i \in [2]$ with probability at least $1 - e^{-\Omega(M)}$. Therefore, for all $j \leq M/\tilde{c}$ for some sufficiently large constant $\tilde{c} > 1$, we have

$$\tilde{\lambda}_j \in [\tilde{c}_3 j^{-a}, \tilde{c}_4 j^{-a}]$$

with probability at least $1 - e^{-\Omega(M)}$ for some constants $\tilde{c}_3, \tilde{c}_4 > 0$. For $j \in [M/\tilde{c}, M]$, by monotonicity of the eigenvalues, we have

$$\tilde{\lambda}_j \leq \tilde{\lambda}_{\lfloor M/\tilde{c} \rfloor} \leq \tilde{c}_4 \left(\left\lfloor \frac{M}{\tilde{c}} \right\rfloor \right)^{-a} \leq \tilde{c}_5 M^{-a} \leq \tilde{c}_5 j^{-a}$$

for some sufficiently large constant $\tilde{c}_5 > \tilde{c}_4$ with probability at least $1 - e^{-\Omega(M)}$. Moreover, using Lemma B.25 with $k = 0$, we obtain

$$\tilde{\lambda}_j \geq \tilde{\lambda}_M \geq \mu_{\min}(\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top) \geq \tilde{c}_6 \tilde{\lambda}_{2M} \geq \tilde{c}_7 (M/\tilde{c})^{-a} \geq \tilde{c}_8 j^{-a}$$

with probability at least $1 - e^{-\Omega(M)}$ for some constants $\tilde{c}_6, \tilde{c}_7, \tilde{c}_8 > 0$. Combining the bounds for $j \leq M/\tilde{c}$ and $j \in [M/\tilde{c}, M]$ completes the proof. $\square$

Lemma B.28 (Ratio of eigenvalues of $\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$ under power-law spectrum). *Suppose Assumption 3.2 hold. There exists some a -dependent constant $c > 0$ such that for any $k \geq 1$, the ratio between the $M/2$ -th and M -th eigenvalues*

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)} \leq c$$

with probability at least $1 - e^{-\Omega(M)}$.

Proof of Lemma B.28. We prove the lemma under two scenarios where k is relatively small (or large) compared with M .

Let $c > 0$ be some sufficiently large constant. Applying Lemma B.24 with $\mathbf{H}_{k:\infty}$ replacing $\mathbf{H}$, for $k_0 = M/c$, we have

$$\begin{aligned} \mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) &\leq \frac{3}{2} \cdot \left(\lambda_{M/2+k} + \frac{\sum_{i>k_0} \lambda_{i+k}}{M} \right) + c_1 \cdot \lambda_{k_0+1+k}, \\ &\lesssim \left(\frac{M}{2} + k \right)^{-a} + \frac{(k_0 + k)^{1-a}}{M} + (k_0 + 1 + k)^{-a} \\ &\lesssim (k \vee M)^{-a} + (k \vee M)^{-a} \left(1 \vee \frac{k}{M} \right) + (k \vee M)^{-a} \\ &\lesssim (k \vee M)^{-a} \left(1 \vee \frac{k}{M} \right) \end{aligned} \tag{B.29}$$

with probability at least $1 - e^{-\Omega(M)}$ for some constant $c_1 > 0$.

Case 1: $k \lesssim M$ From Lemma B.25, we have

$$\mu_{\min}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) \gtrsim \lambda_{k+2M} \gtrsim (k \vee M)^{-a}.$$

with probability at least $1 - e^{-\Omega(M)}$. Therefore

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)} \lesssim 1$$

with probability at least $1 - e^{-\Omega(M)}$ when $k/M \lesssim 1$.

Case 2: $k \gtrsim M$ On the other hand, when k is relatively large, using Lemma B.24 with $\mathbf{H}_{k:\infty}$ replacing $\mathbf{H}$ again, we obtain

$$\begin{aligned} \mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) &\geq \frac{1}{2} \cdot \left(\lambda_{M+k} + \frac{\sum_{i>k_0} \lambda_{i+k}}{M} \right) - c_1 \cdot \lambda_{k_0+1+k}, \\ &\geq c_2 \left[(M+k)^{-a} + \frac{(k_0+k)^{1-a}}{M} \right] - c_3 \cdot (k_0+1+k)^{-a} \end{aligned}$$

with probability at least $1 - e^{-\Omega(M)} - e^{-\Omega(k_0)}$, where $c_1, c_2, c_3 > 0$ are some universal constants. Choosing $k_0 = M/c^2$ for some sufficiently large constant $c > 0$, we further obtain

$$\begin{aligned} \mu_M(\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top) &\geq c_4(M+k)^{-a} \left[1 + \frac{k}{M}\right] - c_5(M+k)^{-a} \\ &\geq c_6(M \vee k)^{-a} \left[1 \vee \frac{k}{M}\right] - c_7(M \vee k)^{-a} \end{aligned} \quad (\text{B.30})$$

with probability at least $1 - e^{-\Omega(M)}$, where $(c_i)_{i=4}^7$ are a -dependent constants. Since

$$c_6(M \vee k)^{-a} \left[1 \vee \frac{k}{M}\right] - c_7(M \vee k)^{-a} \geq \frac{c_6}{2}(k \vee M)^{-a} \left(1 \vee \frac{k}{M}\right)$$

when k is large, i.e., $k/M > \tilde{c}$ for some sufficiently large a -dependent constant $\tilde{c} > 0$ that may depend on $(c_i)_{i=1}^7$, we have from Eq. (B.29) and (B.30) that

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top)} \lesssim \frac{(k \vee M)^{-a} (1 \vee \frac{k}{M})}{(k \vee M)^{-a} (1 \vee \frac{k}{M})} \lesssim 1$$

with probability at least $1 - e^{-\Omega(M)}$. □

B.7.3 Concentration results under logarithmic power-law spectrum

Lemma B.29 (Theorem 6 in Bartlett et al. (2020)). *Suppose Assumption 3.4 hold. Then there exist some a -dependent constants $c, \tilde{c} > 0$ such that, with probability at least $1 - e^{-\Omega(M)}$*

$$\mu_j(\mathbf{SHS}^\top) \in \begin{cases} [c \cdot j^{-1} \log^{-a}(j+1), \tilde{c} \cdot j^{-1} \log^{-a}(j+1)] & j \leq k^*, \\ [c \cdot M^{-1} \log^{1-a}(M), \tilde{c} \cdot M^{-1} \log^{1-a}(M)] & k^* < j \leq M, \end{cases}$$

where $k^* \asymp M/\log(M)$. Also, there exists some a -dependent constants $c_1, c_2 > 0$ such that

$$\frac{c_1}{j \log^{2a}(j+1)} \leq \mu_j(\mathbf{SH}^2 \mathbf{S}^\top) \leq \frac{c_2}{j \log^{2a}(j+1)}$$

with probability at least $1 - e^{-\Omega(M)}$.

Proof of Lemma B.29. The proof is adapted from the proof of Theorem 6 in Bartlett et al. (2020). We include it here for completeness.

First part of Lemma B.29. In Lemma B.24, for some constant $c > 1$, choose

$$k^* := \min \left\{ k \geq 0 : \sum_{i>k} \lambda_i \geq c \cdot M \cdot \lambda_{k+1} \right\}.$$

Then with probability $\geq 1 - e^{-\Omega(M)}$, we have:

$$\text{for every } 1 \leq j \leq M, \quad \frac{1}{c_1} \cdot \left(\lambda_j + \frac{\sum_{i>k^*} \lambda_i}{M} \right) \leq \tilde{\lambda}_j \leq c_1 \cdot \left(\lambda_j + \frac{\sum_{i>k^*} \lambda_i}{M} \right),$$

where $c_1 > 1$ is a constant.

When $\lambda_j \approx j^{-1} \log^{-a}(j+1)$, we have

$$k^* \approx M/\log(M),$$

and

$$\sum_{i>k^*} \lambda_i \approx \log^{1-a}(k^*) \approx \log^{1-a}(M).$$

Therefore, we have

$$\begin{aligned} \tilde{\lambda}_j &\approx \lambda_j + \frac{\sum_{i>k^*} \lambda_i}{M} \\ &\approx \begin{cases} j^{-1} \log^{-a}(j+1) & j \leq k^*, \\ M^{-1} \log^{1-a}(M) & k^* < j \leq M, \end{cases} \end{aligned}$$

where $k^* \approx M/\log(M)$.

Second part of Lemma B.29. Let $\bar{\lambda}_i$ denote the i -th eigenvalue of $\mathbf{S}\mathbf{H}^2\mathbf{S}^\top$ for $i \in [M]$. Using Lemma B.24 with $k = M/c$ for some sufficiently large constant c_0 and noting that $\sum_{i>k} \lambda_i^2 \approx \sum_{i>k} i^{-2} \log^{-2a}(i+1) \lesssim k^{-1} \log^{-2a} k$, we have

$$\begin{aligned} &\frac{1}{2} \cdot j^{-2} \log^{-2a}(j+1) - \tilde{c}_2 \cdot M^{-2} \log^{-2a} M \\ &\leq \bar{\lambda}_j \leq \frac{3}{2} \cdot (j^{-2} \log^{-2a}(j+1) + \tilde{c}_1 M^{-2} \log^{-2a} M) + \tilde{c}_2 \cdot M^{-2} \log^{-2a} M \end{aligned}$$

for every $j \in [M]$ for some constants $\tilde{c}_i, i \in [2]$ with probability at least $1 - e^{-\Omega(M)}$. Therefore, for all $j \leq M/\tilde{c}$ for some sufficiently large constant $\tilde{c} > 1$, we have

$$\bar{\lambda}_j \in [\tilde{c}_3 \cdot j^{-2} \log^{-2a}(j+1), \tilde{c}_4 \cdot j^{-2} \log^{-2a}(j+1)]$$

with probability at least $1 - e^{-\Omega(M)}$ for some constants $\tilde{c}_3, \tilde{c}_4 > 0$. For $j \in [M/\tilde{c}, M]$, by monotonicity of the eigenvalues, we have

$$\bar{\lambda}_j \leq \bar{\lambda}_{\lfloor M/\tilde{c} \rfloor} \leq \tilde{c}_4 \left(\left\lfloor \frac{M}{\tilde{c}} \right\rfloor \right)^{-2} \log^{-2a} \left(\left\lfloor \frac{M}{\tilde{c}} \right\rfloor \right) \leq \tilde{c}_5 M^{-2} \log^{-2a} M \leq \tilde{c}_6 \cdot j^{-2} \log^{-2a}(j+1)$$

for some constants $c_5, c_6 > 0$ with probability at least $1 - e^{-\Omega(M)}$. Moreover, using Lemma B.25 with $k = 0$, we obtain

$$\bar{\lambda}_j \geq \bar{\lambda}_M \geq \mu_{\min}(\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty}^2 \mathbf{S}_{k:\infty}^\top) \geq \tilde{c}_7 \bar{\lambda}_{2M} \geq \tilde{c}_8 \cdot j^{-2} \log^{-2a}(j+1)$$

with probability at least $1 - e^{-\Omega(M)}$ for some constants $\tilde{c}_7, \tilde{c}_8 > 0$ when $j \in [M/\tilde{c}, M]$. Combining the bounds for $j \leq M/\tilde{c}$ and $j \in [M/\tilde{c}, M]$ completes the proof. $\square$

Lemma B.30 (Ratio of eigenvalues of $\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top$ under logarithmic power-law spectrum). *Suppose Assumption 3.4 hold. There exists some a -dependent constant $c > 0$ such that for any $k \geq 1$, the ratio between the $M/2$ -th and M -th eigenvalues*

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)} \leq c$$

with probability at least $1 - e^{-\Omega(M)}$.

Proof of Lemma B.30. Similar to the proof of Lemma B.28, we prove the lemma under two scenarios where k is relatively small (or large) compared with M .

Let $c > 0$ be some sufficiently large constant. Applying Lemma B.24 with $\mathbf{H}_{k:\infty}$ replacing $\mathbf{H}$, for $k_0 = M/c$, we have

$$\begin{aligned} \mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) &\leq \frac{3}{2} \cdot \left(\lambda_{M/2+k} + \frac{\sum_{i>k_0} \lambda_{i+k}}{M} \right) + c_1 \cdot \lambda_{k_0+1+k}, \\ &\lesssim \left(\frac{M}{2} + k \right)^{-1} \log^{-a} \left(\frac{M}{2} + k \right) + \frac{\log^{1-a}(k_0 + k)}{M} + \frac{\log^{-a}(k_0 + 1 + k)}{k_0 + 1 + k} \\ &\lesssim \frac{\log^{-a}(M + k)}{(M + k)} + \frac{\log^{1-a}(M + k)}{M} \lesssim \frac{\log^{1-a}(M + k)}{M} \end{aligned} \quad (\text{B.31})$$

with probability at least $1 - e^{-\Omega(M)}$ for some constant $c_1 > 0$.

Case 1: $k \lesssim M$. Applying Lemma B.24 with $\mathbf{H}_{k:\infty}$ replacing $\mathbf{H}$, for $k_0 = M/c$, we have

$$\begin{aligned} \mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top) &\gtrsim \frac{1}{2} \cdot \left(\lambda_{M+k} + \frac{\sum_{i>k_0} \lambda_{i+k}}{M} \right) - c_1 \cdot \lambda_{k_0+1+k}, \\ &\gtrsim (M + k)^{-1} \log^{-a} (M + k) + \frac{\log^{1-a}(k_0 + k)}{M} - c \frac{\log^{-a}(k_0 + 1 + k)}{k_0 + 1 + k} \\ &\gtrsim \frac{\log^{-a}(M + k)}{(M + k)} + \frac{\log^{1-a}(M + k)}{M} - \tilde{c} \frac{\log^{-a}(M)}{M} \\ &\gtrsim \frac{\log^{1-a} M}{M} \end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$. Therefore,

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty}\mathbf{H}_{k:\infty}\mathbf{S}_{k:\infty}^\top)} \lesssim \left[\frac{\log^{1-a}(M + k)}{M} \right] / \left[\frac{\log^{1-a} M}{M} \right] \lesssim 1$$

with probability at least $1 - e^{-\Omega(M)}$ when $k/M \lesssim 1$.

Case 2: $k \gtrsim M$. On the other hand, when k is relatively large, using Lemma B.24 with $\mathbf{H}_{k:\infty}$ replacing $\mathbf{H}$ and $k_0 = M/c$ again, we obtain

$$\begin{aligned}
\mu_M(\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top) &\geq \frac{1}{2} \cdot \left(\lambda_{M+k} + \frac{\sum_{i>k_0} \lambda_{i+k}}{M} \right) - c_1 \cdot \lambda_{k_0+1+k}, \\
&\gtrsim (M+k)^{-1} \log^{-a}(M+k) + \frac{\log^{1-a}(k_0+k)}{M} - c_2 \frac{\log^{-a}(k_0+1+k)}{k_0+1+k} \\
&\gtrsim k^{-1} \log^{-a}(k) + \frac{\log^{1-a}(M+k)}{M} - c_3 \frac{\log^{-a}(M+k)}{M+k} \\
&\gtrsim \frac{\log^{1-a}(M+k)}{M}
\end{aligned}$$

with probability at least $1 - e^{-\Omega(M)}$, where $c_1, c_2, c_3 > 0$ are some a -dependent constants. Therefore,

$$\frac{\mu_{M/2}(\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top)}{\mu_M(\mathbf{S}_{k:\infty} \mathbf{H}_{k:\infty} \mathbf{S}_{k:\infty}^\top)} \lesssim \left[\frac{\log^{1-a}(M+k)}{M} \right] / \left[\frac{\log^{1-a}(M+k)}{M} \right] \lesssim 1$$

with probability at least $1 - e^{-\Omega(M)}$ when $k/M \gtrsim 1$. □

Appendix C

Appendix for Chapter 4

C.1 Proofs

C.1.1 Proof of Proposition 4.1

Adopt the shorthand $R_i = \log \frac{\pi_\theta(y_i | x_i)}{\pi_{\text{ref}}(y_i | x_i)}$. For any $i \in [n_f]$, we have

$$\begin{aligned}
 \lim_{\beta \rightarrow 0} -\frac{2}{\beta} \cdot \log \sigma \left(-\beta \cdot \log \frac{\pi_\theta(y_i | x_i)}{\pi_{\text{ref}}(y_i | x_i)} \right) - \frac{2}{\beta} \log 2 &= \lim_{\beta \rightarrow 0} -\frac{2}{\beta} \cdot \log \sigma(-\beta R_i) - \frac{2}{\beta} \log 2 \\
 &= \lim_{\beta \rightarrow 0} \frac{2}{\beta} \cdot \log \left(\frac{1 + \exp(\beta R_i)}{2} \right) \\
 &= \lim_{\beta \rightarrow 0} \frac{2}{\beta} \cdot \log \left(1 + \frac{\exp(\beta R_i) - 1}{2} \right) \\
 &= \lim_{\beta \rightarrow 0} \frac{2}{\beta} \cdot \frac{\exp(\beta R_i) - 1}{2} \\
 &= R_i.
 \end{aligned}$$

Averaging over all $i \in [n_f]$ and noting that $\sum_{i=1}^{n_f} R_i / n_f = \mathcal{L}_{\text{GA}}(\theta) - \mathbb{E}_{\mathcal{D}_{\text{FG}}}[\log \pi_{\text{ref}}(y_i | x_i)]$ yields the first part of Proposition 4.1.

For the second part of Proposition 4.1, by definition

$$\begin{aligned}
 \nabla_\theta \mathcal{L}_{\text{NPO}, \beta}(\theta) &= \frac{1}{n_f} \sum_{i=1}^{n_f} -\frac{2}{\beta} \nabla_\theta \log \sigma(-\beta R_i) \\
 &= \frac{1}{n_f} \sum_{i=1}^{n_f} \frac{2}{\beta} \nabla_\theta \log(1 + \exp(\beta R_i)) \\
 &= \frac{1}{n_f} \sum_{i=1}^{n_f} \frac{2}{\beta} \cdot \frac{\beta \exp(\beta R_i)}{1 + \exp(\beta R_i)} \cdot \nabla_\theta R_i
 \end{aligned}$$

$$= \frac{1}{n_f} \sum_{i=1}^{n_f} \frac{2 \exp(\beta R_i)}{1 + \exp(\beta R_i)} \cdot \nabla_{\theta} R_i.$$

Therefore, it follows immediately that

$$\begin{aligned} \lim_{\beta \rightarrow 0} \nabla_{\theta} \mathcal{L}_{\text{NPO}, \beta}(\theta) &= \lim_{\beta \rightarrow 0} \frac{1}{n_f} \sum_{i=1}^{n_f} \frac{2 \exp(\beta R_i)}{1 + \exp(\beta R_i)} \cdot \nabla_{\theta} R_i = \frac{1}{n_f} \sum_{i=1}^{n_f} \nabla_{\theta} R_i \\ &= \frac{1}{n_f} \sum_{i=1}^{n_f} \nabla_{\theta} \log \pi_{\theta}(y_i | x_i) = \nabla_{\theta} \mathcal{L}_{\text{GA}}(\theta). \end{aligned}$$

This completes the proof of Proposition 4.1.

C.1.2 Proof of Theorem 4.2

Let $\eta_0 := \eta/n_f$ denote the normalized learning rate. For $i, j \in [n_f]$, define $\gamma_{i,j} := \langle x_i, x_j \rangle$ and $c_{\text{init},i} := \langle \theta_{\text{init}}, x_i \rangle$. Throughout the proof, we also use $C_k (k = 0, 1, 2, \dots)$ to denote constants that may depend on $(B_{\theta}, b_x, B_x, \beta)$ but not on (t, n_f, d) . By the definition of the logistic model and some algebra, we have

$$\nabla_{\theta} \log \pi_{\theta}(y_i | x_i) = x_i(2y_i - 1)(1 - \pi_{\theta}(y_i | x_i)).$$

Therefore, the gradient of $\mathcal{L}_{\text{GA}}$ and $\mathcal{L}_{\text{NPO}}$ both lie in the span of $x_1, \dots, x_{n_f}$. Consequently, $\theta_{\text{GA}}^{(t)} - \theta_{\text{init}}$ and $\theta_{\text{NPO}}^{(t)} - \theta_{\text{init}}$ can be rewritten as follows:

$$\theta_{\text{GA}}^{(t)} - \theta_{\text{init}} = \sum_{i=1}^{n_f} \alpha_{\text{GA},i}^{(t)} \cdot x_i, \quad \theta_{\text{NPO}}^{(t)} - \theta_{\text{init}} = \sum_{i=1}^{n_f} \alpha_{\text{NPO},i}^{(t)} \cdot x_i,$$

where

$$\begin{aligned} \alpha_{\text{GA},i}^{(t+1)} &:= \alpha_{\text{GA},i}^{(t)} - \eta_0(2y_i - 1)(1 - \pi_{\theta}(y_i | x_i)), \\ \alpha_{\text{NPO},i}^{(t+1)} &:= \alpha_{\text{NPO},i}^{(t)} - \eta_0(2y_i - 1)(1 - \pi_{\theta}(y_i | x_i)) \cdot W_i^{(t)}, \end{aligned}$$

and

$$W_i^{(t)} = \pi_{\theta_{\text{NPO}}^{(t)}}^{\beta}(y_i | x_i) / [\pi_{\theta_{\text{NPO}}^{(t)}}^{\beta}(y_i | x_i) + \pi_{\theta_{\text{init}}}^{\beta}(y_i | x_i)].$$

We also define $\alpha_{\star,i}^{(0)} = 0$ and adopt the shorthand notations $\alpha_{\star}^{(t)} = (\alpha_{1,\star}^{(t)}, \dots, \alpha_{n_f,\star}^{(t)})$ for $\star \in \{\text{GA}, \text{NPO}\}$ and $\gamma_i = (\gamma_{i,1}, \dots, \gamma_{i,n_f})$.

For $\star \in \{\text{GA}, \text{NPO}\}$, we have

$$\pi_{\theta_{\star}^{(t)}}(y_i | x_i) = \frac{1}{1 + \exp\left((1 - 2y_i)\langle x_i, \theta_{\star}^{(t)} \rangle\right)}$$

$$\begin{aligned}
&= \frac{1}{1 + \exp \left((1 - 2y_i)c_{\text{init},i} + (1 - 2y_i)\langle x_i, \sum_{j=1}^{n_f} \alpha_{\star}^{(t)} x_j \rangle \right)} \\
&= \frac{1}{1 + \exp \left((1 - 2y_i)c_{\text{init},i} + (1 - 2y_i)\langle \alpha_{\star}^{(t)}, \gamma_i \rangle \right)} =: \text{pred}_i \left(\langle \alpha_{\star}^{(t)}, \gamma_i \rangle \right),
\end{aligned}$$

where we denote the dependence on $(c_{\text{init},i}, y_i)$ implicitly using pred_i for notational simplicity.

Therefore, letting $b_{\star,i}^{(t)} := \langle \alpha_{\star}^{(t)}, \gamma_i \rangle$ for $\star \in \{\text{GA}, \text{NPO}\}$ and combining all $i \in [n_f]$, we obtain

$$b_{\text{GA}}^{(t+1)} := b_{\text{GA}}^{(t)} - \eta_0 \cdot \Gamma \Delta_{\text{GA}}^{(t)}, \quad b_{\text{NPO}}^{(t+1)} := b_{\text{NPO}}^{(t)} - \eta_0 \cdot \Gamma \text{diag} \left\{ \Delta_{\text{NPO}}^{(t)} \right\} \mathbf{W}^{(t)},$$

where

$$\begin{aligned}
\Gamma &:= (\gamma_{i,j})_{1 \leq i,j \leq n_f}, \\
\Delta_{\star}^{(t)} &:= \left(\Delta_{\star,1}^{(t)}, \dots, \Delta_{\star,n_f}^{(t)} \right)^{\top}, \quad \Delta_{\star,i}^{(t)} := (2y_i - 1)(1 - \text{pred}_i(b_{\star,i}^{(t)})), \text{ for } i \in [n_f], \star \in \{\text{GA}, \text{NPO}\}, \\
\mathbf{W}^{(t)} &:= (\mathbf{W}_1^{(t)}, \dots, \mathbf{W}_{n_f}^{(t)})^{\top}, \quad \mathbf{W}_i^{(t)} = \mathbf{W}_i(b_{\text{NPO}}^{(t)}) := \pi_{\theta_{\text{NPO}}^{(t)}}^{\beta}(y_i | x_i) / [\pi_{\theta_{\text{NPO}}^{(t)}}^{\beta}(y_i | x_i) + \pi_{\theta_{\text{init}}}^{\beta}(y_i | x_i)].
\end{aligned}$$

Again, we hide here the dependence on $(\beta, c_{\text{init},i}, y_i)$ in $\mathbf{W}_i$ for simplicity. We now claim the following results of which the proofs are deferred to Section C.1.3 and C.1.4.

Lemma C.1 (GA converges to infinity linearly). *Under the assumptions in Theorem 4.2 and the notations in the proof of Theorem 4.2, there exist some (B_{θ}, b_x, B_x) -dependent constants $C_0, C_1, C_2 > 0$ such that the GA iterations $\{b_{\text{GA}}^{(t)}\}_{t=1}^{\infty}$ satisfy*

$$\begin{aligned}
C_1 \eta_0 t &\leq b_{\text{GA},i}^{(t)} \leq C_2 \eta_0 t, \text{ when } y_i = 0, \\
-C_2 \eta_0 t &\leq b_{\text{GA},i}^{(t)} \leq -C_1 \eta_0 t, \text{ when } y_i = 1.
\end{aligned}$$

for all $i \in [n_f]$ and $t \geq 1$ when $\max_{i \neq j} |\gamma_{i,j}| \leq C_0/n_f$.

Lemma C.2 (NPO converges to infinity exponentially slow). *Under the assumptions in Theorem 4.2 and the notations in the proof of Theorem 4.2, there exist some $(B_{\theta}, b_x, B_x, \beta)$ -dependent constants $C_0, C_b > 0, C_a \in (0, 1)$ such that the NPO iterations $\{b_{\text{NPO}}^{(t)}\}_{t=1}^{\infty}$ satisfy*

$$\begin{aligned}
b_{\text{NPO},i}^{(t)} &\in \left[-\frac{1}{\beta} \log(C_b \eta_0 t + 1), -\frac{1}{\beta} \log(C_a \eta_0 t + 1) \right] \text{ when } y_i = 1, \\
b_{\text{NPO},i}^{(t)} &\in \left[\frac{1}{\beta} \log(C_a \eta_0 t + 1), \frac{1}{\beta} \log(C_b \eta_0 t + 1) \right] \text{ when } y_i = 0.
\end{aligned}$$

for all $i \in [n_f]$ and $t \geq 1$ when $\max_{i \neq j} |\gamma_{i,j}| \leq C_0/n_f$.

Combining Lemma C.1 and C.2 and noting

$$\begin{aligned}\|\theta_\star^{(t)} - \theta_{\text{init}}\|_{X^\top X} &= \sqrt{(\theta_\star^{(t)} - \theta_{\text{init}})^\top X^\top X (\theta_\star^{(t)} - \theta_{\text{init}})} = \sqrt{\alpha_\star^{(t)\top} X X^\top X X^\top \alpha_\star^{(t)}} \\ &= \sqrt{b_\star^{(t)\top} (X^\top X)^{-1} X X^\top X X^\top (X^\top X)^{-1} b_\star^{(t)}} = \|b_\star^{(t)}\|_2\end{aligned}$$

for $\star \in \{\text{GA}, \text{NPO}\}$ completes the proof.

C.1.3 Proof of Lemma C.1

We prove Lemma C.1 by induction.

Case 1: $t = 1$ When $t = 1$, since $|c_{\text{init}}| \leq \|x_i\|_2 \cdot \|\theta_{\text{init}}\|_2 \leq B_x B_\theta$, it follows from the definition of $\text{pred}_i(\cdot)$ that

$$C_3 \leq \Delta_{\text{GA},i}^{(0)} \leq 1 \text{ when } y_i = 1, \quad -1 \leq \Delta_{\text{GA},i}^{(0)} \leq -C_4 \text{ when } y_i = 0$$

for all $i \in [n_f]$ for some constants $C_3, C_4 \in (0, 1)$ depending only on (B_θ, b_x, B_x) . Note that there exists a constant $C_0 > 0$ depending on C_3, C_4, b_x such that

$$\left| \sum_{j \neq i} \gamma_{i,j} \Delta_{\text{GA},j}^{(0)} \right| \leq \frac{\gamma_{i,i}}{2} \left| \Delta_{\text{GA},i}^{(0)} \right|$$

for all $i \in [n_f]$ when $\max_{i \neq j} |\gamma_{i,j}| \leq C_0/n_f$. It follows that

$$\begin{aligned}-\eta_0 \cdot \gamma_i^\top \Delta_{\text{GA}}^{(0)} &\in \left[-\frac{3}{2} \gamma_{i,i} \eta_0 |\Delta_{\text{GA},i}^{(0)}|, -\frac{1}{2} \gamma_{i,i} \eta_0 |\Delta_{\text{GA},i}^{(0)}| \right] \in [-C_2 \eta_0, -C_1 \eta_0] \text{ when } y_i = 1, \\ -\eta_0 \cdot \gamma_i^\top \Delta_{\text{GA}}^{(0)} &\in \left[\frac{1}{2} \gamma_{i,i} \eta_0 |\Delta_{\text{GA},i}^{(0)}|, \frac{3}{2} \gamma_{i,i} \eta_0 |\Delta_{\text{GA},i}^{(0)}| \right] \in [C_1 \eta_0, C_2 \eta_0] \text{ when } y_i = 0\end{aligned}$$

for some (B_θ, b_x, B_x) -dependent constants $C_1, C_2 > 0$. Therefore,

$$\begin{aligned}b_{\text{GA},i}^{(1)} &= b_{\text{GA},i}^{(0)} - \eta_0 \cdot \gamma_i^\top \Delta_{\text{GA}}^{(0)} \in [-C_2 \eta_0, -C_1 \eta_0] \text{ when } y_i = 1, \\ b_{\text{GA},i}^{(1)} &= b_{\text{GA},i}^{(0)} - \eta_0 \cdot \gamma_i^\top \Delta_{\text{GA}}^{(0)} \in [C_1 \eta_0, C_2 \eta_0] \text{ when } y_i = 0.\end{aligned}$$

As a consequence,

$$\begin{aligned}b_{\text{GA},i}^{(1)} &\leq b_{\text{GA},i}^{(0)} = 0, \quad \Delta_{\text{GA},i}^{(1)} \geq \Delta_{\text{GA},i}^{(0)} \text{ when } y_i = 1, \\ b_{\text{GA},i}^{(1)} &\geq b_{\text{GA},i}^{(0)} = 0, \quad \Delta_{\text{GA},i}^{(1)} \leq \Delta_{\text{GA},i}^{(0)} \text{ when } y_i = 0.\end{aligned}$$

Case 2: $t = K + 1$ Now, suppose we have

$$\begin{aligned}b_{\text{GA},i}^{(t)} &\in [-C_2 \eta_0 t, -C_1 \eta_0 t] \text{ when } y_i = 1, \\ b_{\text{GA},i}^{(t)} &\in [C_1 \eta_0 t, C_2 \eta_0 t] \text{ when } y_i = 0\end{aligned}$$

for $t \in [K]$ and

$$\begin{aligned} b_{\text{GA},i}^{(K)} &\leq \dots \leq b_{\text{GA},i}^{(0)} = 0, \quad \Delta_{\text{GA},i}^{(K)} \geq \dots \geq \Delta_{\text{GA},i}^{(0)} \quad \text{when } y_i = 1, \\ b_{\text{GA},i}^{(K)} &\geq \dots \geq b_{\text{GA},i}^{(0)} = 0, \quad \Delta_{\text{GA},i}^{(K)} \leq \dots \leq \Delta_{\text{GA},i}^{(0)} \quad \text{when } y_i = 0. \end{aligned}$$

By the monotonicity of $\Delta_{\text{GA},i}^{(t)}$, we have

$$C_3 \leq \Delta_{\text{GA},i}^{(K)} \leq 1 \quad \text{when } y_i = 1, \quad -1 \leq \Delta_{\text{GA},i}^{(K)} \leq -C_4 \quad \text{when } y_i = 0$$

for all $i \in [n_f]$. Therefore, following similar arguments as in the $t = 1$ case, we have

$$\begin{aligned} -\eta_0 \cdot \gamma_i^\top \Delta_{\text{GA}}^{(K)} &\in [-C_2\eta_0, -C_1\eta_0] \quad \text{when } y_i = 1, \\ -\eta_0 \cdot \gamma_i^\top \Delta_{\text{GA}}^{(K)} &\in [C_1\eta_0, C_2\eta_0] \quad \text{when } y_i = 0. \end{aligned}$$

Then it follows from the induction assumption that

$$\begin{aligned} b_{\text{GA},i}^{(K+1)} &= b_{\text{GA},i}^{(K)} - \eta_0 \cdot \gamma_i^\top \Delta_{\text{GA}}^{(K)} \in [-C_2\eta_0(K+1), -C_1\eta_0(K+1)] \quad \text{when } y_i = 1, \\ b_{\text{GA},i}^{(K+1)} &= b_{\text{GA},i}^{(K)} - \eta_0 \cdot \gamma_i^\top \Delta_{\text{GA}}^{(K)} \in [C_1\eta_0(K+1), C_2\eta_0(K+1)] \quad \text{when } y_i = 0, \end{aligned}$$

and also

$$\begin{aligned} b_{\text{GA},i}^{(K+1)} &\leq b_{\text{GA},i}^{(K)}, \quad \Delta_{\text{GA},i}^{(K+1)} \geq \Delta_{\text{GA},i}^{(K)} \quad \text{when } y_i = 1, \\ b_{\text{GA},i}^{(K+1)} &\geq b_{\text{GA},i}^{(K)}, \quad \Delta_{\text{GA},i}^{(K+1)} \leq \Delta_{\text{GA},i}^{(K)} \quad \text{when } y_i = 0. \end{aligned}$$

This concludes the induction step and therefore completes the proof.

C.1.4 Proof of Lemma C.2

We prove Lemma C.2 by induction.

Our induction assumption is the following: there exist some constants $C_0 > 0, C_a \in (0, 1), C_b > 0, C_1, C_2$ depending only on $(B_\theta, b_x, B_x, \beta)$ such that when $\max_{i \neq j} |\gamma_{i,j}| \leq C_0/n_f$, for any $t \geq 1$ we have

1.

$$\begin{aligned} b_{\text{NPO},i}^{(t)} &\leq \dots \leq b_{\text{NPO},i}^{(0)} = 0, \quad \Delta_{\text{NPO},i}^{(t)} \geq \dots \geq \Delta_{\text{NPO},i}^{(0)} \quad \text{when } y_i = 1, \\ b_{\text{NPO},i}^{(t)} &\geq \dots \geq b_{\text{NPO},i}^{(0)} = 0, \quad \Delta_{\text{NPO},i}^{(t)} \leq \dots \leq \Delta_{\text{NPO},i}^{(0)} \quad \text{when } y_i = 0. \end{aligned}$$

2.

$$b_{\text{NPO},i}^{(t)} \in \left[-\frac{1}{\beta} \log(C_b\eta_0 t + 1), -\frac{1}{\beta} \log(C_a\eta_0 t + 1)\right] \quad \text{when } y_i = 1, \quad (\text{C.1})$$

$$b_{\text{NPO},i}^{(t)} \in \left[\frac{1}{\beta} \log(C_a\eta_0 t + 1), \frac{1}{\beta} \log(C_b\eta_0 t + 1)\right] \quad \text{when } y_i = 0. \quad (\text{C.2})$$

3.

$$b_{\text{NPO},i}^{(t)} \in [-C_2 - \frac{1}{\beta} \log(\eta_0 t + 1), -C_1 - \frac{1}{\beta} \log(\eta_0 t + 1)] \text{ when } y_i = 1, \quad (\text{C.3})$$

$$b_{\text{NPO},i}^{(t)} \in [C_1 + \frac{1}{\beta} \log(\eta_0 t + 1), C_2 + \frac{1}{\beta} \log(\eta_0 t + 1)] \text{ when } y_i = 0. \quad (\text{C.4})$$

Lemma C.2 follows immediately from the second part of the induction assumption.

In the following, we first specify the parameter-dependent constants C_0, C_1, C_2, C_a, C_b in the $t = 1$ case and prove the induction assumption when $k = 1$. Then given the induction assumption holds when $t \leq K$, we prove that it holds when $t = K + 1$ as well.

Case 1: $t = 1$ When $t = 1$, since $|c_{\text{init}}| \leq \|x_i\|_2 \cdot \|\theta_{\text{init}}\|_2 \leq B_x B_\theta$, it follows from the definition of $\text{pred}_i(\cdot)$ that

$$C_3 \leq \Delta_{\text{NPO},i}^{(0)} \leq 1 \text{ when } y_i = 1, \quad -1 \leq \Delta_{\text{NPO},i}^{(0)} \leq -C_4 \text{ when } y_i = 0 \quad (\text{C.5})$$

for all $i \in [n_f]$ for some constants $C_3, C_4 \in (0, 1)$ depending only on (B_θ, b_x, B_x) . Moreover, we claim that

$$\mathbf{W}_i^{(t)} \in \left[C_5 \cdot \exp\left((2y_i - 1)\beta b_{\text{NPO},i}^{(t)}\right), C_6 \cdot \exp\left((2y_i - 1)\beta b_{\text{NPO},i}^{(t)}\right) \right] \quad (\text{C.6})$$

for all $i \in [n_f]$ and t such that $\text{pred}_i(b_{\text{NPO},i}^{(t)}) \leq \text{pred}_i(b_{\text{NPO},i}^{(0)})$ for some $(B_\theta, b_x, B_x, \beta)$ -dependent constants $C_5, C_6 > 0$.

Now, suppose Eq. (C.3) and (C.4) hold for some $(B_\theta, b_x, B_x, \beta)$ -dependent constants $C_1, C_2 > 0$ which we will specify later. Then, there exists a constant $C_0 > 0$ depending on $C_{1:6}, b_x$ such that

$$\left| \sum_{j \neq i} \gamma_{i,j} \Delta_{\text{NPO},j}^{(0)} \mathbf{W}_j \right| \leq \frac{\gamma_{i,i}}{2} \left| \Delta_{\text{NPO},i}^{(0)} \mathbf{W}_i \right| \quad (\text{C.7})$$

for all $i \in [n_f]$ when $\max_{i \neq j} |\gamma_{i,j}| \leq C_0/n_f$. Furthermore, combining Eq. (C.5), (C.6), (C.7) gives

$$\begin{aligned} -\eta_0 \cdot \gamma_i^\top \text{diag}\{\Delta_{\text{NPO}}^{(0)}\} \mathbf{W}_i^{(0)} &\in \left[-\frac{3}{2} \gamma_{i,i} \eta_0 |\Delta_{\text{NPO},i}^{(0)}| \mathbf{W}_i^{(0)}, -\frac{1}{2} \gamma_{i,i} \eta_0 |\Delta_{\text{NPO},i}^{(0)}| \mathbf{W}_i^{(0)} \right] \\ &\in [-C_8 \eta_0 \exp(\beta b_{\text{NPO},i}^{(0)}), -C_7 \eta_0 \exp(\beta b_{\text{NPO},i}^{(0)})] \text{ when } y_i = 1, \\ -\eta_0 \cdot \gamma_i^\top \text{diag}\{\Delta_{\text{NPO}}^{(0)}\} \mathbf{W}_i^{(0)} &\in \left[\frac{1}{2} \gamma_{i,i} \eta_0 |\Delta_{\text{NPO},i}^{(0)}| \mathbf{W}_i^{(0)}, \frac{3}{2} \gamma_{i,i} \eta_0 |\Delta_{\text{NPO},i}^{(0)}| \mathbf{W}_i^{(0)} \right] \\ &\in [C_7 \eta_0 \exp(-\beta b_{\text{NPO},i}^{(0)}), C_8 \eta_0 \exp(\beta b_{\text{NPO},i}^{(0)})] \text{ when } y_i = 0, \end{aligned}$$

if $\max_{i \neq j} |\gamma_{i,j}| \leq C_0/n_f$. Here $C_7, C_8 > 0$ are some $(C_3, C_4, C_5, C_6, \beta)$ -dependent constants, and we pick C_7 such that $C_7 \beta < 1$. As a consequence,

$$b_{\text{NPO},i}^{(1)} \leq b_{\text{NPO},i}^{(0)} = 0, \quad \Delta_{\text{NPO},i}^{(1)} \geq \Delta_{\text{NPO},i}^{(0)} \text{ when } y_i = 1,$$

$$b_{\text{NPO},i}^{(1)} \geq b_{\text{NPO},i}^{(0)} = 0, \quad \Delta_{\text{NPO},i}^{(1)} \leq \Delta_{\text{NPO},i}^{(0)} \quad \text{when } y_i = 0.$$

This concludes the proof of the first part of the induction assumption.

Now, we start to prove the second part of the induction assumption. For i such that $y_i = 1$, consider the ordinary differential equations

$$\begin{aligned} b_l'(t) &= -C_8\eta_0(1 + \exp(C_8)) \cdot \exp(\beta b_l(t)), & b_l(0) &= 0; \\ b_u'(t) &= -C_7\eta_0 \cdot \exp(\beta b_l(t)), & b_u(0) &= 0. \end{aligned}$$

It can be verified that the ODEs have closed-form solutions

$$\begin{aligned} b_l(t) &= -\frac{1}{\beta} \log(\beta C_8\eta_0(1 + \exp(C_8))t + 1), \\ b_u(t) &= -\frac{1}{\beta} \log(\beta C_7\eta_0 t + 1). \end{aligned}$$

Since

$$b_{\text{NPO},i}^{(1)} = b_{\text{NPO},i}^{(1)} - \eta_0 \cdot \gamma_i^\top \Delta_{\text{NPO},i}^{(0)} \geq b_{\text{NPO},i}^{(1)} - C_8\eta_0 \exp(\beta b_{\text{NPO},i}^{(0)}) \geq b_{\text{NPO},i}^{(1)} - C_8\eta_0,$$

it follows that for any point $b_{\text{NPO},i}^{(\varepsilon)} = \varepsilon b_{\text{NPO},i}^{(1)} + (1 - \varepsilon)b_{\text{NPO},i}^{(0)}$ with $\varepsilon \in [0, 1]$

$$\begin{aligned} -C_8\eta_0(1 + \exp(C_8\eta_0)) \cdot \exp(\beta b_{\text{NPO},i}^{(\varepsilon)}) &\leq -C_8\eta_0 \exp(\beta b_{\text{NPO},i}^{(0)}) \leq -C_7\eta_0 \exp(\beta b_{\text{NPO},i}^{(0)}) \\ &\leq -C_7\eta_0 \cdot \exp(\beta b_{\text{NPO},i}^{(\varepsilon)}). \end{aligned}$$

Therefore, we have by the comparison theorem for ODEs that

$$b_l(\varepsilon) \leq b_{\text{NPO},i}^{(\varepsilon)} \leq b_u(\varepsilon)$$

for $\varepsilon \in [0, 1]$. Setting

$$C_a := C_7\beta, \quad C_b := \beta C_8(1 + \exp(C_8))$$

concludes the second part of the induction assumption.

For the last part of the induction assumption, since $\log(x + 1) + \log(c) \leq \log(cx + 1) \leq \log(x + 1) + \log(c + 1)$ when $c \leq 1$, we have

$$-\frac{1}{\beta} [\log(\eta_0 t + 1) + \log(\beta C_8(1 + \exp(C_8)))] \leq b_l(t) \leq b_u(t) \leq -\frac{1}{\beta} [\log(\eta_0 t + 1) + \log(\beta C_7)],$$

where the last inequality uses $\beta C_7 < 1$. Therefore, we obtain

$$b_{\text{NPO},i}^{(1)} \in [-\frac{1}{\beta} \log(\eta_0 t + 1) - C_2, -\frac{1}{\beta} \log(\eta_0 t + 1) - C_1],$$

where

$$C_1 := \frac{1}{\beta} \log(\beta C_7), \quad C_2 := \frac{1}{\beta} \log(\beta C_8(1 + \exp(C_8)))$$

for i such that $y_i = 1$. Following the same arguments, similarly, we also have

$$b_{\text{NPO},i}^{(1)} \in [\frac{1}{\beta} \log(\eta_0 t + 1) + C_1, \frac{1}{\beta} \log(\eta_0 t + 1) + C_2]$$

for i such that $y_i = 0$. This concludes the last part of the induction assumption.

Case 2: $t = K + 1$ Suppose the induction assumption holds for $t \in [K]$, we now show that the induction assumption holds for $t = K + 1$ as well. Following the proof of $t = 1$ case and using the monotonicity property of $\{\Delta_{\text{NPO},i}^{(t)}\}_{t=0}^K$, we have

$$C_3 \leq \Delta_{\text{NPO},i}^{(K)} \leq 1 \text{ when } y_i = 1, \quad -1 \leq \Delta_{\text{NPO},i}^{(K)} \leq -C_4 \text{ when } y_i = 0$$

for all $i \in [n_f]$. Since $\text{pred}_i(b_{\text{NPO},i}^{(K)}) \leq \text{pred}_i(b_{\text{NPO},i}^{(K)})$ by the definition of pred_i and the monotonicity of $\{b_{\text{NPO},i}^{(t)}\}_{t=0}^K$, it follows from Claim (C.6) that

$$\mathbf{W}_i^{(K)} \in \left[C_5 \cdot \exp\left((2y_i - 1)\beta b_{\text{NPO},i}^{(K)}\right), C_6 \cdot \exp\left((2y_i - 1)\beta b_{\text{NPO},i}^{(K)}\right) \right].$$

Putting the last two displays together and following the same argument as the $t = 1$ case, we find that

$$\begin{aligned} -\eta_0 \cdot \gamma_i^\top \text{diag}\{\Delta_{\text{NPO}}^{(K)}\} \mathbf{W}_i^{(K)} &\in [-C_8 \eta_0 \exp(\beta b_{\text{NPO},i}^{(K)}), -C_7 \eta_0 \exp(\beta b_{\text{NPO},i}^{(K)})] \text{ when } y_i = 1, \\ -\eta_0 \cdot \gamma_i^\top \text{diag}\{\Delta_{\text{NPO}}^{(K)}\} \mathbf{W}_i^{(K)} &\in [C_7 \eta_0 \exp(-\beta b_{\text{NPO},i}^{(K)}), C_8 \eta_0 \exp(\beta b_{\text{NPO},i}^{(K)})] \text{ when } y_i = 0. \end{aligned}$$

The first part of the induction assumption (for $t = K + 1$) follows immediately as the sign of the gradient updates $-\eta_0 \cdot \gamma_i^\top \text{diag}\{\Delta_{\text{NPO}}^{(K)}\} \mathbf{W}_i^{(K)}$ are determined as above.

Note that $b_l(K) \leq b_{\text{NPO},i}^{(K)} \leq b_u(K)$ by the induction assumption. Similarly, using the comparison theorem of ODEs, we obtain

$$b_l(K + \varepsilon) \leq b_{\text{NPO},i}^{(K+\varepsilon)} \leq b_u(K + \varepsilon)$$

for $\varepsilon \in [0, 1]$ and $b_{\text{NPO},i}^{(K+\varepsilon)} := \varepsilon b_{\text{NPO},i}^{(K+1)} + (1 - \varepsilon) b_{\text{NPO},i}^{(K)}$. Choosing $\varepsilon = 1$ gives the second part of the induction assumption for $t = K + 1$. The last part of the induction assumption for $t = K + 1$ follows from the same algebra as in the $t = 1$ case.

Proof of Claim (C.6) By definition

$$\mathbf{W}_i^{(t)} = \frac{\pi_{\theta_{\text{NPO}}^{(t)}}^\beta(y_i \mid x_i)}{\pi_{\theta_{\text{NPO}}^{(t)}}^\beta(y_i \mid x_i) + \pi_{\text{ref}}^\beta(y_i \mid x_i)} = \frac{\text{pred}_i(b_{\text{NPO},i}^{(t)})^\beta}{\text{pred}_i(b_{\text{NPO},i}^{(t)})^\beta + \text{pred}_i(b_{\text{NPO},i}^{(0)})^\beta}.$$

When $\text{pred}_i(b_{\text{NPO},i}^{(t)}) \leq \text{pred}_i(b_{\text{NPO},i}^{(0)})$, we have

$$\mathbf{W}_i^{(t)} \in \left[\frac{\text{pred}_i(b_{\text{NPO},i}^{(t)})^\beta}{2\text{pred}_i(b_{\text{NPO},i}^{(0)})^\beta}, \frac{\text{pred}_i(b_{\text{NPO},i}^{(t)})^\beta}{\text{pred}_i(b_{\text{NPO},i}^{(0)})^\beta} \right] \in [C_l \cdot \text{pred}_i(b_{\text{NPO},i}^{(t)})^\beta, C_u \cdot \text{pred}_i(b_{\text{NPO},i}^{(t)})^\beta] \quad (\text{C.8})$$

for some constants $C_l, C_u > 0$ depending only on $(B_\theta, b_x, B_x, \beta)$. Note that $\text{pred}_i(b_{\text{NPO},i}^{(t)}) \leq \text{pred}_i(b_{\text{NPO},i}^{(0)})$ is equivalent to $(1 - 2y_i)b_{\text{NPO},i}^{(t)} \geq 0$. Therefore, under this condition, we have

$$\text{pred}_i(b_{\text{NPO},i}^{(t)}) = \frac{1}{1 + \exp\left((1 - 2y_i)c_{\text{init},i} + (1 - 2y_i)b_{\text{NPO},i}^{(t)}\right)}$$

$$\begin{aligned}
&= \frac{\exp\left((2y_i - 1)b_{\text{NPO},i}^{(t)}\right)}{\exp\left((2y_i - 1)b_{\text{NPO},i}^{(t)}\right) + \exp\left((1 - 2y_i)c_{\text{init},i}\right)} \\
&\in \left[\frac{\exp\left((2y_i - 1)b_{\text{NPO},i}^{(t)}\right)}{1 + \exp\left((1 - 2y_i)c_{\text{init},i}\right)}, \frac{\exp\left((2y_i - 1)b_{\text{NPO},i}^{(t)}\right)}{\exp\left((1 - 2y_i)c_{\text{init},i}\right)} \right]. \tag{C.9}
\end{aligned}$$

Putting Eq. (C.8) and (C.9) together and recalling that $|c_{\text{init},i}| \leq B_\theta B_x$, we obtain

$$W_i^{(t)} \in \left[C_5 \cdot \exp\left((2y_i - 1)\beta b_{\text{NPO},i}^{(t)}\right), C_6 \cdot \exp\left((2y_i - 1)\beta b_{\text{NPO},i}^{(t)}\right) \right]$$

for some constants $C_5, C_6 > 0$ depending only on $(B_\theta, b_x, B_x, \beta)$. This concludes the proof of Claim (C.6).

C.2 The role of KL Divergence on the Forget Set

In this section, we report *Forget KL*, the KL divergence between the output distributions of the initial model and the unlearned model on the forget set, defined as $\mathbb{E}_{\mathcal{D}_{\text{FG}}} \text{KL}(\pi_{\text{ref}}(\cdot | x) || \pi_{\theta}(\cdot | x))$. In experiments on the synthetic dataset and TOFU dataset, we observe that the models exhibit better unlearning performance when the forget KL is maintained at a moderate level. This suggests that explicitly maximizing the forget KL may not be an ideal objective for unlearning tasks.

C.2.1 Synthetic Experiment

We present the forget KL for the synthetic experiments in Figure C.1 (a) and (b). Combining with Figure 4.4, we find that the unlearned models attain Pareto frontiers when the forget KL is suitably large—an excessively large forget KL (as in GA, GA + RT after 1200 steps) or an excessively small forget KL (as in IDK + RT) may deteriorate the unlearning performance.

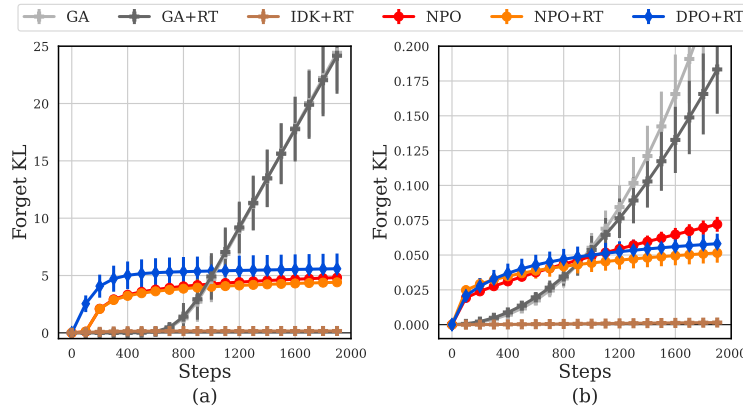


Figure C.1: Forget KL versus optimization steps for all methods in the synthetic experiment. (a): $\alpha = 1$, (b): $\alpha = 0$. The errorbars denote ± 1 standard deviation over 5 runs.

C.2.2 Experiments on the TOFU dataset

We also examine the Forget KL during the unlearning process in the TOFU dataset. We first observed that while GA and GA+RT tend to induce an explosively large KL divergence on the forget set during the unlearning process, the NPO-based approaches induce a much slower growth of Forget KL (Figure 4.8). It stabilizes at a moderate level even after several epochs. One natural insight from this distinction is that even in the context of unlearning, a larger KL divergence on the forget set is not necessarily advantageous. Rather, a moderate and stabilized KL divergence on the forget set is preferable, which ensures the unlearned models generate fluent outputs with reasonable linguistic structures but incorrect content. This also suggests that KL divergence on the forget set may not be a suitable objective function to

maximize for unlearning LLMs, contrary to what was done in some prior literature (Chen and Yang, 2023).

C.3 Ablation experiments on the role of retain loss

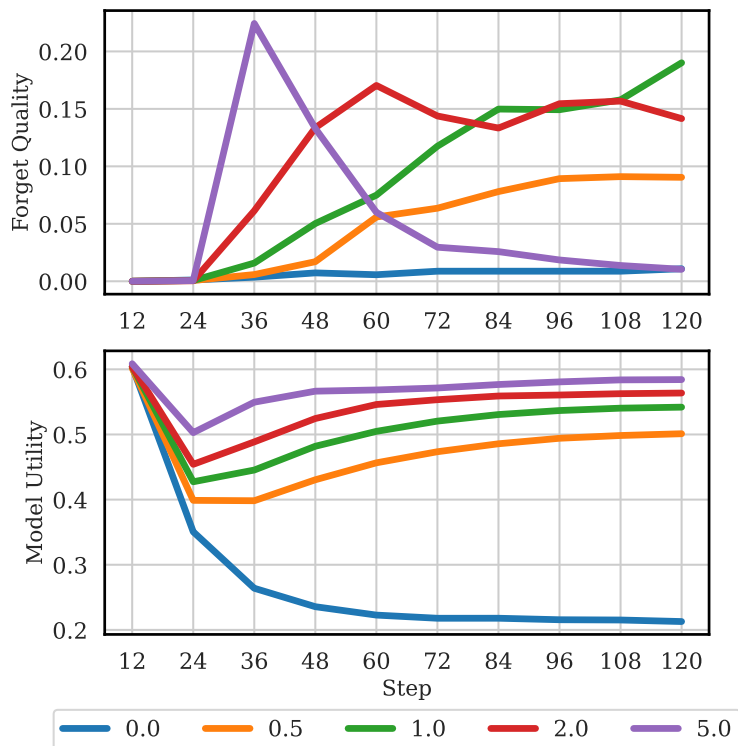


Figure C.2: The evolution of the forget quality and model utility when we adjust the weights of the NPO term and the retain loss term in NPO+RT. The experiments are performed on Forget10.

Maini et al. (2024) demonstrated that methods incorporating a retain set outperform those that solely optimize a loss function based on the forget set. It is natural that a regularization via retain loss will improve the performance on the retain set, which agrees with the tendency shown in Figure C.2 that the model utility monotonically grows as the weight of the retain loss increases. Surprisingly, with the help of the retain loss, the forget quality can also improve. Specifically, NPO+RT shows a much more significant forget quality as the weight of the retain loss grows from 0 to 2 (see the top-right panel of Figure 4.6 and Figure C.2). We conjecture that it is because fine-tuning the retain loss term can help the model to preserve answer templates and their linguistic structure while optimizing the NPO term forces the model to forget some specific facts. The combination of the two pushes the unlearned model

to approximate the retrained model by generating the outputs with similar templates but incorrect entities. However, further increasing the weight of the retain loss (e.g., from 2 to 5, in Figure C.2) leads to a drop in the forget quality, possibly due to the diminished scale of the NPO term. Notably, in our experiments, the retain loss plays a more significant role when we target forgetting a larger fraction of the data (See the middle and right panels of Figure 4.6).

C.4 Experimental details of the synthetic experiments

In this section, we discuss the experimental details of the synthetic experiments studied in Section 4.4.

Initial model and retrained model. We create the initial model π_{ref} and the retrained model π_{retr} via optimizing over θ using the cross-entropy loss on the entire dataset $\mathcal{D} = \mathcal{D}_{\text{FG}} \cup \mathcal{D}_{\text{RT}}$ and the retain dataset $\mathcal{D}_{\text{RT}}$, respectively. Concretely, initializing at $\theta = \mathbf{0}_{128}$, we run gradient descent for 20000 steps with the learning rate equals 0.05 to obtain the initial model π_{ref} and the retrained model π_{retr} .

Unlearning. During unlearning, starting from the initial model π_{ref} , we run gradient descent on each of the loss functions for 2000 steps with the learning rates selected via a grid search. We choose the learning rates so that the training remains stable within 2000 steps. It should be noted that variations of the learning rates may affect the number of steps needed to reach the minimal forget distance (or retain distance) in Figure 4.4 (a1, a2, b). However, they are less likely to alter the Pareto curves shown in Figure 4.4 (c). A grid search is also conducted to select the optimal β for NPO (and DPO)-based methods. The choices of learning rate and β are summarized in Table C.1.

Method	learning rate		β	
	$\alpha = 1$	$\alpha = 0$	$\alpha = 1$	$\alpha = 0$
GA,GA+RT, IDK+RT	5e-4	1e-4	NA	NA
NPO, NPO+RT	5e-3	5e-2	1	10
DPO+RT	5e-3	5e-2	0.1	5

Table C.1: Values of learning rate and β for different methods when $\alpha = 1$ and $\alpha = 0$ in the synthetic experiments.

C.5 Experiments on the TOFU dataset

In this section, we provide details of the experiments on the TOFU dataset (Maini et al., 2024). We first present a detailed explanation of the metrics, the baseline methods and the hyperparameters in the experiments. Then, we provide the full results on different levels of the tasks.

C.5.1 Experiments Setup

Dataset

TOFU Dataset. We evaluate NPO-based methods and all baselines on the TOFU (Task of Fictitious Unlearning) dataset (Maini et al., 2024) designed for measuring the unlearning methods for LLMs. TOFU contains 200 fictitious author profiles, each consisting of 20 question-answering pairs generated by GPT-4 based on a set of predefined attributes. These profiles are fictitious and do not exist in the pre-training data, providing a controlled environment for studying unlearning LLMs. TOFU introduces three levels of unlearning tasks, each aiming at forgetting a subset of 2, 10, and 20 authors (comprising 1%, 5%, and 10% of the training data, respectively), referred to as the *forget set* $\mathcal{D}_{\text{FG}}$, with a computational constraint that scales linearly with the size of the forget set. We refer to these tasks as Forget01, Forget05, and Forget10, respectively.

Dataset for Evaluation. In addition to the forget set, Maini et al. (2024) also introduced other datasets to measure the performance of the unlearned model. The *retain set* $\mathcal{D}_{\text{RT}}$ is the part of the data that we do not hope the model to forget, which is, by definition, the complementary set of the forget set in the full dataset. To evaluate the model performance on the retain set, TOFU earmarks a subset of 400 question-answer pairs, accounting for 10% of the data, as an exclusive retain set that is never included in the forget set for any task. Moreover, to measure the general capacities of the unlearned models, two additional datasets are introduced: the Real Authors set and the Real World set. The Real Authors set includes question-answer pairs about authors in the real world and often deals with neighboring concepts entangled with those in the forget set. The Real World set contains commonsense knowledge about the real world and is designed to examine the general world knowledge of the unlearned model.

Dataset beyond Forget10. In scenarios where the targeted forget set exceeds 10% of the data (Forget20, Forget30, Forget50, and Forget90), we reorganize the original forget and retain sets within the TOFU dataset. To assess the Truth Ratio on an evaluation dataset, it is necessary to utilize the perturbed and paraphrased answers, which in TOFU were generated by properly prompting GPT4. To avoid any potential distribution shift from the newly crafted responses and their original forms, we evaluate the Truth Ratio using the publicly available data within TOFU. Consequently, even for tasks that are beyond forget10, we continue to use the data from the standard forget10 subset to compute the Truth Ratio on

the forget set. This serves as a reasonable proxy for evaluating the Truth Ratio on the full forget set.

Metrics

Model Utility. We measure the general capacities of the unlearned model using *Model Utility*, which aggregates multiple metrics across the retain set, the real-world set and the real-author set. Given a question-answer pair $x = [q, a]$, we compute the normalized conditional probability $\mathbb{P}(a \mid q)^{1/|a|}$, where $|\cdot|$ denotes the number of tokens in a certain sequence. This probability is then averaged over the retain set, the Real Authors set, and the Real World set each. We also compute the averaged ROUGE-L recall score (Lin, 2004) across these datasets, a metric that evaluates the accuracy of the model’s response compared to the reference answers. Finally, we compute the averaged Truth Ratio on the three datasets above. The Truth Ratio defined in (Maini et al., 2024) measures how likely the unlearned model will give a correct answer versus an incorrect one. More specifically, given a question q , Maini et al. (2024) generated a paraphrased (correct) answer $\tilde{a}$ via prompting GPT4. They then generated five perturbed answers with the exactly same templates but incorrect answers $\hat{a}_i, i = 1, 2, 3, 4, 5$ in the same way. The Truth Ratio is defined by

$$R_{truth} := \frac{\frac{1}{5} \sum_{i=1}^5 \mathbb{P}(\hat{a}_i \mid q)^{1/|\hat{a}_i|}}{\mathbb{P}(\tilde{a} \mid q)^{1/|\tilde{a}|}}. \quad (\text{C.10})$$

The model utility is defined as the harmonic average of the nine metrics above (the probability, the ROUGE score, and the Truth Ratio on the retain set, the Real Authors set, and the Real World set).

Forget Quality. Forget quality assesses how well the unlearned model mimics the retrained model (defined as the model trained only on the retain set). This is a rigorous measurement as the ultimate goal for unlearning LLM is not only to stop generating the content related to the forget set but also to make the unlearned model indistinguishable from the retrained one. From a practical view of point, this requires the next-token probability given a prefix of the unlearned model to be as close as possible to that of the retrained model. In TOFU, they compute the Truth Ratio (defined in Eq. C.10) on each question-answer pair from the forget set. Instead of simply averaging them, they test whether the distribution of the Truth Ratio computed from the unlearned and the retrained models are indistinguishable. More specifically, they perform the Kolmogorov-Smirnov (KS) test and compute the p-value of the test. A large p-value indicates that the two models are indistinguishable from the Truth Ratio. When the p-value is above 0.05, we say the forgetting is significant.

Baseline Methods

In this section, we introduce the baseline methods in our experiments.

GA-based Methods. Gradient Ascent (GA) is a key component in many LLM unlearning methods. Performing GA is equivalent to doing Gradient Descent (GD) on the negative cross-entropy loss function, which is denoted as $\mathcal{L}_{\text{GA}}(\theta)$ defined in Eq. (4.1). Based on gradient ascent, a large class of unlearning methods performs gradient-based optimization on a linear combination of the GA loss $\mathcal{L}_{\text{GA}}$ and several other loss functions that encourage unlearning (Jang et al., 2022; Yao et al., 2023; Chen and Yang, 2023; Maini et al., 2024; Eldan and Russinovich, 2023). Such a loss function can be written as

$$\mathcal{L}(\theta) = c_{\text{GA}}\mathcal{L}_{\text{GA}}(\theta) + c_{\text{FG}}\mathcal{L}_{\text{FG}}(\theta) + c_{\text{RT}}\mathcal{L}_{\text{RT}}(\theta) - c_{\text{FGKL}}\mathcal{K}_{\text{FG}}(\theta) + c_{\text{RTKL}}\mathcal{K}_{\text{RT}}(\theta), \quad (\text{C.11})$$

where $c_{\text{GA}}, c_{\text{FG}}, c_{\text{RT}}, c_{\text{FGKL}}, c_{\text{RTKL}}$ are non-negative weights. Here, $\mathcal{L}_{\text{FG}}(\theta) = -\mathbb{E}_{\mathcal{D}_{\text{FG}}}[\log(\pi_{\theta}(\tilde{y}|x))]$ is the Forget loss where $(x, y) \sim \mathcal{D}_{\text{FG}}$ and $\tilde{y} \neq y$ is any response for prompt x which show some extent of ignorance towards the question x . $\mathcal{L}_{\text{RT}}(\theta) = -\mathbb{E}_{\mathcal{D}_{\text{RT}}}[\log(\pi_{\theta}(y|x))]$ is the retain loss. $\mathcal{K}_{\text{FG}}(\theta) = \mathbb{E}_{\mathcal{D}_{\text{FG}}}[\text{D}(\pi_{\theta}(\cdot|x)||\pi_{\text{ref}}(\cdot|x))]$ is the expected KL divergence on the forget set. $\mathcal{K}_{\text{RT}}(\theta) = \mathbb{E}_{\mathcal{D}_{\text{RT}}}[\text{D}(\pi_{\theta}(\cdot|x)||\pi_{\text{ref}}(\cdot|x))]$ is the expected KL divergence on the retain set. In our experiments, we use three GA-based methods reported in (Maini et al., 2024), referred to as GA, GA+RT, GA+KL, which fall in this class of loss function. The weights in Eq. (C.11) are shown in Table C.2.

Loss	c_{GA}	c_{FG}	c_{RT}	c_{FGKL}	c_{RTKL}
GA	1	0	0	0	0
GA+RT	1	0	1	0	0
GA+KL	1	0	0	0	1
IDK+RT	0	1	1	0	0

Table C.2: The weights for different components in GA-based losses and IDK+RT loss.

IDK-based Methods ('I don't know'). Maini et al. (2024) proposed IDK+RT, which is a supervised loss function comprising of the retain loss and IDK loss term. The IDK loss term $\mathcal{L}_{\text{FG}}$ is the averaged cross-entropy loss for question-answer pairs with questions x from the forget set $\mathcal{D}_{\text{FG}}$ and answers y replaced by $\tilde{y} = \text{'I don't know'}$ or a similar sentence showing ignorance towards this question. IDK+RT does not involve GA loss, and in general, IDK+RT loss shows a higher stability than GA-based methods.

DPO-based Methods. We also tested the DPO method (Rafailov et al., 2024) and its variants by adding either the retain loss or the KL divergence on the retain set. In the DPO loss, we take 'I don't know' or its variants as positive responses and the answers in the forget set as negative responses. We use $\beta = 0.1$ in all DPO-based experiments, which is commonly recognized as the optimal inverse temperature in most cases.

KTO-based Methods. We examine Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024), an alignment method with only non-paired preference data. The objective function of KTO is (we use a slightly different version than the original one as in (Ethayarajh et al., 2024))

$$\mathcal{L}_{\text{KTO}} := \frac{2}{\beta} \mathbb{E}_{\mathcal{D}_{\text{FG}}} \left[-\log \sigma \left(z_{\text{ref}} - \beta \log \frac{\pi_{\theta}(y | x)}{\pi_{\text{ref}}(y | x)} \right) \right], \quad (\text{C.12})$$

where $\beta > 0$ is the inverse-temperature, σ is the sigmoid function, and

$$z_{\text{ref}} := \mathbb{E}_{x \sim \mathcal{D}_{\text{FG}}} [\beta \cdot \text{D}(\pi_{\theta}(\cdot | x) || \pi_{\text{ref}}(\cdot | x))]. \quad (\text{C.13})$$

Following Ethayarajh et al. (2024), we estimate the KL term via averaged log probability ratio for questions in the forget set and answers in the "I don't know" set (as the unrelated outputs). We examine both KTO and KTO+RT in our experiments with $\beta = 0.1$.

Hyperparameters

For the experiments TOFU, we use Llama2-7b-chat model (Touvron et al., 2023). All experiments are conducted with two A100 GPUs. We use AdamW with a weight decay of 0.01 and a learning rate of 10^{-5} in all finetuning, retraining, and unlearning experiments, which agrees with the setting in (Maini et al., 2024). We use an effective batch size of 32 for all experiments. In finetuning and retraining, we train for 5 epochs, while we train for 10 epochs in the unlearning process. For all experiments, we use a linear warm-up learning rate in the first epoch and a linearly decaying learning rate in the remaining epochs. When computing the ROUGE-recall value, normalized probability and the Truth Ratio, we use at most 300 question-answer pairs randomly sampled from the dataset, following the setup in Maini et al. (2024).

C.5.2 Full Results

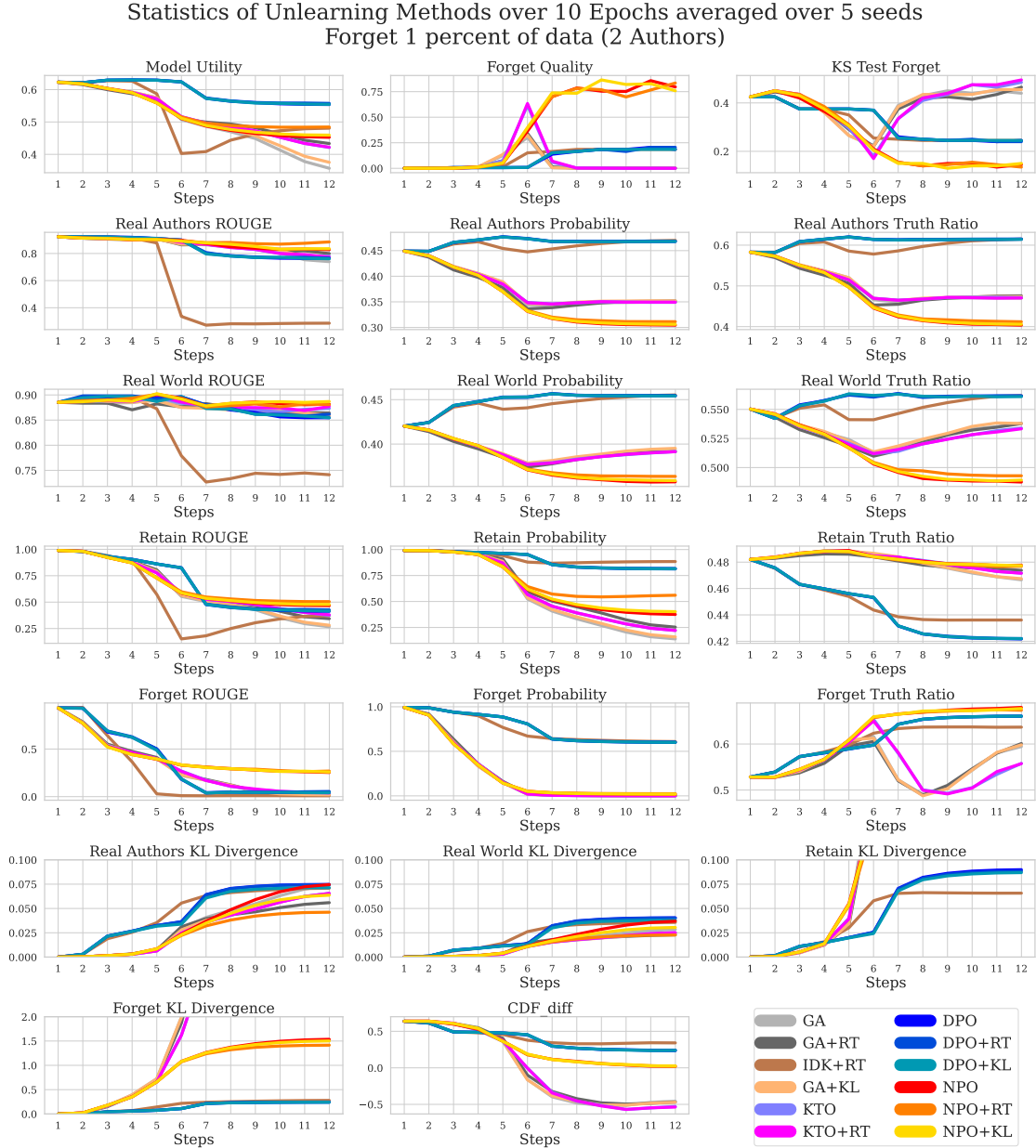


Figure C.3: Statistics for NPO-based methods and baselines on the Forget01 task of TOFU.

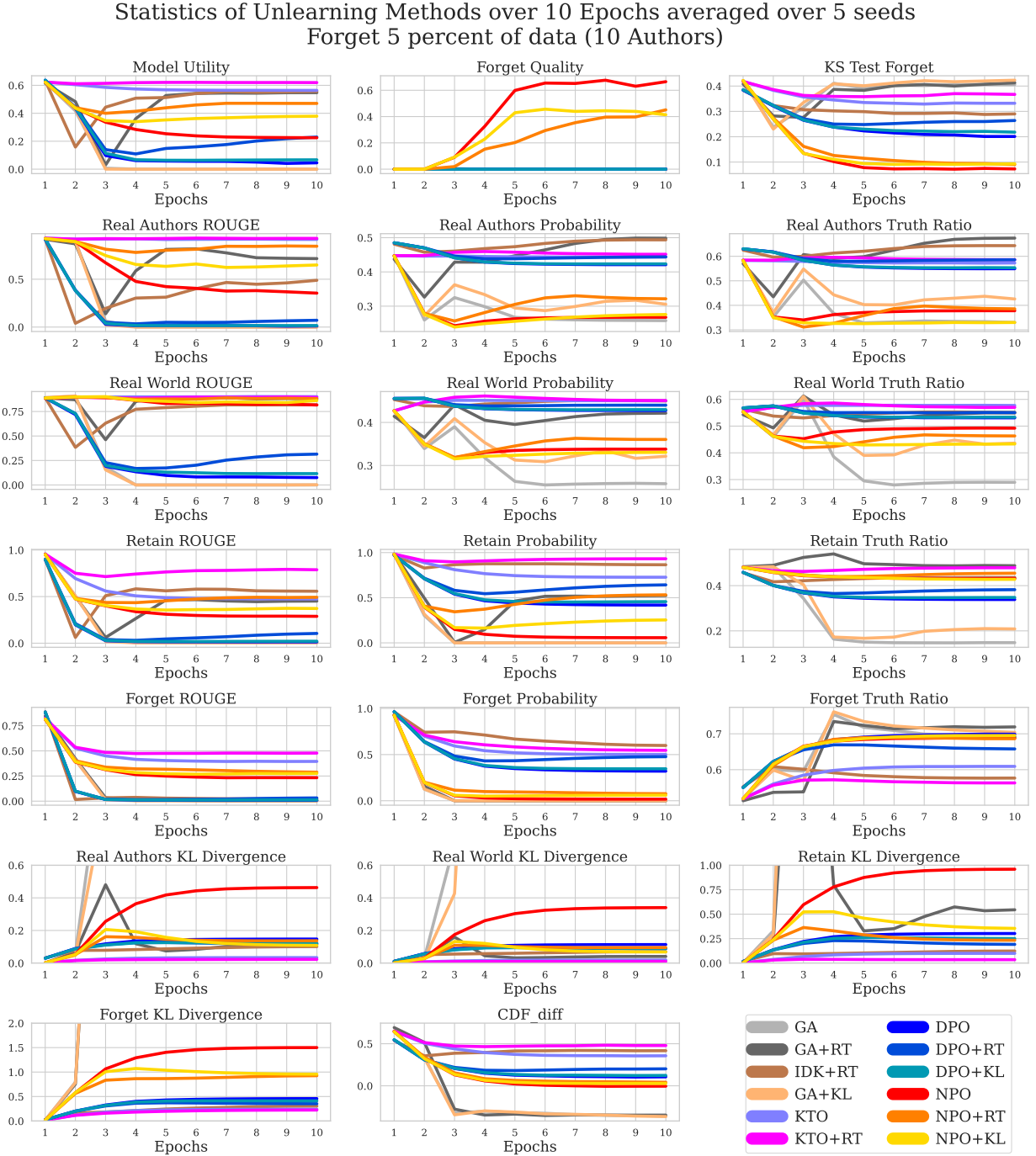


Figure C.4: Statistics for NPO-based methods and baselines on the Forget05 task of TOFU.

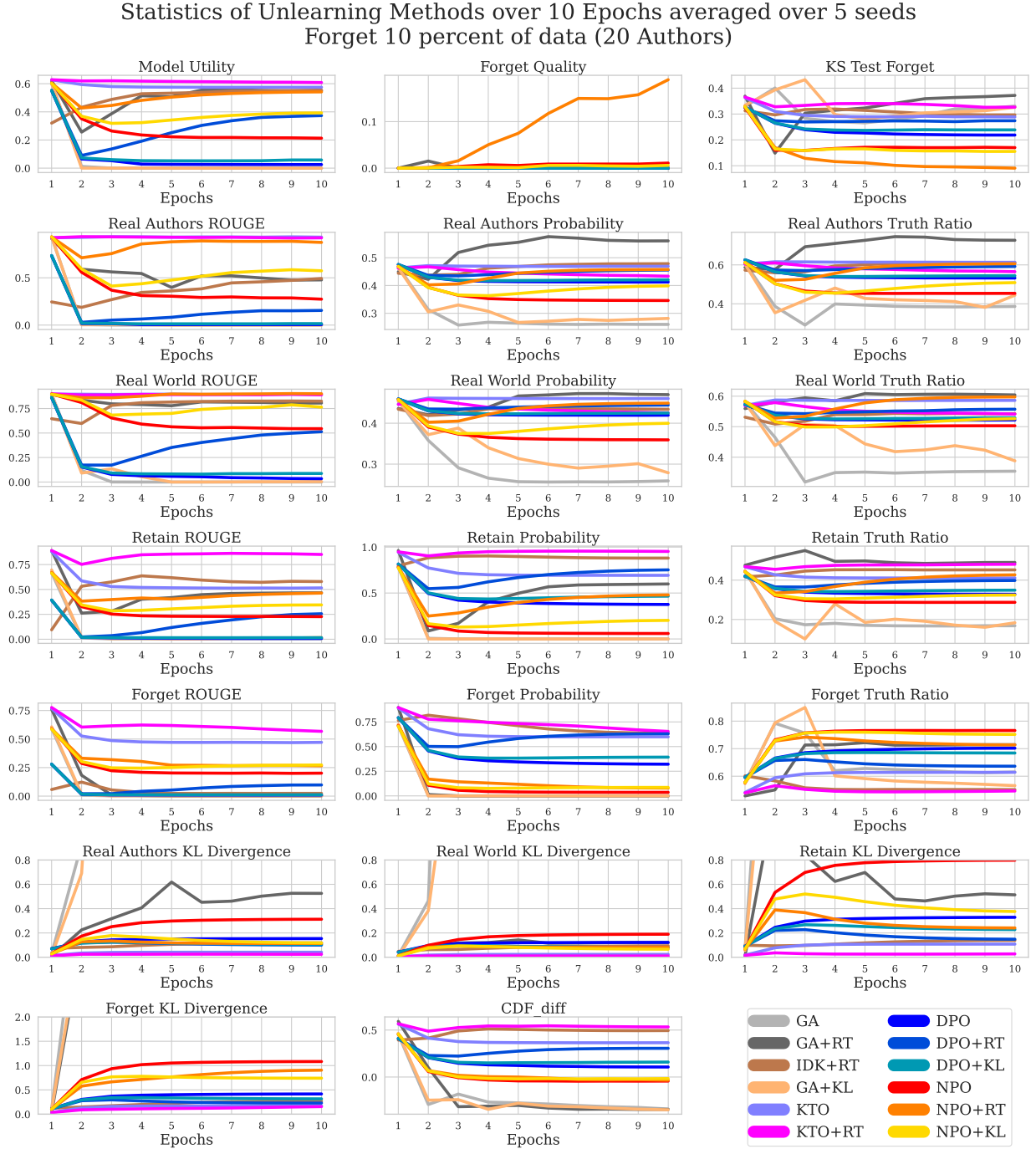


Figure C.5: Statistics for NPO-based methods and baselines on the Forget10 task of TOFU.

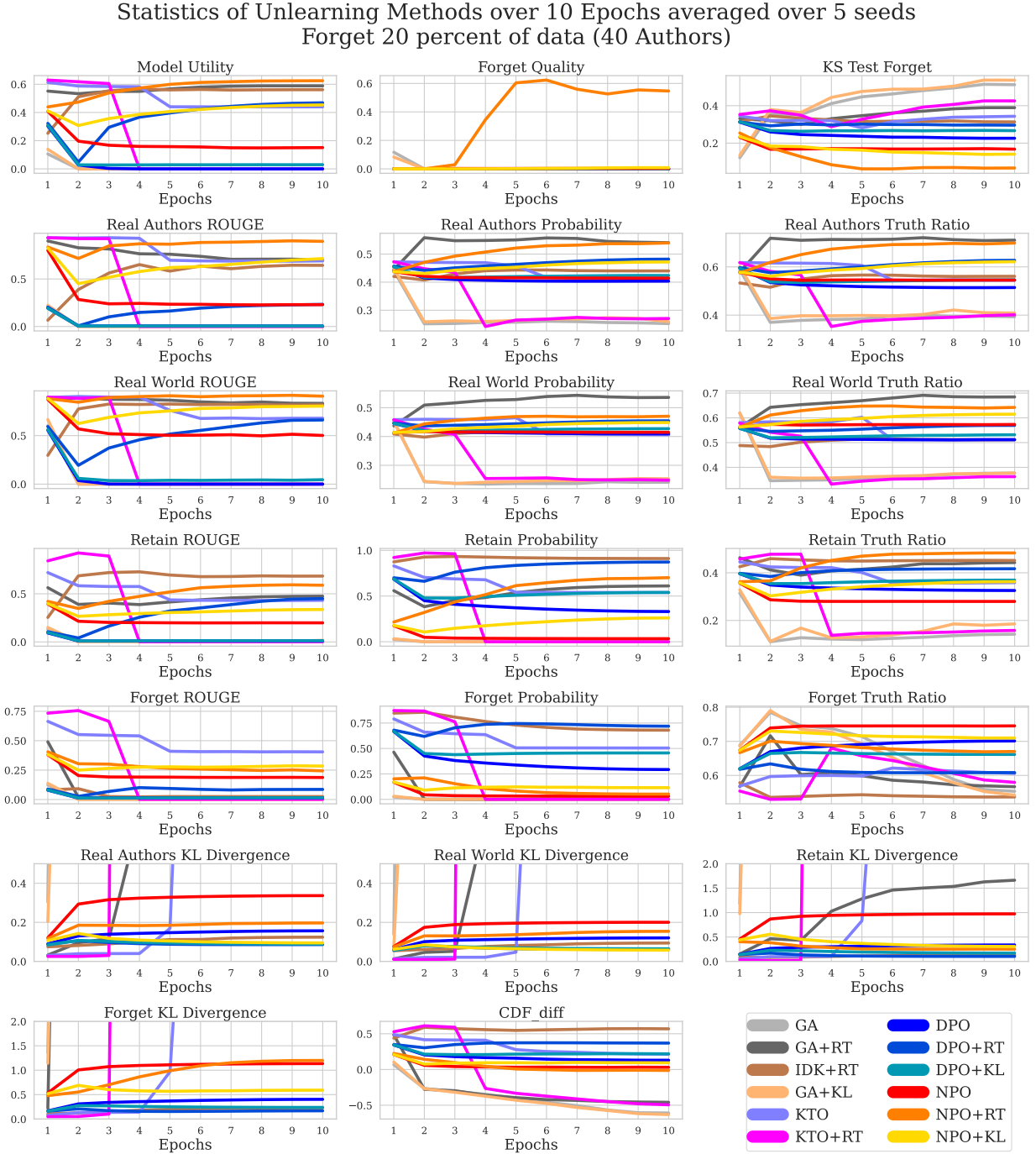


Figure C.6: Statistics for NPO-based methods and baselines on the Forget20 task of TOFU.

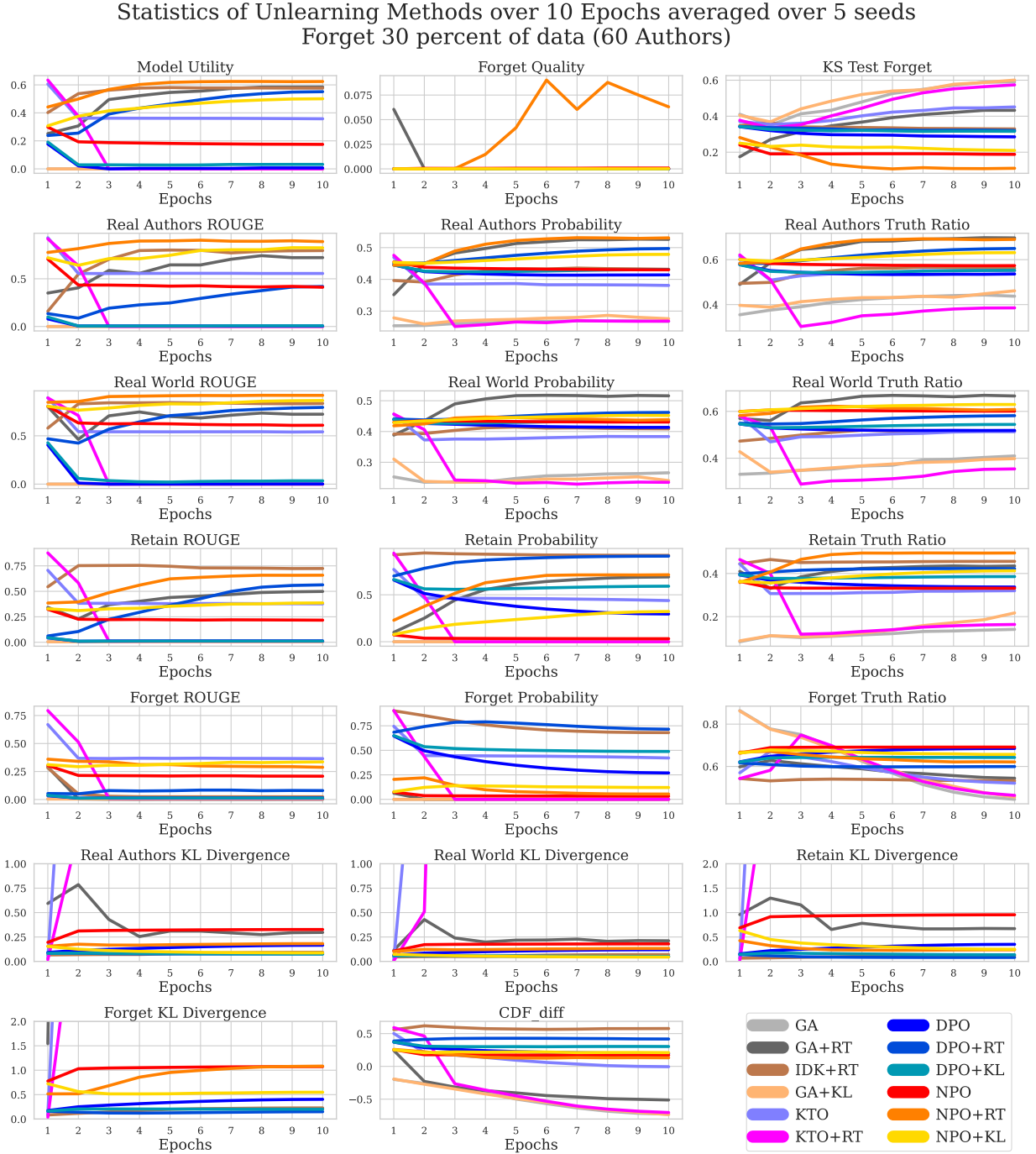


Figure C.7: Statistics for NPO-based methods and baselines on the Forget30 task of TOFU.

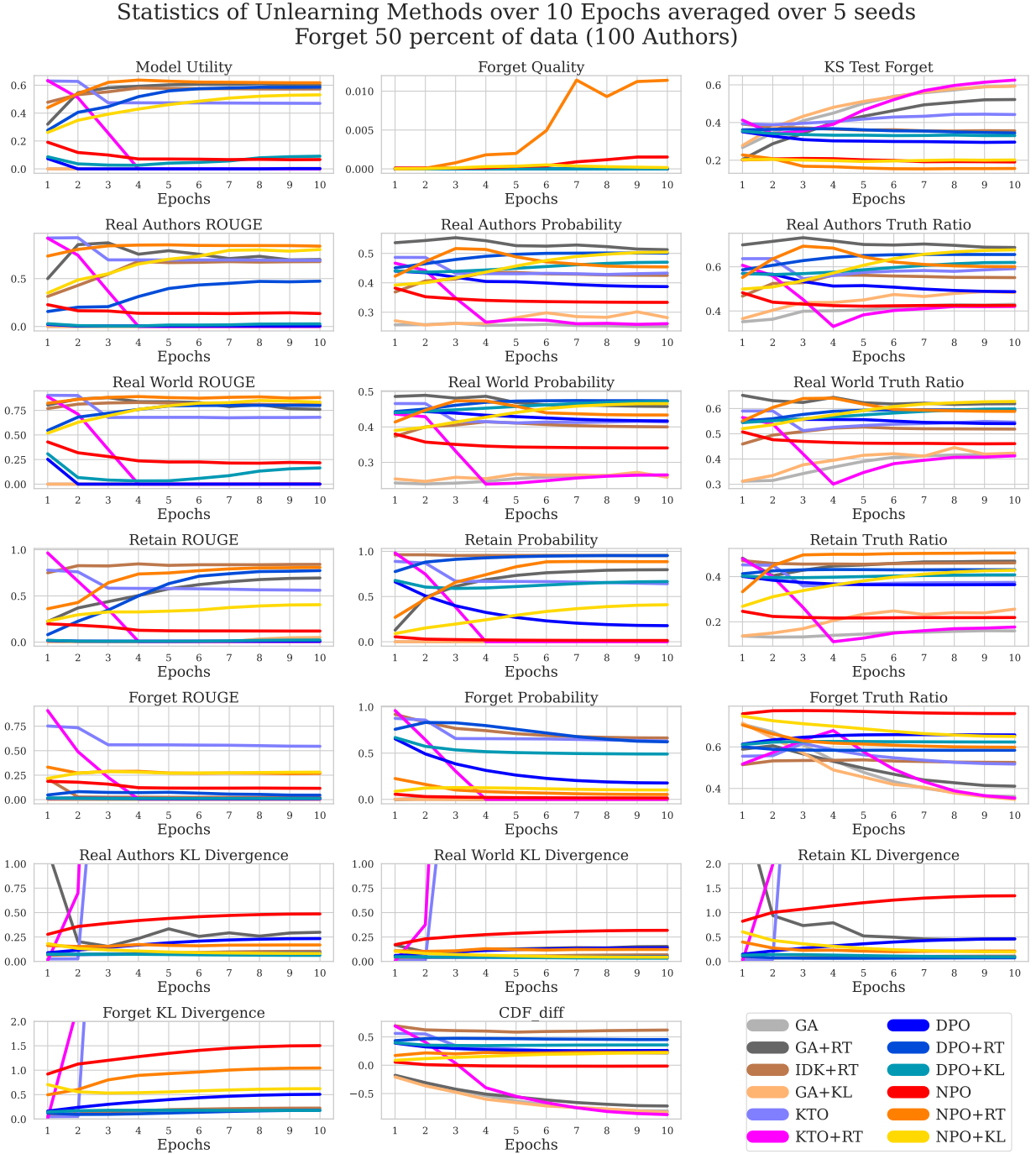


Figure C.8: Statistics for NPO-based methods and baselines on the Forget50 task of TOFU.