

# UC Berkeley

## UC Berkeley Electronic Theses and Dissertations

### Title

Sampling in Statistical Inference and Machine Learning

### Permalink

<https://escholarship.org/uc/item/3241p309>

### ISBN

9798189271120

### Author

Wu, David Xing

### Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

Sampling in Statistical Inference and Machine Learning

By  
David Xing Wu

A dissertation submitted in partial satisfaction of the  
requirements for the degree of  
Doctor of Philosophy  
in  
Computer Science  
in the  
Graduate Division  
of the  
University of California, Berkeley

Committee in charge:  
Professor Prasad Raghavendra, Co-Chair  
Professor Anant Sahai, Co-Chair  
Professor Alistair Sinclair  
Professor Nikhil Srivastava

Spring 2026

# Sampling in Statistical Inference and Machine Learning

Copyright 2026

By

David Xing Wu

Abstract

Sampling in Statistical Inference and Machine Learning

By

David Xing Wu

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Prasad Raghavendra, Co-Chair

Professor Anant Sahai, Co-Chair

Sampling algorithms based on Markov chain Monte Carlo (MCMC) are ubiquitous across theory and practice. The success of MCMC algorithms is typically established by upper bounding their mixing time, and over the past several decades a rich theory for doing so has emerged.

However, classical mixing time based arguments fail to explain why MCMC has been so successful for statistical inference and the scientific simulation. The classical notion of rapid worst-case mixing is often too stringent of a notion to apply to these settings. Instead, one needs to analyze warm start mixing, or analyze the inference and optimization performance of MCMC before it has mixed. In this thesis, we build new theory in these directions to explain MCMC's success in these difficult-to-understand settings.

We then pivot to an entirely different subject matter altogether: that of synthetic data in modern machine learning. We study the counterintuitive success of synthetic data in regimes where the data is generated from models that are weaker than their students. We also probe the limitations of synthetic data to elicit certain reasoning behaviors from language models.



To my family.

## Acknowledgments

My greatest thanks go to my wonderful co-advisors, Anant Sahai and Prasad Raghavendra. They have supported me through thick and thin, especially during the challenging periods of the PhD where I had no clue what direction I should head in. It's easy to feel inspired by their strong sense of research taste and curiosity. I have started to appreciate Prasad's analogy of research as going on a hike: there is always something to appreciate on the path to the summit, if you just remember to look up.

Thanks also to Sidhanth Mohanty, who took me under his wing when I was a first year student. Although we only overlapped for a year at Berkeley, our research journey extended through the rest of my PhD. Between Messenger, Zoom, and visits to Boston, Berkeley, and Chicago, it didn't feel like we were working remotely at all. Thanks for teaching me about the trace method, stochastic localization, and for inspiring me to be a positive and persistent researcher in every research meeting. The lessons that I've learned extend far beyond math, for which I feel extremely fortunate.

I am grateful for the guidance of Alistair Sinclair, from whom I've received a lot of useful advice and learned a lot from various research discussions about sampling. His sampling class also inspired me to keep pushing on those research directions.

Kuikui Liu has always been such a warm, welcoming presence every time I visited MIT. I enjoyed our research chats very much and appreciated the thoughtful, deliberate attitude he brings to his research. It was thanks to him that I got to see the breathtaking New England fall colors for the first time.

I've learned so much about spin glasses and probability theory from collaborations with Brice Huang, and he was always been very kind and patient whenever I asked him about my harebrained ideas.

I am consistently impressed in every interaction I have with Amit Rajaraman—not just because of his research ideas, but also his playful attitude towards life and tackling hard problems. Thanks for coming along for the rollercoaster ride of our RGD project after I pitched it to you in Simons that one summer. We truly oscillated between intense feelings of “it's so back” to “it's so over” within the span of a few days at the end.

Whenever I didn't feel like doing work, I was sure to find much entertainment at Soda Hall, especially when I was spending it with Ansh Nagda, Omar Alrabiah, and Nathan Ju. It was wonderful to stay accountable about reading papers and playing Hanabi and eating good food together. Also, it's impressive that Omar's CSP explanations remain on the board till this day. I bonded over countless balcony lunches with Ishaq Aden-Ali, and will always remember our hot takes (from Twitter or otherwise) and hanging out with the kids. I will look back fondly on the countless outings and hilarious conversations with Chris Ikeokwu and Ricardo Sandoval. Thanks to our hangouts I know more than I ever expected

to know about credit cards and random lore. To Sam Gunn, thanks for accompanying me to many concerts and adventures in Berkeley and beyond. Thanks to the lovely members of the theory group who I've become friends with, who welcomed me with open arms from the very beginning: Zoe, Thiago, Louis, Luke, Meghal, Malvika, Xin, Orr, Angelos, Vamsi, Bhaskar, Mihir, Fran, Nived, Fred, and Hongxun.

Despite not having a clue about what a qubit actually was, I always felt home at Simons. Thanks to Fermi Ma, Ewin Tang, Shivam Nadimpalli, Thiago Bergamaschi, Angelos Pelecanos, and many others, for all the fun and varied conversations about research and random topics, including but not limited to apartments and airplane incident reports. Fermi was a bottomless well of life advice and entertaining stories, and I really admire his intentional and thoughtful life philosophies.

I owe a lot to my roommates Erik Jenner and Shreyas Kapur. It's hard to imagine what the last four years would have been like without their company. Watching hour-long videos about how dishwashers work over Domino's was a strangely effective ointment to the woes of a tough research day. And I absorbed much knowledge about probabilistic programming and AI safety from our living room conversations, which would be hard to replicate elsewhere.

Thanks also to my friends in BAIR: Alex Pan, Chung Min Kim, Grace Luo, Jessica Dai, Kevin Black, and Shreyas Kapur. I remember how someone expressed doubt that we (or any similar group of first year friends) would continue hanging out after our first year, but I'm proud that we kept up with the weekly hangouts and yearly ski trips. Thanks to Grace, Jess, and Chung Min for hosting many fun parties with exciting fruits and non-alcoholic beverage alternatives. Thanks to Kevin for introducing us to the Sonic movie and skiing. Thanks to Alex, for our chronically overlong, rambley conversations about life, which have sharpened my view of the world.

The life of an academic is readily enjoyed when one finds an excuse to travel for work. Thanks to Spencer Compton, Aniket Das, Tselil Schramm, and Nima Anari for many stimulating and warm conversations at Stanford. Thanks to William He, Hoai-An Nguyen, Ting Hsieh, and Bingbin Liu for fond memories in Pittsburgh, and thanks especially to Noah Singer for being the best possible host and long time friend. Thanks to Yeshwanth Cherampanamjeri and Sidhanth Mohanty for innumerable hilarious conversations and cryptic Soviet movies in your slanted Cambridge apartment, and thanks to Aparna Gupte, Megan Le, Fathima Zarin Faizal, and Amit Rajaraman for letting me crash for so long at your apartment in Boston and crush you guys at Cabo.

Thanks to my research mentors from before grad school: David Palmer, Daryl DeFord, Justin Solomon, Chulhee Yun, Suvrit Sra, Ankur Moitra, Robert Burklund, and Lawrence Washington. Their encouragement and guidance greatly influenced my research trajectory.

Many thanks to my undergraduate friends: Daniel, Akshaj, Gongy, Michelle Shen, Kelly,

Pranav, Aniket, Christian, Forest, Alicia, Raj, Michelle Li, Will, and Peter. To Peter, our regular catch-up calls and visits to New York always felt like a breath of fresh air, and, I looked forward to these hangouts every time. Thanks also to my high school friends: Noah Singer, Elliot, Anish, Noah Kim, Arman, David, Ankitha, and Sambuddha. It was a joy to keep everyone updated on the progress of the PhD, and spending time together virtually or in person kept my spirits high.

Thanks to Akshara, whose love and support has made a world of difference.

And finally, thanks to my father Huiquan, mother Meiyu, and sister Angela. I couldn't be here without them.

# Contents

|                                                                             |           |
|-----------------------------------------------------------------------------|-----------|
| <b>Contents</b>                                                             | <b>v</b>  |
| <b>1 Introduction</b>                                                       | <b>1</b>  |
| 1.1 Sampling . . . . .                                                      | 2         |
| 1.2 Synthetic Data in Machine Learning . . . . .                            | 6         |
| <b>I Sampling</b>                                                           | <b>8</b>  |
| <b>Foreword</b>                                                             | <b>9</b>  |
| <b>2 Locally Stationary Distributions: MCMC Before Mixing</b>               | <b>10</b> |
| 2.1 Introduction . . . . .                                                  | 10        |
| 2.2 Preliminaries . . . . .                                                 | 24        |
| 2.3 Properties of locally stationary distributions . . . . .                | 28        |
| 2.4 Warmup: Large independent sets in triangle-free graphs . . . . .        | 32        |
| 2.5 Weak recovery in spiked models . . . . .                                | 35        |
| <b>3 Markov Chains Approximate Message Passing</b>                          | <b>49</b> |
| 3.1 Introduction . . . . .                                                  | 49        |
| 3.2 Technical Overview . . . . .                                            | 58        |
| 3.3 Conjectures and open questions . . . . .                                | 64        |
| 3.4 Preliminaries . . . . .                                                 | 66        |
| 3.5 The dynamics of RGD . . . . .                                           | 70        |
| 3.6 High-temperature mean magnetization estimates in the SK model . . . . . | 78        |
| 3.7 Towards understanding the RGD recursion . . . . .                       | 87        |
| <b>4 Weak Poincaré Inequalities: MCMC Beyond Worst-Case Mixing</b>          | <b>92</b> |
| 4.1 Introduction . . . . .                                                  | 92        |
| 4.2 Technical overview . . . . .                                            | 107       |
| 4.3 Preliminaries . . . . .                                                 | 110       |
| 4.4 Weak functional inequalities . . . . .                                  | 113       |

|           |                                                                             |            |
|-----------|-----------------------------------------------------------------------------|------------|
| 4.5       | Vignette: Sampling from mixture models with advice . . . . .                | 119        |
| 4.6       | Stopped localization schemes . . . . .                                      | 122        |
| 4.7       | Sampling from spherical $p$ -spin models . . . . .                          | 126        |
| 4.8       | High-probability covariance bound of replica symmetric spherical spin glass | 151        |
| 4.9       | Annealed Glauber dynamics on discrete domains . . . . .                     | 173        |
| 4.10      | Deferred calculations for spherical spin glasses . . . . .                  | 175        |
| <b>II</b> | <b>Synthetic Data in Machine Learning</b>                                   | <b>200</b> |
|           | <b>Foreword</b>                                                             | <b>201</b> |
| <b>5</b>  | <b>Weak-to-strong generalization via benign overfitting</b>                 | <b>202</b> |
| 5.1       | Introduction . . . . .                                                      | 202        |
| 5.2       | Preliminaries and setup . . . . .                                           | 205        |
| 5.3       | Main results . . . . .                                                      | 210        |
| 5.4       | Discussion . . . . .                                                        | 216        |
| 5.5       | Preliminaries . . . . .                                                     | 218        |
| 5.6       | Tools for analyzing survival and contamination . . . . .                    | 219        |
| 5.7       | Analyzing the subset ensemble . . . . .                                     | 230        |
| 5.8       | Multilabel classification . . . . .                                         | 233        |
| 5.9       | Improving the bounds for misclassification rate . . . . .                   | 237        |
| 5.10      | Experiments . . . . .                                                       | 242        |
| 5.11      | Heuristic calculations . . . . .                                            | 243        |
| <b>6</b>  | <b>Synthetic Error Injection for Self Correction in Language Models</b>     | <b>251</b> |
| 6.1       | Introduction . . . . .                                                      | 251        |
| 6.2       | Related work . . . . .                                                      | 253        |
| 6.3       | Experimental setup . . . . .                                                | 253        |
| 6.4       | Results . . . . .                                                           | 260        |
| 6.5       | Discussion . . . . .                                                        | 262        |
| 6.6       | Appendix: Training details . . . . .                                        | 265        |
| 6.7       | Appendix: CoT examples . . . . .                                            | 265        |
|           | <b>Bibliography</b>                                                         | <b>269</b> |

# Chapter 1

## Introduction

Modern science is powered by efficient sampling algorithms which infer models of nature through observations. Such sampling methods were instrumental in confirming the existence of gravitational waves [AAA<sup>+</sup>16], determining key cosmological constants [A<sup>+</sup>20], reconstructing phylogenetic trees [RH03], and much more. But why have these algorithms been so successful? Even after decades of intense study, we do not have theory which explains why these algorithms work so well in practice. My research aims to bridge this theory-practice gap in sampling.

My work focuses on the canonical family of sampling algorithms known as Markov chain Monte Carlo (MCMC). MCMC is typically analyzed by establishing *rapid mixing* to the stationary distribution from *worst-case* initializations. However, in many practical settings, rapid worst-case mixing completely fails to explain MCMC’s success. This thesis builds general tools and frameworks to surmount these challenges in MCMC theory. Namely:

- **MCMC Before Mixing.** In inference settings, MCMC is often plagued by *exponentially slow* worst-case mixing. To overcome this challenge, in Chapter 2 (adapted from [LMR<sup>+</sup>24]), we introduce the notion of a *locally stationary distribution*. These distributions can be reached in polynomial time by any reversible Markov chain, and are akin to a first-order stationary point in optimization.

Using this idea, we prove the first efficient guarantees for MCMC in slow-mixing inference and optimization on sparse graphs. In Chapter 3 (adapted from [RW25]), we prove that MCMC performs as well as message passing algorithms for the spiked Wigner inference problem [DKMZ11]. These ideas have subsequently been applied in areas as diverse as LLM reasoning algorithms [RSS<sup>+</sup>25] and quantum Gibbs sampling [BCV25].

- **MCMC Beyond Worst-Case Mixing.** Practitioners have a wide array of tricks to

speed up MCMC convergence, such as *simulated annealing* [KGJV83], which attempts to enable sampling in regimes where worst-case mixing provably fails.

In Chapter 4 (adapted from [HMRW25]), we propose a general framework to prove efficient guarantees for simulated annealing and other average-case initializations. Using this framework, we prove that despite slow worst-case mixing, simulated annealing mixes within a large portion of the high-temperature regime for the *spherical spin glass*. Our work nearly resolves a decades-old conjecture in statistical physics [CHS93].

Beyond inference, this thesis studies the data we use to train large language models (LLMs). LLMs generalize best when trained on expert human data, but such data can be expensive or downright impossible to obtain. This thesis aims to develop and understand new synthetic data approaches to overcome these challenges through a mix of theory and practice. In particular:

- **Weak to Strong Generalization.** Burns et al. [BIK<sup>+</sup>23] empirically find that strong student LLMs can learn from *systematically biased data* generated by weaker teacher LLMs, a phenomenon they term *weak-to-strong generalization*. In Chapter 5 (adapted from [WS25]), we give a theoretical explanation of weak-to-strong generalization based on the techniques of *benign overfitting*. Such understanding may provide a promising lever toward scalable methods for training future LLMs and ensuring their safety.
- **Self-Correction.** How can we teach LLMs to reason and explore new ideas in an efficient manner? To make progress on this question, in Chapter 6 (adapted from [WKS25]), we empirically explore an approach to directly teach reasoning LLMs to self-correct by sampling targeted synthetic training data. We find that such methods fail due to the crucial distribution shift that renders the data off-policy.

We now give a high level overview of these two parts of the thesis, as well as open questions.

## 1.1 Sampling

The classical approach to proving the success of MCMC is to establish rapid convergence, or mixing, to the stationary distribution from any initial distribution, i.e. *rapid worst-case mixing*. MCMC mixing theory has been extensively studied across several disciplines—including probability theory, geometry, theoretical computer science, and physics [HLW06, BGL<sup>+</sup>14, LPW17]. Recently, there has been an explosion of progress in proving worst-case mixing time bounds based on the powerful frameworks of *spectral independence* [ALOG21] and *localization schemes* [CE22]. However, even this new theory cannot hope to fully explain the



success of MCMC at inference, precisely because these posterior distributions suffer from slow worst-case mixing. In certain scenarios, successfully sampling from these distributions would even imply  $NP = RP$  [GŠV15]. This raises the following fundamental question:

*Can we develop a theory of MCMC which explains its empirical success without resorting to worst-case mixing theory?*

The study of metastability, i.e. what happens to the dynamics of MCMC before mixing, was initiated by Kramers [Kra40] over fifty years ago to answer this question; see also [FW98, BDH16] for later developments. It has since been used to explain the behavior of chemical reactions, ferromagnets, and spin glasses. Understanding the long-term behavior of MCMC before mixing therefore also sheds light on the physical world. However, characterizing metastability is often quite challenging, as it requires precise problem-specific control that remains out of reach for many problems of interest. Might there be an alternative notion of metastability which can be analyzed using softer techniques?

## Locally stationary distributions

To overcome the challenges posed by previous notions of metastability, in Chapter 2, we introduce the notion of a *locally stationary distribution*. Drawing inspiration from the theory of optimization, we define a locally stationary distribution as the sampling analog of a first-order stationary point in optimization. Indeed, under standard smoothness assumptions, gradient descent always reaches a first-order stationary point. We prove an analogous property for MCMC: any time-reversible Markov chain on a reasonable state space quickly reaches a locally stationary distribution.

Our framework can be leveraged to analyze MCMC for inference and optimization. Using it, we prove the success of MCMC in inference settings which have exponentially slow mixing times or complexity-theoretic barriers to sampling. In Chapter 2, we make nontrivial progress toward a long-standing physics prediction that MCMC achieves optimal performance on the *stochastic block model* [DKMZ11], a canonical inference model for community detection in large scale social networks. We show that MCMC nontrivially recovers the hidden communities in the stochastic block model, despite having an exponentially long mixing time. Building upon this framework, in Chapter 3, we characterize the optimal performance of MCMC in the related *spiked Wigner model*, a fundamental inference model for low-rank matrix recovery and PCA. We characterize MCMC in all parts of the phase diagram, by showing that MCMC approximates message passing algorithms such as *Approximate Message Passing* (AMP), which is known to have Bayes-optimal performance. Our techniques reveal a generic correspondence between MCMC on spiked matrix inference problems and an effective scalar *state evolution* which characterizes AMP.

## Average-case mixing and simulated annealing

Even if the goal is to efficiently sample from the target distribution, worst-case mixing need not reflect the complexity of the sampling. In many applications, one might expect MCMC to converge from a suitable warm start, even if it gets stuck from worst-case initializations. *Simulated annealing* is a ubiquitous meta-heuristic that constructs a sequence of distributions  $p_1, \dots, p_T$ , and attempts to sample from  $p_{i+1}$  using MCMC warm started from the previous output for  $p_i$ . Despite widespread use in scientific simulations and optimization, provable guarantees for efficient simulated annealing are few and far between. Moreover, previous approaches to the analysis of warm-start mixing tend to be ad hoc or difficult to apply to problems of interest.

In Chapter 4, we introduce a unified framework to bound the average-case mixing time from warm starts. Moreover, we prove that these warm starts can be achieved using simple simulated annealing or data-based initializations. To do so, we generalize previous frameworks based on stochastic localization [CE22] which bound *worst-case mixing times*, to establish certain relaxed functional inequalities which bound *average-case mixing times*.

We apply our framework to achieve state-of-the art guarantees for sampling from various distributions that are more representative of real-life data that is highly multimodal and complex. In Chapter 4, we give the first provable guarantees that simulated annealing efficiently samples from the *spherical  $p$ -spin glass*, in a regime which resolves (up to constant factors) the conjectured sampleable regime [CHS93]. Notably, we establish average-case mixing from warm starts in temperature regimes far beyond where worst-case mixing is provably slow due to metastability.

In real life, one sometimes also has access to data sampled from the distribution of interest. This paves the way for data-driven samplers such as data-based initializations for MCMC and denoising diffusion [HJA20], which is another form of simulated annealing. Indeed, in Chapter 4 we also prove that running simple MCMC algorithms using a data-based initialization succeeds at sampling highly multimodal mixture distributions, despite strong lower bounds against worst-case MCMC sampling in the same setting.

## Future Directions

On the inference side, Chapter 3 makes progress toward understanding the optimality of MCMC for inference by connecting MCMC to message passing for spiked matrix models using the notion of locally stationary distributions. Locally stationary distributions might be useful towards establishing a universal connection between MCMC and message passing for broader classes of inference problems:

*The success of MCMC for high-dimensional statistical inference can be reduced to the success of message passing algorithms.*

Another significant hurdle for closing the gap between MCMC practice and theory is that, despite significant advances in the toolkit for proving MCMC mixing bounds, the concrete theorems we can prove rely on rather technical problem-specific calculations. Furthermore, the Markov chains studied in theory do not fully capture the design choices that practitioners make; such as transition kernels and Markov chains that are learned from data. As a starting point, we can study these questions in a setting where predictions abound: statistical physics.

**A theory of MCMC for statistical physics models.** MCMC is often used as a way to simulate the time evolution of physical systems. One particularly salient class of physical models are the *spin glass models* from statistical physics, which were introduced to study the physical properties of disordered, glassy systems. Physicists have developed several nonrigorous methods for predicting the mixing behavior of MCMC. Generally speaking, physicists believe that a spin glass is tractable to sample from if and only if the temperature is above the *shattering* transition.

This prediction is a perfect testbed for the theory of average-case mixing: spin glasses often exhibit metastability above the shattering transition, so worst-case mixing cannot hope to answer these questions. In Chapter 4, we establish this within a constant factor of the shattering transition for the spherical spin glass. Perhaps ambitiously, one might aim to prove the following:

*Simulated annealing efficiently samples from all spin glasses above their shattering transition.*

Turning now to the question of generalizing the applicability of MCMC theory to practical Markov chains, one step in this direction is to connect standard simulated annealing MCMC to so-called *denoising diffusion* [HJA20] and flow-matching models. Denoising diffusion is the backbone of the modern image generation models such as Stable Diffusion and DALL-E. Notably, denoising diffusion can be viewed as a form of annealed Langevin dynamics with Gaussian annealing.

Moreover, for the spherical spin glass, the simulated annealing algorithm in Chapter 4 matches the exact temperature threshold achieved by diffusion-based algorithms [HMP24]. This is not a coincidence: at a high level, the smoothness guarantees required to prove that denoising diffusion succeeds are closely related to the covariance bounds required to prove rapid mixing. This leads to the following bold question about the optimality of simulated annealing as a sampling algorithm.

*Is temperature-based simulated annealing optimal for spin glasses and real-life distributions more broadly? Can we reduce the success of simulated annealing to that of denoising diffusion?*

There may be a deeper connection at play here, which may lead to further insights.

## 1.2 Synthetic Data in Machine Learning

Modern LLMs like ChatGPT are first pretrained on huge amounts of unlabeled data and then finetuned on human-labeled data [RWC<sup>+</sup>19]. In practice, obtaining high-quality labeled examples is often prohibitively expensive or impossible. An appealing alternative is to generate synthetic labeled data for finetuning, but an understanding of why and how such methods generally work is lacking.

Weak-to-strong (W2S) generalization [BIK<sup>+</sup>23] refers to the learning setting where a strong student model is finetuned on labels synthetically generated by a low-quality teacher model with biased predictions. They observed that W2S generalization occurs for a variety of learning tasks, but offer no explanation as to why. In Chapter 5, we offer a theoretical explanation for W2S generalization: it occurs when the strong model’s representations of the concept are strong enough to distill the weak signal from the bias in the labels. The proof builds on Wu and Sahai [WS24], which proved sharp transitions for the generalization of finetuning for multiclass classification.

In Chapter 6, we propose to generate synthetic data to teach LLMs to recognize and correct their own mistakes for reasoning tasks. We accomplish this by inserting synthetic errors in a chain of thought (CoT), along with an error recognition and correction step. We find that, although the models learn to correct against the synthetic errors, they struggle to reliably self-correct. We find that how closely the synthetic errors match the true *on-policy* errors essentially determines whether the model is able to recover.

### Future Directions

The current paradigm in eliciting strong LLM reasoning performance is to fine-tune models using reinforcement learning (RL). A central question in this area is whether such procedures genuinely discover new knowledge, or whether they simply amplify existing capabilities. One exciting recent work uses MCMC to sample from an annealed version of the LLM’s output distribution to simulate the effects of RL fine-tuning [KD25] without any actual training. I believe that marrying the field of sampling with LLM fine-tuning will lead to useful methods with theoretical grounding. For example, can we cast these reasoning tasks as mixing from a warm start provided by the base model? If so, can we

then design sampling algorithms that surpass the reasoning performance of RL fine-tuning with theoretical guarantees?

To push this idea further, one can look towards the *in-context learning* (ICL) capabilities of LLMs, where LLMs generalize from examples provided in the prompt. ICL can be thought of as an alternate form of fine-tuning, which leads to a natural question:

*Does our understanding of weak-to-strong generalization extend from fine-tuning to ICL?*

It would be interesting to give theoretical foundations for W2S training through ICL. More speculatively, if we model the base LLM itself as weak, and the current context as weak instructions, then perhaps we could use this to explain how reasoning can elicit genuinely new capabilities.

# **Part I**

## **Sampling**

# Foreword

The modern sampler wields a vast array of algorithms and tricks in their toolkit to apply to the sampling problem at hand. Chief among these methods is the canonical family of sampling algorithms known as Markov chain Monte Carlo (MCMC), and the ensuing chapters are dedicated towards understanding its widespread success. In MCMC, one fixes a target distribution  $\mu$  to sample from, and designs a Markov chain whose stationary distribution is precisely  $\mu$ . We will focus on MCMC's particular instantiations in *Glauber dynamics* [Gla63] and *Langevin dynamics* [Lan08], two canonical Markov chains that first arose in physics to study the so-called Ising model and Brownian motion, respectively.

In Chapter 2, we will introduce the framework of *locally stationary distributions*, which can explain the success of Glauber dynamics at various statistical inference problems despite Glauber's exponentially large worst-case mixing times. The capstone result in this chapter is proving the success of Glauber for the so-called *spiked Wigner* inference problem—a toy model of low-rank matrix recovery used as a testbed to study phase transitions for the last two decades [BAP05].

Building upon this result, in Chapter 3 we uncover a surprising connection between Glauber dynamics and Approximate Message Passing (AMP) which fully characterizes the recovery performance of Glauber. We conjecture that such a connection ought to hold more broadly as a potential mechanism that explains Glauber's widespread success in statistical inference.

Finally, in Chapter 4, we explore the sampling success of Langevin dynamics even when worst-case mixing fails. We introduce a framework for proving relaxed functional inequalities which guarantee the mixing success of Langevin (and Glauber) from various warm starts, produced by computationally feasible strategies such as simulated annealing. Using this, we prove warm-start mixing for the spherical spin glass [CHS93] within a constant factor of the conjectured temperature threshold.

# Chapter 2

## Locally Stationary Distributions: MCMC Before Mixing

### 2.1 Introduction

Markov chains are a fundamental algorithmic primitive that are widely applied towards sampling and counting tasks. There is a rich body of literature devoted to understanding worst-case *mixing times* of Markov chains, i.e., the number of steps required for the distribution of the chain to approach its stationary measure started from an arbitrary initialization. For some highlights in this area, see, e.g., the contents and references in [MT06, Dia09, BGJM11, BGL<sup>+</sup>14, LPW17].

Unfortunately, many natural Markov chains fail to mix rapidly from worst-case initializations, in that it takes a super-polynomial number of steps to reach stationarity. Structurally, this is due to the presence of cuts in the state space with very small conductance. Often, slow mixing is inevitable since sampling from the stationary measure is known to be computationally hard (say NP-hard); see, e.g., [Sly10, SS14, GV16]. Nevertheless, it has been empirically observed that certain simple and local Markov chains like *Glauber dynamics* succeed at optimization and inference tasks even when they are not known to mix, such as finding satisfying assignments to SAT formulas [SKC94, BIL<sup>+</sup>16], and clustering stochastic block models [MS12, GBP19]. This suggests that local Markov chains like Glauber dynamics can have algorithmic applications, even if they *fail* to mix rapidly, which raises our main line of inquiry.

**Question 2.1.1.** *What is the long-term behavior of Markov chains that do not mix rapidly? How can slow-mixing Markov chains be harnessed for optimization and inference?*

To lay out the motivation, it is useful to draw an analogy to continuous optimization.



Gradient descent, the canonical algorithm in optimization, converges efficiently to a global minimum if the objective function and parameter space is convex. However, non-convex objective functions and parameter spaces come up often both in theory and practice, and finding global minima can even be provably intractable. On the other hand, gradient descent can be shown to always converge quickly to a local minimum, or more precisely, a first-order stationary point. Moreover, these local minima are useful in practice, and can admit non-trivial theoretical guarantees. See [Nes18] for a comprehensive coverage of analyzing gradient-based optimization methods, and see, e.g., [GJZ17, BLLT20, JT20, JNG<sup>+</sup>21] and the references therein for non-trivial theoretical guarantees on local minima of gradient descent in the context of machine learning.

Analogously, in the context of sampling, certain random walks can be shown to mix rapidly to their stationary measures. However, other random walks can also be shown to mix slowly from worst-case initializations; these are akin to the “hard” nonconvex optimization problems. Despite this, one can show that *any* Markov chain satisfying fairly generic conditions converges to analogs of local minima that we term *locally stationary distributions* (see Definition 2.1.2). Intuitively, a locally stationary distribution corresponds to the stationary measure conditioned on a subset of states that are sparsely connected to the rest of the state space under the Markov chain.

This raises the question of whether these locally stationary measures obey theoretical guarantees that are useful for solving problems in optimization or inference.

## Locally stationary distributions

Let  $P$  be the transition matrix of a time-reversible Markov chain on a state space  $\Omega$ , where  $P[x, y]$  denotes the transition probability from  $x$  to  $y$ . Let  $\pi$  be a stationary distribution w.r.t.  $P$ .

Analogous to local minima in optimization, a locally stationary measure  $\nu$  is one started at which the Markov chain  $P$  remains nearly stationary, i.e., makes little progress. We will use the *KL divergence* to the stationary distribution, denoted  $\text{KL}(\nu \parallel \pi)$ , as a measure of progress. This leads to the following definition of an  $\varepsilon$ -locally stationary measure.

**Definition 2.1.2.** A probability measure  $\nu$  on  $\Omega$  with density  $f$  relative to  $\pi$  is said to be  $\varepsilon$ -locally stationary with respect to  $P$  if

$$\mathcal{E}(f, \log f) := \frac{1}{2} \sum_{x, y \in \Omega} \pi(x) P[x, y] \cdot (f(x) - f(y)) \cdot \log \frac{f(x)}{f(y)} \leq \varepsilon$$

The *Dirichlet form*  $\mathcal{E}(f, \log f)$  measures the rate at which the Markov chain progresses towards the stationary distribution. In particular, for a continuous-time version of the Markov chain  $P$ , if  $\nu_t$  denotes the measure at time  $t$ , we have the following well-known fact [BT06]:

$$\frac{d}{dt} \text{KL}(\nu_t \parallel \pi) = -\mathcal{E}(f_t, \log f_t).$$

As an immediate consequence, we observe that the Markov chain is typically on  $\varepsilon$ -locally stationary measures over time. Formally, we have the following claim.

**Theorem 2.1.3.** *Fix a time-reversible Markov chain  $P$  with a stationary measure  $\pi$ , any starting distribution  $\nu_0$ , and  $\varepsilon, \delta > 0$ . Let  $T = \frac{1}{\delta\varepsilon} \cdot \log\left(\frac{1}{\pi_{\min}}\right)$ . Then, for a time  $t \sim [0, T]$  chosen uniformly at random, the distribution  $\nu_t$  at time  $t$  is  $\varepsilon$ -locally stationary with respect to  $P$  with probability at least  $1 - \delta$ .*

The main conceptual contribution of our work is the following meta-principle for showing that locally stationary distributions solve optimization and inference problems.

*Prove that sampling from the true stationary distribution solves the optimization or inference problem of interest, and additionally, does so for “local” reasons.*

This principle is best illustrated by discussing our algorithmic applications of locally stationary distributions.

**Independent sets in triangle-free graphs.** It is easy to see that any graph  $G$  on  $n$  vertices with degree bounded by  $d$  has an independent set of size  $\frac{n}{d+1}$ , a bound which is tight for the union of  $(d+1)$ -sized cliques. Ajtai, Komlós, and Szemerédi [AKS80] showed that when  $G$  is triangle-free, the size of the maximum independent set guaranteed to exist increases to  $\Omega\left(n \cdot \frac{\log d}{d}\right)$ . Shearer [She83] gave an alternate proof which pins down the leading constant to  $1 - o_d(1)$  and relaxes the assumption of bounded maximum degree to bounded average degree.

It is also well-known that a uniformly random independent set in a triangle-free graph of maximum degree  $d$  has expected size at least  $\Omega\left(n \cdot \frac{\log d}{d}\right)$ ; see, e.g., [AS16, Proposition 1, Page 272]. Hence, it is natural to wonder whether *Glauber dynamics* with respect to the uniform measure over independent sets finds such a large independent set. From a given independent set  $I \subseteq V$ , the transitions of Glauber dynamics can be described as follows:

1. Sample a uniformly random vertex  $v \in V$ .
2. If  $I \cup \{v\}$  is an independent set, then go to  $I' = I \cup \{v\}$  with probability  $1/2$  and  $I' = I \setminus \{v\}$  with probability  $1/2$ .

3. If  $I \cup \{v\}$  is not an independent set, go to  $I' = I \setminus \{v\}$  with probability 1.

Notably, this Markov chain requires  $\exp(\Omega(n))$  steps to mix [MWW07] as soon as  $d \geq 6$ . In fact, the problem of sampling a uniformly random independent set on a graph of maximum degree  $d$  becomes NP-hard [Sly10, SS14] in this regime, even if triangle-freeness is assumed [GŠV15].

Despite these hardness results for the corresponding sampling problems, we show that the above Markov chain can be used to find independent sets of size  $\Omega\left(n \cdot \frac{\log d}{d}\right)$  in triangle-free graphs of maximum degree bounded by  $d$ . Specifically, we show the following result.

**Theorem 2.1.4.** *Let  $G$  be a triangle-free graph on  $n$  vertices with maximum degree bounded by  $d$ . Let  $I$  be an independent set in  $G$  that arises from Glauber dynamics run for  $O(nd^4)$  time. Then the expected size of  $I$  is at least  $\frac{1-o_d(1)}{4} \cdot n \cdot \frac{\log d}{d}$ .*

**Remark 2.1.5.** In fact, one can prove that Glauber dynamics at “fugacity”  $\frac{1}{\log d}$  finds an independent set of size  $(1 - o_d(1)) \cdot n \cdot \frac{\log d}{d}$  by combining our proof method with that of [D]PR18].

As mentioned before, we know that the expected size of a uniformly random independent set satisfies the above lower bound. However, the Glauber dynamics chain does not mix rapidly, and hence does not produce samples from the truly uniform distribution. Instead, it samples from a locally stationary distribution with respect to the Markov chain. Our key insight is that the same proof also goes through for an independent set sampled from a locally stationary distribution with respect to Glauber dynamics.

To give a sense of how local stationarity is used, we briefly discuss the proof. The proof from [AS16] that the expected size of a uniformly random independent set  $\Omega\left(n \cdot \frac{\log d}{d}\right)$  argues that for any vertex  $v$ , and for any pinning  $x_{v \cup N(v)}$  of the independent set outside  $v$  and its neighbors, either:

- the uniform distribution conditioned on the pinning chooses  $v$  with probability  $\gtrsim \frac{\log d}{d}$ , or
- it chooses  $\gtrsim \log d$  neighbors of  $v$  in expectation.

Thus, each vertex can be charged  $\Omega\left(\frac{\log d}{d}\right)$  vertices on average in the independent set. Observe that the above sketch of the argument goes through even if the distribution is not truly uniform but merely has conditional marginals matching the uniform distribution, which is a property we can show holds for locally stationary distributions (Lemma 2.3.5).

Using similar arguments, one can establish that given a triangle-free graph with maximum degree  $d$ , Glauber dynamics run for  $\text{poly}(n)$  many steps on the antiferromagnetic Ising model on  $G$  with inverse temperature  $\frac{1}{\sqrt{d}}$  recovers a cut of relative size  $\frac{1}{2} + \Omega\left(\frac{1}{\sqrt{d}}\right)$ .

**Weak recovery in spiked models.** Beyond independent sets, we also study the performance of Glauber dynamics for statistical inference tasks. Consider the central class of Bayesian models for principal component analysis (PCA) known as *spiked random matrix models*, which consist of a matrix  $M \in \mathbb{R}^{n \times n}$  given by

$$M = \underset{\substack{\uparrow \\ \text{signal} \\ \text{strength}}}{\lambda} \cdot \underset{\substack{\uparrow \\ \text{signal}}}{v} \underset{\substack{\uparrow \\ \text{noise}}}{v}^\top + W.$$

The general algorithmic question is to approximately recover the signal (a unit vector  $v$ ) under appropriate assumptions about the noise ( $W$ ) and the signal strength ( $\lambda$ ). More precisely:

**Problem 2.1.6** (Weak recovery in spiked matrix models). For a unit norm signal vector  $v \in \mathbb{R}^n$ , signal strength  $\lambda \in \mathbb{R}$ , and noise matrix  $W \in \mathbb{R}^{n \times n}$ , given  $M = \lambda \cdot vv^\top + W$ , give an efficient algorithm to extract a unit norm estimate  $\hat{v}$  such that  $\langle v, \hat{v} \rangle \geq \Omega(1)$ .

In the situation where  $W$  is a Wigner matrix, this model, known as the *spiked Wigner model*, has been a subject of extensive study. The work of [BAP05] determined that once  $\lambda > 1$ , a spectral algorithm based on computing the top eigenvector succeeds at weak recovery. There is a fairly large body of work on the spiked Wigner model, towards characterizing optimal estimation error, efficient algorithms and its generalizations to rank larger than one [DMK<sup>+</sup>16, DAM17, EAK18, LM17, Mio17, BM19, MV21]. When the prior distribution over  $v$  is the uniform distribution over  $\left\{\pm \frac{1}{\sqrt{n}}\right\}^n$  and  $\mathbb{S}^{n-1}$ , there are efficient algorithms that even achieve the maximum information-theoretically achievable correlation  $|\langle v, \hat{v} \rangle|$ , based on approximate message passing [FMM21, CFM23], and algorithmic stochastic localization [MW23].

In the case where the prior distribution is on the hypercube, given the matrix  $M$ , this posterior is described by an Ising model: a probability distribution  $\mu_{\beta M}$  over  $\{\pm 1\}^n$  defined by the following proportionality relation for a suitably chosen  $\beta > 0$ :

$$\mu_{\beta M}(x) \propto \exp\left(\frac{1}{2} \langle x, \beta M x \rangle\right) \text{ for all } x \in \{\pm 1\}^n.$$

Sampling from the above posterior distribution is desirable as it achieves the maximum information-theoretically achievable correlation. The canonical algorithm for sampling is

to run the Glauber dynamics Markov chain, but unfortunately, provable guarantees for Glauber dynamics are currently lacking. Thus, a natural question en route is: *does Glauber dynamics for the Ising model  $\mu_{\beta M}$  weakly recover the signal in polynomial-time?*

We make progress towards answering this question affirmatively in this work by showing that Glauber dynamics at a slightly higher temperature than the posterior distribution succeeds for a broad family of settings.

Formally, we show the following result:

**Theorem 2.1.7.** *Let  $W$  be a matrix with  $\kappa \leq W \leq 1 - \kappa$ , and  $v \in \left\{\pm \frac{1}{\sqrt{n}}\right\}^n$ . Let  $P$  denote the kernel of the Glauber dynamics chain with stationary distribution  $\mu_{W+\lambda vv^\top}$ , and  $x_0$  an arbitrary point on  $\{\pm 1\}^n$ . There exists a large enough constant  $\lambda > 0$  such that for  $T = \widetilde{\Theta}(n^5)$ , and for  $t \sim [0, T]$ , with probability  $1 - o(1)$  we have:*

$$\mathbf{E}_{x \sim P^t \delta_{x_0}}[|\langle x, v \rangle|] \geq \left( \kappa \exp\left(-\frac{1}{\kappa}\right) - o(1) \right) \cdot \sqrt{n}.$$

A natural approach to recover the signal  $v$  from the matrix  $M$  is the spectral method, which amounts to computing (even approximately) the eigenvector corresponding to the largest eigenvalue for the matrix  $M$ . At a high-level, the above theorem demonstrates that Glauber dynamics can *simulate* the spectral method in certain regimes. We further expect Glauber to achieve weak recovery when run for  $T = n^{1+o(1)}$  steps, but we leave this open as a direction for future improvement.

**Remark 2.1.8.** The above model of choice captures several commonly considered models of study in the algorithms and complexity of statistical inference, such as the *spiked Wigner model* [BAP05], and *random/planted 2XOR* (see, e.g., [AOW15] and the references within).

**Remark 2.1.9.** In the Rademacher spiked Wigner model, where  $W \sim \text{GOE}(n)$  and  $v \sim \left\{\pm \frac{1}{\sqrt{n}}\right\}^n$ , the posterior has the form

$$\Pr[v|M] \propto \exp\left(-\frac{n}{2} \|M - \lambda vv^\top\|_F^2\right) \propto \exp\left(\frac{\lambda n}{2} \cdot v^\top M v\right).$$

The above is an Ising model, and suggests Glauber dynamics as a natural algorithm for weak recover. We note that the Ising model we shall analyze will be a higher temperature version of the above, that is, a distribution with density proportional to  $\exp\left(\frac{\beta n}{2} \cdot v^\top M v\right)$  for some  $\beta < \lambda$  (as opposed to the “correct” value  $\lambda$ ). Interestingly, such recovery guarantees were not previously known, even if one *information theoretically* samples from the higher temperature Ising model.

**Stochastic block model.** Another case of interest is one where the Ising model  $M$  arises from a stochastic block model. To describe this result, we first begin by defining the two-community stochastic block model.

**Definition 2.1.10** (2-community stochastic block model). Let  $d, \lambda \in \mathbb{R}$  be fixed parameters such that  $\lambda^2 \leq d$ . The distribution  $\text{SBM}(n, d, \lambda)$  is defined over pairs  $(\sigma, G) \in \{\pm 1\}^n \times \{0, 1\}^{n \times n}$  generated as follows.

Let  $\sigma \in \{\pm 1\}^n$  be a *signal* vector drawn uniformly at random (i.e. the prior is uniform). Given  $\sigma$ , we draw a random graph  $G$  by including an edge between  $u, v \in [n]$  independently with probability  $\frac{d+\lambda\sqrt{d}}{n}$  if  $\sigma(u) = \sigma(v)$ , and with probability  $\frac{d-\lambda\sqrt{d}}{n}$  otherwise.

In a general stochastic block model, the signal vector  $\sigma$  can be over a larger finite alphabet  $[q]$ , and the probability of including an edge between  $u, v \in [n]$  is an arbitrary function of  $\sigma(u), \sigma(v)$ . In this work, we will use the term stochastic block model (SBM) to refer exclusively to the special case of two communities as defined above.

**Remark 2.1.11.** The 2-community stochastic block model can be viewed as a special case of a spiked matrix model where  $M$  is a highly sparse matrix. Due to the sparsity of  $M$ , this spiked matrix model falls outside the scope of Theorem 2.1.7, as the “noise” part fails to satisfy the spectral bound.

The weak recovery problem for stochastic block model is that of recovering a labelling  $\hat{\sigma}$  given the graph  $G$  such that  $\hat{\sigma}$  has non-trivial correlation with the true signal  $\sigma$ . More precisely, an algorithm for weak recovery is required to find a  $\hat{\sigma}$  such that  $\frac{1}{n}|\langle \hat{\sigma}, \sigma \rangle| \geq \Omega(1)$ .

Starting with the work of Decelle, Krzakala, Moore, & Zdeborova [DKMZ11] that posited broad conjectures about these models, an extensive body of work has emerged over the past decade. For the case of 2 communities, [DKMZ11] posited that weak-recovery is possible if and only if the signal strength  $\lambda^2 > 1$ . This coincides with the *Kesten–Stigum threshold*, a threshold for broadcast processes on trees studied in the works of Kesten and Stigum [KS66, KS67]. The works of Mossel, Neeman, & Sly [MNS18] and Massoulié [Mas14] confirmed the algorithmic side, namely that weak recovery can be solved efficiently above the KS threshold with a spectral algorithm, while [MNS15] showed impossibility below the threshold. We refer the reader to the survey of Abbe [Abb17] for a detailed treatment of the literature on community detection.

We show that Glauber dynamics succeeds at weak recovery when the signal strength is a constant factor above the Kesten–Stigum threshold.

**Theorem 2.1.12.** *There exist constants  $\lambda_0, \beta, c > 0$  such that for all  $\lambda$  satisfying  $|\lambda| \geq \lambda_0$ , for  $(\sigma, G) \sim \text{SBM}(n, d, \lambda)$ , with probability  $1 - o(1)$  over the randomness of  $(\sigma, G)$ , the following holds.*

Let  $P$  denote the kernel of Glauber dynamics with stationary distribution  $\mu_{\frac{\beta}{\sqrt{d}}}(A_G - \frac{d}{n}\mathbf{1}\mathbf{1}^\top)$  and  $x_0$  an arbitrary point on  $\{\pm 1\}^n$ . For  $T = \tilde{\Theta}(n^{5+o_d(1)})$ , and for  $t \sim [0, T]$ , with probability  $1 - o(1)$ , we have:

$$\mathbf{E}_{x \sim P^t \delta_{x_0}}[|\langle x, \sigma \rangle|] > cn.$$

**Remark 2.1.13.** Similar to Remark 2.1.9, the posterior distribution  $\sigma|G$  to solve the recovery problem in the stochastic block model is an Ising model (see, e.g., [Moo17, Eq. (7)]):

$$\Pr[\sigma|G] \propto \prod_{ij \in E(G)} \left( \frac{d + \lambda\sqrt{d}}{d - \lambda\sqrt{d}} \right)^{(1+\sigma_i\sigma_j)/2} \cdot \prod_{ij \notin E(G)} \left( \frac{1 - \frac{d+\lambda\sqrt{d}}{n}}{1 - \frac{d-\lambda\sqrt{d}}{n}} \right)^{(1+\sigma_i\sigma_j)/2}$$

For large  $d$ , the Ising model that we analyze is approximately equal to a higher temperature version of the above true posterior.

Although spectral algorithms for weak recovery were already known in all the cases listed above, understanding the power of Glauber dynamics is interesting in its own right. It is arguably a more natural algorithm than spectral methods in the context of a Bayesian estimation problem like stochastic block models. In particular, Glauber dynamics remains locally consistent with the underlying probabilistic model at every vertex. On the other hand, a spectral algorithm that computes the top eigenvector maximizes a global objective, while crudely approximating the local features of the probabilistic model.

Finally, our analysis for spiked models establishes a direct correspondence between locally stationary measures for Glauber dynamics and fixed points of a Markov chain over the one-dimensional real line  $\mathbb{R}$  (related to the *restricted Gaussian dynamics* Markov chain). This correspondence may pave the way for a much tighter analysis to establish that Glauber dynamics achieves information theoretically optimal recovery in some of these models. To elucidate further on this correspondence, we will give a brief technical overview here.

## Technical overview

In Section 2.3, we derive a few basic properties of locally-stationary measures. This is followed by the result on independent sets presented in Section 2.4 as a warmup.

In this technical overview, we will focus on the inference problem in spiked matrix models. For a matrix  $M \in \mathbb{R}^{n \times n}$  and a vector  $h \in \mathbb{R}^n$ , we will use  $\mu_{M,h}$  to denote the distribution over  $\{\pm 1\}^n$  defined as

$$\mu_{M,h}(x) \propto \exp\left(\frac{1}{2}\langle x, Mx \rangle + \langle h, x \rangle\right),$$

and  $\mu_M$  to denote the distribution  $\mu_{M,0}$ .

Consider the stationary measure  $\mu_M$  for a spiked matrix  $M = \lambda v v^\top + W$ . We outline the proofs of Theorems 2.1.7 and 2.1.12 here, which consist of two parts.

- First, we show that locally stationary distributions with respect to Glauber dynamics over  $\{\pm 1\}^n$  are also locally stationary with respect to the *restricted Gaussian dynamics* Markov chain.
- Next, we show that samples from locally stationary distributions for RGD achieve weak recovery.

Let us first recall the definitions of Glauber dynamics and Restricted Gaussian dynamics for Ising models.

**Definition 2.1.14** (Glauber dynamics). Glauber dynamics with respect to a distribution  $\pi$  over  $\{\pm 1\}^n$  is a Markov chain on  $\{\pm 1\}^n$ , where a transition from  $x$  is given by the following:

- Sample index  $i$  uniformly from  $[n]$ .
- Transition to  $x^{\oplus i}$  with probability  $\frac{\pi(x^{\oplus i})}{\pi(x) + \pi(x^{\oplus i})}$ , and stay at  $x$  otherwise. Here,  $x^{\oplus i}$  denotes  $x$  with the  $i$ th bit flipped.

The following definition follows the restricted Gaussian dynamics framework of [LST21, STL20, CE22].

**Definition 2.1.15** (Restricted Gaussian dynamics). Consider the joint random variable  $(x, z)$  where  $x \sim \mu_M$ , and  $z|x := (\lambda \langle v, x \rangle + \sqrt{\lambda} g) \cdot v$  for  $g \sim \mathcal{N}(0, 1)$ . *Restricted Gaussian dynamics* (RGD) is a Markov chain on  $\{\pm 1\}^n$  where for any  $x$ , the transition to  $x'$  is sampled as follows:

- Sample  $z|x$ .
- Sample  $x'|z$ .

**Remark 2.1.16.** By definition, the above Markov chains are ergodic and reversible with respect to  $\pi$  and  $\mu_M$  respectively, and so asymptotically converge to them as their stationary distributions.

**Remark 2.1.17.** We should think of  $z|x$  as being a noisy surrogate for how well  $x$  correlates with the hidden direction  $v$ .



Informally, we prove the following correspondence between locally stationary distributions for Glauber dynamics and locally stationary distributions for RGD. In fact, this correspondence is a consequence of a more generic statement; refer to Lemma 2.3.7 for details.

**Lemma 2.1.18** (Informal version of Lemma 2.3.7). *Let  $\nu$  be a distribution over  $\{\pm 1\}^n$  that is  $\varepsilon$ -locally stationary under Glauber dynamics for  $\mu_M$ . Suppose for every  $z \in \mathbb{R}$ , Glauber dynamics for the distribution of  $x|z$  is “well-expanding”, and  $\log \frac{1}{\mu_M(x)} \leq \text{poly}(n)$  for all  $x \in \{\pm 1\}^n$ , then  $\nu$  is  $\varepsilon \cdot \text{poly}(n)$ -locally stationary under restricted Gaussian dynamics.*

To conclude that  $\nu$  is locally stationary under restricted Gaussian dynamics, it suffices to verify the structural properties of  $\mu_M$ . The lower bound on the minimum probability follows from upper and lower bounds on the values that the Hamiltonian can achieve. To show that Glauber dynamics for  $x|z$  is well-expanding, we must investigate the structure of this distribution further. A simple calculation reveals that the distribution of  $x|z$  is, in fact, the Ising model  $\mu_{W,z}$ . In the setting of Theorem 2.1.7, where the spectral diameter of  $W$  is bounded by 1, prior works [BB19, EKZ22, AJK<sup>+</sup>21a, CE22] prove that  $\mu_{W,z}$  always satisfies a “modified log-Sobolev inequality” (our relevant notion of “well-expanding”). In the setting of Theorem 2.1.12, where  $W$  is a centered stochastic block model, a similar result is proved in a companion paper [LMRW24].

**Remark 2.1.19.** This decomposition of  $\mu_M$  into a mixture of other Ising models is well-known in the literature by the name *Hubbard–Stratonovich transformation* [Hub59]; see also [KLR22].

In summary, we showed if  $\nu$  is locally stationary with respect to Glauber dynamics, then it is also locally stationary for the RGD chain

$$x \rightarrow z|x \rightarrow x'|z.$$

Thus, it suffices to prove that  $\mathbf{E}_{x \sim \nu} |\langle x, v \rangle|$  is bounded away from 0 for any distribution  $\nu$  that is locally stationary for RGD. The two ingredients that go into proving this are:

- A generic principle that says: if  $\nu$  is locally stationary for a Markov chain  $P$ , then for any bounded function  $f$ ,  $|\mathbf{E}_{x \sim \nu} f(x) - \mathbf{E}_{x \sim P_\nu} f(x)|$  is small (Corollary 2.3.3).
- If the correlation of a distribution  $\nu$  is too close to 0, then a single step of RGD causes a significant boost in correlation, which means  $\nu$  cannot be locally stationary. In particular, for  $f(x) = |\langle x, v \rangle|$ , if  $\mathbf{E}_{x \sim \nu} f(x)$  is too close to 0, then  $\mathbf{E}_{x \sim P_\nu} f(x) - \mathbf{E}_{x \sim \nu} f(x)$  is nontrivially large, which means any locally stationary distribution  $\nu$  must achieve large correlation.

See Section 2.5 for the details of this argument.

## Related work

Motivated by statistical physics, the phenomenon of *metastability* of random walks has been extensively studied. We refer the reader to the monograph by Bovier & Hollander [BDH16] for related literature. The notion of metastability in [BDH16] appears to be a slightly stricter notion than local stability, and thus does not generically hold for all reversible Markov chains.

In the context of sampling distributions over a continuous domain, Balasubramanian, Chewi, Erdogdu, Salim & Zhang [BCE<sup>+</sup>22] showed that the Langevin Monte Carlo algorithm always outputs a sample from a distribution whose relative Fisher information is small. This is the continuous sampling analog of convergence of gradient descent to approximate first-order stationary points. Building on these ideas, Cheng, Wang, Zhang & Zhu [CWZZ24] study the notion of *conditional mixing* for Langevin and Glauber dynamics and apply it to efficiently sample from Gaussian mixtures.

Our analysis of Glauber dynamics borrows ideas from a recent line of works on sampling from Ising models. Glauber dynamics for an Ising model defined by a matrix  $M$  was shown to mix quickly if eigenvalues of  $M$  lie within an interval of length 1 [AJK<sup>+</sup>21a, EKZ22]. This is sharp, as evidenced by the Curie–Weiss model  $M = \frac{\beta}{n} \mathbf{1}\mathbf{1}^\top$ . Stronger evidence for hardness of sampling beyond this spectral criterion was recently provided by Kunisky [Kun23], based on a reduction to a certain statistical hypothesis testing problem. Koehler, Lee & Risteski [KLR22] devised more sophisticated algorithms based on simulated tempering and variational inference to sample from Ising models when they have constantly many eigenvalues outside an interval of length 1.

Besides the question of fast mixing and metastability, the problem of how well MCMC-based algorithms perform for optimization and inference tasks was recently studied in several works. Chen, Mossel & Zadik [CMZ23] proved that when initialized at the empty set, natural Metropolis chains on cliques fail to find cliques of sublinear size in polynomial time, even if such a clique is planted inside the Erdős–Rényi random graph  $G(n, 1/2)$ . This is despite there being an abundance of algorithms which can recover a planted clique of size down to  $O(\sqrt{n})$ . Nevertheless, MCMC-based algorithms were redeemed in a more recent work of Gheissari, Jagannath & Xu [GJX23] using a more carefully designed chain and initialization. In a recent work, Sellke [Sel24b] proved that low-temperature Langevin dynamics achieves the conjectured computational threshold for optimizing pure spherical spin glass models.

## Open problems

We conclude with several open directions, which we believe may be amenable to the framework of locally stationary distributions.

**Bayesian inference via MCMC.** First, there is the direction of pushing our results further in the setting of SBM. To set the scene, let  $\pi(x) \propto \exp(H(x))$  be the true posterior for SBM, where  $H(x)$  is the SBM Hamiltonian (see Remark 2.1.13 for an explicit formula). It is well known that optimal recovery is achieved information theoretically by sampling from  $\pi$  (see, e.g., [Moo17, Section 4]). However, it takes exponential time to mix to  $\pi$  from a worst-case initialization. On the other hand, in Theorem 2.1.12, we achieve weak recovery by running Glauber dynamics on the density  $\pi_\beta(x) \propto \exp(\beta H(x))$  for some (constant)  $\beta$  strictly smaller than 1. It is natural to investigate whether a sampling algorithm based on *simulated annealing*, i.e. running Glauber dynamics by varying the temperature over time, can succeed at sampling from  $\pi$ . Our main result can be viewed as a modest step in this direction, as we show that running the chain for  $\text{poly}(n)$  steps at a mismatched temperature already gives a warm start for correlation.

**Problem 2.1.20** (Optimal recovery for stochastic block model). Can an instance of simulated annealing sample from  $\pi$ ?

**Computationally optimal inference.** The  $k$ -community stochastic block model is known to undergo an *information-computation gap* when  $k \geq 5$  (see, e.g., [AS16]). Specifically, for every  $k \geq 5$ , there exists a choice of degree  $d$  and signal-to-noise ratio  $\lambda$  for which weak recovery is information-theoretically possible, but likely impossible for efficient algorithms [HS17]. This gap admits the construction of SBM instances where weak recovery is tractable, but information-theoretically optimal recovery is intractable to efficient algorithms.

**Example 2.1.21.** Consider a 10-community block model obtained by taking two disjoint 5-community block model graphs, and planting a sparse bisection between them. The planted bisection should be sparse enough so it is clearly detectable to efficient algorithms. However, the parameters  $d$  and  $\lambda$  for the 5-community models are chosen to be in the intractable regime. An information-theoretically optimal algorithm achieves weak recovery within each 5-community model. At the same time, algorithms for the 2-community block model achieve weak recovery in the 10-community model, since they can find the planted bisection, and correctly classify vertices as belonging to either  $\{1, 2, 3, 4, 5\}$  or  $\{6, 7, 8, 9, 10\}$ .

In settings such as the above, information-theoretically optimal inference is hard, but weak recovery is still tractable. This motivates the study of *computationally optimal inference*

*algorithms*: algorithms that achieve the best guarantees possible in polynomial time. For such problems, the Glauber dynamics chain must necessarily fail to mix rapidly to the posterior distribution, but perhaps the locally stationary distribution it samples from can achieve the computationally optimal recovery guarantees?

**Problem 2.1.22** (Computationally optimal inference). Is (annealed) Glauber dynamics a computationally optimal algorithm for the  $k$ -community SBM and, more generally, for random CSPs with planted solutions?

**Metastable states.** Local stationarity is a generic property of any time-reversible Markov chain, so a priori there is no reason to expect that a locally stationary distribution  $\nu$  has any nice properties. For example, if we run the Markov chain for  $T$  steps, Theorem 2.1.3 guarantees an  $\varepsilon$ -locally stationary distribution  $\nu$  where  $\varepsilon = O(1/T)$ , and the simple random walk on the  $n$ -vertex cycle graph demonstrates that this is tight if  $T = o(n^2)$ . This suggests the following natural questions. Suppose the stationary distribution  $\pi$  is a Gibbs distribution on  $\{\pm 1\}^n$ . Under what additional structural assumptions on  $\pi$  can we both obtain  $\varepsilon = o(1/T)$  for sufficiently large  $T = \text{poly}(n)$  and endow  $\nu$  with a physical or geometric interpretation?

The notion of metastable states for Gibbs distributions in statistical physics [BDH16] suggests a conceptual path forward towards this goal. In particular, one might hope to show that a locally stationary distribution is close to a metastable state, i.e. a conditional Gibbs distribution restricted to a metastable set of configurations.

**Problem 2.1.23** (Metastability). Suppose  $\pi$  is a Gibbs distribution which has a metastable subset  $S$  with exponentially small conductance. Let  $\nu$  be the locally-stationary distribution after running Glauber (or Langevin) dynamics for  $\text{poly}(n)$  steps with uniform initialization in  $S$ . Is  $\nu$  close to the conditional Gibbs distribution  $\pi_S$ , e.g.,  $\text{KL}(\nu \parallel \pi_S) = o(1)$ ?

For a concrete setting, suppose  $\pi$  is a spherical spin glass in the *shattering* regime [EAMS23]. Can one show that in  $\text{poly}(n)$  time, Langevin dynamics with uniform initialization remains stuck in the clusters identified there?

**Cavity method.** The *cavity method*, and the related *replica method*, originated in physics to predict the properties of various Gibbs distributions. Some striking achievements of this heuristic in producing accurate predictions are the Parisi formula [Par80, Tal06], and the  $k$ -SAT satisfiability threshold [MPZ02, MMZ06, DSS15]. It was also employed in the work of Decelle, Krzakala, Moore, & Zdeborová [DKMZ11] to conjecture the Kesten–Stigum threshold as the computational threshold for SBM.

Of particular interest to us are the works of Coja-Oghlan, Krzakala, Perkins & Zdeborová [COKPZ17, COP19a, COP19b], that characterize the recovery rate that the optimal

estimator, namely sampling from the Gibbs distribution, achieves for various planted constraint satisfaction problems. Their proofs use fairly minimal properties of the Gibbs distribution. More concretely, for a graph  $G$  and an assignment  $x$ , let  $H_G(x)$  be a Hamiltonian, and let  $\pi_G$  be the corresponding Gibbs distribution. Their proofs rely on the following properties satisfied by Gibbs distributions.

- (Gibbs ratios) For any graph  $G$  and any vertex  $v$ :

$$\frac{\pi_G(x)}{\pi_{G \setminus v}(x)} \propto \exp(H_G(x) - H_{G \setminus v}(x)),$$

- (Approximate independence) For random  $G$  and  $v$ , the marginals of the neighbors of  $v$  are approximately independent in  $\pi_{G \setminus v}$ .

If one can show that a family of locally stationary distributions also satisfy the Gibbs ratios up to a multiplicative error and approximate independence on a random graph  $G$ , then one could hope to port over the cavity method predictions and their rigorous proofs in a black-box fashion.

**Problem 2.1.24** (Cavity method). Let  $\{\nu_G\}_{G \in \text{graphs}}$  be a family of locally stationary distributions where  $\nu_G$  arises from running the Glauber dynamics for  $\pi_G$  for time- $T$  initialized at the uniform distribution. For a random graph  $G$  and random vertex  $v$ , do  $\nu_G$  and  $\nu_{G \setminus v}$  satisfy, up to small error, the Gibbs ratios and approximate independence properties?

**Beyond average-case models.** Our work proves that Glauber dynamics recovers planted spikes when the input matrix has a clean “signal + noise” structure. Recently, there has been a flurry of work on inference in *semirandom models*; see, e.g., [BKS23, GHKM23] and the references within, where it is possible to extract the hidden signal using semidefinite programming-based algorithms. A natural direction is to investigate whether Glauber dynamics solves semirandom inference problems.

**Problem 2.1.25** (Semirandom models). Does Glauber dynamics succeed at finding solutions to semirandom planted CSPs, or large cliques in semirandom graphs with planted cliques as is done in the works of [GHKM23] and [BKS23] respectively?

In a similar vein, semidefinite programming has been phenomenally successful at solving dense CSPs, and more generally CSPs on graphs with low threshold-rank [BRS11]. A reason to believe that local algorithms perform well is Theorem 2.1.7, where we show that Glauber dynamics can recover rank-1 spikes in threshold-rank-1 matrices.

**Problem 2.1.26** (CSPs on low threshold-rank graphs). Does running Glauber dynamics give a PTAS for Max Cut on a dense graph?

## 2.2 Preliminaries

We begin by setting up some notation.

- Let  $P$  be the transition matrix of a time-reversible Markov chain on state space  $\Omega$  with stationary distribution  $\pi$ , where  $P[i, j]$  denotes the transition probability from  $i$  to  $j$ . Let  $P_t = \exp(-t(I - P))$  denote the time- $t$  transition kernel.
- For a distribution  $\nu$  absolutely continuous with respect to  $\pi$ , we use  $f(x) := \frac{d\nu}{d\pi}(x)$  to refer to its relative density to  $\pi$ .
- We use  $\nu_t$  to denote  $P_t\nu$ , and we assume that  $\nu_t$  is absolutely continuous with respect to  $\pi$  throughout. In particular, we write  $f_t(x) := \frac{d\nu_t}{d\pi}(x)$  to denote its relative density to  $\pi$ ;
- We use  $\mathbf{m}(\nu) := \mathbf{E}_{x \sim \nu} x$  to denote the mean of  $\nu$ .
- We use  $\mathbf{y} \sim_P x$  when  $\mathbf{y}$  is chosen as a random neighbor of  $x$  according to transition probabilities given by  $P$ . We drop the subscript  $P$  from the  $\sim$  when the Markov chain is clear from context.

**Remark 2.2.1.** The way to think of the time- $t$  transition kernel  $P_t$  for a Markov chain with kernel  $P$  on a discrete space is via the process: sample  $t \sim \text{Poisson}(t)$  and take  $t$  steps using  $P$ .

We will require the following simple consequence of the definition of total variation distance.

**Fact 2.2.2.** For any pair of distributions  $\nu$  and  $\pi$  on  $\Omega$ , and any function  $f : \Omega \rightarrow \mathbb{R}$ , we have

$$|\mathbf{E}_\nu f - \mathbf{E}_\pi f| \leq (f_{\max} - f_{\min}) \cdot d_{\text{TV}}(\nu, \pi).$$

**Definition 2.2.3** (Dirichlet form). For functions  $f, g : \Omega \rightarrow \mathbb{R}$ , and  $x, y \in \Omega$ , the *Dirichlet form* of  $f$  and  $g$  with respect to  $P$  is:

$$\mathcal{E}_P(f, g) := \frac{1}{2} \mathbf{E}_{x \sim \pi} \mathbf{E}_{y \sim_P x} (f(x) - f(y)) \cdot (g(x) - g(y)).$$

We drop the  $P$  in the subscript when it is clear from context.

**Remark 2.2.4.** When we use the Glauber dynamics chain for a distribution  $\pi$  on a hypercube, we use  $\mathcal{E}_\pi$  to denote the corresponding Dirichlet form.

The Dirichlet form measures the rate at which a Markov chain makes progress towards the stationary distribution. The following is one way of articulating this notion; see, e.g., [BT06].

**Fact 2.2.5.**  $\frac{d}{dt} \text{KL}(\nu_t \parallel \pi) = -\mathcal{E}(f_t, \log f_t) = -\frac{1}{2} \mathbf{E}_{x \sim \pi} \mathbf{E}_{y \sim x} (f_t(x) - f_t(y)) \cdot \log \frac{f_t(x)}{f_t(y)}.$

**Definition 2.2.6** (Modified log-Sobolev inequality). We say  $P$  satisfies a *modified log-Sobolev inequality* (MLSI) with constant  $C$  if for any function  $f : \Omega \rightarrow \mathbb{R}_{>0}$ ,

$$\mathcal{E}_P(f, \log f) \geq C \cdot \text{Ent}[f].$$

Here,  $\text{Ent}[f] := \mathbf{E}_\pi[f \log f] - \mathbf{E}_\pi f \log \mathbf{E}_\pi f$  is the entropy functional. In particular,  $\rho_{\text{MLSI}}(P)$  is the best (largest) such constant  $C$ .

We will need the following fact concerning the MLSI for product measures; see, e.g., [Goe04, Lemma 2.5].

**Fact 2.2.7** (MLSI for product measures). *Let  $\pi$  be a distribution over  $\{\pm 1\}^n$  with independent coordinates. Then  $\rho_{\text{MLSI}}(\pi) \geq 1/n$ .*

Finally, we need the following lemma showing that an MLSI implies concentration of Lipschitz functions, in particular linear forms [Goe04, Theorem 5.1].

**Lemma 2.2.8** (MLSI concentration). *Let  $\mu$  be an arbitrary distribution over  $\{\pm 1\}^n$  such that  $\rho_{\text{MLSI}}(\mu) \geq \alpha$ . Then, for any function  $f : \{\pm 1\}^n \rightarrow \mathbb{R}$  that is 1-Lipschitz, in that  $|f(x) - f(y)| \leq 1$  if  $\|x - y\|_1 \leq 2$ ,*

$$\Pr_{x \sim \mu} [|f(x) - \mathbf{E}_\mu f| \geq t] \leq 2 \exp\left(-\frac{\alpha t^2}{2}\right).$$

**Remark 2.2.9.** It is well known that the KL divergence to the stationary distribution decays exponentially at a rate dictated by  $\rho_{\text{MLSI}}(\pi)$  (see [BT06, Theorem 2.4] for more details):

$$\text{KL}(\nu_t \parallel \pi) \leq \text{KL}(\nu_0 \parallel \pi) \cdot \exp(-\rho_{\text{MLSI}}(\pi)t).$$

**Measure decompositions.** Some of the key properties of locally stationary distributions rely on the notion of a measure decomposition. These can be defined in great generality, but we will restrict our attention to distributions on subsets of  $\mathbb{R}^n$  for concreteness.

**Definition 2.2.10** (Measure decomposition). Let  $\pi$  be a distribution on  $\mathbb{R}^n$ . Let  $\rho$  be a mixture distribution, also on  $\mathbb{R}^n$ , which indexes into a family of mixture components  $\{\pi_z\}_{z \in \mathbb{R}^n}$ . We say that  $(\rho, \pi_z)$  is a measure decomposition for  $\pi$  if

$$\pi = \mathbf{E}_{z \sim \rho} \pi_z .$$

One should think of the mixture components  $\pi_z$  as being “simpler” distributions than the original measure  $\pi$ . Not all measure decompositions are useful; there is always a trivial measure decomposition where the mixture  $\rho$  is exactly  $\pi$  and the simpler distributions  $\pi_z$  are just Dirac masses at  $z$ .

Associated to each measure decomposition is a natural Markov chain; see, e.g., [CE22, Definition 6].

**Definition 2.2.11** (Markov chain associated to a measure decomposition). Given a measure decomposition  $\pi = \mathbf{E}_{z \sim \rho} \pi_z$ , its associated Markov chain is defined by

$$x \rightarrow z | x \rightarrow x' | z .$$

Notably, Glauber dynamics and restricted Gaussian dynamics can both be viewed as the associated Markov chain to certain measure decompositions. The relevant decomposition for Glauber dynamics represents  $\pi$  as the mixture of its conditional marginals.

**Remark 2.2.12.** Measure decompositions constructed using stochastic localization have recently been used to prove functional inequalities for a wide class of Ising models [EKZ22, CE22, LMRW24]. We will discuss these properties in greater depth in Section 2.5.

**Symmetric KL divergence.** A useful potential function for us is the symmetric KL divergence.

**Definition 2.2.13.** For a pair of distributions  $\pi$  and  $\nu$  on  $\Omega$ , we define their *symmetric KL divergence* as:

$$\text{SKL}(\pi, \nu) := \text{KL}(\nu \parallel \pi) + \text{KL}(\pi \parallel \nu) .$$

**Observation 2.2.14.** For any  $\pi$  and  $\nu$ , setting  $f$  to be the density of  $\nu$  with respect to  $\pi$ , we have

$$\text{SKL}(\pi, \nu) = \frac{1}{2} \cdot \mathbf{E}_{x, y \sim \pi} \left[ (f(x) - f(y)) \cdot \log \frac{f(x)}{f(y)} \right] .$$

Observe that the above quantity is the Dirichlet form for the trivial “one-step” Markov chain with transition matrix  $\mathbf{1}\pi^\top$  with stationary distribution  $\pi$ , which we shall denote  $K(\pi)$ .



**Lemma 2.2.15.** *Let  $\nu$  be an arbitrary distribution with density  $f$  with respect to  $\pi$ , and  $\tau$  such that  $\tau > \max_{x \in \Omega} \log f(x)$  or  $\tau > \max_{x \in \Omega} \log \frac{1}{f(x)}$ . Then, the symmetric KL divergence can be bounded in terms of the KL divergence as follows.*

$$\text{SKL}(\pi, \nu) \leq (6 + 12\tau) \cdot \text{KL}(\nu \parallel \pi).$$

*Proof.* Recall that the Hellinger distance is defined as

$$H^2(\nu, \pi) := \frac{1}{2} \cdot \mathbf{E}_{\substack{x \sim \pi \\ y \sim \pi}} \left[ \left( \sqrt{f(y)} - \sqrt{f(x)} \right)^2 \right].$$

The symmetric KL divergence may be bounded by the Hellinger distance as follows.

$$\begin{aligned} 2 \cdot \text{SKL}(\nu, \pi) &= \mathbf{E}_{\substack{x \sim \pi \\ y \sim \pi}} \left[ (f(y) - f(x)) (\log f(y) - \log f(x)) \right] \\ &= \mathbf{E}_{\substack{x \sim \pi \\ y \sim \pi}} \left[ (f(y) - f(x)) \cdot \log \frac{f(y)}{f(x)} \cdot \mathbf{1} \left[ \frac{f(y)}{f(x)} \in \left[ \frac{1}{2}, 2 \right] \right] \right] \\ &\quad + \mathbf{E}_{\substack{x \sim \pi \\ y \sim \pi}} \left[ (f(y) - f(x)) \cdot \log \frac{f(y)}{f(x)} \cdot \mathbf{1} \left[ \frac{f(y)}{f(x)} \notin \left[ \frac{1}{2}, 2 \right] \right] \right] \end{aligned}$$

To control the first term, we note that for  $t \in [\frac{1}{2}, 2]$ ,  $(t-1) \log t \leq 6 (\sqrt{t} - 1)^2$ . Consequently,

$$\begin{aligned} &\mathbf{E}_{\substack{x \sim \pi \\ y \sim \pi}} \left[ (f(y) - f(x)) \cdot \log \frac{f(y)}{f(x)} \cdot \mathbf{1} \left[ \frac{f(y)}{f(x)} \in \left[ \frac{1}{2}, 2 \right] \right] \right] \\ &\leq 6 \cdot \mathbf{E}_{\substack{x \sim \pi \\ y \sim \pi}} \left[ \left( \sqrt{f(y)} - \sqrt{f(x)} \right)^2 \cdot \mathbf{1} \left[ \frac{f(y)}{f(x)} \in \left[ \frac{1}{2}, 2 \right] \right] \right] \\ &\leq 12 H^2(\nu, \pi). \end{aligned}$$

To control the second term, we first have

$$\mathbf{E}_{\substack{x \sim \pi \\ y \sim \pi}} \left[ (f(y) - f(x)) \cdot \log f(y) \cdot \mathbf{1} \left[ \frac{f(y)}{f(x)} \notin \left[ \frac{1}{2}, 2 \right] \right] \right] \leq \tau \cdot \mathbf{E}_{\substack{x \sim \pi \\ y \sim \pi}} \left[ |f(y) - f(x)| \cdot \mathbf{1} \left[ \frac{f(y)}{f(x)} \notin \left[ \frac{1}{2}, 2 \right] \right] \right]$$

Furthermore, for  $t \notin [\frac{1}{2}, 2]$ ,  $|t-1| \leq 6(\sqrt{t}-1)^2$ . Consequently,

$$\begin{aligned} &\mathbf{E}_{\substack{x \sim \pi \\ y \sim \pi}} \left[ |f(y) - f(x)| \cdot \mathbf{1} \left[ \frac{f(y)}{f(x)} \notin \left[ \frac{1}{2}, 2 \right] \right] \right] \leq 6 \cdot \mathbf{E}_{\substack{x \sim \pi \\ y \sim \pi}} \left[ \left( \sqrt{f(y)} - \sqrt{f(x)} \right)^2 \cdot \mathbf{1} \left[ \frac{f(y)}{f(x)} \notin \left[ \frac{1}{2}, 2 \right] \right] \right] \\ &\leq 12 H^2(\nu, \pi). \end{aligned}$$

Putting this together, we get that

$$\text{SKL}(\nu, \pi) \leq (6 + 12\tau) \cdot H^2(\nu, \pi).$$

To complete the proof, we use the well-known fact that  $H^2(\nu, \pi) \leq \text{KL}(\nu \parallel \pi)$  ([HW58], also see [SV16, Equation (16)]).  $\square$

## 2.3 Properties of locally stationary distributions

We record some useful properties of locally stationary distributions below.

**Random walks yield locally-stationary measures at a typical time.** The following is a generic statement about any time-reversible Markov chain achieving a locally stationary distribution.

**Lemma 2.3.1.** *For any distribution  $\nu$ , any Markov chain transition kernel  $P$  with stationary distribution  $\pi$  and any  $T > 0$ , for  $t$  chosen uniformly at random in  $[0, T]$ :*

$$\mathbf{E}_{t \sim [0, T]} \mathcal{E}_P(f_t, \log f_t) \leq \frac{\text{KL}(\nu \parallel \pi)}{T} \leq \frac{\log \frac{1}{\pi_{\min}}}{T}. \quad (2.1)$$

*Proof.* By Fact 2.2.5,

$$\begin{aligned} 0 &\leq \text{KL}(\nu_T \parallel \pi) \\ &= \text{KL}(\nu \parallel \pi) - \int_0^T \mathcal{E}(f_t, \log f_t) dt \\ &= \text{KL}(\nu \parallel \pi) - T \cdot \mathbf{E}_{t \sim [0, T]} \mathcal{E}(f_t, \log f_t) \\ &\leq \log \frac{1}{\pi_{\min}} - T \cdot \mathbf{E}_{t \sim [0, T]} \mathcal{E}(f_t, \log f_t). \end{aligned}$$

Rearranging the above gives us the desired statement.  $\square$

A simple consequence of Lemma 2.3.1 and Markov's inequality is that for most times in  $[0, T]$ ,  $\nu_t$  is indeed locally stationary.

**Theorem 2.1.3.** *Fix a time-reversible Markov chain  $P$  with a stationary measure  $\pi$ , any starting distribution  $\nu_0$ , and  $\varepsilon, \delta > 0$ . Let  $T = \frac{1}{\delta \varepsilon} \cdot \log \left( \frac{1}{\pi_{\min}} \right)$ . Then, for a time  $t \sim [0, T]$  chosen uniformly at random, the distribution  $\nu_t$  at time  $t$  is  $\varepsilon$ -locally stationary with respect to  $P$  with probability at least  $1 - \delta$ .*

**Stationarity over small time-scales.** We will also require the observation that if the Dirichlet form at a distribution is small, so too is the total variation distance between it and the distribution obtained after one step of the Markov chain.

**Lemma 2.3.2.** *Let  $P$  be a reversible Markov chain with stationary distribution  $\pi$ ,  $\nu$  an arbitrary distribution, and  $f$  its density relative to  $\pi$ . Then,*

$$\mathcal{E}_P(f, \log f) \geq \text{KL}(P\nu \parallel \nu) \geq 2 \cdot d_{\text{TV}}(\nu, P\nu)^2.$$

*Proof.* We recall the definition of the Dirichlet form,

$$\begin{aligned}
\mathcal{E}_P(f, \log f) &= \frac{1}{2} \mathbf{E}_{\substack{x \sim \pi \\ y \sim P x}} (f(x) - f(y)) \log \frac{f(x)}{f(y)} \\
&= \mathbf{E}_{\substack{x \sim \pi \\ y \sim P x}} f(x) \log \left( \frac{f(x)}{f(y)} \right) && \text{(Reversibility of } P) \\
&= \mathbf{E}_{\substack{x \sim \nu \\ y \sim P x}} \log \left( \frac{f(x)}{f(y)} \right) \\
&= \mathbf{E}_{x \sim \nu} \log f(x) - \mathbf{E}_{y \sim P \nu} \log f(y).
\end{aligned}$$

Adding and subtracting a term, we may neatly express the above in terms of KL divergences as

$$\begin{aligned}
\mathcal{E}_P(f, \log f) &= \left[ \left( \mathbf{E}_{x \sim \nu} \log f(x) - \mathbf{E}_{y \sim P \nu} \log \frac{dP \nu}{d\pi}(y) \right) + \left( \mathbf{E}_{y \sim P \nu} \log \frac{dP \nu}{d\pi}(y) - \mathbf{E}_{y \sim P \nu} \log f(y) \right) \right] \\
&= [(\text{KL}(\nu \parallel \pi) - \text{KL}(P \nu \parallel \pi)) + \text{KL}(P \nu \parallel \nu)] \\
&\geq \text{KL}(P \nu \parallel \nu) \geq 2 \cdot d_{\text{TV}}(P \nu, \nu)^2
\end{aligned}$$

as desired, where the second-to-last inequality follows from the fact that the KL divergence to the stationary distribution is non-increasing with time, and the last inequality is Pinsker's.  $\square$

Consequently, averages of bounded functions do not change much after one step of the Markov chain

**Corollary 2.3.3.** *Let  $\phi : \Omega \rightarrow \mathbb{R}$  be a bounded function on the state space of a Markov chain  $P$ , and  $\nu$  be an  $\varepsilon$ -locally stationary measure. Then,*

$$|\mathbf{E}_{x \sim \nu}[\phi(x)] - \mathbf{E}_{x \sim P \nu}[\phi(x)]| \leq \|\phi\|_{\infty} \cdot \sqrt{2\varepsilon}.$$

**Locally stationary measures are close to stationary measures on small neighborhoods.**

As their name suggests, locally stationary distributions locally resemble the true stationary distribution. For example, typical samples from the locally stationary distribution approximately satisfies the detailed balance condition. Even though we do not explicitly employ this in any applications, we include it here as it gives the impression of a fundamental structural property of locally stationary distributions.

**Lemma 2.3.4.** *For an  $\varepsilon$ -locally stationary distribution  $\nu$  with relative density  $f$ , and for  $x \sim \nu$  and  $y \sim x$ , with probability at least  $1 - \delta$ , we have  $\frac{f(x)}{f(y)} = 1 \pm O\left(\sqrt{\frac{\varepsilon}{\delta}}\right)$ , where  $\delta > 2\varepsilon$ .*

*Proof.* Since  $\mathcal{E}(f, \log f) < \varepsilon$ ,

$$\begin{aligned}\mathbf{E}_{x \sim \pi} \mathbf{E}_{y \sim x} (f(x) - f(y)) \cdot \log \frac{f(x)}{f(y)} &< \varepsilon \\ \mathbf{E}_{x \sim \nu} \mathbf{E}_{y \sim x} \left(1 - \frac{f(y)}{f(x)}\right) \cdot \log \frac{f(x)}{f(y)} &< \varepsilon\end{aligned}$$

Since the random variable at hand is always nonnegative, we can apply Markov's inequality, which tells us that with probability at least  $1 - \delta$ :

$$\left(1 - \frac{f(y)}{f(x)}\right) \cdot \log \frac{f(x)}{f(y)} < \frac{\varepsilon}{\delta}.$$

The claim then follows since the above inequality is violated if  $\frac{f(x)}{f(y)}$  deviates from 1 by more than a constant multiple of  $\sqrt{\frac{\varepsilon}{\delta}}$ .  $\square$

Formally, the following can be abstracted out of the proof of Theorem 2.1.4:

**Lemma 2.3.5.** *Let  $P$  be a Glauber dynamics chain for a distribution  $\mu$  on  $\{\pm 1\}^n$ , and let  $\nu$  be an  $\varepsilon$ -locally stationary measure with respect to  $P$  for some  $\varepsilon > 0$ . For a subset of coordinates  $W \subset [n]$ , and an assignment  $x_{\overline{W}}$  of coordinates outside  $W$ , let  $P_{W, x_{\overline{W}}}$  denote the Glauber dynamics chain of  $\mu|_{x_{\overline{W}}}$ . Suppose for every choice of  $W$  and  $x_{\overline{W}}$ , we have  $\rho_{\text{MLSI}}(P_{W, x_{\overline{W}}}) \geq C$ , then:*

$$\mathbf{E}_{x_{\overline{W}} \sim \nu} \left[ d_{\text{TV}} \left( \nu|_{x_{\overline{W}}}, \pi|_{x_{\overline{W}}} \right) \right] \leq \frac{1}{C} \cdot \sqrt{\varepsilon}.$$

**Corollary 2.3.6.** *In the setting of Lemma 2.3.5, suppose  $\phi : \{0, 1\}^W \rightarrow \mathbb{R}$  is a bounded function of  $x_W$  then,*

$$\mathbf{E}_{x \sim \nu} [\phi(x_W)] \geq \mathbf{E}_{x_{\overline{W}}} \mathbf{E}_{x_W \sim \pi|_{x_{\overline{W}}}} [\phi(x_W)] - \frac{1}{C} \cdot \sqrt{\varepsilon} \cdot \|\phi\|_{\infty}.$$

**Local stationarity can be transferred under component MLSI.** Let  $\pi = \mathbf{E}_{z \sim \rho} \pi_z$  be a measure decomposition. In this section we prove that if Glauber dynamics is locally stationary, so too is the Markov chain associated with the measure decomposition, provided that the mixture components  $\pi_z$  all have a good MLSI constant.

**Lemma 2.3.7.** *Let  $P$  be the Markov chain associated to a measure decomposition  $\pi = \mathbf{E}_{z \sim \rho} \pi_z$ . Let  $f : \{\pm 1\}^n \rightarrow \mathbb{R}_{>0}$  be any function and set  $\tau$  such that  $\min_{x \in \{\pm 1\}^n} f(x) > \exp(-\tau)$  or  $\max_{x \in \{\pm 1\}^n} f(x) < \exp(\tau)$ . For  $\delta := \inf_z \rho_{\text{MLSI}}(\pi_z)$ , we have*

$$\mathcal{E}_P(f, \log f) \leq O\left(\frac{\tau}{\delta}\right) \cdot \mathcal{E}_{\pi}(f, \log f).$$

*Proof.* We use  $C_n$  to denote the hypercube graph on vertex set  $\{\pm 1\}^n$  with edge set having pairs of vertices that differ in a single coordinate. For any nonnegative function  $f$  and distribution  $\pi$  with  $\mathbf{E}_\pi f = 1$ , we use  $f \cdot \pi$  to denote the distribution  $\nu$  with  $\frac{d\nu}{d\pi}(x) = f(x)$ .

For any function  $f$  satisfying the assumption of the statement, we have

$$\begin{aligned}
\mathcal{E}_\pi(f, \log f) &= \sum_{\{x,y\} \in C_n} \frac{1}{n} \cdot \frac{\pi(x) \cdot \pi(y)}{\pi(x) + \pi(y)} \cdot (f(x) - f(y)) \log \frac{f(x)}{f(y)} \\
&= \sum_{\{x,y\} \in C_n} \frac{1}{n} \cdot \frac{\mathbf{E}_{z \sim \rho} \pi_z(x) \cdot \mathbf{E}_{z \sim \rho} \pi_z(y)}{\mathbf{E}_{z \sim \rho} \pi_z(x) + \mathbf{E}_{z \sim \rho} \pi_z(y)} \cdot (f(x) - f(y)) \log \frac{f(x)}{f(y)} \\
&\geq \sum_{\{x,y\} \in C_n} \frac{1}{n} \cdot \mathbf{E}_{z \sim \rho} \left[ \frac{\pi_z(x) \cdot \pi_z(y)}{\pi_z(x) + \pi_z(y)} \right] \cdot (f(x) - f(y)) \log \frac{f(x)}{f(y)} \\
&= \mathbf{E}_{z \sim \rho} [\mathcal{E}_{\pi_z}(f, \log f)].
\end{aligned}$$

Above, the inequality follows from the concavity of the function  $(a, b) \mapsto \frac{ab}{a+b}$  in the non-negative quadrant, and all the Dirichlet forms are with respect to the Glauber dynamics chain.

Continuing the above calculation,

$$\begin{aligned}
\mathbf{E}_{z \sim \rho} [\mathcal{E}_{\pi_z}(f, \log f)] &\geq \mathbf{E}_{z \sim \rho} \left[ \rho_{\text{MLSI}}(\pi_z) \cdot \mathbf{E}_{\pi_z}[f] \cdot \text{KL} \left( \frac{f}{\mathbf{E}_{\pi_z} f} \cdot \pi_z \parallel \pi_z \right) \right] \\
&\geq \Omega \left( \frac{\delta}{\tau} \right) \mathbf{E}_{z \sim \rho} \left[ \mathbf{E}_{\pi_z}[f] \cdot \text{SKL} \left( \frac{f}{\mathbf{E}_{\pi_z} f} \cdot \pi_z, \pi_z \right) \right] && \text{(by Lemma 2.2.15 and MLSI)} \\
&= \Omega \left( \frac{\delta}{\tau} \right) \mathbf{E}_{z \sim \rho} \left[ \mathbf{E}_{x,y \sim \pi_z} (f(x) - f(y)) \log \frac{f(x)}{f(y)} \right] && \text{(by Observation 2.2.14)} \\
&= \Omega \left( \frac{\delta}{\tau} \right) \mathcal{E}_P(f, \log f).
\end{aligned}$$

The claim follows. □

The upshot is that we have complete freedom to select the measure decomposition, provided we can establish an MLSI for the components. This can be useful when it is easier to directly analyze the consequences of local stationarity for the associated Markov chain instead of Glauber dynamics.

## 2.4 Warmup: Large independent sets in triangle-free graphs

Observe that any graph  $G$  on  $n$  vertices with maximum degree  $d$  has an independent set of size  $\frac{n}{d+1}$ , a bound which is tight for the disjoint union of  $(d+1)$ -sized cliques. Ajtai, Komlós, and Szemerédi [AKS80] showed that when  $G$  is triangle-free, the size of the maximum independent set increases to  $\Omega\left(n \cdot \frac{\log d}{d}\right)$ . Shearer [She83] gave an alternate proof that shows such an independent set exists with a leading constant of 1, even if  $G$  merely has average-degree bounded by  $d$ . As a warmup, we prove that Glauber dynamics succeeds at finding a large independent set in  $O(nd^4)$  steps.

**Theorem 2.1.4.** *Let  $G$  be a triangle-free graph on  $n$  vertices with maximum degree bounded by  $d$ . Let  $\mathbf{I}$  be an independent set in  $G$  that arises from Glauber dynamics run for  $O(nd^4)$  time. Then the expected size of  $\mathbf{I}$  is at least  $\frac{1-o_d(1)}{4} \cdot n \cdot \frac{\log d}{d}$ .*

To prove Theorem 2.1.4, we will need the following crude bound on the modified log-Sobolev constant for the uniform distribution over independent sets of a star. A short proof is provided at the end of this section.

**Lemma 2.4.1.** *Let  $\pi$  denote the uniform distribution over independent sets of a star with  $\Delta$  many leaves. Then  $\rho_{\text{MLSI}}(\pi) \geq \exp(-O(\Delta))$ .*

**Remark 2.4.2.** The bound can easily be made  $\frac{1}{\text{poly}(\Delta)}$ , but we will not need this here.

We also leverage the following simple and well-known lemma on the local behavior of a uniformly random independent set. For completeness, we include a short proof of it at the end of this section, following the one provided in Alon & Spencer [AS16, Proposition 1, Page 272]. Throughout this section, we write  $N(v) = \{u \in V : u \sim v\}$  for the open neighborhood of  $v \in V$ , and  $N[v] = N(v) \cup \{v\}$  for the closed neighborhood.

**Lemma 2.4.3.** *Let  $G$  be a triangle-free graph of maximum degree  $d$ , and let  $\pi$  denote the uniform measure over independent sets of  $G$ . For every vertex  $v \in V$ , define the following real-valued score function over  $\{0, 1\}^V$ :*

$$\phi_v(\mathbf{x}) := d\mathbf{x}_v + \sum_{u \in N(v)} \mathbf{x}_u. \quad (2.2)$$

*Then for every pinning  $\tau \in \{0, 1\}^{\overline{N[v]}}$ , we have*

$$\mathbf{E}_{\mathbf{x} \sim \pi}[\phi_v(\mathbf{x}) \mid \mathbf{x}_{\overline{N[v]}} = \tau] \geq \frac{\log d}{2}.$$

The key property of this score function is that it readily yields a lower bound on the size of an independent set  $x \in \{0, 1\}^V$ . This follows from the observation that

$$n \cdot \mathbf{E}_{v \sim V} \phi_v(x) \leq 2d \cdot \sum_{v \in V} x_v. \quad (2.3)$$

Note that by averaging over  $\tau \in \{0, 1\}^{\overline{N[v]}}$  drawn from the marginal distribution of  $\pi$  induced on  $\overline{N[v]}$ , the conclusion of Lemma 2.4.3 combined with Eq. (2.3) implies that a uniformly random independent set drawn from  $\pi$  has expected size at least  $\frac{1}{4} \cdot n \cdot \frac{\log d}{d}$ . We observe that the same claim holds even if the distribution over independent sets is merely locally stationary with respect to Glauber dynamics, rather than being truly uniform.

*Proof of Theorem 2.1.4.* As discussed above, a direct application of the law of total expectation combined with Lemma 2.4.3 yields the first claim concerning the expected size of a uniformly random independent set. We now turn to the second claim. Let  $T \geq 0$  be a parameter to be determined later, and for every  $0 \leq t \leq T$ , let  $\nu_t$  denote the distribution over independent sets after running Glauber dynamics for time- $t$  from an arbitrary initialization. Our goal is to establish the lower bound

$$\mathbf{E}_{t \sim [0, T]} \mathbf{E}_{\nu_t} \mathbf{E}_{v \sim V} \phi_v(x) \geq \frac{\log d}{2} - \varepsilon \quad (2.4)$$

for  $0 < \varepsilon < o_d(1)$ , which when combined with Eq. (2.3) immediately implies that the expected size of the independent set discovered by Glauber dynamics is  $\left(\frac{1-o_d(1)}{4}\right) \cdot n \cdot \frac{\log d}{d}$ . For the purpose of analysis, if  $\nu$  is any distribution over independent sets, we shall think of  $x \sim \nu$  as being sampled in the following alternate way.

1. For a fixed vertex  $v$ , sample  $x_{\overline{N[v]}}$  from the marginal distribution induced by  $\nu$  on  $\overline{N[v]}$ . For each  $w \in N(v)$  that has a neighbor in the independent set  $x_{\overline{N[v]}}$ , pin  $x_w$  to 0, since it is deterministically equal to 0 in the conditional measure  $\nu|_{x_{\overline{N[v]}}}$ .
2. If the number of unpinned vertices at this stage is strictly larger than  $\log d$ , sample  $x_v$  from its corresponding conditional marginal distribution.
3. Let  $U$  be the set of remaining unpinned vertices. Sample  $x_U \sim \nu|_{x_{\overline{U}}}$ .

For any vertex  $v \in V$ , we have

$$\begin{aligned} \mathbf{E}_{x \sim \nu} \phi_v(x) &= \mathbf{E}_{U, x_{\overline{U}}|v} \mathbf{E}_{\nu|_{x_{\overline{U}}}} \phi_v(x) \\ &\geq \mathbf{E}_{U, x_{\overline{U}}|v} \left[ \mathbf{E}_{\pi|_{x_{\overline{U}}}} \phi_v(x) - 2d \cdot d_{\text{TV}}(\nu|_{x_{\overline{U}}}, \pi|_{x_{\overline{U}}}) \right] \end{aligned}$$

$$\geq \frac{\log d}{2} - 2d \cdot \mathbf{E}_{U, x_{\bar{U}}|v} d_{\text{TV}}(v|x_{\bar{U}}, \pi|x_{\bar{U}}), \quad (2.5)$$

Note that the random subset of vertices  $U$ , as well as the boundary condition  $x_{\bar{U}}$ , are all drawn from the above process with respect to  $v$ , not  $\pi$ . The first inequality follows by applying Fact 2.2.2 along with  $2d$ -boundedness and nonnegativity of the score function  $\phi_v$ . For the second inequality, note that if  $v \in U$ , then we may invoke Lemma 2.4.3. Now suppose  $v \notin U$ . If  $v$  is pinned 1, then  $\phi_v(x) = d$ . If  $v$  is pinned 0, then by triangle-freeness,  $x_u = 1$  with probability  $1/2$  independently for all  $u \in U$ . Since  $|U| \geq \log d$ , the lower bound follows.

In the rest of this argument, we will show that when  $t \sim [0, T]$ ,  $v$  is equal to  $v_t$  and  $v$  is chosen uniformly at random, we can achieve a strong upper bound on

$$\mathbf{E}_t \mathbf{E}_v \mathbf{E}_{U, x_{\bar{U}}|v} d_{\text{TV}}(v_t|x_{\bar{U}}, \pi|x_{\bar{U}}).$$

For the rest of this proof, we shall abbreviate  $v_t|x_{\bar{U}}$  and  $\pi|x_{\bar{U}}$  as  $v'_t$  and  $\pi'$ , respectively. Furthermore, let  $f'_t$  denote the relative density of  $v'_t$  with respect to  $\pi'$ . By Pinsker's inequality, we can bound the above by:

$$\mathbf{E}_t \mathbf{E}_v \mathbf{E}_{U, x_{\bar{U}}|v} \sqrt{\text{KL}(v'_t \parallel \pi')} \leq \sqrt{\mathbf{E}_t \mathbf{E}_v \mathbf{E}_{U, x_{\bar{U}}|v} \text{KL}(v'_t \parallel \pi')}.$$

We focus our attention on showing an upper bound on the term in the square-root.

$$\begin{aligned} \mathbf{E}_t \mathbf{E}_{v \sim V} \mathbf{E}_{U, x_{\bar{U}}|v} \text{KL}(v'_t \parallel \pi') &\leq \mathbf{E}_t \mathbf{E}_v \mathbf{E}_{U, x_{\bar{U}}|v} \frac{1}{\rho_{\text{MLSI}}(\pi')} \mathcal{E}_{\pi'}(f'_t, \log f'_t) \\ &\leq O(d) \cdot \mathbf{E}_t \mathbf{E}_v \mathbf{E}_{U, x_{\bar{U}}|v} \mathcal{E}_{\pi'}(f'_t, \log f'_t) \\ &\leq O(d^2) \cdot \mathbf{E}_t \mathcal{E}_{\pi}(f_t, \log f_t) \\ &\leq O(d^2) \cdot \frac{n}{T}. \end{aligned}$$

In the above, the first inequality uses the definition of  $\rho_{\text{MLSI}}(\pi')$ . For the second inequality, note that almost surely, either  $G[U]$  is a star centered at  $v$  with at most  $\log d$  many leaves, or  $G[U]$  consists entirely of isolated vertices due to pinning  $v$ . In either case, we have  $\rho_{\text{MLSI}}(\pi') \geq \Omega(\frac{1}{d})$  by appropriately applying Fact 2.2.7 or Lemma 2.4.1. The third inequality is based on comparing Dirichlet forms. The final inequality is a direct application of Lemma 2.3.1. Plugging in the above into (2.5) and setting  $T = O(nd^4/\varepsilon^2)$ , we get Eq. (2.4) as desired.  $\square$

**Remark 2.4.4.** Using a similar argument, one can establish a similar result for Max-Cut on triangle-free graphs with maximum degree  $d$ . In particular, Glauber dynamics run for  $\text{poly}(n)$  many steps on the antiferromagnetic Ising model on  $G$  with inverse temperature  $\frac{1}{\sqrt{d}}$  finds a cut of size  $\frac{1}{2} + \Omega\left(\frac{1}{\sqrt{d}}\right)$ .



*Proof of Lemma 2.4.3.* Let  $S \subseteq N(v)$  denote the subset of neighbors of  $v$  which are not adjacent to any vertex of the independent set  $x_{N[v]} = \tau$ , and write  $k = |S|$ . Observe that the distribution of  $x_{N[v]}$  conditioned on  $x_{N[v]} = \tau$  is given by choosing the singleton  $\{v\}$  with probability  $\frac{1}{2^{k+1}}$ , or a uniformly random subset of  $S$  with the remaining probability. Hence,

$$\mathbb{E}_{x \sim \pi}[\phi_v(x) \mid x_{N[v]} = \tau] = \frac{d}{2^{k+1}} + \frac{k}{2} \cdot \frac{2^k}{2^{k+1}}.$$

The above expression is always at least  $\frac{\log d}{2}$  for any choice of nonnegative integer  $k$ .  $\square$

*Proof of Lemma 2.4.1.* By [DSC96, Corollary A.4] and [BT06, Proposition 3.6], we have that  $\rho_{\text{MLSI}}(\pi) \geq \frac{1-2\pi_*}{\log\left(\frac{1}{\pi_*}-1\right)} \cdot \lambda(\pi)$ , where  $\lambda(\pi)$  denotes the spectral gap of Glauber dynamics for  $\pi$ , and  $\pi_* = \min_{x: \pi(x) > 0} \pi(x) = \frac{1}{2^{\Delta+1}}$ . Hence, it suffices to establish that  $\lambda(\pi) \geq \exp(-O(\Delta))$ . For this, we appeal to the simple fact that random walk on a connected graph with  $n$  vertices has spectral gap at least  $1/\text{poly}(n)$ . A comparison of Dirichlet forms between Glauber dynamics and simple random walk on the  $n = 2^{\Delta+1}$  many independent sets of  $G$  yields the desired lower bound.  $\square$

## 2.5 Weak recovery in spiked models

In this section, we present our main application: using Glauber dynamics on Ising models to achieve weak recovery guarantees for spiked matrix models. Let us recall the definition of an Ising model.

**Definition 2.5.1** (Ising model). Let  $J \in \mathbb{R}^{n \times n}$  be a symmetric *interaction matrix* and  $h \in \mathbb{R}^n$  an *external field*. The *Ising model* corresponding to  $J$  and  $h$  is the probability distribution  $\mu_{J,h}$  on  $\{\pm 1\}^n$ , where

$$\mu_{J,h}(x) \propto \exp\left(\frac{1}{2}x^\top Jx + \langle h, x \rangle\right).$$

We drop the  $h$  from the subscript when it is equal to 0.

We now pose the general algorithmic task that we wish to solve using Glauber dynamics in *polynomial* time.

**Problem 2.5.2.** Let  $W$  be a symmetric matrix in  $\mathbb{R}^{n \times n}$  and let  $v$  be a unit vector in  $\mathbb{R}^n$ . For  $M := W + \lambda vv^\top$ , what is the behavior of Glauber dynamics run for  $\text{poly}(n)$  many steps for the Ising model  $\mu_M$ ? In particular, under what assumptions does Glauber dynamics recover a vector that is well-correlated with  $v$  after  $\text{poly}(n)$  time?

Our main results of this section resolves Problem 2.5.2 affirmatively in the following concrete settings.

- $M = W + \lambda vv^\top$ , where the spectral diameter of  $W$  is at most 1, and  $v \in \{\pm \frac{1}{\sqrt{n}}\}^n$ .
- $M = A_G - \frac{d}{n} \mathbf{1}\mathbf{1}^\top$ , the degree-centered adjacency matrix of a sparse stochastic block model.

In particular, despite a failure to mix to  $\mu$ , Glauber dynamics run for polynomially many steps still manages to recover nontrivial information about the planted signal  $v$ .

**Theorem 2.1.7.** *Let  $W$  be a matrix with  $\kappa \leq W \leq 1 - \kappa$ , and  $v \in \{\pm \frac{1}{\sqrt{n}}\}^n$ . Let  $P$  denote the kernel of the Glauber dynamics chain with stationary distribution  $\mu_{W + \lambda vv^\top}$ , and  $x_0$  an arbitrary point on  $\{\pm 1\}^n$ . There exists a large enough constant  $\lambda > 0$  such that for  $T = \tilde{\Theta}(n^5)$ , and for  $t \sim [0, T]$ , with probability  $1 - o(1)$  we have:*

$$\mathbf{E}_{x \sim P^t \delta_{x_0}} [|\langle x, v \rangle|] \geq \left( \kappa \exp\left(-\frac{1}{\kappa}\right) - o(1) \right) \cdot \sqrt{n}.$$

Similarly, we have the following result for the stochastic block model.

**Theorem 2.1.12.** *There exist constants  $\lambda_0, \beta, c > 0$  such that for all  $\lambda$  satisfying  $|\lambda| \geq \lambda_0$ , for  $(\sigma, G) \sim \text{SBM}(n, d, \lambda)$ , with probability  $1 - o(1)$  over the randomness of  $(\sigma, G)$ , the following holds.*

*Let  $P$  denote the kernel of Glauber dynamics with stationary distribution  $\mu_{\frac{\beta}{\sqrt{d}}(A_G - \frac{d}{n} \mathbf{1}\mathbf{1}^\top)}$  and  $x_0$  an arbitrary point on  $\{\pm 1\}^n$ . For  $T = \tilde{\Theta}(n^{5+o_d(1)})$ , and for  $t \sim [0, T]$ , with probability  $1 - o(1)$ , we have:*

$$\mathbf{E}_{x \sim P^t \delta_{x_0}} [|\langle x, \sigma \rangle|] > cn.$$

Our strategy to study Glauber dynamics in each of these settings is to relate its behavior to that of a different Markov chain called *restricted Gaussian dynamics*, whose definition we recall below.

**Definition 2.5.3** (Restricted Gaussian dynamics). For  $M = W + \lambda vv^\top$ , *restricted Gaussian dynamics* (RGD) is a Markov chain on  $\{\pm 1\}^n$  where a transition from  $x$  to  $y$  is given by the following:

- Sample  $g \sim \mathcal{N}(0, 1)$  and define  $z := (\lambda \langle v, x \rangle + \sqrt{\lambda} g) \cdot v$ .
- Sample  $y$  from the Gibbs distribution  $\mu_{W, z}$ .

Note that the above definition specifies the full joint distribution of  $(x, z)$ , which induces a measure decomposition  $\mu_M = \mathbf{E}_{z \sim \rho} \mu_{W, z}$ . Further, RGD is the associated Markov chain for this decomposition.

**Remark 2.5.4.** Given access to  $M$ , but not  $W$  and  $v$ , it is unclear how to efficiently implement restricted Gaussian dynamics. For example, for the SBM application, one has  $v = \frac{1}{\sqrt{n}} \sigma$ . If one could compute  $v$  (or  $W$ ), then the recovery task would already be solved. Nevertheless, in light of Lemma 2.3.7, it is useful for analysis because we can relate its behavior to that of Glauber dynamics on  $M$ .

In Section 2.5, we will introduce some preliminaries on Ising models that will be crucial in the analysis. Then, in Section 2.5, we show that RGD achieves weak recovery. Finally, in Section 2.5, we prove the main theorems about weak recovery using Glauber dynamics, Theorems 2.1.7 and 2.1.12.

## Entropic stability and conservation of variance

Recall the definition of a measure decomposition from Section 2.2. In order to prove our weak recovery result, we will need more structural properties for our decompositions. We now formalize these requirements.

At a high level, we would like a measure decomposition where the components of the decomposition  $\pi_z$  “inherit” properties of  $\pi$  itself, but are also simpler at the same time. When the individual components  $\pi_z$  inherit the variance of  $\pi$ , the mixture is said to satisfy conservation of variance.

**Definition 2.5.5** (Conservation of variance). We say that an Ising model  $\pi$  on  $\{\pm 1\}^n$  satisfies *conservation of variance* for  $(\rho, \pi_z)$  with parameter  $C_{\text{Var}} \in [0, 1]$ , if for all functions  $f : \{\pm 1\}^n \rightarrow \mathbb{R}$ , we have

$$\mathbf{E}_{z \sim \rho} [\mathbf{Var}_{\pi_z} [f]] \geq C_{\text{Var}} \cdot \mathbf{Var}_{\pi} [f].$$

**Remark 2.5.6.** By the law of total variance, for any  $f : \{\pm 1\}^n \rightarrow \mathbb{R}$ ,

$$\mathbf{Var}_{\pi} [f] = \mathbf{E}_{z \sim \rho} [\mathbf{Var}_{\pi_z} [f]] + \mathbf{Var} [\mathbf{E}_{\pi_z} f] \geq \mathbf{E}_{z \sim \rho} [\mathbf{Var}_{\pi_z} [f]].$$

The notion of conservation of variance captures mixtures where, loosely, a reverse of the above inequality is true.

To control the mean of various Ising models, we will also use another notion, called *entropic stability* [CE22]. To introduce it, we need the notion of a *tilt*. For a measure  $\pi$  on  $\Omega \subseteq \mathbb{R}^n$  and vector  $v \in \mathbb{R}^n$ , we define the *tilted measure*  $\mathcal{T}_v \pi$  on  $\Omega$  by

$$\frac{d\mathcal{T}_v \pi(x)}{d\pi(x)} \propto e^{\langle v, x \rangle}.$$

**Definition 2.5.7** (Entropic stability). Let  $\Omega \subseteq \mathbb{R}^n$  and  $\psi : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$ . For  $\alpha > 0$ , we say that a measure  $\pi$  on  $\Omega$  is  $\alpha$ -entropically stable with respect to  $\psi$  if for all  $v \in \mathbb{R}^n$ ,

$$\psi(\mathbf{m}(\mathcal{T}_v \pi), \mathbf{m}(\pi)) \leq \alpha \cdot \text{KL}(\mathcal{T}_v \pi \parallel \pi),$$

where we recall that  $\mathbf{m}(\pi) = \mathbf{E}_{x \sim \pi} x$ . We denote by  $\alpha_{\text{Ent}}$  the best (smallest) such  $\alpha$ .

**Remark 2.5.8.** It turns out that entropic stability can be used to prove the related notion of conservation of *entropy* for certain measure decompositions, which in turn implies conservation of variance; see, e.g., [BT06, Proof of Proposition 3.5]. However, we will not dwell on this point and refer the interested reader to [CE22].

We now state our special measure decompositions for the two applications. For bounded spectral diameter  $W$ , one can decompose the Ising model into a strongly log-concave mixture of product distributions; these types of decompositions were studied in [BB19, CE22]. See [CE22, Section 5.1] for details.

**Theorem 2.5.9.** Let  $W$  be a symmetric matrix with  $\kappa \leq W \leq 1 - \kappa$  and  $h$  be an arbitrary external field. Then there exists a (strongly log-concave) mixture  $\rho$  on  $\mathbb{R}^n$  such that

$$\mu_{W,h} = \mathbf{E}_{z \sim \rho} [\mu_{0,Wz+h}].$$

Moreover, the following properties hold:

- (9)  $\mu_{W,h}$  is  $\frac{1}{\kappa}$ -entropically stable with respect to  $(x, y) \mapsto \|x - y\|^2$ .
- (9)  $C_{\text{Var}}(\rho, \mu_{0,Wz+h}) \geq \exp(-1/\kappa)$ .
- (9) For every  $z$  in the support of  $\rho$ ,  $\rho_{\text{MLSI}}(\mu_{0,Wz+h}) \geq \frac{1}{n} \cdot \frac{1}{1-\kappa}$ .

Recently, a subset of the authors proved that a similar decomposition exists for an Ising model associated to the stochastic block model [LMRW24].

**Theorem 2.5.10.** *Let  $(\sigma, G) \sim \text{SBM}(n, d, \lambda)$ , and let  $\bar{A}_G = A_G - \mathbf{E}[A_G|\sigma]$  be the centered adjacency matrix. There exists some constant  $\beta > 0$ , a mixture distribution  $\rho$  on  $\mathbb{R}^n$ , and function  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  such that for any external field  $h$ ,*

$$\mu_{\beta \bar{A}_G / \sqrt{d}, h} = \mathbf{E}_{z \sim \rho}[\mu_{H, g(z)+h}],$$

where  $H$  is an interaction matrix supported on the edges of a forest, with at most one additional cycle per connected component. Moreover, the following properties hold:

- (10) *There is a positive constant  $\alpha_{\text{Ent}}$  such that  $\mu_{\beta \bar{A}_G / \sqrt{d}, h}$  is  $\alpha_{\text{Ent}}$ -entropically stable with respect to  $(x, y) \mapsto \|I_{[n] \setminus H}(x - y)\|^2$ .*
- (10)  $C_{\text{Var}}(\mu_{\beta \bar{A}_G / \sqrt{d}, h}) \geq \Omega(1)$ .
- (10) *For every  $z$  in the support of  $\rho$ ,  $\rho_{\text{MLSI}}(\mu_{H, g(z)+h}) \geq \frac{1}{n^{1+o_d(1)}}$ .*
- (10) *The number of non-isolated vertices in  $H$  is at most  $\gamma n$ , where  $\gamma(d) = o_d(1)$  is a sufficiently small constant that also shrinks with  $d$ . Here, a non-isolated vertex is a vertex with at least one distinct neighbor.*

**Remark 2.5.11.** We provide pointers for the reader interested in extracting Theorem 2.5.10 from [LMRW24, CE22].

- Theorem 2.5.10 follows from [LMRW24, Theorem 5.2] and [CE22, Lemma 40] (cf. [LMRW24, Lemma 3.18]).
- For Theorem 2.5.10, conservation of entropy follows from [LMRW24, Theorem 5.2, Lemma 3.18, Lemma 3.29]. This, in turn, implies conservation of variance by the generic fact that  $\lim_{c \rightarrow \infty} \text{Ent}[(f + c)^2] = 2\text{Var}[f]$  (see e.g., [BT06, Proof of Proposition 3.5]).
- Theorem 2.5.10 is [LMRW24, Theorem 5.1].
- Theorem 2.5.10 follows from [LMRW24, Lemma 6.21].

## Restricted Gaussian dynamics achieves weak recovery

In this section we prove versions of Theorems 2.1.7 and 2.1.12 for restricted Gaussian dynamics—our goal will be to show that if the Dirichlet form with respect to the restricted Gaussian dynamics Markov chain is small, then it has non-trivial correlation with the

planted vector  $v$ . Since local stationarity with respect to Glauber dynamics transfers over to local stationarity with respect to RGD (Lemma 2.3.7), this suffices to complete the proof.

Showing that RGD succeeds at correlating with the spike amounts to Lemma 2.5.12, where we show that for “nice” Ising models, a strong external field applied to the Ising model shows itself in its mean. In particular, if the field is aligned with  $v$ , then the mean is correlated with  $v$ . One can then show that for any distribution  $\nu$  that is locally stationary with respect to RGD, samples from  $\nu$  must already have good correlation with the planted  $v$  (Lemma 2.5.17).

**Lemma 2.5.12.** *Let  $W$  be an interaction matrix such that for any external field  $h$ , there is a decomposition*

$$\mu_{W,h} = \mathbf{E}_{z \sim \rho} [\mu_{H,g_h(z)}],$$

*where  $H$  is an arbitrary interaction matrix supported on the edges of a graph that we also denote  $H$ , such that the following hold.*

1. *There is a positive constant  $\alpha_{\text{Ent}}$  such that  $\mu_{W,h}$  is  $\alpha_{\text{Ent}}$ -entropically stable with respect to  $(x, y) \mapsto \|I_{[n] \setminus H}(x - y)\|^2$ ,*
2. *Variance is conserved in this decomposition with constant  $C_{\text{Var}}$ .*
3. *There is a constant  $\gamma < 1$  such that the number of non-isolated vertices in  $H$  is at most  $\gamma n$ .*

Now, let  $v \in \left\{ \pm \frac{1}{\sqrt{n}} \right\}^n$ , and set  $s > 0$ . Then,

$$\mathbf{E}_{x \sim \mu_{W,sv}} |\langle x, v \rangle| \geq (1 - \gamma) \cdot \frac{C_{\text{Var}}}{2} \cdot \min \left\{ s, 2 \cdot \sqrt{\frac{n}{C_{\text{Var}} \alpha_{\text{Ent}}}} \right\}$$

**Remark 2.5.13.** The above assumptions on  $H$  may be a bit confusing. However, in the case of an interaction matrix with bounded spectral diameter, we in fact have  $\gamma = 0$ , so  $H$  is empty.

**Remark 2.5.14.** We further remark that by [LMRW24, Lemma 3.18]—a mild strengthening of [CE22, Lemma 40]—the first condition holds if for all tilts  $h$ ,  $\|\text{Cov}(\mu_{W,h})_{[n] \setminus H}\| \leq \alpha_{\text{Ent}}$ . In fact, only a bound on  $\|\text{Cov}(\mu_{W,tv})_{[n] \setminus H}\|$  for “most”  $0 \leq t \leq s$  is required. A technique to show the second condition, involving stochastic localization, also requires a bound on the covariance.

*Proof.* The proof strategy is to use the fact that:

$$\mathbf{E}_{x \sim \mu_{W,sv}} \langle x, v \rangle = \int_0^s \frac{d}{dt} \mathbf{E}_{x \sim \mu_{W,tv}} \langle x, v \rangle dt,$$

and obtain a lower bound on the correlation by showing a lower bound on the derivative.

The following standard calculation gives us a formula for the derivative as a variance of the correlation.

$$\begin{aligned} \frac{d}{dt} \mathbf{E}_{x \sim \mu_{W,tv}} \langle x, v \rangle &= \frac{d}{dt} \frac{\mathbf{E}_{x \sim \mu_{W,0}} e^{t\langle x, v \rangle} \langle x, v \rangle}{\mathbf{E}_{x \sim \mu_{W,0}} e^{t\langle x, v \rangle}} \\ &= \frac{\mathbf{E}_{x \sim \mu_{W,0}} e^{t\langle x, v \rangle} \langle x, v \rangle^2}{\mathbf{E}_{x \sim \mu_{W,0}} e^{t\langle x, v \rangle}} - \frac{(\mathbf{E}_{x \sim \mu_{W,0}} e^{t\langle x, v \rangle} \langle x, v \rangle)^2}{(\mathbf{E}_{x \sim \mu_{W,0}} e^{t\langle x, v \rangle})^2} = \mathbf{Var}_{x \sim \mu_{W,tv}} \langle x, v \rangle. \end{aligned}$$

Now, consider the measure decomposition provided by the assumptions, of the form

$$\mu_{W,tv} = \mathbf{E}_{z \sim \rho} [\mu_{H,z}]$$

for some measure  $\rho$  over external fields  $z$ . We have by the law of total variance that

$$\mathbf{Var}_{x \sim \mu_{W,tv}} \langle x, v \rangle \geq \mathbf{E}_{z \sim \rho} \mathbf{Var}_{x \sim \mu_{H,z}} \langle x, v \rangle.$$

Note that we can lower bound the above variance by the contribution of the isolated vertices in  $H$ , which are mutually independent of all other vertices. In particular, let  $S_{\text{iso}}$  be the set of isolated vertices in  $H$ . Then, we have:

$$\begin{aligned} \mathbf{E}_{z \sim \rho} \mathbf{Var}_{x \sim \mu_{H,z}} \langle x, v \rangle &\geq \mathbf{E}_{z \sim \rho} \mathbf{Var}_{x \sim \mu_{H,z}} \left[ \sum_{i \in S_{\text{iso}}} x_i v_i \right] \\ &= \mathbf{E}_{z \sim \rho} \sum_{i \in S_{\text{iso}}} v_i^2 \cdot \mathbf{Var}_{x \sim \mu_{H,z}} [x_i]. \end{aligned}$$

By assumption, the decomposition conserves the variance of arbitrary functions with parameter  $C_{\text{Var}}$ . Hence,

$$\mathbf{E}_{z \sim \rho} \mathbf{Var}_{x \sim \mu_{H,z}} [x_i] \geq C_{\text{Var}} \cdot \mathbf{Var}_{x \sim \mu_{W,tv}} [x_i] = C_{\text{Var}} \cdot \left( 1 - \mathbf{E}_{x \sim \mu_{W,tv}} [x_i]^2 \right).$$

Consequently, because  $\|v\|_{\infty}^2 = 1/n$ ,

$$\mathbf{Var}_{x \sim \mu_{W,tv}} \langle x, v \rangle \geq C_{\text{Var}} \cdot \left( (1 - \gamma) - \frac{1}{n} \|I_{[n] \setminus H} \cdot \mathbf{m}(\mu_{W,tv})\|^2 \right).$$

By assumption,  $\mu_{W,tv}$  is  $\alpha_{\text{Ent}}$ -entropically stable with respect to  $(x, y) \mapsto \|I_{[n]\setminus H} \cdot (x - y)\|^2$ . By definition, this implies that

$$\begin{aligned} \|I_{[n]\setminus H}(\mathbf{m}(\mu_{W,tv}) - \mathbf{m}(\mu_{W,0}))\|^2 &\leq \alpha_{\text{Ent}} \cdot \text{KL}(\mu_{W,tv} \parallel \mu_{W,0}) \text{ and} \\ \|I_{[n]\setminus H}(\mathbf{m}(\mu_{W,0}) - \mathbf{m}(\mu_{W,tv}))\|^2 &\leq \alpha_{\text{Ent}} \cdot \text{KL}(\mu_{W,0} \parallel \mu_{W,tv}) \end{aligned}$$

and since  $\mathbf{m}(\mu_{W,0}) = 0$  by symmetry, we get that

$$\|I_{[n]\setminus H}\mathbf{m}(\mu_{W,tv})\|^2 \leq \frac{\alpha_{\text{Ent}}}{2} \cdot \text{SKL}(\mu_{W,tv}, \mu_{W,0}) = \frac{\alpha_{\text{Ent}}}{2} \cdot \mathbf{E}_{x \sim \mu_{W,tv}} \langle tv, x \rangle.$$

Therefore,

$$\frac{d}{dt} \mathbf{E}_{x \sim \mu_{W,tv}} \langle x, v \rangle \geq C_{\text{Var}} \left( (1 - \gamma) - \frac{t\alpha_{\text{Ent}}}{2n} \mathbf{E}_{x \sim \mu_{W,tv}} \langle x, v \rangle \right).$$

Now, set

$$L := \frac{sC_{\text{Var}}}{1 + \frac{s^2C_{\text{Var}}\alpha_{\text{Ent}}}{4n}} \cdot (1 - \gamma).$$

We claim that  $\mathbf{E}_{x \sim \mu_{W,sv}} \langle x, v \rangle \geq L$ . To prove this claim, assume for contradiction that  $\mathbf{E}_{x \sim \mu_{W,sv}} \langle x, v \rangle < L$ . Then, since  $\mathbf{E}_{x \sim \mu_{W,sv}} \langle x, v \rangle$  is non-decreasing with  $s$  (namely, its derivative is variance), we get:

$$\begin{aligned} \mathbf{E}_{x \sim \mu_{W,sv}} \langle x, v \rangle &\geq C_{\text{Var}} \int_0^s \left( (1 - \gamma) - \frac{tL\alpha_{\text{Ent}}}{2n} \right) dt \\ &= C_{\text{Var}} \left( s(1 - \gamma) - \frac{s^2L\alpha_{\text{Ent}}}{4n} \right) = L, \end{aligned}$$

which contradicts the assumption that  $\mathbf{E}_{x \sim \mu_{W,sv}} \langle x, v \rangle < L$ .

We conclude the proof by doing casework on  $s$ ; to this end set  $s_\star := 2 \cdot \sqrt{\frac{n}{C_{\text{Var}}\alpha_{\text{Ent}}}}$ . If  $s \leq s_\star$ , then the definition of  $L$  yields  $L \geq \frac{sC_{\text{Var}}}{2} \cdot (1 - \gamma)$ . On the other hand, the case of  $s \geq s_\star$  immediately reduces to the above calculation because by monotonicity,

$$\mathbf{E}_{x \sim \mu_{W,sv}} \langle x, v \rangle \geq \mathbf{E}_{x \sim \mu_{W,s_\star v}} \langle x, v \rangle \geq (1 - \gamma) \cdot \frac{C_{\text{Var}}}{2} \cdot 2 \cdot \sqrt{\frac{n}{C_{\text{Var}}\alpha_{\text{Ent}}}}. \quad \square$$

**Remark 2.5.15.** The above Lemma 2.5.12 can easily be generalized to relax the assumptions on  $v$  to having bounded  $\ell_\infty$  norm or being subgaussian. However, the resulting bound is a little messier, so we omit it for the sake of readability.

As a corollary of the above and the definition of restricted Gaussian dynamics, we get the following.



**Corollary 2.5.16** (RGD boost). *Given  $W$  as in the previous theorem,  $x$  such that  $|\langle x, v \rangle| = r$ , and  $\mathbf{y} \sim_{P_{\text{RGD}}} x$ , we have:*

$$\mathbf{E}_{\mathbf{y}} |\langle \mathbf{y}, v \rangle| \geq (1 - \gamma) \cdot \frac{C_{\text{Var}}}{2} \cdot \mathbf{E}_{g \sim \mathcal{N}(0,1)} \min \left\{ |\lambda r + \sqrt{\lambda} g|, 2 \sqrt{\frac{n}{C_{\text{Var}} \alpha_{\text{Ent}}}} \right\}.$$

**Lemma 2.5.17** (Correlation of locally stationary distributions under RGD). *Let  $W$  be an interaction matrix satisfying the conditions in Lemma 2.5.12, and set  $M = W + \lambda v v^\top$ , where  $v \in \left\{ \pm \frac{1}{\sqrt{n}} \right\}^n$  and  $\lambda \geq 1$ . Additionally, suppose that the distribution  $\mu_{W,sv}$  satisfies  $\rho_{\text{MLSI}}(\mu_{W,sv}) \geq n^{-1-o(1)}$  for every  $s \in \mathbb{R}$ . Set*

$$\delta_\star = \delta_\star(\lambda) := (1 - \gamma) \cdot \frac{C_{\text{Var}}}{2} \cdot \min \left\{ \sqrt{\frac{2\lambda}{\pi}}, \frac{\lambda}{2} \right\} - 1.$$

*Then for any  $\lambda$  such that  $\delta_\star > 0$ , any  $\varepsilon$ -locally stationary distribution  $v$  under RGD, and any  $\delta \in (0, \delta_\star)$ , it holds that*

$$\mathbf{E}_{x \sim v} |\langle x, v \rangle| \geq 0.99 \cdot (1 - \gamma) \cdot \sqrt{\frac{C_{\text{Var}} n}{\alpha_{\text{Ent}}}} - 2\delta,$$

*for  $\varepsilon < \delta^4 \cdot \frac{\alpha_{\text{Ent}}^2}{C_{\text{Var}}^2 (1-\gamma)^4} \cdot \frac{1}{n^2} - \exp(-\Omega(n))$ .*

*Proof.* Set  $\eta = (1 - \gamma) \cdot \sqrt{\frac{C_{\text{Var}}}{\alpha_{\text{Ent}}}}$ ,  $\bar{\eta} = 0.99\eta$ ,  $\eta^\downarrow = 0.995\eta$ , and define the function  $\phi(x) := \min \{ |\langle x, v \rangle|, \eta \sqrt{n} \}$ .

By Corollary 2.3.3, which states that bounded functions are stable under a single step of restricted Gaussian dynamics,

$$\mathbf{E}_{\mathbf{y} \sim P_{\text{RGD}} v} \phi(\mathbf{y}) - \mathbf{E}_{x \sim v} \phi(x) < \eta \sqrt{\varepsilon n}, \quad (2.6)$$

that is, the average value of  $\phi(x)$  cannot change much under a single step of RGD.

Assume for contradiction that  $\mathbf{E}_{x \sim v} \phi(x) < \bar{\eta} \sqrt{n} - 2\delta$ . We will show that a single step of RGD boosts the expected correlation enough that the resulting value of  $\mathbf{E}_{\mathbf{y} \sim P_{\text{RGD}} v} \phi(\mathbf{y})$  violates (2.6). This follows from two claims. First, observe that by Markov's inequality:

$$\Pr_{x \sim v} [|\langle x, v \rangle| < \bar{\eta} \sqrt{n} - n^{o(1)} - \delta] = \Pr_{x \sim v} [\phi(x) < \bar{\eta} \sqrt{n} - n^{o(1)} - \delta] \geq \Omega \left( \frac{\delta}{\eta \sqrt{n}} \right).$$

Second, we establish that for any  $x$  such that  $|\langle x, v \rangle| = r$  and  $\mathbf{y} \sim_{P_{\text{RGD}}} x$ ,

1. if  $r < \bar{\eta}\sqrt{n} - n^{o(1)} - \delta$ , then  $\mathbf{E}_y \phi(y) - r > \delta$ , and
2.  $\mathbf{E}_y \phi(y) - \min\{r, \bar{\eta}\sqrt{n} - n^{o(1)}\} > -\exp(-n^{1-o(1)})$ .

These two claims together tell us that:

$$\mathbf{E}_{x \sim P_{\text{RGD}}^v} \phi(x) - \mathbf{E}_{x \sim v} \phi(x) > \Omega\left(\frac{\delta^2}{\eta\sqrt{n}}\right) - \exp(-n^{1-o(1)}) \quad (2.7)$$

which gives the desired contradiction, as (2.6) and (2.7) cannot simultaneously be true given the assumption on the relationship between  $\delta$  and  $\varepsilon$ .

It remains to prove the claimed lower bounds on  $\mathbf{E}_y \phi(y) - r$ . Let us begin with the simpler task of proving lower bounds on  $\mathbf{E}_y |\langle y, v \rangle| - r$ : we will show that

1. if  $r < \eta\sqrt{n} - \delta$ , then  $\mathbf{E}_y |\langle y, v \rangle| - r > \delta$ , and
2.  $\mathbf{E}_y |\langle y, v \rangle| - \min\{r, \eta\sqrt{n}\} > -\exp(-\Omega(n))$ .

Let  $x$  be an arbitrary point on the hypercube with  $|\langle x, v \rangle| = r$ . We may assume without loss of generality that  $\langle x, v \rangle \geq 0$ .

For ease of notation, set  $\tilde{C} = \frac{C_{\text{Var}}}{2} \cdot (1 - \gamma)$ . We prove this lower bound by splitting into cases based on the value of  $r$ . By Corollary 2.5.16, the following lower bounds hold for  $y \sim P_{\text{RGD}}$  depending on the value of  $r$ :

**Case  $r < 1$ .** We have

$$\begin{aligned} \mathbf{E}_y |\langle y, v \rangle| &\geq \mathbf{E}_{g \sim \mathcal{N}(0,1)} \min \left\{ \tilde{C} \cdot |\lambda r + \sqrt{\lambda} g|, \eta\sqrt{n} \right\} \\ &\geq \tilde{C} \cdot \left( \mathbf{E}_{g \sim \mathcal{N}(0,1)} |\lambda r + \sqrt{\lambda} g| - \mathbf{E}_{g \sim \mathcal{N}(0,1)} |\lambda r + \sqrt{\lambda} g| \cdot \mathbf{1}_{\tilde{C} \cdot |\lambda r + \sqrt{\lambda} g| > \eta\sqrt{n}} \right) \\ &\geq \tilde{C} \cdot \left( \sqrt{\lambda} \mathbf{E}_{g \sim \mathcal{N}(0,1)} |g| - \exp(-\Omega(n)) \right) \\ &= \tilde{C} \left( \sqrt{\frac{2\lambda}{\pi}} - \exp(-\Omega(n)) \right) \end{aligned}$$

where the final inequality follows from the fact that  $\mathbf{E}_{g \sim \mathcal{N}(0,1)} |g + R|$  is minimized at  $R = 0$ , and from standard Gaussian concentration.

It follows that

$$\mathbf{E}_y |\langle y, v \rangle| - r \geq \tilde{C} \cdot \sqrt{\frac{2\lambda}{\pi}} - 1 - \exp(-\Omega(n)) > \delta.$$

**Case**  $1 \leq r \leq \frac{3\eta\sqrt{n}}{2\tilde{C}\lambda}$ . In this case,

$$\begin{aligned} \mathbf{E}_y |\langle \mathbf{y}, v \rangle| &\geq \mathbf{E}_{g \sim \mathcal{N}(0,1)} \min \left\{ \frac{\tilde{C}}{2} \cdot |\lambda r + \sqrt{\lambda} g|, \eta\sqrt{n} \right\} \\ &\geq \frac{\tilde{C}}{2} \cdot \left( \mathbf{E}_{g \sim \mathcal{N}(0,1)} |\lambda r + \sqrt{\lambda} g| - \mathbf{E}_{g \sim \mathcal{N}(0,1)} |\lambda r + \sqrt{\lambda} g| \cdot \mathbf{1}_{\tilde{C} \cdot |\lambda r + \sqrt{\lambda} g| > 2\eta\sqrt{n}} \right) \\ &\geq \frac{\tilde{C}\lambda}{2} \cdot r - \exp(-\Omega(n)). \end{aligned}$$

Observe that

$$\mathbf{E}_y |\langle \mathbf{y}, v \rangle| - r > \left( \frac{\tilde{C}\lambda}{2} - 1 \right) > \delta.$$

**Case**  $\frac{3\eta\sqrt{n}}{2\tilde{C}\lambda} \leq r \leq \eta\sqrt{n} - \delta$ . In this case, by standard Gaussian concentration, the value of  $\tilde{C} \cdot |\lambda r + \sqrt{\lambda} g|$  exceeds  $\eta\sqrt{n}$  with exponentially high probability, and hence  $\mathbf{E}_y |\langle \mathbf{y}, v \rangle| \geq \eta\sqrt{n} - \exp(-\Omega(n))$ . Thus, in this case too, we have  $\mathbf{E}_y |\langle \mathbf{y}, v \rangle| - r > \delta - \exp(-\Omega(n))$ .

**Case**  $\eta\sqrt{n} - \delta \leq r$ . By identical reasoning to the previous case, in this case  $\mathbf{E}_y |\langle \mathbf{y}, v \rangle| - \eta\sqrt{n} > -\exp(-\Omega(n))$ .

We must now translate the above lower bounds to lower bounds on  $\mathbf{E}_y \phi(\mathbf{y})$ . To do this, we will use Lemma 2.2.8, which says that the correlation of a sample concentrates around its expectation. We will again consider different cases, depending on where the expectation lies. Concretely, we will prove that

$$\mathbf{E} \phi(\mathbf{y}) = \mathbf{E} \min \{ |\langle \mathbf{y}, v \rangle|, \eta\sqrt{n} \} \geq \min \{ \mathbf{E} |\langle \mathbf{y}, v \rangle|, \bar{\eta}\sqrt{n} - n^{o(1)} \} - \exp(-n^{1-o(1)}). \quad (2.8)$$

Towards proving this, we have

$$\mathbf{E} \phi(\mathbf{y}) = \mathbf{E} |\langle \mathbf{y}, v \rangle| \mathbf{1}_{|\langle \mathbf{y}, v \rangle| \leq \eta\sqrt{n}} + \eta\sqrt{n} \cdot \Pr [|\langle \mathbf{y}, v \rangle| \geq \eta\sqrt{n}] \quad (2.9)$$

$$= \mathbf{E} |\langle \mathbf{y}, v \rangle| + \mathbf{E} (\eta\sqrt{n} - |\langle \mathbf{y}, v \rangle|) \mathbf{1}_{|\langle \mathbf{y}, v \rangle| \geq \eta\sqrt{n}}. \quad (2.10)$$

Let  $\eta^\uparrow = 1.005\eta$ , and recall our earlier definitions  $\bar{\eta} = 0.99\eta$  and  $\eta^\downarrow = 0.995\eta$ .

**Case**  $\mathbf{E} |\langle \mathbf{y}, v \rangle| < \eta^\downarrow \sqrt{n}$ .

Here, we have by (2.10) that

$$\mathbf{E} \phi(\mathbf{y}) \geq \mathbf{E} |\langle \mathbf{y}, v \rangle| - \sqrt{n} \cdot \Pr [|\langle \mathbf{y}, v \rangle| \geq \eta\sqrt{n}].$$

Lemma 2.2.8 allows us to bound

$$\Pr [|\langle \mathbf{y}, v \rangle| \geq \eta\sqrt{n}] \leq \exp \left( -\Omega \left( \frac{\eta\sqrt{n} - \mathbf{E} |\langle \mathbf{y}, v \rangle|}{n^{o(1)}} \right)^2 \right) = \exp \left( -n^{1-o(1)} \right).$$

**Case**  $\eta^\downarrow \sqrt{n} \leq \mathbf{E}|\langle \mathbf{y}, v \rangle| \leq \eta^\uparrow \sqrt{n}$ .

(2.10) implies that

$$\begin{aligned}
\mathbf{E}\phi(\mathbf{y}) &\geq \mathbf{E}|\langle \mathbf{y}, v \rangle| - \mathbf{E}|\eta\sqrt{n} - |\langle \mathbf{y}, v \rangle|| \\
&\geq \mathbf{E}|\langle \mathbf{y}, v \rangle| - |\eta\sqrt{n} - \mathbf{E}|\langle \mathbf{y}, v \rangle|| - \mathbf{E}||\langle \mathbf{y}, v \rangle| - \mathbf{E}|\langle \mathbf{y}, v \rangle|| \\
&\geq \mathbf{E}|\langle \mathbf{y}, v \rangle| - |\eta\sqrt{n} - \mathbf{E}|\langle \mathbf{y}, v \rangle|| - n^{o(1)} \\
&\geq \eta^\downarrow \sqrt{n} - (\eta - \eta^\downarrow)\sqrt{n} - n^{o(1)} \\
&= \bar{\eta}\sqrt{n} - n^{o(1)},
\end{aligned}$$

where the third inequality uses Lemma 2.2.8.

**Case**  $\eta^\uparrow \sqrt{n} < \mathbf{E}|\langle \mathbf{y}, v \rangle|$ .

(2.9) implies that

$$\mathbf{E}\phi(\mathbf{y}) \geq \eta\sqrt{n} \cdot \Pr[|\langle \mathbf{y}, v \rangle| \geq \eta\sqrt{n}].$$

We give a lower bound for the above using Lemma 2.2.8 again:

$$\Pr[|\langle \mathbf{y}, v \rangle| \geq \eta\sqrt{n}] \geq 1 - \exp\left(-\Omega\left(\frac{(\mathbf{E}|\langle \mathbf{y}, v \rangle| - \eta\sqrt{n})^2}{n^{o(1)}}\right)\right) = 1 - \exp\left(-n^{1-o(1)}\right),$$

as desired.

Now, let us put the pieces together to obtain a contradiction to (2.6). By (2.8), and the preceding lower bound on the shift in  $\mathbf{E}|\langle \mathbf{y}, v \rangle|$ , if  $|\langle x, v \rangle| < \bar{\eta}\sqrt{n} - \delta - n^{o(1)}$ ,

$$\begin{aligned}
\mathbf{E}_{\mathbf{y} \sim P_{\text{RGD}}\delta_x} \phi(\mathbf{y}) &\geq \min\{\mathbf{E}|\langle \mathbf{y}, v \rangle|, \bar{\eta}\sqrt{n} - n^{o(1)}\} - \exp\left(-n^{1-o(1)}\right) \\
&\geq \min\{r + \delta, \bar{\eta}\sqrt{n} - n^{o(1)}\} - \exp\left(-n^{1-o(1)}\right) \\
&\geq r + \delta - \exp\left(-n^{1-o(1)}\right),
\end{aligned}$$

and similarly, if  $|\langle x, v \rangle| > \bar{\eta}\sqrt{n} - \delta - n^{o(1)}$ , we have

$$\mathbf{E}_{\mathbf{y} \sim P_{\text{RGD}}\delta_x} \phi(\mathbf{y}) \geq \min\{r, \bar{\eta}\sqrt{n} - n^{o(1)}\} - \exp\left(-n^{1-o(1)}\right),$$

which completes the proof. □

## Glauber dynamics achieves weak recovery

We are finally ready to prove the main results of this section.

**Theorem 2.1.7.** *Let  $W$  be a matrix with  $\kappa \leq W \leq 1 - \kappa$ , and  $v \in \left\{\pm \frac{1}{\sqrt{n}}\right\}^n$ . Let  $P$  denote the kernel of the Glauber dynamics chain with stationary distribution  $\mu_{W+\lambda vv^\top}$ , and  $x_0$  an arbitrary point on  $\{\pm 1\}^n$ . There exists a large enough constant  $\lambda > 0$  such that for  $T = \tilde{\Theta}(n^5)$ , and for  $t \sim [0, T]$ , with probability  $1 - o(1)$  we have:*

$$\mathbf{E}_{x \sim P^t \delta_{x_0}} [|\langle x, v \rangle|] \geq \left( \kappa \exp\left(-\frac{1}{\kappa}\right) - o(1) \right) \cdot \sqrt{n}.$$

*Proof.* The proof strategy is to establish that  $\nu_t$  is locally stationary for restricted Gaussian dynamics, and then use Lemma 2.5.17 to conclude. Let  $\nu_t$  denote the distribution after running Glauber dynamics for time  $t$ . For  $t \sim [0, T]$ , by Theorem 2.1.3, the distribution  $\nu_t$  is  $O\left(\frac{n}{\zeta T}\right)$ -locally stationary, except with probability at most  $\zeta$ . Let us assume that this event occurs.

We now compute the parameters we can plug into Lemma 2.3.7.

- First, due to local stationarity, we have  $\mathcal{E}(f_t, \log f_t) < O\left(\frac{n}{\zeta T}\right)$ .
- By Theorem 2.5.9 of Theorem 2.5.9, the value of  $\delta$  from Lemma 2.3.7 is at least  $\frac{1}{(1-\kappa) \cdot n}$ .
- Finally, it remains to bound the one-sided density parameter  $\tau$  from Lemma 2.3.7. Since the Hamiltonian of  $\mu$  is bounded in absolute value by  $O(n)$ , we may take  $\tau = O(n)$ .

Once we plug in these parameters, we get:

$$\mathcal{E}_{\text{RGD}}(f_t, \log f_t) < \tilde{O}\left(\frac{n^3}{\zeta T}\right).$$

When  $\zeta T \gg n^5$ , we get

$$\mathcal{E}_{\text{RGD}}(f_t, \log f_t) < o\left(\frac{1}{n^2}\right).$$

By Lemma 2.5.17 with parameters  $\varepsilon = o\left(\frac{1}{n^2}\right)$  and  $\delta = o(1)$ , and the values for  $C_{\text{Var}}$  and  $\alpha_{\text{Ent}}$  from Theorem 2.5.9, we get that:

$$\mathbf{E}_{x \sim \nu_t} |\langle x, v \rangle| \geq \left( \sqrt{\frac{C_{\text{Var}}}{\alpha_{\text{Ent}}}} - o(1) \right) \cdot \sqrt{n} \geq (\kappa \exp(-1/\kappa) - o(1)) \sqrt{n}.$$

We treat  $\kappa$  as a constant, and can set  $\zeta = o(1)$  and  $T = \tilde{\Omega}(n^5)$  to finish the proof.  $\square$

The proof of Theorem 2.1.12 for SBM weak recovery is very similar, but we provide the explicit parameter dependencies for completeness. To be explicit, in the notation of Problem 2.5.2, for SBM recovery we have  $v = \frac{1}{\sqrt{n}}\sigma$ ,  $M = A_G - \frac{d}{n}\mathbf{1}\mathbf{1}^\top$ , and  $W = \bar{A}_G := A_G - \mathbf{E}[A_G|\sigma]$ . One can verify that  $M = W + \lambda\sqrt{d} \cdot vv^\top$  by using the fact that  $\mathbf{E}[A_G|\sigma] = \frac{d}{n}\mathbf{1}\mathbf{1}^\top + \frac{\lambda\sqrt{d}}{n}\sigma\sigma^\top$ .

**Theorem 2.1.12.** *There exist constants  $\lambda_0, \beta, c > 0$  such that for all  $\lambda$  satisfying  $|\lambda| \geq \lambda_0$ , for  $(\sigma, G) \sim \text{SBM}(n, d, \lambda)$ , with probability  $1 - o(1)$  over the randomness of  $(\sigma, G)$ , the following holds.*

*Let  $P$  denote the kernel of Glauber dynamics with stationary distribution  $\mu_{\frac{\beta}{\sqrt{d}}(A_G - \frac{d}{n}\mathbf{1}\mathbf{1}^\top)}$  and  $x_0$  an arbitrary point on  $\{\pm 1\}^n$ . For  $T = \tilde{\Theta}(n^{5+o_d(1)})$ , and for  $t \sim [0, T]$ , with probability  $1 - o(1)$ , we have:*

$$\mathbf{E}_{x \sim P^t \delta_{x_0}}[|\langle x, \sigma \rangle|] > cn.$$

*Proof.* Again, let  $\nu_t$  denote the distribution after running Glauber dynamics on  $\mu_{M,0}$  until time  $t$ . As before, let us assume that  $\nu_t$  is  $O\left(\frac{n}{\zeta T}\right)$ -locally stationary, which occurs with probability  $1 - \zeta$ . We now compute the parameters we can plug into Lemma 2.3.7.

- By Theorem 2.5.10 of Theorem 2.5.10, the value of  $\delta$  from Lemma 2.3.7 is at least  $\frac{1}{n^{1+o_d(1)}}$ .
- To bound the one-sided density parameter  $\tau$  from Lemma 2.3.7, we have  $\|Wx\|_\infty \leq O(\log n)$ , so that we can take  $\tau \leq O(n \log n)$ .

Once we plug in these parameters, we get:

$$\mathcal{E}_{\text{RGD}}(f_t, \log f_t) < \tilde{O}\left(\frac{n^{3+o_d(1)}}{\zeta \cdot T}\right).$$

When  $\zeta \cdot T \gg n^{5+o_d(1)}$ , we get that

$$\mathcal{E}_{\text{RGD}}(f_t, \log f_t) < o\left(\frac{1}{n^2}\right).$$

By Lemma 2.5.17 with parameters  $\varepsilon$  as above and  $\delta = O(1)$ , for some universal constant  $c$ ,

$$\mathbf{E}_{x \sim \nu_t}|\langle x, \sigma \rangle| \geq cn.$$

Setting  $\zeta = o(1)$  and  $T = \tilde{\Omega}(n^{5+o_d(1)})$  completes the proof.  $\square$

# Chapter 3

## Markov Chains Approximate Message Passing

### 3.1 Introduction

Markov chains are fundamental objects of study which model the dynamics of complex systems arising in statistical physics, computational biology, and social networks. Besides modeling physical systems, they are a highly successful algorithmic tool to sample from high-dimensional distributions in Bayesian inference, statistics, and theoretical computer science.

It has long been empirically observed that canonical Markov chains such as *Glauber dynamics* which attempt to sample from the posterior distribution are quite successful for Bayesian inference. For example, [DKMZ11] observed that for community detection in the stochastic block model, the output of the Glauber dynamics is Bayes optimal. However, there has been a dearth of rigorous results confirming or explaining the algorithmic success of Markov chains in statistical inference. One fundamental reason for the lack of rigorous progress is that, for many posterior distributions of interest, Glauber dynamics provably suffers from slow worst-case mixing. This leads to the following natural question.

**Question 3.1.1.** *If we run Glauber dynamics for polynomial time on the posterior distribution of a natural Bayesian inference problem, does it achieve Bayes-optimal performance?*

To make progress on this, we revisit early empirical studies which compared the performance of Glauber dynamics to that of local *message passing* algorithms. Our starting point is [DKMZ11], which found that the performance of Glauber dynamics matches that of the Belief Propagation (BP) algorithm which directly estimates the marginals of the posterior in the stochastic block model.

Since then, message passing algorithms such as BP and its dense variant Approximate Message Passing (AMP) [DMM09, BM11] have been widely applied in a variety of high-dimensional statistics problems, to great success (for a few examples, see [DAM16, MV21, EAMS22, HMP24] and references therein). A particularly appealing property of AMP is that its asymptotic performance can be rigorously characterized by a finite-dimensional recursion known as *state evolution*.

For certain inference problems, one can use state evolution to design AMP algorithms to construct Bayes-optimal estimators. Similarly, we *a priori* know that Glauber achieves the same guarantees in exponential time, since by then the dynamics would be able to produce Gibbs samples. Motivated by this, we refine our initial question to ask about connecting the performance of these two classes of algorithms if Glauber is only run for polynomial time.

**Question 3.1.2.** *If we run Glauber dynamics for polynomial time, can we prove it performs as well as AMP?*

Our main results make progress on Question 3.1.2 by proving a formal connection between Markov chains and AMP for the spiked Wigner inference problem, which we define below. We show that the dynamics of the correlation with the spike—the observable of primary interest—are governed by fixed point relations which also characterize the asymptotic performance of AMP for the same problem. See Figure 3.1 for a schematic diagram of our main results.

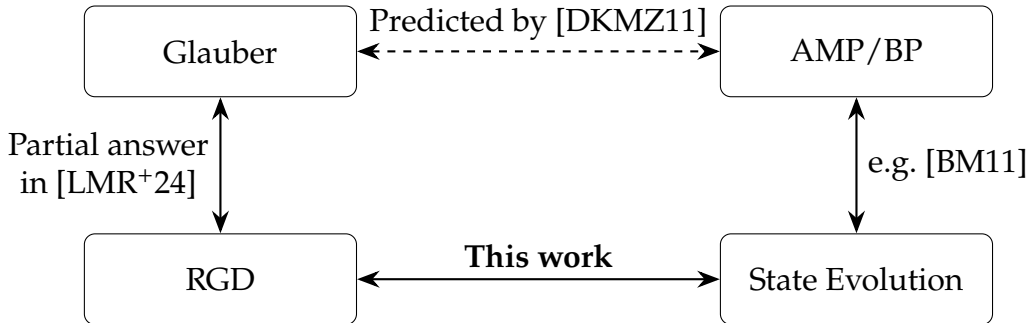


Figure 3.1: An illustration of our results.

The precision of our results requires model-specific calculations that we did not attempt to generalize beyond the spiked Wigner setting. However, our high-level proof strategy is fairly generic and we believe it should be applicable to general spiked matrix inference problems.



In the spiked Wigner model, one aims to recover an unknown *signal*  $\mathbf{x}$  drawn from the prior  $\text{Unif}(\{\pm 1\})^{\otimes N}$  given a noisy matrix measurement

$$\mathbf{M} = \mathbf{W} + \frac{\lambda}{N} \mathbf{x} \mathbf{x}^\top, \quad (3.1)$$

for some signal-to-noise ratio (or SNR)  $\lambda \geq 0$ , and a *noise* matrix  $\mathbf{W} \sim \text{GOE}(N)$  with  $W_{ij} = W_{ji} \sim \mathcal{N}(0, \frac{1}{N})$  for  $i \neq j$  and  $W_{ii} \sim \mathcal{N}(0, \frac{2}{N})$ .

The spiked Wigner model has been extensively studied over the last two decades, providing a rich test bed for average-case complexity. Perhaps the first interesting phenomenon uncovered by this line of work is the famous BBP phase transition [HR04, BAP05], which marks the information-theoretic and computational phase transition at  $\lambda = 1$ . Indeed, it is known that for  $\lambda < 1$ ,  $\mathbf{M}$  is mutually contiguous with  $\mathbf{W}$ , whereas for  $\lambda > 1$  the top eigenvector of  $\mathbf{M}$  is nontrivially correlated with  $\mathbf{x}$ .

A straightforward calculation shows that the posterior distribution in the spiked Wigner model is of the form

$$\mu_{\beta \mathbf{M}}(\sigma) \propto \exp\left(\frac{\beta}{2} \sigma^\top \mathbf{M} \sigma\right) = \exp\left(\frac{\beta}{2} \sigma^\top \mathbf{W} \sigma + \frac{\beta \lambda}{2N} \langle \sigma, \mathbf{x} \rangle^2\right) \quad (3.2)$$

for  $\beta = \lambda$ .

**Beyond worst-case mixing.** [LMR<sup>+</sup>24] introduced the framework of locally stationary distributions to analyze the performance of Markov chains before they have mixed. Roughly speaking, a locally stationary distribution is any distribution which is approximately invariant under another step of the Markov chain. They are also very easy to sample from—any reversible Markov chain such as Glauber dynamics produces a locally stationary distribution in  $\text{poly}(N)$  steps.

For the spiked Wigner problem, they show that for sufficiently large (constant)  $\lambda$  and  $\beta < \frac{1}{4}$ , the Glauber dynamics run on  $\mu_{\beta \mathbf{M}}$  for a large polynomial time from an arbitrary initialization reaches a vector that is nontrivially correlated with the spike. To prove this result, they combined the properties of locally stationary distributions with an analysis of a different Markov chain known as *Restricted Gaussian Dynamics* (RGD), which we define below.<sup>1</sup>

The following definition follows the restricted Gaussian dynamics framework of [STL20, LST21, CE22].

---

<sup>1</sup>The precise definition is not terribly important, so readers seeking intuition can skip the definition without much concern.

**Definition 3.1.3** (Restricted Gaussian dynamics). For  $M = W + \frac{\lambda}{N} \mathbf{x} \mathbf{x}^\top$ , *restricted Gaussian dynamics* (RGD) at inverse temperature  $\beta$  is a Markov chain on  $\{\pm 1\}^N$  where a transition from  $\sigma$  to  $\sigma'$  is given by the following:

- Sample  $g \sim \mathcal{N}(0, 1)$  and define  $z := \beta \lambda \frac{1}{N} \langle \sigma, \mathbf{x} \rangle + \sqrt{\frac{\beta \lambda}{N}} g$ .
- Sample  $\sigma'$  from the distribution  $\mu_{\beta W, z \mathbf{x}}$ .

It is straightforward to show that RGD has  $\mu_{\beta M}$  as its unique stationary distribution. However, the same bottleneck that precludes fast mixing for Glauber also prevents RGD from mixing quickly. In spite of this, we achieve a precise and nearly distributional understanding of the dynamics of RGD, which allows us to deduce several interesting properties of its dynamics. Although RGD is not an implementable algorithm, the tools developed by [LMR<sup>+</sup>24] allow us to transfer this understanding over to Glauber dynamics in certain temperature regimes.

Before we state our main results, we take a brief excursion to give some standard spin glass definitions.

**Definition 3.1.4.** Let  $W \sim \text{GOE}(N)$ . The Sherrington–Kirkpatrick model (SK model) at inverse temperature  $\beta > 0$  and external field  $v \in \mathbb{R}^N$  is the distribution over  $\{\pm 1\}^N$  with density

$$\mu_{\beta W, v}(\sigma) \propto \exp \left( \frac{\beta}{2} \cdot \sigma^\top W \sigma + \langle v, \sigma \rangle \right).$$

Our results for the performance of Glauber on the spiked Wigner posterior rely on the rapid mixing of Glauber dynamics for the SK model. To simplify the discussion, we will state our results whenever the SK model satisfies the following mixing condition based on modified log-Sobolev inequalities.

**Condition 1** (MLSI). We say that  $\beta < 1$  satisfies (MLSI) if with probability  $1 - o(1)$  over the random matrix  $W$ , for all  $h \in \mathbb{R}$ ,  $\mu_{\beta W, h \mathbf{1}}$  satisfies a modified log-Sobolev inequality with constant  $\Omega(1/N)$ .

A recent breakthrough line of work [EKZ22, CE22, AKV24] has made significant progress towards establishing rapid mixing of Glauber dynamics in the entire high-temperature regime for the SK model. In particular, they establish that (MLSI) holds for  $\beta < 0.295$ ; it is conjectured that (MLSI) holds for the entire high-temperature regime  $\beta < 1$  [MPV87]. We are now poised to state our main results.

## Main results

The following theorem establishes that the short-term behavior of the correlation of Glauber iterates is an approximate fixed point of an AMP-like fixed point equation in the entire high SNR regime  $\lambda > 1$ . Furthermore, it establishes a sharp phase transition at  $\beta = \frac{1}{\lambda}$  for the success of Glauber dynamics at weak recovery of the spike.

**Theorem 3.1.5** (Informal, see Theorem 3.5.2). *Let  $\sigma_0 \in \{\pm 1\}^N$  be arbitrary, and let  $(\sigma_t)_{t \geq 0}$  be the trajectory of the Glauber dynamics run with stationary distribution  $\mu_{\beta M}$  initialized at  $\sigma_0$ . Let  $T \geq \tilde{\Omega}(N^4)$ , and  $\widehat{\sigma} = \sigma_t$  for a uniformly random  $t \sim [0, T]$ .*

*Then, for any  $\varepsilon > 0$  and any  $\beta \in (0, 1)$  such that  $\beta \neq \frac{1}{\lambda}$  and (MLSI) holds, with high probability over the noise  $\mathbf{W}$ , we have*

$$\Pr \left[ \left| \frac{1}{N} |\langle \widehat{\sigma}, \mathbf{x} \rangle| - \text{OPT}_{\beta, \lambda} \right| > \frac{1}{N^{1/2-\varepsilon}} \right] = o_N(1),$$

*for some constant  $\text{OPT}_{\beta, \lambda}$  that is the largest fixed point of an explicit AMP-like recursion. Here,  $\text{OPT}_{\beta, \lambda} > 0$  if  $\beta > \frac{1}{\lambda}$  and  $\text{OPT}_{\beta, \lambda} = 0$  if  $\beta < \frac{1}{\lambda}$ .*

*Furthermore,  $\text{OPT}_{\beta, \lambda}$  characterizes the correlation of a Gibbs sample:*

$$\Pr_{\sigma \sim \mu_{\beta M}} \left[ \left| \frac{1}{N} |\langle \sigma, \mathbf{x} \rangle| - \text{OPT}_{\beta, \lambda} \right| > \frac{1}{N^{1/2-\varepsilon}} \right] \leq e^{-cN^\varepsilon}.$$

**Remark 3.1.6.** If one only cared about obtaining a correlation that is within  $o_N(1)$  of  $\text{OPT}_{\beta, \lambda}$ , the formal version of the above result will show that it suffices to set  $T = \tilde{\Omega}(N^3)$ .

The key technical innovation in our work is a much more fine-grained understanding of the RGD Markov chain, which is essential for proving the above theorem. As mentioned earlier, the connection between RGD and the Glauber dynamics is present in prior work [LMR<sup>+</sup>24]. Although the form of their connection breaks when  $\beta > 1$ , our understanding of RGD extends to this regime. We believe that a connection between RGD and the Glauber dynamics should also exist in this low-temperature regime, but do not attempt to answer this question in the present work.

Towards illustrating our fine-grained understanding of RGD, we present the following theorem, which provides a direct comparison between RGD and AMP. Indeed, this connection is precisely what yields the fixed point equations alluded to in Theorem 3.1.5.

**Theorem 3.1.7** (Informal, see Section 3.2). *Let  $\lambda \geq \beta > 1$  and  $\varepsilon > 0$  independent of  $N$ . There exists some threshold  $\tau \in (0, 1)$  (depending on  $\beta, \lambda$ ) such that the following is true.*

Let  $x \in \{\pm 1\}^N$  be an arbitrary spike, and fix  $\sigma \in \{\pm 1\}^N$  such that  $\frac{1}{N} |\langle \sigma, x \rangle| > \tau$ . Let  $\sigma^{\text{RGD}}$  be obtained by taking one step of RGD from  $\sigma$ , and  $\sigma^{\text{AMP}}$  be obtained by taking one step of an AMP-like update from  $\sigma$ . Then, with high probability over the noise  $\mathbf{W}$ , the following holds:

$$\left| \frac{1}{N} \langle \sigma^{\text{RGD}}, x \rangle - \frac{1}{N} \langle \sigma^{\text{AMP}}, x \rangle \right| = o_N(1).$$

**Remark 3.1.8.** The formal version of this theorem depends on verifying an explicit numerical condition which can easily be simulated on a computer; see Section 3.2 for more details.

In other words, if both algorithms are provided a sufficiently correlated warm start, one step of RGD is close to one step of AMP in terms of correlation. In the formal version, the instantiation  $\sigma$  is also allowed to slightly depend on  $\mathbf{W}$ .

One is naturally led to wonder whether a version of Theorem 3.1.5 holds true for RGD for  $\beta > 1$ , so that we can relate the correlation of a locally stationary point of RGD to that of a sample from the posterior? Unfortunately, we expect such results to be false from worst-case initializations. However, we *can* still hope for such results from non-worst-case initializations, specifically from warm starts that already have some non-trivial correlation.

**Theorem 3.1.9** (Informal, see Theorem 3.5.1). Fix  $\lambda > 1$ ,  $K > 0$  large,  $\varepsilon > 0$  small, and  $\beta > \frac{1}{\lambda}$ . There exists a threshold  $\tau \in (0, 1)$  such that the following holds. With high probability over the noise  $\mathbf{W}$ , the following holds. Let  $\sigma_0 \in \{\pm 1\}^N$ , and assume that  $|R(\sigma_0, \mathbf{1})| > \tau$ . Suppose that we run the RGD Markov chain at inverse temperature  $\beta$  from  $\sigma_0$  for  $T \geq \omega(\log N)$  steps to arrive at distribution  $\nu_T$ . Suppose  $T \leq N^K$ . Then,

$$\mathbf{E}_{\sigma_T \sim \nu_T} [||R(\sigma_T, \mathbf{x})| - \text{OPT}_{\beta, \lambda}|] \leq O\left(\frac{1}{N^{1/2-\varepsilon}}\right).$$

While the requirement that the initialization has sufficiently large correlation may seem like too strong a requirement, we emphasize that this is not the case. If one ran an annealed version of the RGD Markov chain where  $\beta$  is slowly incremented, the resulting trajectory of correlations is large enough that this lower bound condition is satisfied. Accordingly, we believe this might provide a path to showing that annealed *Glauber* dynamics attains a correlation of  $\text{OPT}_{\beta, \lambda}$ .

**Remark 3.1.10.** Again, the formal version of the above theorem requires establishing two explicit numerical conditions, which can easily be simulated on a computer.

**Remark 3.1.11.** We draw attention here to the (perhaps) surprising fact that simulations suggest the correlation  $\text{OPT}_{\beta,\lambda}$  with the spike increases *past*  $\beta = \lambda$ . This is a rather subtle point: while the posterior mean, corresponding to  $\mu_{\lambda M}$  (and AMP for that matter) attains the MMSE, this does not imply that samples from the posterior are Bayes-optimal with respect to correlation. What *is* true is that the posterior mean is Bayes optimal for the *normalized squared correlation*, i.e. it is the estimator  $\hat{x}$  which maximizes  $\mathbf{E}\left[\frac{\langle \hat{x}, x \rangle^2}{\|\hat{x}\|^2}\right]$ .

En route to proving the above result, we prove a new high-precision estimate for the mean magnetization for the SK model under weak external field, which may be of independent interest. These estimates enable us to overcome technical difficulties with worst-case analysis of the RGD dynamics for  $\beta < 1$ .

**Theorem 3.1.12** (Informal, see Lemma 3.6.1). *Let  $\beta < 1$  and  $\varepsilon > 0$  sufficiently small. With probability  $1 - o(1)$  over  $W \sim \text{GOE}(N)$ , for all  $h = O(N^{-(1/4+\varepsilon)})$ ,*

$$\frac{1}{N} \mathbf{E}_{x \sim \mu_{\beta W, h1}} \langle x, \mathbf{1} \rangle = \mathbf{E}_{g \sim \mathcal{N}(0,1)} [\tanh(\beta g \sqrt{q} + h)] \left(1 + O\left(\frac{1}{N^\varepsilon}\right)\right) + O\left(\frac{1}{N^{1/2+\varepsilon}}\right),$$

where  $q$  is the unique solution to  $q = \mathbf{E}[\tanh^2(\beta g \sqrt{q} + h)]$ .

In particular, this result establishes the correct asymptotics for the mean magnetization up to the phase transition in weak external field established by [DW23].

## Related work

**Spiked Wigner model.** The spiked Wigner model has provided a rich test bed for average-case algorithms and complexity. Indeed, the basic question of efficient detection and recovery of the spike has been studied through the lens of spectral algorithms [BAP05], low-degree polynomials [PWBM18], and message passing [DAM16]. Interestingly, these efficient algorithms all succeed in polynomial time once  $\lambda > 1$ . Furthermore, when  $\lambda < 1$ , the mutual information between the observation  $M$  and the signal  $x$  turns out to be  $o(1)$  [LM17], indicating that one cannot hope to design an estimator  $\hat{x} \in \{\pm 1\}^N$  that is non-trivially correlated with the spike  $x$ .

For more general priors beyond  $\text{Unif}(\{\pm 1\}^N)$ , the story becomes significantly richer. For example, there exist sparse priors where there exists a statistical-computational gap: there is an intermediate regime of SNRs  $(\lambda_c, \lambda)$  where recovery is information theoretically possible, but no known efficient algorithm succeeds at weak recovery. [EAKJ20] pin down the information-theoretic threshold at  $\lambda = \lambda_c$  for general bounded i.i.d. priors. The algorithmic side of special cases such as sparse PCA [DM14] and non-negative PCA [MR15]

have also been extensively studied. The problem has also been generalized to the tensor setting [MR14, WEM19, BGJ20, KX25], where there is an interesting gap in performance between naive local algorithms and methods based on the sum-of-squares hierarchy or the Kikuchi method.

In terms of sampling from the posterior, due to the sign symmetry of the posterior creating a bottleneck in distributions, Glauber suffers an exponential worst-case mixing time. On the flip side, it has been observed that random initialization or simulated annealing appears to empirically succeed at the goal of sampling from the posterior [DKMZ11]. Nevertheless, it has remained challenging to give provable guarantees about the performance of sampling algorithms, even under algorithmic warm starts. A notable positive sampling result in this direction is [MW23], which uses a different diffusion-based algorithm to sample from the posterior in Wasserstein distance when the SNR  $\lambda$  is a sufficiently large constant.

**Bayes optimal estimation.** There is by now a large body of work characterizing Bayes optimal estimation for mean-squared error. Indeed, past work has designed algorithms that attain mean-squared error within  $\frac{\text{polylog}(N)}{\sqrt{N}}$  of the MMSE [RV18, LW22], as well as generalizations to constant-rank spikes, and other priors [Mio17, MV21]. These aforementioned algorithms are based on AMP—one first computes the top eigenvector of  $M$  to obtain a vector that has non-trivial correlation with the spike  $x$ , then refines this estimate by iteratively applying the AMP update. To our knowledge, these algorithms based on AMP are the only algorithms that are known to obtain a mean-squared error within  $o(1)$  of the MMSE in the entire high SNR regime  $\lambda > 1$ .

Typically, one considers the inference setting where knowledge of the full generative model is assumed. Perhaps confusingly, this is also referred to as the “Bayes-optimal setting”. In [AKUZ19], they study a mismatched setting for spiked Wigner where the knowledge of  $\lambda$  is not assumed. In fact, in Section 3.2 we will show that the performance of their AMP algorithm corresponds with the fixed point equations for RGD.

In certain scenarios, Markov chain-based algorithms (or a sample from the posterior) might be preferred to AMP, since they have the ability to capture second-order information, like the variance of the correlation with the spike. Furthermore, samples can be used more flexibly to compute general posterior expectations.

**Dynamics of Markov chains before mixing.** Many physically and algorithmically interesting phenomena manifest only when Markov chain dynamics do *not* mix quickly, such as *aging* [CK93] and *metastability* [BDH16, BJ24]. In the context of *spin glasses*, physicists have made many predictions about the nonequilibrium behavior of natural Markov chains such as Langevin dynamics and Glauber dynamics. For the pure spherical  $p$ -spin glass,

the seminal work of Cugliandolo and Kurchan [CK93] derived a heuristic set of integro-differential equations which govern the behavior of Langevin dynamics when run for  $O(1)$  time. This was later rigorously proven for  $T = O(1)$  by [BDG06] and the dynamics of the energy observable was eventually established for  $T = \exp(O(N))$  by [Sel24b]. In recent work, [DGPZ25] rigorously prove that the energy and overlap of the sequential scan block dynamics for the SK model are dictated by an analogous set of integro-difference equations for  $O(N)$  timescales.

[LSS22] characterized the overlap of Langevin dynamics for spiked matrix posteriors in the large system limit  $N \rightarrow \infty$  for *fixed* times  $t$ . The analogous result in discrete time would correspond to asymptotically controlling the overlap for  $O(N)$  steps of Glauber dynamics. In contrast, we give control on any sufficiently large polynomial time scale for Glauber. [BGP24] give an extremely precise characterization of trajectory of Langevin diffusion for the posterior of the multi-spike tensor PCA problem given polynomially many samples of the spiked tensor (the spike being fixed, and the noise tensor i.i.d. across all samples). Concretely, in the rank-one matrix setting, they establish strong recovery of the spike  $x$  when the SNR  $\lambda = N^\alpha$  for any  $\alpha > 0$ . Their proofs are based on careful analysis of the SDE governing the correlations, which are inaccessible in the discrete setting.

**Locally stationary distributions.** Before a recent work of Liu, Mohanty, Raghavendra, Rajaraman, and Wu [LMR<sup>+</sup>24], it was not rigorously known whether Glauber dynamics on the spiked Wigner posterior achieves nontrivial recovery of the spike. As mentioned earlier in the introduction, they introduce the framework of locally stationary distributions to analyze the performance of slow-mixing Markov chains on inference problems such as spiked Wigner.

Our results directly improve on theirs in the following concrete ways:

- (i) We achieve tight temperature and SNR thresholds for our analysis, conditional on widely believed mixing conditions for the Sherrington–Kirkpatrick (SK) model.
- (i) Wherever SK mixing holds, we nail down the exact correlation that Glauber dynamics achieves for any SNR  $\lambda$ , as the solution to a simple fixed point equation. In particular, widely believed mixing conditions for the SK model recover the BBP transition  $\lambda > 1$ .
- (i) In particular, we unconditionally recover weak recovery in the regime  $\lambda > \frac{1}{0.295}$  for the success of Glauber, based on the rigorous SK mixing results of [AKV24].
- (i) As in Remark 3.1.6, we improve the runtime guarantees by an  $N^2$  factor.

Locally stationary distributions have since been used in other settings to understand the behavior of Markov chains before mixing. Notably, [BCV25] introduces a definition

of local stationarity for quantum Hamiltonians, and uses it to show that all metastable states satisfy certain structural guarantees. [RSS<sup>+</sup>25] also find an interesting application of locally stationary distributions in the context of stochastic backtracking algorithms for LLM reasoning.

## Organization

In Section 3.2, we provide a technical overview of our techniques to understand the behavior of RGD, and its connection to AMP. In Section 3.3, we list a few open questions that seem important to fleshing out our understanding of RGD and Glauber. In Section 3.4, we cover some basic preliminaries that will be useful.

In Section 3.5, we prove our main theorems about restricted Gaussian dynamics. Next, in Section 3.6, we establish novel high-precision estimates for the mean magnetization of a high-temperature SK model under weak external field, which are crucial for analyzing RGD. Finally, in Section 3.7, we describe some results analyzing the one-dimensional recursion governing the behavior of RGD.

## 3.2 Technical Overview

Our main results crucially depend on a very precise understanding of the restricted Gaussian dynamics. As it turns out, its study transparently clarifies the fixed point structure and its relationship to AMP. Using the machinery of locally stationary distributions [LMR<sup>+</sup>24, Lemma 3.7], we can then transfer our understanding of RGD to Glauber dynamics; see Theorem 3.5.2 for more details.

The technical overview is organized as follows. We will first define RGD and describe the high-level reduction showing why one should expect RGD to boil down to a one-dimensional recursion on correlation. Then, in Section 3.2, we show how spin glass machinery can be leveraged to explicitly understand what this one-dimensional recursion is. Once we have written down the explicit recursion, we can readily connect it to AMP; this is carried out in Section 3.2. Finally, we describe some qualitative properties of this recursion, such as the nature of its fixed points, in Section 3.2.

First off, observe that due to the rotational symmetry of the noise matrix and prior, we may assume without loss of generality that the spike is the all-ones vector  $\mathbf{1}$ . In this case, the scaled posterior (3.2) of the spiked Wigner model becomes

$$\mu_{\beta \mathbf{M}}(\sigma) \propto \exp\left(\frac{\beta}{2} \sigma^\top \mathbf{M} \sigma\right) = \exp\left(\frac{\beta}{2} \sigma^\top \mathbf{W} \sigma + \frac{\beta \lambda}{2N} \langle \sigma, \mathbf{1} \rangle^2\right).$$



We start by explaining where RGD arises from, and the connection between the posterior and the SK model. It arises from the Hubbard–Stratonovich transform [Str57, Hub59] (also see [LMRW24, Theorem 3.12]). Let  $\sigma \sim \mu_{\beta W + \frac{\beta\lambda}{N} \mathbf{1}\mathbf{1}^\top}$ , and  $\mathbf{g} \sim \mathcal{N}(0, 1)$ . Set  $z = \frac{\beta\lambda}{N} \langle \sigma, \mathbf{1} \rangle + \sqrt{\frac{\beta\lambda}{N}} \mathbf{g}$ . Then, the posterior distribution on  $\sigma$  conditioned on  $z$  is simply

$$\begin{aligned} \Pr[\sigma \mid z] &\propto \mu_{\beta M}(\sigma) \cdot \exp\left(-\frac{N}{2\beta\lambda} \left(z - \frac{\beta\lambda}{N} \langle \sigma, \mathbf{1} \rangle\right)^2\right) \\ &\propto \exp\left(\frac{\beta}{2} \sigma^\top \mathbf{W} \sigma + z \langle \sigma, \mathbf{1} \rangle\right) = \mu_{\beta W, z\mathbf{1}}, \end{aligned}$$

and the quadratic term from the spike in the Hamiltonian is now an external field. In other words,

$$\mu_{\beta W + \frac{\beta\lambda}{N} \mathbf{1}\mathbf{1}^\top} = \mathbf{E}_z \mu_{\beta W, z\mathbf{1}}.$$

The RGD Markov chain (Definition 3.1.3) is then simply the noising-denoising process associated to the above measure decomposition: if at  $\sigma \in \{\pm 1\}^N$ , draw  $\mathbf{g} \sim \mathcal{N}(0, 1)$ , then move to  $\sigma' \sim \mu_{\beta W, \left(\frac{\beta\lambda}{N} \langle \sigma, \mathbf{1} \rangle + \sqrt{\frac{\beta\lambda}{N}} \mathbf{g}\right) \mathbf{1}}$ . It is not difficult to show, given this perspective of a noising-denoising process, that RGD has stationary distribution equal to  $\mu_{\beta M}$ . The usefulness of RGD is that it brings forth the importance of the correlation with the spike, which is the observable we ultimately care about.

**Reduction to one dimension.** So far, we have written down the  $N$ -dimensional RGD chain, whose evolution explicitly depends on the correlation with the spike. To make RGD more tractable and eventually recover the AMP update, we would like to reduce the dynamics to a finite-dimensional one.

This motivates us to consider the related  $\mathcal{P}^{\text{RGD}}$  Markov chain on  $[-1, 1]$ , which tracks the (normalized) correlation of the  $N$ -dimensional RGD iterate with the spike. In this Markov chain, if starting at  $y \in \mathbb{R}$ , we draw  $\mathbf{g} \sim \mathcal{N}(0, 1)$ , set  $z = \beta\lambda y + \sqrt{\frac{\beta\lambda}{N}} \mathbf{g}$ , draw  $\sigma \sim \mu_{\beta W, z\mathbf{1}}$ , and finally move to  $z' = \frac{1}{N} \langle \sigma, \mathbf{1} \rangle$ . The stationary distribution of this Markov chain is just the (normalized) marginal of  $\mu_{\beta M}$  along the all-ones direction.

So far, this 1-dimensional chain still goes through an  $N$ -dimensional SK Gibbs sample  $\sigma$ ; in particular, it depends on the  $N \times N$  matrix  $\mathbf{W}$ . Therefore, to reduce  $\mathcal{P}^{\text{RGD}}$  to an effective 1-dimensional recursion, we must establish two things:

- Concentration of  $z'$  around the mean correlation. More explicitly, we would like to show that with high probability, the disorder  $\mathbf{W}$  is such that for all the encountered

external fields  $z$ , the correlation of the random draw  $\sigma$  is concentrated around the mean correlation  $\langle \sigma \rangle$ .

- A scalar formula for the mean correlation of a Gibbs sample from an SK model with external field.

Both of these points are tractable using spin glass machinery in the high temperature regime for SK, which we describe shortly. As we shall soon see, the scalar formula that governs the mean correlation is what crucially leads to state evolution equations.

## The mean recursion and fixed point structure

Let us now see more concretely how the fixed point relation arises from the analysis of  $\mathcal{P}^{\text{RGD}}$ .

Define the mean magnetization  $\mathbf{m}_h = \frac{1}{N} \mathbf{E}_{\sigma \sim \mu_{\beta W, h}} \langle \sigma, \mathbf{1} \rangle$ , and for  $\sigma \sim \mu_{\beta W, h}$ , let  $\tilde{\mathbf{g}}_h$  be the random variable  $\sqrt{N} \cdot (\frac{1}{N} \langle \sigma, \mathbf{1} \rangle - \mathbf{m}_h)$ . Under this notation, we can write down the following description of  $\mathcal{P}^{\text{RGD}}$ : from  $z \in \mathbb{R}$ ,

1. draw  $g \sim \mathcal{N}(0, 1)$ , set  $\mathbf{h} = \beta \lambda z + \sqrt{\frac{\beta \lambda}{N}} g$ ,
2. move to  $\mathbf{z}' = \mathbf{m}_h + \frac{1}{\sqrt{N}} \tilde{\mathbf{g}}_h$ .

For simplicity of exposition, let us first assume  $z = \Omega(1)$ ; we will comment on the case of small  $z$  later. First, observe by gaussian concentration, with very high probability we have  $\mathbf{h} \approx \beta \lambda z$ .

To simplify the recursion further, we need concentration of  $\mathbf{z}'$  around  $\mathbf{m}_h$ , as well as an explicit formula for  $\mathbf{m}_h$ . The key observation is that such properties are tractable to prove if the SK model  $\mu_{\beta W, h}$  is in the high temperature regime. We first need to introduce some standard spin glass quantities.

**Definition 3.4.11** (Overlap constant). For  $\beta \geq 0$  and  $h > 0$ , we define  $q = q_{\beta, h}$  as the unique solution to the recursion

$$q = \mathbf{E}_{g \sim \mathcal{N}(0, 1)} \left[ \tanh^2 (\beta g \sqrt{q} + h) \right].$$

We also define  $q_{\beta, h} = 0$  for  $h = 0$ .

We can now define the high-temperature region for the SK model, first predicted by de Almeida and Thouless [dAT78].<sup>2</sup>

---

<sup>2</sup>It is not rigorously known whether the AT line exactly characterizes the high temperature regime. However, this has been numerically verified, so in the interest of exposition we brush this issue aside for now.

**Definition 3.4.12** (AT line). A non-negative pair  $(\beta, h)$  is said to satisfy the Almeida–Thouless (AT) condition, or to be above the AT line, if

$$\beta^2 \mathbf{E} \operatorname{sech}^4 \left( \beta z \sqrt{q_{\beta,h}} + h \right) < 1. \quad (\text{AT})$$

We show that if (AT) is satisfied, the correlation  $\langle \sigma, 1 \rangle$  of a Gibbs sample  $\sigma \sim \mu_{\beta W, h}$  concentrates around the mean correlation. In other words, we have  $z' \approx m_h$  with high probability. Hence, we conclude that a single step of  $\mathcal{P}^{\text{RGD}}$  can be approximated by the map  $z \mapsto m_{\beta \lambda z}$ .

Turning now to the mean formula, in Lemma 3.6.1, we establish that with high probability over  $W$ , for all  $h = O(1)$  where (AT) is satisfied,

$$m_h \approx \mathbf{E}_{g \sim \mathcal{N}(0,1)} [\tanh(\beta g \sqrt{q_{\beta,h}} + h)], \quad (3.3)$$

where we recall  $q$  from Definition 3.4.11. In particular, we establish tighter control on  $m_h$  when  $h = o(1)$  than is implied by existing literature, which is crucial for analyzing the dynamics near  $h = 0$ .

Thus, we have reduced a single step of  $\mathcal{P}^{\text{RGD}}$  to the deterministic iteration

$$z_{t+1} \leftarrow \mathbf{E}_{g \sim \mathcal{N}(0,1)} [\tanh(\beta g \sqrt{q} + \beta \lambda z_t)],$$

where  $q = q_{\beta, \beta \lambda z_t}$ .

This suggests that the limiting behavior of  $\mathcal{P}^{\text{RGD}}$  is given by fixed point solutions to the following set of equations:

$$\begin{aligned} z &= \mathbf{E}_{g \sim \mathcal{N}(0,1)} [\tanh(\beta g \sqrt{q} + \beta \lambda z)] , \\ q &= \mathbf{E}_{g \sim \mathcal{N}(0,1)} [\tanh^2(\beta g \sqrt{q} + \beta \lambda z)] . \end{aligned} \quad (3.4)$$

Whether  $\mathcal{P}^{\text{RGD}}$  provably converges to these fixed points is not completely clear, and we defer further discussion on this matter to Section 3.2.

Let us now address the case of small  $z$ . When  $z = o(1)$ , and especially if  $z$  is  $\Theta(\frac{1}{\sqrt{N}})$ , the picture is more delicate, as then the random quantities fluctuate at the same scale as  $z$  itself. In Theorem 3.5.1 we use the anticoncentration of  $g$  to show that in  $\Omega(\log N)$  steps, we reach the region  $z = \Omega(1)$  with constant probability.

## RGD simulates AMP

Let us now draw the explicit connection to AMP fixed points. The state evolution fixed points (see Lemma 3.4.10 for a precise statement) that govern the late-stage performance of Bayes AMP for spiked Wigner [DAM16, MV21] read

$$\mu = \lambda \mathbf{E}_{g \sim \mathcal{N}(0,1)} [\tanh(\lambda g \sigma + \lambda^2 \sigma^2)]$$

$$\sigma^2 = \mathbf{E}_{g \sim \mathcal{N}(0,1)} [\tanh^2(\lambda g \sigma + \lambda^2 \sigma^2)].$$

Let  $(\mu, \sigma^2)$  be the unique fixed point to the SE equations. We claim that  $(z, q) = (\frac{\mu}{\lambda}, \sigma^2)$  is also a solution to (3.4) for  $\beta = \lambda$ . Indeed, using that  $\mathbf{E} \tanh(\sqrt{\gamma} g + \gamma) = \mathbf{E} \tanh^2(\sqrt{\gamma} g + \gamma)$  for any  $\gamma \geq 0$ , we see that  $\mu = \lambda \sigma^2$ . Therefore our putative assignment satisfies  $z = q$ , and some straightforward algebra yields that  $(z, q)$  is a valid solution to the fixed point equations.

As it turns out, the update equations for RGD given by (3.4) are nearly identical to the state evolution equations for mismatched AMP for the spiked Wigner model [AKUZ19]. Specifically, we have  $\Delta_0 = \frac{1}{\lambda^2}$ ,  $\Delta = \frac{1}{\beta \lambda}$ ,  $\eta_t(\cdot) = \tanh(\cdot)$ . Then some straightforward algebra, combined with [AKUZ19, Theorem 1], yields the SE recursion

$$\begin{aligned} M_{t+1} &= \mathbf{E} \left[ \tanh(\beta g \sqrt{Q_t} + \beta \lambda M_t) \right] \\ Q_{t+1} &= \mathbf{E} \left[ \tanh^2(\beta g \sqrt{Q_t} + \beta \lambda M_t) \right]. \end{aligned}$$

The direct comparison is just identifying  $M_t \leftarrow z_t$  and  $Q_t \leftarrow q_t$ . The only difference is that  $(M_t, Q_t)$  update synchronously, whereas our recursion sets  $q_t = q_t(z_t)$  and therefore collapses into an effective one-dimensional recursion on just  $z_t$ .

**Remark 3.2.1.** A subtle point here is that even for  $\beta = \lambda$ , the mismatched AMP is *not* syntactically equivalent to Bayes AMP; in order to derive Bayes AMP, one must artificially enforce that  $z = q$ .

## Qualitative behavior of the fixed point equations

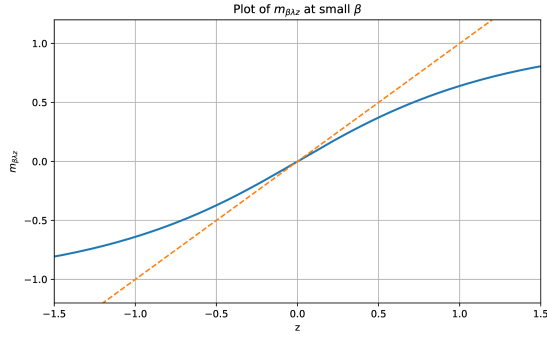
The qualitative behavior of the correlation recursion is determined by the fixed points of the recursion, as well as the convergence to said fixed points.

Convergence is nontrivial to guarantee, and indeed the proof of convergence for Bayes AMP does not immediately generalize to the mismatched setting where  $\beta \neq \lambda$ . To this end, in Proposition 3.7.1 we establish that so long as  $(\beta, \beta \lambda z)$  satisfies (AT), the map  $z \mapsto \mathbf{m}_{\beta \lambda z}$  is strictly increasing in  $z$ . For any  $\beta < 1$ , (AT) is automatically satisfied, and in this setting we are able to fully characterize convergence for the recursion starting from any  $z$ .

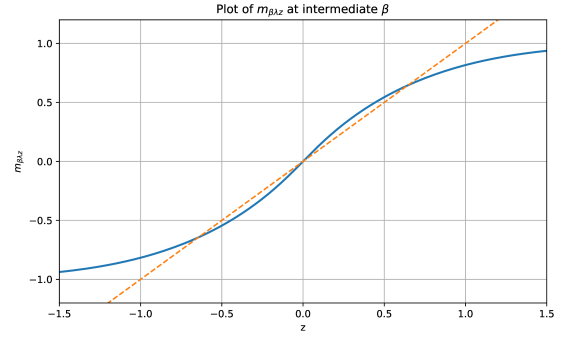
Turning now to the fixed points, one can immediately identify  $z = 0$  as one fixed point of this recursion. As it turns out, this map  $z \mapsto \mathbf{m}_{\beta \lambda z}$  behaves differently depending on  $\beta$ , as illustrated in Figure 3.2:

- (ii) If  $\beta$  is too small (smaller than  $\frac{1}{\lambda}$ ), the signal is washed out and 0 is the only fixed point of this recursion. Furthermore, it is a stable fixed point.

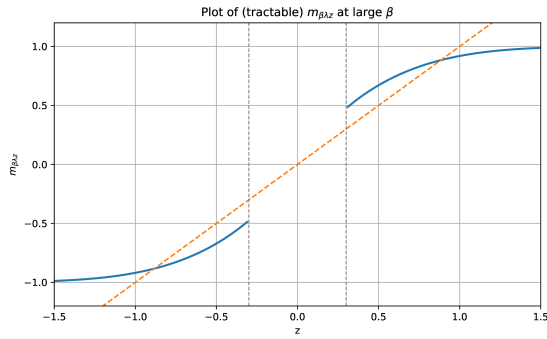
- (ii) If  $\beta$  is of intermediate magnitude (between  $\frac{1}{\lambda}$  and 1), in addition to an unstable fixed point at 0, this function has stable fixed points at  $\pm \text{OPT}_{\beta, \lambda}$ .
- (ii) If  $\beta > 1$ , the formula for the mean correlation  $m_{\beta \lambda z}$  is only valid if  $z$  is sufficiently large. Indeed, if  $z$  is close to 0, (AT) is violated, and the corresponding SK model is known to be in the low-temperature replica-symmetry breaking (RSB) regime. Hence, the recursion is only valid for large  $z$ , and in this region there are at most two stable fixed points  $\pm \text{OPT}_{\beta, \lambda}$ .
- (ii) Finally, if  $\beta$  is very large, the RSB region for  $z$  is so large that it swallows the fixed points that are present at smaller  $\beta$ .



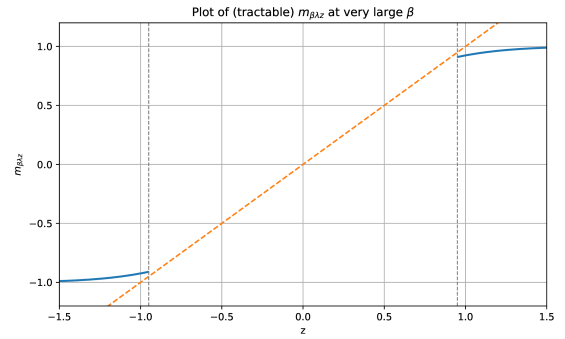
(a)  $m_{\beta \lambda z}$  at  $\beta < \frac{1}{\lambda}$ : 0 is the only fixed point.



(b)  $m_{\beta \lambda z}$  at intermediate  $\frac{1}{\lambda} < \beta < 1$ : non-trivial fixed points appear.



(c)  $m_{\beta \lambda z}$  at  $1 < \beta < \lambda$ : the 0 fixed point is swallowed by intractability.



(d)  $m_{\beta \lambda z}$  at  $\beta \gg \lambda$ : all fixed points are swallowed by intractability.

Figure 3.2: Behavior of the RGD recursion for different values of  $\beta$ .

We make the above predictions for the convergence and fixed points rigorous in Lemma 3.7.5. Figure 3.2 suggests two kinds of results one could hope to show.

- (a) First, that for any  $\beta$ , the RGD Markov chain reaches the stable  $\text{OPT}_{\beta,\lambda}$  fixed point from a sufficiently warm start which puts  $z_0$  in the high-temperature, replica-symmetric (RS) regime. That is, if the initialization  $z_0$  of  $\mathcal{P}^{\text{RGD}}$  is such that  $(\beta, \beta\lambda z_0)$  satisfies (AT), then the correlation shoots to the stable fixed point  $\text{OPT}_{\beta,\lambda}$  in a small constant number of steps of RGD. This is essentially immediate from classical analyses of fixed point recursions; see Lemma 3.5.4 for details.
- (b) Second, that for  $\beta < 1$ , the RGD Markov chain reaches the stable  $\text{OPT}_{\beta,\lambda}$  fixed point from an *arbitrary* initialization. In the case that  $\beta < \frac{1}{\lambda}$ , 0 is the only stable fixed point, so for now we assume that  $\beta \in (\frac{1}{\lambda}, 1)$ . Since the previous point implies that the correlation goes to  $\text{OPT}_{\beta,\lambda}$  once it becomes sufficiently large, the interesting part here is showing that the Markov chain successfully escapes the unstable fixed point at 0. We establish this in Lemma 3.5.6, where we show that the noise  $g$  and  $\tilde{g}$ , which are typically  $O\left(\frac{1}{\sqrt{N}}\right)$ , are anticoncentrated enough to cause the correlation to escape 0.

**Remark 3.2.2.** As a simple illustrative example, we can carry out the same analysis on the RGD dynamics for the Curie-Weiss model at inverse temperature  $\beta$  and external field  $h$ . In this case  $m_h$  can be computed exactly as  $\tanh(h)$ , and so we see that RGD converges to the stable fixed point of the mean-field recursion  $m = \tanh(\beta m + h)$ .

### 3.3 Conjectures and open questions

One natural question is whether we can extend our analysis to other inference problems. At least for the case of spiked matrix problems, the general recipe seems fairly adaptable. For example, consider a more general spiked matrix inference problem with a different prior and noise model (i.e. not GOE). The main obstacles here are (1) establishing replica symmetry (specifically overlap concentration) for this more general spin glass, and using it to show sample magnetization concentration à la Lemma 3.6.2 and (2) proving a mean correlation formula, analogous to Lemma 3.6.1.

**Question 3.3.1.** *Can one establish a kind of universality result for our fixed point characterization that does not rely on model-specific calculations?*

It would also be interesting to see if one could push the limits of our analysis to the multi-spike setting and achieve results analogous to [BGP24] for Glauber dynamics.

Another interesting (but open-ended) question is: what other classes of inference problems can be analyzed using locally stationary distributions? The success of RGD in the present analysis is rather contingent on the spiked matrix structure; in other inference problems, it is unclear what Markov chain should be used as a proxy to understand Glauber dynamics.

**Question 3.3.2.** *Identify broader inference problems that can be analyzed through the lens of locally stationary distributions.*

We next turn to open questions regarding the spiked Wigner problem in particular. One fundamental barrier to applying our techniques to understand Glauber for  $\beta > 1$  is that the transfer of local stationarity (Lemma 3.4.8) does not work in this setting, even if we assume mixing guarantees in the entire conjectured high-temperature region (AT). In particular, it is unclear how to prove a version of Lemma 3.4.8 where we only assume that the component measure satisfies an MLSI with high probability, under the assumption that the initialization from which the locally stationary distribution is obtained lives in this high-temperature region.

**Question 3.3.3.** *Assuming SK mixes fast for all  $(\beta, h)$  satisfying (AT), can we extend our analysis to characterize the recovery performance for annealed Glauber dynamics up to  $\beta \in (1, \beta_\lambda)$  for some  $\beta_\lambda > \lambda$ ?*

It seems likely that if the above question were answered affirmatively, this would provide an alternate algorithm that attains Bayes-optimal performance in the entire high-SNR regime  $\lambda > 1$ .

We note that [AKUZ19] makes some predictions for the behavior of these models in the RSB regime, i.e. beyond  $\beta_\lambda$ .

For technical reasons, the current techniques which involve a transfer of local stationarity between Glauber and RGD necessarily incur some  $\text{poly}(N)$  overhead in the runtime. On the other hand, based on physics heuristics and the rigorous results of [Sel24b, DGPZ25], one would expect the dynamics to achieve nonzero correlation in even  $O(N)$  steps.

**Conjecture 3.3.4.** *Glauber dynamics succeeds at weak recovery in (near-)linear runtime.*

One path towards achieving such a guarantee is to establish exponential contraction of the local stationarity parameter, perhaps by establishing a weak Poincaré inequality [HMRW25] for symmetric test functions. Indeed, we believe that our techniques to analyze RGD might help in showing that annealed Glauber dynamics succeeds at sampling from the posterior (and potentially even from lower-temperature models, up to the aforementioned threshold  $\beta_\lambda > \lambda$ ). Furthermore, while the “conservation of variance”

property used to show sampling guarantees for annealed Glauber dynamics is typically reliant on proving certain covariance bounds (e.g. [HMRW25, Lemma 6.8]) which are often technically challenging, perhaps this can be sidestepped if one instead directly tried to establish a weak Poincaré inequality for the (one-dimensional!) RGD Markov chain.

**Conjecture 3.3.5.** *For  $\beta \leq \beta_\lambda$ , annealed Glauber dynamics samples from the scaled posterior  $\mu_{\beta M}$  in  $\text{poly}(N)$ , or even  $O(N \log N)$  time.*

This threshold  $\beta_\lambda$  is defined as the largest temperature  $\beta$  at which there is a nonzero fixed point  $z$  to the recursion (3.4) such that  $(\beta, \beta\lambda z)$  satisfies (AT). As shown in Proposition 3.7.1, this threshold is in fact greater than  $\lambda$ .

## 3.4 Preliminaries

### Notation

- Given  $\sigma_1, \sigma_2 \in \{\pm 1\}^N$ , we denote  $R(\sigma_1, \sigma_2) = \frac{1}{N} \langle \sigma_1, \sigma_2 \rangle$  to be the normalized correlation between the two.
- We use  $c$  to denote small positive constants whose value may change from line to line, and  $C$  to denote similarly fickle large constants.

### Markov chain basics

**Definition 3.4.1** (Ising model). Let  $J \in \mathbb{R}^{n \times n}$  be a symmetric *interaction matrix* and  $h \in \mathbb{R}^n$  an *external field*. The *Ising model* corresponding to  $J$  and  $h$  is the probability distribution  $\mu_{J,h}$  on  $\{\pm 1\}^n$ , where

$$\mu_{J,h}(x) \propto \exp\left(\frac{1}{2}x^\top Jx + \langle h, x \rangle\right).$$

We drop the  $h$  from the subscript when it is equal to 0.

We also recall the standard definitions of functional inequalities which imply rapid mixing. For a reversible Markov chain  $P$  on state space  $\Omega$  with stationary distribution  $\pi$ , and any functions  $f, g : \Omega \rightarrow \mathbb{R}$ , define the *Dirichlet form*  $\mathcal{E}(f, g) := \mathbf{E}_{x \sim \pi} \mathbf{E}_{y \sim P_x}(f(x) - f(y))(g(x) - g(y))$ .

**Definition 3.4.2** (Poincaré inequality). We say  $P$  satisfies a *Poincaré inequality* with constant  $C$  if for any function  $f : \Omega \rightarrow \mathbb{R}$ :

$$\mathcal{E}(f, f) \geq C \cdot \text{Var}[f].$$



**Definition 3.4.3** (Modified log-Sobolev inequality). We say  $P$  satisfies a *modified log-Sobolev inequality* (MLSI) with constant  $C$  if for any function  $f : \Omega \rightarrow \mathbb{R}_{>0}$ ,

$$\mathcal{E}_P(f, \log f) \geq C \cdot \text{Ent}[f],$$

where  $\text{Ent}[f] := \mathbf{E}_\pi[f \log f] - \mathbf{E}_\pi f \log \mathbf{E}_\pi f$ .

## Locally stationary distributions

We now give the relevant definitions for locally stationary distributions. Following [LMR<sup>+</sup>24, Definition 1.2], we use the following notion of local stationarity.

**Definition 3.4.4** (Locally stationary distribution). Let  $P$  be a time-reversible ergodic Markov chain on a finite state space  $\Omega$ , with associated stationary distribution  $\pi$ . A probability measure  $\nu$  on  $\Omega$  is said to be  $\varepsilon$ -*locally stationary* with respect to  $P$  if, with  $f = \frac{d\nu}{d\pi}$ ,

$$\mathcal{E}(f, \log f) := \sum_{x, y \in \Omega} P_{x \rightarrow y} (f(x) - f(y)) \cdot \log \frac{f(x)}{f(y)} \leq \varepsilon.$$

One can also show that at a typical time  $t$ , *any* reversible Markov chain yields a locally stationary distribution.

**Lemma 3.4.5.** Let  $\nu_0$  be an arbitrary distribution,  $\nu_t = P^t \nu_0$ , and set  $f_t = \frac{d\nu_t}{d\pi}$ . Also set  $f = \mathbf{E}_{t \sim [0, T]} f_t$ , and  $\nu := \mathbf{E}_{t \sim [0, T]} \nu_t$ . Then,

$$\mathcal{E}(f, \log f) \leq \frac{\text{KL}(\nu_0 \parallel \pi)}{T} \leq \frac{\log(1/\pi_{\min})}{T}.$$

In particular,  $\nu$  is  $\frac{\log(1/\pi_{\min})}{T}$ -locally stationary with respect to  $\pi$ .

*Proof.* This immediately follows from [LMR<sup>+</sup>24, Lemma 3.1] (also see [BCV25, Lemma II.1]), coupled with the observation that the Dirichlet form  $f \mapsto \mathcal{E}(f, \log f)$  is convex.  $\square$

We will also need to use the fact that a locally stationary distribution does not change much in TV distance if we apply  $P$  for a few steps.

The following TV-stability lemma is essentially [LMR<sup>+</sup>24, Lemma 3.2].

**Lemma 3.4.6** (TV stability). Let  $P$  be a reversible Markov chain with stationary distribution  $\pi$ ,  $\nu$  an arbitrary distribution which is  $\varepsilon$ -locally stationary. Then,

$$d_{\text{TV}}(\nu, P^T \nu) \leq \frac{T \sqrt{\varepsilon}}{2}.$$

*Proof.* When  $T = 1$ , this is [LMR<sup>+</sup>24, Lemma 3.2]. The data processing inequality implies that for any  $k \geq 1$ ,  $d_{\text{TV}}(P^k \nu, P^{k+1} \nu) \leq \frac{\sqrt{\varepsilon}}{2}$ . The claim follows on applying the triangle inequality.  $\square$

**Corollary 3.4.7.** *Let  $\phi : \Omega \rightarrow \mathbb{R}$  be a bounded function on the state space of a Markov chain  $P$ , and  $\nu$  an  $\varepsilon$ -locally stationary measure. Then,*

$$|\mathbf{E}_{x \sim \nu}[\phi(x)] - \mathbf{E}_{x \sim P^T \nu}[\phi(x)]| \leq \|\phi\|_{\infty} \cdot T \sqrt{\varepsilon}.$$

We will also need to transfer local stationarity between Glauber and RGD, using [LMR<sup>+</sup>24, Lemma 3.7].

**Lemma 3.4.8** (Local stationarity transfer). *Let  $P$  be the Markov chain associated to a measure decomposition  $\pi = \mathbf{E}_{z \sim \rho} \pi_z$ . Let  $f : \{\pm 1\}^n \rightarrow \mathbb{R}_{>0}$  be any function and set  $\tau$  such that  $\min_{x \in \{\pm 1\}^n} f(x) > \exp(-\tau)$  or  $\max_{x \in \{\pm 1\}^n} f(x) < \exp(\tau)$ . For  $\delta := \inf_z \rho_{\text{MLSI}}(\pi_z)$ , we have*

$$\mathcal{E}_P(f, \log f) \leq O\left(\frac{\tau}{\delta}\right) \cdot \mathcal{E}_{\pi}(f, \log f).$$

## Approximate Message Passing and State Evolution

For simplicity, we only recall the specific version of AMP used for spiked Wigner [DAM16].

**Definition 3.4.9.** Set  $m^{-1} = x^0 = 0$ , and for  $t \geq 0$  we update

$$\begin{aligned} m^t &= \tanh(x^t), \\ \mathbf{b}_t &= \frac{\lambda^2}{N} \sum_{i=1}^N (1 - \tanh^2(x_i^t)) \\ x^{t+1} &= \lambda \mathbf{M} m^t - \mathbf{b}_t m^{t-1} \end{aligned}$$

where  $\tanh(\cdot)$  is applied entrywise.

State evolution [BM11] predicts the empirical distribution of the coordinates of  $x^t$ . We record the special case relevant to our setting: the correlation of the AMP iterate with the planted signal  $x$  can be predicted by a scalar recursion.

**Lemma 3.4.10** (State evolution for correlation). *Fix a time  $t \geq 0$ . Let  $(\mu_t, \sigma_t)$  be defined by the scalar recursion initialized with  $(\mu_0, \sigma_0) = (0, 0)$ :*

$$\begin{aligned}\mu_{t+1} &= \lambda \mathbf{E}_{g \sim \mathcal{N}(0,1)} [\tanh(\lambda g \sigma_t + \lambda^2 \sigma_t^2)] \\ \sigma_{t+1}^2 &= \mathbf{E}_{g \sim \mathcal{N}(0,1)} [\tanh^2(\lambda g \sigma_t + \lambda^2 \sigma_t^2)].\end{aligned}$$

Let  $X_0 \sim \text{Unif}\{\pm 1\}$ ,  $Z_0 \sim \mathcal{N}(0, 1)$ . As  $N \rightarrow \infty$ , we have the following convergence in  $W_2$ :

$$\frac{1}{N} \langle m^t, \mathbf{x} \rangle \rightarrow \mathbf{E}[X_0(\mu_t X_0 + \sigma_t Z_0)].$$

## The Sherrington–Kirkpatrick model

In this subsection, we introduce several important parameters in the SK model.

**Definition 3.4.11** (Overlap constant). For  $\beta \geq 0$  and  $h > 0$ , we define  $q = q_{\beta,h}$  as the unique solution to the recursion

$$q = \mathbf{E}_{g \sim \mathcal{N}(0,1)} [\tanh^2(\beta g \sqrt{q} + h)].$$

We also define  $q_{\beta,h} = 0$  for  $h = 0$ .

It is known that in the high-temperature replica-symmetric phase  $\beta < 1$ , the overlap between two samples from the corresponding SK model concentrates around the above number.

**Definition 3.4.12** (AT line). A non-negative pair  $(\beta, h)$  is said to satisfy the Almeida–Thouless (AT) condition, or to be above the AT line, if

$$\beta^2 \mathbf{E} \text{sech}^4(\beta z \sqrt{q_{\beta,h}} + h) < 1. \quad (\text{AT})$$

Observe that if  $\beta < 1$ ,  $(\beta, h)$  satisfies (AT) for any  $h \in \mathbb{R}_{\geq 0}$ . A very important question in the study of the SK model is when the model is replica-symmetric: when does the overlap between two independent samples from the Gibbs measure concentrate?

**Definition 3.4.13.** We say that  $(\beta, h)$  satisfies *exponential overlap concentration* if for some constant  $C$  independent of  $N$ , with high probability over  $\mathbf{W} \sim \text{GOE}(N)$ ,

$$\mathbf{E}_{\sigma^1, \sigma^2 \sim \mu_{\beta \mathbf{W}, h1}} \left[ \exp \left( \frac{N \left( \frac{1}{N} \langle \sigma^1, \sigma^2 \rangle - q_{\beta,h} \right)^2}{C} \right) \right] \leq 2. \quad (\text{overlap-conc})$$

It is believed that the replica-symmetric region coincides with the set of  $(\beta, h)$  satisfying (AT).

**Condition 2 (RS-AT).**  $\beta$  is said to satisfy (RS-AT) if the following holds with high probability over  $W$ . Let  $h_* = \inf\{h \geq 0 : (\beta, h) \text{ satisfies (AT)}\}$ . Then for all  $h > h_*$ ,  $\mu_{\beta W, h}$  satisfies (overlap-conc).

**Remark 3.4.14.** At a high level, this condition is saying that being above the AT line implies a strong form of replica symmetry, and that the AT region is monotonic in  $h$ . The latter condition has been numerically verified, but the authors are unaware of a rigorous statement of this form.

**Remark 3.4.15.** Significant progress has been made towards establishing that for the SK model,  $(\beta, h)$  satisfies (AT) if and only if the corresponding model is replica-symmetric. On the negative side, it was shown by Toninelli [Ton02] that if  $(\beta, h)$  does *not* satisfy (AT), the corresponding model exhibits replica symmetry breaking. On the positive side, it was shown by Talagrand [Tal11, Theorem 13.7.1] that exponential overlap concentration holds for a very large subregion of the area above the AT line, which is conjectured to coincide with the entirety of the region above the AT line. Making this conjecture rigorous requires proving an explicit numerical condition. We refrain from elaborating as the details are not directly relevant to our work, and instead refer the interested reader to Talagrand’s text, as well as the discussion in [JT17, p.8]. We remark that it is known that Talagrand’s subregion contains the entire high-temperature regime  $\beta < 1$ , see e.g. [BXY23, p.3] and [Che21].

## 3.5 The dynamics of RGD

The goal of this section will be to understand the trajectory of RGD—we will establish that RGD rapidly converges to fixed points of the one-dimensional recursion of interest. Recall from the technical overview that for  $\lambda > 1$  and  $\beta > 0$ , this recursion is given by the update function  $f_\beta$  defined by

$$f_{\beta, \lambda}(z) = \mathbf{E} \left[ \tanh \left( \beta g \sqrt{q_{\beta, \beta \lambda z}} + \beta \lambda z \right) \right]. \quad (3.5)$$

When  $\lambda$  is clear from context, we refer to  $f_{\beta, \lambda}$  as simply  $f_\beta$ .

Often in this section, we will make the following assumption about the structure of  $f_\beta$ .

**Condition 3 (AT-fixed-pt).** For  $\lambda > 1$  and  $\beta > 0$ , we say that  $(\beta, \lambda)$  satisfies (AT-fixed-pt) if there exists a fixed point  $\text{OPT}_{\beta, \lambda}$  of  $f_{\beta, \lambda}$  such that  $(\beta, \beta \lambda \text{OPT}_{\beta, \lambda})$  satisfies (AT).

This condition has some direct consequences about the recursion, which we state below in Lemma 3.5.3. One might wonder when (AT-fixed-pt) holds; in fact from the definition of the AT line one immediately concludes that it holds for  $\beta < 1$ . In Lemma 3.7.6, we are also able to show that it holds at the posterior temperature where  $\beta = \lambda$ . It is an interesting open question whether one can rigorously prove that (AT-fixed-pt) holds for  $\beta \in (1, \lambda)$ .

Let us now state the main result for this section.

**Theorem 3.5.1** (Fixed-temperature RGD). *Fix  $\lambda > 1$ ,  $K > 0$  large,  $\varepsilon > 0$  small, and  $\beta > \frac{1}{\lambda}$ . With high probability over the noise  $\mathbf{W}$ , the following holds. Let  $x_0 \in \{\pm 1\}^N$ , and assume that  $(\beta, \beta\lambda|R(x_0, 1)|)$  satisfies (AT),  $\beta$  satisfies (RS-AT), and  $(\beta, \lambda)$  satisfies (AT-fixed-pt). Suppose that we run the RGD Markov chain at inverse temperature  $\beta$  from  $x_0$  for  $T \geq \omega(\log N)$  steps to arrive at distribution  $\nu_T$ . Suppose  $T \leq N^K$ . Then,*

$$\mathbf{E}_{x_T \sim \nu_T} [||R(x_T, 1)| - \text{OPT}_{\beta, \lambda}|] \leq O\left(\frac{1}{N^{1/2-\varepsilon}}\right).$$

Combining Theorem 3.5.1 with the properties of locally stationary distributions, we obtain the following control over the recovery performance of Glauber when  $\beta < 1$ .

**Theorem 3.5.2** (Fixed-temperature Glauber). *Let  $\beta \in (\frac{1}{\lambda}, 1)$  satisfy (MLSI) and  $\varepsilon > 0$ . With high probability over  $\mathbf{W}$ , the following holds. Let  $\sigma_0 \in \{\pm 1\}^N$  be arbitrary, and let  $(\sigma_t)_{t \geq 0}$  be the trajectory of the Glauber dynamics run with stationary distribution  $\mu_{\beta M}$  initialized at  $\sigma_0$ . Let  $T \geq \tilde{\Omega}(N^{4-2\varepsilon})$ , and  $\widehat{\sigma} = \sigma_t$  for a uniformly random  $t \sim [0, T]$ . Then, there exists an explicit constant  $\text{OPT}_{\beta, \lambda} > 0$  such that*

$$\Pr \left[ \left| \frac{1}{N} |\langle \widehat{\sigma}, x \rangle| - \text{OPT}_{\beta, \lambda} \right| > \frac{C}{N^{1/2-\varepsilon}} \right] = o_N(1).$$

*Proof.* Let us condition on the very high probability event that  $\|\mathbf{W}\|_{\text{op}} = O(1)$ . Let  $\nu$  denote the distribution  $\mathbf{E}_{t \sim [0, T]} \nu_t$ , and  $f$  its relative density with respect to  $\mu_{\beta M}$ . By Lemma 3.4.5, we have that  $\mathcal{E}(f, \log f) \leq O(\frac{N}{T})$ . Applying Lemma 3.4.8 yields that  $\mathcal{E}_{\text{RGD}}(f, \log f) \leq O(\frac{\tau}{\delta} \cdot \frac{N}{T})$ . We can crudely bound  $\tau = O(N)$  because  $\|\beta M\|_{\text{op}} = O(1)$ , and since we assumed that  $\beta$  satisfies (MLSI), we have  $\delta = \Omega(\frac{1}{N})$ , yielding that

$$\mathcal{E}_{\text{RGD}}(f, \log f) \leq O\left(\frac{N^3}{T}\right).$$

We can now apply Corollary 3.4.7 to  $\nu$  with the test function  $\phi(x) = ||R(x, \mathbf{1})| - \text{OPT}_{\beta, \lambda}|$ , to conclude that

$$\left| \mathbf{E}_{\nu} [||R(x, \mathbf{1})| - \text{OPT}_{\beta, \lambda}|] - \mathbf{E}_{(\mathcal{P}^{\text{RGD}})^{\text{C log } N_{\nu}}} [||R(x, \mathbf{1})| - \text{OPT}_{\beta, \lambda}|] \right| \leq \tilde{O}\left(\sqrt{\frac{N^3}{T}}\right).$$

Now since  $\beta < 1$ , all three of the conditions (AT), (RS-AT), and (AT-fixed-pt) are known to hold. Therefore Theorem 3.5.1 yields the appropriate control on the latter expectation, and we conclude by applying Markov.  $\square$

We shall prove Theorem 3.5.1 in two steps. First, in Lemma 3.5.4, we shall prove the claim assuming a warm start initialization where  $|R(x_0, \mathbf{1})| = \Omega(1)$ ; such a warm start is necessary for  $\beta > 1$  since the AT condition is nontrivial there. We then show in Lemma 3.5.6 that when  $\beta < 1$ , RGD escapes the unstable fixed point where  $R(x_0, \mathbf{1}) = 0$ .

For use later in this section, we also note the following consequences of (AT-fixed-pt), proved in Section 3.7.

**Lemma 3.5.3.** *Let  $\lambda > 1$ . If  $\beta < 1$  or  $\beta = \lambda$ ,  $(\beta, \lambda)$  satisfies (AT-fixed-pt). Suppose that  $(\beta, \lambda)$  satisfies (AT-fixed-pt). Then,*

- (iii) *if  $h$  is such that  $(\beta, \beta\lambda h)$  satisfies (AT),  $f'_{\beta, \lambda}(h) > 0$ ,*
- (iii)  *$\text{OPT}_{\beta, \lambda}$  is the unique fixed point  $h$  of  $f_{\beta, \lambda}$  such that  $(\beta, \beta\lambda h)$  satisfies (AT), and*
- (iii)  *$\text{OPT}_{\beta, \lambda}$  is a stable fixed point of  $f_{\beta, \lambda}$ .*

## RGD from a warm start reaches the stable fixed point

In this section, we will show that running RGD on the posterior of spiked Wigner attains the Gibbs correlation, provided that the Markov chain is initialized at a point (or distribution) with typically large correlation. Our analysis will require RGD to be run for at least a sufficiently large constant number of steps, but the analysis extends to any large polynomial number of steps. In contrast, existing nonasymptotic analyses of AMP only work for sublinearly many steps.

**Lemma 3.5.4.** *Fix  $\lambda > 1$ ,  $K > 0$  large,  $\varepsilon > 0$  small, and  $\beta > \frac{1}{\lambda}$ . With high probability over the noise  $\mathbf{W}$ , the following holds. Let  $x_0 \in \{\pm 1\}^n$ , and assume that  $(\beta, \beta\lambda|R(x_0, \mathbf{1})|)$  satisfies (AT), and  $|R(x_0, \mathbf{1})| = \Omega(1)$ . Also assume that  $\beta$  satisfies (RS-AT) and  $(\beta, \lambda)$  satisfies (AT-fixed-pt).<sup>3</sup> Suppose*

---

<sup>3</sup>Here and below, satisfying (AT) means we satisfy the AT line inequality with some constant gap

that we run the RGD Markov chain at inverse temperature  $\beta$  from  $x_0$  for  $T \geq \omega(\log N)$  steps to arrive at distribution  $v_T$ . Suppose  $T \leq N^K$ . Then,

$$\mathbf{E}_{x_T \sim v_T} [|R(x_T, \mathbf{1})| - \text{OPT}_{\beta, \lambda}] \leq O\left(\frac{1}{N^{1/2-\varepsilon}}\right).$$

**Remark 3.5.5.** If one only desired to have correlation within a(n arbitrarily small) constant of  $\text{OPT}_{\beta}$ , our proof (specifically (3.7)) will show that it suffices to run RGD for a (large) constant number of steps.

*Proof.* For ease of notation, let  $c > 0$  be a sufficiently small constant such that  $(\beta, \beta\lambda c)$  satisfies (AT), and let

$$\mathcal{Z} = \{z \in [-2, 2] : (\beta, \beta\lambda z) \text{ satisfies (AT) and } |z| > c\}.$$

Observe that  $\mathcal{Z}$  contains  $R(x_0, \mathbf{1})$ . Consider the one-dimensional projection of the RGD walk, initialized at  $z_0 = R(x_0, \mathbf{1})$ . One step of this one-dimensional walk is defined by walking from  $z_t$  as

$$z_{t+1} = R(x_{t+1}, \mathbf{1}), \text{ where } x_{t+1} \sim \mu_{\beta W, \beta\lambda z_t + \sqrt{\frac{\beta\lambda}{n}} g_t},$$

where each  $g_t$  is iid drawn from  $\mathcal{N}(0, 1)$ . Let us condition on the following high-probability events:

- (v) For every  $t < T$ ,  $|g_t| < N^\varepsilon$ . By standard Gaussian tail bounds, this occurs with probability  $T \cdot e^{-\Omega(N^\varepsilon)} = e^{-cN^\varepsilon}$ .
- (v) For all external fields  $h$  with  $h/(\beta\lambda) \in \mathcal{Z}$  and constant  $k$ ,

$$\mathbf{E}_{\sigma \sim \mu_{\beta W, h}} \left( R(\sigma, \mathbf{1}) - \mathbf{E} \left[ \tanh \left( \beta \sqrt{q_{\beta, h}} g + h \right) \right] \right)^{2k} < \frac{1}{N^k}.$$

This follows from Lemma 3.6.2, along with the assumption that  $\beta$  satisfies (RS-AT).

We claim that, on the above events, for all  $t < T$ ,  $\beta\lambda z_t$  is bounded away from the AT line by a constant. The base case  $t = 0$  is true by assumption. Next, since  $\beta\lambda z_t$  is bounded away from the AT line by a constant, (i) implies that  $\beta\lambda z_t + \sqrt{\frac{\beta\lambda}{n}} g_t$  does so as well. (ii) then implies that with high probability, for all  $t < T$ ,

$$\left| z_{t+1} - f_\beta \left( z_t + \sqrt{\frac{1}{\beta\lambda N}} g_t \right) \right| \leq O\left(\frac{1}{N^{1/2-\varepsilon}}\right).$$

The Lipschitzness of  $f_{\beta,\lambda}$  with (i) implies that

$$|z_{t+1} - f_{\beta}(z_t)| \leq O\left(\frac{1}{N^{1/2-\varepsilon}}\right). \quad (3.6)$$

Note in particular that by the monotonicity assumption in (RS-AT), if the above is true, then  $\beta\lambda z_{t+1}$  also satisfies (AT) with a constant gap. Indeed, by Lemma 3.5.3(i), if  $z_t < \text{OPT}_{\beta,\lambda}$ , we have  $f_{\beta}(z_t) > z_t$  so it satisfies (AT). On the other hand, if  $z_t > \text{OPT}_{\beta,\lambda}$ , then  $f_{\beta}(z_t) > \text{OPT}_{\beta,\lambda}$ , so it satisfies (AT) due to (AT-fixed-pt). As a result, we conclude that for all  $t < T$ , using  $f_{\beta}$  is valid as an approximation to the RGD update at step  $t$ .

Next, we show that  $z_t$  converges to a small constant interval around  $\text{OPT}_{\beta}$  in a constant number of steps. For starters, since we have assumed (AT-fixed-pt), Lemma 3.5.3(iii) implies that  $\text{OPT} = \text{OPT}_{\beta,\lambda}$  is the unique nonzero fixed point of  $f_{\beta}$  in the AT region, and that it is stable. Let  $\eta > 0$  such that  $f'_{\beta}$  is bounded by some  $L < 1$  on  $[\text{OPT} - \eta, \text{OPT} + \eta]$ .

There exists a large constant  $C$  and random variables  $(\delta_t)_{0 \leq t \leq T}$  such that  $\delta_t < \frac{C}{N^{1/2-\varepsilon}}$  almost surely for all  $t < T$  and  $z_{t+1} = f_{\beta}(z_t + \delta_t)$ . We claim that after some constant number of steps  $\tau$ ,  $z_{\tau} \in [\text{OPT} - \eta/2, \text{OPT} + \eta/2]$ . Indeed, set

$$\text{gap} = \inf_{z \in \mathcal{Z} \setminus [\text{OPT} - \eta/2, \text{OPT} + \eta/2]} |f_{\beta}(z) - z|,$$

a constant bounded away from 0 due to the definition of  $\mathcal{Z}$  and the boundedness of  $f'_{\beta}$  near  $\text{OPT}$ . Then, while  $z_t \in \mathcal{Z} \setminus [\text{OPT} - \eta/2, \text{OPT} + \eta/2]$ , since  $f_{\beta}$  is increasing (Lemma 3.5.3(i)), we have

$$|z_{t+1} - \text{OPT}| < |z_t - \text{OPT}| - \text{gap}.$$

Furthermore, once  $z_t$  reaches  $[\text{OPT} - \eta/2, \text{OPT} + \eta/2]$ , because  $\delta_t = o(1)$ , it is easy to check that  $f'_{\beta}(z_t) < L$  for all future times  $t > \tau$ . For all such times, we have

$$|z_{t+1} - \text{OPT}| = |f_{\beta}(z_t + \delta_t) - f_{\beta}(\text{OPT})| \leq L|z_t + \delta_t - \text{OPT}| \leq L|z_t - \text{OPT}| + L \cdot \frac{C}{N^{1/2-\varepsilon}}.$$

Iterating this inequality yields that

$$|z_{\tau+t} - \text{OPT}| \leq L^t |z_{\tau} - \text{OPT}| + \frac{L}{1-L} \cdot \frac{C}{N^{1/2-\varepsilon}}. \quad (3.7)$$

It follows that once we set  $T = \Omega(\log N)$ ,

$$|z_T - \text{OPT}| \leq O\left(\frac{1}{N^{1/2-\varepsilon}}\right)$$

as desired. □



## High-temperature RGD escapes the unstable fixed point

This section is dedicated to proving the following lemma about RGD escaping from the trivial fixed point at 0 correlation.

**Lemma 3.5.6.** *Let  $\lambda > 1$ ,  $\beta \in (\frac{1}{\lambda}, 1)$ , and  $\varepsilon > 0$  be a sufficiently small constant. Let  $x_0 \in \{\pm 1\}^N$  arbitrary, and  $(x_t)_{t \geq 0}$  the (random) trajectory of RGD initialized at  $x_0$ . Then,*

$$\Pr \left[ \max_{0 \leq t \leq T} \frac{1}{N} |\langle x_t, \mathbf{1} \rangle| \leq \varepsilon \right] \leq e^{-\Omega(T/\log N)}.$$

Lemma 3.5.6 almost immediately follows from the following more general lemma about appropriate stochastic processes escaping unstable fixed points.

**Lemma 3.5.7.** *Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be a smooth function such that for some constants  $L > 1$  and  $\varepsilon > 0$ ,  $|g(x)| > L|x|$  for all  $|x| < 2\varepsilon$ . Let  $X_0 \sim \nu_0$  for some arbitrary distribution  $\nu_0$  on  $\mathbb{R}$ . Suppose we have three sequences  $(X_t, Y_t, Z_t)_{t \geq 0}$  of random variables on a common probability space. Assume that*

(vii) *for all  $t \geq 0$ ,  $Z_t$  has law  $\mathcal{N}(0, \Theta(\frac{1}{N}))$ , and is independent of  $\sigma((X_s)_{0 \leq s \leq t-1}, (Y_s)_{0 \leq s \leq t})$ ,*

(vii)  $\mathbf{E}[Y_t^2 \mid (X_s, Y_s, Z_s)_{0 \leq s \leq t-1}] \leq \frac{C}{N}$ .

(vii)  $(X_t)$  is defined by

$$X_{t+1} = g(X_t) + Y_t + Z_t.$$

Then,

$$\Pr \left[ \max_{0 \leq t \leq T} |X_t| \leq \varepsilon \right] \leq e^{-\Omega(T/\log N)}.$$

*Proof.* Our proof proceeds in two stages. We first show that the dynamics escape zero to a “lukewarm” start, and after this exponential growth kicks in to make  $X_t$  grow to  $\Omega(1)$ .

To be more precise, let  $\kappa > 0$  be a (large) constant to be fixed later. We shall show the following two statements, where  $c > 0$  is a sufficiently small constant:

$$\Pr \left[ |X_{t+1}| \geq \frac{\kappa}{\sqrt{N}} \mid (X_s, Y_s, Z_s)_{0 \leq s \leq t} \right] \geq c, \quad (3.8)$$

$$\Pr \left[ \max_{t \leq s \leq t+C \log N} |X_s| \geq \varepsilon \mid (X_s, Y_s, Z_s)_{0 \leq s \leq t-1}, |X_t| \geq \frac{\kappa}{\sqrt{N}} \right] \geq c. \quad (3.9)$$

Putting these two together with the independence structure of the  $(Y_s, Z_s)$  immediately yields the result.

Let us start by proving (3.8), which is a consequence of anticoncentration and concentration of  $Z_t$ . Indeed, we have

$$\begin{aligned} \Pr \left[ |X_{t+1}| \geq \frac{\kappa}{\sqrt{N}} \mid (X_s, Y_s)_{0 \leq s \leq t}, (Z_s)_{0 \leq s \leq t-1}, |g(X_t) + Y_t| \geq \frac{2\kappa}{\sqrt{N}} \right] \\ \geq \Pr \left[ |Z_t| \leq \frac{\kappa}{\sqrt{N}} \mid (X_s, Y_s)_{0 \leq s \leq t}, (Z_s)_{0 \leq s \leq t-1}, |g(X_t) + Y_t| \geq \frac{2\kappa}{\sqrt{N}} \right] \geq c, \text{ and} \\ \Pr \left[ |X_{t+1}| \geq \frac{\kappa}{\sqrt{N}} \mid (X_s, Y_s)_{0 \leq s \leq t}, (Z_s)_{0 \leq s \leq t-1}, |g(X_t) + Y_t| \leq \frac{2\kappa}{\sqrt{N}} \right] \\ \geq \Pr \left[ |Z_t| \geq \frac{3\kappa}{\sqrt{N}} \mid (X_s, Y_s)_{0 \leq s \leq t}, (Z_s)_{0 \leq s \leq t-1}, |g(X_t) + Y_t| \leq \frac{2\kappa}{\sqrt{N}} \right] \geq c, \end{aligned}$$

implying (3.8).

Let us next prove (3.9). Consider the following event, parametrized by some constant  $\iota > 0$  that we shall fix later:

$$\mathcal{E} = \left\{ (Y_s, Z_s)_{s \geq t} : |Y_s + Z_s| \leq \frac{\iota \kappa}{\sqrt{N}} \cdot (s - t + 1) \right\}.$$

Observe that if  $\iota \kappa$  is sufficiently large, the above holds with nonzero probability. Indeed, we have  $\mathbf{E} \left[ (Y_t + Z_t)^2 \mid (Y_s, Z_s)_{0 \leq s \leq t-1} \right] \leq \frac{C}{N}$ , so Markov's inequality followed by a union bound gives that

$$\Pr [\mathcal{E} \mid (X_s, Y_s, Z_s)_{0 \leq s \leq t-1}] \geq 1 - \sum_{j \geq 0} \frac{C}{\iota^2 \kappa^2 (j+1)^2}, \quad (3.10)$$

which is bounded away from 0 for  $\iota \kappa$  sufficiently large.

Now, towards (3.9), we claim the following for sufficiently small  $\iota$  and some constant  $\tilde{L} \in (1, L)$  bounded away from 1 (to be set later): if  $\mathcal{E}$  holds,  $|X_{t+j}| \leq \varepsilon$ , and  $|X_{t+j}| \geq \tilde{L}^j \cdot \frac{\kappa}{\sqrt{N}}$ , then  $|X_{t+j+1}| \geq \tilde{L}^{j+1} \cdot \frac{\kappa}{\sqrt{N}}$ . It is not difficult to see that this implies (3.9): all the  $X_s$  being smaller than  $\varepsilon$  would contradict the exponential growth prescribed by the previous sentence if the constant  $C$  is taken large enough that  $\tilde{L}^{C \log N} \cdot \frac{\kappa}{\sqrt{N}} \geq \varepsilon$ . To complete the proof, we have that if the events described earlier hold,

$$\begin{aligned} |X_{t+j+1}| &= |g(X_{t+j}) + Y_{t+j} + Z_{t+j}| \\ &\geq |g(X_{t+j})| - |Y_{t+j} + Z_{t+j}| \end{aligned}$$

$$\begin{aligned}
&\geq L \cdot \tilde{L}^j \cdot \frac{\kappa}{\sqrt{N}} - \frac{\iota \kappa}{\sqrt{N}} \cdot (j+1) \\
&\geq \tilde{L}^j \cdot \frac{\kappa}{\sqrt{N}} \cdot \left( L - \iota \cdot \sup_{j \geq 0} \frac{j+1}{\tilde{L}^j} \right).
\end{aligned}$$

To conclude, choose the constants such that

- $\tilde{L} = \frac{L+1}{2}$  is bounded away from 1,
- $\iota$  is small enough that  $L - \iota \cdot \sup_{j \geq 0} \frac{j+1}{\tilde{L}^j} \geq \tilde{L}$ , and
- $\kappa$  is large enough that the probability in (3.10) is bounded away from 0.  $\square$

Let us go back and prove Lemma 3.5.6.

*Proof of Lemma 3.5.6.* We shall prove a similar statement not for the correlation, but for the external fields  $(\mathbf{h}_t)_{t \geq 0}$  encountered along the trajectory of RGD (see Item 2). Concretely, this process is defined as follows, starting from arbitrary  $\mathbf{h} = \mathbf{h}_0 \in \mathbb{R}$ .

1. sample  $\mathbf{z} = f_{\beta, \lambda}(\mathbf{h}) + \frac{1}{\sqrt{N}} \tilde{\mathbf{g}}_{\mathbf{h}}$ , where we recall  $\tilde{\mathbf{g}}_{\mathbf{h}}$  is the deviation from the expected magnetization of a Gibbs sample from  $\mu_{\beta W, \mathbf{h}}$ .
2. draw  $\mathbf{g} \sim \mathcal{N}\left(0, \frac{1}{\beta \lambda N}\right)$ , and move to  $\mathbf{h}' = \mathbf{z} + \mathbf{g}$ .

Lemma 3.6.1 yields that with probability  $1 - o(1)$ , we may write

$$\mathbf{h}_{t+1} = f_{\beta, \lambda}(\mathbf{h}_t) \left( 1 + O\left(\frac{1}{N^\varepsilon}\right) \right) + O\left(\frac{1}{N^{1/2+\varepsilon}}\right) + \tilde{\mathbf{g}}_{\mathbf{h}_t} + \mathbf{g}_t.$$

Let us verify that we can apply Lemma 3.5.7 with  $X_t = \mathbf{h}_t$ ,  $Y_t = O\left(\frac{1}{N^{1/2+\varepsilon}}\right) + \tilde{\mathbf{g}}_{\mathbf{h}_t}$ , and  $Z_t = \mathbf{g}_t$ .

By definition,  $\mathbf{g}_t$  is independent of the collection  $\{\tilde{\mathbf{g}}_{\mathbf{h}_t}, (\tilde{\mathbf{g}}_{\mathbf{h}_s}, \mathbf{g}_s)_{0 \leq s \leq t-1}\}$ . The bound on the conditional second moment of  $\tilde{\mathbf{g}}_{\mathbf{h}}$  follows from Lemma 3.6.2—recall that for  $\beta < 1$ , as mentioned in Remark 3.4.15,  $(\beta, \mathbf{h})$  satisfies (overlap-conc) for all  $\mathbf{h}$ . The lower bound of  $|f_{\beta, \lambda}(x)| (1 + O(\frac{1}{N^\varepsilon})) > L|x|$  for small  $x$  follows from the instability of  $f_{\beta, \lambda}$  at 0 for  $\beta \in (\frac{1}{\lambda}, 1)$ : this is proved in Proposition 3.7.1. The desideratum follows on using Lemma 3.5.7.  $\square$

We may now prove Theorem 3.5.1, restated for convenience.

**Theorem 3.5.1** (Fixed-temperature RGD). *Fix  $\lambda > 1$ ,  $K > 0$  large,  $\varepsilon > 0$  small, and  $\beta > \frac{1}{\lambda}$ . With high probability over the noise  $\mathbf{W}$ , the following holds. Let  $x_0 \in \{\pm 1\}^N$ , and assume that  $(\beta, \beta\lambda|R(x_0, \mathbf{1})|)$  satisfies (AT),  $\beta$  satisfies (RS-AT), and  $(\beta, \lambda)$  satisfies (AT-fixed-pt). Suppose that we run the RGD Markov chain at inverse temperature  $\beta$  from  $x_0$  for  $T \geq \omega(\log N)$  steps to arrive at distribution  $\nu_T$ . Suppose  $T \leq N^K$ . Then,*

$$\mathbf{E}_{x_T \sim \nu_T} [||R(x_T, \mathbf{1})| - \text{OPT}_{\beta, \lambda}|] \leq O\left(\frac{1}{N^{1/2-\varepsilon}}\right).$$

*Proof.* If  $\beta > 1$ , this immediately follows from Lemma 3.5.4, since the AT line is non-trivial there. For  $\beta < 1$ , this follows from Lemmas 3.5.4 and 3.5.6.  $\square$

## 3.6 High-temperature mean magnetization estimates in the SK model

In the RGD recursion for spiked Wigner, the measure decomposition is

$$\mu_{\beta M, 0} = \mathbf{E}_{\substack{x \sim \mu_{\beta M, 0} \\ g \sim N(0, 1)}} [\mu_{\beta W, (\frac{\beta\lambda}{n}\langle \mathbf{1}, x \rangle + \sqrt{\beta\lambda/n}g)\mathbf{1}}].$$

The following explicit formula allows us to determine the deterministic scalar recursion which determines the behavior of locally stationary RGD, in the high-temperature regime  $\beta < 1$ .

**Lemma 3.6.1** (High-temperature magnetization concentration for all fields). *Let  $\beta < 1$ ,  $\varepsilon > 0$  sufficiently small, and  $K > 0$  large independent of  $N$ . With probability  $1 - o(1)$  over  $W \sim \text{GOE}(N)$ , for all  $0 \leq h \leq K$ ,*

(i) if  $h < N^{-(1/4+\varepsilon)}$ ,

$$\frac{1}{N} \mathbf{E}_{x \sim \mu_{\beta W, h\mathbf{1}}} \langle x, \mathbf{1} \rangle = \mathbf{E}_{g \sim N(0, 1)} [\tanh(\beta g \sqrt{q} + h)] \left(1 + O\left(\frac{1}{N^\varepsilon}\right)\right) + O\left(\frac{1}{N^{1/2+\varepsilon}}\right).$$

(i) if  $h > N^{-(1/4+\varepsilon)}$ ,

$$\frac{1}{N} \mathbf{E}_{x \sim \mu_{\beta W, h\mathbf{1}}} \langle x, \mathbf{1} \rangle = \mathbf{E}_{g \sim N(0, 1)} [\tanh(\beta g \sqrt{q} + h)] + O\left(\frac{1}{N^{1/2-\varepsilon}}\right).$$

where  $q = q_h$  is the functional order parameter defined by

$$q = \mathbf{E}_{g \sim N(0, 1)} [\tanh^2(\beta g \sqrt{q} + h)].$$

Note that for  $h > N^{-(1/4+\varepsilon)}$ , a bound of the form of (i) is strictly stronger than that in (ii).

The above theorem will be a corollary of two estimates for the mean magnetization. The first is essentially already present in [Han07] (also see [CT21]).

**Lemma 3.6.2** (essentially [Han07]). *Let  $(\beta, h)$  satisfy (overlap-conc). Then, for any constant  $k$ ,*

$$\mathbf{E} \left[ (R(\sigma, \mathbf{1}) - q_1)^{2k} \right] \leq O \left( \frac{1}{N^k} \right),$$

where we denote  $q_1 = \mathbf{E}_{g \sim N(0,1)} [\tanh(\beta g \sqrt{q} + h)]$ .

The proof is identical to that of Corollary 4.1, Theorem 5.1, and Theorem 6.1 in [Han07]—we observe that all these proofs of upper bounds on the moments go through assuming overlap concentration, even if we do not have a central limit theorem for the overlap (such a CLT is required for the stronger Theorem 1.2 in Hanen’s paper, which proves a CLT for the magnetization).

However, this lemma is not useful in the regime where  $h$  is tiny, say  $\Theta(N^{-1/2})$ , since  $q_1$  itself is  $O(h)$ , and the error is at the same scale as the estimate. To get around this, we prove a mean magnetization formula that is more precise in the small  $h$  regime, albeit only with a lower moment that will suffice for our purposes.

**Lemma 3.6.3** (Magnetization estimates under weak external field). *Let  $\delta > 0$  and  $h < N^{-\alpha}$  for some  $\alpha > 1/4$ . Then,*

$$\mathbf{E} \left[ \left| \frac{1}{N} \sum_{i=1}^N \langle \sigma_i \rangle - q_1 \right|^{2-\delta} \right] \leq O \left( \max \left\{ \frac{1}{N}, \frac{h}{\sqrt{N}} \right\} \right)^{2-\delta}.$$

We will defer the proof of Lemma 3.6.3 to Section 3.6. We may now prove Lemma 3.6.1.

*Proof.* For ease of notation, let  $\mathbf{m}(h)$  be the random variable  $\frac{1}{N} \sum_{i=1}^N \langle \sigma_i \rangle$  under external field  $h$ . It is not difficult to show that  $\mathbf{m}(h)$  is non-decreasing in  $h$ : indeed, its derivative with respect to  $h$  is the variance of the magnetization in the corresponding model. For  $h > N^{-(\frac{1}{4}+\varepsilon)}$ , Markov’s inequality applied to Lemma 3.6.2 yields that

$$\Pr \left[ |\mathbf{m}(h) - q_1(h)| > \frac{1}{N^{\frac{1}{2}-\varepsilon}} \right] < \frac{1}{N^2}. \quad (3.11)$$

On the other hand, for  $h < N^{-(\frac{1}{4}+\varepsilon)}$ , Markov’s inequality applied to Lemma 3.6.3 yields that for some  $c_1$  we will fix shortly,

$$\Pr \left[ |\mathbf{m}(h) - q_1(h)| > \max \left\{ \frac{h}{N^{\frac{1}{2}-\frac{c_1}{2-\delta}}}, \frac{1}{N^{1-\frac{c_1}{2-\delta}}} \right\} \right] < O(N^{-c_1}).$$

Now, choose  $\delta = \varepsilon$  and  $c_1$  such that  $\frac{c_1}{2-\varepsilon} = \frac{1}{2} - \varepsilon$ , so appealing to Lemma 3.7.4, the above reduces to

$$\Pr \left[ |\mathbf{m}(h) - q_1(h)| > \max \left\{ \frac{q_1(h)}{N^\varepsilon}, \frac{1}{N^{1/2+\varepsilon}} \right\} \right] < O(N^{-c_1}). \quad (3.12)$$

For  $c_2 = \frac{1}{2} + \varepsilon$ , let  $S = \{q_1^{-1}(iN^{-c_2}) : 0 \leq i \leq KN^{c_2}\}$ ; note that  $S$  is well defined because  $q_1$  is strictly increasing in  $h$ . We shall perform a union bound over  $S$ . For sufficiently small  $\varepsilon$ , it is plainly true that  $N^{-c_1} \cdot N^{c_2} = o(1)$ , since  $c_1 \approx 1$  and  $c_2 \approx \frac{1}{2}$ . Consequently, (3.12) yields that with probability  $1 - o(1)$ , for all  $h \in S$  with  $h < N^{-(\frac{1}{4}+\varepsilon)}$ ,

$$|\mathbf{m}(h) - q_1(h)| < O \left( \max \left\{ \frac{q_1(h)}{N^\varepsilon}, \frac{1}{N^{1/2+\varepsilon}} \right\} \right).$$

On the other hand, (3.11) yields that with probability  $1 - o(1)$ , for all  $h \in S$  with  $h > N^{-(\frac{1}{4}+\varepsilon)}$ ,

$$|\mathbf{m}(h) - q_1(h)| < \frac{1}{N^{1/2-\varepsilon}}.$$

This bound further extends from  $S$  to all  $h$  bounded by  $K$ : indeed, suppose  $h \leq K$  and  $h_1 < h < h_2$  such that  $h_1, h_2 \in S$ ,  $|q_1(h_1) - q_1(h)| < N^{-c_2}$  and  $|q_1(h_2) - q_1(h)| < N^{-c_2}$ . Then, if  $h < N^{-(\frac{1}{4}+\varepsilon)}$ ,

$$\mathbf{m}(h) \geq \mathbf{m}(h_1) \geq q_1(h_1) - O \left( \frac{q_1(h_1)}{N^\varepsilon} \right) - O \left( \frac{1}{N^{1/2+\varepsilon}} \right) \geq q_1(h) - O \left( \frac{q_1(h)}{N^\varepsilon} \right) - O \left( \frac{1}{N^{1/2+\varepsilon}} \right),$$

where we have used the fact that  $q_1$  is strictly increasing in  $h$ . Similarly, we have

$$\mathbf{m}(h) \leq q_1(h) + O \left( \frac{q_1(h)}{N^\varepsilon} \right) + O \left( \frac{1}{N^{1/2+\varepsilon}} \right).$$

An identical argument works for the alternate error bound when  $h > N^{-(\frac{1}{4}+\varepsilon)}$ .  $\square$

## High-precision mean magnetization estimates under weak external field

For this section, let  $h < N^{-\alpha}$  for some  $\alpha > 1/4$ , and  $\beta < 1$ . We also define  $q, q_1$  by the recursions

$$\begin{aligned} q &= \mathbf{E}_{\mathbf{z} \sim \mathcal{N}(0,1)} \left[ \tanh^2(\beta \sqrt{q} \mathbf{z} + h) \right], \\ q_1 &= \mathbf{E}_{\mathbf{z} \sim \mathcal{N}(0,1)} \left[ \tanh(\beta \sqrt{q} \mathbf{z} + h) \right]. \end{aligned}$$

We also denote  $Z$  to be the partition function of the SK model with external field  $h$ .

In this section, we will prove the following lemma.

**Lemma 3.6.3** (Magnetization estimates under weak external field). *Let  $\delta > 0$  and  $h < N^{-\alpha}$  for some  $\alpha > 1/4$ . Then,*

$$\mathbf{E} \left[ \left| \frac{1}{N} \sum_{i=1}^N \langle \sigma_i \rangle - q_1 \right|^{2-\delta} \right] \leq O \left( \max \left\{ \frac{1}{N}, \frac{h}{\sqrt{N}} \right\} \right)^{2-\delta}.$$

**Lemma 3.6.4.**  $\mathbf{E} \left[ Z^2 \left( \frac{1}{N} \sum_{i=1}^N \langle \sigma_i \rangle - q_1 \right)^2 \right] \leq \mathbf{E}[Z]^2 \cdot O \left( \max \left\{ \frac{1}{N}, \frac{h}{\sqrt{N}} \right\} \right)^2.$

*Proof of Lemma 3.6.3.* We have by Hölder's inequality and [DW23, Corollary 3.5] (using the fact that  $\alpha > 1/4$ ) that

$$\begin{aligned} \mathbf{E} \left[ \left| \frac{1}{N} \sum_{i=1}^N \langle \sigma_i \rangle - q_1 \right|^{2-\delta} \right] &\leq \mathbf{E} \left[ Z^2 \cdot \left( \frac{1}{N} \sum_{i=1}^N \langle \sigma_i \rangle - q_1 \right)^2 \right]^{\frac{2-\delta}{2}} \cdot \mathbf{E} \left[ \frac{1}{Z^{4/\delta-2}} \right]^{\delta/2} \\ &\leq \mathbf{E}[Z]^{2-\delta} \cdot O \left( \max \left\{ \frac{1}{N}, \frac{h}{\sqrt{N}} \right\} \right)^{2-\delta} \cdot \frac{1}{\mathbf{E}[Z]^{2-\delta}} \\ &= O \left( \max \left\{ \frac{1}{N}, \frac{h}{\sqrt{N}} \right\} \right)^{2-\delta}, \end{aligned}$$

as desired. □

We dedicate the remainder of this section to proving Lemma 3.6.4. Using replicas, we can write

$$\mathbf{E} \left[ Z^2 \left( \frac{1}{N} \sum_{i=1}^N \langle \sigma_i \rangle - q_1 \right)^2 \right] = \sum_{\sigma, \rho} \mathbf{E} [e^{H_N(\sigma) + H_N(\rho)} (m(\sigma) - q_1)(m(\rho) - q_1)],$$

where  $H_N(\sigma) = \frac{\beta}{2} \langle \sigma, \mathbf{W} \sigma \rangle + h \langle \sigma, \mathbf{1} \rangle$ , so that  $\mathbf{Cov}(H_N(\sigma), H_N(\rho)) = \frac{\beta^2}{2} N R(\sigma, \rho)^2$ . It follows that  $H_N(\sigma) + H_N(\rho)$  is distributed as  $\mathcal{N}(h \langle \sigma + \rho, \mathbf{1} \rangle, N \beta^2 (1 + R(\sigma, \rho)^2))$ , and so,

$$\mathbf{E} [e^{H_N(\sigma) + H_N(\rho)}] = e^{\frac{N \beta^2}{2} (1 + R(\sigma, \rho)^2) + h \langle \sigma + \rho, \mathbf{1} \rangle}.$$

Since the summand in fact has no dependence on the disorder, we have reduced the moment calculation to an evaluation of the double sum

$$\sum_{\sigma, \rho \in \{\pm 1\}^N} e^{\frac{N \beta^2}{2} (1 + R(\sigma, \rho)^2) + h \langle \sigma + \rho, \mathbf{1} \rangle} (m(\sigma) - q_1)(m(\rho) - q_1)$$

$$= (\mathbf{EZ})^2 (2 \cosh(h))^{-2N} \sum_{\sigma, \rho} e^{\frac{N\beta^2}{2} R(\sigma, \rho)^2 + h \langle \sigma + \rho, \mathbf{1} \rangle} (m(\sigma) - q_1)(m(\rho) - q_1),$$

where we have used that  $\mathbf{E}[Z] = (2 \cosh(h))^N e^{\frac{\beta^2 N}{4}}$ . This explains the dependence on  $(\mathbf{EZ})^2$ .

It thus remains to compute the double sum. To this end, we will use the HS transform, followed by the Laplace method.

**Fact 3.6.5** (Hubbard–Stratonovich transform). *For any  $c > 0$ , we have*

$$e^{cx^2} = \frac{1}{\sqrt{c\pi}} \int_{\mathbb{R}} e^{-\frac{z^2}{c} + 2zx} dz$$

Applying the HS transform with  $c = \frac{2\beta^2}{N}$  and  $x = \frac{\langle \sigma, \rho \rangle}{2}$ , we obtain that

$$e^{\frac{N\beta^2}{2} R(\sigma, \rho)^2} = e^{\frac{\beta^2}{2N} \langle \sigma, \rho \rangle^2} = \sqrt{\frac{N}{2\pi\beta^2}} \int_{\mathbb{R}} e^{-\frac{N}{2\beta^2} z^2 + z \langle \sigma, \rho \rangle} dz,$$

so that the double sum becomes

$$\sqrt{\frac{N}{2\pi\beta^2}} \cdot \int_{\mathbb{R}} \exp\left(-\frac{N}{2\beta^2} z^2\right) \sum_{\sigma, \rho} \exp(z \langle \sigma, \rho \rangle + h \langle \sigma + \rho, \mathbf{1} \rangle) (m(\sigma) - q_1)(m(\rho) - q_1) dz.$$

Observe that given  $z$ , the pairs  $(\sigma_i, \rho_i)$  are independent of each other, and hence this sum over  $\sigma, \rho$  is tractable to explicitly compute. In particular, consider the distribution  $p$  on  $\{\pm 1\}^2$ , with  $p(a, b) = \frac{1}{C(z, h)} \exp(zab + ha + hb)$ . Then one can check the normalization constant is  $C(z, h) = 2(e^z \cosh(2h) + e^{-z})$ , and we have

$$\sum_{\sigma, \rho} \exp(z \langle \sigma, \rho \rangle + h \langle \sigma + \rho, \mathbf{1} \rangle) (m(\sigma) - q_1)(m(\rho) - q_1) = C(z, h)^N \cdot \mathbf{E}_{p^{\otimes N}} [(m(\sigma) - q_1)(m(\rho) - q_1)].$$

Now, expanding it out, we have

$$(m(\sigma) - q_1)(m(\rho) - q_1) = \frac{1}{N^2} \sum_{i \neq j} (\sigma_i - q_1)(\rho_j - q_1) + \frac{1}{N^2} \sum_i (\sigma_i - q_1)(\rho_i - q_1),$$

and therefore,

$$\begin{aligned} f_N(z) &:= \mathbf{E}_{p^{\otimes N}} [(m(\sigma) - q_1)(m(\rho) - q_1)] \\ &= \frac{N(N-1)}{N^2} (m(z) - q_1)^2 + \frac{1}{N} (k(z) - 2q_1 m(z) + q_1^2) \end{aligned}$$



$$\begin{aligned}
&= (m(z) - q_1)^2 + \frac{1}{N} (k(z) - m(z))^2 \\
&= (m(z) - q_1)^2 + \frac{1}{N} (k(z) - q) + \frac{1}{N} (q - q_1^2) + \frac{1}{N} (q_1^2 - m(z)^2)
\end{aligned}$$

where  $m(z) := \mathbf{E}_p[a] = \mathbf{E}_p[b]$  and  $k(z) := \mathbf{E}_p[ab]$ . Thus, our goal is to show that

$$(2 \cosh(h))^{-2N} \cdot \sqrt{\frac{N}{2\pi\beta^2}} \cdot \int_{\mathbb{R}} \exp\left(N\left[-\frac{z^2}{2\beta^2} + \log(2(e^z \cosh(2h) + e^{-z}))\right]\right) f_N(z) dz \leq O\left(\frac{h^2}{N}\right).$$

Rearranging, simplifying, and summarizing, we would like to show that

$$\mathbf{E}_{z \sim \mathcal{N}\left(0, \frac{\beta^2}{N}\right)} \left[ \left( \frac{e^z \cosh(2h) + e^{-z}}{\cosh(2h) + 1} \right)^N f_N(z) \right] \leq O\left(\frac{h^2}{N}\right),$$

where an explicit calculation yields that

$$\begin{aligned}
m(z) &= \frac{e^z \sinh(2h)}{e^z \cosh(2h) + e^{-z}} \text{ and} \\
k(z) &= \frac{e^z \cosh(2h) - e^{-z}}{e^z \cosh(2h) + e^{-z}}.
\end{aligned}$$

Let us massage this expression to deal with the first part of the integrand. Set

$$g_N(z) = \frac{\cosh(z) + \sinh(h)^2 e^z}{\left(1 + \sinh(h)^2\right) e^{\tanh(h)^2 z} \left(e^{\frac{z^2}{2} (1 - \tanh(h)^4)}\right)}.$$

Then,

$$(*) = \mathbf{E}_{z \sim \mathcal{N}\left(0, \frac{\beta^2}{N}\right)} \left[ g_N(z)^N \cdot e^{N \tanh(h)^2 z} \cdot e^{\frac{N z^2}{2} (1 - \tanh(h)^4)} \cdot f_N(z) \right]$$

Define  $\gamma > 0$  by  $\frac{1}{\gamma^2} = \frac{1}{\beta^2} - (1 - \tanh(h)^4)$ , and  $\mu = \gamma^2 \tanh(h)^2$ , so

$$\begin{aligned}
(*) &= \frac{\gamma}{\beta} \cdot \mathbf{E}_{z \sim \mathcal{N}\left(0, \frac{\gamma^2}{N}\right)} \left[ g_N(z)^N \cdot e^{N \tanh(h)^2 z} \cdot f_N(z) \right] \\
&= \frac{\gamma}{\beta} \cdot e^{\frac{1}{2} N \gamma^2 \tanh(h)^4} \cdot \mathbf{E}_{z \sim \mathcal{N}\left(\mu, \frac{\gamma^2}{N}\right)} \left[ g_N(z)^N \cdot f_N(z) \right] \\
&\asymp \mathbf{E}_{z \sim \mathcal{N}\left(\mu, \frac{\gamma^2}{N}\right)} \left[ g_N(z)^N \cdot f_N(z) \right],
\end{aligned}$$

where the final line uses the fact that  $h \ll N^{-1/4}$ . Why is this useful? By the choice of  $g$ , we in fact have  $g(0) = 1$ ,  $g^{(1)}(0) = g^{(2)}(0) = 0$ , and  $g^{(3)}(0) = -2 \tanh(h)^2 (1 - \tanh(h)^4) = O(h^2)$ .

Thus, since  $z$  is typically  $\lesssim \frac{1}{\sqrt{N}}$ ,  $g_N(z)$  is typically  $1 + O\left(\frac{h^2}{N^{3/2}}\right)$ , and so  $g_N(z)^N$  is typically  $1 + O\left(\frac{h^2}{\sqrt{N}}\right)$ , essentially allowing us to ignore it in the integral.

More concretely, for our purposes (where we do not care about precise constants), it suffices to use the looser Cauchy–Schwarz bound, yielding

$$(*) \lesssim \mathbf{E} [g_N(z)^{2N}]^{1/2} \cdot \mathbf{E} [f_N(z)^2]^{1/2}$$

To control this, we will show

- (vi)  $\mathbf{E} \left[ (m(z) - q_1)^{2k} \right] \leq O\left(\frac{h^{2k}}{N^k}\right)$  for all constant  $k$ , proved in Proposition 3.6.6,
- (vi)  $\mathbf{E} \left[ (k(z) - q)^{2k} \right] \leq O\left(\frac{1}{N^k}\right)$  for all constant  $k$ , proved in Proposition 3.6.7, and
- (vi)  $\mathbf{E} [g_N(z)^{2N}] = O(1)$ , proved in Proposition 3.6.8.

These essentially immediately yield Lemma 3.6.4.

## A series of Gaussian expectations

**Proposition 3.6.6** (Controlling the magnetization). *For any constant  $k$ ,*

$$\mathbf{E}_{z \sim \mathcal{N}(\mu, \frac{\gamma^2}{N})} \left[ \left( \frac{e^z \sinh(2h)}{e^z \cosh(2h) + e^{-z}} - q_1 \right)^{2k} \right] \leq O\left(\frac{h^2}{N}\right)^k.$$

*Proof.* Let us write

$$\begin{aligned} & \mathbf{E} \left[ \left( \frac{e^z \sinh(2h)}{e^z \cosh(2h) + e^{-z}} - q_1 \right)^{2k} \right] \\ & \lesssim \mathbf{E} \left[ \left( \frac{e^z \sinh(2h)}{e^z \cosh(2h) + e^{-z}} - \mathbf{E} \left[ \frac{e^z \sinh(2h)}{e^z \cosh(2h) + e^{-z}} \right] \right)^{2k} \right] + \left( \mathbf{E} \left[ \frac{e^z \sinh(2h)}{e^z \cosh(2h) + e^{-z}} \right] - q_1 \right)^{2k} \\ & = O\left(h^{2k}\right) \mathbf{E} \left[ \left( \frac{e^z}{e^z \cosh(2h) + e^{-z}} - \mathbf{E} \left[ \frac{e^z}{e^z \cosh(2h) + e^{-z}} \right] \right)^{2k} \right] + \left( \mathbf{E} \left[ \frac{e^z \sinh(2h)}{e^z \cosh(2h) + e^{-z}} \right] - q_1 \right)^{2k}. \end{aligned}$$

The first term is easy to deal with: because  $z \mapsto \frac{e^z}{e^z \cosh(2h) + e^{-z}}$  is an  $O(1)$ -Lipschitz function, Gaussian concentration of Lipschitz functions yields that the associated  $2k$ th moment is bounded by  $O\left(\frac{\gamma^2}{N}\right)^k$ .

For the second term, because  $q = O(h^2)$ ,

$$q_1 = \mathbf{E}_{g \sim \mathcal{N}(0,1)} \left[ \tanh(\beta g \sqrt{q} + h) \right] = h + O(h^3).$$

On the other hand, setting  $f(z) = \frac{e^z \sinh(2h)}{e^z \cosh(2h) + e^{-z}}$ , so

$$\begin{aligned} f(0) &= \tanh(h) = h + O(h^3), \\ f'(0) &= O(h), \text{ and} \\ f''(0) &= O(h^3), \end{aligned}$$

we have

$$\begin{aligned} \mathbf{E}[f(z)] &= f(0) + f'(0)\mu + \frac{1}{2}f''(0)\left(\mu^2 + \frac{\gamma^2}{N}\right) + O\left(\frac{h}{\sqrt{N}}\right) \\ &= h + O(h) \cdot \gamma^2 \tanh(h)^2 - O(h^3) \cdot \left(\mu^2 + \frac{\gamma^2}{N}\right) + O\left(\frac{h}{\sqrt{N}}\right) \\ &= h + O\left(\frac{h}{\sqrt{N}}\right). \end{aligned}$$

Consequently,

$$\left| \mathbf{E} \left[ \frac{e^z \sinh(2h)}{e^z \cosh(2h) + e^{-z}} \right] - q_1 \right| = O\left(\frac{h}{\sqrt{N}}\right),$$

completing the proof. □

**Proposition 3.6.7** (Controlling the overlap). *For any constant  $k$ ,*

$$\mathbf{E}_{z \sim \mathcal{N}\left(\mu, \frac{\gamma^2}{N}\right)} \left[ \left( \frac{e^z \cosh(2h) - e^{-z}}{e^z \cosh(2h) + e^{-z}} - q \right)^{2k} \right] \leq O\left(\frac{1}{N^k}\right).$$

*Proof.* Our strategy will essentially be identical to that of the previous proof. Again, because  $z \mapsto k(z)$  is an  $O(1)$ -Lipschitz function, we have

$$\mathbf{E} \left[ \left( \frac{e^z \cosh(2h) - e^{-z}}{e^z \cosh(2h) + e^{-z}} - q \right)^{2k} \right] \lesssim O\left(\frac{1}{N^k}\right) + \left( \mathbf{E} \left[ \frac{e^z \cosh(2h) - e^{-z}}{e^z \cosh(2h) + e^{-z}} \right] - q \right)^{2k}$$

We have  $k(z) = \tanh\left(z + \frac{1}{2} \ln \cosh(2h)\right) = \tanh\left(z + h^2 + O(h^4)\right)$ . As a result,

$$\mathbf{E}[k(z)] = \mathbf{E} \left[ z + h^2 - \frac{1}{3}z^3 \right] + O(h^4)$$

$$\begin{aligned}
&= \mu + h^2 - \frac{1}{3} \cdot \mu^3 - \mu \cdot \frac{\gamma^2}{N} + O(h^4) \\
&= \mu + h^2 + O\left(\frac{1}{N}\right).
\end{aligned}$$

Let us also figure out the leading terms in the Taylor series of  $q$ . We have  $q = O(h^2)$ , so suppose that  $q = ah^2 + O(h^4)$ . Expanding out the recursion defining  $q$ , using the Taylor series expansion  $\tanh^2(x) = x^2 + O(x^4)$ , we have

$$q = \mathbf{E}_{g \sim \mathcal{N}(0,1)} [\beta \sqrt{q} g + h] = \beta^2 q + h^2 + O(h^4).$$

It follows that  $q = \frac{1}{1-\beta^2} \cdot h^2 + O(h^4)$ . We also have

$$\begin{aligned}
\mu + h^2 &= \tanh(h)^2 \cdot \frac{1}{\frac{1}{\beta^2} - (1 - \tanh(h)^4)} + h^2 \\
&= h^2 \left( \frac{1}{\frac{1}{\beta^2} - 1} + 1 \right) + O\left(\frac{1}{N}\right) \\
&= \frac{1}{1 - \beta^2} \cdot h^2 + O\left(\frac{1}{N}\right).
\end{aligned}$$

Therefore,

$$|\mathbf{E}[k(z)] - q| = O\left(\frac{1}{N}\right),$$

completing the proof. □

**Proposition 3.6.8.** *It holds that  $\mathbf{E}_{z \sim \mathcal{N}\left(0, \frac{\beta^2}{N}\right)} [g_N(z)^{2N}] = 1 + o_N(1)$ .*

*Proof.* Recall

$$g_N(z) = \frac{\cosh(z) + \sinh(h)^2 e^z}{\left(1 + \sinh(h)^2\right) e^{\tanh(h)^2 \cdot z} \left(e^{\frac{z^2}{2} (1 - \tanh(h)^4)}\right)}.$$

Let  $\kappa = \tanh(h)^2 = \frac{\sinh(h)^2}{1 + \sinh(h)^2}$ , so we have

$$g_N(z) = \frac{\cosh(z) + \kappa \sinh(z)}{e^{\kappa z + \frac{z^2}{2} \cdot (1 - \kappa^2)}}.$$

First note that the terms in the Taylor expansion up to order 2 cancel out—indeed, for small  $z$ , we have  $\cosh(z) + \kappa \sinh(z) = 1 + \kappa z + \frac{z^2}{2} + O(z^3)$ , and

$$e^{\kappa z + \frac{z^2}{2} \cdot (1 - \kappa^2)} = 1 + \kappa z + (1 - \kappa^2) \cdot \frac{z^2}{2} + \frac{1}{2}(\kappa z)^2 + O(z^3) = 1 + \kappa z + \frac{z^2}{2} + O(z^3).$$

The claim near-immediately follows. For large  $z$ , the  $e^{(1-\kappa^2)z^2/2}$  term in the denominator dominates, so we may establish a uniform bound of  $g_N(z) \leq 1 + O(|z|^3)$ . We may thus bound  $\mathbf{E} [g_N(z)^{2N}] \leq \mathbf{E} \left[ (1 + O(|z|^3))^{2N} \right] \leq 1 + O\left(\frac{1}{\sqrt{N}}\right)$ , as desired.  $\square$

## 3.7 Towards understanding the RGD recursion

### Towards understanding the fixed points in the (AT) region

In this section, we analyze some salient properties of the fixed point equations governing RGD and AMP. While we do not show that there exists a unique fixed point to  $f_{\beta,\lambda}$  defined in (3.5) in the (AT) region, we are able to show that there is *at most* one fixed point in this region, and that this fixed point must be stable.

**Proposition 3.7.1.** *Fix  $\lambda > 1$  and  $\beta \geq 0$ . Set  $h_* = \inf\{h \geq 0 : (\beta, h) \text{ satisfies (AT)}\}$ . Suppose that for all  $h > h_*$ ,  $(\beta, h)$  satisfies (AT). Then, for  $f_\beta = f_{\beta,\lambda}$  defined as in (3.5),*

- (i) *If  $\beta < \frac{1}{\lambda}$ , then  $h = 0$  is the only fixed point for  $f_\beta(h)$ . Furthermore,  $h = 0$  is a stable fixed point.*
- (i) *If  $\frac{1}{\lambda} < \beta < 1$ , there are exactly two non-negative fixed points for  $f_\beta(h)$  at 0 and  $\text{OPT}_{\beta,\lambda}$ . Furthermore,  $h = 0$  is an unstable fixed point and  $h = \text{OPT}_{\beta,\lambda}$  is a stable fixed point.*
- (i) *If  $\beta > 1$ , there is at most one positive fixed point  $h > 0$  such that  $(\beta, h)$  satisfies (AT). Furthermore, if  $h$  exists, it is a stable fixed point of  $f_\beta$ . Finally, for  $\beta = \lambda$ , such a fixed point does indeed exist.*
- (i) *For  $h > \frac{h_*}{\beta\lambda}$ ,  $f'_\beta(h) > 0$ .*

Given the above, Lemma 3.5.3 is near-immediate.

**Lemma 3.5.3.** *Let  $\lambda > 1$ . If  $\beta < 1$  or  $\beta = \lambda$ ,  $(\beta, \lambda)$  satisfies (AT-fixed-pt). Suppose that  $(\beta, \lambda)$  satisfies (AT-fixed-pt). Then,*

- (clvi) *if  $h$  is such that  $(\beta, \beta\lambda h)$  satisfies (AT),  $f'_{\beta,\lambda}(h) > 0$ ,*
- (clvi)  *$\text{OPT}_{\beta,\lambda}$  is the unique fixed point  $h$  of  $f_{\beta,\lambda}$  such that  $(\beta, \beta\lambda h)$  satisfies (AT), and*
- (clvi)  *$\text{OPT}_{\beta,\lambda}$  is a stable fixed point of  $f_{\beta,\lambda}$ .*

*Proof.* (i, ii, iii) near-immediately follow from Proposition 3.7.1(iii, iv). The fact that (AT-fixed-pt) is satisfied if  $\beta < 1$  or  $\beta = \lambda$  is a consequence of Proposition 3.7.1(i, ii, iii).  $\square$

We shall prove Proposition 3.7.1 over the course of this section. We start by noting the following simple fact.

**Fact 3.7.2.** *For any  $a, b > 0$ ,*

$$2a\mathbf{E}[\operatorname{sech}^2(ag + b) \tanh(ag + b)] = -\mathbf{E}_{g \sim \mathcal{N}(0,1)}[\operatorname{sech}^2(ag + b) \cdot g] > 0.$$

*Proof.* The first equality is just Stein's lemma. For the second, we use symmetry to express

$$\mathbf{E}_{g \sim \mathcal{N}(0,1)}[\operatorname{sech}^2(ag + b) \cdot g] = \frac{1}{2} \cdot \mathbf{E}[g(\operatorname{sech}^2(b + ag) - \operatorname{sech}^2(b - ag))] < 0.$$

If  $g > 0$  we have  $|b + ag| > |b - ag|$ , and if  $g < 0$ , we have  $|b - ag| > |b + ag|$ . Since  $\operatorname{sech}^2(\cdot)$  is even and decreasing for positive arguments, the claim follows.  $\square$

We also need the following facts about the recursion defining the overlap constant  $q$ .

**Lemma 3.7.3.** *Let  $\beta > 0$  and  $h \geq 0$ . Consider the solution to the fixed point equation*

$$q = q(h) = \mathbf{E}[\tanh^2(\beta g \sqrt{q} + h)].$$

*Then the following hold.*

- (iii) *If  $h > 0$ , there is a unique solution  $q = q(h) > 0$ . If  $h = 0$  and  $\beta < 1$ , then  $q = 0$  is the unique solution to the above recursion, and  $q'(0) = 0$ .*
- (iii) *If  $(\beta, h)$  satisfy (AT), then  $q$  is differentiable. Denoting  $\mathbf{X} = \beta g \sqrt{q} + h$ , we have*

$$q'(h) = \frac{2\mathbf{E}[\tanh(\mathbf{X}) \operatorname{sech}^2(\mathbf{X})]}{1 - \beta^2 \mathbf{E}[\operatorname{sech}^4(\mathbf{X}) - 2 \operatorname{sech}^2(\mathbf{X}) \tanh^2(\mathbf{X})]}.$$

*Consequently,  $q'(h) > 0$ .*

*Proof.* The uniqueness of  $q$  in the mentioned regimes is [Gue01] (also see [Tal10, Proposition 1.3.8]). For the second point, define the shorthand  $T = \tanh(\beta g \sqrt{q} + h)$  and  $S = \operatorname{sech}(\beta g \sqrt{q} + h)$ . We can implicitly differentiate through  $h$  to obtain

$$\begin{aligned} q' &= \mathbf{E}[2TS^2(\beta g(\sqrt{q})' + 1)] \\ &= 2\mathbf{E}[TS^2] + \frac{\beta q'}{\sqrt{q}} \mathbf{E}[gTS^2] \\ &= 2\mathbf{E}[TS^2] + \beta^2 q' \mathbf{E}[S^4 - 2S^2T^2]. \end{aligned} \quad (\text{Stein's lemma})$$

Solving for  $q'$  yields

$$q' = \frac{2\mathbf{E}[TS^2]}{1 - \beta^2\mathbf{E}[S^4 - 2S^2T^2]}.$$

For  $(\beta, h)$  satisfying (AT), the denominator is strictly positive (thus justifying the well-definedness of  $q'$ ). The numerator is strictly positive by Fact 3.7.2.  $\square$

We will also need the following estimates for the behavior of  $q$  and  $q_1$  for small  $h$ .

**Lemma 3.7.4.** *Let  $h \geq 0$  and  $\beta < 1$ . Also let  $q(h)$  be the unique solution to  $q = \mathbf{E}[\tanh^2(\beta g\sqrt{q} + h)]$  and define  $q_1(h) = \mathbf{E}[\tanh(\beta g\sqrt{q(h)} + h)]$ . Then*

$$\begin{aligned} \frac{h^2}{1 - \beta^2} - \frac{2h^4}{(1 - \beta^2)^3} &\leq q(h) \leq \frac{h^2}{1 - \beta^2} \\ h - \frac{h^2}{3(1 - \beta^2)} &\leq q_1(h) \leq h \end{aligned}$$

*Proof.* The inequality for  $q$  is [DW23, Lemma 1.2]. For the inequality on  $q_1$ , we first note that  $\tanh x \leq x$ , so with  $Y = \beta\sqrt{q}z + h$ , we have  $q_1 = \mathbf{E} \tanh Y \leq \mathbf{E}Y = h$ . Furthermore, since  $x - \tanh x \leq \frac{x^2}{3}$ , we obtain  $\mathbf{E}[Y - \tanh Y] \leq \frac{1}{3}\mathbf{E}Y^2 = \frac{1}{3}(h^2 + \beta^2q)$ . Combining with the upper bound on  $q$ , we obtain

$$q_1 \geq h - \frac{h^2}{3(1 - \beta^2)},$$

as desired.  $\square$

We are now ready to prove some nice properties of  $f_\beta$ .

**Lemma 3.7.5.** *Fix  $\lambda > 1$  and  $\beta > 0$ . Consider the function  $f_\beta$  defined as in (3.5):*

$$f_\beta(h) = \mathbf{E}_{g \sim \mathcal{N}(0,1)} \tanh(\beta g\sqrt{q_{\beta,\beta\lambda h}} + \beta\lambda h).$$

*Then,*

- (v) *If  $(\beta, \beta\lambda h)$  satisfies (AT), the function  $m \mapsto \frac{f_\beta(m)}{m}$  is strictly decreasing at  $h$ .*
- (v) *If  $(\beta, \beta\lambda h)$  satisfies (AT),  $f_\beta$  is increasing at  $h$ .*
- (v) *0 is a stable fixed point  $f_\beta$  if  $\beta < \frac{1}{\lambda}$ , and is an unstable fixed point if  $\frac{1}{\lambda} < \beta < 1$ .*

*Proof.* Let  $X = X(h) = \beta\lambda h + \beta g\sqrt{q_{\beta,\beta\lambda h}}$ , and introduce the function  $r = r(h) = q_{\beta,\beta\lambda h}$ . By Stein's lemma, we have

$$\begin{aligned} f'_\beta(h) &= \mathbf{E}_{g \sim \mathcal{N}(0,1)} \left[ \operatorname{sech}^2(X) \cdot \left( \beta g \cdot \frac{d}{dh} \sqrt{r} + \beta\lambda \right) \right] \\ &= \mathbf{E} \left[ \operatorname{sech}^2(X) \left( \beta g \frac{r'}{2\sqrt{r}} + \beta\lambda \right) \right] \\ &= \beta\lambda \mathbf{E}[\operatorname{sech}^2(X)] - \beta^2 r' \mathbf{E}[\operatorname{sech}^2(X) \tanh(X)]. \end{aligned}$$

Here, we have implicitly used the differentiability of  $\sqrt{r(h)}$ , which we justify now.

First, by Lemma 3.7.3(ii), if  $\beta < 1$ ,  $r$  is differentiable and strictly positive for all  $h > 0$ . Let us now compute the limit as  $h \rightarrow 0^+$ . By Lemmas 3.7.3 and 3.7.4, we have  $r'(h) = \beta^2 \lambda^2 \cdot (1 + o_h(1)) \frac{2h}{1-\beta^2}$  and  $r(h) = \beta^2 \lambda^2 \cdot \frac{h^2}{1-\beta^2} (1 + o_h(1))$ . Hence,

$$\lim_{h \rightarrow 0^+} \frac{r'(h)}{\sqrt{r(h)}} = \frac{2\beta\lambda}{\sqrt{1-\beta^2}}.$$

It follows by symmetry that  $f'_\beta(h)$  is well defined for  $\beta < 1$  and all  $h \in \mathbb{R}$ .

If  $\beta < 1$  and  $h = 0$ , then  $r(h) = 0$  by Lemma 3.7.3(i), so  $X = 0$  almost surely and the above simplifies to

$$f'_\beta(0) = \beta\lambda,$$

which implies (iii).

Let us next establish (ii). Abbreviate  $S = \operatorname{sech}(X)$  and  $T = \tanh(X)$ . We have by Lemma 3.7.3(ii) and (AT) that

$$\begin{aligned} f'_\beta(h) &= \beta\lambda \mathbf{E}S^2 - \beta^2 \cdot \beta\lambda \cdot \frac{2\mathbf{E}[TS^2]}{1 - \beta^2 \mathbf{E}[S^4 - 2S^2T^2]} \cdot \mathbf{E}[TS^2] \\ &\geq \beta\lambda \left( \mathbf{E}S^4 - \frac{\mathbf{E}[S^2T]^2}{\mathbf{E}[S^2T^2]} \right) \geq 0, \end{aligned}$$

where the final inequality is the Cauchy–Schwarz inequality.

Finally, consider (i). Set  $g(h) = \frac{f_\beta(h)}{h}$ . We have

$$\begin{aligned} h^2 \cdot g'(h) &= f'_\beta(h)h - f_\beta(h) \\ &= \mathbf{E}[\beta\lambda h S^2] - \beta^2 h r' \mathbf{E}[S^2T] - \mathbf{E}[T]. \end{aligned} \tag{3.13}$$

For fixed  $h$ , define the function  $F(z) = \mathbf{E} \tanh(\beta\lambda z + \beta g\sqrt{q_{\beta,\beta\lambda h}})$ , and denote  $Y = \beta\lambda z + \beta g\sqrt{q_{\beta,\beta\lambda h}}$ . Note that  $F$  is *not* the same as  $f_\beta$ , since we have fixed  $q = q_{\beta,\beta\lambda h}$  (in particular,  $q \neq q_{\beta,\beta\lambda z}$ ). We may then compute for  $z > 0$  that

$$F'(z) = \beta\lambda \mathbf{E} \operatorname{sech}^2(Y) > 0$$



$$F''(z) = -2\beta^2\lambda^2\mathbf{E}[\operatorname{sech}^2(\mathbf{Y})\tanh(\mathbf{Y})] < 0,$$

where the final inequality is Fact 3.7.2. Plugging this back into (3.13),

$$h^2 \cdot g'(h) = -\beta^2 h \cdot r' \cdot \mathbf{E}[S^2 T] + hF'(h) - F(h).$$

Lemma 3.7.3(ii) and Fact 3.7.2 imply that the first term is negative. For the other two terms, note that if we set  $G(z) = zF'(z) - F(z)$ , then because  $G'(z) = F''(z) < 0$  and  $G(0) = 0$ , we have  $G(h) < 0$ , completing the proof.  $\square$

**Lemma 3.7.6.** *Suppose that  $\beta = \lambda > 1$ . Then, there exists a fixed point  $h > 0$  of the function  $f_\beta$  such that  $(\lambda, \lambda^2 h)$  satisfies (AT).*

*Proof.* Consider the auxiliary function  $\psi$  defined by

$$\psi(\gamma) = \lambda^2 \mathbf{E} \left[ \tanh^2(\gamma + g\sqrt{\gamma}) \right].$$

[DAM16, Lemma 6.1] shows that for  $\lambda > 1$ ,  $\psi$  has a nonzero stable fixed point  $\gamma_\star$ . Also, as observed in [DAM16, Eq. (223)], for any  $\gamma > 0$  and  $k \in \mathbb{Z}_{>0}$ ,

$$\mathbf{E} \left[ \tanh^{2k}(\gamma + g\sqrt{\gamma}) \right] = \mathbf{E} \left[ \tanh^{2k-1}(\gamma + g\sqrt{\gamma}) \right]. \quad (3.14)$$

Essentially by definition coupled with the above identity, we have that  $\frac{\gamma_\star}{\lambda^2}$  is a fixed point of  $f_{\lambda, \lambda}$  with  $q_{\lambda, \gamma_\star} = \frac{\gamma_\star}{\lambda^2}$ .

To see why, let us first introduce the shorthand  $\mathbf{Y} = \gamma_\star + g\sqrt{\gamma_\star}$ . Then the stability of  $\gamma_\star$  as a fixed point of  $\psi$  implies that

$$\begin{aligned} 1 > \psi'(\gamma_\star) &= \lambda^2 \mathbf{E} \left[ 2 \tanh(\mathbf{Y}) \cdot \operatorname{sech}^2(\mathbf{Y}) \cdot \left( 1 + \frac{g}{2\sqrt{\gamma_\star}} \right) \right] \\ &= \lambda^2 \mathbf{E} \left[ 2 \tanh(\mathbf{Y}) \operatorname{sech}^2(\mathbf{Y}) + \operatorname{sech}^4(\mathbf{Y}) - 2 \operatorname{sech}^2(\mathbf{Y}) \tanh^2(\mathbf{Y}) \right], \end{aligned}$$

where the last line is Stein's lemma. We also have by (3.14) that

$$\mathbf{E} \left[ \tanh(\mathbf{Y}) \operatorname{sech}^2(\mathbf{Y}) \right] = \mathbf{E} \left[ \tanh(\mathbf{Y}) - \tanh^3(\mathbf{Y}) \right] = \mathbf{E} \left[ \tanh^2(\mathbf{Y}) - \tanh^4(\mathbf{Y}) \right] = \mathbf{E} \left[ \tanh^2(\mathbf{Y}) \operatorname{sech}^2(\mathbf{Y}) \right],$$

so  $\lambda^2 \mathbf{E}[\operatorname{sech}^4(\mathbf{Y})] < 1$  and  $(\lambda, \gamma_\star)$  satisfies (AT), as claimed.  $\square$

To conclude this section, let us put the pieces together and prove Proposition 3.7.1.

*Proof of Proposition 3.7.1.* For  $\beta < 1$ , note that  $(\beta, \beta\lambda h)$  satisfies (AT) for all  $h > 0$ . Consequently, Lemma 3.7.5(i) shows that in this regime, it suffices to understand whether  $\lim_{h \rightarrow 0^+} \frac{f_\beta(h)}{h}$  is greater or less than 1: if it is greater than 1, then 0 is an unstable fixed point and there is a unique fixed point at some positive  $h$  which is stable; if it is smaller than 1, then 0 is the unique fixed point and it is stable. Now Lemma 3.7.5(iii) establishes precisely where 0 is an unstable fixed point, which yields (i) and (ii). Next, (iii) follows from Lemma 3.7.5(i) and Lemma 3.7.6. Finally, (iv) is Lemma 3.7.5(ii).  $\square$

# Chapter 4

## Weak Poincaré Inequalities: MCMC Beyond Worst-Case Mixing

### 4.1 Introduction

A common task of interest in computer science, probability, and physics is to efficiently sample from Gibbs distributions. For a Hamiltonian energy function  $H : \Omega \rightarrow \mathbb{R}$  over state space  $\Omega \subseteq \mathbb{R}^N$ , the associated Gibbs distribution  $\mu_H$  is defined by  $d\mu_H(x) \propto \exp(H(x)) dx$ .

The class of *Markov chain Monte Carlo* (MCMC) algorithms is arguably the most widely used tool for sampling from Gibbs distributions. In this paradigm, one sets up a Markov chain  $P_H$  whose stationary distribution is  $\mu_H$ , and outputs the final state of a  $\text{poly}(N)$ -time random walk according to  $P_H$ . Common choices include the Glauber dynamics, for discrete state spaces such as  $\Omega = \{\pm 1\}^N$ , and the Langevin diffusion, for continuous state spaces such as  $\Omega = \mathbb{R}^N$  or  $\sqrt{N} \cdot \mathbb{S}^{N-1}$ .

To prove that such an algorithm indeed correctly samples from  $\mu_H$ , one bounds the *mixing time* of the Markov chain. A common route to prove a bound on the mixing time is to establish functional inequalities, such as *Poincaré* inequalities. There are now powerful frameworks for proving such functional inequalities, such as *spectral independence* [ALOG21] and *localization schemes* [CE22]. The development of these frameworks has led to a flurry of activity in analyzing mixing times of Markov chains, including the resolution of several long-standing open problems in the algorithmic theory of counting and sampling [ALGV24, ALOG21, AJK<sup>+</sup>22, EKZ22, CE22].

The implications of these inequalities are quite strong. In particular, they imply that for *any* initial distribution  $\nu$ , for an appropriate divergence function, a single step of the Markov

chain shrinks the distance to the stationary distribution by a significant multiplicative factor:

$$\text{Divergence}(P_H v \| \mu_H) \leq \left(1 - \frac{1}{\text{poly}(N)}\right) \text{Divergence}(v \| \mu_H).$$

The presence of such a functional inequality typically implies that a Markov chain mixes rapidly from a *worst-case initialization*.

**Sampling from random initializations.** Many natural Markov chains are expected to produce approximate samples from the Gibbs measure when started at a *random* initialization, but fail to mix rapidly from a worst-case initialization. Often, this is because the Gibbs measure contains pathological clusters (termed *metastable states* in the physics literature) that are essentially disconnected from most of the measure, and account for a vanishing fraction of the total mass. A Markov chain initialized in such a cluster will remain trapped inside it and fail to mix, and therefore methods that show mixing from worst-case initializations cannot give effective guarantees in such settings.

However, one may still hope to show that from a random initialization, the Markov chain samples from the non-pathological part of the Gibbs measure, which is statistically indistinguishable from the true Gibbs measure. In our work, we prove that under suitable conditions, the *simulated annealing* algorithm samples from a distribution close to the Gibbs measure.

**Simulated annealing.** In the simulated annealing algorithm, one defines a “schedule” of inverse temperatures, i.e. for  $i = 0, \dots, T$ , let  $\beta_i := i/T$ . The algorithm initializes at a sample from the uniform distribution  $\mu_{\beta_0 H}$ . Then, for  $i = 1, \dots, T$ , the  $i$ -th stage of the algorithm runs the Markov chain  $P_{\beta_i H}$  corresponding to  $\mu_{\beta_i H}$  for  $\text{poly}(N)$  time, initialized at the output of the previous stage.

The underlying idea of this algorithm is that, for  $T$  sufficiently large, the Gibbs distribution  $\mu_{\beta_{i-1} H}$  is a “warm start” for  $\mu_{\beta_i H}$ , i.e. an initialization with suitably bounded likelihood ratio with  $\mu_{\beta_i H}$ . So, if one could show that each of the Markov chains  $P_{\beta_i H}$  (approximately) mixes rapidly from a warm start, one may inductively argue that the output of the  $i$ -th stage of the algorithm is an approximate sample from  $\mu_{\beta_i H}$ . In other words, simulated annealing chains a sequence of warm starts together into a sampling algorithm.

This algorithmic idea is widely used empirically, and has also been employed to obtain algorithms for approximating the volumes of convex bodies [DFK91, DF91, LS90, KLS97], approximating the number of perfect matchings in a bipartite graph [JSV04], and sampling from the random field Ising model at sufficiently high temperatures [AEGP23], among others. However, we lack a general theory for why simulated annealing achieves provable guarantees beyond the settings of sampling from log-concave distributions and convex

bodies. Indeed, in contrast to the general recipes available to prove mixing from worst-case initialization, proofs of rapid mixing from warm starts often employ ad-hoc arguments.

One of the main contributions of this work is a framework for proving mixing from warm starts, which combined with the above discussion provides general sufficient conditions under which simulated annealing samples from the Gibbs measure. We achieve this by generalizing the frameworks of spectral independence and localization schemes, previously employed to prove mixing from worst-case initialization, to show mixing from warm starts (see Section 4.6 for details). As we discuss just below, our framework gives sampling guarantees for simulated annealing in regimes where mixing from worst-case initializations is provably false.

**Main application: spherical spin glasses.** In a *spherical mixed  $p$ -spin glass*,  $H : \sqrt{N} \cdot \mathbb{S}^{N-1} \rightarrow \mathbb{R}$  is a random Hamiltonian parameterized by coefficients  $\beta, \gamma_2, \dots, \gamma_{p_*} \geq 0$  where:

$$H(\sigma) = \beta \sum_{p \geq 2} \frac{\gamma_p}{N^{(p-1)/2}} \sum_{i_1, \dots, i_p} \mathbf{g}_{i_1, \dots, i_p} \sigma_{i_1} \cdots \sigma_{i_p}, \quad (4.1)$$

for i.i.d.  $\mathbf{g}_{i_1, \dots, i_p} \sim \mathcal{N}(0, 1)$ . The Gibbs distribution  $\mu_H$  is very well-studied in probability, statistical physics, and average-case algorithm design, as it simultaneously exhibits rich behavior and is amenable to analytic tools. Notably, this model undergoes numerous sharp phase transitions as one increases  $\beta$ . For small  $\beta$ , the model satisfies a Poincaré inequality [GJ19]. Beyond a *uniqueness* transition  $\beta_{\text{uniq}}$ , small isolated clusters in  $\mu_H$  known as *metastable states* start to appear [BJ24]. In particular, the natural Markov chain *Langevin diffusion* initialized from such states mixes slowly, thereby precluding a Poincaré inequality. However, these states account for a vanishing fraction of the measure under  $\mu_H$ , and the Langevin diffusion with a random initialization is expected to still mix rapidly over a  $1 - o_N(1)$  fraction of  $\mu_H$ , thereby producing a sample with vanishing total variation distance from  $\mu_H$ .

The threshold for efficient algorithmic sampling is believed to occur at the *shattering* transition  $\beta_{\text{shatter}}$ —beyond this transition, the Gibbs measure shatters into an exponential number of poorly-connected clusters with exponentially small mass, and mixing is provably slow [CHS93, EAMS23, GJK23]. It is expected that all efficient algorithms fail to sample from the Gibbs measure above  $\beta_{\text{shatter}}$ , and recently [EAMS23] gave rigorous evidence for this picture by showing that all *stable* algorithms fail.

We use our framework to prove that *annealed* Langevin diffusion, where one begins by running Langevin diffusion for  $\beta_0 = 0$ , and slowly increases the inverse temperature to the target  $\beta$ , samples from the spherical mixed  $p$ -spin glass. This leads to the first rigorous guarantee in this problem for a Markov chain beyond the uniqueness threshold.

**Theorem 4.1.1 (Informal).** *For any choice of  $\gamma_2, \dots, \gamma_{p_*}$ , there is a threshold  $\beta_{\text{SL}} \leq \beta_{\text{shatter}}$  such that for any  $\beta < \beta_{\text{SL}}$ , with probability at least  $1 - e^{-\Omega(N^{1/5})}$  over the randomness of  $H$ , annealed Langevin diffusion run for  $\text{poly}(N)$  time samples from a distribution whose total variation distance to  $\mu_H$  is at most  $e^{-\Omega(N^{1/5})}$ .*

The thresholds  $\beta_{\text{SL}}$  and  $\beta_{\text{shatter}}$  are formally defined as the supremal  $\beta$  such that the inequalities (SL) and (Non-shattering) below hold. The recent work [HMP24] produces a different sampling algorithm based on algorithmic stochastic localization, which succeeds to the same threshold  $\beta_{\text{SL}}$ ; see below for further discussion. This threshold is a fundamental barrier for stochastic localization based approaches, and we explain its physical significance in Remark 4.7.4.

**Remark 4.1.2.** In many models, we have  $\beta_{\text{uniq}} < \beta_{\text{SL}} < \beta_{\text{shatter}}$ , and  $\beta_{\text{SL}}$  is close to  $\beta_{\text{shatter}}$ . For example, for the pure  $p$ -spin models (where  $\gamma_p = 1$  and all the other  $\gamma_i$  are equal to 0),  $\beta_{\text{uniq}} \asymp (\log p)^{-1/2}$ , while  $\beta_{\text{SL}}, \beta_{\text{shatter}} \asymp 1$  and  $\beta_{\text{SL}}/\beta_{\text{shatter}}$  is bounded from below by the universal constant  $\sqrt{e}/2$ . See [HMP24, Remark 1.1, Eq. 1.8].

## Weak Poincaré inequalities and localization schemes

The starting point of our work is a relaxation of Poincaré inequalities, known as *weak Poincaré inequalities*, which can be leveraged to prove mixing from warm starts. To simplify the discussion, we restrict here to the setting of discrete Markov chains. Our main application is to a continuous Markov chain, namely the Langevin diffusion for a spherical spin glass, and we outline the differences in Remark 4.1.7 below.

Let  $P_H$  be a time-reversible Markov chain with stationary distribution  $\mu_H$ . For any functions  $f, g : \Omega \rightarrow \mathbb{R}$ , define the *Dirichlet form* as  $\mathcal{E}(f, g) := \mathbf{E}_{x \sim \mu_H} \mathbf{E}_{y \sim P_H} (f(x) - f(y))(g(x) - g(y))$ . We say  $P_H$  satisfies a *C-Poincaré inequality* if for any function  $f : \Omega \rightarrow \mathbb{R}$ :

$$\mathcal{E}(f, f) \geq C \cdot \text{Var}[f],$$

for  $C \geq 1/\text{poly}(N)$ . A Poincaré inequality has a classic implication for rapid mixing. In particular, for  $\nu_t$  as the distribution obtained by running  $P_H$  for continuous time  $t$  starting at a distribution  $\nu_0$ , we have:

$$\chi^2(\nu_t \| \mu_H) \leq \exp(-Ct) \cdot \chi^2(\nu_0 \| \mu_H).$$

We say  $P_H$  satisfies a  $(C, \varepsilon)$ -weak Poincaré inequality if for any function  $f : \Omega \rightarrow \mathbb{R}$ :

$$\mathcal{E}(f, f) \geq C \cdot \text{Var}[f] - \varepsilon \cdot \|f - \mathbf{E}f\|_\infty^2$$

One can derive the following mixing guarantee from a weak Poincaré inequality; see, e.g., [RW01, Theorem 2.1].

$$\chi^2(v_t \| \mu_H) \leq \exp(-Ct) \cdot \chi^2(v_0 \| \mu_H) + \varepsilon \cdot \left\| \frac{dv_0}{d\mu_H} - 1 \right\|_\infty^2. \quad (4.2)$$

In particular, if  $v_0$  is a warm start for  $\mu_H$  in the sense that  $\left\| \frac{dv_0}{d\mu_H} - 1 \right\|_\infty$  is suitably small, this implies that the Markov chain's output distribution  $v_t$  approximates  $\mu_H$ .

Since the target measure in one stage of simulated annealing is a warm start for that of the next stage, such a guarantee allows one to inductively argue that simulated annealing succeeds at sampling. We summarize this implication below.

**Theorem 4.1.3** (Informal, see Theorem 4.4.12). *If  $P_{\beta H}$  satisfies a weak Poincaré inequality with suitable parameters for every  $\beta \in [0, 1]$ , then simulated annealing succeeds at sampling from  $\mu_H$ .*

**Localization schemes for weak Poincaré inequalities.** We restrict to the following simple setting:  $\mu_H$  is a distribution on  $\{\pm 1\}^N$ . Let  $P_H$  be the Glauber dynamics Markov chain where in a single step from  $x$ , we pick a uniformly random coordinate  $i \sim [N]$ , and toggle  $x_i$  with probability:

$$\frac{\mu_H(x^{\oplus i})}{\mu_H(x) + \mu_H(x^{\oplus i})}.$$

A special case of the localization schemes framework is the *spectral independence* framework of Anari, Liu, and Oveis Gharan [AL20, ALOG21].

**Theorem 4.1.4** (Spectral independence). *The following local-to-global principle reduces proving a Poincaré inequality to establishing bounds on the spectrum of influence matrices. Suppose for every  $S \subseteq [N]$ , and every pinning  $x_S$  of coordinates in  $S$ , the spectral norm of its influence matrix  $\Psi_{S, x_S}$  is at most  $\alpha$ , then the Glauber dynamics chain satisfies a  $n^{-O(\alpha)}$ -Poincaré inequality. Here, the influence matrix  $\Psi_{S, x_S}$  is an  $(n - |S|) \times (n - |S|)$  matrix indexed by vertices  $v \notin S$ , where*

$$\Psi_{S, x_S}[a, b] := \Pr[x_a = +1 | x_b = +1] - \Pr[x_a = -1 | x_b = +1].$$

While the above theorem has been influential and useful in proving mixing time bounds for a variety of Markov chains relevant to sampling combinatorial structures, the “for every” requirement in the above theorem is quite punishing in average-case settings. For example, in the presence of metastable states,  $P_H$  does not satisfy a Poincaré inequality, but may nevertheless satisfy a weak Poincaré inequality. In such cases, there are choices of  $S$  and  $x_S$  for which  $\Psi_{S, x_S}$  has large spectral norm, and the above statement has no implications for the mixing time of  $P_H$ .

We give a general local-to-global principle to prove weak Poincaré inequalities. A one-line summary of this local-to-global principle is:

*Bounded influence over **all** pinnings implies a Poincaré inequality.*

An analogous summary of the local-to-global principle in the present paper is:

*Bounded influence over **typical** pinnings implies a **weak** Poincaré inequality.*

To give a more concrete instantiation of our message, our result implies a “softer” version of Theorem 4.1.4, tolerant to some “bad” pinnings, which we state below.

**Theorem 4.1.5** (Special case of Lemma 4.6.8 and Remark 4.6.9). *Let  $i_1, \dots, i_N$  be a random permutation of  $[N]$ , let  $S_t := \{i_1, \dots, i_t\}$ , and let  $\mathbf{x} \sim \mu_H$ . Suppose with probability  $1 - \varepsilon$  over the randomness of  $\mathbf{x}$  and the permutation  $i_1, \dots, i_N$ , we have that for every  $t \in [N]$ , the influence matrix  $\Psi_{S_t, \mathbf{x}_{S_t}}$  has spectral norm bounded by  $\alpha$ . Then,  $P_H$  satisfies a  $(n^{-O(\alpha)}, O(\varepsilon))$ -weak Poincaré inequality.*

**Remark 4.1.6.** The reader should think of the spectral norm of  $\Psi_{S_t, \mathbf{x}_{S_t}}$  as quantifying how much variance of the distribution  $\mu_H|_{\mathbf{x}_{S_t}}$  is “explained” by revealing  $\mathbf{x}_{i_{t+1}}$ .

**Remark 4.1.7.** Theorem 4.1.5 holds at a more general level, for a large family of *localization schemes*; see [CE22] for examples of localization schemes and further discussion. The localization scheme at play in the above local-to-global principle is process of revealing coordinates of a Gibbs sample  $\mathbf{x}$  in random order.

In our main application of sampling from a spherical spin glass using simulated annealing of Langevin diffusion, we consider a different localization scheme, *stochastic localization*, where the revealed information at time  $t$  is  $\mathbf{y}_t = t\mathbf{x} + B_t$  where  $(B_t)_{t \geq 0}$  is a standard Brownian motion. Analyzing this localization scheme requires studying *exponential tilts* rather than pinnings of  $\mu_H$ . The analogous local-to-global principle in this setting is:

*Bounded covariance over **typical** exponential tilts implies a **weak** Poincaré inequality.*

We defer a technical discussion to Section 4.2, and refer to Lemma 4.6.8 and Remark 4.6.9 for a formal statement.

## Sampling from spherical spin glasses

We now state our main results for sampling from spherical spin glasses. We will encode the coefficients  $\gamma_2, \dots, \gamma_{p^*}$  in (4.1) into the *mixture function*  $\xi(q) = \sum_{p=2}^{p^*} \gamma_p^2 q^p$ . Note that the parameter  $\beta$  in (4.1) can of course be absorbed into the  $\gamma_p$ , so we can state thresholds directly

in terms of the function  $\xi$ . Physics heuristics [CHS93] suggest that Glauber dynamics and Langevin diffusion, with random initialization, sample from  $\mu_H$  with vanishing total variation error under the following condition. Note that this and the below conditions take the form of an upper bound on  $\xi$  or its derivatives, and therefore demarcate a region of sufficiently high temperature.

$$\xi'(q) < \frac{q}{1-q} \text{ for all } q \in (0, 1). \quad (\text{Non-shattering})$$

Recent work by one of the authors, Montanari, and Pham [HMP24] gives an algorithm based on simulating Eldan's stochastic localization process [Eld13, Eld20] (see below), which samples from  $\mu_H$  with vanishing total variation error under the following condition.

$$\xi''(q) < \frac{1}{(1-q)^2} \text{ for all } q \in [0, 1). \quad (\text{SL})$$

Note that this condition implies (Non-shattering), which can be seen by integrating the inequality. [HMP24] also shows a matching hardness result, that for any strictly replica symmetric model (see (Strict RS) below) *not* satisfying (SL), a generalized family of stochastic localization algorithms fails to sample from  $\mu_H$ .

Our main result is that simulated annealing samples from  $\mu_H$  in the same regime.

**Theorem 4.1.8** (See Theorem 4.7.2). *Under (SL), with probability at least  $1 - e^{-\Omega(N^{1/5})}$  over the randomness of  $H$ , annealed Langevin dynamics produces a sample whose total variation distance to  $\mu_H$  is at most  $e^{-\Omega(N^{1/5})}$ .*

As alluded to in the above discussion, the main input to our framework is a high-probability covariance bound on the *random* exponential tilts of the Gibbs measure encountered along the stochastic localization process. Combined with our weak Poincaré inequality framework, this implies that simulated annealing samples from the Gibbs measure. On the way to proving these covariance bounds, we establish a high-probability covariance bound on all spherical spin glasses in the (*strictly*) *replica symmetric* phase, a high-temperature phase where the model enjoys a certain notion of correlation decay.

$$\xi''(0) < 1 \text{ and } \xi(q) + q + \log(1-q) < 0 \text{ for all } q \in (0, 1). \quad (\text{Strict RS})$$

**Theorem 4.1.9** (Informal, see Theorem 4.7.27). *Under (Strict RS), with probability  $1 - e^{-\Omega(N^{1/5})}$  over the randomness of  $H$ ,  $\|\text{Cov}(\mu_H)\|_{\text{op}} = O(1)$ .*



This is the first covariance bound to cover the entire replica symmetric phase with higher order interactions, and we believe it is interesting in its own right. This result is sharp: in the complement of the replica symmetric regime, arguing as in [AG24, Proposition 4.2] shows that  $\mathbf{E}\|\text{Cov}(\mu_H)\|_{\text{op}}$  is diverging, of order  $\Omega(\sqrt{N})$ .

The relation between (SL) and (Strict RS) is as follows. First, (Strict RS) follows from (SL) by integrating twice. Second, (SL) is equivalent to the condition that random exponential tilts of  $\mu_H$  of any magnitude are typically replica symmetric. This is needed for the algorithmic stochastic localization approach of [HMP24], and arises in the current work (where stochastic localization appears as an analysis tool, rather than as an algorithm) for a similar reason, see Remark 4.7.4.

The connection from Theorem 4.1.9 to high-probability covariance bounds on the tilted measures encountered along the localization process relies on a reduction developed in [HMP24]. This reduction implies that typically, the vast majority of the mass of these tilted measures live near a certain codimension-2 band passing through a *TAP fixed point*, which behaves like a spin glass in two fewer dimensions. The proof of Theorem 4.1.9 also builds on tools developed in [HMP24], and by one of the authors and Sellke in [HS23], which together provide high-precision control of partition functions in the replica symmetric regime.

On the other hand, our approach also leads to several improvements over earlier results. First, we obtain a sampler with total variation error  $e^{-\Omega(N^{1/5})}$  with probability  $1 - e^{-\Omega(N^{1/5})}$ , whereas [HMP24] obtains total variation error  $N^{-\varepsilon}$  with probability  $1 - N^{-\varepsilon}$ , for small constant  $\varepsilon$ . Our total variation error is close to the best possible, as beyond the uniqueness threshold, at least a  $e^{-O(N)}$  fraction of  $\mu_H$  is typically trapped in metastable states [BJ24], which are hard to reach. Moreover, there is no longer a need to encode a mean estimator for the stochastic localization process (see below) directly in the algorithm; running a natural Markov chain is sufficient.

More conceptually, our work gives the first analysis of a Markov chain for this problem that “sees” the benignness of a random initialization and overcomes the uniqueness threshold.

## Weak Poincaré inequalities beyond annealing

The discussion thus far has been focused on proving mixing time bounds for Markov chains initialized at warm starts. In fact, our framework extends beyond this and can be used to prove rapid mixing of a Markov chain initialized at a distribution that “sees” the different components of the target distribution. For instance, consider the simple scenario where the target distribution  $\pi$  is a mixture of two disconnected component distributions, each of which satisfies a (true) Poincaré inequality. The disconnectedness means that the full

distribution  $\pi$  does not satisfy a true Poincaré inequality. However, if we initialize at a distribution that splits its mass equally between the two components, we would expect a Markov chain to rapidly mix to the target distribution.

How does one convert this belief to a (generalizable) proof? The key is that while the distribution may not satisfy a Poincaré inequality for *all* functions, a variant of such an inequality does hold for functions encountered along the trajectory of the Markov chain. More concretely, we may prove the following theorem.

**Theorem 4.1.10** (Informal, see Theorem 4.4.6). *Consider the trajectory  $(v_t)_{t \geq 0}$  of a Markov chain with stationary distribution  $\pi$ , initialized at a distribution  $v_0$ . Suppose that for all  $s \leq T$ ,*

$$\mathcal{E} \left( \frac{dv_s}{d\pi}, \frac{dv_s}{d\pi} \right) \geq \rho_{\text{PI}} \left( \mathbf{Var}_{\pi} \left[ \frac{dv_s}{d\pi} \right] - \delta \right).$$

Then,

$$\chi^2(v_T \parallel \pi) \leq e^{-2\rho_{\text{PI}}T} \chi^2(v_0 \parallel \pi) + \delta.$$

We remark that our earlier equation (4.2) is a near-immediate consequence of the above. Returning to the above example with two disconnected components, if  $v_s$  placed exactly half its mass on each of the two components, the error  $\delta$  can be taken to be 0.

For our first application in Section 4.5, we use this picture of how the initialization can capture “symmetries” in the distribution.

**Sampling from mixtures of log-concave distributions with advice.** An example of a distribution where we can take advantage of “symmetries” is the following. Suppose we have a distribution  $\pi$  which is a mixture of  $K$  distributions

$$\pi = \sum_{i=1}^K p_i \pi_i,$$

each of which is well-connected (e.g., satisfies a Poincaré inequality). We do not expect a Markov chain to rapidly mix to  $\pi$  from a worst-case initialization. Does the scenario change if we initialize more cleverly? To be concrete, suppose we are given  $m$  samples  $x_1, \dots, x_m$  from  $\pi$ , and initialize our Markov chain at the empirical distribution  $\sum_{i=1}^m \delta_{x_i}$ . If the component measures  $(\pi_i)$  are “far apart” and do not interact with each other, we would expect the Markov chain to rapidly mix from this initialization if the fraction of points in each cluster is (approximately) equal to the correct fraction  $p_i$ . On the other hand, if the component measures were very close together, we would expect their mixture to also satisfy a Poincaré inequality.

However, it is unclear how to translate this intuition to a proof. In previous work [KV23], sampling guarantees are provided for this algorithm, but the running time has a doubly exponential dependence on the number of components  $K$ . Our second illustration of weak Poincaré inequalities provides high-probability sampling guarantees for this problem, by running Langevin diffusion for time that is polynomial in all parameters involved. We refer the reader to Section 4.5 for the details of the theorem statement and its (self-contained) proof.

This problem is studied extensively in an independent work of Koehler, Lee, and Vuong [KLV24]. Motivated by the success of *score matching* methods in modern machine learning, they prove that Langevin dynamics and Glauber dynamics converge to the stationary distribution when initialized from the above empirical distribution under similar conditions to our setting, even if the Markov chain updates come from a slightly perturbed distribution (i.e. if they were learned by a score matching algorithm). They also use their techniques to give an efficient algorithm for learning approximately low-rank Ising models.

## Related work

**Markov chain mixing and localization schemes.** The first use of the local-to-global phenomenon in mixing was in the work of [ALGV24] on establishing rapid mixing of the “basis exchange” walk on bases of a matroid, which used the local-to-global theorem for simplicial complexes from [KO20]. Their approach was later formalized into the framework of *spectral independence* [ALOG21], which was widely successful in resolving numerous problems in algorithmic sampling and counting; see [Liu23] for a comprehensive literature survey.

In the world of sampling from continuous distributions, most recent progress on the KLS conjecture on the Poincaré constant of isotropic log-concave distributions (see [LV24] and the recent survey [KL24]) has employed Eldan’s stochastic localization [Eld13]. Later, stochastic localization was used in the work of Eldan, Koehler, and Zeitouni [EKZ22] to analyze the Poincaré constant for Glauber dynamics on Ising models. The seemingly unrelated techniques of spectral independence and stochastic localization approaches to analyzing mixing times were unified under the framework of localization schemes [CE22], which, as an application, also simplified the proof of [EKZ22].

**Weak Poincaré inequalities.** The study of weak Poincaré inequalities was initiated in the work of Aida [Aid98] and Mathieu [Mat06] in the context of proving other functional inequalities. The work of Röckner and Wang [RW01] observed the connection between a Markov chain satisfying a weak Poincaré inequality, and rapid mixing from “sufficiently warm starts”. We refer the reader to the monograph of Wang [Wan06, Chapter 4] for a

comprehensive treatment of weak Poincaré inequalities and their implications to mixing and concentration.

Weak Poincaré inequalities are also related to the notion of  $s$ -conductance, a weakened version of conductance introduced in [LS93] which has been used frequently in the literature on sampling from convex bodies (see [Che23, Section 7.4.2] for a textbook treatment). This connection is explained in [GMT06]. We also refer the reader to [CGG07], which defines a notion of weak log-Sobolev inequality and uses it to derive a rapid mixing result.

The work [AEGP23] gives a sampling algorithm for the ferromagnetic random-field Ising model on a finite domain  $D \subseteq \mathbb{Z}^d$ , which follows an approach of chaining warm starts similar to the present work, inspired by convex body sampling literature [LS93]. [AEGP23] shows that in a certain parameter regime, the Glauber dynamics for this model satisfy a weak Poincaré inequality. They then construct an increasing sequence of sub-domains  $D_0 \subset D_1 \subset \dots \subset D_T = D$  and show that a sample from the model on  $D_i$  can be converted to a warm start for the model on  $D_{i+1}$ . Since the weak Poincaré inequality implies mixing from a warm start, this yields a sampling algorithm based on running the Glauber dynamics on this increasing sequence of models.

The work [AJK<sup>+</sup>21b] introduces a related notion of restricted modified log-Sobolev inequality, which implies entropy contraction (without an additive error, in contrast to a weak Poincaré inequality) for all warm starts. This is used to derive optimal mixing times for several Markov chains.

**Sampling from random initializations.** The separation between worst-case mixing times and mixing from a random initialization has been studied in a variety of other settings. [CDL<sup>+</sup>12, BGZ24] characterize which product measure initializations enjoy rapid mixing in a temperature range where worst-case mixing is exponential for the Curie-Weiss Potts model. Notably, as discussed in [BGZ24, Section 1.3], their analysis also characterizes mixing from initializations constructed by simulated annealing, [LS16, LS17] show that a uniform initialization halves the mixing time for Glauber dynamics for the ferromagnetic Ising model on bounded degree graphs, such as the 1D lattice. [GS22] introduces the notion of weak spatial mixing in a phase, and proves that Glauber dynamics for the ferromagnetic Ising model on the 2D lattice has rapid mixing when initialized uniformly at  $\pm \vec{1}$ . [GS24] uses the same notion to study mixing from a similar random initialization for a certain natural Markov chain for the random cluster model. [BNN24] show rapid mixing for Glauber dynamics for the exponential random graph model when initialized from a carefully chosen Erdős–Rényi random graph.

**Sampling from spherical spin glasses and algorithmic stochastic localization.** There is a long history of work studying Markov chain dynamics on spin glasses. An important line of work [CHS93, CK93, BCKM98, BDG06, BAGJ20, CCM21, Sel24b] studies the Langevin

dynamics for spherical spin glasses on an  $N$ -independent time scale. While the Langevin dynamics do not mix on this time scale, these works capture important statistics of the trajectory such as the energy attained by the Langevin dynamics after a given time, and uncover deep phenomena such as *aging*.

Rapid mixing guarantees at sufficiently high temperature were obtained in [GJ19] for the Langevin dynamics for spherical spin glasses, and in [BB19, EKZ22, AJK<sup>+</sup>22, ABXY24, AJK<sup>+</sup>24, AKV24] for the Glauber dynamics for the Sherrington–Kirkpatrick model [SK75] and Ising spin glasses. These approaches show mixing from a worst-case initialization via a functional inequality.

Recently, [EAMS22, AMS23] introduced a new sampling algorithm based on simulating Eldan’s stochastic localization scheme [Eld13, Eld20]. This approach has since been used in applications such as Bayesian posterior sampling [MW23, MW24], and is closely related to the denoising diffusions method in machine learning [SDWGM15, HJA20, SSDK<sup>+</sup>21] (see [Mon23] for details). The resulting algorithm samples in a wider range of temperatures, though with the weaker guarantee of vanishing *Wasserstein* rather than total variation error. The recent work [HMP24] improved this guarantee to total variation, and the resulting algorithm succeeds to the same threshold (SL) as in the present work.

Within the algorithmic stochastic localization approach, the main task is to estimate the means of a sequence of exponential tilts of the Gibbs measure, which appear as the drift process of a stochastic differential equation parametrizing the localization process. In [EAMS22], this is achieved with an estimator based on approximate message passing (AMP), which is accurate to leading order. [HMP24] develops an improved estimator with a suitable correction term, which improves the algorithm’s guarantee from *Wasserstein* to total variation error.

**Covariance bounds for spin glasses.** There has been a great deal of recent work on covariance bounds for spin glasses [BXY23, AG24, BSXY24], in part due to the connection between covariance bounds and functional inequalities developed in the localization schemes literature. In particular, [AG24, BSXY24] address the case of the Sherrington–Kirkpatrick (SK) model, and [BXY23] addresses the SK model with external field.

## Open problems

We conclude with several open problems.

**Non-sampling guarantees for simulated annealing.** While we initiate a study of simulated annealing to attain sampling guarantees, one could ask how to analyze simulated annealing beyond sampling. In recent work [LMR<sup>+</sup>24], three of the authors, Liu, and Raghavendra introduce the framework of *locally stationary distributions* to analyze slow-

mixing Markov chains, and leverage it to obtain recovery guarantees for the spiked Wigner and stochastic block model inference problems. We start by reiterating [LMR<sup>+</sup>24, Problems 1.20 and 1.21]—is simulated annealing computationally optimal for random CSPs with planted solutions?

Further, consider the problem of optimizing the Hamiltonian (4.1) of the mixed  $p$ -spin model. Historically, simulated annealing was one of the earliest algorithms developed for this problem [CHKW23]. The works [Mon25, Sub21, AMS21, Sel24a] develop algorithms that are optimal among suitably Lipschitz algorithms [HS22] and conjecturally among all efficient algorithms. The limiting energy obtained by natural Markov chain dynamics is a long-standing question in its own right [CK93], which was solved for pure models in [Sel24b] but is otherwise open. We ask:

**Problem 4.1.11.** What energy does simulated annealing obtain when run on the Hamiltonian (4.1)?

We refer the reader to [MRT04, FFRT21] and references therein for relevant discussion. We also ask the following question, which seems instrumental to making progress towards the above.

**Problem 4.1.12.** How does a non-worst-case initialization (such as one constructed by simulated annealing) affect the locally stationary distribution that is reached by a Markov chain?

Along similar lines, we have the following concrete question about understanding Markov chains from non-worst-case initializations.

**Worst-case combinatorial optimization via simulated annealing.** The paradigm of solving a semidefinite program and rounding its solution has been extremely successful at achieving optimal approximation guarantees for a wide variety of combinatorial optimization problems, especially constraint satisfaction problems [KKMO07, Rag08].

However, on large families of instances (sparse ones for instance), the solutions produced by these SDPs can be refined locally to improve the approximation ratio, but these improvements do not match the corresponding hardness thresholds. For example, for the problem of Max Cut, the classical SDP algorithm [GW95] gives an  $\alpha_{\text{GW}}$ -approximation for  $\alpha_{\text{GW}} \approx 0.878$ , and a local refinement [HK22] produces an  $\alpha_{\text{GW}} + \Omega\left(\frac{1}{d^2}\right)$ -approximation. On the other hand, it is (UG-)hard [Tre01] to approximate the max-cut better than  $\alpha_{\text{GW}} + O\left(\frac{1}{\sqrt{d}}\right)$ .

**Problem 4.1.13.** Does a Markov chain initialized at the SDP solution attain a  $\alpha_{\text{GW}} + \Omega\left(\frac{1}{\sqrt{d}}\right)$ -approximation to the max-cut in a bounded degree graph?

**Sampling from spin glasses up to the shattering threshold.** It is conjectured that the Langevin diffusion with uniform random initialization samples from spherical  $p$ -spin models for inverse temperatures up to the shattering threshold (Non-shattering) [CHS93, CK93]. Similarly, this is conjectured for the Glauber dynamics Markov chain for models over the hypercube  $\{\pm 1\}^N$  instead of the sphere  $S_N$ , for an analogous shattering threshold. As a start, can we show such guarantees for simulated annealing (as opposed to a fixed-temperature Markov chain from uniform initialization)?

**Problem 4.1.14.** Does simulated annealing sample from  $p$ -spin models up to the shattering threshold?

The failure of algorithmic stochastic localization beyond the (SL) condition [HMP24, Section 10] suggests that ideas beyond our proof strategy are required to prove the above.

**Simulated annealing in more general models.** For sampling from the spherical  $p$ -spin model, our results show that simulated annealing succeeds in the regime (SL) where algorithmic stochastic localization succeeds. At the level of proofs, these methods are also closely related, as both revolve around suitable control of the localization process: in the algorithmic stochastic localization approach, this is used to construct a mean estimator for the localized measures, and in our approach it is used to bound the localized measures' covariances. These tasks are closely linked; see Remark 4.7.4.

One question is whether simulated annealing succeeds in more general models. In particular, samplers based on algorithmic stochastic localization have been developed for the Sherrington–Kirkpatrick model in the replica symmetric regime [EAMS22, Cel24],  $p$ -spin models over the hypercube [AMS23], and posteriors of spiked matrix models [MW23]. These samplers are proven to have vanishing Wasserstein error, and sampling with vanishing total variation error remains an open problem in these models. It would be interesting to show that simulated annealing achieves this. More speculatively, we may ask if there is a general reduction from a sampling guarantee for algorithmic stochastic localization to one for simulated annealing.

**#BIS.** A major open problem in the field of approximate counting is settling the complexity of #BIS: where the algorithmic task is to approximate the number of independent sets in a bipartite graph. So far, algorithmic progress for this problem has been limited to restricted classes of graphs, such as lattices & tori [HPR19], and expander graphs [JKP20]. Numerous interesting approximate counting problems have been shown to be #BIS-hard [CGM12, GJ12, CGG<sup>+</sup>16, GSVY16]. While vanilla Glauber dynamics fails at the corresponding sampling task, it is plausible that a variant of simulated annealing succeeds.

**Problem 4.1.15.** Does (a simple variant of) simulated annealing succeed at sampling a uniformly random independent set in a bipartite graph?

**Structural guarantees from weak Poincaré inequalities.** According to physics heuristics, the Gibbs measure of a spherical mixed  $p$ -spin glass between  $\beta_{\text{uniq}}$  and  $\beta_{\text{shatter}}$  consists of one main cluster accounting for nearly all the mass, and metastable states with exponentially small mass that are poorly connected to the main cluster and each other. We do not prove this picture, but the weak Poincaré inequality we obtain (up to  $\beta_{\text{SL}}$ ) is sufficient to imply a sampling guarantee for simulated annealing. One open direction is to show that the above picture holds, and that the main cluster satisfies a genuine Poincaré inequality. More generally, one may ask:

**Problem 4.1.16.** If a distribution satisfies a weak Poincaré inequality, is it TV-close to a distribution satisfying a true Poincaré inequality?

We note that Lemmas 4.4.9 and 4.9.5 show a converse of this statement, that if we perturb a distribution satisfying a true Poincaré inequality (for the Langevin diffusion or Glauber dynamics Markov chains), the resulting distribution satisfies a weak Poincaré inequality.

## Organization

In Section 4.2, we give a technical overview of how we use weak Poincaré inequalities to analyze simulated annealing for our main application of sampling from spherical  $p$ -spin distributions.

In Section 4.3, we cover some basic preliminaries that will be useful. Then, in Section 4.4, we formally define weak functional inequalities and establish some of their basic properties.

In Section 4.5, we demonstrate the effectiveness of this framework by showing how to sample from a mixture of distributions satisfying Poincaré inequalities from data-based initializations.

Our main application to spherical  $p$ -spin models spans Sections 4.6 to 4.8, and requires more background in stochastic localization and spin glass theory. In Section 4.6, we review some basic properties of stochastic localization and show how to adapt the framework of localization schemes from [CE22] to prove weak functional inequalities. Then, in Section 4.7, we initiate the discussion of weak Poincaré inequalities for spherical  $p$ -spin models. To assist the reader in understanding the proof of a weak Poincaré inequality, we provide a separate technical overview in Section 4.7. The rest of Section 4.7 reduces the proof to proving high-probability covariance bounds for strictly replica-symmetric models with small external field, which is then established in Section 4.8.



**Acknowledgements.** BH is extremely grateful to Andrea Montanari and Huy Tuan Pham for early discussions on this problem, and to Ahmed El Alaoui, Sinho Chewi, and Mark Sellke for enlightening conversations. We would also like to thank Sitan Chen, Jason Gaitonde, Kuikui Liu, and Francisco Pernice for insightful discussions. We would like to thank Thiago Bergamaschi for pointing out an error in an application to the ferromagnetic Potts model in an earlier version of this paper.

## 4.2 Technical overview

Let  $H$  be a Hamiltonian on state space  $\Omega$ , and let  $\mu_H$  be its Gibbs distribution. Our goal in this section is to describe our strategy to prove that simulated annealing succeeds at sampling. In our application,  $\Omega$  is the scaled sphere  $S_N := \sqrt{N} \cdot \mathbb{S}^{N-1}$ , and  $\mu_H$  comes with an associated Markov chain known as *Langevin diffusion*, which we denote with  $P_H$ . For ease of exposition, we restrict the discussion to this setting, though much of it holds in a more general setting.

**Definition 4.2.1** (Simulated annealing, informal). Initialize at the uniform distribution on  $S_N$  (which is equal to  $\mu_0$ ), and for each  $i \in [m]$ , run  $P_{\frac{i}{m}H}$  for time  $T$ .

For the sequel, we abbreviate  $P_{\frac{i}{m}H}$  and  $\mu_{\frac{i}{m}H}$  as  $P_i$  and  $\mu_i$ , and we use  $P_{i,t}$  to denote running  $P_i$  for time  $t$ . The strategy to prove that simulated annealing succeeds at sampling is to establish a weak Poincaré inequality for  $P_i$  for all  $i$ .

Let  $L$  be the infinitesimal generator of  $P_i$ . For functions  $f, g : \Omega \rightarrow \mathbb{R}$ , we define the *Dirichlet form*  $\mathcal{E}(f, g)$  as  $\mathbf{E}_{\mu_i}[f L g]$ .

**Remark 4.2.2.** In the case of Langevin diffusion for a distribution  $\pi$ , the Dirichlet form can be evaluated as

$$\mathcal{E}(f, g) = \mathbf{E}_{\mu_i}[\langle \nabla f, \nabla g \rangle],$$

where  $\nabla$  denotes the Euclidean gradient if  $\pi$  is supported on  $\mathbb{R}^N$ , and the Riemannian gradient on  $S_N$  if  $\pi$  is supported on  $S_N$ .

As discussed in Section 4.1, we say a Markov chain satisfies a *weak Poincaré inequality* with parameters  $(C, \varepsilon)$  if

$$\mathcal{E}(f, f) \geq C \cdot \mathbf{Var}[f] - \varepsilon \cdot \|f - \mathbf{E}f\|_\infty^2 - \varepsilon \cdot \sup_{x \in \Omega} \|\nabla f(x)\|^2,$$

which implies the following mixing result Theorem 4.4.6 for the chi-squared divergence; see also [RW01, Theorem 2.1]. Defining  $\nu_t$  as the distribution after running the Markov

chain for time  $t$  from initial distribution  $\nu_0$ , we have

$$\chi^2(\nu_t \parallel \mu_i) \leq \exp(-Ct) \cdot \chi^2(\nu_0 \parallel \mu_i) + \varepsilon \cdot \left\| \frac{d\nu_0}{d\mu_i} - 1 \right\|_\infty^2 + \varepsilon \cdot \sup_{x \in \Omega} \left\| \nabla \frac{d\nu_0}{d\mu_i}(x) \right\|^2.$$

**Analyzing simulated annealing with weak Poincaré inequalities.** To see why the above statement plays well with simulated annealing, imagine plugging in initialization  $\nu_0 = \mu_{i-1}$ . By selecting the number of annealing steps  $m = \text{poly}(N)$ , we can ensure  $\left\| \frac{d\nu_0}{d\mu_i} - 1 \right\|_\infty$  and  $\sup_{x \in \Omega} \left\| \nabla \frac{d\nu_0}{d\mu_i}(x) \right\|$  are  $O(1)$ . The guarantee after running the Markov chain for some sufficiently large polynomial time  $T$  is then

$$\chi^2(P_{i,T} \mu_{i-1} \parallel \mu_i) \leq O(\varepsilon),$$

which in particular implies

$$d_{\text{TV}}(P_{i,T} \mu_{i-1}, \mu_i) \leq O(\sqrt{\varepsilon}).$$

When we combine the above with the data processing inequality, we then get the following guarantee for  $\nu_{m,T} := P_{m,T} \cdots P_{2,T} P_{1,T} \nu_0$ , the distribution that simulated annealing samples from.

$$\begin{aligned} d_{\text{TV}}(P_{m,T} \cdots P_{1,T} \mu_0, \mu_m) &\leq d_{\text{TV}}(P_{m,T} \cdots P_{1,T} \mu_0, P_{m,T} \mu_{m-1}) + d_{\text{TV}}(P_{m,T} \mu_{m-1}, \mu_m) \\ &\leq d_{\text{TV}}(P_{m-1,T} \cdots \mu_0, \mu_{m-1}) + O(\sqrt{\varepsilon}). \end{aligned}$$

Applying the above inequality  $m$  times tells us that  $d_{\text{TV}}(\nu_{m,T}, \mu_m) \leq O(\sqrt{\varepsilon} \cdot m)$ .

We now turn our attention to the proof technique for showing a weak Poincaré inequality.

**How to prove weak Poincaré inequalities.** Suppose our goal is to prove a weak Poincaré inequality for a measure  $\pi$ . The high-level strategy in the localization schemes approach for proving a weak Poincaré inequality is to design a *measure decomposition* of  $\pi$ : for some mixture distribution  $\rho$ , express  $\pi$  as  $\mathbf{E}_{z \sim \rho} \pi_z$ . Refer to Section 4.3 for a brief review of measure decompositions. Once we have a measure decomposition in hand, establishing the following simple set of inequalities forms the crux of the argument. Let  $f$  be a function such that  $\mathbf{E}_\pi f = 1$ .

1. *Conservation of Dirichlet form.*

$$\mathcal{E}_\pi(f, f) \geq \mathbf{E}_{z \sim \rho} \mathcal{E}_{\pi_z}(f, f).$$

In the case of Langevin diffusion, this is an equality, and in the case of Glauber dynamics, the inequality is true by a generic concavity argument; see, e.g. [AJK<sup>+</sup>22, Page 19] or [LMRW24, Page 5].

2. *Weak Poincaré inequality for good component measures.*

$$\mathbf{E}_{z \sim \rho} \mathcal{E}_{\pi_z}(f, f) \geq C \cdot \mathbf{E}_{z \sim \rho} \mathbf{Var}_{\pi_z}[f] \cdot \mathbf{1}[z \text{ “good”}] - \varepsilon \|f - 1\|_{\infty}^2 - \varepsilon \|\nabla f\|_{\infty}^2,$$

where the “good”  $\pi_z$  are those which satisfy a  $(c, \varepsilon)$ -weak Poincaré inequality. This inequality follows from the nonnegativity of norms and Dirichlet forms.

3. *Approximate conservation of variance.*

$$\mathbf{E}_{z \sim \rho} \mathbf{Var}_{\pi_z}[f] \geq \alpha \cdot \mathbf{Var}_{\pi}[f].$$

This is one of the parts that depends nontrivially on  $\pi$  and the decomposition  $\rho$ , and we discuss the general proof strategy for this portion based on localization schemes.

4. *High-probability goodness of component measures.*

$$\Pr_{z \sim \rho}[z \text{ “good”}] \geq 1 - \varepsilon.$$

This part also requires analyzing the measure decomposition we design. Ideally, the measure decomposition presents us with “simpler” measures than  $\pi$  itself.

Once we have the above inequalities at hand, we get a  $(c\alpha, 2\varepsilon)$ -weak Poincaré inequality; see Lemma 4.4.10 for details.

**How to construct a good measure decomposition.** Henceforth, we restrict our attention to the case where  $\pi = \mu_H$ , the Gibbs distribution for a spherical mixed  $p$ -spin glass model. In the discussion below, we fix  $H$  as a typical Hamiltonian, and drop the phrase “with high probability” for events that occur with high probability over the randomness of  $H$ .

To construct our measure decomposition, we rely on Eldan’s stochastic localization [Eld13]. Our inspiration is the use of stochastic localization as a tool for measure decomposition for proving Poincaré inequalities in the work of Chen and Eldan [CE22]. Stochastic localization is a measure-valued random process  $(\mu_t)_{t \geq 0}$  described by:

$$d\mu_t(x) \propto \exp\left(\langle \mathbf{y}_t, x \rangle - \frac{t}{2} \|x\|^2\right) d\mu_H(x),$$

where  $\mathbf{y}_t = \sigma + B_t$  where  $\sigma \sim \mu_H$  and  $(B_t)_{t \geq 0}$  is a standard Brownian motion; see [AM22, Theorem 2] for a proof of why the above description of stochastic localization is equivalent to the more traditional definition via a stochastic differential equation that  $\mu_t$  obeys.

We run stochastic localization up to a stopping time  $\tau$ , defined as

$$\tau := \min\{t : 0 \leq t \leq T, \|\mathbf{Cov}(\mu_t)\| \geq K \text{ or } t = T\},$$

where  $T$  is chosen as a sufficiently large constant, independent of  $N$ . We impose the constraint on the covariance matrix as it is relevant to satisfying approximate conservation of variance: [CE22, Eq. (20)] proves that a measure decomposition based on stochastic localization run for time at most  $T$  satisfies approximate conservation of variance with parameter  $\alpha = \exp(-KT)$  if  $\|\text{Cov}(\mu_t)\|_{\text{op}}$  is bounded by  $K$  almost surely. Hence, by construction, we automatically ensure that our measure decomposition satisfies approximate conservation of variance.

For the measure decomposition to ultimately be useful, we also need to argue that the component measures satisfy a weak Poincaré inequality with high probability. Building on technical results in Huang, Montanari, and Pham [HMP24, Section 9.2], we show that the stochastic localization process run up to time  $T$  starting at  $\mu_H$  gives a distribution satisfying an  $(\Omega(1), \exp(-\Omega(n)))$ -weak Poincaré inequality with probability  $1 - \exp(-\Omega(n))$  over the randomness of the stochastic localization path; see Lemma 4.7.7 for details. Unfortunately, in the situation where the stochastic localization process stops before  $T$ , we do not have a simple way to show a weak Poincaré inequality, and for our analysis, treat  $z$  arising from early stopping as “bad”.

Thus, we have:  $\Pr_{z \sim \rho}[z \text{ “good”}] \geq 1 - \exp(-\Omega(n)) - \Pr[\tau < T]$ . To bound  $\Pr[\tau < T]$ , it is sufficient to prove a high-probability covariance norm bound on the entire stochastic localization path for  $0 \leq t \leq T$ . Most of the technical work in this paper is devoted to proving this covariance norm bound.

**Theorem 4.2.3** (Informal version of Lemma 4.7.6). *For a typical  $H$ , with probability  $1 - e^{-\Omega(n^{1/5})}$  over the randomness of the stochastic localization path, we have  $\|\text{Cov}(\mu_t)\|_{\text{op}} < K$ .*

The proof of the covariance bound spans Sections 4.7 and 4.8; we give a detailed technical overview of how it is proved in Section 4.7.

## 4.3 Preliminaries

### Notation

- We use  $S_N$  to denote the scaled  $(N - 1)$ -sphere,  $\sqrt{N} \cdot \mathbb{S}^{N-1}$ .
- We use  $\rho$  to denote the uniform measure over  $S_N$ .
- Given  $\sigma_1, \sigma_2 \in S_N$ , we use  $R(\cdot, \cdot)$  to denote the normalized inner product (i.e. the *overlap*)

$$R(\sigma_1, \sigma_2) := \frac{\langle \sigma_1, \sigma_2 \rangle}{N}.$$

- For an interval  $I \subseteq [-1, 1]$  and  $\mathbf{x} \in S_N$ , define  $\text{Band}(\mathbf{x}, I) := \{\sigma \in S_N : R(\sigma, \mathbf{x}) \in I\}$ .
- We use  $c$  to denote small constants whose values may change from line to line, and  $C$  to denote similarly fickle large constants.
- Let  $f : \Omega \rightarrow \mathbb{R}$  be any function. We define  $\text{osc}(f) := \sup f - \inf f$ .
- Let  $f : \Omega \rightarrow \mathbb{R}$  be a smooth function. If  $\Omega \subseteq \mathbb{R}^N$ , then  $\nabla f$  denotes its Euclidean gradient. If  $\Omega \subseteq S_N$ , then  $\nabla_{\text{sp}} f$  denotes the Riemannian gradient on  $S_N$ . When the correct notion of gradient is clear from context, by an abuse of notation we will suppress this distinction and simply write  $\nabla f$ .

## Measure decompositions

Our framework for proving weak functional inequalities relies on the notion of a measure decomposition.

**Definition 4.3.1** (Measure decomposition). Let  $\pi$  be a distribution on  $\mathbb{R}^N$ . Let  $\rho$  be a mixture distribution, also on  $\mathbb{R}^N$ , which indexes into a family of mixture components  $\{\pi_z\}_{z \in \mathbb{R}^N}$ . We say that  $(\rho, \pi_z)$  is a measure decomposition for  $\pi$  if

$$\pi = \mathbf{E}_{z \sim \rho} \pi_z.$$

One reason measure decompositions are useful is that they compose nicely with worst-case functional inequalities, as shown in the following lemma [BB19, AJK<sup>+</sup>22, CE22].

**Lemma 4.3.2** (Measure decomposition composition). *Let  $\pi$  be a distribution over  $\Omega \subseteq \mathbb{R}^N$ , and  $\pi = \mathbf{E}_{z \sim \rho} \pi_z$  a measure decomposition of  $\pi$  such that*

- *for all functions  $f$ ,  $\mathbf{E}_{z \sim \rho} \text{Var}_{\pi_z}[f] \geq C_{\text{Var}} \text{Var}_{\pi}[f]$ , and*
- *Every  $\pi_z$  satisfies a  $\rho_{\text{PI}}$ -Poincaré inequality with respect to Langevin diffusion.*

*Then,  $\pi$  satisfies a  $\rho_{\text{PI}} C_{\text{Var}}$ -Poincaré inequality.*

In Lemma 4.4.10, we will show an average-case relaxation of the above result, that  $\pi$  satisfies a weak Poincaré inequality if *most* measures in the decomposition satisfy weak Poincaré inequalities. Then, in Section 4.6, we construct explicit measure decompositions using the localization schemes framework introduced in [CE22]. This will show weak Poincaré inequalities for our measures of interest.

Besides proving functional inequalities, measure decompositions have also been directly used for sampling and inference (see, e.g., [MW24, LMR<sup>+</sup>24]).

## Langevin diffusion

In this paper, we study Langevin diffusion on  $\mathbb{R}^N$  and the scaled sphere  $S_N$ . These definitions can be directly generalized to the setting of Riemannian manifolds, but we do not comment further on this.

**Definition 4.3.3** (Langevin diffusion on  $\mathbb{R}^N$ ). Let  $\pi$  be a distribution on  $\mathbb{R}^N$  with density at  $x$  proportional to  $e^{-V(x)}$  for some function  $V$ . The *Langevin diffusion* process with stationary distribution  $\pi$  is the solution to the stochastic differential equation

$$dZ_t = -\nabla V(Z_t) dt + \sqrt{2} dB_t,$$

where  $(B_t)_{t \geq 0}$  is a standard Brownian motion.

**Definition 4.3.4** (Langevin diffusion on  $S_N$ ). Let  $\pi$  be a distribution on  $S_N$  with  $d\pi(x) \propto e^{-V(x)} d\rho(x)$ , where  $V : S_N \rightarrow \mathbb{R}$ . The Langevin diffusion process with stationary distribution  $\pi$  is the solution to the stochastic differential equation

$$dZ_t = -\nabla_{\text{sp}} V(Z_t) dt + \sqrt{2} dB_t,$$

where  $(B_t)_{t \geq 0}$  is a standard spherical Brownian motion. (For a textbook introduction to spherical Brownian motion, see [Hsu02].)

We will use the following standard reversibility fact for Langevin diffusion [Che23, Example 1.2.17].

**Fact 4.3.5** (Reversibility of Langevin diffusion). *The Langevin diffusion SDE with stationary distribution  $\pi$  is reversible with respect to  $\pi$ . In particular, the ergodicity of the process implies that  $\text{KL}(\text{Law}(Z_t) \parallel \pi) \xrightarrow{t \rightarrow \infty} 0$ .*

Furthermore, it is well-known that Langevin diffusion on  $\mathbb{R}^N$  with respect to a strongly log-concave stationary distribution converges rapidly.

**Definition 4.3.6.** Let  $\pi$  be a distribution over  $\mathbb{R}^N$  with density proportional to  $e^{-V}$ .  $\pi$  is said to be  $\alpha$ -strongly log-concave if  $V$  is  $\alpha$ -strongly convex, that is,  $\nabla^2 V \geq \alpha I$ .

We also use the following log-Sobolev mixing bound [Che23, Theorem 1.2.24].

**Fact 4.3.7** (Log-Sobolev mixing). *Let  $\pi$  be a distribution satisfying a log-Sobolev inequality with constant  $\rho_{\text{LS}}$ , in that for any differentiable function  $f : \mathbb{R}^N \rightarrow \mathbb{R}_{>0}$ ,*

$$\mathbf{E}_\pi \|\nabla \sqrt{f}\|^2 \geq \rho_{\text{LS}} \text{Ent}_\pi[f].$$

Then, if  $\pi_t$  is the distribution at time  $t$  of Langevin diffusion,

$$\text{KL}(\pi_t \parallel \pi) \leq \text{KL}(\pi_0 \parallel \pi) e^{-\rho_{\text{LS}} \cdot t}.$$

Furthermore,  $\alpha$ -strongly log-concave distributions  $\pi$  satisfy a log-Sobolev inequality with constant  $\alpha$ .

## 4.4 Weak functional inequalities

In this paper, we study continuous-time Markov chains.

**Definition 4.4.1** (Markov semigroup). Let  $(X_t)_{t \geq 0}$  denote a continuous-time Markov process on state space  $\Omega$ . Let  $(P_t)_{t \geq 0}$  be the associated Markov semigroup operator;  $P_t$  acts on functions  $f : \Omega \rightarrow \mathbb{R}$  via  $P_t f(x) = \mathbf{E}[f(X_t) | X_0 = x]$ . Throughout, we assume that the semigroup is reversible with respect to stationary distribution  $\pi$ . Furthermore, let  $L$  denote the infinitesimal generator of  $P_t$ , i.e.,  $P_t = e^{-tL}$ . For functions  $f, g : \Omega \rightarrow \mathbb{R}$ , we define the Dirichlet form as  $\mathcal{E}(f, g) = \mathbf{E}_\pi[f L g]$ .

See e.g. [Che23, Section 1.2] for a textbook treatment. Of particular interest to us are the two settings where the semigroup corresponds to a discrete-time Markov chain or the Langevin diffusion defined in Section 4.3. In these cases, the Dirichlet form satisfies the following explicit identities.

**Fact 4.4.2** (Dirichlet form from discrete-time Markov chain). *Let  $P$  be the transition matrix of a reversible discrete-time Markov chain with stationary distribution  $\pi$ . We can define an associated continuous-time semigroup operator  $(P_t)_{t \geq 0}$  by setting  $L = I - P$ . The Dirichlet form for the continuous-time dynamics satisfies*

$$\mathcal{E}(f, g) := \mathbf{E}_{x \sim \pi} \mathbf{E}_{y \sim P x} (f(x) - f(y))(g(x) - g(y)).$$

Here, for a probability distribution  $\mu$ , we say  $x \sim \mu$  to denote a sample  $x$  from  $\mu$ , and we use  $y \sim_P x$  for a single transition from  $x$  according to  $P$ .

**Fact 4.4.3** (Dirichlet form for Langevin diffusion). *We will need the following explicit identities for the Dirichlet form for Langevin diffusion.*

- (3) When  $(P_t)_{t \geq 0}$  corresponds to Langevin diffusion on  $\mathbb{R}^N$  with stationary distribution  $\pi$ , the Dirichlet form is  $\mathcal{E}(f, g) = \mathbf{E}_\pi[\langle \nabla f, \nabla g \rangle]$ .
- (3) When  $(P_t)_{t \geq 0}$  corresponds to Langevin diffusion on  $S_N$  with stationary distribution  $\pi$ , the Dirichlet form is  $\mathcal{E}(f, g) = \mathbf{E}_\pi[\langle \nabla_{\text{sp}} f, \nabla_{\text{sp}} g \rangle]$ .

**Definition 4.4.4.** We say  $\pi$  satisfies a *weak Poincaré inequality* if for some error functional  $\text{Error} : \mathbb{R}_{>0}^\Omega \rightarrow \mathbb{R}_{\geq 0}$  and  $\rho_{\text{PI}} > 0$ ,

$$\text{Var}_\pi[f] \leq \frac{1}{\rho_{\text{PI}}} \cdot \mathcal{E}(f, f) + \text{Error}(f).$$

Similarly, we say  $\pi$  satisfies a *weak modified log-Sobolev inequality* if for some error functional  $\text{Error} : \mathbb{R}_{>0}^\Omega \rightarrow \mathbb{R}_{\geq 0}$  and  $C_{\text{MLSI}} \geq 0$ ,

$$\text{Ent}_\pi[f] \leq \frac{1}{C_{\text{MLSI}}} \cdot \mathcal{E}(f, \log f) + \text{Error}(f).$$

**Theorem 4.4.5.** Consider the trajectory  $(v_t)_{t \geq 0}$  of a reversible continuous-time Markov chain with stationary distribution  $\pi$ , initialized at the distribution  $v_0$ , and suppose that  $\pi$  satisfies a weak MLSI with parameters  $\text{Error}$  and  $C_{\text{MLSI}}$ . Fix  $T > 0$ , and set  $\Lambda_T$  to be the distribution on  $[0, T]$  with density  $\Lambda_T(s) = \frac{C_{\text{MLSI}}}{e^{C_{\text{MLSI}}T} - 1} \cdot e^{C_{\text{MLSI}}s}$ . Then,

$$\text{KL}(v_T \parallel \pi) \leq e^{-C_{\text{MLSI}}T} \text{KL}(v_0 \parallel \pi) + \mathbf{E}_{s \sim \Lambda_T}[\text{Error}(\frac{dv_s}{d\pi})].$$

*Proof.* Let  $f_0 = \frac{dv_0}{d\pi}$ , and let  $f_t = P_t f_0 = \frac{dv_t}{d\pi}$  (this last equality holds due to reversibility). For ease of notation, set  $\text{Error}_t = \text{Error}(f_t)$  for  $t \geq 0$ . Recalling that  $\mathcal{E}(f_t, \log f_t) = -\frac{d}{dt} \text{KL}(v_t \parallel \pi)$ , the weak MLSI says that

$$-\mathcal{E}(f_t, \log f_t) + C_{\text{MLSI}} \cdot \text{KL}(v_t \parallel \pi) - C_{\text{MLSI}} \cdot \text{Error}_t \leq 0,$$

so

$$\frac{d}{dt} \left( e^{C_{\text{MLSI}}t} \cdot \text{KL}(v_t \parallel \pi) - C_{\text{MLSI}} \int_0^t e^{C_{\text{MLSI}}s} \text{Error}_s ds \right) \leq 0.$$

Therefore,

$$e^{C_{\text{MLSI}}T} \cdot \text{KL}(v_T \parallel \pi) - C_{\text{MLSI}} \int_0^T e^{C_{\text{MLSI}}s} \text{Error}_s ds \leq \text{KL}(v_0 \parallel \pi),$$

and

$$\text{KL}(v_T \parallel \pi) \leq e^{-C_{\text{MLSI}}T} \text{KL}(v_0 \parallel \pi) + C_{\text{MLSI}} \int_0^T e^{C_{\text{MLSI}}(s-T)} \text{Error}_s ds.$$

Noting that  $\Lambda_T(s) = \frac{C_{\text{MLSI}}}{e^{C_{\text{MLSI}}T} - 1} \cdot e^{C_{\text{MLSI}}s} \geq C_{\text{MLSI}} e^{C_{\text{MLSI}}(s-T)}$ , the above implies that

$$\text{KL}(v_T \parallel \pi) \leq e^{-C_{\text{MLSI}}T} \text{KL}(v_0 \parallel \pi) + \mathbf{E}_{s \sim \Lambda_T}[\text{Error}_s],$$

as desired. □



By essentially the same proof, we obtain the analogous result for weak Poincaré inequalities.

**Theorem 4.4.6.** *Consider the trajectory  $(v_t)_{t \geq 0}$  of a (continuous-time) Markov chain with stationary distribution  $\pi$ , initialized at the distribution  $v_0$ , and suppose that  $\pi$  satisfies a weak Poincaré inequality with parameters  $\text{Error}$  and  $\rho_{\text{PI}}$ . Fix  $T > 0$ , and set  $\Lambda_T$  to be the distribution on  $[0, T]$  with density  $\Lambda_T(s) = \frac{2\rho_{\text{PI}}}{e^{2\rho_{\text{PI}}T} - 1} \cdot e^{2\rho_{\text{PI}}s}$ . Then,*

$$\chi^2(v_T \parallel \pi) \leq e^{-2\rho_{\text{PI}}T} \chi^2(v_0 \parallel \pi) + \mathbf{E}_{s \sim \Lambda_T}[\text{Error}(\frac{dv_s}{ds})].$$

For the analysis of the annealed Langevin dynamics, we will also require the following definition. For  $f : \Omega \rightarrow \mathbb{R}$ , let  $\text{osc}(f) := \sup(f) - \inf(f)$ , and let  $\nabla f$  denote the Riemannian gradient.

**Definition 4.4.7** (Weak functional inequalities for Langevin). We say a distribution  $\pi$  on  $\Omega \subseteq \mathbb{R}^N$  or  $\Omega \subseteq S_N$  satisfies a  $(\rho_{\text{PI}}, \varepsilon)$ -weak Poincaré inequality if for all differentiable functions  $f$ ,

$$\text{Var}_{\pi}[f] \leq \frac{1}{\rho_{\text{PI}}} \cdot \mathcal{E}(f, f) + \varepsilon \cdot (\text{osc}(f)^2 + \sup_{x \in \Omega} \|\nabla f\|^2).$$

Similarly, we say  $\pi$  satisfies a  $(C_{\text{MLSI}}, \varepsilon)$ -weak modified log-Sobolev inequality if for all differentiable functions  $f$ ,

$$\text{Ent}_{\pi}[f] \leq \frac{1}{C_{\text{MLSI}}} \cdot \mathcal{E}(f, \log f) + \varepsilon \cdot (\text{osc}(\sqrt{f})^2 + \sup_{x \in \Omega} \|\nabla f\|^2).$$

**Remark 4.4.8.** As mentioned in the beginning of this section, by replacing the Riemannian gradient with the discrete gradient, an analogous theory can be developed for annealed Glauber dynamics; see Definition 4.9.1.

We shall typically use weak Poincaré inequalities with functions  $f$  that have expectation 1, where we bound  $\text{osc}(f) \leq 2\|f - 1\|_{\infty}$ .

## Properties of weak functional inequalities

In this section, we state some crucial properties of weak functional inequalities for Langevin diffusion on  $\mathbb{R}^N$  or  $S_N$ . With minor modifications, the same results hold for Glauber dynamics on finite state spaces; see Section 4.9 for formal details.

**Lemma 4.4.9.** *Let  $\pi$  be a distribution on  $\mathbb{R}^N$  or  $S_N$  satisfying a  $\rho_{\text{PI}}$ -Poincaré inequality for Langevin diffusion, and  $\pi'$  a distribution such that  $d_{\text{TV}}(\pi, \pi') \leq \delta$ . Then,  $\pi'$  satisfies a  $(\rho_{\text{PI}}, \delta \max(\rho_{\text{PI}}^{-1}, 1))$ -weak Poincaré inequality for Langevin diffusion.*

*Proof.* There exists a coupling  $C$  of  $(\pi, \pi')$  such that for  $(x, x') \sim C$ ,  $\Pr[x \neq x'] \leq \delta$ . Thus,

$$\begin{aligned} \mathcal{E}_{\pi'}(f, f) &= \mathbf{E}_{\pi'} \|\nabla f\|^2 \\ &\geq \mathbf{E}_{\pi} \|\nabla f\|^2 - \delta \sup \|\nabla f\|^2 \\ &\geq \rho_{\text{PI}} \mathbf{Var}_{\pi}[f] - \delta \sup \|\nabla f\|^2. \end{aligned}$$

Let  $I = [\inf f, \sup f]$ . Note that  $\mathbf{Var}_{\pi}[f] = \inf_{a \in I} \mathbf{E}_{\pi}[(f - a)^2]$ . For each  $a \in I$ ,

$$\mathbf{E}_{\pi}[(f - a)^2] \geq \mathbf{E}_{\pi'}[(f - a)^2] - \delta \cdot \text{osc}(f)^2,$$

and therefore

$$\mathbf{Var}_{\pi}[f] \geq \mathbf{Var}_{\pi'}[f] - \delta \cdot \text{osc}(f)^2. \quad (4.3)$$

Combining with the above shows

$$\mathcal{E}_{\pi'}(f, f) \geq \rho_{\text{PI}} \mathbf{Var}_{\pi'}[f] - \delta (\rho_{\text{PI}} \cdot \text{osc}(f)^2 + \sup \|\nabla f\|^2). \quad \square$$

As foreshadowed previously, measure decompositions compose well with weak functional inequalities. Indeed, the following lemma can be viewed as a relaxation of the setup to prove genuine functional inequalities (cf. Lemma 4.3.2).

**Lemma 4.4.10.** *Let  $\pi$  be a distribution over  $\mathbb{R}^N$  or  $S_N$ , and  $\pi = \mathbf{E}_{z \sim \rho} \pi_z$  a measure decomposition of  $\pi$  such that*

- *for all functions  $f$ ,  $\mathbf{E}_{z \sim \rho} \mathbf{Var}_{\pi_z}[f] \geq C_{\text{Var}} \mathbf{Var}_{\pi}[f]$ , and*
- *with probability  $1 - \eta$  over  $z \sim \rho$ ,  $\pi_z$  satisfies a  $(\rho_{\text{PI}}, \delta)$ -weak Poincaré inequality with respect to Langevin diffusion.*

*Then,  $\pi$  satisfies a  $(\rho_{\text{PI}} C_{\text{Var}}, \frac{\delta + \eta}{C_{\text{Var}}})$ -weak Poincaré inequality.*

*Proof.* Let us say that  $z$  is *good* if  $\pi_z$  satisfies a weak Poincaré inequality, and  $f$  be a function. Then,

$$\begin{aligned} \mathcal{E}_{\pi}(f, f) &= \mathbf{E}_{z \sim \rho} \mathcal{E}_{\pi_z}(f, f) \\ &\geq \mathbf{E}_{z \sim \rho} \mathcal{E}_{\pi_z}(f, f) \mathbf{1}_{z \text{ is good}} \end{aligned}$$

$$\begin{aligned}
&\geq \mathbf{E}_{z \sim \rho} \rho_{\text{PI}} \mathbf{Var}_{\pi_z}[f] \mathbf{1}_{z \text{ is good}} - \delta \rho_{\text{PI}} \cdot (\text{osc}(f)^2 + \sup \|\nabla f\|^2) \\
&= \mathbf{E}_{z \sim \rho} \rho_{\text{PI}} \mathbf{Var}_{\pi_z}[f] - \delta \rho_{\text{PI}} \cdot (\text{osc}(f)^2 + \sup \|\nabla f\|^2) - \mathbf{E}_{z \sim \rho} \rho_{\text{PI}} \mathbf{Var}_{\pi_z}[f] \mathbf{1}_{z \text{ is not good}} \\
&\geq \rho_{\text{PI}} \mathbf{E}_{z \sim \rho} \mathbf{Var}_{\pi_z}[f] - (\delta \rho_{\text{PI}} + \eta \rho_{\text{PI}}) \cdot (\text{osc}(f)^2 + \sup \|\nabla f\|^2) \\
&\geq C_{\text{Var}} \rho_{\text{PI}} \mathbf{Var}_{\pi}[f] - (\delta \rho_{\text{PI}} + \eta \rho_{\text{PI}}) \cdot (\text{osc}(f)^2 + \sup \|\nabla f\|^2).
\end{aligned}$$

The desired follows.  $\square$

## Weak Poincaré inequalities and annealed Markov chains

The notion of weak functional inequalities defined in Definition 4.4.7 can be naturally applied in the context of simulated annealing, which we now define.

**Definition 4.4.11** (Annealing scheme). Let  $H$  be a Hamiltonian over  $\Omega$ , and  $(\mu_\beta)_{\beta \geq 0}$  the class of distributions over  $\Omega$  with  $\mu_\beta(\sigma) \propto e^{\beta H(\sigma)}$ . For each  $\beta \leq \beta_0$ , let  $P = P_\beta$  be a (reversible and ergodic) Markov chain with stationary distribution  $\mu_\beta$ .

An (inverse) temperature schedule is any function  $\beta : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ . An annealing scheme  $\mathcal{A}$  is the time-inhomogeneous Markov chain such that at time  $t$ , one applies the Markov chain  $P_{\beta(t)}$ .

Of interest is the temperature schedule of the form  $t \mapsto \delta \cdot \lfloor \frac{t}{T} \rfloor$ , with the chain being run for time  $T \cdot \left( \frac{\beta_0}{\delta} + 1 \right)$ .

**Theorem 4.4.12.** Let  $T, \delta > 0$  such that  $k_0 := \frac{\beta_0}{\delta}$  is an integer. Suppose that for each  $\beta = k\delta$  for  $0 \leq k \leq k_0$ ,  $\mu_\beta$  satisfies a  $(\rho_{\text{PI}}, \varepsilon)$ -weak Poincaré inequality for  $P_\beta$ . Consider the annealing scheme given by schedule  $t \mapsto \delta \cdot \lfloor \frac{t}{T} \rfloor$ , run for total time  $T \cdot \left( \frac{\beta_0}{\delta} + 1 \right)$ . Let  $\nu$  be the output distribution of this annealing scheme. Then,

$$d_{\text{TV}}(\nu, \mu_{\beta_0}) \leq \frac{\beta_0}{\delta} \cdot \left[ (1 + \delta \sup \|\nabla H\|) e^{2\delta \|H\|_\infty} - 1 \right] \cdot O \left( e^{-2\rho_{\text{PI}} T} + \varepsilon \right)^{1/2}.$$

**Remark 4.4.13.** Setting  $\varepsilon = 0$  and  $\delta = \beta_0$  matches the guarantees of [CE22] (after applying Pinsker's inequality).

*Proof.* We shall prove the above using a simple inductive argument – our goal will be to show that initialized at  $\mu_\beta$ , the  $P_{\beta+\delta}$  Markov chain run for time  $T$  yields a distribution sufficiently close (in total variation distance) to  $\mu_{\beta+\delta}$ . The total variation distance between the distribution that the annealed Markov chain outputs and the true distribution  $\mu_{\beta_0}$  is then upper bounded by the sum of these total variation errors.

Let  $\nu^{(r,k)}$  be the distribution obtained by running the annealed Markov chain initialized with  $\mu_{r\delta}$  until inverse temperature  $k\delta$ . In particular,  $\nu^{(r,k)}$  corresponds to the result of running our annealed Markov chain for  $T(k-r)$  time, and  $\nu^{(k,k)} = \mu_{k\delta}$ . We are interested in bounding  $d_{\text{TV}}(\nu^{(0,k_0)}, \mu_{\beta_0})$ . We have

$$\begin{aligned}
d_{\text{TV}}(\nu^{(0,k_0)}, \mu_{\beta_0}) &= d_{\text{TV}}(\nu^{(k_0-1,k_0)}, \mu_{\beta_0}) + \sum_{1 \leq r \leq k_0-1} \left( d_{\text{TV}}(\nu^{(r-1,k_0)}, \mu_{\beta_0}) - d_{\text{TV}}(\nu^{(r,k_0)}, \mu_{\beta_0}) \right) \\
&\leq d_{\text{TV}}(\nu^{(k_0-1,k_0)}, \mu_{\beta_0}) + \sum_{1 \leq r \leq k_0-1} d_{\text{TV}}(\nu^{(r-1,k_0)}, \nu^{(r,k_0)}) \\
&\hspace{15em} \text{(Triangle inequality)} \\
&\leq d_{\text{TV}}(\nu^{(k_0-1,k_0)}, \mu_{\beta_0}) + \sum_{1 \leq r \leq k_0-1} d_{\text{TV}}(\nu^{(r-1,r)}, \nu^{(r,r)}) \quad \text{(Data processing)} \\
&= \sum_{1 \leq r \leq k_0} d_{\text{TV}}(\nu^{(r-1,r)}, \mu_{r\delta}).
\end{aligned}$$

We now turn to controlling the error functional  $\text{osc}(f)^2 + \sup \|\nabla f\|^2$ . Fix an arbitrary  $\beta$ , and set  $f$  to be the likelihood ratio  $\frac{d\mu_\beta}{d\mu_{\beta+\delta}}$ . Then,

$$\begin{aligned}
\|f - 1\|_\infty &\leq \left\| \frac{e^{-\delta H}}{\mathbf{E}_{\mu_{\beta+\delta}} e^{-\delta H}} - 1 \right\|_\infty \\
&\leq \left\| \frac{e^{-\delta H} - 1}{\mathbf{E}_{\mu_{\beta+\delta}} e^{-\delta H}} \right\|_\infty + \left| \frac{1}{\mathbf{E}_{\mu_{\beta+\delta}} e^{-\delta H}} - 1 \right| \\
&\leq \frac{e^{\delta \|H\|_\infty} - 1}{e^{-\delta \|H\|_\infty}} + \frac{e^{\delta \|H\|_\infty} - 1}{e^{-\delta \|H\|_\infty}} \leq 2 \cdot (e^{2\delta \|H\|_\infty} - 1).
\end{aligned}$$

Hence,  $\text{osc}(f) \leq 4 \cdot (e^{2\delta \|H\|_\infty} - 1)$ . Next, a simple computation yields

$$\begin{aligned}
\|\nabla f\| &= \frac{\delta e^{-\delta H}}{\mathbf{E}_{\mu_{\beta+\delta}} e^{-\delta H}} \|\nabla H\| \\
&\leq 2\delta \cdot e^{2\delta \|H\|_\infty} \|\nabla H\|,
\end{aligned}$$

so we have  $\sup \|\nabla f\| \leq 2\delta \cdot e^{2\delta \|H\|_\infty} \sup \|\nabla H\|$ .

Since each  $\mu_{r\delta}$  satisfies a  $(\rho_{\text{PI}}, \varepsilon)$ -weak Poincaré inequality, Theorem 4.4.6 with the above calculation implies that

$$d_{\text{TV}}(\nu^{(r-1,r)}, \mu_{r\delta})^2 \leq \chi^2(\nu^{(r-1,r)} \parallel \mu_{r\delta})$$

$$\begin{aligned}
&\leq e^{-2\rho_{\text{PI}}T} \cdot \chi^2(\mu_{(r-1)\delta} \parallel \mu_{r\delta}) + \varepsilon \cdot (16(e^{2\delta\|H\|_\infty} - 1)^2 + 4(\delta e^{2\delta\|H\|_\infty} \sup \|\nabla H\|)^2) \\
&\leq (16(e^{2\delta\|H\|_\infty} - 1)^2 + 4(\delta e^{2\delta\|H\|_\infty} \sup \|\nabla H\|)^2) \left( e^{-2\rho_{\text{PI}}T} + \varepsilon \right).
\end{aligned}$$

Plugging this back into the earlier sequence of equations completes the proof.  $\square$

**Remark 4.4.14.** While the proof above has been stated for the annealing scheme where at time  $t$  the Hamiltonian is of the form  $\sigma \mapsto \beta(t) \cdot H(\sigma)$ , the proof immediately extends to essentially any annealing scheme that changes the Hamiltonian “slowly”, in that if  $H_t$  is the Hamiltonian at time  $t$ ,  $\|H_{t+T} - H_t\|_\infty \leq \delta$  for all  $t$ . A concrete example of such a scheme that might work better than the vanilla annealing is that which at time  $t$  has as Hamiltonian  $\sigma \mapsto H(\beta(t) \cdot \sigma)$ .

## 4.5 Vignette: Sampling from mixture models with advice

We are interested in the following question.

Let  $\pi$  be a distribution over  $\mathbb{R}^N$  with density proportional to  $e^{-V}$ . Given oracle access to the gradient  $\nabla V$ , when is it possible to efficiently produce samples that are close (in total variation distance) to  $\pi$ ?

We begin with an overview of existing results towards the above question. Recall from Fact 4.3.7 that for distributions satisfying a Poincaré inequality, such as strongly log-concave distributions, Langevin diffusion enjoys rapid mixing. Beyond this setting, however, very little is known. [BCE<sup>+</sup>22, CWZZ24] prove certain “local mixing” guarantees for Langevin diffusion on non-log-concave distributions, but these do not immediately translate to sampling guarantees. The works [GLR18, LRG18, GTC24] use Langevin diffusion-based algorithms to sample from mixtures of log-concave distributions. Furthermore, the first of these papers proves that it is hard to sample from a mixture of two Gaussian distributions with distinct covariance matrices given access to just the gradient  $\nabla V$ .

In [KV23], the first theoretical guarantees are provided for a new model designed to circumvent this issue, where in addition to being given access to the gradient  $\nabla V$ , we are also given “advice” in the form of  $m$  samples from the distribution (also see [NHH<sup>+</sup>20] and [Hin10, GLZ<sup>+</sup>18, XLZW16] for related discussion). In particular, they show that when the stationary distribution is a mixture of constantly many strongly log-concave distributions, Langevin diffusion initialized at the empirical measure on the advice gets close to the stationary distribution. However, their dependence on the number of components  $K$  is doubly exponential. The main result in this section improves the doubly exponential dependence to a polynomial one for any mixture of distributions satisfying Poincaré inequalities. Similar results are obtained by Koehler, Lee, & Vuong [KLV24].

**Theorem 4.5.1.** *Let  $\varepsilon, \delta \in (0, 1)$ , and let  $\pi$  a mixture*

$$\pi = \sum_{i=1}^K p_i \pi_i$$

*of distributions  $(\pi_i)_{i=1}^K$ , where each  $\pi_i$  satisfies a Poincaré inequality with constant (at least)  $\rho_{\text{PI}}$ . Further assume that  $p_i \geq p_*$  for all  $i$ . Let  $\nu_0$  be a random distribution over  $\mathbb{R}^N$  such that  $\mathbf{E}\nu_0 = \pi$ , in that for any measurable subset  $A$  of  $\mathbb{R}^N$ ,  $\mathbf{E}\nu_0(A) = \pi(A)$ . Set*

$$m = \Omega\left(\frac{\log(1/\delta)}{p_* \varepsilon^2}\right).$$

*Let  $\nu_1, \dots, \nu_m$  be iid draws from  $\nu_0$ , and  $\nu$  the uniform mixture over the  $(\nu_i)_{i=1}^m$ . Further suppose that with probability at least  $1 - \delta$ ,  $\chi^2(\nu \parallel \pi) \leq M$ . Denoting by  $\mu_T$  the distribution attained by running Langevin diffusion for time  $T$  initialized  $\nu$ , it holds that*

$$\Pr[\chi^2(\mu_T \parallel \pi) \leq \varepsilon] \geq 1 - O(\delta),$$

*for  $T = \Omega\left(\frac{1}{\rho_{\text{PI}}} \log\left(\frac{M}{\varepsilon}\right)\right)$ , where the probability is over the draws of  $\nu_i$ .*

**Remark 4.5.2.** One should think of  $\nu_0$  as being the point mass distribution supported on a (random) sample drawn from  $\pi$ . Alternatively, one can think of  $\nu_0$  as being the distribution obtained by drawing a sample  $x_0$  according to  $\pi$ , then running Langevin diffusion for a short amount of time — doing this would make the  $\chi^2$ -divergence  $\chi^2(\nu_0 \parallel \pi)$  finite. We also remark that a version of this proof goes through if we have that each  $\pi_i$  satisfies a log-Sobolev inequality instead of a Poincaré inequality, working with KL divergences instead.

*Proof of Theorem 4.5.1.* The idea of the proof will be to show that up to some additive error depending on the samples,  $\pi$  does satisfy a Poincaré inequality with respect to the distributions along the path of Langevin diffusion initialized at the empirical distribution. This error corresponds to how imbalanced the samples are in terms of the mixture weights — a straightforward concentration argument using Bernstein's inequality then shows that this error is small, so the  $\chi^2$  divergence essentially decays exponentially fast, as if  $\pi$  satisfied a true Poincaré inequality.

Let  $f_t$  be the Radon-Nikodym derivative of  $\mu_t$  (obtained by running Langevin diffusion initialized at  $\nu$ ) with respect to  $\pi$ . By definition, we have

$$\chi^2(\mu_t \parallel \pi) = \mathbf{E}_\pi[f_t^2] - 1$$

$$\begin{aligned}
&= \sum_{i=1}^K p_i (\mathbf{E}_{\pi_i}[f_t^2] - 1) \\
&= \sum_{i=1}^K p_i \mathbf{Var}_{\pi_i}[f_t] + \sum_{i=1}^K p_i (\mathbf{E}_{\pi_i}[f_t]^2 - 1).
\end{aligned}$$

Because each  $\pi_i$  satisfies a Poincaré inequality, the first term is bounded as

$$\sum_{i=1}^K p_i \mathbf{Var}_{\pi_i}[f_t] \leq \frac{1}{\rho_{\text{PI}}} \sum_{i=1}^K p_i \mathbf{E}_{\pi_i} \|\nabla f_t\|^2 = \frac{1}{\rho_{\text{PI}}} \mathbf{E}_{\pi} \|\nabla f_t\|^2.$$

Consequently,

$$\chi^2(\mu_t \parallel \pi) \leq \frac{1}{\rho_{\text{PI}}} \cdot \mathbf{E}_{\pi} \|\nabla f_t\|^2 + \sum_{i=1}^K p_i (\mathbf{E}_{\pi_i}[f_t]^2 - 1). \quad (4.4)$$

Theorem 4.4.6 then yields that

$$\begin{aligned}
\chi^2(\mu_T \parallel \pi) &\leq \chi^2(\mu_0 \parallel \pi) \cdot e^{-2\rho_{\text{PI}}T} + \mathbf{E}_{s \sim \Lambda_T} \left[ \sum_{i=1}^K p_i (\mathbf{E}_{\pi_i}[f_s]^2 - 1) \right] \\
&\leq M e^{-2\rho_{\text{PI}}T} + \mathbf{E}_{s \sim \Lambda_T} \left[ \sum_{i=1}^K p_i (\mathbf{E}_{\pi_i}[f_s]^2 - 1) \right].
\end{aligned}$$

Above, we use that because the chi-squared divergence to  $\pi$  of each of the  $\nu_i$  is at most  $M$ , so is that of the mixture  $\mu_0 = \nu$ .

To conclude, we shall establish tail bounds on

$$\mathbf{E}_{s \sim \Lambda_T} \left[ \sum_{i=1}^K p_i (\mathbf{E}_{\pi_i}[f_s]^2 - 1) \right].$$

For  $1 \leq j \leq m$ , let  $f_s^{(j)}$  be the Radon-Nikodym derivative of  $\mu_s^{(j)}$  with respect to  $\pi$ , where  $\mu_s^{(j)}$  is the distribution obtained by running Langevin diffusion for time  $s$  initialized at  $\nu_j$ . It is not difficult to see that  $f_s = \frac{1}{m} \sum_{j=1}^m f_s^{(j)}$ .

First, for fixed  $s$  and  $j$ , we use the fact that the  $(\mathbf{E}_{\pi_i}[f_s^{(j)}])_i$  are independent mean 1 random variables, with Hoeffding's inequality, to get tail bounds for  $\mathbf{E}_{\pi_i}[f_s^{(j)}]^2 - 1$ . We may use this to bound a certain Orlicz norm of this random variable — this bound on the norm also transfers to  $\mathbf{E}_{s \sim \Lambda_T} [\sum_{i=1}^K p_i (\mathbf{E}_{\pi_i}[f_s]^2 - 1)]$  as it is a convex combination of random variables with bounded Orlicz norm. This immediately yields the desired tail bound.

Fix  $s$  and  $i$ . To start, we have the almost sure bounds

$$\frac{1}{p_*} = \frac{1}{p_*} \mathbf{E}_\pi[f_s^{(j)}] = \frac{1}{p_*} \sum_{r=1}^K p_r \mathbf{E}_{\pi_r}[f_s^{(j)}] \geq \mathbf{E}_{\pi_i}[f_s^{(j)}] \geq 0.$$

Note that because the expected  $v_j$  is equal to  $\pi$ ,  $\mathbf{E}_{v_j} \mathbf{E}_{\pi_i}[f_s^{(j)}] = 1$  for any  $j$ . Furthermore, because  $\mathbf{E}_{\pi_i}[f_s^{(j)}]$  is a mean 1 random variable which is bounded in  $[0, \frac{1}{p_*}]$ , its variance is at most  $\frac{1}{p_*}$  (see e.g. [BD00]). Bernstein's inequality implies that

$$\Pr \left[ |\mathbf{E}_{\pi_i} f_s - 1| > t \right] = \Pr \left[ \left| \frac{1}{m} \sum_{j=1}^m \mathbf{E}_{\pi_i} [f_s^{(j)}] - 1 \right| > t \right] \leq 2 \exp \left( -\frac{mp_*}{2} \cdot \frac{t^2}{1+t} \right).$$

Thus, for any  $t > 0$ ,

$$\begin{aligned} \Pr \left[ |\mathbf{E}_{\pi_i}[f_s]^2 - 1| > t \right] &\leq \Pr \left[ |\mathbf{E}_{\pi_i} f_s - 1| > \frac{t}{2(1+\sqrt{t})} \right] \\ &\leq 2 \exp \left( -\frac{mp_*}{8} \cdot \frac{\left( \frac{t}{1+\sqrt{t}} \right)^2}{1 + \frac{t}{1+\sqrt{t}}} \right). \end{aligned}$$

Now, consider the Orlicz norm  $\|\cdot\|_\psi$  associated to the above family of tail bounds. As mentioned earlier, standard machinery may be used to go from the above tail bounds to a bound on the norm  $\|\mathbf{E}_{\pi_i}[f_s]^2 - 1\|_\psi$ . Convexity of the norm yields the same bound on  $\|\mathbf{E}_{s \sim \Lambda_T} \sum_{i=1}^K p_i (\mathbf{E}_{\pi_i}[f_s]^2 - 1)\|_\psi$ . Translating this back to a tail bound, we get that

$$\Pr \left[ \left| \mathbf{E}_{s \sim \Lambda_T} \sum_{i=1}^K p_i (\mathbf{E}_{\pi_i}[f_s]^2 - 1) \right| > \frac{\varepsilon}{2} \right] \leq 2 \exp \left( -\frac{mp_* \varepsilon^2}{10} \right) \leq \delta.$$

Conditioning on the above event not happening, we get that

$$\chi^2(\mu_T \parallel \pi) \leq \chi^2(\mu_0 \parallel \pi) \cdot e^{-2\rho_{\text{PI}} T} + \frac{\varepsilon}{2} \leq \varepsilon,$$

as desired. □

## 4.6 Stopped localization schemes

### Localization schemes

We review some basic notions for the localization schemes framework introduced in [CE22].



**Definition 4.6.1** (Linear-tilt localization scheme). Let  $\mu = \mu_0$  be a probability measure,  $(\mu_t)_{t \in \mathbb{Z}_{\geq 0}}$  be a localization process. A linear-tilt localization scheme is one where  $\mu_t$  is defined by

$$\mu_{t+1}(x) = \mu_t(x) (1 + \langle x - \mathbf{m}(\mu_t), Z_t \rangle)$$

where  $Z_t$  is a random variable with  $\mathbf{E}[Z_t | \mu_t] = 0$  and  $\mathbf{m}(\mu_t)$  denotes the mean of  $\mu_t$ .

For our main application to  $p$ -spin models, we will focus on a continuous-time version of linear-tilt localization known as stochastic localization [Eld13].

**Definition 4.6.2** (Stochastic localization). Let  $\mu$  be a probability measure on  $\Omega \subseteq \mathbb{R}^N$ ,  $(B_t)_{t \geq 0}$  be a standard Brownian motion on  $\mathbb{R}^N$ . The stochastic localization process with driving matrix  $(C_t)_{t \geq 0}$  is a localization process  $(\mu_t)_{t \geq 0}$  with  $\mu_0 = \mu$  and

$$\mu_t(x) \propto \mu_0(x) \exp\left(-\frac{1}{2} \langle x, \Sigma_t x \rangle + \langle y_t, x \rangle\right),$$

where  $\Sigma_t = \int_0^t C_s^2 ds$  and  $y_t = \int_0^t C_s^2 \mathbf{m}(\mu_s) ds + C_s dB_s$ .

A crucial property of these localization schemes is that establishing (approximate) conservation of variance reduces to bounding the covariance matrices of the intermediate distributions  $\mu_t$ . The following conservation of variance estimate is from [CE22, Claim 22].

**Lemma 4.6.3** (Conservation of variance for linear-tilt localization). Let  $(\mu_t)_{t \in \mathbb{Z}_{\geq 0}}$  be a linear-tilt localization process. Suppose that for all  $t \leq T$  we have

$$\|\mathbf{Cov}(Z_t | \mu_t)^{1/2} \cdot \mathbf{Cov}(\mu_t) \cdot \mathbf{Cov}(Z_t | \mu_t)^{1/2}\|_{\text{op}} \leq K_t,$$

where  $K_t \in [0, 1]$ . Then for any function  $\varphi$ ,

$$\frac{\mathbf{EVar}_{\mu_T}[\varphi]}{\mathbf{Var}_{\mu}[\varphi]} \geq \prod_{t=0}^{T-1} (1 - K_t).$$

**Lemma 4.6.4** (Conservation of variance for stochastic localization). Let  $(\mu_t)_{t \geq 0}$  be a stochastic localization process with driving matrix  $(C_t)_{t \geq 0}$ . Suppose that for all  $t \leq T$  we have

$$\|C_t^{1/2} \cdot \mathbf{Cov}(\mu_t) \cdot C_t^{1/2}\|_{\text{op}} \leq K_t$$

where  $K_t \in [0, 1]$ . Then for any function  $\varphi$ ,

$$\frac{\mathbf{EVar}_{\mu_T}[\varphi]}{\mathbf{Var}_{\mu}[\varphi]} \geq e^{-\int_0^T K_t dt}.$$

## Proving weak Poincaré inequalities using stopped localization schemes

To apply Theorem 4.4.12, we required weak Poincaré inequalities for the measures of interest. To show these, we next introduce a generic tool to prove these using Lemma 4.4.10, building on the localization schemes framework introduced in Section 4.6. Let  $\mu$  be a distribution. Using a localization scheme, we would like to design a measure decomposition  $\mu = \mathbf{E}_{z \sim \rho} \mu_z$  such that

- for all functions  $f$ ,  $\mathbf{Var}_\pi[f] \leq C_{\text{Var}} \mathbf{E}_{z \sim \rho} \mathbf{Var}_{\pi_z}[f]$ , and
- with probability  $1 - \eta$  over  $z \sim \rho$ ,  $\pi_z$  satisfies a  $(\rho_{\text{PI}}, \delta)$ -weak Poincaré inequality.

One way to ensure the first condition — approximate conservation of variance — is to simply stop the localization scheme whenever it fails to hold. Indeed, the following lemma immediately follows from Lemma 4.6.3.

**Lemma 4.6.5.** *Let  $\mu = \mu_0$  be a measure, and let  $(\mu_t)_{t \in \mathbb{Z}_{\geq 0}}$  be a linear-tilt localization process defined by*

$$\mu_{t+1}(x) = \mu_t(x) (1 + \langle x - \mathbf{m}(\mu_t), Z_t \rangle)$$

*for some random variable  $Z_t$  with  $\mathbf{E}[Z_t | \mu_t] = 0$ . Let  $T > 0$  be an arbitrary stopping time and  $0 \leq K_t < 1$  for each  $t \geq 0$ , and consider the stopping time*

$$\tau = T \wedge \inf_{t \geq 0} \left\{ \|\mathbf{Cov}(Z_t | \mu_t)^{1/2} \cdot \mathbf{Cov}(\mu_t) \cdot \mathbf{Cov}(Z_t | \mu_t)^{1/2}\|_{\text{op}} \geq K_t \right\}.$$

*Then, for any function  $\varphi$ ,*

$$\frac{\mathbf{EVar}_{\mu_\tau}[\varphi]}{\mathbf{Var}_\mu[\varphi]} \geq \prod_{t \geq 0} (1 - K_t).$$

Similarly, we have the following lemma for stochastic localization, which follows from Lemma 4.6.4.

**Lemma 4.6.6.** *Let  $\mu = \mu_0$  be a measure, and  $(\mu_t)_{t \geq 0}$  be a stochastic localization process with driving matrix  $(C_t)_{t \geq 0}$ . Let  $T, K > 0$  be constant parameters, and consider the stopping time*

$$\tau = T \wedge \inf_{t \geq 0} \left\{ \|C_t^{1/2} \cdot \mathbf{Cov}(\mu_t) \cdot C_t^{1/2}\|_{\text{op}} \geq K \right\}.$$

*Then,*

$$\frac{\mathbf{EVar}_{\mu_\tau}[\varphi]}{\mathbf{Var}_\mu[\varphi]} \geq e^{-TK}.$$

**Remark 4.6.7.** The localization process in the above lemmas can depend on  $\varphi$ , and need not be a linear-tilt localization. The more general requirement is that

$$\frac{\mathbf{E} [\mathbf{Var}_{\mu_{t+1}}[\varphi] | \mu_t]}{\mathbf{Var}_{\mu_t}[\varphi]} \geq K_t \quad \text{or} \quad \frac{1}{\mathbf{Var}_{\mu_t}[\varphi]} \cdot \frac{d}{ds} \mathbf{E} [\mathbf{Var}_{\mu_s}[\varphi] | \mu_t] \Big|_{s=t} \geq K.$$

This can always be achieved by stopping the localization process whenever these conditions fail to hold.

With these elements in hand, we now show how to prove a weak Poincaré inequality using stopped localization schemes.

**Lemma 4.6.8.** *Let  $\mu = \mu_0$  be a measure, and  $(\mu_t)_{t \geq 0}$  be a stochastic localization process with driving matrix  $(C_t)_{t \geq 0}$ . Let  $T, K > 0$  be constant parameters. Suppose that with probability  $1 - \eta_1$ , it holds that  $\|C_t^{1/2} \cdot \mathbf{Cov}(\mu_t) \cdot C_t^{1/2}\|_{\text{op}} < K$  for all  $t \in [0, T]$ . Further suppose that with probability  $1 - \eta_2$ ,  $\mu_T$  satisfies a  $(\rho_{\text{PI}}, \delta)$ -weak Poincaré inequality. Then,  $\mu$  satisfies a  $(\rho_{\text{PI}} e^{-TK}, e^{TK}(\delta + \eta_1 + \eta_2))$ -weak Poincaré inequality.*

*Proof.* As in Lemma 4.6.6, define the stopping time

$$\tau = T \wedge \inf_{t \geq 0} \left\{ \|C_t^{1/2} \cdot \mathbf{Cov}(\mu_t) \cdot C_t^{1/2}\|_{\text{op}} \geq K \right\}.$$

Consider the measure decomposition  $\mu = \mathbf{E} \mu_\tau$ . By Lemma 4.6.6, this decomposition is variance-conserving with parameter  $e^{-TK}$ . By the hypothesis of the lemma,  $\tau = T$  with probability  $1 - \eta_1$ , and  $\mu_T$  satisfies a weak Poincaré inequality with probability  $1 - \eta_2$ . Consequently,  $\mu_\tau$  satisfies a weak Poincaré inequality with probability at least  $1 - \eta_1 - \eta_2$ . Lemma 4.4.10 completes the proof.  $\square$

**Remark 4.6.9.** An analogous lemma to the above holds if  $(\mu_t)_{t \in \mathbb{Z}_{\geq 0}}$  is any linear-tilt localization process.

While it will not be used in this paper, we note that a similar method proves a weak Poincaré inequality for a natural Markov chain associated to a localization scheme. This includes for example the restricted Gaussian dynamics; see [CE22] for several other examples.

**Lemma 4.6.10.** *Let  $\mu = \mu_0$  be a measure, and  $(\mu_t)_{t \geq 0}$  be a stochastic localization process with driving matrix  $(C_t)_{t \geq 0}$ . Let  $T, K > 0$  be constant parameters. Consider the Markov chain  $P$  given by  $P_{x \rightarrow y} = \mathbf{E} \left[ \frac{\mu_T(x) \mu_T(y)}{\mu_0(x)} \right]$ . Define the stopping time*

$$\tau = T \wedge \inf_{t \geq 0} \left\{ \|C_t^{1/2} \cdot \mathbf{Cov}(\mu_t) \cdot C_t^{1/2}\|_{\text{op}} \geq K \right\}.$$

*If  $\tau = T$  with probability at least  $1 - \delta$ , then  $P$  satisfies a  $(e^{-TK}, \delta e^{TK})$ -weak Poincaré inequality.*

*Proof.* For the Markov chain  $P$ , the Dirichlet form is given by  $\mathcal{E}_P(f, f) = \mathbf{E}\mathbf{Var}_{\mu_T}[f]$  (see, e.g., [CE22, Proposition 19]). We then have the chain of inequalities

$$\begin{aligned} \mathbf{E}\mathbf{Var}_{\mu_T}[f] &\geq \mathbf{E}\mathbf{Var}_{\mu_\tau}[f] - \delta_{\text{osc}}(f)^2 \\ &\geq e^{-TK} \mathbf{Var}_{\mu_0}[f] - \delta_{\text{osc}}(f)^2. \end{aligned}$$

The first inequality here is immediate since

$$\mathbf{E}\mathbf{Var}_{\mu_\tau}[f] - \mathbf{E}\mathbf{Var}_{\mu_T}[f] = \mathbf{E}\mathbf{Var}_{\mu_\tau}[f] \mathbf{1}_{\tau \neq T} \leq \text{osc}(f)^2 \Pr[\tau \neq T].$$

The second inequality follows from Lemma 4.6.4.  $\square$

**Remark 4.6.11.** As in Remark 4.6.7, the above lemma can be generalized to localization schemes other than stochastic localization.

## 4.7 Sampling from spherical $p$ -spin models

In this section, we prove that simulated annealing samples from spherical spin glass models for models satisfying (SL). Recall that  $S_N = \sqrt{N} \cdot \mathbb{S}^{N-1}$ . For  $\gamma_2, \gamma_3, \dots, \gamma_{p^*} \geq 0$ , the mixed  $p$ -spin Hamiltonian  $H_N : S_N \rightarrow \mathbb{R}$  is defined by

$$H_N(\sigma) := \sum_{p \geq 2} \frac{\gamma_p}{N^{(p-1)/2}} \sum_{i_1, \dots, i_p=1}^N g_{i_1, \dots, i_p} \sigma_{i_1} \cdots \sigma_{i_p}, \quad (4.5)$$

for i.i.d. samples  $g_{i_1, \dots, i_p}$  from  $\mathcal{N}(0, 1)$ . This is the gaussian process on  $\mathbb{R}^N$  with covariance

$$\mathbf{E}H_N(\sigma^1) \cdot H_N(\sigma^2) = N \cdot \xi(R(\sigma^1, \sigma^2)),$$

where we recall the mixture function  $\xi$  is defined by  $\xi(s) = \sum_{p=2}^{p^*} \gamma_p^2 s^p$ . The algorithm we will study is the following simple annealing scheme for Langevin diffusion.

**Definition 4.7.1** (Annealed Langevin diffusion). Let  $\delta_N, T_N > 0$  be parameters possibly depending on  $N$ . For any  $\beta \geq 0$ , let  $\mu_\beta := \mu_{\beta H_N}$ , where  $H_N$  is the  $p$ -spin Hamiltonian. Annealed Langevin diffusion is the annealing scheme  $\mathcal{A}$  where  $\beta(t) = \delta_N \lfloor t/T_N \rfloor$  and  $P_\beta$  is the Langevin semigroup operator for Gibbs distribution  $\mu_\beta$ . In words,  $\mathcal{A}$  keeps  $\beta$  constant for time  $T_N$  and then increments  $\beta$  by  $\delta_N$ .

**Theorem 4.7.2.** Let  $H_N$  be a mixed  $p$ -spin Hamiltonian whose mixture function  $\xi$  satisfies (SL), which we recall below:

$$\xi''(q) < \frac{1}{(1-q)^2} \text{ for all } q \in [0, 1).$$

Let  $\mu$  be the associated Gibbs measure over the scaled sphere  $S_N$ , with

$$d\mu(\sigma) \propto \exp(H_N(\sigma)) d\rho(\sigma).$$

With probability  $1 - e^{-cN^{1/5}}$  over the randomness of  $H_N$ , the following holds. For some parameters  $\delta_N = O(N^{-4/5})$ ,  $T_N = \Omega(N^{1/5})$ , the output measure  $\nu$  of the annealed Langevin diffusion scheme with these parameters satisfies

$$d_{\text{TV}}(\nu, \mu) \leq e^{-cN^{1/5}}.$$

**Remark 4.7.3.** We expect that the error  $e^{-cN^{1/5}}$  can be improved to  $e^{-cN}$ , matching the fact that a  $e^{-O(N)}$  fraction of the Gibbs measure is typically trapped in metastable states between the uniqueness and shattering thresholds [BJ24]. However, we will not pursue this improvement in this paper.

**Remark 4.7.4.** The condition (SL) is a fundamental barrier for stochastic localization, both as an algorithm and a proof technique. As was essentially shown in [HMP24, Section 10], for models satisfying (Strict RS) but not (SL), the means  $\mathbf{m}(\mu_t)$  along the localization process do not move stably, in the sense that there exist time intervals of width  $o(1)$  in which  $\mathbf{m}(\mu_t)$  moves by  $\Omega(N^{1/2})$ . (The condition (Strict RS) is an artifact of the proof, and it is expected that the mean continues to move non-stably beyond the regime (Strict RS)). In the setting of [HMP24], this implies that their algorithmic simulation of the localization process fails, because approximate message passing will not estimate the mean at some times. In our setting, this implies that the covariance  $\text{Cov}(\mu_t)$ , which arises as the derivative of  $\mathbf{m}(\mu_t)$ , is genuinely not bounded in operator norm at some times, and thus the main input to our framework does not hold.

**Remark 4.7.5.** While the result above is stated for the continuous time Langevin *diffusion*, the results therein can be adapted to the discretized Langevin Monte Carlo algorithm using standard tools, à la [Che23, Part II], to obtain a polynomial time sampling algorithm.

To prove the above, we shall use Theorem 4.4.12 in conjunction with Lemma 4.6.8. For the remainder of this section, let  $(\gamma_p)_{p \geq 2}$  be a sequence of weights such that the associated mixture function  $\xi$  satisfies the condition (SL).

**Notation** (Measure decomposition for  $p$ -spin models). Let  $\mu_{H_N} = \mu_0$ . For a large constant time  $T$ , let  $(\mu_t)_{0 \leq t \leq T}$  be the stochastic localization process with driving matrix  $\text{Id}$  (see Definition 4.6.2).

**Lemma 4.7.6** (Covariance bound on stochastic localization path). *There exist constants  $c, K$ , depending only on  $\xi$ , such that for any constant  $T > 0$  the following holds with probability at least  $1 - e^{-cN^{1/5}}$  over the randomness of  $H_N$ . If  $(\mu_t)_{0 \leq t \leq T}$  is the (random) trajectory of stochastic localization initialized at  $\mu_0 = \mu_{H_N}$ , with probability  $1 - e^{-cN^{1/5}}$ ,  $\|\mathbf{Cov}(\mu_t)\|_{\text{op}} < K$  for all  $0 \leq t \leq T$ . In other words,*

$$\Pr_{H_N} \left[ \Pr_{(\mu_t)|H_N} \left[ \|\mathbf{Cov}(\mu_t)\|_{\text{op}} < K \text{ for all } 0 \leq t \leq T \right] \geq 1 - e^{-cN^{1/5}} \right] \geq 1 - e^{-cN^{1/5}}.$$

**Lemma 4.7.7** (Weak Poincaré inequality for endpoint distributions). *There exists a constant  $T$  depending only on  $\xi$  such that the following holds with probability at least  $1 - e^{-cN}$  over the randomness of  $H_N$ . With probability at least  $1 - e^{-cN}$ , the (random) measure  $\mu_T$  satisfies a  $(c, e^{-cN})$ -weak Poincaré inequality. In other words,*

$$\Pr_{H_N} \left[ \Pr_{\mu_T|H_N} \left[ \mu_T \text{ satisfies a } (c, e^{-cN})\text{-weak Poincaré inequality} \right] \geq 1 - e^{-cN} \right] \geq 1 - e^{-cN}.$$

Let us first see how these two lemmas imply the main theorem.

*Proof of Theorem 4.7.2.* Fix some  $0 \leq \beta \leq 1$ . Note that the Hamiltonian  $\beta H_N$  has mixture function  $\xi_\beta(s) = \xi(\beta^2 s)$ , and if  $\xi$  satisfies (SL) then  $\xi_\beta$  does as well. Plugging in Lemmas 4.7.6 and 4.7.7 into Lemma 4.6.8 implies that with probability at least  $1 - e^{-cN^{1/5}}$ ,  $\mu_{\beta H_N}$  satisfies a  $(c, e^{-cN^{1/5}})$ -weak Poincaré inequality. A union bound implies that with probability  $1 - e^{-cN^{1/5}}$ , for all  $\beta$  encountered along the annealing schedule,  $\mu_{\beta H_N}$  satisfies a  $(c, e^{-cN^{1/5}})$ -weak Poincaré inequality.

By [HS22, Proposition 2.3], with probability  $1 - e^{-cN}$ ,  $\|\nabla H_N\|_\infty = O(\sqrt{N})$ . The same argument implies that with probability  $1 - e^{-cN}$ ,  $\|H_N\|_\infty = O(N)$ . With probability  $1 - e^{-cN^{1/5}}$ , all three of these events occur, and Theorem 4.4.12 completes the proof.  $\square$

We conclude this subsection by proving Lemma 4.7.7.

*Proof of Lemma 4.7.7.* Let  $d\mu_t(\sigma) \propto e^{H_{N,T}(\sigma)} d\sigma$  for  $H_{N,T}(\sigma) = H_N(\sigma) + \langle \mathbf{y}, \sigma \rangle$ . Let

$$S_N(\mathbf{y}) = \{\sigma \in S_N : R(\mathbf{y}, \sigma) > 0\}.$$

Let  $\mathbf{U} \in \mathbb{R}^{N \times (N-1)}$  be a matrix whose columns are an orthonormal basis of the orthogonal complement of  $\mathbf{y}$ . Let  $\hat{\mathbf{y}} = \sqrt{N} \mathbf{y} / \|\mathbf{y}\|$  be  $\mathbf{y}$  (which is a.s. nonzero) scaled to length  $\sqrt{N}$ , and define the map  $\sigma_{\mathbf{y}}(\rho) : \mathbb{R}^{N-1} \rightarrow S_N(\mathbf{y})$  by

$$\sigma_{\mathbf{y}}(\rho) = \frac{\hat{\mathbf{y}} + \mathbf{U} \rho}{\sqrt{1 + R(\rho, \rho)}}.$$

This is the inverse of the map that first stereographically projects  $S_N(\mathbf{y})$  from the origin to  $\widehat{\mathbf{y}} + \mathbf{U}\mathbb{R}^{N-1}$ , the plane tangent to  $S_N$  at  $\widehat{\mathbf{y}}$ , and then maps the resulting point to coordinates given by  $\mathbf{U}$ . Let  $\varepsilon_0 = 0.1$ , and  $A = \{\rho \in \mathbb{R}^{N-1} : \|\rho\|^2 \leq \varepsilon_0 N\}$ , and note that

$$\sigma_{\mathbf{y}}(A) = \{\sigma \in S_N : R(\sigma, \widehat{\mathbf{y}}) \geq (1 + \varepsilon_0)^{-1/2}\}$$

is a spherical cap around  $\widehat{\mathbf{y}}$ . Let  $A' := \sigma_{\mathbf{y}}(A)$ . By arguments in [HMP24, Subsection 9.2], there exists a measure  $\nu$  (denoted  $\tilde{\nu}_{H_N, \mathbf{y}'}^{\text{proj}}$ , see Eq. 2.10 therein) such that the following holds with probability  $1 - e^{-cN}$ .

- The push-forward of  $\nu|_A$  through  $\sigma_{\mathbf{y}}$  coincides with  $(\mu_T)_{|A'}$ . (Lemma 9.5 therein.)
- $\mu_T(A') = 1 - e^{-cN}$ . (Lemma 9.6 therein states this with  $1 - o_N(1)$  in place of  $1 - e^{-cN}$ , but the proof implies bound  $1 - e^{-cN}$ , as this is the bound given by Proposition 5.12 used therein.)
- $\nu(A) = 1 - e^{-cN}$ . (Corollary 9.7 therein, modulo the same issue of  $1 - o_N(1)$  versus  $1 - e^{-cN}$ , which is addressed the same way.)
- $\nu$  is  $\Omega(1)$ -strongly log-concave. (Proposition 9.8 therein.)

By the well-known Bakry-Émery condition (see, e.g., [Che23, Section 1.2.3]), on this event  $\nu$  satisfies a  $\rho_{\text{PI}}$ -Poincaré inequality for some  $\rho_{\text{PI}} = \Omega(1)$ . We will transfer this inequality to a  $(\rho_{\text{PI}}, e^{-cN})$ -weak Poincaré inequality for  $\mu_T$ . Consider a smooth test function  $f : S_N \rightarrow \mathbb{R}$  and let  $\tilde{f} : \mathbb{R}^{N-1} \rightarrow \mathbb{R}$  be defined by  $\tilde{f} = f \circ \sigma_{\mathbf{y}}$ . Since  $d_{\text{TV}}(\mu_T, (\mu_T)_{|A'}) = e^{-cN}$  and  $d_{\text{TV}}(\nu, \nu|_A) = e^{-cN}$ , and  $\text{osc}(\tilde{f}) \leq \text{osc}(f)$ , arguing as in (4.3) shows

$$\begin{aligned} \text{Var}_{\mu_T}(f) &\leq \text{Var}_{(\mu_T)_{|A'}}(f) + e^{-cN} \text{osc}(f)^2 \\ &= \text{Var}_{\nu|_A}(\tilde{f}) + e^{-cN} \text{osc}(f)^2 \\ &\leq \text{Var}_{\nu}(\tilde{f}) + 2e^{-cN} \text{osc}(f)^2. \end{aligned}$$

By the Poincaré inequality for  $\nu$  and the definition of the Dirichlet form for Langevin diffusion,

$$\text{Var}_{\nu}(\tilde{f}) \leq \frac{1}{\rho_{\text{PI}}} \cdot \mathcal{E}_{\nu}(\tilde{f}, \tilde{f}) = \frac{1}{\rho_{\text{PI}}} \cdot \mathbf{E}_{\nu}[\|\nabla \tilde{f}\|^2]$$

By [HMP24, Proof of Lemma 9.5], the map  $\sigma_{\mathbf{y}}$  has Jacobian  $J_{\sigma_{\mathbf{y}}}$  satisfying  $\|J_{\sigma_{\mathbf{y}}}\|_{\text{op}} \leq 1$ , and thus for all  $\rho \in \mathbb{R}^{N-1}$ ,

$$\|\nabla \tilde{f}(\rho)\| = \|\nabla(f \circ \sigma_{\mathbf{y}})(\rho)\| \leq \|\nabla f(\sigma_{\mathbf{y}}(\rho))\|.$$

It follows that

$$\begin{aligned}
\mathbf{E}_\nu[\|\nabla \tilde{f}\|^2] &\leq \mathbf{E}_{\nu|_A}[\|\nabla \tilde{f}\|^2] + e^{-cN} \sup \|\nabla \tilde{f}\|^2 \\
&\leq \mathbf{E}_{(\mu_T)|_{A'}}[\|\nabla f\|^2] + e^{-cN} \sup \|\nabla f\|^2 \\
&\leq \mathbf{E}_{\mu_T}[\|\nabla f\|^2] + 2e^{-cN} \sup \|\nabla f\|^2
\end{aligned}$$

Combining the above shows

$$\mathbf{Var}_{\mu_T}(f) \leq \frac{1}{\rho_{\text{PI}}} \mathcal{E}_{\mu_T}(f, f) + 2e^{-cN} \left( \text{osc}(f)^2 + \frac{1}{\rho_{\text{PI}}} \sup \|\nabla f\|^2 \right).$$

The result follows by adjusting  $c$ . □

## Technical overview for covariance bounds

The proof of the main theorem has boiled down to Lemma 4.7.6 — we now give a high-level overview of our proof strategy for this. We wish to show that with very high probability ( $1 - e^{-\Omega(N^{1/5})}$ ), the covariance is bounded along the entire path  $(\mu_t)_{0 \leq t \leq T}$  of stochastic localization. By performing a union bound over time and a standard perturbation argument, it suffices to show that for a fixed time  $t \in [0, T]$ ,  $\mu_t$  has bounded covariance with very high probability.

To do this, we recall an alternate view of stochastic localization [AM22]. The measure at time  $t$  of stochastic localization (with the identity driving matrix) is given as follows. First, draw  $\sigma \sim \mu_{H_N}$ , and independently  $\mathbf{g} \sim \mathcal{N}(0, I_N)$ . Then,  $\mu_t$  has the same law as  $\mu_{H_N, t\sigma + \sqrt{t}\mathbf{g}}$  in that

$$\mu_{H_N, t\sigma + \sqrt{t}\mathbf{g}}(\tilde{\sigma}) \propto \exp \left( H_N(\tilde{\sigma}) + \langle t\sigma + \sqrt{t}\mathbf{g}, \tilde{\sigma} \rangle \right).$$

As written, the covariance of this distribution is difficult to analyze — the sample  $\sigma$  has very complicated correlations with the disorder of the Hamiltonian  $H_N$ , making it intractable.

**The planting trick.** To deal with this, we will use the *planting trick* introduced by Achlioptas and Coja-Oghlan [ACO08]. The application of this method in the context of stochastic localization is by now standard [EAMS22, EAMS23, HMP24], and we review the main ideas for the reader's convenience.

**Definition 4.7.8** (Planted  $p$ -spin model). The planted measure  $\mu_{\text{pl}}$  is a joint law over a Hamiltonian  $H_N$  and a *spike*  $x \in S_N$  given by

$$d\mu_{\text{pl}}(H_N, x) \propto \exp(H_N(x)) \cdot d\rho(x) \cdot d\mu_{\text{null}}(H_N),$$

where  $\rho$  is the uniform measure over  $S_N$  and  $\mu_{\text{null}}$  is the law over  $p$ -spin Hamiltonians with mixture function  $\xi$ . We frequently abuse notation to let  $\mu_{\text{pl}}(H_N)$  denote the marginal of  $\mu_{\text{pl}}$  on  $H_N$ .



To provide further intuition for the above definition, consider the following alternate sampling interpretation of the planted model, which describes the distribution of  $\mathbf{x}$  conditioned on  $H_N$ .

**Fact 4.7.9.** *Consider the following inference problem. We start by sampling the spike  $\mathbf{x} \sim S_N$ , sample  $G^{(p)}$  as a rank- $p$  tensor with iid  $\mathcal{N}(0, 1)$  entries for  $p \geq 2$ , and for each  $p$  let  $M^{(p)} = -G^{(p)} + \frac{\gamma_p}{N^{(p-1)/2}} \mathbf{x}^{\otimes p}$ .*

*Then, the posterior on  $\mathbf{x}$  after observing the tensors  $(M^{(p)})_{p \geq 2}$  is of the form  $\mu(\mathbf{x} = \sigma \mid (M^{(p)})) \propto \exp(H_N(\sigma))$ , where*

$$H_N(\sigma) = \sum_{p \geq 2} \frac{\gamma_p}{N^{(p-1)/2}} \langle M^{(p)}, \sigma^{\otimes p} \rangle.$$

*Then, the joint law of  $(H_N, \mathbf{x})$  is  $\mu_{\text{pl}}$ .*

The above says that conditioned on  $H_N$ , the distribution of  $\mathbf{x}$  (according to  $\mu_{\text{pl}}$ ) is simply distributed as a sample according to  $\mu_{H_N}$ . That is, the spike  $\mathbf{x}$  resulting in a Hamiltonian  $H_N \sim \mu_{\text{pl}}$  is exchangeable with a sample from  $\mu_{H_N}$ .

The latter of these interpretations will be very useful for us. When dealing with the measure at time  $t$  of stochastic localization applied to the  $p$ -spin model, the primary issue was that it was unclear how to deal with the sample  $\sigma$  drawn from the Gibbs distribution. However, if we could work with the planted  $p$ -spin model, this issue would be absent. Indeed, the exchangeability of the spike and a sample implies that the law of  $\mu_t$  applied to the planted model is given by

$$\mu_{H_N, t\mathbf{x} + \sqrt{t}\mathbf{g}}(\tilde{\sigma}) \propto \exp \left( H_N(\tilde{\sigma}) + \langle t\mathbf{x} + \sqrt{t}\mathbf{g}, \tilde{\sigma} \rangle \right),$$

where  $\mathbf{x}$  is the spike hidden in  $H_N$ . This decouples the randomness of the external field  $t\mathbf{x} + \sqrt{t}\mathbf{g}$  and the disorder of the Hamiltonian  $H_N$  that arises from the Gaussians  $(G^{(p)})_{p \geq 2}$ .

As was shown in [HMP24, Corollary 3.5] and recalled just below, the planted and null models are mutually contiguous. Thus high-probability statements from one model transfer to the other, and it suffices to study the planted model.

For all models satisfying (Strict RS), the measures  $\mu_{\text{null}}(H_N)$  and  $\mu_{\text{pl}}(H_N)$  from Definition 4.7.8 are mutually contiguous, i.e., for any sequence of events  $\mathcal{E}_N$ ,  $\mu_{\text{null}}(\mathcal{E}_N) \rightarrow 0$  whenever  $\mu_{\text{pl}}(\mathcal{E}_N) \rightarrow 0$ .

The transfer from the  $p$ -spin model to the planted model may then be carried out by setting

$$\mathcal{E}_N = \left\{ H_N : \Pr_{\substack{\sigma \sim \mu_{H_N} \\ \mathbf{g} \sim \mathcal{N}(0, I_N)}} \left[ \|\text{Cov}(\mu_{H_N, t\sigma + \sqrt{t}\mathbf{g}})\| > K \right] < e^{-cN^{1/5}} \right\}.$$

This event is very complicated in the null model, but exchangeability makes it tractable in the planted model. In the actual proof, we will require a stronger (quantitative) version of mutual contiguity; see Proposition 4.7.15 for details.

Now, we must understand what the Hamiltonian in the planted model looks like conditioned on the spike.

**Fact 4.7.10.** *Consider the following process: sample  $\mathbf{x} \sim S_N$ ,  $\tilde{H}_N \sim \mu_{\text{null}}$ , and define  $H_N$  by  $H_N(\sigma) = \tilde{H}_N(\sigma) + N \cdot \xi(R(\mathbf{x}, \sigma))$ . Then, the joint law of  $(H_N, \mathbf{x})$  is  $\mu_{\text{pl}}$ .*

Consequently, our goal is to bound the covariance of the distribution

$$\mu_t(\sigma) \propto \exp \left( \tilde{H}_N(\sigma) + N \cdot \xi(R(\mathbf{x}, \sigma)) + \langle t\mathbf{x} + \sqrt{t}\mathbf{g}, \sigma \rangle \right).$$

for  $\tilde{H}_N \sim \mu_{\text{null}}$  with mixture function  $\xi$ . Now, define  $\xi_t$  by  $\xi_t(s) = \xi(s) + ts$ , and extend the definition of the  $p$ -spin model (4.5) to allow a random linear term. Then,

$$\mu_t(\sigma) \propto \exp \left( \underbrace{\tilde{H}_{N,t}(\sigma) + N \cdot \xi_t(R(\mathbf{x}, \sigma))}_{H_{N,t}(\sigma)} \right),$$

where  $\tilde{H}_{N,t} \sim \mu_{\text{null}}$  with mixture function  $\xi_t$ .

**The TAP planted model.** We now turn to controlling the covariance matrix of these models. As we will see below, it is relatively easier to bound the covariance matrix (in fact, the second moment matrix) of a model with zero or small external field. However, for any time  $t > 0$ ,  $H_{N,t}$  has an external field. We will use a method developed in [HMP24] to reduce to the case of a model with zero or small external field.

Let  $\mathbf{m}^{\text{true}} = \mathbf{m}(\mu_t)$ . The main intuition of this reduction is that the Gibbs measure concentrates near a codimension-2 band passing through  $\mathbf{m}^{\text{true}}$  and orthogonal to  $\mathbf{m}^{\text{true}}$  and  $\mathbf{x}$ , and furthermore *the model on this band is essentially a replica symmetric model with no external field*. Moreover, one expects that both  $R(\mathbf{m}^{\text{true}}, \mathbf{m}^{\text{true}})$  and  $R(\mathbf{m}^{\text{true}}, \mathbf{x})$  concentrate near a value  $q_* = q_*(t)$  defined by  $\xi'_t(q_*) = \frac{q_*}{1-q_*}$ .

However,  $\mathbf{m}^{\text{true}}$  is a complicated function of  $H_{N,t}$ , so it is a priori difficult to reason about the joint distribution of  $(\mathbf{m}^{\text{true}}, H_{N,t})$ . Thus, this reduction is formally carried out by conditioning on a TAP fixed point  $\mathbf{m}^{\text{TAP}}$ , which will serve as a proxy for  $\mathbf{m}^{\text{true}}$ . Define the TAP free energy

$$\mathcal{F}_{\text{TAP}}(\mathbf{m}) = H_{N,t}(\mathbf{m}) + \frac{N}{2} \cdot \theta(R(\mathbf{m}, \mathbf{m})) + \frac{N}{2} \log(1 - R(\mathbf{m}, \mathbf{m})),$$

where

$$\theta(s) = \xi(1) - \xi(s) - (1-s)\xi'(s).$$

As shown in [HMP24], for sufficiently small constant  $\iota > 0$ , with probability  $1 - e^{-cN}$   $\mathcal{F}_{\text{TAP}}$  has a unique critical point  $\mathbf{m}^{\text{TAP}}$  in the region  $\mathcal{S}_\iota$  defined by  $R(\mathbf{m}, \mathbf{m}), R(\mathbf{m}, \mathbf{x}) \in [q_* - \iota, q_* + \iota]$ . Due to the existence and uniqueness of  $\mathbf{m}^{\text{TAP}}$ , it becomes possible to relate  $H_{N,t}$  to a “TAP-planted model” where one samples  $\mathbf{m}^{\text{TAP}}$  first, and then samples  $H_{N,t}$  conditional on  $\nabla \mathcal{F}_{\text{TAP}}(\mathbf{m}) = 0$ :

**Lemma 4.7.11** (Planted-to-TAP reduction; see Lemma 4.7.20). *For any small constant  $\iota > 0$ , the following holds. For any  $H_{N,t}$ -measurable event  $\mathcal{E}$ , if*

$$\sup_{\mathbf{m}^{\text{TAP}} \in \mathcal{S}_\iota} \Pr(\mathcal{E} | \nabla \mathcal{F}_{\text{TAP}}(\mathbf{m}^{\text{TAP}}) = 0) \rightarrow 0,$$

*then  $\Pr(\mathcal{E}) \rightarrow 0$ .*

Crucially, the conditional law of  $H_{N,t}$  in the TAP-planted model is very tractable, as (for a fixed  $\mathbf{m}^{\text{TAP}}$ )  $\nabla \mathcal{F}_{\text{TAP}}(\mathbf{m}^{\text{TAP}}) = 0$  amounts to a linear constraint on the Gaussian process  $H_{N,t}$ . The resulting explicit conditional law of  $H_{N,t}$  is described in Lemma 4.7.22.

**Remark 4.7.12.** While it will not be relevant to our purposes, [EAMS22, AMS23, HMP24] have shown that  $\mathbf{m}^{\text{TAP}}$  typically approximates  $\mathbf{m}^{\text{true}}$  well, in the sense that  $\|\mathbf{m}^{\text{true}} - \mathbf{m}^{\text{TAP}}\|^2 = O(1)$ , thereby justifying the heuristic that  $\mathbf{m}^{\text{TAP}}$  is a proxy for  $\mathbf{m}^{\text{true}}$ .

**Remark 4.7.13.** The idea of reducing to a TAP-planted model has also been used beyond the setting of sampling from spherical spin glasses. In the recent work [Hua24], an analogous reduction is used to obtain the capacity of the Ising perceptron. In this application, passage to the TAP-planted model is used to tightly control a partition function rather than to bound a covariance matrix.

Consequently, we can now work within the TAP-planted model. Let  $H_{\text{TAP}}$  denote the Hamiltonian  $H_{N,t}$  after conditioning on  $\mathbf{x}$  and  $\nabla \mathcal{F}_{\text{TAP}}(\mathbf{m}^{\text{TAP}}) = 0$ . As

$$\text{Cov}(\mu_{H_{\text{TAP}}}) \leq \mathbf{E}_{\sigma \sim \mu_{H_{\text{TAP}}}} (\sigma - \mathbf{v})(\sigma - \mathbf{v})^\top$$

for any  $\mathbf{v} \in \mathbb{R}^N$  (with equality at  $\mathbf{v} = \mathbf{m}^{\text{true}}$ ), it suffices to control the operator norm of

$$\mathbf{E}_{\sigma \sim \mu_{H_{\text{TAP}}}} (\sigma - \mathbf{m}^{\text{TAP}})(\sigma - \mathbf{m}^{\text{TAP}})^\top.$$

**Reduction to slices of the sphere.** Next, to control the covariance, let us decompose the sphere into codimension-2 slices

$$S(a, b) := \left\{ \sigma \in S_N : R(\sigma, \mathbf{m}) = \left(1 + \frac{a}{\sqrt{N}}\right) R(\mathbf{m}, \mathbf{m}), R(\sigma, \mathbf{x}) = \left(1 + \frac{b}{\sqrt{N}}\right) R(\mathbf{x}, \mathbf{m}) \right\},$$

with the central slice centered at  $\mathbf{m} = \mathbf{m}^{\text{TAP}}$ . Let  $\mu_{a,b}$  be the measure  $\mu_{H_{\text{TAP}}}$  conditioned to lie in the codimension-2 slice  $S(a, b)$ .

The concentration of the Gibbs measure described in the previous section implies that, viewed as random variables of a sample  $\sigma \sim \mu_{H_{\text{TAP}}}$ ,  $a$  and  $b$  are well-concentrated around 0. Let  $v_{a,b}$  be the center of  $S(a, b)$ . The covariance of the distribution may be bounded as

$$\begin{aligned} \text{Cov}(\mu) &\leq \mathbf{E}(\sigma - \mathbf{m})(\sigma - \mathbf{m})^\top \\ &= \mathbf{E}_{(a,b)} \mathbf{E}_{\sigma \sim \mu_{a,b}} (\sigma - v_{a,b} + v_{a,b} - \mathbf{m})(\sigma - v_{a,b} + v_{a,b} - \mathbf{m})^\top \\ &\leq 2\mathbf{E}_{(a,b)} \mathbf{E}_{\sigma \sim \mu_{a,b}} (\sigma - v_{a,b})(\sigma - v_{a,b})^\top + 2\mathbf{E}_{(a,b)} (v_{a,b} - \mathbf{m})(v_{a,b} - \mathbf{m})^\top \\ &\leq 2\mathbf{E}_{(a,b)} \mathbf{E}_{\sigma \sim \mu_{a,b}} (\sigma - v_{a,b})(\sigma - v_{a,b})^\top + 2\mathbf{E}_{(a,b)} O(a^2 + b^2). \end{aligned} \quad (4.6)$$

One can interpret  $v_{a,b}$  as explaining the variation within the slice originating from the  $\mathbf{m}$  and  $\mathbf{x}$  directions. Hence, as alluded to in the previous discussion about the TAP planted model, the key fact is that under  $\mu_{a,b}$ , the recentered sample  $\sigma - v_{a,b}$  is a sample from a spherical spin glass in two lower dimensions, as can be shown by calculating the covariance of the (conditioned) Gaussian process  $H_{\text{TAP}}$  restricted to this slice. This verification is carried out in Corollary 4.7.24.

These codimension-2 models have the crucial property that the spherical spin glass on the slice  $a = b = 0$  is a model satisfying (Strict RS) with no external field (i.e. degree-1 term), while nearby slices have a small (random) external field of magnitude  $\sqrt{a^2 + b^2}$ . In particular, the first term of (4.6) requires bounding the second moment of a Gibbs sample from a strictly RS model with small (random) external field. As a result, (4.6) would be bounded if we proved the following.

1. Let  $H_N$  be the Hamiltonian of a slightly generalized mixed  $p$ -spin model, where we allow the mixture function  $\xi$  to have a small linear term  $\gamma_1 q$  (in our proofs we allow  $\gamma_1^2 \leq N^{-4/5}$ ), such that the non-degree-1 part  $\xi_{\sim 1}$  of  $\xi$  satisfies (Strict RS). Then, with high probability,

$$\|\mathbf{E}_{\mu_{H_N}} \sigma \sigma^\top\|_{\text{op}} = O(1 + \gamma_1^2 N).$$

Much of Section 4.8 is dedicated to showing this.

2. The second moments of  $a$  and  $b$  are  $O(1)$ . In fact, we will show in Lemma 4.7.35 that they are essentially  $O(1)$ -subgaussian.

Let us start by explaining how to show subgaussianity.

**Subgaussianity of  $a, b$ .** The distribution  $\nu$  of  $(a, b)$  is given by

$$\nu(a, b) \propto \exp \left( \log \widehat{Z}_{a,b} + \frac{N-4}{2} \log \left( 1 - \frac{\|v_{a,b}\|^2}{N} \right) + H_{\text{TAP}}(v_{a,b}) \right).$$

Here, the first term  $\log \widehat{Z}_{a,b}$  is the free energy of the  $(N-2)$ -dimensional  $p$ -spin model  $\mu_{a,b}$ , obtained by restricting  $\mu_{H_{\text{TAP}}}$  to the slice  $S(a, b)$  and rescaling the distribution to lie on  $S_{N-2}$ . The second term is an effective decrement in the free energy caused by the radius of the sphere  $S(a, b)$  shrinking for larger values of  $a$  and  $b$ . The third term is an effective increment in the free energy coming from the energy of  $H_{\text{TAP}}$  at the center of the slice  $S(a, b)$ .

For a fixed  $(a, b)$ , the only random quantities in the definition of  $\nu$  are the first and third terms. In Theorem 4.7.27, proved in Section 4.8, we show that the first term may essentially be approximated by a deterministic function of the mixture function of  $\mu_{a,b}$ , at the cost of incurring a small  $O(1)$  error. We do not elaborate on the details of this proof in the technical overview; it is similar to that used to bound the covariance (which we explain shortly). The third term is similar, and is a deterministic function plus a small Gaussian, whose variance is  $O(a^2 + b^2)$ .

Given these bounds, we may show that the distribution  $\nu$  is strongly log-concave at 0 with high probability over the randomness of  $H_{\text{TAP}}$ . A simple perturbation argument then implies that  $\nu$  is strongly log-concave in a macroscopic neighborhood of 0, implying subgaussianity.

**Covariance bound for strictly RS models with small external fields.** The covariance bound has now boiled down to bounding  $\|M\|_{\text{op}}$ , for  $M = \mathbf{E}_{\mu_{H_N}}[\sigma\sigma^\top]$  the second moment matrix of model satisfying Eq. (Strict RS) with small external field. Note that  $M$  is a  $H_N$ -measurable random variable.

The proof proceeds in two high level steps, which we carry out in Section 4.8.

1. We show using the second moment method that with positive probability over  $H_N$ ,  $\|M\|_{\text{op}}$  is bounded.
2. Using a much simpler argument, we can show that  $\|M\|_{\text{op}}$  is essentially  $O(N^{-1/10})$ -Lipschitz in the disorder. Hence, by gaussian concentration, it concentrates very well around its expectation (which is  $O(1)$  by the positive probability bound).

Let us elaborate a bit more on the proof of the first point above. It turns out that, under the condition (Strict RS) with small external field, the leading order contribution to  $M$  comes from the degree-2 part of the Hamiltonian  $H_{N,2}(\sigma) = \frac{\gamma_2}{N^{1/2}} \sum_{i,j} g_{i,j} \sigma_i \sigma_j$ . We will

ultimately reduce the study of the covariance matrix of  $\mu_{H_N}$  to that of  $\mu_{H_{N,2}}$ , and then show boundedness of  $\text{Cov}(\mu_{H_{N,2}})$  using random matrix theory. A similar strategy of isolating the degree-2 component of  $H_N$  was used to study the partition function and magnetization of strictly RS models in [HMP24].

**Degree-2 behavior.** Let us discuss the typical behavior of the covariance of  $\mu_{H_{N,2}}$ . Define the degree-2 Gibbs measure

$$d\mu_{H_{N,2}}(\sigma) \propto \exp(H_{N,2}(\sigma)) d\rho(\sigma),$$

with corresponding partition function  $Z_{N,2} = \int \exp(H_{N,2}(\sigma)) d\rho(\sigma)$ . This is the spherical Sherrington-Kirkpatrick model with interaction matrix  $A = \nabla^2 H_N(0)$ ; note that  $A$  is a scaled GOE matrix. Observe that if we shift  $A$  by a constant multiple of the identity  $\gamma \text{Id}_N$ , the measure does not change, as it is supported on  $S_N$ . The crucial observation is the following:

For a careful choice of  $\gamma$ , the measure  $d\mu_{H_{N,2}}(\sigma) \propto \exp(-\frac{1}{2}\langle \sigma, (\gamma \text{Id}_N - A)\sigma \rangle) d\rho(\sigma)$  looks like a Gaussian with covariance  $(\gamma \text{Id}_N - A)^{-1}$ .

In fact, we will see that it suffices to pick  $\gamma = 1 + \xi''(0)$ . The typical value of  $\|x\|_2^2$ , where  $x \sim \mathcal{N}(0, \gamma \text{Id}_N - A)^{-1}$ , is equal to  $\text{Tr}(\gamma \text{Id}_N - A)^{-1}$ . By approximating this trace using the semicircle law for the eigenvalues of  $A$  and the explicit choice of  $\gamma$ , we see that  $\|x\|_2^2 \approx N$ , which justifies the heuristic that this Gaussian approximates the spherical distribution  $\mu_{H_{N,2}}$ .

For the above discussion to be well-defined, we require that  $\gamma \text{Id}_N - A$  is positive definite, which can only occur if the maximum eigenvalue of  $A$  is bounded above by  $\gamma = 1 + \xi''(0)$ . By standard concentration inequalities about the maximum eigenvalue of a GOE matrix, this holds with a constant margin with exponentially good probability. Thus, at least for typical realizations of  $H_{N,2}$ , the covariance will have bounded operator norm. To make this rigorous, we will use the Laplace transform to precisely control the moments of the overlaps, as was previously done in [BL16, HMP24].

**Reduction to degree-2.** Below, we give some justification for why one should expect to be able to reduce to the degree-2 behavior. We will heuristically argue this by showing that the partition function  $Z_N$  is essentially controlled by the degree-2 portion.

To simplify the discussion, let us assume that we are in a  $2+p$  spin model, so that  $H_N(\sigma) = H_{N,2}(\sigma) + H_{N,p}(\sigma)$ , where  $H_{N,p}(\sigma) = \frac{\gamma_p}{N^{(p-1)/2}} \sum_{i_1, \dots, i_p=1}^N g_{i_1, \dots, i_p} \sigma_{i_1} \cdots \sigma_{i_p}$ . The corresponding mixture function decomposes as  $\xi(q) = \gamma_2^2 q^2 + \xi_{\sim 2}(q)$ , so that  $\xi_{\sim 2}(q) = \gamma_p^2 q^p$  corresponds to the non degree-2 part of the mixture function. It turns out that, once we condition on  $H_{N,2}$  (and hence the value of  $Z_{N,2}$ ), the full partition function  $Z_N$  is essentially deterministic. Indeed, we will show in Proposition 4.8.2 that with very high probability,

$$Z_N \approx Z_{N,2} e^{N \xi_{\sim 2}(1)/2}$$

To see why this is reasonable, let us consider the first two moments of  $Z_N$  conditioned on the degree-2 Hamiltonian  $H_{N,2}$ . Indeed, let  $\mathbf{E}_{\sim 2}$  denote expectation with respect to  $H_{N,p}$  conditioned on  $H_{N,2}$ . Standard gaussian MGF calculations yield  $\mathbf{E}_{\sim 2} Z_N = e^{N\xi_{\sim 2}(1)/2} Z_{N,2}$  and

$$\begin{aligned}\mathbf{E}_{\sim 2}[Z_N^2] &= Z_{N,2}^2 e^{N\xi_{\sim 2}(1)} \int \exp(N\xi_{\sim 2}(R(\sigma^1, \sigma^2))) d\rho^{\otimes 2}(\sigma^1, \sigma^2) \\ &= (\mathbf{E}_{\sim 2} Z_N)^2 \int \exp(N\gamma_p^2 R(\sigma^1, \sigma^2)^p) d\rho^{\otimes 2}(\sigma^1, \sigma^2).\end{aligned}$$

At sufficiently high temperatures, the typical overlap behavior  $R(\sigma^1, \sigma^2) \asymp N^{-1/2}$ , where  $\sigma^1, \sigma^2$  are iid draws from  $\mu_{H_N}$ . This matches the overlap behavior at infinite temperature, where the Gibbs distribution is uniform. Then, pretending that  $R(\sigma^1, \sigma^2) = cN^{-1/2}$  for all  $\sigma^1, \sigma^2$ , we obtain that

$$\mathbf{E}_{\sim 2}[Z_N^2] \approx (\mathbf{E}_{\sim 2} Z_N)^2 \exp\left(c^p \gamma_p^2 N^{1-p/2}\right).$$

Since  $p \geq 3$ , it follows that, conditional on  $H_{N,2}$ , the conditional variance of  $Z_N$  is tiny compared to its conditional expectation. In summary, we see that the higher degree portions of the partition function have negligible contributions to the fluctuations of  $Z_N$ , so that the typical behavior of  $Z_N$  is controlled by  $Z_{N,2}$ .

Turning now to the covariance bound, we will control the  $(i, j)$ th covariance entry  $M_{i,j} := \int \sigma_i \sigma_j e^{H_N(\sigma)} d\rho(\sigma)$ . A crucial fact is that, by rotational invariance of the sphere and gaussians, we can rotate to the eigenbasis of  $A = \nabla^2 H_N(0)$  so that  $A$  becomes diagonal. When  $A$  is diagonal, one can in fact show that

$$\mathbf{E}_{\sim 2}[M_{i,j}^2] \lesssim \frac{1}{N^2} (\mathbf{E}_{\sim 2} M_{i,j})^2,$$

where  $\mathbf{E}_{\sim 2} M_{i,j}$  can be interpreted as (up to normalization) the predicted  $(i, j)$ th covariance entry by just looking at the degree-2 randomness; see Propositions 4.8.17 and 4.8.18 for details. It follows that the Frobenius norm error of the true covariance compared to the degree-2 covariance is  $O(1)$ . Combined with the typical behavior of the degree-2 covariance being essentially the diagonal matrix  $((1 + \xi''(0))\text{Id}_N - A)^{-1}$ , we conclude an  $O(1)$  covariance bound for  $\mu_{H_N}$ .

Although this direct moment approach can be made rigorous at sufficiently high temperature, it will not cover the entire regime (Strict RS) of our main theorem. To deal with this, we will use the *free energy typical truncation* recently introduced by [HS23]. The main idea is that, while pairs  $\sigma^1, \sigma^2$  with overlap  $R(\sigma^1, \sigma^2) \asymp N^{-1/2}$  do not necessarily dominate the second moment  $\mathbf{E}[Z_N^2]$  throughout the regime (Strict RS), there is a truncation  $\tilde{Z}_N$  accounting for nearly all of  $Z_N$ , whose second moment is dominated by such pairs. We defer the details to the following sections.

## Null models, planted models, and contiguity

As described in the technical overview, we will need a quantitative strengthening of contiguity between the null and planted models. For convenience, let us restate the definition of the planted model.

**Definition 4.7.8** (Planted  $p$ -spin model). The planted measure  $\mu_{\text{pl}}$  is a joint law over a Hamiltonian  $H_N$  and a *spike*  $x \in S_N$  given by

$$d\mu_{\text{pl}}(H_N, x) \propto \exp(H_N(x)) \cdot d\rho(x) \cdot d\mu_{\text{null}}(H_N),$$

where  $\rho$  is the uniform measure over  $S_N$  and  $\mu_{\text{null}}$  is the law over  $p$ -spin Hamiltonians with mixture function  $\xi$ . We frequently abuse notation to let  $\mu_{\text{pl}}(H_N)$  denote the marginal of  $\mu_{\text{pl}}$  on  $H_N$ .

**Remark 4.7.14** (Interpretation of planted model). Equivalently, the planted measure  $\mu_{\text{pl}}$  can be described as follows.

- Sample  $x \sim S_N$ .
- Sample  $\tilde{H}_N \sim \mu_{\text{null}}$ .
- Define  $H_N$  by  $H_N(\sigma) = \tilde{H}_N(\sigma) + N \cdot \xi(R(x, \sigma))$ .

The following Bayesian interpretation of  $\mu_{\text{pl}}$  will make the planted model amenable to explicit calculation. For  $(x, H_N)$  sampled from  $\mu_{\text{pl}}$ , the posterior distribution  $x|H_N$  is described by the density:

$$d\mu_{x|H_N}(\sigma) \propto \exp(H_N(\sigma)) d\rho(\sigma).$$

Therefore, the distribution of  $(H_N, \sigma)$  for  $\sigma \sim \mu_{H_N}$  is identical to that of  $(H_N, x)$ .

In order to show the probability bound of  $1 - e^{-cN^{1/5}}$  in Lemma 4.7.6, we will prove the following quantitative strengthening of mutual contiguity, under the following quantitative strict RS condition. Note that, since the proof of Theorem 4.7.2 union bounds over  $\text{poly}(N)$  many values of  $\beta$ , quantitative control of the error in Lemma 4.7.6 is needed to carry out the proof.

**Condition 4** ( $\varepsilon$ -strict replica symmetry). We say  $\xi$  is  $\varepsilon$ -strictly replica symmetric if for all  $q \in (0, 1)$ ,

$$\frac{1}{q^2} \cdot (\xi(q) + q + \log(1 - q)) \leq -\varepsilon/2. \quad (4.7)$$



Under this assumption, we prove the following quantitative contiguity result in Section 4.8.

**Proposition 4.7.15** (Quantitative contiguity). *Under Condition 4, there exists  $c = c(\varepsilon) > 0$  such that for any event  $\mathcal{E}$ , if  $\mu_{\text{pl}}(\mathcal{E}) = p$ , then  $\mu_{\text{null}}(\mathcal{E}) \leq e^{-cN^{1/5}} + e^{\frac{1}{c}\sqrt{\log \frac{2}{p}}} p$ .*

Thus, from now on, we work under the planted model. One reason the planted model is easier to work with is because of the following lemma, which provides a simple description of the distribution of  $\mu_t$  (by describing the distribution of the external field at time  $t$ ) in the planted model.

**Lemma 4.7.16.** *Let  $\mu_t$  be the distribution after running stochastic localization with the Id driving matrix for time  $t$  initialized at  $\mu_{H_N}$ . Then  $\mu_t$  arises as the Gibbs distribution of the Hamiltonian  $H_{N,t}(\sigma)$ :*

$$H_{N,t}(\sigma) = H_N(\sigma) + \langle \mathbf{y}_t, \sigma \rangle,$$

where

$$(H_N, \mathbf{y}_t) \stackrel{\text{law}}{=} (H_N, t\mathbf{x} + \sqrt{t}\mathbf{g}),$$

where  $\mathbf{x} \sim S_N$ ,  $H_N \sim \mu_{\text{pl}}(\cdot|\mathbf{x})$ , and  $\mathbf{g} \sim \mathcal{N}(0, \text{Id}_N)$ .

**Notation** ( $\mu_{\text{pl},t}$ ,  $\xi_t(q)$ ,  $\gamma(q)$ ). We will use  $\mu_{\text{pl},t}$  to denote the distribution of the pair  $(H_{N,t}, \mathbf{x})$ ,  $\xi_t(q) = \xi(q) + tq$  to refer to the mixture function of  $H_{N,t}$ , and  $\gamma(q)$  to refer to the function  $q\xi'_t(q)$ .

In the subsequent sections, we will prove a high probability covariance bound for  $\mu_t$  at a fixed time  $t$  under the planted model.

**Lemma 4.7.17.** *There exist universal constants  $c, T, K$ , such that for any  $t \in [0, T]$ , with probability at least  $1 - e^{-cN^{1/5}}$  over the randomness of  $H_N$  drawn from  $\mu_{\text{pl},t}$ , we have  $\|\mathbf{Cov}(\mu_t)\| \leq K$ .*

We now have all the necessary ingredients to prove the covariance bound along the entire localization path for the null model.

*Proof of Lemma 4.7.6.* Define  $\mathcal{T}$  as the discrete set  $\{iT/\delta : 1 \leq i \leq 1/\delta, i \in \mathbb{Z}\}$  for  $\delta = N^{-100}$ . We will prove:

$$\Pr_{H_N} \left[ \Pr_{(\mu_t)|H_N} \left[ \|\mathbf{Cov}(\mu_t)\|_{\text{op}} < K \text{ for all } t \in \mathcal{T} \right] \geq 1 - e^{-cN^{1/5}} \right] \geq 1 - e^{-cN^{1/5}}.$$

A simple continuity argument can be used to derive the desired statement from the above. By taking a union bound over all elements of  $\mathcal{T}$ , along with Proposition 4.7.15 and Lemma 4.7.17, we can conclude:

$$\mathbf{E}_{H_N} \Pr_{(\mu_t)|H_N} \left[ \|\mathbf{Cov}(\mu_t)\|_{\text{op}} > K \text{ for some } t \in \mathcal{T} \right] \leq e^{-2cN^{1/5}}.$$

The resulting statement then follows from Markov's inequality on the random variable

$$\Pr_{(\mu_t)|H_N} \left[ \|\mathbf{Cov}(\mu_t)\|_{\text{op}} > K \text{ for some } t \in \mathcal{T} \right]. \quad \square$$

## TAP planted models

In this section, we formally introduce the TAP planted model and relate it to the planted model from the previous section.

**Definition 4.7.18.** Let  $H_N$  be a planted Hamiltonian with mixture function  $\xi$ , and define

$$\theta(s) = \xi(1) - \xi(s) - (1-s)\xi'(s).$$

The associated *TAP free energy* is defined by

$$\mathcal{F}_{\text{TAP}}(\mathbf{m}) = H_N(\mathbf{m}) + \frac{N}{2} \cdot \theta(R(\mathbf{m}, \mathbf{m})) + \frac{N}{2} \cdot \log(1 - R(\mathbf{m}, \mathbf{m})).$$

While the TAP free energy is interesting for a multitude of reasons, we will be interested in it because its fixed points provide a good proxy for the mean. Furthermore, the linearity of the TAP free energy in the Gaussian coefficients of the Hamiltonian provides certain desirable properties (that using the true mean would not allow).

We use the following TAP fixed-point fact from [HMP24, Fact 4.2].

**Fact 4.7.19** (TAP fixed-point equation). *Let  $\xi$  be a mixture function satisfying the condition (SL). For any  $t \in [0, \infty)$ , let  $\xi_t(q) = \xi(q) + tq$ . Then there is a unique solution in  $[0, 1)$ , which we denote  $q_* = q_*(t)$ , to*

$$\xi'_t(q_*) = \frac{q_*}{1 - q_*}.$$

**Lemma 4.7.20.** For any  $K > 0$ , sufficiently small (constant)  $\iota > 0$  and  $\mathbf{x} \in S_N$ :

$$\begin{aligned} & \Pr_{\mu_{\text{pl},t}} \left[ \|\text{Cov}(\mu_{H_{N,t}})\|_{\text{op}} \geq K \right] \\ & \leq C \cdot \sup_{\mathbf{m} \in \mathcal{S}_\iota} \Pr_{\mu_{\text{pl},t}|\mathbf{x}} \left[ \|\text{Cov}(\mu_{H_{N,t}})\|_{\text{op}} \geq K \wedge \mathcal{E}_\iota \mid \nabla \mathcal{F}_{\text{TAP}}(\mathbf{m}) = 0 \right]^{1/2} + 2e^{-cN}, \end{aligned}$$

where

$$\mathcal{S}_\iota = \mathcal{S}_\iota(\mathbf{x}) := \left\{ \mathbf{m} \in \mathbb{R}^N : |R(\mathbf{m}, \mathbf{m}) - q_*|, |R(\mathbf{m}, \mathbf{x}) - q_*| < \iota \right\},$$

and  $\mathcal{E}_\iota$  is the event that  $\mathcal{F}_{\text{TAP}}$  has a unique critical point  $\mathbf{m}^{\text{TAP}}$  in  $\mathcal{S}_\iota$ , and that

$$\Pr_{\sigma \sim \mu_{H_{N,t}}} \left[ R(\sigma, \mathbf{m}^{\text{TAP}}), R(\sigma, \mathbf{x}) \in [q_* - \iota, q_* + \iota] \right] \geq 1 - e^{-cN}.$$

*Proof.* The above statement is effectively due to [HMP24, Propositions 4.4(d) and 4.5(a)]. For the reader's convenience, we include the steps to arriving at the above statement. For any event  $\mathcal{E}$  (and in particular, for the event  $\mathcal{E}$  defined in [HMP24, Proposition 4.4]), we have:

$$\Pr \left[ \|\text{Cov}(\mu_{H_{N,t}})\|_{\text{op}} \geq K \right] \leq \Pr \left[ \|\text{Cov}(\mu_{H_{N,t}})\|_{\text{op}} \geq K \wedge \mathcal{E}_\iota \wedge \mathcal{E} \right] + \Pr \left[ \overline{\mathcal{E}} \right] + \Pr \left[ \overline{\mathcal{E}_\iota} \right].$$

The desired statement follows by observing that  $\Pr \left[ \overline{\mathcal{E}_\iota} \right] \leq e^{-cN}$  by [HMP24, Proposition 4.5(a)],  $\Pr \left[ \overline{\mathcal{E}} \right] \leq e^{-cN}$  by [HMP24, Proposition 4.4], and applying [HMP24, Proposition 4.4(d)] with  $X = \mathbf{1} \left[ \|\text{Cov}(\mu_{H_{N,t}})\|_{\text{op}} \geq K \wedge \mathcal{E}_\iota \right]$ .  $\square$

Lemma 4.7.20 reduces our task to studying the covariance matrix in a *conditional* planted model.

**Notation** ( $\mu_{\text{TAP}}, H_{\text{TAP}}, q_m, q_x$ ). For  $\mathbf{x} \sim S_N$  and  $\mathbf{m} \in \mathbb{R}^N$ , we consider the distribution  $\mu_{\text{TAP}, \mathbf{x}, \mathbf{m}}$  of  $H_{\text{TAP}}$  for  $H_{\text{TAP}} \sim (\mu_{\text{pl},t}|\mathbf{x}, \nabla \mathcal{F}_{\text{TAP}}(\mathbf{m}) = 0)$ . We use  $q_m$  and  $q_x$  to refer to  $R(\mathbf{m}, \mathbf{m})$  and  $R(\mathbf{m}, \mathbf{x})$  respectively.

**Lemma 4.7.21.** Let  $\mathbf{x} \in S_N$ , let  $\mathcal{S}_\iota$  be as in Lemma 4.7.20, and let  $\mathbf{m} \in \mathcal{S}_\iota$ . Then for an absolute constant  $K > 0$ ,

$$\Pr_{H_{\text{TAP}} \sim \mu_{\text{TAP}, \mathbf{x}, \mathbf{m}}} \left[ \|\text{Cov}(\mu_{H_{\text{TAP}}})\| \geq K \wedge \mathcal{E}_\iota \right] \leq e^{-cN^{1/5}}.$$

*Proof of Lemma 4.7.17.* The statement is immediate from Lemmas 4.7.20 and 4.7.21.  $\square$

The rest of this section is dedicated to proving Lemma 4.7.21. As a first step, we determine the law of the typical Hamiltonian sampled from  $\mu_{\text{TAP}}$ . We prove the following lemma in Section 4.10 — it follows by routine calculations, using the form of the law of a Gaussian process conditioned on the value of a linear function of it. Recall that  $\xi_t(q) = \xi(q) + tq$ .

**Lemma 4.7.22.** *The law of Hamiltonian  $H_{\text{TAP}} \sim \mu_{\text{TAP},x,m}$  is described by a Gaussian process  $(H_{\text{TAP}}(\sigma))_{\sigma \in \mathcal{S}_N}$  defined by*

$$\begin{aligned} \mathbf{E} H_{\text{TAP}}(\sigma) &= N \xi_t(R(x, \sigma)) - \langle x, v(\sigma) \rangle \cdot \xi'_t(q_x) - \frac{\xi'_t(R(\mathbf{m}, \sigma))}{\gamma'(q_m)} \cdot \langle \mathbf{m}, \sigma \rangle \cdot \left( \theta'(q_m) - \frac{1}{1 - q_m} \right) \\ \frac{1}{N} \mathbf{Cov}(H_{\text{TAP}}(\sigma), H_{\text{TAP}}(\sigma')) &= \xi_t(R(\sigma, \sigma')) - R(\sigma, \sigma') \frac{\xi'_t(R(\mathbf{m}, \sigma)) \xi'_t(R(\mathbf{m}, \sigma'))}{\xi'_t(q_m)} \\ &\quad + \frac{\xi''_t(q_m)}{\gamma'(q_m) \xi'_t(q_m)} \gamma(R(\mathbf{m}, \sigma)) \gamma(R(\mathbf{m}, \sigma')), \end{aligned}$$

where

$$\begin{aligned} v(\sigma) &:= \frac{\xi'_t(R(\mathbf{m}, \sigma))}{\xi'_t(q_m)} \left[ I - \frac{\xi''_t(q_m)}{\gamma'(q_m)} \cdot \frac{\mathbf{m} \mathbf{m}^\top}{N} \right] \sigma \\ \gamma(q) &:= q \cdot \xi'_t(q). \end{aligned}$$

For the proofs below, it will also be helpful to consider Hamiltonians with a linear term representing an external field. For a sequence  $\gamma_1, \gamma_2, \dots, \gamma_{p_*}$ , consider the following generalization of  $H_N$  from (4.5):

$$H_N(\sigma) := \sum_{p \geq 1} \frac{\gamma_p}{N^{(p-1)/2}} \sum_{i_1, \dots, i_p=1}^N \mathbf{g}_{i_1, \dots, i_p} \sigma_{i_1} \cdots \sigma_{i_p}. \quad (4.8)$$

This has mixture function

$$\xi(s) = \sum_{p \geq 1} \gamma_p^2 s^p.$$

We will write  $\xi_{\sim 1}(s) = \sum_{p \geq 2} \gamma_p^2 s^p$  for the part of  $\xi$  with degree at least 2, and extend Condition 4 to such  $\xi$  as follows.

**Condition 5** ( $\varepsilon$ -strict replica symmetry). *We say  $\xi$  is  $\varepsilon$ -strictly replica symmetric if  $\gamma_1^2 \leq N^{-4/5}$  and  $\xi_{\sim 1}$  satisfies Condition 4.*

## Slices in TAP planted models

For succinctness, we shall fix  $x \in S_N$  and  $m \in \mathcal{S}_l(x)$ , and use  $\mu_{\text{TAP}}$  to refer to the distribution  $\mu_{\text{TAP},x,m}$ . For  $H_{\text{TAP}} \sim \mu_{\text{TAP}}$ , we are interested in bounding the covariance of  $\mu_{H_{\text{TAP}}}$ . To reason about  $\mu_{H_{\text{TAP}}}$ , we write it as a mixture of distributions over  $(N-2)$ -dimensional slices of  $S_N$ . For  $a, b \in \mathbb{R}$ , we define

$$S(a, b) := \left\{ \sigma \in S_N : R(\sigma, m) = \left(1 + \frac{a}{\sqrt{N}}\right) q_m, R(\sigma, x) = \left(1 + \frac{b}{\sqrt{N}}\right) q_x \right\}.$$

Let  $r_{a,b}$  refer to the radius of this slice, which is equal to

$$r_{a,b} = \sqrt{1 - q_m \left(1 + \frac{a}{\sqrt{N}}\right)^2 - \frac{q_m q_x^2}{q_m - q_x^2} \left(\frac{a-b}{\sqrt{N}}\right)^2}.$$

Note in particular that

$$q_m \left(1 + \frac{a}{\sqrt{N}}\right)^2 + \frac{q_m q_x^2}{q_m - q_x^2} \left(\frac{a-b}{\sqrt{N}}\right)^2 \geq q_m \left(1 + \frac{a}{\sqrt{N}}\right)^2. \quad (4.9)$$

We refer to the uniform distribution on this slice as  $\rho_{a,b}$ , and the partition function on the slice as

$$Z_{a,b} := \mathbf{E}_{\sigma \sim \rho_{a,b}} \exp(H_{\text{TAP}}(\sigma)).$$

With this definition, the partition function of the original Hamiltonian is given by

$$Z = \Lambda_N \int Z_{a,b} r_{a,b}^{N-4} d(a, b)$$

for some fixed number  $\Lambda_N$  depending only on  $N$ .

**Remark 4.7.23.** To see why we scale by  $r_{a,b}^{N-4}$ , observe that when  $H_{\text{TAP}}$  is the constant-0 Hamiltonian, the resulting distribution on the sphere should be uniform. The distribution restricted to each slice must also be uniform. However, not all slices are weighted equally — slice that have smaller radii must be downweighted accordingly, with this weighting proportional to  $r_{a,b}^{N-4}$  for  $S(a, b)$ .<sup>1</sup>

Use  $\nu$  to refer to the distribution over  $(a, b)$  where  $d\nu(a, b) \propto Z_{a,b} r_{a,b}^{N-4} d(a, b)$ , and  $\mu_{a,b}$  to refer to the distribution  $\mu_{H_{\text{TAP}}}$  restricted to  $S(a, b)$ . Now, we can write  $\mu_{H_{\text{TAP}}}$  as the following mixture:

$$\mu_{H_{\text{TAP}}} = \mathbf{E}_{(a,b) \sim \nu} \mu_{a,b}.$$

---

<sup>1</sup>The constant of proportionality here is something depending only on  $N$ .

We will need coarse understanding of the tails of  $\nu$ , and fine understanding of the distribution  $\mu_{a,b}$  for small  $a$  and  $b$ .

Now, let us probe the distribution  $\mu_{a,b}$ .

$$\frac{d\mu_{a,b}}{d\rho_{a,b}}(\sigma) = \frac{\exp(H_{\text{TAP}}(\sigma))}{\mathbf{E}_{\sigma \sim \rho_{a,b}} \exp(H_{\text{TAP}}(\sigma))}$$

Since  $S(a, b)$  can be naturally identified with  $S_{N-2}$ , the first step to understanding  $\mu_{a,b}$  is to express it as a  $p$ -spin model on  $S_{N-2}$ . To do so, we will verify that *some* Hamiltonian that gives rise to  $\mu_{a,b}$  has a mixture function that is given by a polynomial in the overlap. We can write any  $\sigma \in S(a, b)$  as:

$$\sigma = \sqrt{N} \cdot v(a, b) + \underbrace{\sqrt{1 - \|v(a, b)\|^2}}_{r_{a,b}} \sigma_{\perp}$$

for  $\sigma_{\perp} \in S_N$  orthogonal to  $\mathbf{m}$  and  $\mathbf{x}$ , and for  $v(a, b)$  in the span of  $\mathbf{m}$  and  $\mathbf{x}$ . Let  $Q$  be an isometric linear transformation that maps  $S_{N-2}$  to  $S_N \perp \{\mathbf{m}, \mathbf{x}\}$ . We can write  $H_{\text{TAP}}(\sigma) = H_{\text{TAP}}(\sqrt{N} \cdot v(a, b) + r_{a,b} Q\tau)$  (where  $\tau \in S_{N-2}$ ). The following is a consequence of Lemma 4.7.22, and is proved in Section 4.10.

**Corollary 4.7.24.** *For a fixed choice of  $a$  and  $b$ , the Gaussian process  $\left(H_{\text{TAP}}(\sqrt{N} \cdot v(a, b) + r_{a,b} Q\tau)\right)_{\tau \in S_{N-2}}$  is described by the following law.*

- Let  $H_{a,b}$  be a spherical  $p$ -spin Hamiltonian with mixture function  $\xi_{a,b}$  given by:

$$\xi_{a,b}(s) := \xi_t\left(\|v(a, b)\|^2 + r_{a,b}^2 s\right) - \xi_t\left(\|v(a, b)\|^2\right) - s \cdot \frac{r_{a,b}^2 \xi'_t\left(q_{\mathbf{m}} \cdot \left(1 + \frac{a}{\sqrt{N}}\right)\right)^2}{\xi'_t(q_{\mathbf{m}})}.$$

- Let  $V(a, b) := \xi_t\left(\|v(a, b)\|^2\right) - \|v(a, b)\|^2 \cdot \frac{\xi'_t\left(\left(1 + \frac{a}{\sqrt{N}}\right)q_{\mathbf{m}}\right)^2}{\xi'_t(q_{\mathbf{m}})} + \frac{\xi''_t(q_{\mathbf{m}})}{\gamma'(q_{\mathbf{m}})\xi'_t(q_{\mathbf{m}})} \cdot \gamma\left(\left(1 + \frac{a}{\sqrt{N}}\right)q_{\mathbf{m}}\right)^2$ .

The law of  $H_{\text{TAP}}\left(\sqrt{N} \cdot v(a, b) + r_{a,b} Q\tau\right)$  is the same as that of  $H_{a,b}(\tau) + \sqrt{N} \cdot g_{a,b} + \mathbf{E}_{\mu_{\text{TAP}}} H_{\text{TAP}}\left(\sqrt{N} \cdot v(a, b) + r_{a,b} Q\tau\right)$  where  $g_{a,b}$  is a centered Gaussian of variance  $V(a, b)$  independent of  $H_{a,b}$ .

Now,  $H_{\text{TAP}}$  is described by the collection  $(H_{a,b}, g_{a,b})_{a,b}$ . This is *not* an independent collection of random variables. The only structural properties of this collection we will use are:

- For each  $a, b \in \mathbb{R}$ , we have  $H_{a,b}$  and  $g_{a,b}$  are independent.

- For any  $a, b$ , we have  $g_{a,b} = g_{0,0} + \widehat{g}_{a,b}$ , where  $\widehat{g}_{a,b}$  is a centered Gaussian of variance  $O\left(\frac{a^4+b^4}{N^2}\right)$ .

We will first give an explicit form for  $v(a, b)$ .

**Fact 4.7.25.** *We have*

$$\sqrt{N} \cdot v(a, b) = m \cdot \left(1 + \frac{aq_m - bq_x^2}{\sqrt{N}(q_m - q_x^2)}\right) + \frac{q_m q_x}{q_m - q_x^2} x \cdot \left(\frac{b-a}{\sqrt{N}}\right).$$

**Lemma 4.7.26.** *There exists  $\varepsilon = \varepsilon(\xi) > 0$  such that for all  $t \geq 0$  and all  $|a|, |b| \leq \varepsilon N^{1/10}$ , the mixture function  $\xi_{a,b}$  defined in Corollary 4.7.24 (recall this implicitly depends on  $t$ ) is  $\varepsilon$ -strictly replica symmetric (Condition 5).*

*Proof.* We will first show  $\xi'_{a,b}(0) \leq N^{-4/5}$ . We calculate:

$$\begin{aligned} \xi'_{a,b}(0) &= r_{a,b}^2 \xi'_t(\|v(a, b)\|^2) - \frac{r_{a,b}^2 \xi'_t\left(q_m \left(1 + \frac{a}{\sqrt{N}}\right)\right)^2}{\xi'_t(q_m)} \\ &\leq \xi'_t\left(q_m + \frac{2aq_m}{\sqrt{N}} + O\left(\frac{a^2+b^2}{N}\right)\right) - \frac{\xi'_t\left(q_m \left(1 + \frac{a}{\sqrt{N}}\right)\right)^2}{\xi'_t(q_m)}. \end{aligned}$$

Here, the inequality follows from the fact that  $\|v(a, b)\|^2 = q_m \left(1 + \frac{a}{\sqrt{N}}\right)^2 + O\left(\frac{a^2+b^2}{N}\right)$ ,  $\xi'_t$  is non-decreasing, and  $r_{a,b}^2 \leq 1$ . Now, we have, using the  $O(1)$ -Lipschitzness of  $\xi'_t$ ,

$$\xi'_t(\|v(a, b)\|^2) = \xi'_t(q_m) + \xi''_t(q_m) \cdot \frac{2aq_m}{\sqrt{N}} + O\left(\frac{a^2+b^2}{N}\right)$$

and

$$\xi'_t\left(q_m \left(1 + \frac{a}{\sqrt{N}}\right)\right)^2 = \xi'_t(q_m)^2 + \frac{2aq_m}{\sqrt{N}} \cdot \xi''_t(q_m) \cdot \xi'_t(q_m) + O\left(\xi'_t(q_m) \cdot \frac{a^2+b^2}{N}\right).$$

Thus  $\xi'_{a,b}(0) = O\left(\frac{a^2+b^2}{N}\right)$ . Setting  $\varepsilon$  sufficiently small ensures  $\xi'_{a,b}(0) \leq N^{-4/5}$ . Next, we show  $(\xi_{a,b})_{\sim 1}$  satisfies Condition 4. Since  $\xi$  satisfies (SL), there exists sufficiently small  $\varepsilon = \varepsilon(\xi)$  such that  $\xi''(q) \leq \frac{1-\varepsilon}{(1-q)^2}$  for all  $q \in [0, 1)$ . Then,

$$\xi''_{a,b}(q) = \left(1 - \|v(a, b)\|^2\right)^2 \xi''_t\left(\|v(a, b)\|^2 + \left(1 - \|v(a, b)\|^2\right)q\right)$$

$$\begin{aligned}
&\leq \left(1 - \|v(a, b)\|^2\right)^2 \cdot \frac{1 - \varepsilon}{\left(1 - \|v(a, b)\|^2 - \left(1 - \|v(a, b)\|^2\right)q\right)^2} \\
&= \frac{1 - \varepsilon}{(1 - q)^2} \leq \frac{1}{(1 - q)^2} - \varepsilon.
\end{aligned}$$

Integrating twice shows

$$(\xi_{a,b})_{\sim 1}(q) + q + \log(1 - q) \leq -\frac{1}{2}\varepsilon q^2. \quad \square$$

We are now ready to bound  $\mathbf{Cov}(\mu_{H_{\text{TAP}}})$ . First, recall that for any distribution  $\mu$  over  $\mathbb{R}^N$  and any vector  $v \in \mathbb{R}^N$ , we have  $\mathbf{Cov}(\mu) \leq \mathbf{E}_{\sigma \sim \mu}(\sigma - v)(\sigma - v)^\top$ . Thus it suffices to bound

$$\begin{aligned}
&\mathbf{E}_{\sigma \sim \mu_{H_{\text{TAP}}}}(\sigma - \mathbf{m})(\sigma - \mathbf{m})^\top \\
&= \mathbf{E}_{(a,b) \sim \nu} \mathbf{E}_{\sigma \sim \mu_{a,b}} \left( \sigma - \sqrt{N} \cdot v(a, b) + \sqrt{N} \cdot v(a, b) - \mathbf{m} \right) \left( \sigma - \sqrt{N} \cdot v(a, b) + \sqrt{N} \cdot v(a, b) - \mathbf{m} \right)^\top \\
&\leq 2\mathbf{E}_{(a,b) \sim \nu} \mathbf{E}_{\sigma \sim \mu_{a,b}} \left( \sigma - \sqrt{N} \cdot v(a, b) \right) \left( \sigma - \sqrt{N} \cdot v(a, b) \right)^\top + 2\mathbf{E}_{(a,b) \sim \nu} (\sqrt{N} \cdot v(a, b) - \mathbf{m})(\sqrt{N} \cdot v(a, b) - \mathbf{m})^\top
\end{aligned} \tag{4.10}$$

We will bound the spectral norm of each of the above terms below. We employ the following statement for proving the desired bounds, proved in Section 4.8.

**Theorem 4.7.27.** *Suppose  $\varepsilon > 0$ , and  $\xi$  is  $\varepsilon$ -strictly replica symmetric (Condition 5). Then, there exist  $c = c(\varepsilon)$  and  $C = C(\varepsilon)$  such that the following hold with probability  $1 - e^{-cN^{1/5}}$ .*

(a) *We have  $\nabla^2 H_N(0) \leq (1 + \xi''(0) - \varepsilon^2/8)I_N$  and*

$$\left| \log Z_N - \frac{N\xi(1)}{2} - \frac{N\xi''(0)}{4} - \frac{\log(1 - \xi''(0))}{2} + \frac{1}{2} \log \det((1 + \xi''(0))I_N - \nabla^2 H_N(0)) \right| \leq 1.$$

(b) *The Gibbs measure satisfies  $\|\mathbf{E}_{\mu_{H_N}} \sigma \sigma^\top\|_{\text{op}} \leq C(1 + \gamma_1^2 N)$ .*

**Bounding the first term.** Observe that:

$$\mathbf{E}_{\sigma \sim \mu_{a,b}} \left( \sigma - \sqrt{N} \cdot v(a, b) \right) \left( \sigma - \sqrt{N} \cdot v(a, b) \right)^\top = r_{a,b}^2 Q_{a,b} \mathbf{E}_{\tau \sim \mu_{H_{a,b}}} \tau \tau^\top Q_{a,b}^\top.$$

Thus,

$$\left\| \mathbf{E}_{\sigma \sim \mu_{a,b}} \left( \sigma - \sqrt{N} \cdot v(a, b) \right) \left( \sigma - \sqrt{N} \cdot v(a, b) \right)^\top \right\| = r_{a,b}^2 \left\| \mathbf{E}_{\tau \sim \mu_{H_{a,b}}} \tau \tau^\top \right\| \leq C r_{a,b}^2 (1 + a^2 + b^2). \tag{4.11}$$



where the inequality follows from Lemma 4.7.26 and Theorem 4.7.27.

Our next goal is to control the fluctuations of  $(a, b)$ . By definition, if  $\widehat{Z}_{a,b}$  is the partition function of the distribution with Hamiltonian  $H_{a,b}$  defined by  $H_{a,b}(\tau) = H_{\text{TAP}}(\sqrt{N} \cdot v(a, b) + r_{a,b} Q\tau)$ , we have

$$v(a, b) \propto \exp \left( \log \widehat{Z}_{a,b} + (N-4) \log r_{a,b} + \mathbf{E}_{\mu_{\text{TAP}}} H_{\text{TAP}}(\sqrt{N} \cdot v(a, b)) + \sqrt{N} g_{a,b} \right). \quad (4.12)$$

We will show that  $(a, b) \sim \nu$ , conditioned on  $|a|, |b| \leq \varepsilon N^{1/10}$  for  $\varepsilon$  as in Lemma 4.7.26, is subgaussian with variance  $O(1)$ . We will also show in Lemma 4.7.34 that  $\nu$  places very little mass outside the set  $|a|, |b| \leq \varepsilon N^{1/10}$ .

**Lemma 4.7.28.** *Let  $\varepsilon$  be as in Lemma 4.7.26. On an event with probability  $1 - e^{-cN^{1/5}}$ , the following holds. The density of  $(a, b)$  under  $\nu$ , conditioned on  $|a|, |b| \leq \varepsilon N^{1/10}$ , is given by*

$$v(a, b) \propto \exp \left( N \widehat{E}_{a,b} + \sqrt{N} g_{a,b} + \text{Error}_{a,b}^{(1)} + \text{Error}_{a,b}^{(2)} + \Delta_{a,b} \right),$$

where  $|\Delta_{a,b}| \leq 1$  and

$$\begin{aligned} \widehat{E}_{a,b} &= \frac{\xi_{a,b}(1)}{2} + \log r_{a,b} + \frac{1}{N} \mathbf{E}_{\mu_{\text{TAP}}} H_{\text{TAP}}(\sqrt{N} \cdot v(a, b)) - \frac{\xi_t(1)}{2} \\ &= \frac{1}{2} \left( \log r_{a,b}^2 - \xi_t(\|v(a, b)\|^2) - r_{a,b}^2 \cdot \frac{\xi'_t \left( q_m \left( 1 + \frac{a}{\sqrt{N}} \right) \right)^2}{\xi'_t(q_m)} \right) \\ &\quad - \gamma(q_x) \cdot \frac{\xi'_t \left( q_m \left( 1 + \frac{a}{\sqrt{N}} \right) \right)}{\xi'_t(q_m)} \cdot \left( \left( 1 + \frac{b}{\sqrt{N}} \right) - q_m \cdot \frac{\xi''_t(q_m)}{\gamma'(q_m)} \cdot \left( 1 + \frac{a}{\sqrt{N}} \right) \right) \\ &\quad + \xi_t \left( q_x \left( 1 + \frac{b}{\sqrt{N}} \right) \right) + \frac{\gamma \left( q_m \left( 1 + \frac{a}{\sqrt{N}} \right) \right)}{\gamma'(q_m)} \cdot \left( (1 - q_m) \xi''_t(q_m) + \frac{1}{1 - q_m} \right). \end{aligned}$$

for the error terms

$$\text{Error}_{a,b}^{(1)} = \left( \log \widehat{Z}_{a,b} - \frac{N \xi_{a,b}(1)}{2} - \frac{N \xi''_{a,b}(0)}{4} + \frac{1}{2} \log \det \left( (1 + \xi''_{a,b}(0)) \text{Id} - \nabla^2 H_{a,b}(0) \right) \right) - 4 \log r_{a,b}$$

and

$$\text{Error}_{a,b}^{(2)} = \frac{N \xi''_{a,b}(0)}{4} - \frac{1}{2} \log \det \left( (1 + \xi''_{a,b}(0)) \text{Id} - \nabla^2 H_{a,b}(0) \right)$$

The above follows by expanding out all the terms in the expression (4.12) for the density of  $\nu$  and evaluating the term  $\log \widehat{Z}_{a,b}$  using Theorem 4.7.27, which applies because the models  $\xi_{a,b}$  for  $|a|, |b| \leq \varepsilon N^{1/10}$  are  $\varepsilon$ -strictly replica symmetric by Lemma 4.7.26. Because the conclusion of Theorem 4.7.27 holds with probability  $1 - e^{-cN^{1/5}}$ , we may evaluate  $\log \widehat{Z}_{a,b}$  over a  $1/\text{poly}(N)$ -net of such  $(a, b)$  via a union bound, and then infer the estimate for all such  $a, b$  by a standard continuity argument.

**Lemma 4.7.29.**  $\nabla \widehat{E}_{a,b} \Big|_{(a,b)=(0,0)} = 0$ .

**Lemma 4.7.30.** *There exist constants  $\eta, \varepsilon > 0$  such that for all  $|a|, |b| \leq \varepsilon \sqrt{N}$ ,  $N \nabla^2 \widehat{E}_{a,b} \leq -\eta \text{Id}$ .*

The above follow from routine calculations, which we defer to Section 4.10.

**Lemma 4.7.31.** *For every constant  $\iota > 0$ , there is a constant  $c$  such that with probability  $1 - e^{-cN}$ , for all  $a, b$ , we have  $|g_{a,b} - g_{0,0}| \leq \iota \frac{a^2 + b^2}{\sqrt{N}}$ .*

**Lemma 4.7.32.** *With probability  $1 - e^{-cN^{1/5}}$ ,  $|\text{Error}_{a,b}^{(1)}| = O(1)$  uniformly for all  $|a|, |b| < \varepsilon N^{1/10}$ , for  $\varepsilon$  as in Lemma 4.7.26.*

*Proof.* We shall show this very high probability bound for a fixed  $a, b$ . Constructing a net over the relevant  $a, b$  and performing a union bound over this net allows us to extend this to a uniform bound for all  $a, b$ ; we omit the details. We may write the error term as

$$\begin{aligned} \text{Error}_{a,b}^{(1)} = \log \widehat{Z}_{a,b} - \frac{N \xi_{a,b}(1)}{2} - \frac{\log(1 - \xi_{a,b}''(0))}{2} \\ - \frac{N \xi_{a,b}''(0)}{4} + \frac{1}{2} \log \det \left( (1 + \xi_{a,b}''(0)) \text{Id} - \nabla^2 H_{a,b}(0) \right) + O(1). \end{aligned}$$

Due to the bound on  $a$  and  $b$ , the above is  $O(1)$  with very high probability by Theorem 4.7.27(a). The desideratum follows.  $\square$

**Lemma 4.7.33.** *For any sufficiently small  $\iota > 0$ , with probability at least  $1 - e^{-cN}$ ,  $\left| \text{Error}_{a,b}^{(2)} - \text{Error}_{0,0}^{(2)} \right| = O(1)$  for all  $|a|, |b| < \iota N^{1/4}$ .*

We relegate the proof of the above to the appendix Section 4.10. The idea of the proof is that the Hessian  $\nabla^2 H_{a,b}(0)$  does not deviate too much for small variations in  $a, b$  – the first order terms in the deviation end up being cancelled by the  $\xi_{a,b}''(0)/2$  term, while the second order terms are  $O(1)$ .

**Lemma 4.7.34.** *Let  $\mathcal{E}_t$  be as in Lemma 4.7.20. With probability  $1 - e^{-cN^{1/5}}$ , either  $\mathcal{E}_t$  does not hold, or the following holds. For  $\varepsilon$  as in Lemma 4.7.26,*

$$\Pr_{(a,b) \sim \nu} [|a| \leq \varepsilon N^{1/10} \text{ and } |b| \leq \varepsilon N^{1/10}] \geq 1 - e^{-cN^{1/5}}.$$

*Proof.* On the event  $\mathcal{E}_t$ , we have

$$\Pr_{(a,b) \sim \nu} [|a| > \iota N^{1/2} \text{ or } |b| > \iota N^{1/2}] \leq e^{-cN}.$$

Thus, let

$$T = \{(a, b) \in \mathbb{R}^2 : |a| \in [\varepsilon N^{1/10}, \iota N^{1/2}] \text{ or } |b| \in [\varepsilon N^{1/10}, \iota N^{1/2}]\}.$$

It suffices to show that  $\Pr_{(a,b) \sim \nu} [(a, b) \in T] \leq e^{-cN^{1/5}}$  with probability  $1 - e^{-cN^{1/5}}$ . Recall the density of  $(a, b) \sim \nu$  is given by (4.12), and that  $\mathbf{E}\widehat{Z}_{a,b} = e^{N\xi_{a,b}(1)/2}$ . Thus, for  $\widehat{\mathbf{E}}$  denoting expectation with respect to the  $\widehat{Z}_{a,b}$  alone,

$$\begin{aligned} & \widehat{\mathbf{E}} \int_T \exp \left( \log \widehat{Z}_{a,b} + (N-4) \log r_{a,b} + \mathbf{E}_{\mu_{\text{TAP}}} H_{\text{TAP}}(\sqrt{N} \cdot v(a, b)) + \sqrt{N} g_{a,b} \right) d(a, b) \\ &= \int_T \exp \left( N\widehat{E}_{a,b} + \frac{N\xi_t(1)}{2} + \sqrt{N} g_{a,b} - 4 \log r_{a,b} \right) d(a, b). \end{aligned}$$

On the event in Lemma 4.7.31, we have, for any constant  $\iota > 0$ ,

$$\sqrt{N} g_{a,b} \leq \sqrt{N} g_{0,0} + \iota(a^2 + b^2).$$

By Lemma 4.7.30,

$$N\widehat{E}_{a,b} \leq N\widehat{E}_{0,0} - \eta(a^2 + b^2).$$

Combining shows that

$$N\widehat{E}_{a,b} + \sqrt{N} g_{a,b} - 4 \log r_{a,b} \leq N\widehat{E}_{0,0} + \sqrt{N} g_{0,0} - \frac{\eta}{2}(a^2 + b^2) + O(1).$$

Combining shows

$$\begin{aligned} & \widehat{\mathbf{E}} \int_T \exp \left( \log \widehat{Z}_{a,b} + (N-4) \log r_{a,b} + \mathbf{E}_{\mu_{\text{TAP}}} H_{\text{TAP}}(\sqrt{N} \cdot v(a, b)) + \sqrt{N} g_{a,b} \right) d(a, b) \\ & \leq e^{-cN^{1/5}} \exp \left( N\widehat{E}_{0,0} + \frac{N\xi_t(1)}{2} + \sqrt{N} g_{0,0} \right) \end{aligned}$$

and therefore with probability  $1 - e^{-cN^{1/5}/2}$  over the  $\widehat{Z}_{a,b}$ ,

$$\int_T \exp \left( \log \widehat{Z}_{a,b} + (N-4) \log r_{a,b} + \mathbf{E}_{\mu_{\text{TAP}}} H_{\text{TAP}}(\sqrt{N} \cdot v(a, b)) + \sqrt{N} g_{a,b} \right) d(a, b)$$

$$\leq e^{-cN^{1/5}/2} \exp\left(N\widehat{E}_{0,0} + \frac{N\xi_t(1)}{2} + \sqrt{N}g_{0,0}\right) \quad (4.13)$$

On the other hand, Lemma 4.7.28 implies that with probability  $1 - e^{-cN^{1/5}}$ ,

$$\begin{aligned} & \log \widehat{Z}_{0,0} + (N-4) \log r_{0,0} + \mathbf{E}_{\mu_{\text{TAP}}} H_{\text{TAP}}(\sqrt{N} \cdot v(0,0)) + \sqrt{N}g_{0,0} \\ &= N\widehat{E}_{0,0} + \frac{N\xi_t(1)}{2} + \sqrt{N}g_{0,0} + \text{Error}_{0,0}^{(1)} + \text{Error}_{0,0}^{(2)} + O(1), \end{aligned}$$

and Lemma 4.7.32 implies  $|\text{Error}_{0,0}^{(1)}| = O(1)$  with probability  $1 - e^{-cN^{1/5}}$ . Furthermore, Lemma 4.8.3 below implies that  $|\text{Error}_{0,0}^{(2)}| \leq N^{1/10}$  with probability  $1 - e^{-cN^{1/5}}$ . Thus

$$\begin{aligned} & \log \widehat{Z}_{0,0} + (N-4) \log r_{0,0} + \mathbf{E}_{\mu_{\text{TAP}}} H_{\text{TAP}}(\sqrt{N} \cdot v(0,0)) + \sqrt{N}g_{0,0} \\ & \geq N\widehat{E}_{0,0} + \frac{N\xi_t(1)}{2} + \sqrt{N}g_{0,0} - 2N^{1/10}, \end{aligned}$$

and standard continuity arguments imply that for  $T' = \{(a, b) : |a|, |b| \leq N^{-10}\}$ ,

$$\begin{aligned} & \int_{T'} \exp\left(\log \widehat{Z}_{a,b} + (N-4) \log r_{a,b} + \mathbf{E}_{\mu_{\text{TAP}}} H_{\text{TAP}}(\sqrt{N} \cdot v(a,b)) + \sqrt{N}g_{a,b}\right) d(a,b) \\ & \geq e^{-3N^{1/10}} \exp\left(N\widehat{E}_{0,0} + \frac{N\xi_t(1)}{2} + \sqrt{N}g_{0,0}\right). \end{aligned}$$

Comparing with (4.13) implies the conclusion, after adjusting  $c$ .  $\square$

**Lemma 4.7.35.** *With probability  $1 - e^{-cN^{1/5}}$ , either  $\mathcal{E}_t$  does not hold or the following holds. There exists a random variable  $X$  over  $\mathbb{R}^2$  (coupled with  $v$ ) such that the following holds for  $(a, b) \sim v$ .*

(a) *With probability at least  $1 - e^{-cN^{1/5}}$ ,  $X = (a, b)$ .*

(b)  *$X$  has mean  $O(1)$  and is  $O(1)$ -subgaussian.*

*Proof.* This is an immediate corollary of Lemmas 4.7.29 to 4.7.34, setting  $X$  to be the random variable that is equal to  $(a, b)$  if  $|a|, |b| < \varepsilon N^{1/10}$ , and 0 otherwise.  $\square$

We are now finally prepared to bound  $\text{Cov}(\mu)$ .

**Lemma 4.7.21.** *Let  $x \in S_N$ , let  $\mathcal{S}_t$  be as in Lemma 4.7.20, and let  $m \in \mathcal{S}_t$ . Then for an absolute constant  $K > 0$ ,*

$$\Pr_{H_{\text{TAP}} \sim \mu_{\text{TAP}, x, m}} [\|\text{Cov}(\mu_{H_{\text{TAP}}})\| \geq K \wedge \mathcal{E}_t] \leq e^{-cN^{1/5}}.$$

*Proof.* Note that  $|a|, |b| \leq 2\sqrt{N}$  almost surely. Let  $X$  be as in Lemma 4.7.35. This lemma implies that with probability at least  $1 - e^{-cN^{1/5}}$  over the randomness of the Hamiltonian,

$$\begin{aligned}\mathbf{E}_{(a,b) \sim \nu} [a^2 + b^2] &= \mathbf{E}[\|X\|^2] + \mathbf{E}_{(a,b) \sim \nu} [\mathbf{1}[X \neq (a, b)](a^2 + b^2)] \\ &\leq O(1) + \Pr(X \neq (a, b)) \cdot 8N = O(1).\end{aligned}$$

Thus, by plugging in (4.11) along with this observation into (4.10), we get that the following holds with probability at least  $1 - e^{-cN^{1/5}}$

$$\begin{aligned}\|\mathbf{Cov}(\mu)\| &\leq 2C\mathbf{E}_{(a,b) \sim \nu}(1 + a^2 + b^2) + \|2\mathbf{E}_{(a,b) \sim \nu}(\sqrt{N} \cdot v(a, b) - \mathbf{m})(\sqrt{N} \cdot v(a, b) - \mathbf{m})^\top\| \\ &\leq O(1) + 2\mathbf{E}_{(a,b) \sim \nu} \|\sqrt{N} \cdot v(a, b) - \mathbf{m}\|^2 \\ &\leq O(1) + 2\mathbf{E}_{(a,b) \sim \nu} [O(a^2 + b^2)] \\ &\leq O(1).\end{aligned}$$

□

## 4.8 High-probability covariance bound of replica symmetric spherical spin glass

In this section we prove the main technical input to the proofs in Section 4.7. This takes the form of a high-probability bound on the partition function and covariance matrix (in fact, second moment matrix) of a spherical spin glass in the replica symmetric phase.

In this section, we let  $H_N$  be defined as in (4.8), with a linear term corresponding to an external field:

$$H_N(\sigma) := \sum_{p \geq 1} \frac{\gamma_p}{N^{(p-1)/2}} \sum_{i_1, \dots, i_p=1}^N \mathbf{g}_{i_1, \dots, i_p} \sigma_{i_1} \cdots \sigma_{i_p}.$$

We recall  $\xi_{\sim 1}(q) = \sum_{p \geq 2} \gamma_p^2 q^p$  denotes the part of  $\xi$  without the linear term, and let

$$\xi_{\sim 2}(q) = \gamma_1^2 q + \sum_{p \geq 3} \gamma_p^2 q^p$$

denote the part of  $\xi$  excluding the degree 2 term.

The results in this section hold under the following condition, which we restate for reference.

**Condition 5** ( $\varepsilon$ -strict replica symmetry). *We say  $\xi$  is  $\varepsilon$ -strictly replica symmetric if  $\gamma_1^2 \leq N^{-4/5}$  and  $\xi_{\sim 1}$  satisfies Condition 4.*

Throughout this section, we treat  $\varepsilon > 0$  as a constant and let  $O_\varepsilon(1)$  denote a quantity bounded depending on  $\varepsilon$ .

**Theorem 4.7.27.** *Suppose  $\varepsilon > 0$ , and  $\xi$  is  $\varepsilon$ -strictly replica symmetric (Condition 5). Then, there exist  $c = c(\varepsilon)$  and  $C = C(\varepsilon)$  such that the following hold with probability  $1 - e^{-cN^{1/5}}$ .*

(a) *We have  $\nabla^2 H_N(0) \leq (1 + \xi''(0) - \varepsilon^2/8)I_N$  and*

$$\left| \log Z_N - \frac{N\xi(1)}{2} - \frac{N\xi''(0)}{4} - \frac{\log(1 - \xi''(0))}{2} + \frac{1}{2} \log \det ((1 + \xi''(0))I_N - \nabla^2 H_N(0)) \right| \leq 1.$$

(b) *The Gibbs measure satisfies  $\|\mathbf{E}_{\mu_{H_N}} \sigma \sigma^\top\|_{\text{op}} \leq C(1 + \gamma_1^2 N)$ .*

In the below proofs, we allow the constants  $c$  and  $C$  to change from line to line, but they will always be uniform in  $\varepsilon$ . We always set  $C$  sufficiently large depending on  $\varepsilon$ , and then  $c$  sufficiently small depending on  $\varepsilon, C$ .

Theorem 4.7.27 will be proved through the following pair of propositions. We introduce the degree-2 Hamiltonian

$$H_{N,2}(\sigma) := \frac{\gamma_2}{N^{1/2}} \sum_{i_1, i_2=1}^N \mathbf{g}_{i_1, i_2} \sigma_{i_1} \sigma_{i_2} = \frac{1}{2} \langle \nabla^2 H_N(0) \sigma, \sigma \rangle. \quad (4.14)$$

Similarly let  $H_{N,\sim 2}(\sigma) = H_N(\sigma) - H_{N,2}(\sigma)$  be the non degree-2 part of  $H_N(\sigma)$ . Define the degree-2 Gibbs measure and partition function by

$$d\mu_{H_{N,2}}(\sigma) = \frac{\exp(H_{N,2}(\sigma))}{Z_{N,2}} d\rho(\sigma), \quad Z_{N,2} = \int_{S_N} \exp(H_{N,2}(\sigma)) d\rho(\sigma).$$

Throughout this section, we will let  $\mathbf{E}$  denote expectation with respect to the disorder coefficients  $\mathbf{g}_{i_1, \dots, i_p}$ , while  $\langle \cdot \rangle$  denotes averaging with respect to  $\sigma \sim \mu_{H_N}$  (or several i.i.d. samples  $\sigma^1, \sigma^2, \dots$  from this measure). Similarly, let  $\langle \cdot \rangle_2$  denote Gibbs average with respect to  $\mu_{H_{N,2}}$ .

Note that  $\nabla^2 H_N(0)$  depends on  $H_N$  only through  $H_{N,2}$ .

**Proposition 4.8.1** (Concentration of degree-2 partition function; proved in Section 4.8).

*With probability  $1 - e^{-cN}$  over  $H_{N,2}$ , we have  $\nabla^2 H_N(0) \leq (1 + \xi''(0) - \varepsilon^2/8)I_N$  and*

$$\left| \log Z_{N,2} - \frac{N\xi''(0)}{2} - \frac{\log(1 - \xi''(0))}{2} + \frac{1}{2} \log \det ((1 + \xi''(0))I - \nabla^2 H_N(0)) \right| \leq 1/2.$$

The following is proved in Section 4.8.

**Proposition 4.8.2.** *There is a  $H_{N,2}$ -measurable event with probability  $1 - e^{-cN^{1/5}}$  on which the following holds with probability  $1 - e^{-cN^{1/5}}$  over  $H_{N,\sim 2}$ .*

1. The partition functions  $Z_N, Z_{N,2}$  satisfy

$$\left| \log \frac{Z_N}{Z_{N,2}} - \frac{N\xi_{\sim 2}(1)}{2} \right| \leq 1/2.$$

2. The Gibbs measure satisfies  $\|\langle \sigma \sigma^\top \rangle\|_{\text{op}} \leq C(1 + \gamma_1^2 N)$ .

*Proof of Theorem 4.7.27.* Immediate from Propositions 4.8.1 and 4.8.2, since  $\xi(1) = \xi_{\sim 2}(1) + \frac{1}{2}\xi''(0)$ .  $\square$

We also show the following concentration of the log determinant in Theorem 4.7.27.

**Lemma 4.8.3** (Proved in Section 4.8). *There exists a  $H_{N,2}$ -measurable random variable  $X$  that the following holds.*

1. With probability  $1 - e^{-cN}$ ,  $X = \log \det((1 + \xi''(0))I - \nabla^2 H_N(0)) - N\xi''(0)/2$ .
2.  $X$  has mean  $O_\varepsilon(1)$  and is  $O_\varepsilon(1)$ -subgaussian.

This implies the quantitative contiguity between the planted and null models, which we restate below for convenience.

**Proposition 4.7.15** (Quantitative contiguity). *Under Condition 4, there exists  $c = c(\varepsilon) > 0$  such that for any event  $\mathcal{E}$ , if  $\mu_{\text{pl}}(\mathcal{E}) = p$ , then  $\mu_{\text{null}}(\mathcal{E}) \leq e^{-cN^{1/5}} + e^{\frac{1}{c}\sqrt{\log \frac{2}{p}}} p$ .*

*Proof.* Let  $\mathcal{E}_{\text{good}}$  be intersection of the event in Theorem 4.7.27, the event

$$X = \log \det((1 + \xi''(0))I - \nabla^2 H_N(0)) - N\xi''(0)/2$$

from Lemma 4.8.3, and the event  $X \leq t$ , for some  $t > 0$  to be determined. Then, after adjusting  $c = c(\varepsilon)$  as necessary,

$$\mu_{\text{null}}(\overline{\mathcal{E}_{\text{good}}}) \leq e^{-cN^{1/5}} + \mathbb{P}(X > t) \leq e^{-cN^{1/5}} + e^{-c(t - \frac{1}{c})_+^2}.$$

Note that  $\log \mathbb{E} Z_N = N\xi(1)/2$ , while on the event  $\mathcal{E}_{\text{good}}$ ,

$$\log Z_N = \frac{N\xi(1) - X + O_\varepsilon(1)}{2} \geq \frac{N\xi(1) - t - \frac{1}{c}}{2}.$$

Thus  $\frac{\mathbf{E}Z_N}{Z_N} \leq e^{\frac{1}{2}(t+\frac{1}{c})}$ . So,

$$\begin{aligned}\mu_{\text{null}}(\mathcal{E}) &\leq \mu_{\text{null}}(\overline{\mathcal{E}_{\text{good}}}) + \int \frac{\mathbf{E}Z_N}{Z_N} \mathbf{1}[H_N \in \mathcal{E} \cap \mathcal{E}_{\text{good}}] d\mu_{\text{pl}}(H_N) \\ &\leq e^{-cN^{1/5}} + e^{-c(t-\frac{1}{c})_+^2} + e^{\frac{1}{2}(t+\frac{1}{c})} p.\end{aligned}$$

We then take  $t = \frac{1}{c} + \sqrt{\frac{1}{c} \log \frac{1}{p}}$ , so that this is bounded by

$$e^{-cN^{1/5}} + \left(1 + e^{\frac{1}{c} + \sqrt{\frac{1}{c} \log \frac{1}{p}}}\right) p.$$

Further adjusting  $c$  proves the desired bound.  $\square$

## Concentration of degree-2 partition function

We write  $A = \frac{1}{2} \nabla^2 H_N(0) = \frac{\sqrt{\xi''(0)}}{2} M$ . It is straightforward to check that  $M$  is distributed as a sample from  $\text{GOE}(N)$ .

**Fact 4.8.4.** *We have  $\xi''(0) \leq 1 - \varepsilon$ .*

*Proof.* Writing (4.7) as

$$\xi_{\sim 1}(q) + q + \log(1 - q) \leq -\varepsilon q^2/2$$

and Taylor expanding around  $q = 0$  implies the result.  $\square$

**Fact 4.8.5.** *With probability  $1 - e^{-cN}$ ,  $\nabla^2 H_N(0) \leq (1 + \xi''(0) - \varepsilon^2/8)I$ .*

*Proof.* With probability  $1 - e^{-cN}$ , we have  $\lambda_{\max}(M) \leq 2 + \varepsilon^2/8$ . Then,

$$\begin{aligned}\lambda_{\max}((1 + \xi''(0) - \varepsilon^2/8)I - \nabla^2 H_N(0)) &\geq 1 + \xi''(0) - \varepsilon^2/8 - \sqrt{\xi''(0)}(2 + \varepsilon^2/8) \\ &= (1 - \sqrt{\xi''(0)})^2 - \varepsilon^2(1 + \sqrt{\xi''(0)})/8 \\ &> \varepsilon^2/4 - \varepsilon^2/4 = 0\end{aligned}$$

by Fact 4.8.4.  $\square$

For  $\gamma \in (\lambda_{\max}(A), +\infty)$ , define

$$G(\gamma) = \gamma - \frac{1}{2N} \log \det(\gamma I - A). \quad (4.15)$$



Note that

$$G'(\gamma) = 1 - \frac{1}{2N} \operatorname{Tr}(\gamma I - A)^{-1}$$

is continuous and increasing, with  $\lim_{\gamma \downarrow \lambda_{\max}(A)} G'(\gamma) = -\infty$  and  $\lim_{\gamma \uparrow +\infty} G'(\gamma) = 1$ . Thus  $G'$  has a unique root  $\gamma_*$  in  $(\lambda_{\max}(A), +\infty)$ . The following lemma is a consequence of [HMP24, Lemma 7.3], which is proved by an analysis of a Laplace transform of the free energy also used in [BL16].

**Lemma 4.8.6.** *With probability  $1 - e^{-cN}$  over  $H_{N,2}$ ,*

$$Z_{N,2} = (1 + O(N^{-c})) \sqrt{\frac{2}{G''(\gamma_*)}} (2e)^{-N/2} \exp(NG(\gamma_*)). \quad (4.16)$$

*Proof.* Recalling (4.14), we have

$$H_{N,2}(\sigma) = \frac{\sqrt{\xi''(0)}}{2} \langle M\sigma, \sigma \rangle,$$

and Fact 4.8.4 implies the factor  $\sqrt{\xi''(0)}/2$  is bounded away from  $1/2$ . Then [HMP24, Lemma 7.3] (with  $u = 0$ ) implies the result.  $\square$

Define  $\gamma_0 = (1 + \xi''(0))/2$ . The next lemma shows that, although the variable  $\gamma_*$  in (4.16) is random, we may approximate it deterministically by  $\gamma_0$ .

**Lemma 4.8.7.** *For sufficiently large  $C$  depending on  $\varepsilon$ , and sufficiently small  $c$  depending on  $\varepsilon, C$ , with probability  $1 - e^{-cN}$  the following holds for all  $\gamma \in [\gamma_0 - N^{-1/2}, \gamma_0 + N^{-1/2}]$ .*

1.  $|G'(\gamma_0)| \leq 1/(C\sqrt{N})$ .
2.  $|\frac{G''(\gamma)}{2/(1-\xi''(0))} - 1| \leq N^{-1/3}$ .

*Proof.* Let  $d\rho_{\text{sc}}(x) = \frac{1}{2\pi} \mathbf{1}[|x| \leq 2] \sqrt{4 - x^2} \, dx$  denote Wigner's semicircle law, and

$$f_1(x) = 1 - \frac{1}{1 + \xi''(0) - 2\sqrt{\xi''(0)}x}, \quad f_2(x) = \frac{2}{(1 + \xi''(0) - 2\sqrt{\xi''(0)}x)^2}.$$

For  $k \in [2]$ , let

$$L_k = \int f_k(x) \, d\rho_{\text{sc}}(x).$$

We will show that with probability  $1 - e^{-cN}$ , for each  $k \in [2]$ ,

$$|G^{(k)}(\gamma_0) - L_k| \leq \frac{1}{C\sqrt{N}}. \quad (4.17)$$

Recall that  $M \sim \text{GOE}(N)$ . For  $f : \mathbb{R} \rightarrow \mathbb{R}$ , define the spectral trace

$$\text{Tr } f(M) = \sum_{i=1}^N f(\lambda_i(M)).$$

Note that  $G^{(k)}(\gamma_0) = N^{-1} \cdot \text{Tr } f_k(M)$ . Define

$$\tilde{f}_k(x) = f_k(\min(x, 2 + \varepsilon^2/8)).$$

By the proof of Fact 4.8.5,  $1 + \xi''(0) - 2\sqrt{\xi''(0)}x \geq \varepsilon^2/8$  for  $x \leq 2 + \varepsilon^2/8$ , so  $\tilde{f}_k$  is  $O_\varepsilon(1)$ -Lipschitz. Moreover,  $\lambda_{\max}(M) \leq 2 + \varepsilon^2/8$  with probability  $1 - e^{-cN}$ , and on this event  $\text{Tr } f_k(M) = \text{Tr } \tilde{f}_k(M)$ .

By [GZ00, Lemma 1.2(b)], if we write  $M_{i,i} = \sqrt{2/N}Z_{i,i}$ ,  $M_{i,j} = \sqrt{1/N}Z_{i,j}$ , then  $\text{Tr } \tilde{f}_k(M)$  is a  $O_\varepsilon(1)$ -Lipschitz function of the standard gaussians  $(Z_{i,j})_{1 \leq i \leq j \leq N}$ . Thus  $\text{Tr } \tilde{f}_k(M)$  is  $O_\varepsilon(1)$ -subgaussian, i.e.

$$\mathbb{P}(|\text{Tr } \tilde{f}_k(M) - \mathbf{E} \text{Tr } \tilde{f}_k(M)| \geq t) \leq 2e^{-t^2/C}$$

for some  $C = O_\varepsilon(1)$ . By [BY05, Theorem 1.1],

$$\text{Tr } \tilde{f}_k(M) - NL_k$$

converges in distribution to a gaussian with mean and variance  $O_\varepsilon(1)$ . Combined with subgaussianity of  $\text{Tr } \tilde{f}_k(M)$ , this implies

$$|\mathbf{E} \text{Tr } \tilde{f}_k(M) - NL_k| = O_\varepsilon(1).$$

It follows that (after possibly increasing  $C = O_\varepsilon(1)$ ),

$$\mathbb{P}(|\text{Tr } \tilde{f}_k(M) - NL_k| \geq t) \leq 2e^{-(t-C)_+^2/C}.$$

Thus

$$\begin{aligned} \mathbb{P}(|G^{(k)} - L_k| \geq t) &\leq \mathbb{P}(\text{Tr } f_k(M) \neq \text{Tr } \tilde{f}_k(M)) + \mathbb{P}(|\text{Tr } \tilde{f}_k(M) - NL_k| \geq Nt) \\ &\leq e^{-cN} + 2e^{-(Nt-C)_+^2/C}. \end{aligned}$$

Plugging in  $t = 1/(C\sqrt{N})$  proves (4.17). Next, direct calculations show  $L_1 = 0$ ,  $L_2 = \frac{2}{1-\xi''(0)}$ . The former directly implies conclusion (1), and the latter implies

$$\left| G''(\gamma_0) - \frac{2}{1-\xi''(0)} \right| \leq \frac{1}{C\sqrt{N}}.$$

Moreover, on the probability  $1 - e^{-cN}$  event that  $\lambda_{\max}(M) \leq 2 + \varepsilon^2/8$ ,  $G^{(3)}(\gamma) = O_\varepsilon(1)$  for all  $\gamma \in [\gamma_0 - N^{-1/2}, \gamma_0 + N^{-1/2}]$ . This implies the conclusion (2).  $\square$

Lemma 4.8.3 is proved by the same method, and we present the proof here.

*Proof of Lemma 4.8.3.* Let

$$f_0(x) = \log\left(1 + \xi''(0) - \sqrt{\xi''(0)}x\right).$$

An elementary calculation shows that

$$L_0 := \int f_0(x) \, d\rho_{\text{sc}}(X) = \xi''(0)/2.$$

Proceeding as in the above proof, we have

$$\log \det \left( (1 + \xi''(0))I - \sqrt{\xi''(0)} \nabla^2 H_N(0) \right) = \text{Tr } f_0(M).$$

If we take  $\tilde{f}_0(x) = f_0(\min(x, 2 + \varepsilon^2/8))$ , then  $\text{Tr } f_0(M) = \text{Tr } \tilde{f}_0(M)$  with probability  $1 - e^{-cN}$ . The same proof shows  $\text{Tr } \tilde{f}_0(M)$  is  $O_\varepsilon(1)$ -subgaussian, and

$$|\mathbf{E} \text{Tr } \tilde{f}_0(M) - NL_0| = O_\varepsilon(1).$$

Thus we may take  $X = \text{Tr } \tilde{f}_0(M) - NL_0 = \text{Tr } \tilde{f}_0(M) - N\xi''(0)/2$ . □

*Proof of Proposition 4.8.1.* The assertion  $\nabla^2 H_N(0) \leq (1 + \xi''(0) - \varepsilon^2/8)I_N$  is proved in Fact 4.8.5. Suppose the events in Lemmas 4.8.6 and 4.8.7 occur. Since  $\gamma_*$  is the solution to  $G'(\gamma_*) = 0$ , we have

$$|\gamma_0 - \gamma_*| \leq |G'(\gamma_0)| \cdot \frac{1 - \xi''(0)}{2(1 - N^{-1/3})} \leq \frac{1}{C\sqrt{N}}.$$

So,

$$N|G(\gamma_0) - G(\gamma_*)| \leq \frac{N}{2} |\gamma_0 - \gamma_*|^2 \sup_{\gamma \in [\gamma_0 - N^{-1/2}, \gamma_0 + N^{-1/2}]} G''(\gamma) \leq \frac{1}{C^2} \cdot \frac{2(1 + N^{-1/3})}{1 - \xi''(0)} \leq \frac{3}{C^2\varepsilon}.$$

Moreover,  $\xi''(\gamma_0)/\xi''(\gamma_*) = 1 + O(N^{-1/3})$ . Combining with Lemma 4.8.6 shows that, for some  $\Delta$  satisfying  $|\Delta| \leq \frac{3}{C^2\varepsilon}$ ,

$$\begin{aligned} Z_{N,2} &= (1 + O(N^{-c}))e^\Delta \sqrt{\frac{2}{G''(\gamma_0)}} (2e)^{-N/2} \exp(NG(\gamma_0)) \\ &= (1 + O(N^{-c}))e^\Delta \sqrt{1 - \xi''(0)} (2e)^{-N/2} \exp(N\gamma_0) \det \left( \gamma_0 I - \frac{1}{2} \nabla^2 H_N(0) \right)^{-1/2} \\ &= (1 + O(N^{-c}))e^\Delta \sqrt{1 - \xi''(0)} \exp(N\xi''(0)/2) \det \left( (1 + \xi''(0))I - \nabla^2 H_N(0) \right)^{-1/2}. \end{aligned}$$

Taking a logarithm and setting  $C$  sufficiently large concludes the proof. □

Finally, the following concentration estimates for samples from  $\mu_{H_{N,2}}$  will be useful in the sequel. This is proved similarly to [HMP24, Lemma 7.5], and we defer the proof to Section 4.10.

Let  $v_1, \dots, v_N$  denote the (unit) eigenvectors of  $\nabla^2 H_N(0)$ . These are well defined on the almost sure event that all eigenvalues of  $\nabla^2 H_N(0)$  have multiplicity 1.

**Proposition 4.8.8.** *With probability  $1 - e^{-cN}$  over  $H_{N,2}$ , the following holds. Let  $\sigma, \sigma^1, \sigma^2 \sim \mu_{H_{N,2}}$ , and let  $W = \langle \sigma, v_i \rangle$  for any  $i \in [N]$ , or  $W = \langle \sigma^1, \sigma^2 \rangle / \sqrt{N}$ . Then:*

1. *For any  $0 \leq t \leq N^{1/5}$ ,  $\mathbb{P}(|W| \geq t) \leq 3e^{-ct^2}$ .*
2. *For any  $k \in 2\mathbb{N}$ , there exists  $C_k > 0$  independent of  $N$  such that  $\langle W^k \rangle_2 \leq C_k$ .*

*In particular, part (2) implies  $\|\langle \sigma \sigma^\top \rangle_2\|_{\text{op}} \leq C$ .*

## Conditional positive probability bounds for non degree-2 part

In this subsection, we prove the following propositions, which establish a weaker version of Proposition 4.8.2 with positive instead of high probability.

**Proposition 4.8.9.** *There is a  $H_{N,2}$ -measurable event with probability  $1 - e^{-cN^{1/5}}$  on which, with probability  $1 - N^{-1/15}$  over  $H_{N,\sim 2}$ ,*

$$\left| \log \frac{Z_N}{Z_{N,2}} - \frac{N \xi_{\sim 2}(1)}{2} \right| = O(N^{-1/15}).$$

**Proposition 4.8.10.** *There is a  $H_{N,2}$ -measurable event with probability  $1 - e^{-cN^{1/5}}$  on which, with probability  $1/2$  over  $H_{N,\sim 2}$ ,*

$$\left\| \frac{Z_N}{Z_{N,2} e^{N \xi_{\sim 2}(1)/2}} \langle \sigma \sigma^\top \rangle - \langle \sigma \sigma^\top \rangle_2 \right\|_F^2 \leq C(1 + \gamma_1^4 N^2).$$

In conjunction with Propositions 4.8.8 and 4.8.9, the above immediately implies a positive probability bound on the second moment matrix  $\langle \sigma \sigma^\top \rangle$ .

Both propositions rely on the following truncation to  $Z_N$  developed in [HS23], which allows one to estimate  $Z_N$  via the second moment method throughout the strictly RS regime. As shown in the following lemma, this truncation does not significantly affect the first moment; at the same time, it will force the second moment to be dominated by pairs of nearly-orthogonal points.

**Lemma 4.8.11.** *The following holds for sufficiently small  $c > 0$  depending on  $\varepsilon$ . Let*

$$T = T(H_N) := \left\{ \sigma \in S_N : \int_{S_N} \mathbf{1}[|R(\sigma, \tau)| \geq N^{-2/5}] e^{H_N(\tau)} d\rho(\tau) \leq e^{N\xi(1)/2 - cN^{1/5}} \right\}.$$

*Then, we have:*

$$\mathbf{E} \int_{S_N} \mathbf{1}[\sigma \notin T] e^{H_N(\sigma)} d\rho(\sigma) \leq e^{N\xi(1)/2 - cN^{1/5}}, \quad (4.18)$$

$$\mathbf{E} \int_{S_N} \mathbf{1}[\sigma \notin T] e^{H_{N,2}(\sigma)} d\rho(\sigma) \leq e^{N\xi''(0)/4 - cN^{1/5}}, \quad (4.19)$$

$$\mathbf{E} \int_{S_N} \mathbf{1}[\sigma^1 \notin T, |R(\sigma^1, \sigma^2)| \leq 3N^{-2/5}] e^{H_N(\sigma^1) + H_N(\sigma^2)} d\rho^{\otimes 2}(\sigma^1, \sigma^2) \leq e^{N\xi(1) - cN^{1/5}}, \quad (4.20)$$

$$\mathbf{E} \int_{S_N} \mathbf{1}[\sigma^1 \notin T, |R(\sigma^1, \sigma^2)| \leq 3N^{-2/5}] e^{H_{N,2}(\sigma^1) + H_N(\sigma^2)} d\rho^{\otimes 2}(\sigma^1, \sigma^2) \leq e^{N\xi(1)/2 + N\xi''(0)/4 - cN^{1/5}}. \quad (4.21)$$

The proof of the above lemma is very similar to [HS23, Proposition 3.1] and [HMP24, Lemma 7.9], and we defer it to Section 4.10. As a corollary, we can get control on the first two moments of  $Z_N$  with respect to the randomness in  $H_{N,\sim 2}$ . To this end, let  $\mathbf{E}_{\sim 2}$  denote expectation with respect to  $H_{N,\sim 2}$ .

**Corollary 4.8.12.** *There is a  $H_{N,2}$ -measurable event with probability  $1 - e^{-cN^{1/5}}$  on which the following holds. For*

$$\tilde{T} = \tilde{T}(H_{N,\sim 2}) := \left\{ \sigma \in S_N : \langle \mathbf{1}[|R(\sigma, \tau)| \geq N^{-2/5}] e^{H_{N,\sim 2}(\tau)} \rangle_2 \leq e^{N\xi_{\sim 2}(1)/2 - cN^{1/5}} \right\}, \quad (4.22)$$

*where the Gibbs average is with respect to  $\tau \sim \langle \cdot \rangle_2$ , we have*

$$\begin{aligned} \mathbf{E}_{\sim 2} \left\langle \mathbf{1}[\sigma \notin \tilde{T}] e^{H_{N,\sim 2}(\sigma)} \right\rangle_2 &\leq e^{N\xi_{\sim 2}(1)/2 - cN^{1/5}}, \\ \mathbf{E}_{\sim 2} \left\langle \mathbf{1}[\sigma \notin \tilde{T}] \right\rangle_2 &\leq e^{-cN^{1/5}}, \\ \mathbf{E}_{\sim 2} \left\langle \mathbf{1}[\sigma^1 \notin \tilde{T}, |R(\sigma^1, \sigma^2)| \leq 3N^{-2/5}] e^{H_{N,\sim 2}(\sigma^1) + H_{N,\sim 2}(\sigma^2)} \right\rangle_2 &\leq e^{N\xi_{\sim 2}(1) - cN^{1/5}}, \\ \mathbf{E}_{\sim 2} \left\langle \mathbf{1}[\sigma^1 \notin \tilde{T}, |R(\sigma^1, \sigma^2)| \leq 3N^{-2/5}] e^{H_{N,\sim 2}(\sigma^2)} \right\rangle_2 &\leq e^{N\xi_{\sim 2}(1)/2 - cN^{1/5}}. \end{aligned}$$

*Proof.* By Proposition 4.8.1 and Lemma 4.8.3, with probability  $1 - e^{-cN^{2/5}}$  over  $H_{N,2}$ ,

$$Z_{N,2} \geq e^{N\xi''(0)/4 - cN^{1/5}/2}.$$

On this event, for  $\sigma \in T$  where  $T$  is as in Lemma 4.8.11,

$$\begin{aligned} \langle \mathbf{1}[|R(\sigma, \tau)| \geq N^{-2/5}] e^{H_{N, \sim 2}(\tau)} \rangle_2 &= \frac{1}{Z_{N, 2}} \int_{S_N} \mathbf{1}[|R(\sigma, \tau)| \geq N^{-2/5}] e^{H_N(\tau)} d\rho(\tau) \\ &\leq e^{N\xi_{\sim 2}(1)/2 - cN^{1/5}/2}. \end{aligned}$$

Here we recall  $\xi(1)/2 - \xi''(0)/4 = \xi_{\sim 2}(1)/2$ . So,  $\sigma \in \tilde{T}(H_{N, \sim 2}, c/2)$ , where this denotes  $\tilde{T}$  defined with  $c/2$  in place of  $c$ . Therefore  $T \subseteq \tilde{T}(H_{N, \sim 2}, c/2)$ .

By Markov's inequality and Lemma 4.8.11, with probability  $1 - e^{-cN^{1/5}/4}$  over  $H_{N, 2}$ ,

$$\mathbf{E}_{\sim 2} \int_{S_N} \mathbf{1}[\sigma \notin T] e^{H_N(\sigma)} d\rho(\sigma) \leq e^{N\xi(1)/2 - 3cN^{1/5}/4}$$

On the intersection of these events,

$$\begin{aligned} \mathbf{E}_{\sim 2} \left\langle \mathbf{1}[\sigma \notin \tilde{T}(H_{N, \sim 2}, c/2)] e^{H_{N, \sim 2}(\sigma)} \right\rangle_2 &\leq \mathbf{E}_{\sim 2} \left\langle \mathbf{1}[\sigma \notin T] e^{H_{N, \sim 2}(\sigma)} \right\rangle_2 \\ &= \frac{1}{Z_{N, 2}} \mathbf{E}_{\sim 2} \int_{S_N} \mathbf{1}[\sigma \notin T] e^{H_N(\sigma)} d\rho(\sigma) \\ &\leq e^{N\xi_{\sim 2}(1)/2 - cN^{1/5}/4}. \end{aligned}$$

The first conclusion follows by adjusting  $c$ , and the other two conclusions follow similarly.  $\square$

For the rest of this subsection, we condition on a realization of  $H_{N, 2}$  satisfying the following good event.

**Definition 4.8.13.** Let  $E_2$  denote the  $H_{N, 2}$ -measurable event that the events in Proposition 4.8.8 and Corollary 4.8.12 hold. This occurs with probability  $1 - e^{-cN^{1/5}}$ .

We can now prove Proposition 4.8.9.

*Proof of Proposition 4.8.9.* Let  $\tilde{T}$  be as in Corollary 4.8.12. We can write

$$\frac{Z_N}{Z_{N, 2}} = \langle e^{H_{N, \sim 2}(\sigma)} \rangle_2 = X_1 + X_2, \quad (4.23)$$

where

$$X_1 = \left\langle \mathbf{1}[\sigma \in \tilde{T}] e^{H_{N, \sim 2}(\sigma)} \right\rangle_2, \quad X_2 = \left\langle \mathbf{1}[\sigma \notin \tilde{T}] e^{H_{N, \sim 2}(\sigma)} \right\rangle_2.$$

We will show that that  $X_2$  is much smaller than  $\mathbf{E}_{\sim 2} X_1$  with high probability, and then control the fluctuations of  $X_1$ . For all  $\sigma \in S_N$ ,  $\mathbf{E}_{\sim 2}[e^{H_{N,\sim 2}(\sigma)}] = e^{N\xi_{\sim 2}(1)/2}$ , so Corollary 4.8.12 implies

$$(1 - e^{-cN^{1/5}})e^{N\xi_{\sim 2}(1)/2} \leq \mathbf{E}_{\sim 2}[X_1] \leq e^{N\xi_{\sim 2}(1)/2}. \quad (4.24)$$

On the other hand, by Corollary 4.8.12 and Markov's inequality, with probability  $1 - e^{-cN^{1/5}/2}$  over  $H_{N,\sim 2}$ ,

$$X_2 \leq e^{N\xi_{\sim 2}(1)/2 - cN^{1/5}/2}, \quad (4.25)$$

so  $X_2 \leq e^{-cN^{1/5}/2} \mathbf{E}_{\sim 2} X_1$ , as desired.

We now control the fluctuations of  $X_1$  by estimating

$$\mathbf{Var}_{\sim 2}[X_1] := \mathbf{E}_{\sim 2}[X_1^2] - \mathbf{E}_{\sim 2}[X_1]^2.$$

Then, for  $\sigma^1, \sigma^2 \sim \mu_{H_{N,2}}$ ,

$$\mathbf{E}_{\sim 2}[X_1^2] = \mathbf{E}_{\sim 2} \left\langle \mathbf{1}[\sigma^1, \sigma^2 \in \tilde{T}] e^{H_{N,\sim 2}(\sigma^1) + H_{N,\sim 2}(\sigma^2)} \right\rangle_2 \leq \mathbf{E}_{\sim 2}[Y_1] + \mathbf{E}_{\sim 2}[Y_2],$$

where

$$\begin{aligned} Y_1 &= \left\langle \mathbf{1}[|R(\sigma^1, \sigma^2)| \leq N^{-2/5}] e^{H_{N,\sim 2}(\sigma^1) + H_{N,\sim 2}(\sigma^2)} \right\rangle_2, \\ Y_2 &= \left\langle \mathbf{1}[\sigma^1 \in \tilde{T}, |R(\sigma^1, \sigma^2)| \geq N^{-2/5}] e^{H_{N,\sim 2}(\sigma^1) + H_{N,\sim 2}(\sigma^2)} \right\rangle_2. \end{aligned} \quad (4.26)$$

By the definition of  $\tilde{T}$  and (4.24),

$$\mathbf{E}_{\sim 2}[Y_2] \leq \mathbf{E}_{\sim 2} \left\langle \mathbf{1}[\sigma^1 \in \tilde{T}] e^{H_{N,\sim 2}(\sigma^1)} \right\rangle_2 e^{N\xi_{\sim 2}(1)/2 - cN^{1/5}} \leq e^{N\xi_{\sim 2}(1) - cN^{1/5}}. \quad (4.27)$$

We further calculate

$$\mathbf{E}_{\sim 2}[Y_1] = e^{N\xi_{\sim 2}(1)} \left\langle \mathbf{1}[|R(\sigma^1, \sigma^2)| \leq N^{-2/5}] e^{N\xi_{\sim 2}(R(\sigma^1, \sigma^2))} \right\rangle_2.$$

Recall that in Theorem 4.7.27, we assumed  $\gamma_1^2 \leq N^{-4/5}$ . Thus, for  $|R| \leq N^{-2/5}$ ,

$$\xi_{\sim 2}(R) = \gamma_1^2 R + O(R^3) = O(N^{-6/5}).$$

It follows that  $\mathbf{E}_{\sim 2}[Y_1] \leq (1 + O(N^{-1/5}))e^{N\xi_{\sim 2}(1)}$ .

Combining the above estimates shows  $\mathbf{E}_{\sim 2}[X_1^2] \leq (1 + O(N^{-1/5}))e^{N\xi_{\sim 2}(1)}$ . Further combining with the lower bound in (4.24) shows

$$\mathbf{Var}_{\sim 2}[X_1] = O(N^{-1/5})e^{N\xi_{\sim 2}(1)}.$$

By Chebyshev's inequality, with probability  $1 - N^{-1/15}/2$ ,

$$|X_1 - \mathbf{E}_{\sim 2}[X_1]| = O(N^{-1/15})e^{N\xi_{\sim 2}(1)/2}$$

Union bounding with the event in (4.25), and recalling (4.24), we conclude that with probability  $1 - N^{-1/15}$ ,

$$\frac{Z_N}{Z_{N,2}} = (1 + O(N^{-1/15}))e^{N\xi_{\sim 2}(1)/2}. \quad \square$$

The above proof also implies the following estimate, which will be useful in the sequel.

**Corollary 4.8.14.** *On the event  $E_2$  (Definition 4.8.13), with probability  $1 - e^{-cN^{1/5}}$  over  $H_{N,\sim 2}$ ,*

$$\int \mathbf{1}[|R(\sigma^1, \sigma^2)| \geq N^{-2/5}] e^{H_N(\sigma^1) + H_N(\sigma^2)} d\rho^{\otimes 2}(\sigma^1, \sigma^2) \leq Z_{N,2}^2 e^{N\xi_{\sim 2}(1) - cN^{1/5}}.$$

*Proof.* Dividing through by  $Z_{N,2}^2$ , it suffices to show, for  $\sigma^1, \sigma^2 \sim \mu_{H_{N,2}}$ ,

$$\left\langle \mathbf{1}[|R(\sigma^1, \sigma^2)| \geq N^{-2/5}] e^{H_{N,\sim 2}(\sigma^1) + H_{N,\sim 2}(\sigma^2)} \right\rangle_2 \leq e^{N\xi_{\sim 2}(1) - cN^{1/5}}.$$

The left-hand side is bounded by  $Y_2 + Y_3$ , where  $Y_2$  is as in (4.26) and

$$Y_3 := \left\langle \mathbf{1}[\sigma^1 \notin \tilde{T}] e^{H_{N,\sim 2}(\sigma^1) + H_{N,\sim 2}(\sigma^2)} \right\rangle_2 = \left\langle \mathbf{1}[\sigma \notin \tilde{T}] e^{H_{N,\sim 2}(\sigma)} \right\rangle_2 \left\langle e^{H_{N,\sim 2}(\sigma)} \right\rangle_2.$$

By (4.27),  $\mathbf{E}_{\sim 2}[Y_2] \leq e^{N\xi_{\sim 2}(1) - cN^{1/5}}$ . By Corollary 4.8.12,

$$\mathbf{E}_{\sim 2} \left\langle \mathbf{1}[\sigma \notin \tilde{T}] e^{H_{N,\sim 2}(\sigma)} \right\rangle_2 \leq e^{N\xi_{\sim 2}(1)/2 - cN^{1/5}}, \quad \mathbf{E}_{\sim 2} \langle e^{H_{N,\sim 2}(\sigma)} \rangle_2 \leq e^{N\xi_{\sim 2}(1)/2}.$$

So, the following estimates each hold with probability  $1 - e^{-cN^{1/5}/4}$  over  $H_{N,\sim 2}$ :

$$Y_2 \leq e^{N\xi_{\sim 2}(1) - cN^{1/5}/2}, \quad \left\langle \mathbf{1}[\sigma \notin \tilde{T}] e^{H_{N,\sim 2}(\sigma)} \right\rangle_2 \leq e^{N\xi_{\sim 2}(1)/2 - 3cN^{1/5}/4},$$

$$\langle e^{H_{N,\sim 2}(\sigma)} \rangle_2 \leq e^{N\xi_{\sim 2}(1)/2 + cN^{1/5}/4}.$$

The conclusion follows on the intersection of these events, after adjusting  $c$ .  $\square$

We now turn to the proof of Proposition 4.8.10. By rotational invariance of gaussians, we may assume  $\nabla^2 H_N(0)$  is diagonal while keeping the law of  $H_{N,\sim 2}$  unchanged. For  $i, j \in [N]$ , define

$$X_{i,j} = \left\langle \sigma_i \sigma_j \left( e^{H_{N,\sim 2}(\sigma) - N\xi_{\sim 2}(1)/2} - \mathbf{1}[i = j] \right) \right\rangle_2,$$



and note that this equals the  $(i, j)$  entry of the matrix appearing in Proposition 4.8.10. For  $\sigma \in \mathbb{R}^N$  and  $i \in [N]$ , let  $\sigma_{\sim i} \in \mathbb{R}^{N-1}$  denote  $\sigma$  with coordinate  $i$  omitted. Similarly, for  $i \neq j$ , let  $\sigma_{\sim i, j} \in \mathbb{R}^{N-2}$  denote  $\sigma$  with coordinates  $i$  and  $j$  omitted, and by slight abuse of notation let  $\sigma_{\sim i, i} = \sigma_{\sim i}$ . For  $i, j \in [N]$  (possibly with  $i = j$ ) define analogously to  $\tilde{T}$

$$\tilde{T}_{i,j} := \left\{ \sigma \in S_N : \left\langle \mathbf{1}[|R(\sigma_{\sim i, j}, \tau_{\sim i, j})| \geq 2N^{-2/5}] e^{H_{N, \sim 2}(\tau)} \right\rangle_2 \leq e^{N\xi_{\sim 2}(1)/2 - cN^{1/5}} \right\}, \quad (4.28)$$

where we recall the Gibbs average is with respect to  $\tau \sim \langle \cdot \rangle_2$ . Then define

$$\begin{aligned} \tilde{X}_{i,j} &= \left\langle \mathbf{1}[|\sigma_i|, |\sigma_j| \leq \log N, \sigma \in \tilde{T}_{i,j}] \sigma_i \sigma_j \left( e^{H_{N, \sim 2}(\sigma) - N\xi_{\sim 2}(1)/2} - \mathbf{1}[i = j] \right) \right\rangle_2 \\ \widehat{X}_{i,j} &= \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R(\sigma_{\sim i, j}^1, \sigma_{\sim i, j}^2)| \leq 2N^{-2/5}] \right. \\ &\quad \left. \sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2 \left( e^{H_{N, \sim 2}(\sigma^1) - N\xi_{\sim 2}(1)/2} - \mathbf{1}[i = j] \right) \left( e^{H_{N, \sim 2}(\sigma^2) - N\xi_{\sim 2}(1)/2} - \mathbf{1}[i = j] \right) \right\rangle_2. \end{aligned}$$

Note that  $\widehat{X}_{i,j}$  is the contribution to  $X_{i,j}^2$  coming from  $\sigma^1, \sigma^2$  that are both not localized to coordinate  $i$  or  $j$  and have small overlap. The following two lemmas reduce the task of controlling  $X_{i,j}^2$  to bounding  $\mathbf{E}_{\sim 2} \widehat{X}_{i,j}$ . They are proved by manipulating the typicality truncations  $\tilde{T}$  and  $\tilde{T}_{i,j}$  similarly to the proofs above; we defer these proofs to Section 4.10.

**Lemma 4.8.15.** *For each  $i, j \in [N]$ , with probability  $1 - e^{-c \log^2 N}$  over  $H_{N, \sim 2}$ ,*

$$X_{i,j}^2 \leq 2\tilde{X}_{i,j}^2 + e^{-c \log^2 N}.$$

**Lemma 4.8.16.** *For each  $i, j \in [N]$ ,*

$$\mathbf{E}_{\sim 2} \tilde{X}_{i,j}^2 \leq \mathbf{E}_{\sim 2} \widehat{X}_{i,j} + e^{-cN^{1/5}}.$$

We now turn to bounding the  $\mathbf{E}_{\sim 2} \widehat{X}_{i,j}$ . This is achieved by the following pair of propositions.

**Proposition 4.8.17.** *For any  $i \in [N]$ , we have  $\mathbf{E}_{\sim 2} \widehat{X}_{i,i} \leq C(N\gamma_1^4 + N^{-1})$ .*

**Proposition 4.8.18.** *For any distinct  $i, j \in [N]$ , we have  $\mathbf{E}_{\sim 2} \widehat{X}_{i,j} \leq C(\gamma_1^4 + N^{-2})$ .*

Throughout the next two proofs,  $\langle \cdot \rangle_2$  denotes expectation w.r.t.  $\sigma^1, \sigma^2 \sim \langle \cdot \rangle_2$ , and we write  $R = R(\sigma^1, \sigma^2)$ ,  $R_{\sim i} = R(\sigma_{\sim i}^1, \sigma_{\sim i}^2)$ , and  $R_{\sim i, j} = R(\sigma_{\sim i, j}^1, \sigma_{\sim i, j}^2)$ .

*Proof of Proposition 4.8.17.* By direct calculation,

$$\begin{aligned}\mathbf{E}_{\sim 2} \widehat{X}_{i,i} &= \mathbf{E}_{\sim 2} \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_i^2| \leq \log N, |R_{\sim i}| \leq 2N^{-2/5}] \right. \\ &\quad \left. (\sigma_i^1)^2 (\sigma_i^2)^2 \left( e^{H_{N,\sim 2}(\sigma^1) - N\xi_{\sim 2}(1)/2} - 1 \right) \left( e^{H_{N,\sim 2}(\sigma^2) - N\xi_{\sim 2}(1)/2} - 1 \right) \right\rangle_2 \\ &= \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_i^2| \leq \log N, |R_{\sim i}| \leq 2N^{-2/5}] (\sigma_i^1)^2 (\sigma_i^2)^2 \left( e^{N\xi_{\sim 2}(R)} - 1 \right) \right\rangle_2.\end{aligned}$$

In view of Proposition 4.8.8,  $\sigma_i^1$  and  $\sigma_i^2$  are subgaussian of scale  $O(1)$  and  $R$  is subgaussian of scale  $O(N^{-1/2})$ . We will see that the above integral is dominated by  $|\sigma_i^1| \asymp |\sigma_i^2| \asymp 1$  and  $|R| \asymp N^{-1/2}$ , in which case Taylor expanding  $e^{N\xi_{\sim 2}(R)}$  shows this integral has the desired scale. Formally, we can write the above integral as  $Y_{i,i}^{(1)} + Y_{i,i}^{(2)}$ , where

$$\begin{aligned}Y_{i,i}^{(1)} &= \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_i^2| \leq \log N, |R_{\sim i}| \leq 2N^{-1/2} \log N] (\sigma_i^1)^2 (\sigma_i^2)^2 \left( e^{N\xi_{\sim 2}(R)} - 1 \right) \right\rangle_2 \\ Y_{i,i}^{(2)} &= \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_i^2| \leq \log N, 2N^{-1/2} \log N \leq |R_{\sim i}| \leq 2N^{-2/5}] (\sigma_i^1)^2 (\sigma_i^2)^2 \left( e^{N\xi_{\sim 2}(R)} - 1 \right) \right\rangle_2\end{aligned}$$

are the contributions from  $|R_{\sim i}|$  smaller and larger than  $2N^{-1/2} \log N$ .

We first address  $Y_{i,i}^{(2)}$ . Note that on the event in the indicator in  $Y_{i,i}^{(2)}$ ,  $|R| \leq 3N^{-2/5}$ . Thus, as  $\gamma_1^2 \leq N^{-4/5}$ ,

$$N|\xi_{\sim 2}(R)| \leq N\gamma_1^2 R + O(NR^3) = O(N^{-1/5}).$$

It follows that  $|e^{N\xi_{\sim 2}(R)} - 1| \leq 1$ . Then, by Cauchy–Schwarz,

$$\begin{aligned}Y_{i,i}^{(2)} &\leq \left\langle \mathbf{1}[|R| \geq N^{-1/2} \log N] (\sigma_i^1)^2 (\sigma_i^2)^2 \right\rangle_2 \\ &\leq \left\langle \mathbf{1}[|R| \geq N^{-1/2} \log N] \right\rangle_2^{1/2} \left\langle (\sigma_i^1)^2 (\sigma_i^2)^2 \right\rangle_2^{1/2} \\ &\leq e^{-c \log^2 N} \cdot O(1) = e^{-c \log^2 N},\end{aligned}\tag{4.29}$$

where we have used Proposition 4.8.8(1) for the tail probability and Proposition 4.8.8(2) for the coordinate moments.

Next we turn to  $Y_{i,i}^{(1)}$ . On the event in the indicator in  $Y_{i,i}^{(1)}$ ,  $|R| \leq 3N^{-1/2} \log N$ , so

$$N|\xi_{\sim 2}(R)| \leq N\gamma_1^2 R + O(NR^3) = O(N^{-3/10} \log N),$$

where we recall  $\gamma_1^2 \leq N^{-4/5}$ . Thus, Taylor expanding the exponential and  $\xi_{\sim 2}$ ,

$$Y_{i,i}^{(1)} = \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_i^2| \leq \log N, |R_{\sim i}| \leq 2N^{-1/2} \log N] (\sigma_i^1)^2 (\sigma_i^2)^2 \right\rangle_2$$

$$\begin{aligned}
& \left( N\xi_{\sim 2}(R) + \frac{1}{2}N^2\xi_{\sim 2}(R)^2 + \frac{1}{6}N^3\xi_{\sim 2}(R)^3 \right) \Bigg\rangle_2 + O(N^{-6/5} \log^8 N) \\
&= \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_i^2| \leq \log N, |R_{\sim i}| \leq 2N^{-1/2} \log N] (\sigma_i^1)^2 (\sigma_i^2)^2 \right. \\
& \quad \left. (N(\gamma_1^2 R + \gamma_3^2 R^3 + \gamma_4^2 R^4) + \frac{1}{2}N^2(\gamma_1^2 R + \gamma_3^2 R^3)^2 + \frac{1}{6}N^3\gamma_1^6 R^3) \right\rangle_2 + O(N^{-11/10} \log^9 N).
\end{aligned}$$

By exchangeability of  $(\sigma_i^1, -\sigma_i^1)$ ,  $(\sigma_i^2, -\sigma_i^2)$ , and  $(R_{\sim i}, -R_{\sim i})$ , all the odd degree in  $R$  terms vanish, leaving

$$Y_{i,i}^{(1)} = \frac{1}{2}N^2\gamma_1^4 Q_2 + (N\gamma_4^2 + N^2\gamma_1^2\gamma_3^2)Q_4 + \frac{1}{2}N^2\gamma_3^4 Q_6 + o(N^{-1}),$$

and where we have introduced the notation

$$Q_k := \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_i^2| \leq \log N, |R_{\sim i}| \leq 2N^{-1/2} \log N] (\sigma_i^1)^2 (\sigma_i^2)^2 R^k \right\rangle_2.$$

By Cauchy–Schwarz and Proposition 4.8.8, for each  $k \in \{2, 4, 6\}$ ,

$$Q_k \leq \left\langle (\sigma_i^1)^4 (\sigma_i^2)^4 \right\rangle_2^{1/2} \left\langle R^{2k} \right\rangle_2^{1/2} = O(N^{-k/2}).$$

This implies  $Y_{i,i}^{(1)} \leq C(N\gamma_1^4 + N^{-1})$ . Combining with the bound (4.29) on  $Y_{i,i}^{(2)}$  implies the result.  $\square$

*Proof of Proposition 4.8.18.* We calculate as above

$$\begin{aligned}
\mathbf{E}_{\sim 2} \widehat{X}_{i,j} &\leq \mathbf{E}_{\sim 2} \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-2/5}] \right. \\
& \quad \left. \sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2 e^{H_{N,\sim 2}(\sigma^1) + H_{N,\sim 2}(\sigma^2) - N\xi_{\sim 2}(1)} \right\rangle_2 \\
&= \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-2/5}] \sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2 e^{N\xi_{\sim 2}(R)} \right\rangle_2.
\end{aligned}$$

Our strategy for evaluating this will be similar as above, except that because this integral contains  $\sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2$  instead of  $(\sigma_i^1)^2 (\sigma_i^2)^2$ , we will need to expand the exponential more carefully to obtain cancellations in these terms. Formally, we write the above integral as  $Y_{i,j}^{(1)} + Y_{i,j}^{(2)}$  for

$$\begin{aligned}
Y_{i,j}^{(1)} &= \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-1/2} \log N] \sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2 e^{N\xi_{\sim 2}(R)} \right\rangle_2, \\
Y_{i,j}^{(2)} &= \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, 2N^{-1/2} \log N \leq |R_{\sim i,j}| \leq 2N^{-2/5}] \sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2 e^{N\xi_{\sim 2}(R)} \right\rangle_2.
\end{aligned}$$

Identically to the previous proof, on the event in the indicator in  $Y_{i,j}^{(2)}$  we have  $N|\xi_{\sim 2}(R)| = O(N^{-1/5})$ , so  $e^{N\xi_{\sim 2}(R)} \leq 2$ . Then, by Cauchy–Schwarz and Proposition 4.8.8,

$$\begin{aligned} |Y_{i,j}^{(2)}| &\leq 2 \left\langle \mathbf{1}[|R| \geq N^{-1/2} \log N] |\sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2| \right\rangle_2 \\ &\leq 2 \langle \mathbf{1}[|R| \geq N^{-1/2} \log N] \rangle_2^{1/2} \left\langle (\sigma_i^1)^4 (\sigma_i^2)^4 \right\rangle_2^{1/4} \left\langle (\sigma_j^1)^4 (\sigma_j^2)^4 \right\rangle_2^{1/4} \\ &\leq 2e^{-c \log^2 N} \cdot O(1) \cdot O(1) = e^{-c \log^2 N}. \end{aligned}$$

To address  $Y_{i,j}^{(1)}$ , define  $\Delta_i = \sigma_i^1 \sigma_i^2 / N$  and  $\Delta_j = \sigma_j^1 \sigma_j^2 / N$ , the contributions to  $R$  coming from the  $i$ th and  $j$ th coordinate, respectively. Then, by exchangeability of  $(\sigma_i^1, -\sigma_i^1)$  and  $(\sigma_j^1, -\sigma_j^1)$ ,

$$\begin{aligned} 4Y_{i,j}^{(1)} &= \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-1/2} \log N] \right. \\ &\quad \left. \sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2 (e^{N\xi_{\sim 2}(R_{\sim i,j} + \Delta_i + \Delta_j)} - e^{N\xi_{\sim 2}(R_{\sim i,j} + \Delta_i - \Delta_j)} - e^{N\xi_{\sim 2}(R_{\sim i,j} - \Delta_i + \Delta_j)} + e^{N\xi_{\sim 2}(R_{\sim i,j} - \Delta_i - \Delta_j)}) \right\rangle_2. \end{aligned}$$

Note that on the event in this indicator,  $|R_{\sim i,j} \pm \Delta_i \pm \Delta_j| \leq 3N^{-1/2} \log N$ . Define

$$\kappa(x) = e^{N\xi_{\sim 2}(x)},$$

and note that

$$\begin{aligned} \sup_{|x| \leq 3N^{-1/2} \log N} \kappa^{(4)}(x) &= \sup_{|x| \leq 3N^{-1/2} \log N} \left( N\xi_{\sim 2}^{(4)}(x) + 4N^2 \xi'_{\sim 2}(x) \xi_{\sim 2}^{(3)}(x) + 3N^2 \xi''_{\sim 2}(x)^2 + \right. \\ &\quad \left. + 6N^3 \xi'_{\sim 2}(x)^2 \xi''_{\sim 2}(x) + N^4 \xi'_{\sim 2}(x)^4 \right) \kappa(x) = O(N^{6/5}), \end{aligned}$$

where we have used that  $\sup_{|x| \leq 3N^{-1/2} \log N} \kappa(x) \leq 2$  and  $\gamma_1^2 \leq N^{-4/5}$ . Since  $|\Delta_i|, |\Delta_j| \leq N^{-1} \log^2 N$  on the event in the indicator, for  $s_i, s_j \in \{\pm 1\}$ ,

$$\begin{aligned} e^{N\xi_{\sim 2}(R_{\sim i,j} + s_i \Delta_i + s_j \Delta_j)} &= \kappa(R_{\sim i,j}) + \kappa'(R_{\sim i,j})(s_i \Delta_i + s_j \Delta_j) + \frac{1}{2} \kappa''(R_{\sim i,j})(s_i \Delta_i + s_j \Delta_j)^2 \\ &\quad + \frac{1}{6} \kappa^{(3)}(R_{\sim i,j})(s_i \Delta_i + s_j \Delta_j)^3 + O(N^{6/5}) \cdot (2N^{-1} \log^2 N)^4. \end{aligned}$$

It follows that

$$\begin{aligned} &e^{N\xi_{\sim 2}(R_{\sim i,j} + \Delta_i + \Delta_j)} - e^{N\xi_{\sim 2}(R_{\sim i,j} + \Delta_i - \Delta_j)} - e^{N\xi_{\sim 2}(R_{\sim i,j} - \Delta_i + \Delta_j)} + e^{N\xi_{\sim 2}(R_{\sim i,j} - \Delta_i - \Delta_j)} \\ &= 4\kappa''(R_{\sim i,j})\Delta_i \Delta_j + O(N^{-14/5} \log^8 N), \end{aligned}$$

and thus

$$\begin{aligned}
Y_{i,j}^{(1)} &= \left\langle \mathbf{1}[\sigma_i^1, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-1/2} \log N] \sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2 \Delta_i \Delta_j \kappa''(R_{\sim i,j}) \right\rangle_2 \\
&\quad + O(N^{-14/5} \log^{12} N) \\
&= \left\langle \mathbf{1}[\sigma_i^1, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-1/2} \log N] \right. \\
&\quad \left. (\sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2)^2 (N^{-1} \xi''_{\sim 2}(R_{\sim i,j}) + \xi'_{\sim 2}(R_{\sim i,j})^2) e^{N \xi_{\sim 2}(R_{\sim i,j})} \right\rangle_2 + o(N^{-2}).
\end{aligned}$$

On the event in this indicator,  $e^{N \xi_{\sim 2}(R_{\sim i,j})} \leq 2$ , and therefore  $\xi'_{\sim 2}$  and  $\xi''_{\sim 2}$  can be Taylor expanded to obtain

$$Y_{i,j}^{(1)} \leq Y_{i,j}^{(3)} + Y_{i,j}^{(4)} + o(N^{-2}),$$

where

$$\begin{aligned}
Y_{i,j}^{(3)} &= 6\gamma_3^2 N^{-1} \left\langle \mathbf{1}[\sigma_i^1, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-1/2} \log N] \right. \\
&\quad \left. (\sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2)^2 R_{\sim i,j} e^{N \xi_{\sim 2}(R_{\sim i,j})} \right\rangle_2, \\
Y_{i,j}^{(4)} &= \mathbf{E}_{\mu_{H_{N,2}}} \left\langle \mathbf{1}[\sigma_i^1, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-1/2} \log N] \right. \\
&\quad \left. (\sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2)^2 \left( 12\gamma_4^2 N^{-1} R_{\sim i,j}^2 + (\gamma_1^2 + 3\gamma_3^2 R_{\sim i,j}^2)^2 \right) e^{N \xi_{\sim 2}(R_{\sim i,j})} \right\rangle_2
\end{aligned}$$

On the event in these indicators, we further have

$$|R^2 - R_{\sim i,j}^2| = |R - R_{\sim i,j}| |R + R_{\sim i,j}| \leq (|\Delta_i| + |\Delta_j|) \cdot 5N^{-1/2} \log N = O(N^{-3/2} \log^3 N). \quad (4.30)$$

From this it readily follows that

$$\begin{aligned}
Y_{i,j}^{(4)} &\leq 2 \left\langle \mathbf{1}[\sigma_i^1, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-1/2} \log N] \right. \\
&\quad \left. (\sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2)^2 (12\gamma_4^2 N^{-1} R^2 + (\gamma_1^2 + 3\gamma_3^2 R^2)^2) \right\rangle_2 + o(N^{-2}) \\
&= 2\gamma_1^4 \tilde{Q}_0 + (12\gamma_1^2 \gamma_3^2 + 24\gamma_4^2 N^{-1}) \tilde{Q}_2 + 18\gamma_3^2 \tilde{Q}_4 + o(N^{-2}),
\end{aligned}$$

where

$$\tilde{Q}_k = \left\langle \mathbf{1}[\sigma_i^1, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-1/2} \log N] (\sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2)^2 R^k \right\rangle_2.$$

By Cauchy–Schwarz and Proposition 4.8.8, for each  $k \in \{0, 2, 4\}$ ,

$$\widetilde{Q}_k \leq \langle (\sigma_i^1)^8 (\sigma_i^2)^8 \rangle_2^{1/4} \left\langle (\sigma_j^1)^8 (\sigma_j^2)^8 \right\rangle_2^{1/4} \langle R^{2k} \rangle_2^{1/2} = O(N^{-k/2}).$$

This implies  $Y_{i,j}^{(4)} \leq C(\gamma_1^4 + N^{-2})$ . To control  $Y_{i,j}^{(3)}$ , we recall that  $|N\xi_{\sim 2}(R_{\sim i,j})| = O(N^{-3/10} \log N)$  and Taylor expand the exponential:

$$\begin{aligned} Y_{i,j}^{(3)} &= 6\gamma_3^2 N^{-1} \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-1/2} \log N] \right. \\ &\quad \left. (\sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2)^2 R_{\sim i,j} (1 + N\xi_{\sim 2}(R_{\sim i,j})) \right\rangle_2 + o(N^{-2}) \\ &= 6\gamma_3^2 N^{-1} \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-1/2} \log N] \right. \\ &\quad \left. (\sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2)^2 (R_{\sim i,j} + N\gamma_1^2 R_{\sim i,j}^2 + N\gamma_3^2 R_{\sim i,j}^4) \right\rangle_2 + o(N^{-2}). \end{aligned}$$

By exchangeability of  $(R_{\sim i,j}, -R_{\sim i,j})$ , the contribution of the term  $R_{\sim i,j}$  vanishes. By (4.30), we can further estimate  $R_{\sim i,j}^2$  with  $R^2$ , obtaining

$$\begin{aligned} Y_{i,j}^{(3)} &= 6\gamma_3^2 \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R_{\sim i,j}| \leq 2N^{-1/2} \log N] \right. \\ &\quad \left. (\sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2)^2 (\gamma_1^2 R^2 + \gamma_3^2 R^4) \right\rangle_2 + o(N^{-2}) \\ &= 6\gamma_1^2 \gamma_3^2 \widetilde{Q}_2 + 6\gamma_3^4 \widetilde{Q}_4 + o(N^{-2}) \leq C(\gamma_1^2 N^{-1} + N^{-2}). \end{aligned}$$

Combining all of the above estimates concludes the proof.  $\square$

*Proof of Proposition 4.8.10.* By a union bound, the event in Lemma 4.8.15 holds for all  $i, j \in [N]$  with probability  $1 - e^{-c \log^2 N}$  (over  $H_{N, \sim 2}$ ). On this event,

$$\left\| \frac{Z_N}{Z_{N,2} e^{N\xi_{\sim 2}(1)/2}} \langle \sigma \sigma^\top \rangle - \langle \sigma \sigma^\top \rangle_2 \right\|_F^2 = \sum_{i,j=1}^N X_{i,j}^2 \leq 2 \sum_{i,j=1}^N \widetilde{X}_{i,j}^2 + e^{-c \log^2 N}. \quad (4.31)$$

Combining Lemma 4.8.16 and Propositions 4.8.17 and 4.8.18 shows that

$$\mathbf{E}_{\sim 2} \sum_{i,j=1}^N \widetilde{X}_{i,j}^2 \leq 2C(N^2 \gamma_1^4 + 1) + e^{-cN^{1/5}} \leq 3C(N^2 \gamma_1^4 + 1).$$

Thus, with probability  $2/3$  over  $H_{N,\sim 2}$ ,  $\sum_{i,j=1}^N \tilde{X}_{i,j}^2 \leq 9C(N^2\gamma_1^4 + 1)$ . Combining with (4.31) and taking a final union bound shows that with probability  $1/2$  over  $H_{N,\sim 2}$ ,

$$\left\| \frac{Z_N}{Z_{N,2} e^{N\xi_{\sim 2}(1)/2}} \langle \sigma \sigma^\top \rangle - \langle \sigma \sigma^\top \rangle_2 \right\|_F^2 \leq 18C(N^2\gamma_1^4 + 1) + e^{-c \log^2 N} \leq 20C(N^2\gamma_1^4 + 1).$$

The result follows after adjusting  $C$ . □

## From positive to very high probability

In this section, we boost the positive probability bound on the second moment matrix to a very high probability bound. To this end, we will show that an appropriate proxy function for the second moment matrix is very Lipschitz. This will imply the desired concentration by standard gaussian concentration.

Let  $\mathbf{g} \in \mathbb{R}^{N+N^3+N^4+\dots+N^{p^*}}$  be the vectorized collection of all gaussian interactions corresponding to  $H_{N,\sim 2}$ . Throughout, as in the previous subsection, we will condition on the event  $E_2$  from Definition 4.8.13 over  $H_{N,2}$ , which holds with probability  $1 - e^{-cN^{1/5}}$ . We define the following functions of  $\mathbf{g}$ :

$$\begin{aligned} F_1(\mathbf{g}) &= B(1 + \gamma_1^2 N) \cdot \left( 1 - \left| \log \frac{Z_N}{Z_{N,2}} - \frac{N\xi_{\sim 2}(1)}{2} \right| \right) \\ F_2(\mathbf{g}) &= B(1 + \gamma_1^2 N) - \|\langle \sigma \sigma^\top \rangle\|_{\text{op}} \\ F(\mathbf{g}) &= \max(\min(F_1(\mathbf{g}), F_2(\mathbf{g})), 0), \end{aligned}$$

where  $B$  is a sufficiently large constant (specified in the proof of Lemma 4.8.19). If we can show that with high probability over  $\mathbf{g}$ ,  $\min(F_1(\mathbf{g}), F_2(\mathbf{g})) \geq 0$ , then the conclusion follows. Indeed, from  $F_2(\mathbf{g}) \geq 0$ , we immediately obtain  $\|\langle \sigma \sigma^\top \rangle\|_{\text{op}} \leq B(1 + \gamma_1^2 N)$ .  $F_1$  allows control over the free energy of the  $p$ -spin model in terms of that of the corresponding 2-spin model. This gives good control over the overlaps (in a manner to be made precise shortly), which is crucial for establishing the high probability statement. It is also important earlier in this section, in showing that the free energy of the  $p$ -spin model concentrates well.

Towards this, we first start with the positive probability statement, which was essentially established in the previous subsections.

**Lemma 4.8.19.** *There exists a constant  $B > 0$  such that with probability at least  $\frac{1}{3}$ , we have  $F(\mathbf{g}) \geq \frac{B}{2}(1 + \gamma_1^2 N)$ .*

*Proof.* By Proposition 4.8.9, with probability  $1 - O(N^{-1/15})$  over  $\mathbf{g}$ , we have  $F_1(\mathbf{g}) \geq \frac{B}{2}(1 + \gamma_1^2 N)$ , so  $\frac{Z_N}{Z_{N,2} e^{N\xi_{\sim 2}(1)/2}} \geq e^{-1/2}$ . Intersecting this with the event from Proposition 4.8.10

implies that with probability at least  $\frac{1}{3}$ ,

$$\begin{aligned} \|\langle \sigma \sigma^\top \rangle\|_{\text{op}} &\leq e^{1/2} \|\langle \sigma \sigma^\top \rangle_2\|_{\text{op}} + e^{1/2} \sqrt{C(1 + \gamma_1^4 N^2)} \\ &\leq \frac{B}{2}(1 + \gamma_1^2 N), \end{aligned}$$

where we have used  $E_2$  to apply Proposition 4.8.8 and after appropriately picking  $B$ .  $\square$

Let  $\mathcal{E}$  denote the  $H_{N,2}$ -measurable event from Corollary 4.8.14:

$$\left\{ \mathbf{g} : \int \mathbf{1}[|R(\sigma^1, \sigma^2)| \geq N^{-2/5}] e^{H_N(\sigma^1) + H_N(\sigma^2)} d\rho^{\otimes 2}(\sigma^1, \sigma^2) \leq Z_{N,2}^2 e^{N\xi_{\sim 2}(1) - cN^{1/5}} \right\},$$

which holds with probability  $1 - e^{-cN^{1/5}}$  over  $\mathbf{g}$ . The key observation is that this gives us good control on the overlaps.

**Lemma 4.8.20.** *On  $\mathcal{E}$ , if  $F_1(\mathbf{g}) \geq 0$ , then for any  $0 \leq p \leq \log^2 N$ , we have*

$$\|\langle \sigma^{\otimes p} \rangle\|_F^2 \leq O(N^{3p/5}).$$

*Proof.* By splitting up the expectation based on whether  $|R(\sigma^1, \sigma^2)| \geq N^{-2/5}$ , on  $\mathcal{E}$  we have

$$\begin{aligned} \|\langle \sigma^{\otimes p} \rangle\|_F^2 &= \langle N^p R(\sigma^1, \sigma^2)^p \rangle \\ &\leq N^{3p/5} + N^{2p} \frac{1}{Z_N^2} \int \mathbf{1}[|R(\sigma^1, \sigma^2)| \geq N^{-2/5}] e^{H_N(\sigma^1) + H_N(\sigma^2)} d\rho^{\otimes 2}(\sigma^1, \sigma^2) \\ &\leq N^{3p/5} + N^{2p} \frac{Z_{N,2}^2}{Z_N^2} e^{N\xi_{\sim 2}(1) - cN^{1/5}} \quad (\text{Definition of } \mathcal{E}) \\ &\leq N^{3p/5} + N^{2p} e^{1 - cN^{1/5}}, \end{aligned}$$

where the last line used  $F_1(\mathbf{g}) \geq 0$ . Since  $p \leq \log^2 N$ , the above quantity is  $O(N^{3p/5})$ , as desired.  $\square$

The above is a crucial input to prove Lipschitzness of  $F$  on  $\mathcal{E}$ .

**Lemma 4.8.21.** *The function  $F$  is  $O((1 + \gamma_1^2 N)N^{-1/10})$ -Lipschitz restricted to  $\mathcal{E}$ .*

Before we prove this, let us see how it implies Proposition 4.8.2, restated for convenience.

**Proposition 4.8.2.** *There is a  $H_{N,2}$ -measurable event with probability  $1 - e^{-cN^{1/5}}$  on which the following holds with probability  $1 - e^{-cN^{1/5}}$  over  $H_{N,\sim 2}$ .*



1. The partition functions  $Z_N, Z_{N,2}$  satisfy

$$\left| \log \frac{Z_N}{Z_{N,2}} - \frac{N\xi_{\sim 2}(1)}{2} \right| \leq 1/2.$$

2. The Gibbs measure satisfies  $\|\langle \sigma \sigma^\top \rangle\|_{\text{op}} \leq C(1 + \gamma_1^2 N)$ .

*Proof of Proposition 4.8.2.* By Kirszbraun's extension theorem, we can extend  $F$  to  $\tilde{F}$  such that each  $\tilde{F}$  has the same Lipschitz constant as  $F$  and agrees with  $F$  on  $\mathcal{E}$ . We can now apply gaussian concentration to  $\tilde{F}$  to conclude that

$$\Pr \left[ |\tilde{F}(\mathbf{g}) - \mathbf{E}\tilde{F}(\mathbf{g})| \geq \frac{B}{4}(1 + \gamma_1^2 N) \right] \leq e^{-cN^{1/5}}. \quad (4.32)$$

By Lemma 4.8.19, with probability at least  $\frac{1}{3}$ , we have  $F(\mathbf{g}) \geq \frac{B}{2}(1 + \gamma_1^2 N)$ . Upon further intersection with  $\mathcal{E}$  (where  $\tilde{F}(\mathbf{g}) = F(\mathbf{g})$ ) and the complement of the event in (4.32), we conclude  $\mathbf{E}\tilde{F}(\mathbf{g}) \geq \frac{B}{4}(1 + \gamma_1^2 N)$ . Thus,

$$\begin{aligned} \Pr[F(\mathbf{g}) = 0] &\leq \Pr[\mathcal{E}^c] + \Pr[\tilde{F}(\mathbf{g}) = 0] \\ &\leq e^{-cN^{1/5}} + \Pr \left[ |\tilde{F}(\mathbf{g}) - \mathbf{E}\tilde{F}(\mathbf{g})| \geq \frac{B(1 + \gamma_1^2 N)}{4} \right] \\ &\leq e^{-cN^{1/5}}, \end{aligned}$$

after adjusting  $c$ .

□

Finally, let us prove the Lipschitz bound.

*Proof of Lemma 4.8.21.* The set  $\mathcal{E}$  is a convex set in  $\mathbf{g}$ . Indeed,  $e^{H_N(\sigma)}$  is a convex function of  $\mathbf{g}$ , so the LHS of the inequality  $\mathcal{E}$  is convex in  $\mathbf{g}$ , whereas the RHS does not depend on  $\mathbf{g}$ , and sublevel sets of convex functions are convex. Furthermore,  $F$  is absolutely continuous (hence differentiable almost everywhere), so to prove  $F$  is Lipschitz on  $\mathcal{E}$  it suffices to bound  $\|\nabla F\|$  on  $\mathcal{E}$ , wherever it is defined.

The easier case is if  $\min(F_1(\mathbf{g}), F_2(\mathbf{g})) < 0$ . In this case,  $F(\mathbf{g}) = 0$  in an open neighborhood of  $\mathbf{g}$ , so  $\nabla F(\mathbf{g}) = 0$  identically. Therefore, for the rest of the proof, assume  $\min(F_1(\mathbf{g}), F_2(\mathbf{g})) \geq 0$ . We will compute the gradient of the  $F_i$ 's, and to simplify the calculation, we will take the gradient with respect to  $\mathbf{g}_p \in \mathbb{R}^{N^p}$  corresponding to the degree- $p$  disorder in  $H_{N,\sim 2}$ .

For  $F_1(\mathbf{g})$ , note that its only dependence on  $\mathbf{g}$  is via  $\log Z_N$ , we have

$$\|\nabla_{g_p} F_1(\mathbf{g})\| = B(1 + \gamma_1^2 N) \|\nabla_{g_p} \log Z_N\| = B(1 + \gamma_1^2 N) \cdot \frac{\gamma_p}{N^{(p-1)/2}} \|\langle \sigma^{\otimes p} \rangle\|_F,$$

and since  $F_1(\mathbf{g}) \geq 0$ , we can apply Lemma 4.8.20 to conclude that

$$\begin{aligned} \|\nabla_{\mathbf{g}} \log Z_N\|^2 &\lesssim \sum_{p \in [p_*] \setminus \{2\}} \gamma_p^2 N^{-(p-1)} \cdot N^{3p/5} \\ &\leq \gamma_1^2 \cdot N^{3/5} + \sum_{p \geq 3} \gamma_p^2 N^{1-2p/5} \\ &\lesssim \gamma_1^2 \cdot N^{3/5} + N^{-1/5} \end{aligned} \quad (4.33)$$

Since  $\gamma_1^2 \leq N^{-4/5}$ , we conclude that  $\|\nabla_{\mathbf{g}} F_1(\mathbf{g})\| \lesssim (1 + \gamma_1^2 N) N^{-1/10}$ , as desired.

Turning now to  $F_2(\mathbf{g})$ , we observe that  $\|\langle \sigma \sigma^\top \rangle\|_{\text{op}} = \langle \langle u, \sigma \rangle^2 \rangle$ , where  $u$  is the top eigenvector of  $\langle \sigma \sigma^\top \rangle$  with  $\|u\|_2 = 1$ . By the envelope theorem, we can evaluate the gradient with  $u$  fixed. For any  $v \in \mathbb{R}^{N^p}$  with  $\|v\|_2 = 1$ , we will upper bound  $\langle v, \nabla_{g_p} \langle \langle u, \sigma \rangle^2 \rangle \rangle$ . Applying the quotient rule yields

$$\begin{aligned} \langle v, \nabla_{g_p} \langle \langle u, \sigma \rangle^2 \rangle \rangle &= \langle \langle u, \sigma \rangle^2 \langle v, \nabla_{g_p} H(\sigma) \rangle \rangle - \langle \langle u, \sigma \rangle^2 \rangle \langle \langle v, \nabla_{g_p} H(\sigma) \rangle \rangle \\ &= \frac{\gamma_p}{N^{(p-1)/2}} (\langle \langle u, \sigma \rangle^2 \langle v, \sigma^{\otimes p} \rangle \rangle - \langle \langle u, \sigma \rangle^2 \rangle \langle \langle v, \sigma^{\otimes p} \rangle \rangle). \end{aligned}$$

Consider the first term  $\langle \langle u, \sigma \rangle^2 \langle v, \sigma^{\otimes p} \rangle \rangle$ . Using Hölder's inequality with  $q = 1 + \log N$  and  $q' = 1 + \frac{1}{\log N}$ , we see

$$\begin{aligned} \langle \langle u, \sigma \rangle^2 \langle v, \sigma^{\otimes p} \rangle \rangle &\leq \langle \langle u, \sigma \rangle^{2q'} \rangle^{1/q'} \langle \langle v, \sigma^{\otimes p} \rangle^q \rangle^{1/q} \\ &\leq N^{1/\log N} \langle \langle u, \sigma \rangle^2 \rangle \langle v^{\otimes q}, \langle \sigma^{\otimes p q} \rangle \rangle^{1/q} \\ &\lesssim (1 + \gamma_1^2 N) \langle v^{\otimes q}, \langle \sigma^{\otimes p q} \rangle \rangle^{1/q} \\ &\leq (1 + \gamma_1^2 N) \cdot N^{3p/10}, \end{aligned}$$

where in the second to last line we have used  $F_2(\mathbf{g}) \geq 0$  to apply the bound  $\langle \langle u, \sigma \rangle^2 \rangle \leq O(1 + \gamma_1^2 N)$ , and in the last line we have used  $F_1(\mathbf{g}) \geq 0$ , along with  $p q \leq O(\log N)$ , to apply Lemma 4.8.20. The same argument upper bounds the contribution of the second term as  $O((1 + \gamma_1^2 N) N^{3p/10})$ . These bounds (aside from the common factor of  $O(1 + \gamma_1^2 N)$ , which we can pull out), exactly match the ones used in the calculation as carried out for  $F_1(\mathbf{g})$  in (4.33). Hence, the same argument ultimately yields  $\|\nabla_{\mathbf{g}} F_2(\mathbf{g})\| \lesssim (1 + \gamma_1^2 N) N^{-1/10}$ , completing the proof.  $\square$

## 4.9 Annealed Glauber dynamics on discrete domains

In this section, we collect the analogous results for weak functional inequalities for Glauber dynamics.

**Definition 4.9.1** (Weak Poincaré for the hypercube). We say  $\pi$  on  $\{\pm 1\}^n$  satisfies a  $(\rho_{\text{PI}}, \varepsilon)$ -weak Poincaré inequality for Glauber dynamics if for all functions  $f$ ,

$$\text{Var}_\pi[f] \leq \frac{1}{\rho_{\text{PI}}} \cdot \mathcal{E}(f, f) + \varepsilon \cdot \text{osc}(f)^2.$$

Similarly, we say  $\pi$  satisfies a  $(C_{\text{MLSI}}, \varepsilon)$ -weak modified log-Sobolev inequality if for all functions  $f$ ,

$$\text{Ent}_\pi[f] \leq \frac{1}{C_{\text{MLSI}}} \cdot \mathcal{E}(f, \log f) + \varepsilon \cdot \text{osc}(\sqrt{f})^2.$$

**Remark 4.9.2.** The above definition is related to the continuous setting by using the discrete gradient, which can be bounded by  $\text{osc}(f)^2$ .

**Remark 4.9.3.** A weak Poincaré inequality with sufficiently good parameters implies a true Poincaré inequality for Glauber dynamics. Indeed, any low conductance cut limits on the region of valid  $(\rho_{\text{PI}}, \varepsilon)$ . Hence, by Cheeger, one can conclude that Glauber satisfies a true Poincaré inequality, with some loss in parameters.

First, we will need a concavity property for the Dirichlet form for Glauber dynamics, which is well-known. We provide a proof for the sake of self-containedness.

**Fact 4.9.4.** Let  $\pi$  be a distribution on  $\{\pm 1\}^n$ , and  $\pi = \mathbf{E}_{z \sim \rho} \pi_z$  a measure decomposition of  $\pi$ . Then  $\mathcal{E}_\pi(f, f) \geq \mathbf{E}_{z \sim \rho} \mathcal{E}_{\pi_z}(f, f)$ .

*Proof.* For Glauber dynamics on the hypercube, we have

$$\begin{aligned} \mathcal{E}_\pi(f, f) &= \frac{1}{n} \sum_{\|x-y\|_1=2} \frac{\pi(x)\pi(y)}{\pi(x) + \pi(y)} (f(x) - f(y))^2 \\ &= \frac{1}{n} \sum_{\|x-y\|_1=2} \frac{\mathbf{E}_{z \sim \rho} \pi_z(x) \mathbf{E}_{z \sim \rho} \pi_z(y)}{\mathbf{E}_{z \sim \rho} \pi_z(x) + \mathbf{E}_{z \sim \rho} \pi_z(y)} (f(x) - f(y))^2 \\ &\geq \mathbf{E}_{z \sim \rho} \mathcal{E}_{\pi_z}(f, f), \end{aligned}$$

where the last line follows from concavity of the map  $(a, b) \mapsto \frac{ab}{a+b}$  for  $a, b > 0$ .  $\square$

The following lemma transfers a true Poincaré inequality on  $\pi$  to a weak Poincaré inequality on  $\pi'$  for Glauber dynamics on the hypercube.

**Lemma 4.9.5.** *Let  $\pi, \pi'$  be distributions on  $\{\pm 1\}^n$  such that  $\pi$  satisfies a  $\rho_{\text{PI}}$ -Poincaré inequality for Glauber dynamics and  $d_{\text{TV}}(\pi, \pi') \leq \delta$ . Then,  $\pi'$  satisfies a weak  $(\rho_{\text{PI}}, \delta(1 + \rho_{\text{PI}}^{-1}))$ -Poincaré inequality for Glauber dynamics.*

*Proof.* By definition, there exists a coupling  $C$  of  $(\pi, \pi')$  such that for  $(x, x') \sim C$ ,  $\Pr[x \neq x'] \leq \delta$ . The main difference in the proof compared to the Langevin case Lemma 4.4.9 is that the Dirichlet form comparison is now bounded in terms of  $\text{osc}(f)^2$  rather than  $\sup \|\nabla f\|^2$ . Indeed, by Fact 4.4.2 we have  $\mathcal{E}_\pi(f, f) = \mathbf{E}_{x \sim \pi} \mathbf{E}_{y \sim p_x} (f(x) - f(y))^2$ , and for any fixed  $x$  the function  $\mathbf{E}_{y \sim p_x} (f(x) - f(y))^2 \leq \text{osc}(f)^2$ , so the coupling allows to conclude that  $\mathcal{E}_{\pi'}(f, f) \geq \mathcal{E}_\pi(f, f) - \delta \cdot \text{osc}(f)^2$ . Similarly, we can deduce that  $\text{Var}_\pi[f] \geq \text{Var}_{\pi'}[f] - \delta \cdot \text{osc}(f)^2$ . Hence, we have

$$\begin{aligned} \mathcal{E}_{\pi'}(f, f) &\geq \mathcal{E}_\pi(f, f) - \delta \cdot \text{osc}(f)^2 \\ &\geq \rho_{\text{PI}} \cdot \text{Var}_\pi[f] - \delta \cdot \text{osc}(f)^2 && (\pi \text{ satisfies PI}) \\ &\geq \rho_{\text{PI}} \cdot \text{Var}_{\pi'}[f] - \delta \cdot \text{osc}(f)^2 (1 + \rho_{\text{PI}}). \end{aligned}$$

□

**Remark 4.9.6.** The above two results also hold more generally if  $P$  is the Markov chain associated to a Doob localization scheme (cf. [CE22, Section 2.3]), such as when  $P$  is Glauber dynamics for a general product domain.

**Lemma 4.9.7.** *Let  $\pi$  be a distribution over  $\{\pm 1\}^n$ , and  $\pi = \mathbf{E}_{z \sim \rho} \pi_z$  a measure decomposition of  $\pi$  such that*

- *for all functions  $f$ ,  $\mathbf{E}_{z \sim \rho} \text{Var}_{\pi_z}[f] \geq C_{\text{Var}} \text{Var}_\pi[f]$ , and*
- *with probability  $1 - \eta$  over  $z \sim \rho$ ,  $\pi_z$  satisfies a  $(\rho_{\text{PI}}, \delta)$ -weak Poincaré inequality with respect to Glauber.*

*Then,  $\pi$  satisfies a  $(\rho_{\text{PI}} C_{\text{Var}}, \frac{\delta + \eta}{C_{\text{Var}}})$ -weak Poincaré inequality.*

*Proof.* The proof is the same as that of Lemma 4.4.10, except in the Langevin case we have  $\mathcal{E}_\pi(f, f) = \mathbf{E}_{z \sim \rho} \mathcal{E}_{\pi_z}(f, f)$ , whereas here we apply Fact 4.9.4 to get the desired inequality. □

Finally, we record the following simple observation connecting weak functional inequalities in discrete domains.

**Fact 4.9.8.** Let  $\pi$  be a distribution on a finite state space  $\Omega$ , and set  $C_\pi = \frac{1-2\pi_{\min}}{\log(1/\pi_{\min}-1)}$ . If  $\pi$  satisfies a  $(\rho_{\text{PI}}, \varepsilon)$ -weak Poincaré inequality, then  $\pi$  also satisfies a  $\left(4\rho_{\text{PI}}C_\pi, \frac{\varepsilon}{C_\pi}\right)$ -weak MLSI and a  $(\rho_{\text{PI}}C_\pi, \frac{\varepsilon}{C_\pi})$ -weak LSI.

*Proof.* For finite state spaces, it is well-known that the LSI of the complete graph Markov chain  $P_{K(\pi)}$  has  $C_{\text{MLSI}} = \frac{1-2\pi_{\min}}{\log(1/\pi_{\min}-1)}$  (see e.g., [DSC96]). Furthermore, observe that  $\mathcal{E}_{P_{K(\pi)}}(f, f) = \text{Var}_\pi[f]$ . Hence,

$$\begin{aligned} \frac{1}{\rho_{\text{PI}}} \mathcal{E}(f, f) &\geq \text{Var}_\pi[f] - \varepsilon \cdot \text{osc}(f)^2 && \text{(Weak PI)} \\ &\geq C_\pi \text{Ent}_\pi[f^2] - \varepsilon \cdot \text{osc}(f)^2, && \text{(LSI of } P_{K(\pi)}) \end{aligned}$$

which establishes the weak LSI. For the weak MLSI, one applies the inequality  $4\mathcal{E}(f, f) \leq \mathcal{E}(f^2, \log f^2)$ , whose proof reduces to checking the two-variable inequality  $4(\sqrt{u} - \sqrt{v})^2 \leq (u - v) \log \frac{u}{v}$  for positive  $u, v$ .  $\square$

## 4.10 Deferred calculations for spherical spin glasses

### The TAP Hamiltonian

In this subsection, we will prove Lemma 4.7.22, which we restate for convenience.

**Lemma 4.7.22.** The law of Hamiltonian  $H_{\text{TAP}} \sim \mu_{\text{TAP}, x, m}$  is described by a Gaussian process  $(H_{\text{TAP}}(\sigma))_{\sigma \in \mathcal{S}_N}$  defined by

$$\begin{aligned} \mathbf{E} H_{\text{TAP}}(\sigma) &= N \xi_t(R(x, \sigma)) - \langle x, v(\sigma) \rangle \cdot \xi'_t(q_x) - \frac{\xi'_t(R(\mathbf{m}, \sigma))}{\gamma'(q_m)} \cdot \langle \mathbf{m}, \sigma \rangle \cdot \left( \theta'(q_m) - \frac{1}{1 - q_m} \right) \\ \frac{1}{N} \text{Cov}(H_{\text{TAP}}(\sigma), H_{\text{TAP}}(\sigma')) &= \xi_t(R(\sigma, \sigma')) - R(\sigma, \sigma') \frac{\xi'_t(R(\mathbf{m}, \sigma)) \xi'_t(R(\mathbf{m}, \sigma'))}{\xi'_t(q_m)} \\ &\quad + \frac{\xi''_t(q_m)}{\gamma'(q_m) \xi'_t(q_m)} \gamma(R(\mathbf{m}, \sigma)) \gamma(R(\mathbf{m}, \sigma')), \end{aligned}$$

where

$$\begin{aligned} v(\sigma) &:= \frac{\xi'_t(R(\mathbf{m}, \sigma))}{\xi'_t(q_m)} \left[ I - \frac{\xi''_t(q_m)}{\gamma'(q_m)} \cdot \frac{\mathbf{m} \mathbf{m}^\top}{N} \right] \sigma \\ \gamma(q) &:= q \cdot \xi'_t(q). \end{aligned}$$

To prove the above, we will need the following formulas for any  $p$ -spin Hamiltonian  $H_N$  with mixture function  $\xi$ .

**Fact 4.10.1.** *For any  $u, v, m \in \mathbb{R}^N$ , we have:*

$$\frac{1}{N} \mathbf{E} \langle u, \nabla H_N(m) \rangle \langle v, \nabla H_N(m) \rangle = R(u, v) \xi'(R(m, m)) + R(m, u) R(m, v) \xi''(R(m, m)).$$

*Proof.* Once we write the derivative as its definition as a limit, the order of the limit and the expectation operator can be swapped by the dominated convergence theorem.

$$\begin{aligned} \frac{1}{N} \mathbf{E} \langle u, \nabla H_N(m) \rangle \langle v, \nabla H_N(m) \rangle &= \frac{1}{N} \mathbf{E} \lim_{\delta, \varepsilon \rightarrow 0} \frac{H_N(m + \delta u) - H_N(m)}{\delta} \cdot \frac{H_N(m + \varepsilon v) - H_N(m)}{\varepsilon} \\ &= \frac{1}{N} \lim_{\delta, \varepsilon \rightarrow 0} \frac{1}{\delta \varepsilon} \mathbf{E} (H_N(m + \delta u) - H_N(m)) \cdot (H_N(m + \varepsilon v) - H_N(m)) \\ &= \lim_{\delta, \varepsilon \rightarrow 0} \frac{1}{\delta \varepsilon} [\xi(R(m + \delta u, m + \varepsilon v)) - \xi(R(m + \delta u, m)) - \xi(R(m, m + \varepsilon v)) + \xi(R(m, m))] \\ &= R(u, v) \xi'(R(m, m)) + R(m, u) R(m, v) \xi''(R(m, m)). \quad \square \end{aligned}$$

**Fact 4.10.2.** *For any  $u, v, m \in \mathbb{R}^N$ , we have:*

$$\frac{1}{N} \mathbf{E} \langle u, \nabla H_N(m) \rangle H_N(v) = R(u, v) \xi'(R(m, v)).$$

The proof of the above is analogous to the proof of Fact 4.10.1, and hence omitted. We now prove Lemma 4.7.22.

*Proof of Lemma 4.7.22.* The distribution of  $H_{\text{TAP}}(\sigma)$  is the same as that of  $H_{N,t}(\sigma) | \mathbf{x}, \nabla \mathcal{F}_{\text{TAP}}(\mathbf{m}) = 0$ . Recall that  $H_{N,t}(\sigma) = N \xi_t(R(\mathbf{x}, \sigma)) + \tilde{H}(\sigma)$  where  $\tilde{H}(\sigma)$  is a centered Gaussian process. Next, observe that conditioning on  $\nabla \mathcal{F}_{\text{TAP}}(\mathbf{m}) = 0$  is the same as conditioning on

$$\nabla \tilde{H}(\mathbf{m}) = -\mathbf{x} \cdot \xi'_t(q_x) - \mathbf{m} \cdot \left( \theta'(q_m) - \frac{1}{1 - q_m} \right). \quad (4.34)$$

Observe that  $(H_{\text{TAP}}(\sigma))_{\sigma \in S_N}$  is a Gaussian process, since it is obtained by conditioning on another Gaussian process satisfying affine constraints. First, observe that we can write

$$H_{\text{TAP}}(\sigma) = N \xi_t(R(\mathbf{x}, \sigma)) + \tilde{H}_{\text{TAP}}(\sigma) \quad (4.35)$$

where  $\tilde{H}_{\text{TAP}}(\sigma) = \tilde{H}(\sigma)|_{\mathbf{x}}, \nabla \mathcal{F}_{\text{TAP}}(\mathbf{m}) = 0$ . To understand the behavior of  $\tilde{H}_{\text{TAP}}(\sigma)$ , we break  $\tilde{H}(\sigma)$  into a sum of two terms: one term for its projection onto the space  $U := \left\{ \left\langle \nabla \tilde{H}(\mathbf{m}), u \right\rangle : u \in \mathbb{R}^N \right\}$ , and the part that is orthogonal to  $U$ , and thus independent of  $\nabla \tilde{H}(\mathbf{m})$ . Concretely, let us write

$$\tilde{H}(\sigma) = \left\langle \nabla \tilde{H}(\mathbf{m}), v(\sigma) \right\rangle + \left( \tilde{H}(\sigma) - \left\langle \nabla \tilde{H}(\mathbf{m}), v(\sigma) \right\rangle \right). \quad (4.36)$$

This is true for any  $v(\sigma)$ , but we have set up the definition such that the second summand is independent of  $\nabla \tilde{H}(\mathbf{m})$ . To verify this, since these two random variables are each mean 0, it suffices to check that for any  $u \in S_N$ ,

$$\mathbb{E} \left\langle \nabla \tilde{H}(\mathbf{m}), u \right\rangle \left( \tilde{H}(\sigma) - \left\langle \nabla \tilde{H}(\mathbf{m}), v(\sigma) \right\rangle \right) = 0.$$

By Facts 4.10.1 and 4.10.2, the left-hand-side of the above is:

$$R(u, \sigma) \xi'_t(R(\mathbf{m}, \sigma)) - R(u, v(\sigma)) \xi'_t(q_m) - R(\mathbf{m}, u) R(\mathbf{m}, v(\sigma)) \xi''(q_m).$$

We would like  $v(\sigma)$  to be such that this is 0 for *all*  $u$ . Setting  $u$  orthogonal to  $\mathbf{m}$  and  $\sigma$  shows that we must have  $v(\sigma)$  in the subspace spanned by  $\mathbf{m}$  and  $\sigma$ .

Suppose that  $v(\sigma) = \alpha \sigma + \beta \mathbf{m}$ . Then, plugging this into the above requires that

$$\begin{aligned} 0 &= R(\sigma, u) \xi'_t(R(\mathbf{m}, \sigma)) - (\alpha R(\sigma, u) + \beta R(\mathbf{m}, u)) \xi'_t(q_m) - R(\mathbf{m}, u) (\alpha R(\mathbf{m}, \sigma) + \beta q_m) \xi''_t(q_m) \\ &= R(\sigma, u) (\xi'_t(R(\mathbf{m}, \sigma)) - \alpha \xi'_t(q_m)) \\ &\quad - R(\mathbf{m}, u) (\beta \xi'_t(q_m) - \alpha R(\mathbf{m}, \sigma) \xi''_t(q_m) - \beta q_m \xi''_t(q_m)). \end{aligned}$$

Since this is true for all  $u$ , each of these two terms must be 0. That is,

$$\alpha = \frac{\xi'_t(R(\mathbf{m}, \sigma))}{\xi'_t(q_m)}$$

and

$$\beta = -\alpha \cdot \frac{R(\mathbf{m}, \sigma) \xi''_t(q_m)}{\xi'_t(q_m) + q_m \xi''_t(q_m)},$$

so

$$\begin{aligned} v(\sigma) &= \frac{\xi'_t(R(\mathbf{m}, \sigma))}{\xi'_t(q_m)} \left( \sigma - \mathbf{m} \cdot \frac{R(\mathbf{m}, \sigma) \xi''_t(q_m)}{\xi'_t(q_m) + q_m \xi''_t(q_m)} \right) \\ &= \frac{\xi'_t(R(\mathbf{m}, \sigma))}{\xi'_t(q_m)} \left( \text{Id} - \frac{R(\mathbf{m}, \sigma) \xi''_t(q_m)}{\xi'_t(q_m) + q_m \xi''_t(q_m)} \cdot \frac{\mathbf{m} \mathbf{m}^\top}{N} \right) \sigma \end{aligned}$$

as defined.

Now returning to (4.36), when we condition on  $x$  and  $\nabla \mathcal{F}_{\text{TAP}}(\mathbf{m}) = 0$ , by plugging in (4.34), we get

$$\begin{aligned} \tilde{H}_{\text{TAP}}(\sigma) = & -\langle x, v(\sigma) \rangle \cdot \xi'_t(q_x) - \langle \mathbf{m}, v(\sigma) \rangle \cdot \left( \theta'(q_m) - \frac{1}{1 - q_m} \right) \\ & + \left( \tilde{H}(\sigma) - \left\langle \nabla \tilde{H}(\mathbf{m}), v(\sigma) \right\rangle \right) \end{aligned} \quad (4.37)$$

We use  $\hat{H}(\sigma)$  to denote the random variable  $\tilde{H}(\sigma) - \left\langle \nabla \tilde{H}(\mathbf{m}), v(\sigma) \right\rangle$ , whose distribution remains unaffected by the conditioning, as this random variable is independent of  $x$  and the event  $\nabla \mathcal{F}_{\text{TAP}}(\mathbf{m}) = 0$ . Since  $\hat{H}(\sigma)$  is centered, our expression for  $\mathbf{E} H_{\text{TAP}}(\sigma)$  follows from (4.35) and (4.37), and the observation that

$$R(\mathbf{m}, v(\sigma)) = \frac{\xi'_t(R(\mathbf{m}, \sigma))}{\gamma'(q_m)} \cdot R(\mathbf{m}, \sigma).$$

It remains to compute  $N^{-1} \mathbf{Cov}(H_{\text{TAP}}(\sigma), H_{\text{TAP}}(\sigma'))$  for any  $\sigma, \sigma' \in S_N$ . Observe that this is equal to  $N^{-1} \mathbf{E} \hat{H}(\sigma) \hat{H}(\sigma')$ . By Facts 4.10.1 and 4.10.2, we have that this is equal to:

$$\begin{aligned} & \xi_t(R(\sigma, \sigma')) - R(v(\sigma), \sigma') \xi'_t(R(\mathbf{m}, \sigma')) - R(v(\sigma'), \sigma) \xi'_t(R(\mathbf{m}, \sigma)) \\ & + R(v(\sigma), v(\sigma')) \xi'_t(q_m) + R(\mathbf{m}, v(\sigma)) R(\mathbf{m}, v(\sigma')) \xi''(q_m). \end{aligned}$$

The formula for the covariance can be obtained from the above by expanding  $v(\sigma)$ . □

Next, we look at the mixture function of these “TAP planted distributions on slices”.

**Corollary 4.7.24.** *For a fixed choice of  $a$  and  $b$ , the Gaussian process  $\left( H_{\text{TAP}}(\sqrt{N} \cdot v(a, b) + r_{a,b} Q\tau) \right)_{\tau \in S_{N-2}}$  is described by the following law.*

- Let  $H_{a,b}$  be a spherical  $p$ -spin Hamiltonian with mixture function  $\xi_{a,b}$  given by:

$$\xi_{a,b}(s) := \xi_t\left(\|v(a, b)\|^2 + r_{a,b}^2 s\right) - \xi_t\left(\|v(a, b)\|^2\right) - s \cdot \frac{r_{a,b}^2 \xi'_t\left(q_m \cdot \left(1 + \frac{a}{\sqrt{N}}\right)\right)^2}{\xi'_t(q_m)}.$$

- Let  $V(a, b) := \xi_t\left(\|v(a, b)\|^2\right) - \|v(a, b)\|^2 \cdot \frac{\xi'_t\left(\left(1 + \frac{a}{\sqrt{N}}\right) q_m\right)^2}{\xi'_t(q_m)} + \frac{\xi''_t(q_m)}{\gamma'(q_m) \xi'_t(q_m)} \cdot \gamma\left(\left(1 + \frac{a}{\sqrt{N}}\right) q_m\right)^2$ .

The law of  $H_{\text{TAP}}(\sqrt{N} \cdot v(a, b) + r_{a,b} Q\tau)$  is the same as that of  $H_{a,b}(\tau) + \sqrt{N} \cdot g_{a,b} + \mathbf{E}_{\mu_{\text{TAP}}} H_{\text{TAP}}(\sqrt{N} \cdot v(a, b) + r_{a,b} Q\tau)$  where  $g_{a,b}$  is a centered Gaussian of variance  $V(a, b)$  independent of  $H_{a,b}$ .



*Proof.* Let  $\tau, \tau' \in S_{N-2}$ , and

$$\sigma = \sqrt{N} \cdot v(a, b) + r_{a,b} Q \tau \text{ and } \sigma' = \sqrt{N} \cdot v(a, b) + r_{a,b} Q \tau'.$$

Recall from Lemma 4.7.22 that

$$\begin{aligned} N^{-1} \mathbf{Cov}(H_{\text{TAP}}(\sigma), H_{\text{TAP}}(\sigma')) \\ = \xi_t(R(\sigma, \sigma')) - R(\sigma, \sigma') \frac{\xi'_t(R(\mathbf{m}, \sigma)) \xi'_t(R(\mathbf{m}, \sigma'))}{\xi'_t(q_m)} + \frac{\xi''_t(q_m)}{\gamma'(q_m) \xi'_t(q_m)} \gamma(R(\mathbf{m}, \sigma)) \gamma(R(\mathbf{m}, \sigma')). \end{aligned}$$

By the definition of  $\sigma$  and  $\sigma'$ , we have  $R(\mathbf{m}, \sigma) = R(\mathbf{m}, \sigma') = R(\mathbf{m}, \sqrt{N} \cdot v(a, b)) = \left(1 + \frac{a}{\sqrt{N}}\right) q_m$ , and  $R(\sigma, \sigma') = \|v(a, b)\|^2 + r_{a,b}^2 R(\tau, \tau')$ , since  $Q$  is an isometry. As a result,

$$\begin{aligned} N^{-1} \mathbf{Cov}(H_{\text{TAP}}(\sigma), H_{\text{TAP}}(\sigma')) \\ = \xi_t(\|v(a, b)\|^2 + r_{a,b}^2 R(\tau, \tau')) - \left(\|v(a, b)\|^2 + r_{a,b}^2 R(\tau, \tau')\right) \frac{\xi'_t\left(\left(1 + \frac{a}{\sqrt{N}}\right) q_m\right)^2}{\xi'_t(q_m)} \\ + \frac{\xi''_t(q_m)}{\gamma'(q_m) \xi'_t(q_m)} \gamma\left(\left(1 + \frac{a}{\sqrt{N}}\right) q_m\right)^2. \end{aligned}$$

This may be written as

$$N^{-1} \mathbf{Cov}(H_{\text{TAP}}(\sigma), H_{\text{TAP}}(\sigma')) = \xi_{a,b}(R(\tau, \tau')) + V(a, b).$$

This implies that  $H_{\text{TAP}}(\sigma)$  is equal to  $H_{a,b}(\tau) + g_{a,b}$  for some Gaussian process  $(H_{a,b}(\tau))_{\tau \in S_{N-2}}$ , where  $g_{a,b}$  is a centered Gaussian of variance  $V(a, b)$ . To complete the proof, we must show that the correlation structure of  $H_{a,b}$  can be achieved by a  $p$ -spin model with mixture function  $\xi_{a,b}$ . To do this, it suffices to show that  $\xi_{a,b}$  is indeed a valid mixture function, in that  $\xi_{a,b}^{(p)}(0) \geq 0$  for all  $p \geq 1$ , and  $\xi_{a,b}(0) = 0$ . The latter of these is clearly true by construction. The former is easily seen to be true for  $p \geq 2$ , since for such  $p$ ,

$$\xi_{a,b}^{(p)}(0) = r_{a,b}^{2p} \xi_t^{(p)}(\|v(a, b)\|^2) \geq 0$$

since  $\xi_t$  is a valid mixture function. For  $p = 1$ ,

$$\xi'_{a,b}(0) = r_{a,b}^2 \cdot \left( \xi'_t(\|v(a, b)\|^2) - \frac{\xi'_t\left(q_m \left(1 + \frac{a}{\sqrt{N}}\right)\right)^2}{\xi'_t(q_m)} \right)$$

$$\begin{aligned}
&\stackrel{(4.9)}{\geq} r_{a,b}^2 \cdot \left( \xi'_t \left( q_m \left( 1 + \frac{a}{\sqrt{N}} \right)^2 \right) - \frac{\xi'_t \left( q_m \left( 1 + \frac{a}{\sqrt{N}} \right) \right)^2}{\xi'_t(q_m)} \right) \\
&= \frac{r_{a,b}^2}{\xi'_t(q_m)} \cdot \left( \xi'_t \left( q_m \left( 1 + \frac{a}{\sqrt{N}} \right)^2 \right) \xi'_t(q_m) - \xi'_t \left( q_m \left( 1 + \frac{a}{\sqrt{N}} \right) \right)^2 \right) \geq 0.
\end{aligned}$$

In the first inequality above, we use the fact that  $\xi'_t$  is non-decreasing. The final inequality is an application of Cauchy-Schwarz.  $\square$

## Understanding concentration around the codimension-2 slice

Next, we bound the variance of  $g_{a,b} - g_{0,0}$ .

**Lemma 4.7.31.** *For every constant  $\iota > 0$ , there is a constant  $c$  such that with probability  $1 - e^{-cN}$ , for all  $a, b$ , we have  $|g_{a,b} - g_{0,0}| \leq \iota \frac{a^2 + b^2}{\sqrt{N}}$ .*

*Proof of Lemma 4.7.31.* The strategy is to prove that for any  $a, b, a', b' \in \mathbb{R}$  of magnitude  $\ll N^{1/4}$ ,  $g_{a,b} - g_{a',b'}$  is a Gaussian of variance  $O(\|v(a, b) - v(a', b')\|^4) = O\left(\frac{(a-a')^4 + (b-b')^4}{N^2}\right)$ . The desideratum then immediately follows by applying Slepian's lemma on  $(|g_{a,b} - g_{0,0}|)_{a,b}$  comparing it to the Gaussian process  $\langle G, (v(a, b) - v(0, 0))(v(a, b) - v(0, 0))^T \rangle$  for a standard Gaussian matrix  $G$ .

We carry out the calculation for  $a', b' = 0$ ; the general case follows similarly. We have  $g_{a,b} = N^{-1/2} \left( H_{\text{TAP}}(\sqrt{N} \cdot v(a, b)) - \mathbf{E} H_{\text{TAP}}(\sqrt{N} \cdot v(a, b)) \right)$ , and  $g_{0,0} = N^{-1/2} (H_{\text{TAP}}(\mathbf{m}) - \mathbf{E} H_{\text{TAP}}(\mathbf{m}))$ . Clearly,  $g_{a,b} - g_{0,0}$  is a centered Gaussian process. As in the proof of Lemma 4.7.22, we have that the distribution of  $H_{\text{TAP}}(\sigma) - \mathbf{E} H_{\text{TAP}}(\sigma)$  is the same as that of

$$\tilde{H}(\sigma) - \langle \nabla \tilde{H}(\mathbf{m}), v(\sigma) \rangle,$$

where  $\tilde{H}$  is a Hamiltonian distributed according to the mixture function  $\xi_t$ . Thus, the distribution of  $g_{a,b} - g_{0,0}$  is:

$$\tilde{H}(u(a, b)) - \langle \nabla \tilde{H}(\mathbf{m}), v(u(a, b)) \rangle - \tilde{H}(\mathbf{m}) + \langle \nabla \tilde{H}(\mathbf{m}), v(\mathbf{m}) \rangle.$$

For brevity, we denote  $v(a, b)$  as  $\mathbf{m} + \varepsilon$ . We express  $\tilde{H}(\mathbf{m} + \varepsilon)$  in its Taylor expansion, and we get:

$$\sqrt{N}(g_{a,b} - g_{0,0}) = \sum_{i \geq 1} \frac{1}{i!} \langle D_i \tilde{H}(\mathbf{m}), \varepsilon^{\otimes i} \rangle - \langle \nabla \tilde{H}(\mathbf{m}), v(\mathbf{m} + \varepsilon) - v(\mathbf{m}) \rangle. \quad (4.38)$$

Expanding out  $v(\mathbf{m} + \varepsilon) - v(\mathbf{m})$  ultimately yields:

$$v(\mathbf{m} + \varepsilon) - v(\mathbf{m}) = \varepsilon + R(\mathbf{m}, \varepsilon) \varepsilon \frac{\xi_t''(q_m)}{\gamma'(q_m)} - \frac{R(\mathbf{m}, \varepsilon)^2 \xi_t''(q_m)^2}{\xi_t'(q_m) \gamma'(q_m)} \mathbf{m}.$$

Plugging in the above into Eq. (4.38) gives:

$$\begin{aligned} \sqrt{N}(g_{a,b} - g_{0,0}) &= \sum_{i \geq 2} \frac{1}{i!} \left\langle D_i \tilde{H}(\mathbf{m}), \varepsilon^{\otimes i} \right\rangle \\ &\quad - \left\langle \nabla \tilde{H}(\mathbf{m}), \varepsilon \right\rangle R(\mathbf{m}, \varepsilon) \frac{\xi_t''(q_m)}{\gamma'(q_m)} - \left\langle \nabla \tilde{H}(\mathbf{m}), \mathbf{m} \right\rangle \frac{R(\mathbf{m}, \varepsilon)^2 \xi_t''(q_m)^2}{\xi_t'(q_m) \gamma'(q_m)} \end{aligned}$$

We have an explicit expression for  $\varepsilon$ :

$$\varepsilon = \frac{a q_m - b q_x^2}{\sqrt{N}(q_m - q_x^2)} \mathbf{m} + \frac{q_m q_x}{q_m - q_x^2} \left( \frac{b - a}{\sqrt{N}} \right) \mathbf{x}.$$

This explicit expression can be used to obtain the following bounds on the variances of the above terms:

$$\begin{aligned} \text{Var} \left[ \frac{1}{i!} \left\langle D_i \tilde{H}(\mathbf{m}), \varepsilon^{\otimes i} \right\rangle \right] &\leq \frac{O(a^{2i} + b^{2i})}{N^{i-1}} \\ \text{Var} \left[ \left\langle \nabla \tilde{H}(\mathbf{m}), \varepsilon \right\rangle R(\mathbf{m}, \varepsilon) \frac{\xi_t''(q_m)}{\gamma'(q_m)} \right] &= \frac{O(a^4 + b^4)}{N} \\ \text{Var} \left[ \left\langle \nabla \tilde{H}(\mathbf{m}), \mathbf{m} \right\rangle \frac{R(\mathbf{m}, \varepsilon)^2 \xi_t''(q_m)^2}{\xi_t'(q_m) \gamma'(q_m)} \right] &\leq \frac{O(a^4 + b^4)}{N} \end{aligned}$$

The expression for  $\sqrt{N}(g_{a,b} - g_{0,0})$  only involves a constant number of terms, and since the first term enumerates over  $i \geq 2$ , and since  $|a|, |b| \leq \sqrt{N}$ , we have an overall bound of  $\frac{O(a^4 + b^4)}{N}$ . Dividing by  $\sqrt{N}$  gives the desired variance bound.  $\square$

**Lemma 4.7.29.**  $\nabla \widehat{E}_{a,b} \Big|_{(a,b)=(0,0)} = 0.$

*Proof.* Recall

$$\widehat{E}_{a,b} = \frac{1}{2} \underbrace{\left( \log r_{a,b}^2 - \xi_t(\|v(a,b)\|^2) - r_{a,b}^2 \cdot \frac{\xi_t' \left( q_m \left( 1 + \frac{a}{\sqrt{N}} \right) \right)^2}{\xi_t'(q_m)} \right)}_{(I)}$$

$$\begin{aligned}
& + \underbrace{\xi_t \left( q_x \left( 1 + \frac{b}{\sqrt{N}} \right) \right) + \frac{\gamma \left( q_m \left( 1 + \frac{a}{\sqrt{N}} \right) \right)}{\gamma'(q_m)} \cdot \left( (1 - q_m) \xi_t''(q_m) + \frac{1}{1 - q_m} \right)}_{\text{(II)}} \\
& - \underbrace{\frac{\gamma(q_x)}{\xi_t'(q_m)} \cdot \xi_t' \left( q_m \left( 1 + \frac{a}{\sqrt{N}} \right) \right) \cdot \left( \left( 1 + \frac{b}{\sqrt{N}} \right) - q_m \cdot \frac{\xi_t''(q_m)}{\gamma'(q_m)} \cdot \left( 1 + \frac{a}{\sqrt{N}} \right) \right)}_{\text{(III)}}.
\end{aligned}$$

Because

$$\|v(a, b)\|^2 = q_m \left( 1 + \frac{a}{\sqrt{N}} \right)^2 + \frac{q_m q_x}{q_m - q_x^2} \cdot \left( \frac{a - b}{\sqrt{N}} \right)^2,$$

we have

$$\nabla \|v(a, b)\|^2|_{(a,b)=(0,0)} = -\nabla r_{a,b}^2|_{(a,b)=(0,0)} = \left( \frac{2q_m}{\sqrt{N}}, 0 \right).$$

We also have  $r_{0,0}^2 = 1 - q_m$  and  $\|v(0, 0)\|^2 = q_m$ . Let us start by computing the derivative with respect to  $a$ . We have

$$\begin{aligned}
\sqrt{N} \cdot \partial_a \text{(III)}|_{(a,b)=(0,0)} &= \xi_t''(q_m) \cdot q_m \cdot \left( 1 - \frac{q_m \xi_t''(q_m)}{\gamma'(q_m)} \right) + \xi_t'(q_m) \cdot \left( -q_m \cdot \frac{\xi_t''(q_m)}{\gamma'(q_m)} \right) \\
&= \frac{q_m \xi_t''(q_m)}{\gamma'(q_m)} (\gamma'(q_m) - q_m \xi_t''(q_m) - \xi_t'(q_m)) = 0.
\end{aligned}$$

Next,

$$\begin{aligned}
& \sqrt{N} \cdot \partial_a \text{(I)}|_{(a,b)=(0,0)} \\
&= \frac{1}{r_{0,0}^2} \cdot (-2q_m) - \xi_t'(\|v(0, 0)\|^2) \cdot (2q_m) - (-2q_m) \cdot \left( \frac{\xi_t'(q_m)^2}{\xi_t'(q_m)} \right) - r_{0,0}^2 \cdot \frac{2\xi_t'(q_m) \cdot \xi_t''(q_m) \cdot q_m}{\xi_t'(q_m)} \\
&= \frac{-2q_m}{1 - q_m} - 2q_m \xi_t'(q_m) + 2q_m \xi_t'(q_m) - 2q_m(1 - q_m) \xi_t''(q_m) \\
&= \frac{-2q_m}{1 - q_m} - 2q_m(1 - q_m) \xi_t''(q_m).
\end{aligned}$$

Finally,

$$\begin{aligned}
& \sqrt{N} \cdot \partial_a \text{(II)}|_{(a,b)=(0,0)} \\
&= \frac{\gamma'(q_m) \cdot q_m}{\gamma'(q_m)} \cdot \left( (1 - q_m) \xi_t''(q_m) + \frac{1}{1 - q_m} \right)
\end{aligned}$$

$$= -\frac{1}{2} \cdot \sqrt{N} \cdot \partial_a(\text{I})|_{(a,b)=(0,0)}$$

as desired. The derivative with respect to  $b$  is much simpler, since the derivative of  $r_{a,b}^2$  with respect to  $b$  is 0 at  $(0,0)$ . Consequently,  $\partial_b(\text{I})|_{(a,b)=(0,0)} = 0$ ,  $\partial_b(\text{II})|_{(a,b)=(0,0)} = q_x \xi'_t(q_x) = \gamma(q_x)$ , and  $\partial_b(\text{III})|_{(a,b)=(0,0)} = \xi'_t(q_m)$ , completing the proof.  $\square$

**Lemma 4.7.30.** *There exist constants  $\eta, \varepsilon > 0$  such that for all  $|a|, |b| \leq \varepsilon \sqrt{N}$ ,  $N \nabla^2 \widehat{E}_{a,b} \leq -\eta \text{Id}$ .*

*Proof.* For ease of notation, define  $\widetilde{E}_{a,b} = \widehat{E}_{\sqrt{N}a, \sqrt{N}b}$ ,  $\widetilde{r}_{a,b} = r_{\sqrt{N}a, \sqrt{N}b}$ , and  $\widetilde{v}(a, b) = v(\sqrt{N}a, \sqrt{N}b)$ . As in the previous lemma, recall

$$\begin{aligned} \widetilde{E}_{a,b} = & \frac{1}{2} \left( \log \widetilde{r}_{a,b}^2 - \xi_t(\|\widetilde{v}(a, b)\|^2) - \widetilde{r}_{a,b}^2 \cdot \frac{\xi'_t(q_m(1+a))^2}{\xi'_t(q_m)} \right) \\ & + \xi_t(q_x(1+b)) + \frac{\gamma(q_m(1+a))}{\gamma'(q_m)} \cdot \left( (1-q_m)\xi''_t(q_m) + \frac{1}{1-q_m} \right) \\ & - \frac{\gamma(q_x)}{\xi'_t(q_m)} \cdot \xi'_t(q_m(1+a)) \cdot \left( (1+b) - q_m \cdot \frac{\xi''_t(q_m)}{\gamma'(q_m)} \cdot (1+a) \right). \end{aligned}$$

Because the Hessian is Lipschitz in all the parameters involved, it suffices to prove the negative definiteness of the Hessian at  $(0,0)$ , under the assumption that  $q_m = q_x = q$ , where  $q$  (formerly denoted  $q_*(t)$ ) satisfies  $\xi'_t(q) = \frac{q}{1-q}$ . Under these constraints, we have  $\gamma(q) = q\xi'_t(q) = \frac{q^2}{1-q}$  and  $\gamma'(q) = \xi'_t(q) + q\xi''_t(q) = \frac{q}{1-q} (1 + (1-q)\xi''_t(q))$ .  $\widetilde{E}_{a,b}$  simplifies as

$$\begin{aligned} \widetilde{E}_{a,b} = & \underbrace{\frac{1}{2} \left( \log \widetilde{r}_{a,b}^2 - \xi_t(\|\widetilde{v}(a, b)\|^2) - \widetilde{r}_{a,b}^2 \cdot \frac{\xi'_t(q(1+a))^2}{\xi'_t(q)} \right)}_{\text{(I)}} \\ & + \underbrace{\xi_t(q(1+b)) + \frac{\gamma(q(1+a))}{q} \cdot \frac{1 + (1-q)^2 \xi''_t(q)}{1 + (1-q)\xi''_t(q)}}_{\text{(II)}} \\ & - \underbrace{q(1+b) \cdot \xi'_t(q(1+a)) + q^2(1+a) \cdot \frac{\xi''_t(q)}{\gamma'(q)}}_{\text{(III)}}. \end{aligned}$$

We have  $\partial_b^2(\text{III})|_{(0,0)} = 0$ , and

$$\partial_b^2(\text{II})|_{(0,0)} = \xi''_t(q) \cdot q^2.$$

We have that  $\partial_b \tilde{r}_{a,b}^2|_{(0,0)} = 0$ , and  $\partial_b^2 \tilde{r}_{a,b}^2|_{(0,0)} = \frac{-2q^2}{1-q}$ . Consequently,

$$\begin{aligned} \partial_b^2(\text{I})|_{(0,0)} &= \frac{1}{2} \left( \frac{1}{r_{0,0}^2} \cdot \partial_b^2 \tilde{r}_{a,b}^2|_{(0,0)} - \xi'_t(\|v(0,0)\|^2) \cdot \partial_b^2 \|\tilde{v}(a,b)\|^2|_{(0,0)} - \partial_b^2 \tilde{r}_{a,b}^2|_{(0,0)} \cdot \xi'_t(q) \right) \\ &= \frac{-q^2}{(1-q)^2}. \end{aligned}$$

It follows that

$$\partial_b^2 \widehat{E}_{a,b}|_{(0,0)} = \xi_t''(q) \cdot q^2 - \frac{q^2}{(1-q)^2}.$$

Similarly, we have  $\partial_a \partial_b(\text{II})|_{(0,0)} = 0$ , and

$$\partial_a \partial_b(\text{III})|_{(0,0)} = \xi_t''(q) \cdot q^2.$$

We have that  $\partial_a \partial_b \tilde{r}_{a,b}^2|_{(0,0)} = \frac{-2q^2}{1-q}$ . Consequently,

$$\begin{aligned} \partial_a \partial_b(\text{I})|_{(0,0)} &= \frac{1}{2} \left( \frac{1}{r_{0,0}^2} \cdot \partial_a \partial_b \tilde{r}_{a,b}^2|_{(0,0)} - \xi'_t(\|v(0,0)\|^2) \cdot \partial_a \partial_b \|\tilde{v}(a,b)\|^2|_{(0,0)} - \partial_a \partial_b \tilde{r}_{a,b}^2|_{(0,0)} \cdot \xi'_t(q) \right) \\ &= \frac{-q^2}{(1-q)^2}. \end{aligned}$$

It follows that

$$\partial_a \partial_b \widehat{E}_{a,b}|_{(0,0)} = \xi_t''(q) \cdot q^2 - \frac{q^2}{(1-q)^2}.$$

It remains to compute the second derivative with respect to  $a$ . We have

$$\partial_a^2(\text{III})|_{(0,0)} = q^3 \xi_t'''(q).$$

We also have

$$\begin{aligned} \partial_a^2(\text{II})|_{(0,0)} &= q \gamma''(q) \cdot \frac{1 + (1-q)^2 \xi_t''(q)}{1 + (1-q) \xi_t''(q)} \\ &= q \cdot \frac{1 + (1-q)^2 \xi_t''(q)}{1 + (1-q) \xi_t''(q)} \cdot (2\xi_t''(q) + q \xi_t'''(q)) \\ &\leq q^2 \xi_t'''(q) + 2q \xi_t''(q). \end{aligned}$$

We have  $\partial_a^2 \tilde{r}_{a,b}^2|_{(0,0)} = -\left(2q + \frac{2q^2}{1-q}\right) = -\frac{2q}{1-q}$  and  $\partial_a \tilde{r}_{a,b}^2|_{(0,0)} = -2q$ . Finally,

$$\partial_a^2(\text{I})|_{(0,0)} = \frac{1}{r_{0,0}^2} \cdot \frac{-2q}{1-q} - \frac{1}{r_{0,0}^4} \cdot (-2q)^2 - \xi'_t(q) \cdot \frac{-2q}{1-q} - \xi_t''(q) \cdot (-2q)^2 - \frac{-2q}{1-q} \cdot \xi'_t(q)$$

$$\begin{aligned}
& -2 \cdot (-2q) \cdot \frac{2\xi'_t(q)\xi''_t(q)q}{\xi'_t(q)} - (1-q) \cdot \frac{2q^2(\xi'_t(q)\xi'''_t(q) + \xi''_t(q)^2)}{\xi'_t(q)} \\
& = \frac{-2q}{(1-q)^2} - \frac{4q^2}{(1-q)^2} - 4q^2\xi''_t(q) + 8q^2\xi''_t(q) - 2q^2(1-q)\xi'''_t(q) - 2q^2(1-q) \cdot \frac{\xi''_t(q)^2}{\xi'_t(q)} \\
& = \frac{-2q(2q+1)}{(1-q)^2} + 4q^2\xi''_t(q) - 2q^2(1-q)\xi'''_t(q) - 2q(1-q)^2\xi''_t(q)^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\partial_a^2 \widehat{E}_{a,b} \Big|_{(0,0)} & \leq -q^3\xi'''_t(q) + q^2\xi'''_t(q) + 2q\xi''_t(q) - \frac{q(2q+1)}{(1-q)^2} \\
& \quad + 2q^2\xi''_t(q) - q^2(1-q)\xi'''_t(q) - q(1-q)^2\xi''_t(q)^2 \\
& = \frac{-q(2q+1)}{(1-q)^2} + 2q\xi''_t(q) + 2q^2\xi''_t(q) - q(1-q)^2\xi''_t(q)^2.
\end{aligned}$$

To conclude, let us check that the Hessian is negative definite. Because  $\xi''_t(q) \cdot q^2 - \frac{q^2}{(1-q)^2} < 0$  by the SL condition (SL), it suffices to check that

$$\partial_a^2 \widehat{E}_{a,b} \Big|_{(0,0)} < q^2\xi''_t(q) - \frac{q^2}{(1-q)^2}.$$

This is true if and only if

$$q(1-q)^2\xi''_t(q)^2 - q(q+2)\xi''_t(q) + \frac{q(q+1)}{(1-q)^2} > 0.$$

It is not difficult to see that this is true if  $\xi''_t(q)$  is less than the smaller root of the above quadratic, which is equal to

$$\frac{(q+2) - \sqrt{(q+2)^2 - 4(q+1)}}{2(1-q)^2} = \frac{1}{(1-q)^2}.$$

This is true by the SL condition (SL), concluding the proof.  $\square$

Next, we shall prove Lemma 4.7.33. Recall the definition

$$\text{Error}_{a,b}^{(2)} = \frac{N\xi''_{a,b}(0)}{4} - \frac{1}{2} \log \det \left( (1 + \xi''_{a,b}(0))\text{Id} - \nabla^2 H_{a,b}(0) \right),$$

where  $\nabla^2 H_{a,b}(0)$  is equal to the restriction of  $r_{a,b}^2 \cdot \nabla^2 H_{\text{TAP}}(\sqrt{N} \cdot v(a, b))$  restricted to the codimension-2 subspace orthogonal to  $\mathbf{m}$  and  $\mathbf{x}$ .

**Lemma 4.7.33.** *For any sufficiently small  $\iota > 0$ , with probability at least  $1 - e^{-cN}$ ,  $\left| \text{Error}_{a,b}^{(2)} - \text{Error}_{0,0}^{(2)} \right| = O(1)$  for all  $|a|, |b| < \iota N^{1/4}$ .*

Let us start by computing the correlation structure of the random matrices  $\nabla^2 H_{a,b}(0)$ . Note that  $\nabla^2 H_{a,b}(0)$  is an  $(N-2)$ -dimensional GOE matrix scaled by  $\sqrt{\xi''_{a,b}(0)}$ .

**Fact 4.10.3.** *For  $\sigma^1, \sigma^2$ ,*

$$\begin{aligned} \frac{1}{N} \mathbf{E} \langle \nabla^2 H_{\text{TAP}}(\sigma^1), u^1 \otimes u^2 \rangle \langle \nabla^2 H_{\text{TAP}}(\sigma^2), v^1 \otimes v^2 \rangle \\ = \xi_t'' (R(\sigma^1, \sigma^2)) \cdot (R(u^1, v^2) \cdot R(u^2, v^1) + R(u^1, v^1) \cdot R(u^2, v^2)). \end{aligned}$$

*In particular,*

$$\frac{1}{N} \mathbf{E} \langle \nabla^2 H_{\text{TAP}}(\sigma^1), \nabla^2 H_{\text{TAP}}(\sigma^2) \rangle = \xi_t'' (R(\sigma^1, \sigma^2)).$$

The above follows from calculations similar to those involved in the proofs of Facts 4.10.1 and 4.10.2; we omit the details.

*Proof of Lemma 4.7.33.* We shall prove the statement for a fixed  $a, b$ ; a union bound over  $a, b$  implies the boundedness for all  $a, b \leq \iota N^{1/4}$ .

Recalling that  $\nabla^2 H_{a,b}(0) = r_{a,b}^2 \nabla^2 H_{\text{TAP}}(\sqrt{N} \cdot v(a, b))$  is a GOE matrix scaled by  $\sqrt{\xi''_{a,b}(0)}$ . By Fact 4.10.3,

$$\begin{aligned} \frac{1}{N} \mathbf{E} \langle \nabla^2 H_{a,b}(0), \nabla^2 H_{0,0}(0) \rangle &= r_{a,b}^2 \cdot r_{0,0}^2 \cdot \frac{1}{N} \mathbf{E} \langle \nabla^2 H_{\text{TAP}}(\sqrt{N} \cdot v(a, b)), \nabla^2 H_{\text{TAP}}(\sqrt{N} \cdot v(0, 0)) \rangle \\ &= r_{a,b}^2 \cdot r_{0,0}^2 \cdot \xi_t'' (\langle v(a, b), v(0, 0) \rangle). \end{aligned}$$

For comparison, we have

$$\frac{1}{N} \mathbf{E} \|\nabla^2 H_{a,b}(0)\|_F^2 = \xi''_{a,b}(0) = r_{a,b}^4 \cdot \xi_t'' (\|v(a, b)\|^2).$$

For succinctness of notation, let

$$\begin{pmatrix} \alpha_1 & \rho \\ \rho & \alpha_2 \end{pmatrix} = \begin{pmatrix} r_{0,0}^4 \xi_t'' (\|v(0, 0)\|^2) & r_{a,b}^2 r_{0,0}^2 \xi_t'' (\langle v(a, b), v(0, 0) \rangle) \\ r_{a,b}^2 r_{0,0}^2 \xi_t'' (\langle v(a, b), v(0, 0) \rangle) & r_{a,b}^4 \xi_t'' (\|v(a, b)\|^2) \end{pmatrix}$$

be the covariance structure of the scaled GOE matrices  $\nabla^2 H_{0,0}(0)$  and  $\nabla^2 H_{a,b}(0)$ . It is not difficult to see that  $\alpha_2 = \alpha_1 + O\left(\frac{a^2 + b^2}{N}\right)$ , and  $\rho$  is between  $\alpha_1$  and  $\alpha_2$ . Also note that



$\alpha_1 = \xi''_{0,0}(0)$  and  $\alpha_2 = \xi''_{a,b}(0)$ . Then, for some choice of GOE matrices  $G$  and  $\tilde{G}$ , we may write

$$\begin{aligned}\nabla^2 H_{0,0}(0) &= \sqrt{\alpha_1} G \\ \nabla^2 H_{a,b}(0) &= \frac{\rho}{\sqrt{\alpha_1}} G + \sqrt{\alpha_2 - \frac{\rho^2}{\alpha_1}} \tilde{G}.\end{aligned}$$

We thus have

$$\begin{aligned}2 \left( \text{Error}_{0,0}^{(2)} - \text{Error}_{a,b}^{(2)} \right) &= \frac{N\alpha_1}{2} - \frac{N\alpha_2}{2} - \log \det \left( \underbrace{(1 + \alpha_1) \text{Id} - \sqrt{\alpha_1} G}_{M_1} \right) \\ &\quad + \log \det \left( (1 + \alpha_2) \text{Id} - \frac{\rho}{\sqrt{\alpha_1}} G - \sqrt{\alpha_2 - \frac{\rho^2}{\alpha_1}} \tilde{G} \right).\end{aligned}$$

We may write the matrix inside the final log det as

$$\begin{aligned}(1 + \alpha_2) \text{Id} - \frac{\rho}{\sqrt{\alpha_1}} G - \sqrt{\alpha_2 - \frac{\rho^2}{\alpha_1}} \tilde{G} \\ = \underbrace{((1 + \alpha_1) \text{Id} - \sqrt{\alpha_1} G)}_{M_2} + \underbrace{(\alpha_2 - \alpha_1) \text{Id} - \left( \frac{\rho}{\sqrt{\alpha_1}} - \sqrt{\alpha_1} \right) G - \sqrt{\alpha_2 - \frac{\rho^2}{\alpha_1}} \tilde{G}}_{M_3}.\end{aligned}$$

The difference of the two log det terms is thus equal to

$$\log \det \left( \text{Id} + \underbrace{M_1^{-1/2} M_2 M_1^{-1/2} + M_1^{-1/2} M_3 M_1^{-1/2}}_M \right).$$

Observe that  $M_3$  is a scaled GOE matrix independent of  $M_1$  (and  $M_2$ ). Taylor expanding the above, we shall control the trace and Frobenius norm of  $M$ . It may be verified that the higher-order terms, corresponding to higher Schatten norms, are  $O(1)$ . To control the trace and Frobenius norm, we shall essentially control their values in expectation. Standard concentration arguments for GOE matrices, along the lines of Lemma 4.8.6 using [GZ00, Lemma 1.2(b) and Corollary 1.6(b)], allow us to assume (with probability  $1 - e^{-cN}$ ) that the eigenvalues of  $G$  are distributed according to the semicircular distribution up to some

small Wasserstein perturbation. That is, with probability  $1 - e^{-cN}$ , denoting by  $\lambda_i(G)$  the eigenvalues of  $G$ ,

$$\begin{aligned} \text{Tr } M_1^{-1/2} M_2 M_1^{-1/2} &= \sum_{1 \leq i \leq N} \frac{(\alpha_2 - \alpha_1) - \left(\frac{\rho}{\sqrt{\alpha_1}} - \sqrt{\alpha_1}\right) \lambda_i(G)}{(1 + \alpha_1) - \sqrt{\alpha_1} \lambda_i(G)} \\ &= N \cdot \int \frac{(\alpha_2 - \alpha_1) - \left(\frac{\rho}{\sqrt{\alpha_1}} - \sqrt{\alpha_1}\right) u}{(1 + \alpha_1) - \sqrt{\alpha_1} u} d\mu_{\text{sc}}(u) + O(1) \\ &= N(\alpha_2 - \rho) + O(1), \end{aligned}$$

where the final equality follows from the standard semicircle integral  $\int \frac{1}{x-u} d\mu_{\text{sc}}(u) = \frac{1}{2} \left( x - \sqrt{x^2 - 4} \right)$ . On the other hand, because  $\tilde{G}$  is independent of  $G$ ,

$$\text{Tr } M_1^{-1/2} M_3 M_1^{-1/2} = O(1)$$

with probability  $1 - e^{-cN}$ . Let us next control the Frobenius norms of these matrices. Again, because  $\tilde{G}$  is independent of  $G$ , with very high probability,

$$\|M\|_F^2 = O(1) + \|M_1^{-1/2} M_2 M_1^{-1/2}\|_F^2 + \|M_1^{-1/2} M_3 M_1^{-1/2}\|_F^2.$$

The first squared Frobenius norm is equal to

$$\sum \left( \frac{(\alpha_2 - \alpha_1) - \left(\frac{\rho}{\sqrt{\alpha_1}} - \sqrt{\alpha_1}\right) \lambda_i(G)}{(1 + \alpha_1) - \sqrt{\alpha_1} \lambda_i(G)} \right)^2.$$

Let  $\iota$  such that  $(1 + \alpha_1) - (2 + \iota)\sqrt{\alpha_1} > \iota$  (this uses strict replica symmetry). Then, with probability  $1 - e^{-cN}$ ,  $|\lambda_i(G)| \leq 2 + \iota$  for all  $i$ . Conditioned on this event happening, and recalling that  $\alpha_2 - \alpha_1 = O\left(\frac{a^2 + b^2}{N}\right)$ , the above is  $\frac{O(a^2 + b^2)^2}{N}$ . This is  $O(1)$  for all choices of  $a, b \leq \iota N^{1/4}$ .

We must next control the squared Frobenius norm of  $M_1^{-1/2} M_3 M_1^{-1/2}$ . Let us condition on a typical realization of  $M_1$ : all its eigenvalues are smaller than  $2 + \iota$  in magnitude, and the empirical spectral distribution is Wasserstein-close to the semicircle law in the same sense as the previous section (where we controlled the trace), in that

$$\sum \frac{1}{(1 + \alpha_1) - \sqrt{\alpha_1} \lambda_i(G)} = N \cdot \int \frac{1}{(1 + \alpha_1) - \sqrt{\alpha_1} u} d\mu_{\text{sc}}(u) + O(1) = N + O(1)$$

Because  $M_3$  is independent of  $M_1$ , it suffices to control the expected Frobenius norm of the matrix – the true realization concentrates around its expectation to additive  $O(1)$  factors. It is not difficult to see that this expectation is equal to

$$\begin{aligned}
& \frac{1}{N} \cdot \left( \alpha_2 - \frac{\rho^2}{\alpha_1} \right) \cdot \left( \sum \frac{1}{(1 + \alpha_1) - \sqrt{\alpha_1} \lambda_i(G)} \right)^2 \\
&= N \cdot \left( \alpha_2 - \frac{\rho^2}{\alpha_1} \right) \cdot \left( \int \frac{1}{(1 + \alpha_1) - \sqrt{\alpha_1} u} d\mu_{\text{sc}}(u) \right)^2 + O(1) \\
&= N \cdot \left( \alpha_2 - \frac{\rho^2}{\alpha_1} \right) + O(1).
\end{aligned}$$

Putting the pieces together and returning to the Taylor expansion, we get that with very high probability,

$$\begin{aligned}
& 2 \left( \text{Error}_{0,0}^{(2)} - \text{Error}_{a,b}^{(2)} \right) \\
&= \frac{N\alpha_1}{2} - \frac{N\alpha_2}{2} + \log \det (\text{Id} + M) \\
&= O(1) + \frac{N\alpha_1}{2} - \frac{N\alpha_2}{2} + \text{Tr}(M) - \frac{1}{2} \|M\|_F^2 \\
&= O(1) + N \cdot \left( \frac{\alpha_1}{2} - \frac{\alpha_2}{2} + (\alpha_2 - \rho) - \frac{1}{2} \left( \alpha_2 - \frac{\rho^2}{\alpha_1} \right) \right) \\
&= O(1) + N \cdot \left( \frac{\alpha_1}{2} - \rho + \frac{\rho^2}{2\alpha_1} \right) \\
&= O(1) + \frac{N}{2} \cdot \frac{(\alpha_1 - \rho)^2}{\alpha_1}.
\end{aligned}$$

Because  $\alpha_1 - \rho = O\left(\frac{a^2+b^2}{N}\right)$ , this is  $O(1)$  for  $a, b \leq \iota N^{1/4}$ , completing the proof.  $\square$

## Moment calculations for covariance bounds

We first prove subgaussian concentration for the covariance of the degree-2 part.

*Proof of Proposition 4.8.8.* Part (2) follows from part (1) by a standard tail integration argument. Indeed, the random variable  $W$  is bounded by  $N^{1/2}$ , so the contribution to  $\mathbb{E}[W^k]$  from the event  $|W| \geq N^{1/5}$  is bounded by

$$N^{k/2} \mathbb{P}(|W| \geq N^{1/5}) \leq N^{k/2} e^{-cN^{2/5}},$$

which is vanishing for any constant  $k$ . So, we focus on proving part (1). In the case  $W = \langle \sigma^1, \sigma^2 \rangle / \sqrt{N}$ , this is a special case of [HMP24, Lemma 7.5] (where we take  $u = 0$ ). We consider the case  $W = \langle \sigma, v_i \rangle$ . Recall that

$$H_{N,2}(\sigma) = \langle A\sigma, \sigma \rangle = \frac{\sqrt{\xi''(0)}}{2} \langle M\sigma, \sigma \rangle$$

where  $M \sim \text{GOE}(N)$ . For  $0 \leq s \leq N^{1/5} \log N$ , we will evaluate

$$\int e^{H_{N,2}(\sigma)} d\rho(\sigma) \quad \text{and} \quad \int e^{H_{N,2}(\sigma) + s\langle v_i, \sigma \rangle} d\rho(\sigma)$$

using [HMP24, Lemma 7.3]. We recall the function  $G : (\lambda_{\max}(A), +\infty)$  defined in (4.15), which we copy below for convenience, and define  $\tilde{G}$  by

$$G(\gamma) = \gamma - \frac{1}{2N} \log \det(\gamma I - A), \quad \tilde{G}(\gamma) = G(\gamma) + \frac{s^2}{4N(\gamma - \lambda_i(A))}.$$

Recall from below (4.15) that  $G'$  has a unique root  $\gamma_*$  on  $(\lambda_{\max}(A), +\infty)$ . By the same argument,  $\tilde{G}'$  has a unique root  $\tilde{\gamma}_*$  on the same interval. As argued in the proof of Lemma 4.8.6, the conditions of [HMP24, Lemma 7.3] apply. Applying this lemma with  $u = 0$  and  $u = sv_i$ , respectively, shows that with probability  $1 - e^{-cN}$ ,

$$\begin{aligned} \int e^{H_{N,2}(\sigma)} d\rho(\sigma) &= (1 + O(N^{-c})) \sqrt{\frac{2}{G''(\gamma_*)}} (2e)^{-N/2} \exp(NG(\gamma_*)), \\ \int e^{H_{N,2}(\sigma) + s\langle v_i, \sigma \rangle} d\rho(\sigma) &= (1 + O(N^{-c})) \sqrt{\frac{2}{\tilde{G}''(\tilde{\gamma}_*)}} (2e)^{-N/2} \exp(N\tilde{G}(\tilde{\gamma}_*)). \end{aligned} \quad (4.39)$$

Suppose further the probability  $1 - e^{-cN}$  event in Lemma 4.8.7 holds. As argued in the proof of Proposition 4.8.1,  $|\gamma_* - \gamma_0| \leq \frac{1}{C\sqrt{N}}$ . Also, with probability  $1 - e^{-cN}$ ,  $\lambda_{\max}(A) \leq \frac{\sqrt{\xi''(0)}}{2}(2 + \varepsilon^2/8)$ .

We will show that on the intersection of these events  $|\gamma_* - \tilde{\gamma}_*| = O(N^{-3/5} \log^2 N)$ . First note that

$$\gamma_* - \lambda_{\max}(A) \geq \gamma_0 - \lambda_{\max}(A) - \frac{1}{C\sqrt{N}} \geq \frac{1 + \xi''(0) - (2 + \varepsilon^2/8)\sqrt{\xi''(0)}}{2} - \frac{1}{C\sqrt{N}} \geq \varepsilon^2/32$$

is bounded below by a constant, as in the proof of Fact 4.8.5. Since  $G'(\gamma_*) = 0$  and  $\tilde{G}'(\gamma) = G'(\gamma) - \frac{s^2}{4N(\gamma - \lambda_i(A))^2}$ , we have  $\tilde{\gamma}_* \geq \gamma_*$  and so  $\tilde{\gamma}_* - \lambda_i(A)$  is also bounded below by a constant. Thus, as  $s \leq N^{1/5} \log N$ , we have

$$0 \geq \tilde{G}'(\gamma_*) = -\frac{s^2}{4N(\gamma_* - \lambda_i(A))^2} = O(N^{-3/5} \log^2 N).$$

By direct computation,

$$\tilde{G}''(\gamma) = G''(\gamma) + \frac{s^2}{2N(\gamma - \lambda_i(A))^3}.$$

Since  $\tilde{\gamma}_* \geq \gamma_*$ , we have  $\gamma - \lambda_i(A)$  is bounded below by a constant for all  $\gamma \in [\gamma_*, \tilde{\gamma}_*]$ . We claim that  $\tilde{\gamma}_* \in [\gamma_0 - N^{-1/2}, \gamma_0 + N^{-1/2}]$ . Combining both conclusions of Lemma 4.8.7, we obtain that  $G''(\gamma) \geq \Omega_\varepsilon(1)$  for  $\gamma \in [\gamma_0 - N^{-1/2}, \gamma_0 + N^{-1/2}]$ , and consequently the desired claim on  $\tilde{\gamma}_*$  holds.

We thus have  $\tilde{G}''(\gamma) = O_\varepsilon(1)$  in the same interval, and consequently we can conclude the stronger statement  $|\gamma_* - \tilde{\gamma}_*| = O(N^{-3/5} \log^2 N)$ . Furthermore, since  $\tilde{G}^{(3)}(\gamma) = O_\varepsilon(1)$ , the same logic allows us to also conclude

$$\tilde{G}''(\tilde{\gamma}_*)/\tilde{G}''(\gamma_*) = 1 + O(N^{-3/5} \log^2 N).$$

By Taylor expanding  $\tilde{G}$  around  $\tilde{\gamma}_*$ , we see

$$N|\tilde{G}(\gamma_*) - \tilde{G}(\tilde{\gamma}_*)| \leq \frac{N}{2} |\gamma_* - \tilde{\gamma}_*|^2 \sup_{\gamma \in [\gamma_0 - N^{-1/2}, \gamma_0 + N^{-1/2}]} \tilde{G}''(\gamma) = O(N^{-1/5} \log^4 N).$$

The above two displays allow us to replace instances of  $\tilde{\gamma}_*$  with  $\gamma_*$  in (4.39), yielding

$$\int e^{H_{N,2}(\sigma) + s\langle v_i, \sigma \rangle} d\rho(\sigma) = (1 + O(N^{-c})) \sqrt{\frac{2}{\tilde{G}''(\gamma_*)}} (2e)^{-N/2} \exp\left(N\tilde{G}(\gamma_*)\right),$$

and thus

$$\begin{aligned} \langle e^{s\langle v_i, \sigma \rangle} \rangle_2 &= \frac{\int e^{H_{N,2}(\sigma) + s\langle v_i, \sigma \rangle} d\rho(\sigma)}{\int e^{H_{N,2}(\sigma)} d\rho(\sigma)} \\ &= (1 + O(N^{-c})) \sqrt{\frac{G''(\gamma_*)}{\tilde{G}''(\gamma_*)}} \exp\left(N(\tilde{G}(\gamma_*) - G(\gamma_*))\right) \\ &= (1 + O(N^{-c})) \exp\left(s^2/(4(\gamma_* - \lambda_i(A)))\right) = (1 + O(N^{-c})) \exp(cs^2), \end{aligned}$$

where the last two steps again use that  $\gamma_* - \lambda_i(A)$  is bounded away from 0. The tail estimate on  $W = \langle v_i, \sigma \rangle$  now follows from a standard Chernoff bound.  $\square$

We next turn to Lemma 4.8.11, proving each part in turn.

*Proof of Lemma 4.8.11, Eq. (4.18).* We reproduce Eq. (4.18) below for convenience:

$$\mathbf{E} \int_{S_N} \mathbf{1}[\sigma \notin T(H_N)] e^{H_N(\sigma)} d\rho(\sigma) \leq e^{N\xi(1)/2 - cN^{1/5}}.$$

The proof follows [HS23, Proposition 3.1], except with more precise control of overlaps between  $N^{-2/5}$  and a small constant. By symmetry of the sphere, for any deterministic  $\mathbf{x} \in S_N$ ,

$$\mathbf{E} \int_{S_N} \mathbf{1}[\sigma \notin T(H_N)] e^{H_N(\sigma)} d\rho(\sigma) = \mathbf{E} [\mathbf{1}[\mathbf{x} \notin T(H_N)] e^{H_N(\mathbf{x})}]. \quad (4.40)$$

Let  $\mu_{\text{pl}}(\cdot|\mathbf{x})$  denote the planted model (Definition 4.7.8) conditional on spike  $\mathbf{x}$ . A Gaussian change of measure calculation implies that the right-hand side of (4.40) equals

$$e^{N\xi(1)/2} \mathbb{P}_{H_N^{\mathbf{x},\text{I}} \sim \mu_{\text{pl}}(\cdot|\mathbf{x})}[\mathbf{x} \notin T(H_N^{\mathbf{x},\text{I}})].$$

Thus it suffices to show

$$\mathbb{P}_{H_N^{\mathbf{x},\text{I}} \sim \mu_{\text{pl}}(\cdot|\mathbf{x})}[\mathbf{x} \notin T(H_N^{\mathbf{x},\text{I}})] \leq e^{-cN^{1/5}}.$$

Recall (Remark 4.7.14) that a sample  $H_N^{\mathbf{x},\text{I}} \sim \mu_{\text{pl}}(\cdot|\mathbf{x})$  can be generated by

$$H_N^{\mathbf{x},\text{I}}(\sigma) = N\xi(R(\mathbf{x}, \sigma)) + \tilde{H}_N(\sigma), \quad (4.41)$$

where  $\tilde{H}_N \sim \mu_{\text{null}}$ . Furthermore, from the definition,  $\mathbf{x} \in T(H_N^{\mathbf{x},\text{I}})$  is equivalent to

$$\int_{S_N} \mathbf{1}[|R(\mathbf{x}, \tau)| \geq N^{-2/5}] e^{H_N^{\mathbf{x},\text{I}}(\tau)} d\rho(\tau) \leq e^{N\xi(1)/2 - cN^{1/5}} \quad (4.42)$$

We will show this occurs with probability at least  $1 - e^{-cN^{1/5}}$ . Let  $\psi$  denote the probability density of  $R(\mathbf{x}, \tau) \in [-1, 1]$ , where  $\tau$  is sampled from the Haar measure on  $S_N$ . Then it is known that

$$\psi(q) = \frac{1}{Z_\psi} (1 - q^2)^{(N-3)/2}$$

where  $Z_\psi = \Theta(N^{-1/2})$ . Define the codimension-1 band

$$\text{Band}(q) = \text{Band}(q; \mathbf{x}) := \{\tau \in S_N : R(\mathbf{x}, \tau) = q\}$$

and let

$$Z^{\mathbf{x},\text{I}}(q) = \int_{\text{Band}(q)} e^{H_N^{\mathbf{x},\text{I}}(\tau)} d\rho_q(\tau),$$

where  $\rho_q$  is the Haar measure on  $\text{Band}(q)$ , normalized so that  $\rho_q(\text{Band}(q)) = \psi(q)$ . Then the left-hand side of (4.42) is equal to

$$\int_{N^{-2/5} \leq |q| \leq 1} Z^{\mathbf{x},\text{I}}(q) dq.$$

An application of Guerra's interpolation as in [HS23, Lemma 3.3] shows that for any  $q \in [-1, 1]$  and constant  $\eta > 0$ , with probability  $1 - e^{-cN}$ ,

$$\frac{1}{N} \log Z^{x,I}(q) \leq \frac{1}{2} (\xi(1) + \xi(|q|) + |q| + \log(1 - |q|)) + \eta.$$

Since  $\xi_{\sim 1}$  is  $\varepsilon$ -strictly replica symmetric and  $\gamma_1^2 \leq N^{-4/5}$ , this implies

$$\frac{1}{N} \log Z^{x,I}(q) \leq \frac{\xi(1)}{2} - \frac{\varepsilon q^2}{4} + 2\eta.$$

Let  $\delta > 0$  be small depending on  $\varepsilon$ , and  $\eta$  small depending on  $\delta$ . This implies that for any  $|q| \geq \delta$ , with probability  $1 - e^{-cN}$ ,

$$\frac{1}{N} \log Z^{x,I}(q) \leq \frac{\xi(1)}{2} - \frac{\varepsilon \delta^2}{8}.$$

Taking a union bound over a  $N^{-1}$ -net of  $|q| \geq \delta$  as in [HS23, Lemma 3.4] implies that with probability  $1 - e^{-cN}$ ,

$$\int_{\delta \leq |q| \leq 1} Z^{x,I}(q) dq \leq e^{N\xi(1)/2 - cN}. \quad (4.43)$$

We address the remaining range of  $q$  by a first moment bound. Note that

$$\mathbf{E} \int_{N^{-2/5} \leq |q| \leq \delta} Z^{x,I}(q) dq = e^{N\xi(1)/2} \int_{N^{-2/5} \leq |q| \leq \delta} e^{N\xi(q)} \psi(q) dq. \quad (4.44)$$

Recall from Fact 4.8.4 that  $\xi''(0) \leq 1 - \varepsilon$ . Thus, for sufficiently small  $\delta$ , for all  $|q| \leq \delta$ ,

$$\xi_{\sim 1}(q) + \frac{1}{2} \log(1 - q^2) \leq -\varepsilon q^2/4.$$

Thus, for all  $N^{-2/5} \leq |q| \leq \delta$ ,

$$\begin{aligned} \frac{1}{N} \log \left( e^{N\xi(q)} \psi(q) \right) &= \xi(q) + \frac{1}{N} \log \psi(q) = \gamma_1^2 q + \xi_{\sim 1}(q) + \frac{1}{2} \log(1 - q^2) + O(N^{-1} \log N) \\ &\leq \gamma_1^2 q - \varepsilon q^2/4 + O(N^{-1} \log N) \leq -\varepsilon N^{-4/5}/8. \end{aligned}$$

Combining with (4.44) shows

$$\mathbf{E} \int_{N^{-2/5} \leq |q| \leq \delta} Z^{x,I}(q) dq \leq e^{N\xi(1)/2 - cN^{1/5}},$$

so by Markov's inequality, with probability  $1 - e^{-cN^{1/5}/2}$ ,

$$\int_{N^{-2/5} \leq |q| \leq \delta} Z^{x,I}(q) dq \leq e^{N\xi(1)/2 - cN^{1/5}/2}.$$

Combining with (4.43) proves (4.42) after adjusting  $c$ . □

*Proof of Lemma 4.8.11, Eq. (4.19).* By the same argument leading to (4.42), it suffices to prove

$$\int_{S_N} \mathbf{1}[|R(\mathbf{x}, \tau)| \geq N^{-2/5}] e^{H_N^{x, \text{II}}(\tau)} d\rho(\tau) \leq e^{N\xi(1)/2 - cN^{1/5}} \quad (4.45)$$

holds with probability at least  $1 - e^{-cN^{1/5}}$ , where now

$$H_N^{x, \text{II}}(\sigma) = N\gamma_2^2 R(\mathbf{x}, \sigma)^2 + \tilde{H}_N(\sigma),$$

i.e. we have replaced the spike in  $H_N^{x, \text{I}}$  with only its degree-2 part. Then  $H_N^{x, \text{II}}(\sigma) \leq H_N^{x, \text{I}}(\sigma)$  almost surely for all  $\sigma$  such that  $R(\mathbf{x}, \sigma) \geq 0$ , so (4.42) implies

$$\int_{S_N} \mathbf{1}[R(\mathbf{x}, \tau) \geq N^{-2/5}] e^{H_N^{x, \text{II}}(\tau)} d\rho(\tau) \leq e^{N\xi(1)/2 - cN^{1/5}}$$

with probability  $1 - e^{-cN}$ . Moreover, by symmetry of the degree-2 spike,

$$\int_{S_N} \mathbf{1}[R(\mathbf{x}, \tau) \geq N^{-2/5}] e^{H_N^{x, \text{II}}(\tau)} d\rho(\tau) \stackrel{d}{=} \int_{S_N} \mathbf{1}[R(\mathbf{x}, \tau) \leq -N^{-2/5}] e^{H_N^{x, \text{II}}(\tau)} d\rho(\tau).$$

This implies (4.45) and thus (4.19).  $\square$

*Proof of Lemma 4.8.11, Eq. (4.20).* This follows from a slightly more complex form of the same strategy, where the planted Hamiltonian now has two spikes. For  $\mathbf{x}^1, \mathbf{x}^2 \in S_N$ , let

$$H_N^{x^1, x^2, \text{III}}(\sigma) = N\xi(R(\mathbf{x}^1, \sigma)) + N\xi(R(\mathbf{x}^2, \sigma)) + \tilde{H}_N(\sigma).$$

By the same gaussian change of measure argument as above, the expectation in the left-hand side of (4.20) equals

$$e^{N\xi(1)} \int \mathbf{1}[|R(\sigma^1, \sigma^2)| \leq 3N^{-2/5}] \mathbb{P}(\sigma^1 \notin T(H_N^{\sigma^1, \sigma^2, \text{III}})) d\rho^{\otimes 2}(\sigma^1, \sigma^2).$$

Thus it suffices to show that for all  $\mathbf{x}^1, \mathbf{x}^2 \in S_N$  with  $|R(\mathbf{x}^1, \mathbf{x}^2)| \leq 3N^{-2/5}$ ,

$$\int_{S_N} \mathbf{1}[|R(\mathbf{x}^1, \tau)| \geq N^{-2/5}] e^{H_N^{x^1, x^2, \text{III}}(\tau)} d\rho(\tau) \leq e^{N\xi(1)/2 - cN^{1/5}} \quad (4.46)$$

with probability at least  $1 - e^{-cN^{1/5}}$ . Let  $\lambda = R(\mathbf{x}^1, \mathbf{x}^2) \in [-3N^{-2/5}, 3N^{-2/5}]$  and

$$\mathbf{x}^2 = \lambda \mathbf{x}^1 + \sqrt{1 - \lambda^2} \mathbf{x}_\perp^2,$$



where  $\mathbf{x}_\perp^2 \in S_N$  and  $R(\mathbf{x}^1, \mathbf{x}_\perp^2) = 0$ . Let  $\psi_2$  denote the probability density of  $(R(\mathbf{x}^1, \tau), R(\mathbf{x}_\perp^2, \tau)) \in [1, 1]^2$ , where  $\tau$  is sampled from the Haar measure on  $S_N$ . It is known that

$$\psi_2(q) = \frac{\mathbf{1}[q_1^2 + q_2^2 \leq 1]}{Z_{\psi_2}} (1 - q_1^2 - q_2^2)^{(N-4)/2}$$

where  $Z_\psi = \Theta(N^{-1})$ . Define the codimension-2 band

$$\text{Band}(q_1, q_2) = \text{Band}(q_1, q_2; \mathbf{x}^1, \mathbf{x}^2) := \{\tau \in S_N : R(\mathbf{x}^1, \tau) = q_1, R(\mathbf{x}_\perp^2, \tau) = q_2\}.$$

and let

$$Z^{\mathbf{x}^1, \mathbf{x}^2, \text{III}}(q_1, q_2) = \int_{\text{Band}(q_1, q_2)} e^{H_N^{\mathbf{x}^1, \mathbf{x}^2, \text{III}}(\tau)} d\rho_{q_1, q_2}(\tau),$$

where  $\rho_{q_1, q_2}$  is the Haar measure on  $\text{Band}(q_1, q_2)$ , normalized so that  $\rho_{q_1, q_2}(\text{Band}(q_1, q_2)) = \psi_2(q_1, q_2)$ . Then the left-hand side of (4.46) is equal to

$$\int \mathbf{1}[|q_1| \geq N^{-2/5}] Z^{\mathbf{x}^1, \mathbf{x}^2, \text{III}}(q_1, q_2) d(q_1, q_2). \quad (4.47)$$

Note that

$$\frac{1}{N} \log Z^{\mathbf{x}^1, \mathbf{x}^2, \text{III}}(q_1, q_2) = \xi(q_1) + \xi\left(\lambda q_1 + \sqrt{1 - \lambda^2} q_2\right) + \frac{1}{N} \log \int_{\text{Band}(q_1, q_2)} e^{\tilde{H}_N(\tau)} d\rho_{q_1, q_2}(\tau).$$

Let  $\tilde{q} = \tilde{q}(q_1, q_2) := \sqrt{q_1^2 + q_2^2}$ . Applying Guerra's interpolation as in [HS23, Lemma 3.3] shows that for any  $\eta > 0$ , with probability  $1 - e^{-cN}$

$$\frac{1}{N} \log \int_{\text{Band}(q_1, q_2)} e^{\tilde{H}_N(\tau)} d\rho_{q_1, q_2}(\tau) \leq \frac{1}{2} (\xi(1) - \xi(\tilde{q}) + \tilde{q} + \log(1 - \tilde{q})) + \eta.$$

and thus

$$\begin{aligned} \frac{1}{N} \log Z^{\mathbf{x}^1, \mathbf{x}^2, \text{III}}(q_1, q_2) &\leq \xi(q_1) + \xi\left(\lambda q_1 + \sqrt{1 - \lambda^2} q_2\right) + \frac{1}{2} (\xi(1) - \xi(\tilde{q}) + \tilde{q} + \log(1 - \tilde{q})) + \eta \\ &\leq \xi_{\sim 1}(|q_1|) + \xi_{\sim 1}(|q_2|) + \frac{1}{2} (\xi(1) - \xi_{\sim 1}(\tilde{q}) + \tilde{q} + \log(1 - \tilde{q})) + 2\eta. \end{aligned}$$

Since  $\xi_{\sim 1}$  only includes terms that are degree 2 or larger,  $\xi_{\sim 1}(|q_1|) + \xi_{\sim 1}(|q_2|) \leq \xi_{\sim 1}(\tilde{q})$ . Thus the last display is bounded by

$$\frac{1}{2} (\xi(1) + \xi_{\sim 1}(\tilde{q}) + \tilde{q} + \log(1 - \tilde{q})) + 2\eta \leq \frac{\xi(1)}{2} - \frac{\varepsilon \tilde{q}^2}{4} + 2\eta.$$

Arguing as in the proof of equation (4.18) then shows that for any  $\delta > 0$  depending only on  $\varepsilon$ ,

$$\int \mathbf{1}[\tilde{q}(q_1, q_2) > \delta] Z^{x^1, x^2, \text{III}}(q_1, q_2) d(q_1, q_2) \leq e^{N\xi(1)/2 - cN} \quad (4.48)$$

with probability  $1 - e^{-cN}$ . The remaining part of the integral (4.47) has expectation

$$\begin{aligned} & \mathbf{E} \int \mathbf{1}[|q_1| \geq N^{-2/5}, \tilde{q}(q_1, q_2) \leq \delta] Z^{x^1, x^2, \text{III}}(q_1, q_2) d(q_1, q_2) \\ &= e^{N\xi(1)/2} \int \mathbf{1}[|q_1| \geq N^{-2/5}, \tilde{q}(q_1, q_2) \leq \delta] e^{N\xi(q_1) + N\xi(\lambda q_1 + \sqrt{1-\lambda^2}q_2)} \psi_2(q_1, q_2) d(q_1, q_2). \end{aligned} \quad (4.49)$$

Recall  $\gamma_1^2 \leq N^{-4/5}$ . For all  $(q_1, q_2)$  in this indicator,

$$\begin{aligned} & \frac{1}{N} \log \left( e^{N\xi(q_1) + N\xi(\lambda q_1 + \sqrt{1-\lambda^2}q_2)} \psi_2(q_1, q_2) \right) \\ &= \xi(q_1) + \xi \left( \lambda q_1 + \sqrt{1-\lambda^2}q_2 \right) + \frac{1}{2} \log(1 - \tilde{q}^2) + O(N^{-1} \log N) \\ &\leq 2N^{-4/5} \tilde{q} - \frac{1}{2} (1 - \xi''(0)) \tilde{q}^2 + O(\tilde{q}^3 + N^{-1} \log N) \\ &\leq -\frac{\varepsilon \tilde{q}^2}{2} + 2N^{-4/5} \tilde{q} + O(\tilde{q}^3 + N^{-1} \log N). \end{aligned}$$

Since  $N^{-2/5} \leq \tilde{q} \leq \delta$ , for  $\delta$  sufficiently small depending on  $\varepsilon$  this is bounded by  $-cN^{-4/5}$ . Combining with (4.49) shows that with probability  $1 - e^{-cN^{1/5}}$ ,

$$\int \mathbf{1}[|q_1| \geq N^{-2/5}, \tilde{q}(q_1, q_2) \leq \delta] Z^{x^1, x^2, \text{III}}(q_1, q_2) d(q_1, q_2) \leq e^{N\xi(1)/2 - cN^{1/5}}.$$

Further combining with (4.48) completes the proof.  $\square$

*Proof of Lemma 4.8.11, Eq. (4.21).* This is proved identically to equation (4.20), except with spiked Hamiltonian

$$H_N^{x^1, x^2, \text{IV}}(\sigma) = N\gamma_2^2 R(x^1, \sigma)^2 + N\xi(R(x^2, \sigma)) + \tilde{H}_N(\sigma),$$

i.e. the spike involving  $x^1$  is replaced with just its degree-2 component. The same argument applies and we omit details.  $\square$

We turn to the proofs of Lemmas 4.8.15 and 4.8.16. The following fact will be useful in the proofs of both lemmas.

**Fact 4.10.4.** *If  $\sigma \in S_N$  satisfies  $|\sigma_i|, |\sigma_j| \leq \log N$  and  $\sigma \notin \tilde{T}_{i,j}$ , then  $\sigma \notin \tilde{T}$ .*

*Proof.* Consider any  $\tau \in S_N$  satisfying  $|R(\sigma_{\sim i,j}, \tau_{\sim i,j})| \geq 2N^{-2/5}$ . Since  $|\tau_i|, |\tau_j| \leq N^{1/2}$ ,

$$|R(\sigma, \tau)| \geq |R(\sigma_{\sim i,j}, \tau_{\sim i,j})| - \frac{|\sigma_i||\tau_i| + |\sigma_j||\tau_j|}{N} \geq N^{-2/5}.$$

Thus the expectation over  $\tau$  in (4.22) is larger than the expectation over  $\tau$  in (4.28), and so  $\sigma \notin \widetilde{T}$ .  $\square$

*Proof of Lemma 4.8.15.* We can write  $X_{i,j} = \widetilde{X}_{i,j} + \widetilde{X}_{i,j}^{(1)} + \widetilde{X}_{i,j}^{(2)}$ , where

$$\begin{aligned}\widetilde{X}_{i,j}^{(1)} &= \left\langle \mathbf{1}[|\sigma_i|, |\sigma_j| \leq \log N, \sigma \notin \widetilde{T}_{i,j}] \sigma_i \sigma_j \left( e^{H_{N,\sim 2}(\sigma) - N\xi_{\sim 2}(1)/2} - \mathbf{1}[i = j] \right) \right\rangle_2, \\ \widetilde{X}_{i,j}^{(2)} &= \left\langle \mathbf{1}[|\sigma_i| \vee |\sigma_j| > \log N] \sigma_i \sigma_j \left( e^{H_{N,\sim 2}(\sigma) - N\xi_{\sim 2}(1)/2} - \mathbf{1}[i = j] \right) \right\rangle_2.\end{aligned}$$

By using  $(a + b)^2 \leq 2a^2 + 2b^2$ , we deduce that

$$X_{i,j}^2 \leq 2\widetilde{X}_{i,j}^2 + 2(\widetilde{X}_{i,j}^{(1)} + \widetilde{X}_{i,j}^{(2)})^2,$$

so the rest of the proof is dedicated to showing that  $|\widetilde{X}_{i,j}^{(1)} + \widetilde{X}_{i,j}^{(2)}| \leq e^{-c \log^2 N}$  with probability  $1 - e^{-c \log^2 N}$ . To do so, we will simply apply Markov to control  $|\widetilde{X}_{i,j}^{(1)}|$  and  $|\widetilde{X}_{i,j}^{(2)}|$ .

By Fact 4.10.4 and using  $|\sigma_i|, |\sigma_j| \leq \sqrt{N}$ ,

$$\begin{aligned}|\widetilde{X}_{i,j}^{(1)}| &\leq N \left\langle \mathbf{1}[|\sigma_i|, |\sigma_j| \leq \log N, \sigma \notin \widetilde{T}_{i,j}] \left( e^{H_{N,\sim 2}(\sigma) - N\xi_{\sim 2}(1)/2} + \mathbf{1}[i = j] \right) \right\rangle_2 \\ &\leq N \left\langle \mathbf{1}[\sigma \notin \widetilde{T}] \left( e^{H_{N,\sim 2}(\sigma) - N\xi_{\sim 2}(1)/2} + \mathbf{1}[i = j] \right) \right\rangle_2.\end{aligned}$$

By the first two equations from Corollary 4.8.12,  $\mathbf{E}_{\sim 2}[\widetilde{X}_{i,j}^{(1)}] \leq e^{-cN^{1/5}}$ . Furthermore,

$$\begin{aligned}\mathbf{E}_{\sim 2}|\widetilde{X}_{i,j}^{(2)}| &\leq N \mathbf{E}_{\sim 2} \left\langle (\mathbf{1}[|\sigma_i| > \log N] + \mathbf{1}[|\sigma_j| > \log N]) \left( e^{H_{N,\sim 2}(\sigma) - N\xi_{\sim 2}(1)/2} + \mathbf{1}[i = j] \right) \right\rangle_2 \\ &\leq 2N \left\langle \mathbf{1}[|\sigma_i| > \log N] + \mathbf{1}[|\sigma_j| > \log N] \right\rangle_2 \leq e^{-c \log^2 N}\end{aligned}$$

by Proposition 4.8.8(1). By Markov's inequality, with probability  $1 - e^{-c \log^2 N}$ ,  $|\widetilde{X}_{i,j}^{(1)}| + |\widetilde{X}_{i,j}^{(2)}| \leq e^{-c \log^2 N}$ , after adjusting  $c$  as necessary. This concludes the proof.  $\square$

*Proof of Lemma 4.8.16.* We can write  $\widetilde{X}_{i,j}^2 = \widehat{X}_{i,j} + \widehat{X}_{i,j}^{(1)} - \widehat{X}_{i,j}^{(2)}$ , where

$$\widehat{X}_{i,j}^{(1)} = \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, \sigma^1, \sigma^2 \in \widetilde{T}_{i,j}, |R(\sigma_{\sim i,j}^1, \sigma_{\sim i,j}^2)| > 2N^{-2/5}] \right\rangle$$

$$\begin{aligned}
& \sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2 \left( e^{H_{N,\sim 2}(\sigma^1) - N\xi_{\sim 2}(1)/2} - \mathbf{1}[i = j] \right) \left( e^{H_{N,\sim 2}(\sigma^2) - N\xi_{\sim 2}(1)/2} - \mathbf{1}[i = j] \right) \Bigg\rangle_2, \\
\widehat{X}_{i,j}^{(2)} &= \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, (\sigma^1 \notin \widetilde{T}_{i,j} \vee \sigma^2 \notin \widetilde{T}_{i,j}), |R(\sigma_{\sim i,j}^1, \sigma_{\sim i,j}^2)| \leq 2N^{-2/5}] \right. \\
& \quad \left. \sigma_i^1 \sigma_j^1 \sigma_i^2 \sigma_j^2 \left( e^{H_{N,\sim 2}(\sigma^1) - N\xi_{\sim 2}(1)/2} - \mathbf{1}[i = j] \right) \left( e^{H_{N,\sim 2}(\sigma^2) - N\xi_{\sim 2}(1)/2} - \mathbf{1}[i = j] \right) \right\rangle_2.
\end{aligned}$$

By the same argument as before, it suffices to control  $\mathbf{E}_{\sim 2}|\widehat{X}_{i,j}^{(1)}|$  and  $\mathbf{E}_{\sim 2}|\widehat{X}_{i,j}^{(2)}|$ . Note that almost surely,

$$\begin{aligned}
|\widehat{X}_{i,j}^{(1)}| &\leq N^2 \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, \sigma^1, \sigma^2 \in \widetilde{T}_{i,j}, |R(\sigma_{\sim i,j}^1, \sigma_{\sim i,j}^2)| > 2N^{-2/5}] \right. \\
& \quad \left. \left( e^{H_{N,\sim 2}(\sigma^1) - N\xi_{\sim 2}(1)/2} + 1 \right) \left( e^{H_{N,\sim 2}(\sigma^2) - N\xi_{\sim 2}(1)/2} + 1 \right) \right\rangle_2 \leq N^2 (\widehat{X}_{i,j}^{(3)} + \widehat{X}_{i,j}^{(4)})
\end{aligned}$$

where

$$\begin{aligned}
\widehat{X}_{i,j}^{(3)} &= \left\langle \mathbf{1}[\sigma^1 \in \widetilde{T}_{i,j}, |R(\sigma_{\sim i,j}^1, \sigma_{\sim i,j}^2)| > 2N^{-2/5}] e^{H_{N,\sim 2}(\sigma^1) + H_{N,\sim 2}(\sigma^2) - N\xi_{\sim 2}(1)} \right\rangle_2, \\
\widehat{X}_{i,j}^{(4)} &= \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, |R(\sigma_{\sim i,j}^1, \sigma_{\sim i,j}^2)| > 2N^{-2/5}] \right. \\
& \quad \left. (e^{H_{N,\sim 2}(\sigma^1) - N\xi_{\sim 2}(1)/2} + e^{H_{N,\sim 2}(\sigma^2) - N\xi_{\sim 2}(1)/2} + 1) \right\rangle_2.
\end{aligned}$$

By definition of  $\widetilde{T}_{i,j}$ ,

$$\widehat{X}_{i,j}^{(3)} \leq \left\langle \mathbf{1}[\sigma^1 \in \widetilde{T}_{i,j}] e^{H_{N,\sim 2}(\sigma^1) - N\xi_{\sim 2}(1)/2 - cN^{1/5}} \right\rangle_2,$$

and thus  $\mathbf{E}_{\sim 2}\widehat{X}_{i,j}^{(3)} \leq e^{-cN^{1/5}}$ . Furthermore,

$$\widehat{X}_{i,j}^{(4)} \leq \left\langle \mathbf{1}[|R(\sigma^1, \sigma^2)| > N^{-2/5}] (e^{H_{N,\sim 2}(\sigma^1) - N\xi_{\sim 2}(1)/2} + e^{H_{N,\sim 2}(\sigma^2) - N\xi_{\sim 2}(1)/2} + 1) \right\rangle_2,$$

and thus

$$\mathbf{E}_{\sim 2}\widehat{X}_{i,j}^{(4)} \leq 3 \left\langle \mathbf{1}[|R(\sigma^1, \sigma^2)| > N^{-2/5}] \right\rangle_2 \leq e^{-cN^{1/5}}$$

by Proposition 4.8.8(1). Combining shows  $\mathbf{E}_{\sim 2}|\widehat{X}_{i,j}^{(1)}| \leq e^{-cN^{1/5}}$ . Similarly

$$|\widehat{X}_{i,j}^{(2)}| \leq N^2 \left\langle \mathbf{1}[|\sigma_i^1|, |\sigma_j^1|, |\sigma_i^2|, |\sigma_j^2| \leq \log N, (\sigma^1 \notin \widetilde{T}_{i,j} \vee \sigma^2 \notin \widetilde{T}_{i,j}), \right.$$

$$|R(\sigma_{\sim i,j}^1, \sigma_{\sim i,j}^2)| \leq 2N^{-2/5} \left[ e^{H_{N,\sim 2}(\sigma^1) - N\xi_{\sim 2}(1)/2} + 1 \right] \left( e^{H_{N,\sim 2}(\sigma^2) - N\xi_{\sim 2}(1)/2} + 1 \right) \Bigg\rangle_2$$

By Fact 4.10.4, on the indicator in this expectation,  $\sigma^1, \sigma^2 \in \widetilde{T}$ . Moreover

$$|R(\sigma^1, \sigma^2)| \leq |R(\sigma_{\sim i,j}^1, \sigma_{\sim i,j}^2)| + \frac{|\sigma_i^1| |\sigma_i^2| + |\sigma_j^1| |\sigma_j^2|}{N} \leq 3N^{-2/5}.$$

Thus

$$\begin{aligned} |\widehat{X}_{i,j}^{(2)}| &\leq 2N^2 \left\langle \mathbf{1}[\sigma^1 \notin \widetilde{T}, |R(\sigma^1, \sigma^2)| \leq 3N^{-2/5}] \right. \\ &\quad \left. \left( e^{H_{N,\sim 2}(\sigma^1) - N\xi_{\sim 2}(1)/2} + 1 \right) \left( e^{H_{N,\sim 2}(\sigma^2) - N\xi_{\sim 2}(1)/2} + 1 \right) \right\rangle_2 \end{aligned}$$

and Corollary 4.8.12 implies  $\mathbf{E}_{\sim 2} |\widehat{X}_{i,j}^{(2)}| \leq e^{-cN^{1/5}}$ . □

## **Part II**

# **Synthetic Data in Machine Learning**

# Foreword

Large Language Models (LLMs) have a ravenous appetite for training data, one that has led to modern training routines consuming entire internet’s worth of text data. A natural response here is to rely on LLMs to generate data for training, in the form of *synthetic data*. This part of the thesis will study the success and limitations of synthetic data applied to LLM finetuning.

In Chapter 5, we study the paradigm of weak-to-strong generalization. In it, one generates synthetic labels from a weak teacher LLM and uses them to finetune a strong student LLM, inverting the classic strong teacher weak student setup. Despite the teacher labels having extremely low ground-truth accuracy, we identify regimes in a toy model where such weak-to-strong training generalizes very well.

In Chapter 6, we attempt to directly teach models to catch their own mistakes in their reasoning traces. For completely synthetic reasoning tasks, we find that models trained with this method can successfully correct the synthetic distribution of errors they were trained on. However, we find that such techniques are prone to failure from distribution shift to on-policy errors, and propose methods to mitigate such effects. Our study helps explain why on-policy reinforcement learning methods have been widely successful in training reasoning LLMs.

# Chapter 5

## Weak-to-strong generalization via benign overfitting

### 5.1 Introduction

Motivated by the problem of aligning increasingly capable models, [BIK<sup>+</sup>23] introduced the framework of *weak-to-strong generalization*<sup>1</sup>, which draws an analogy between humans supervising superhuman AIs and weaker teacher models supervising stronger student models. This inverts the classic teacher-student model framework, which typically assumes that the teacher is stronger than the student. [BIK<sup>+</sup>23] found that using GPT-2 to finetune GPT-4 can recover most of the performance of standard supervised finetuning with human-annotated data across standard NLP benchmarks, but struggles on more difficult tasks such as chess puzzles or reward modeling. One failure mode observed by [BIK<sup>+</sup>23] is that the strong student sometimes learns to mimic the weak teacher, i.e. the strong model overfits to the limitations of the weaker one. Since [BIK<sup>+</sup>23], this phenomenon has been studied in a variety of other empirical settings (see e.g. [JCL<sup>+</sup>24, GCW<sup>+</sup>24, LA24, YSS<sup>+</sup>24, TL24]).

These empirical observations naturally lead to the question that we aim to answer in this paper: Can we identify a simple, concrete theoretical setting where we can provably exhibit different phases of weak-to-strong generalization? Below, we survey closely related areas of research.

---

<sup>1</sup>This situation is related, but slightly different, from the earlier identified problem of easy-to-hard generalization (see e.g. [SBG<sup>+</sup>21, HBCW24]) wherein a model is trained on “easy” cases but has to generalize to “hard” cases. The most important distinction is that in weak-to-strong generalization, the weak teacher actually gets things wrong.



**Pseudolabeling and synthetic data.** In semi-supervised learning, *pseudolabeling* refers to the method of using one model’s outputs to generate pseudolabels for unlabeled data [L<sup>+</sup>13, AOA<sup>+</sup>20, RDRS21, ZWH<sup>+</sup>21, CBTQO21, HLZ<sup>+</sup>24]. These pseudolabels are then used to supervise the target model. We consider using the weak model to generate pseudolabels for the strong model in our concrete setting.

Synthetic data generation is another widely popular paradigm for scaling up models in the absence of human-annotated data; see, e.g. [Nik21, LWL<sup>+</sup>24, CLC<sup>+</sup>21, FV22] and references therein. In this approach, one uses generative models to generate synthetic data, which can then be labeled and used to supervise other models. The use of synthetic data to train models resembles the weak-to-strong setup that we study, although we assume the unlabeled datapoints are generated from the ground-truth distribution. Nevertheless, we believe that extending our techniques may yield interesting insights on the success and failure modes of synthetic data.

**Supervised finetuning and scaling laws.** In contemporary deep learning, the dominant paradigm is to pretrain large (e.g. billions of parameters) models on copious (e.g. trillions of tokens) unlabeled data in a self-supervised fashion and then finetune the model on relatively tiny (e.g. less than 100K) finetuning datasets. The prevailing wisdom suggests that (1) pretraining helps the model learn useful features that can be repurposed for specific tasks and (2) the optimal size of models should scale with the amount of data (see, e.g. [KMH<sup>+</sup>20, WBZ<sup>+</sup>21, WKM<sup>+</sup>22, LTM<sup>+</sup>22, HBM<sup>+</sup>22]). Much effort has gone into improving the effectiveness and efficiency of finetuning [HGJ<sup>+</sup>19, HSW<sup>+</sup>21, LARC21, DPHZ24].

According to NTK theory, if the model weights do not move far from initialization during finetuning, we can approximate finetuning as training a generalized linear model using (tangent) features. These features arise from the gradients of the network’s outputs with respect to the parameters at finetuning initialization. This motivates our studying the behavior of interpolating linear models in the context of supervised finetuning. Our concrete theoretical setting sits within the Gaussian-features linear-model style of [WHS22, BHX20, MM22, BLLT20, MVSS20, MNS<sup>+</sup>21, CL21, WT21, SAS22, WDY22, WMT21, CMV<sup>+</sup>23].

**Knowledge distillation.** The topic of weak-to-strong generalization is related to the extensive literature on knowledge distillation, which originated in the desire to compress large powerful models (teachers) into smaller ones (students) [HVD15, BCNM06]. See [GYMT21] for a general survey. Theoretical perspectives for the essentially underparameterized case were developed in [PL19, JZ20], and the theoretically engaged literature in this area has continued to expand; see [YSX<sup>+</sup>24, OLSR<sup>+</sup>23, SPA23, AC24, HBCC24, SBAH23, ZZ23, DS23, XRH<sup>+</sup>24, NMB<sup>+</sup>23, BA23, Har23, SIK<sup>+</sup>21, MFB20] for a few representative more recent examples.

The question of how knowledge distillation can improve the student’s generalization to surpass the teacher (especially when they have the same architecture) has led to the identification of several different underlying mechanisms [YSX<sup>+</sup>24, SPA23, SBAH23, DS23, NMB<sup>+</sup>23, MFB20]. Of these, the closest in spirit to our approach is the regularization viewpoint of [MFB20], which studies a kernel-regression model. They call out the crucial role of the spectrum of the Gram matrix — more specifically, self-distillation accentuates the importance of the larger eigenvalues, which has a regularizing effect by making the corresponding basis functions more prominent in the learned pattern.

**Concurrent theoretical work.** [ZZS<sup>+</sup>24] engages with generative models and observes that temperature can play an important role in allowing a trained model to surpass its training sources. [SPB<sup>+</sup>24] takes a transfer-learning perspective and asserts that naive finetuning on weak pseudolabels tends not to work; our results show conditions where this does in fact succeed. [CPS24] takes a representation-centric perspective in a regression setting and zooms in on the question of how much better the representation is for the stronger model. [LSV24] studies the classification setting and takes a neighborhood perspective that posits that the stronger model’s neighborhood structure allows it to average over the weak labels to get generalization. At a high level, our work along with [MFB20, CPS24, LSV24] all circle around the idea that weak-to-strong generalization works when the cascading learning process purifies representations in the true direction and contracts in false directions.

## Contributions

In this paper, we explore weak-to-strong generalization in a stylized theoretical model that captures the dynamics of finetuning with weak supervision studied in [BIK<sup>+</sup>23]. Under a simple overparameterized spiked covariance model for the pretraining features, we prove that finetuning an overparameterized linear classifier using minimum  $\ell_2$  norm interpolation on top of these features provably exhibits *two distinct phases* of weak-to-strong generalization. In particular, under certain scalings, as we increase the number of weakly labeled finetuning examples, the strong learner’s asymptotic accuracy transitions from (1) random guessing to (2) perfect generalization; see Section 5.3 for a precise statement.

To be specific, we study the generalization of interpolating linear models with and without weak supervision. Although our results can be generalized to any weak teacher which produces logits using a linear head, we assume for the sake of concreteness that the weak teacher is also an interpolating linear model which can be fully expressed by the strong student. We discuss how our theoretical results connect to realistic supervised finetuning scenarios in Section 5.4.

Our results strongly hinge on the tight analysis for benign overfitting in multiclass classification in [WS24], although the study of benign overfitting for regression and binary classification was already carried out by [BLLT20, MNS<sup>+</sup>21].

## 5.2 Preliminaries and setup

Below, we set up the weak-to-strong learning task in Section 5.2, along with the data assumptions in Section 5.2, and finally specify the concrete end-to-end learning algorithm we study in Section 5.2.

### Weak-to-strong setup

To study weak-to-strong generalization, we will consider a simple setup which encapsulates the dynamics of standard supervised finetuning as well as training linear probes on intermediate activations. We will now give a high level description of the weak-to-strong setting, and in subsequent sections formally define the specific assumptions we make to theoretically study weak supervision.

Suppose we have  $n$  labeled datapoints and  $m = n^u$  unlabeled datapoints, where  $u > 1$ . This matches the modern ML paradigm where labeled data is scarce and unlabeled data is abundant.

We assume that we have access to two sets of features extracted from the datapoints: the weak and strong features. Using these features, we will create a weak-to-strong setup with a weak model and a strong model, where the weak model is used to generate hard pseudolabels to train the strong model (see Procedure 5.2.6 for a formal definition). The learning task at hand is binary classification, where each model uses its respective features obtained from the datapoints — see Section 5.4 for how our techniques apply to the multiclass setting.

Clearly, to get nontrivial learning guarantees, there needs to be some relationship between the weak and strong features. To specify this, we will assume that the weak and strong features come from an appropriate *weak-to-strong ensemble* of features, which we formally define in Section 5.2. One way to interpret the weak-to-strong ensemble is that the true hidden direction is highlighted more in the strong features than the weak features; see Figure 5.1.

In supervised finetuning and linear probing, the strong and weak features come from pretraining via the neural tangent features and intermediate activations, respectively. For example, in the GPT2 to GPT4 weak-to-strong setup of [BIK<sup>+</sup>23], the weak features come from GPT2 pretraining, whereas the strong features come from GPT4 pretraining.

Broadly speaking, we study the case where the true labels are generated by a distinguished (but unknown) low-rank subspace hidden in very high dimensional space.

To be concrete, we will study two different classifiers in this setup:

- (2)  $f_{\text{weak}}$ : train/finetune on  $n$  datapoints using weak features in  $\mathbb{R}^{d_{\text{weak}}}$  and ground-truth labels.
- (2)  $f_{\text{w2s}}$ : train/finetune on  $m \gg n$  datapoints using strong features in  $\mathbb{R}^d$  and hard pseudolabels generated from  $f_{\text{weak}}$ .

We study a scheme where  $f_{\text{weak}}$  and  $f_{\text{w2s}}$  are linear models trained by performing minimum  $\ell_2$  interpolation (MNI) on their respective training sets; see Section 5.2 for a formal definition of MNI.

**Remark 5.2.1.** Instead of training with hard (categorical) pseudolabels, one could use the real-valued scores from  $f_{\text{weak}}$ . This would only affect constants which are not crucial to any of our results.

We will measure generalization via the test accuracy of these classifiers. In particular, let  $\ell(\cdot)$  be the 0-1 loss function, and let  $\mathbf{E}[\ell(\cdot)]$  be the expected test error over a fresh test sample. We introduce the following desiderata to define weak-to-strong generalization for binary classification. For multiclass settings with  $k$  different classes, the natural extension would require  $\mathbf{E}[\ell(f_{\text{weak}})] = 1 - \Theta(\frac{1}{k})$ .

**Desiderata 1.** *The main desiderata are the following:*

- (i) *The strong model asymptotically generalizes when trained on the  $m$  weakly labeled datapoints:  $\mathbf{E}[\ell(f_{\text{w2s}})] = o_n(1)$ .*
- (ii) *The strong model can fully represent the weak model.*
- (iii) *The weak model asymptotically does not generalize:  $\mathbf{E}[\ell(f_{\text{weak}})] = \frac{1}{2} - o_n(1)$ .*

In Section 5.3, we show that the above desiderata are achievable in a simple toy model; see Theorem 5.3.3 for a formal statement. This paints a rather striking picture: there are situations where the weak labels asymptotically look like random guessing (Desiderata 1), the strong model can perfectly imitate the weak labels (Desiderata 1), yet the strong model still asymptotically generalizes by extracting enough signal out of the plentiful weak labels (Desiderata 1).<sup>2</sup>

---

<sup>2</sup>In particular, Desiderata 1 allows the failure mode of the strong model imitating the weak model and also rules out more trivial sources of weak labels, such as independent label noise. If the strong model cannot represent the noise in the weak labels, then weak-to-strong generalization is intuitively much simpler.

We also include some bonus desiderata, which paint a comparison to natural alternative training methods. We can also provably achieve these bonus desiderata in certain regimes; see Remark 5.3.4.

**Desiderata 2.** *The extra desiderata are the following:*

- (i) *PCA cannot recover the low-rank structure from  $n + m$  observations of the strong features.*
- (ii) *Let  $f_{\text{strong}}$  be the strong model when trained on  $n$  datapoints using strong features in  $\mathbb{R}^d$  and ground-truth labels. Then  $f_{\text{strong}}$  asymptotically fails:  $\mathbf{E}[\ell(f_{\text{strong}})] = \frac{1}{2} - o_n(1)$ .*

## Data model

Throughout, we will consider data with zero mean Gaussian covariates, which can be viewed as an idealized version of the pretraining features or representations. These covariates are generated from an ambient standard Gaussian vector  $\mathbf{g} = (g_1, \dots, g_D) \in \mathbb{R}^D$ . We make Gaussianity assumptions for the sake of theoretical tractability; in Section 5.4 we discuss potential extensions to other settings.

**Covariates.** A learner observes iid features  $\mathbf{x}_i \sim N(0, \Sigma)$ , where we emphasize that the feature covariances  $\Sigma \in \mathbb{R}^{d \times d}$  are unknown and different for each learner. The different  $\Sigma$ 's capture how weak or strong the features are for each model; we make this precise below. Note that  $\mathbf{x}_i$  is a linear transformation of the underlying randomness  $\mathbf{g}_i \sim N(0, I_D)$ . We will often refer to the eigendecomposition  $\Sigma = U \Lambda U^\top$ , where  $U \in \mathbb{R}^{d \times d}$  is orthogonal and  $\Lambda \in \mathbb{R}^{d \times d}$  is diagonal.

**Labels.** For binary classification, we consider hard labels generated by the signs of Gaussians, so that  $y = \text{sgn}(\langle \mathbf{g}, \mathbf{v}_* \rangle)$ , where  $\mathbf{v}_*$  is an unknown unit-norm direction. To analyze the generalization of various learners, we will allow ourselves to study labels generated by various directions  $\mathbf{v}$ , not just the true  $\mathbf{v}_*$ , all of which we assume are unknown.<sup>3</sup> For multiclass classification, we assume the labels are generated via  $y = \arg \max_{j \in [k]} \langle \mathbf{g}, \mathbf{v}_*^{(j)} \rangle$ , where  $\mathbf{v}_*^{(1)}, \dots, \mathbf{v}_*^{(k)}$  are all unknown.

Following [MNS<sup>+</sup>21, SAS22, WS24], we will make the following assumption on the true label directions to simplify the analysis. However, to study weak-to-strong generalization, we will eventually have to analyze a weak label direction which is *not* 1-sparse. One of our main contributions is showing how to analyze this case.

---

<sup>3</sup>To study weak supervision, we take  $\mathbf{v} = \mathbf{w}_{\text{weak}}$ , the direction that the weak model learns.

**Assumption 5.2.2** (1-sparse assumption). We say that the labels satisfy the 1-sparse assumption relative to covariance  $\Sigma$  if the following holds. The label defining direction  $v_*$  (directions  $v_*^{(1)}, \dots, v_*^{(k)}$  for multiclass) is aligned with a top eigenvector (top- $k$  eigenvectors for multiclass) of  $\Sigma$  such that, in an eigenbasis where  $v_* = e_1$  and  $v_*^{(i)} = e_i$ , respectively, we have

$$y = \text{sgn}(x_1) \quad (\text{Binary})$$

$$y = \arg \max_{j \in [k]} x_j \quad (\text{Multiclass})$$

For example, if the 1-sparse assumption holds for the strong covariance, then in a strong eigenbasis where the strong features are independent, the labels are generated by axis-aligned directions.

**Bi-level ensemble and weak-to-strong ensemble.** To simplify the analysis, we follow [MNS<sup>+</sup>21, SAS22, WS24] and assume the eigenvalues  $\Lambda$  of the covariance are parameterized by the bi-level ensemble defined shortly. The bi-level ensemble is a simple overparameterized version of the well-known spiked covariance model for PCA.

**Definition 5.2.3** (Bi-level ensemble). Let  $x \sim N(0, \Sigma)$ , where the covariance  $\Sigma = U\Lambda U^\top$ . The bi-level ensemble parameterizes  $\Lambda = \Lambda(p, q, r)$ , where  $p > 1$ ,  $0 \leq r < 1$ , and  $0 < q < (p - r)$ . The number of features ( $d$ ), number of spiked directions ( $s$ ), and degree of favoring ( $a$ ) all scale with the number of training points ( $n$ ) as follows:

$$d = \lfloor n^p \rfloor, s = \lfloor n^r \rfloor, a = n^{-q}. \quad (5.1)$$

Then  $\Lambda = \text{diag}(\lambda_i)_{i \in [d]}$ , where

$$\lambda_j = \begin{cases} \frac{ad}{s}, & 1 \leq j \leq s \\ \frac{(1-a)d}{d-s}, & \text{otherwise} \end{cases}. \quad (5.2)$$

If the above holds, we refer to  $x$ ,  $\Lambda$ , or  $\Sigma$  as being drawn from the bi-level ensemble. We use the shorthands  $\lambda_F \triangleq \frac{ad}{s}$  and  $\lambda_U \triangleq \frac{(1-a)d}{d-s}$  to denote the favored and unfavored eigenvalues, respectively.

In particular, the parameterization controls the total number of features,  $d = n^p \gg n$ , as well as the dimension of the low-rank subspace  $s = n^r \ll n$ . For multiclass classification, we allow the number of classes to also scale with the number of datapoints  $n$ .

**Definition 5.2.4** (Scaling for multiclass). For multiclass classification with  $k$  classes, we have  $k = c_k \lfloor n^t \rfloor$ , where  $0 \leq t < r$  and  $c_k$  is a positive integer.

We now pin down a concrete weak-to-strong ensemble, which boils down to specifying the joint distribution of the weak and strong features. We impose a subset relationship between the weak and strong features in the strong basis; see Figure 5.1 for a diagram.

**Assumption 5.2.5** (Weak-to-strong subset ensemble). Let  $\Lambda = \Lambda(p, q, r) \in \mathbb{R}^{d \times d}$  denote the strong eigenvalues and  $\Lambda_{\text{weak}} = \Lambda(p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) \in \mathbb{R}^{d_{\text{weak}} \times d_{\text{weak}}}$  denote the weak eigenvalues, both drawn from the bi-level ensemble. Let  $\lambda_{F,\text{weak}} \triangleq \frac{a_{\text{weak}} d_{\text{weak}}}{s_{\text{weak}}}$  and  $\lambda_{U,\text{weak}} \triangleq \frac{(1-a_{\text{weak}}) d_{\text{weak}}}{d_{\text{weak}} - s_{\text{weak}}}$  denote the weak favored and unfavored eigenvalues, respectively. Let  $U$  be any distinguished eigenbasis of  $\Sigma$  where  $v_* = e_1$  (Assumption 5.2.2). The weak and strong features in the basis  $U$  are related as follows.

$$(5) \quad \mathbf{x}_{\text{strong}} \sim N(0, \Lambda), \text{ where } \Lambda = \lambda_F I_{[s]} + \lambda_U I_{[d] \setminus [s]}.$$

$$(5) \quad \text{There exists subsets of coordinates } S \subseteq [s], T \subseteq [d] \setminus [s], \text{ with } 1 \in S, |S| = s_{\text{weak}}, \text{ and } |T| = d_{\text{weak}} - s_{\text{weak}}, \text{ such that}$$

$$\mathbf{x}_{\text{weak}} = \left( \sqrt{\frac{\lambda_{F,\text{weak}}}{\lambda_F}} \Pi_S + \sqrt{\frac{\lambda_{U,\text{weak}}}{\lambda_U}} \Pi_T \right) \mathbf{x}_{\text{strong}} \stackrel{d}{=} N(0, \lambda_{F,\text{weak}} I_S + \lambda_{U,\text{weak}} I_T).$$

Here,  $\Pi_S$  denotes the projection onto the axis-aligned subspace indexed by  $S$ .

We will often abuse notation and restrict  $\mathbf{x}_{\text{weak}} \in \mathbb{R}^d$  to the coordinates in  $S \cup T$ , viewing  $\mathbf{x}_{\text{weak}} \in \mathbb{R}^{d_{\text{weak}}}$ . Hence, the weak favored features are a subset of the strong favored features, the weak unfavored features are a subset of the strong unfavored features, with different bi-level scalings, and furthermore the 1-sparse assumption holds for both  $\Sigma$  and  $\Sigma_{\text{weak}}$ .

Clearly, under Assumption 5.2.5, Desiderata 1 is satisfied, and the subset ensemble is essentially the simplest relationship between the features one could impose to achieve the desideratum. In Section 5.4, we discuss how we expect the results to change if we relax these data modeling assumptions.

## Learning algorithm

Our models are all trained using minimum  $\ell_2$ -norm interpolation (MNI), which corresponds to the asymptotic behavior of gradient descent with zero initialization [JT21]. Define the data matrix  $\mathbf{X} \in \mathbb{R}^{n \times d}$  by  $\mathbf{X}^\top \triangleq [\mathbf{x}_1 \ \cdots \ \mathbf{x}_n]$ . Let  $\mathbf{y}^{(i)} \in \mathbb{R}^n$  be the (centered) one-hot vector for class  $i \in [k]$ . In MNI, we learn a linear score function  $\mathbf{f}^{(i)}$  via the optimization problem

$$\mathbf{f}^{(i)} = \arg \min_{\mathbf{f}} \|\mathbf{f}\|_2 \quad \text{s.t. } \mathbf{X} \mathbf{f} = \mathbf{y}^{(i)}. \quad (\text{MNI})$$

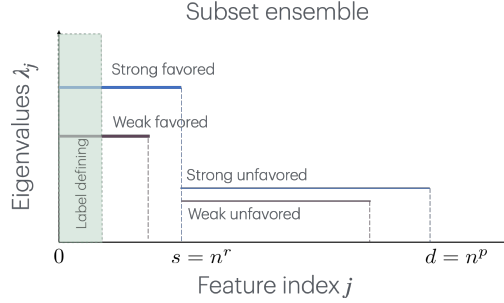


Figure 5.1: Visualization of subset ensemble (Assumption 5.2.5) relating weak and strong features. Notice the decreased favoring for the weak features, and how the weak features are a subset of the strong features in their respective category. Hence, a linear model on the strong features can simulate one on the weak features (Desiderata 1). Moreover, the label defining directions, represented by the labeled shaded box, are in the span of both the weak and strong features.

The coefficients of  $f^{(i)}$  can be easily computed to be  $f^{(i)} = X^\top (XX^\top)^{-1} y^{(i)} \triangleq X^\top A^{-1} y^{(i)}$ , where the matrix  $A \triangleq XX^\top \in \mathbb{R}^{n \times n}$  is the unnormalized Gram matrix and invertible almost surely.

At test time, given a fresh sample  $x_{\text{test}} \sim N(0, \Sigma)$ , in the binary case we predict  $\hat{y} = \text{sgn}(\langle f, x_{\text{test}} \rangle)$ , and for multiclass we predict the class with the highest score:  $\hat{y} = \arg \max_{i \in [k]} \langle f^{(i)}, x_{\text{test}} \rangle$ . To be very explicit, let us formally define the end-to-end training procedure for weak-to-strong binary classification (cf. the description before Remark 5.2.1), with an obvious extension to multiclass settings.

**Procedure 5.2.6** (Weak-to-strong training). The weak learner observes an initial dataset of  $n$  datapoints  $(\tilde{x}_{i,\text{weak}}, y_i)_{i \in [n]}$ , where  $\tilde{x}_{i,\text{weak}}$  are the weak features for the  $i$ th datapoint and  $y_i = \text{sgn}(\langle g_i, v_* \rangle)$  is the corresponding clean hard label. We train  $f_{\text{weak}} \in \mathbb{R}^{d_{\text{weak}}}$  using MNI on these  $n$  clean datapoints.

Then, both learners observe  $m$  extra unlabeled datapoints, where the weak model sees weak features  $(x_{j,\text{weak}})_{j \in [m]}$  and the strong model sees the corresponding strong features  $(x_{j,\text{strong}})_{j \in [m]}$ . Generate  $m$  hard pseudolabels via  $\hat{y}_{j,\text{weak}} = \text{sgn}(\langle f_{\text{weak}}, x_{j,\text{weak}} \rangle)$ , and use MNI to train  $f_{\text{w2s}} \in \mathbb{R}^d$  on  $(x_{j,\text{strong}}, \hat{y}_{j,\text{weak}})_{j \in [m]}$ .

### 5.3 Main results

In this section, we state our main results for binary classification (Theorem 5.3.3) and multilabel classification (Theorem 5.3.5). We focus on the regime  $q+r > u$  and  $q_{\text{weak}}+r_{\text{weak}} >$



1, since [MNS<sup>+</sup>21, Theorem 13] implies that these are the only nontrivial regimes for binary classification under the bi-level ensemble.

In this regime, we also tighten the previous error rates for binary and multiclass classification in the bi-level ensemble from [WS24, Theorem 3.2]; the proof can be found in Section 5.9. To that end, we establish a new concentration inequality for the lower tail of the maximum of correlated Gaussians, which may be of independent interest; we defer its statement to the end of this section.

**Theorem 5.3.1** (Regimes with clean labels). *Suppose the strong features have bi-level covariance  $\Sigma = \Sigma(p, q, r)$  (Definition 5.2.3), where  $q + r > 1$ , the true multiclass labels are 1-sparse (Assumption 5.2.2), and the number of classes  $k = \lfloor n^t \rfloor$  (Definition 5.2.4). Then the test error for  $f_{\text{strong}}$  MNI-trained with  $n$  clean multiclass labels satisfies*

$$\mathbf{E}[\ell(f_{\text{strong}})] = \begin{cases} o_n(1), & \text{if } t < \min \{1 - r, \tau_{\text{strong}}\} \\ 1 - \Theta(\frac{1}{k}), & \text{if } t > \min \{1 - r, \tau_{\text{strong}}\} \end{cases}, \quad (5.3)$$

where  $\tau_{\text{strong}} \triangleq p + 1 - 2(q + r)$ . Furthermore, for binary classification (i.e.  $k = 2$ ), the explicit error rates are given by

$$\mathbf{E}[\ell(f_{\text{strong}})] = \frac{1}{2} - \frac{1}{\pi} \arctan(\Theta(n^{\tau_{\text{strong}}}). \quad (5.4)$$

By generalizing other results from [WS24] to handle imperfect labels, we formally establish weak-to-strong generalization for binary classification in the subset ensemble. We give a brief technical overview of the proof in Section 5.3 and prove it formally in Section 5.7.

**Theorem 5.3.2** (Weak-to-strong generalization for subset ensemble). *Consider the setup in Procedure 5.2.6 where the weak model  $f_{\text{weak}}$  is MNI-trained on  $n$  correctly labeled examples and the strong model  $f_{\text{w2s}}$  is MNI-trained on  $m = n^u$  weakly pseudolabeled examples, where  $q + r > u$  and  $q_{\text{weak}} + r_{\text{weak}} > 1$ . In addition, we make the following data assumptions:*

- (2) *The true binary labels are 1-sparse (Assumption 5.2.2) for the strong covariance.*
- (2) *The weak and strong features follow the subset ensemble (Assumption 5.2.5) with bi-level eigenvalues  $\Lambda(p, q, r)$  and  $\Lambda(p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}})$ , respectively, scaled relative to  $n$ .*
- (2) *There are not too many weakly labeled examples:  $u < \frac{p+1+q+r-(q_{\text{weak}}+r_{\text{weak}})}{2}$ .*

Recall  $\tau_{\text{strong}} \triangleq p + 1 - 2(q + r)$ . Then, the resulting test error for  $f_{\text{w2s}}$  satisfies

$$\mathbf{E}[\ell(f_{\text{w2s}})] = \begin{cases} o_n(1), & \text{if } u > q_{\text{weak}} + r_{\text{weak}} - \min\{1 - r, \tau_{\text{strong}}\} \\ \frac{1}{2} - o_n(1), & \text{if } u < q_{\text{weak}} + r_{\text{weak}} - \min\{1 - r, \tau_{\text{strong}}\}. \end{cases} \quad (5.5)$$

In other words, under the stated parameter regimes, weak-to-strong generalization occurs when the strong model is trained on sufficiently many weak binary labels.<sup>4</sup> Figure 5.2 demonstrates that the conditions for weak-to-strong generalization are not vacuous. We make the third enumerated assumption for technical reasons, but we believe this condition is essentially tight.

By combining Theorems 5.3.1 and 5.3.2, we obtain the exact conditions where the main desiderata from Section 5.2 are satisfied in the bi-level ensemble. The regimes where our theorem applies are depicted in Figures 5.2a and 5.2b, and we validated our theory with numerical simulations of MNI with  $n = 50$  in Figures 5.2c and 5.2d; see Section 5.10 for more details on the experiments.

**Theorem 5.3.3** (Main result). *Under the same setup as Theorem 5.3.2, all the core weak-to-strong desiderata Desiderata 1 are satisfied under the following conditions:*

$$\begin{aligned} u &> q_{\text{weak}} + r_{\text{weak}} - \min\{1 - r, \tau_{\text{strong}}\} && \text{(Weak-to-strong generalization)} \\ 0 &> \tau_{\text{weak}}, && \text{(Weak model does not generalize)} \end{aligned}$$

where  $\tau_{\text{weak}} \triangleq p_{\text{weak}} + 1 - 2(q_{\text{weak}} + r_{\text{weak}})$  and  $\tau_{\text{strong}} \triangleq p + 1 - 2(q + r)$ . In particular, the conditions on the number of weak examples  $m = n^u$  are non-vacuous as long as

$$\tau_{\text{w2s}} \triangleq p + 1 - (q + r + q_{\text{weak}} + r_{\text{weak}}) > 0.$$

Note that under Assumption 5.2.5, Desiderata 1 is immediately satisfied since the weak features are spanned by the strong features. Also, as a sanity check, we confirm that whenever weak-to-strong generalization occurs with  $m$  weak labels, the strong model would generalize with  $m$  clean labels. By Theorem 5.3.1, if the strong model was instead trained on  $m = n^u$  clean examples, it would generalize whenever  $u > 2(q + r) - p$ . By expanding the definition of  $\tau_{\text{strong}}$  and using  $q_{\text{weak}} + r_{\text{weak}} > 1$ ,

$$u > q_{\text{weak}} + r_{\text{weak}} - \tau_{\text{strong}} = 2(q + r) - p + (q_{\text{weak}} + r_{\text{weak}} - 1) > 2(q + r) - p,$$

which recovers the condition with  $m$  clean labels stated above.

---

<sup>4</sup>We expect weak imitation to occur as soon as  $m \gg d$  if the models are trained using gradient descent, i.e. the classic underparameterized regime with more weakly labeled datapoints than features.

**Remark 5.3.4** (Bonus desiderata). If  $q + r > u$ , as in the setup of Theorem 5.3.3, PCA fails to recover the spiked subspace [SSZM13, FW15], so the bonus Desiderata 2 automatically holds. Additionally, by using the condition for where the strong model fails with  $n$  clean examples from Theorem 5.3.1 ( $\tau_{\text{strong}} < 0$ ), one can construct parameter regimes where the bonus Desiderata 2 also holds. The theorem also implies the strong model cannot bootstrap its own performance. Indeed, some basic algebra shows  $\tau_{\text{w2s}} = \tau_{\text{strong}} = \tau_{\text{weak}}$  if  $(p, q, r) = (p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}})$ .

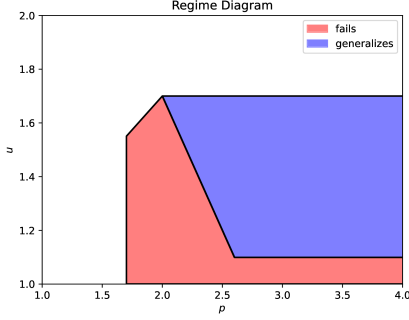
We can extend the above result to the *multilabel* setting, which is a variant of multiclass classification where a datapoint can have multiple positive labels. The basic idea is that multilabel training can be approximated as several independent binary classification problems, and then we can apply the binary analysis. Since this requires additional setup, we defer the formal details to Section 5.8.

**Theorem 5.3.5** (Informal, see Theorem 5.8.5). *Under the analogous assumptions as Theorem 5.3.3 for multilabel classification, there is weak-to-strong generalization for  $f_{\text{w2s}}$  trained on  $m$  weakly hard multilabeled examples from  $f_{\text{weak}}$  in the exact same regimes.*

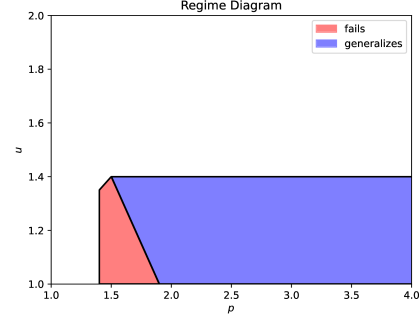
The above result also suggests an interesting result for true multiclass problems with one-hot labels. In particular, one can use the multilabel scheme to generate weak multilabels for a *multiclass* weak-to-strong classifier  $f_{\text{w2s}}$ . One can argue that whenever multilabel weak-to-strong generalization occurs, so too does multiclass weak-to-strong generalization (see Section 5.8 for more justification). By comparing to the regimes for clean multiclass labels due to Theorem 5.3.1, there exists regimes where the strong model would fail if trained using the same number of *clean multiclass* labels (i.e. only one positive label per example). At a high level, this can happen because there are too many classes, which sparsifies the label vectors. In contrast, the weak multilabels do not suffer from the sparsity issue, even though the underlying signal is noisier. This appears to be related to the strategy of using soft labels or logits in the pseudolabeling and knowledge distillation literature.

Finally, we state the aforementioned concentration inequality for the maximum of correlated Gaussians, which may be of independent interest. Our result complements the sharp bound of [LY22] for more moderate deviations; the proof can be found in Section 5.9.

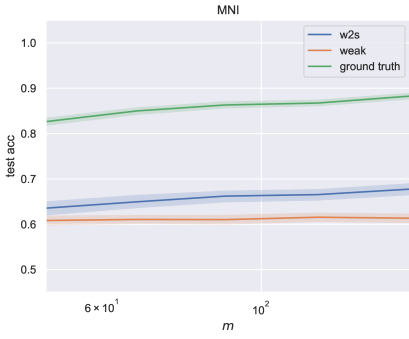
**Theorem 5.3.6** (Lower tail for correlated Gaussians). *Let  $\rho_0 \in (0, 1)$  be a parameter bounded away from 0 and 1, and let  $(g_i)_{i \in [N]}$  be jointly Gaussian with zero mean and unit variance. Suppose  $\mathbf{E}[g_i g_j] \leq \rho_0$  for all distinct  $i, j \in [N]$ . For any  $0 \leq t_N = \delta_0 \sqrt{2(1 - \rho_0) \log N}$  where  $\delta_0 \in [0, 1)$*



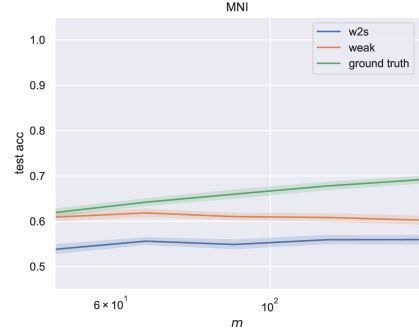
(a)  $(q, r) = (0.9, 0.8), (p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) = (1.4, 0.9, 0.4)$ .



(b)  $(q, r) = (0.9, 0.5), (p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) = (1.1, 0.9, 0.2)$



(c)  $(p, q, r) = (2, 0.6, 0.6), (p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) = (1.4, 0.9, 0.5)$ .



(d)  $(p, q, r) = (1.5, 0.6, 0.8), (p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) = (1.4, 0.9, 0.5)$ .

Figure 5.2: **(Top)**: Regime plots for weak-to-strong generalization based on Theorem 5.3.3. The blue region is the successful w2s regime, and the red region is where w2s training fails. The white region corresponds to regimes where either the hypotheses of the theorem fail to hold, or invalid settings of parameters for the bi-level ensemble. **(Bottom)**: Comparison of simulations of MNI test accuracies for two different regimes where  $n = 50$ . Observe how the weak accuracy is close to random guessing and how the weak-to-strong accuracy increases as  $m$  increases. As corroborated by the plots, Theorem 5.3.3 predicts w2s success in Figure 5.2c and failure in Figure 5.2d.

is bounded away from 1, there is a constant  $C > 0$  depending only on  $\rho_0$  such that

$$\Pr \left[ \max_{i \in [N]} g_i \leq t_N \right] \leq C \cdot N^{(1-\delta_0)^2(1-\frac{1}{\rho_0})} (\log N)^{\frac{1-\rho_0(2-\delta_0)-\delta_0}{2\rho_0}}.$$

In particular, one can take  $C = \sqrt{\frac{\rho_0}{1-\rho_0}}$ . If we further have  $\mathbf{E}[g_i g_j] = \rho_0$  for all distinct  $i, j$  and  $t_N = O\left(\frac{\log \log N}{\sqrt{\log N}}\right)$ , then  $\Pr[\max_{i \in [N]} g_i \leq t_N] = \Theta\left(N^{1-\frac{1}{\rho_0}} (\log N)^{\frac{1}{2\rho_0}-1}\right)$ .

## Technical overview

In this section, we illustrate the main ideas of the proof of Theorem 5.3.2. Recall that the true label is  $\text{sgn}(x_*)$ , where  $x_* = \langle g_{\text{test}}, v_* \rangle$ . Thus, studying weak-to-strong generalization amounts to analyzing

$$\Pr[\text{sgn}(\langle f_{w_{2s}}, x_{\text{test}} \rangle) = \text{sgn}(x_*)], \quad (5.6)$$

where the fresh test point  $x_{\text{test}} \sim N(0, \Sigma)$ . Observe that conditioned on the training data  $\mathbf{X}$ , the true Gaussian  $x_*$  and  $\langle f_{w_{2s}}, x_{\text{test}} \rangle$  are jointly Gaussian random variables, so there exist  $\lambda \in \mathbb{R}$  and  $g \sim N(0, 1)$  independent of  $x_*$  such that

$$\langle f_{w_{2s}}, x_{\text{test}} \rangle = \lambda x_* + g. \quad (5.7)$$

The exact value of  $\lambda$  depends on exactly how much of  $x_*$ 's signal survives through the weak-to-strong training process. Intuitively, the larger  $\lambda$  is, the higher the accuracy. Indeed, it is well-known from the classical study of noise stability that the probability in Eq. (5.6) is precisely  $\frac{1}{2} + \frac{1}{\pi} \arctan(\lambda)$ .

To compute  $\lambda$ , we use Gram-Schmidt to decompose  $f_{w_{2s}}$  to obtain the signal and noise with respect to  $x_*$ . We term these quantities the survival and contamination of  $x_*$ , respectively. Without loss of generality, we can rotate to the basis  $U$  where the strong features are drawn iid from  $N(0, \Lambda)$ , where  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$ . After applying the basis change, one easily computes  $\langle f_{w_{2s}}, x_{\text{test}} \rangle \stackrel{d}{=} N(0, \|\Lambda^{1/2} f_{w_{2s}}\|_2^2)$ . The survival and contamination are then defined as follows.

**Definition 5.3.7** (Survival and contamination). Let  $v \in \mathbb{R}^D$  be a unit vector and let  $f \in \mathbb{R}^d$  be a linear classifier for features drawn from  $N(0, \sum_{i \in [d]} \lambda_i v_i v_i^\top)$ , where  $v_1, \dots, v_d \in \mathbb{R}^D$  are orthonormal. The survival and contamination of  $v$  in  $f$  is defined as

$$\text{SU}(v) \triangleq \sum_{i \in [d]} \sqrt{\lambda_i} f[i] \langle v_i, v \rangle, \quad \text{CN}(v) \triangleq \sqrt{\|\Lambda^{1/2} f\|_2^2 - \text{SU}(v)^2}.$$

If  $f$  is trained by MNI on  $y = \text{sgn}(\langle g, w \rangle)$ , we write  $\text{SU}_n(v|w)$  for the survival of  $v$  given  $n$  labels generated from  $w$ , and similarly for  $\text{CN}_n(v|w)$ . If  $n$  or  $w$  is clear from context, we omit them.

To interpret these notions, consider the simple case where  $v = e_1$  and  $v_i = e_i$ , which is the 1-sparse setting studied by previous papers. There, the survival is simply just  $\sqrt{\lambda_1} f[1]$  and the contamination is  $\sqrt{\sum_{i>1} \lambda_i f[i]^2}$ ; this was used to analyze binary classification in [MNS<sup>+</sup>21].

In the weak-to-strong setting, we have  $\langle f_{w2s}, x_{\text{test}} \rangle = \text{SU}_m(v_*|v_{\text{weak}})x_* + \text{CN}_m(v_*|v_{\text{weak}})g$ , where  $v_{\text{weak}} = f_{\text{weak}}/\|f_{\text{weak}}\|$  is the unit-norm direction learned by the weak model, and  $g$  is a standard Gaussian independent of  $x_*$ . Plugging in  $x_* = \langle g, v_* \rangle$  into (5.7) yields

$$\Pr[\text{sgn}(\text{SU}_m(v_*|v_{\text{weak}})x_* + \text{CN}_m(v_*|v_{\text{weak}})g) = \text{sgn}(x_*)], \quad (5.8)$$

from which we can read off  $\lambda = \frac{\text{SU}_m(v_*|v_{\text{weak}})}{\text{CN}_m(v_*|v_{\text{weak}})}$ . Thus, the explicit formula for the probability yields that  $f_{w2s}$  approaches perfect accuracy if and only if  $\frac{\text{SU}_m(v_*|v_{\text{weak}})}{\text{CN}_m(v_*|v_{\text{weak}})} = \omega(1)$ , whereas its accuracy approaches random guessing if and only if  $\frac{\text{SU}_m(v_*|v_{\text{weak}})}{\text{CN}_m(v_*|v_{\text{weak}})} = o(1)$ .

We now turn to the strategy for controlling the survival and contamination. Using the explicit MNI formula for  $f_{\text{weak}}$  and  $f_{w2s}$ , this analysis reduces to controlling certain random bilinear forms. Applying standard eigenvalue concentration results and a version of the Hanson-Wright concentration inequality for bilinear forms developed by [WS24], we can precisely control the coordinates of  $f_{\text{weak}}$ , and hence  $v_{\text{weak}}$ . Once we have control over  $v_{\text{weak}}$ , we can then study  $f_{w2s}$ .

One technical difficulty is that the weak labels are no longer 1-sparse, even if the true labels are. This marks a departure from previous analyses of benign overfitting in classification, which all assumed 1-sparse labels [MNS<sup>+</sup>21, WMT21, WS24]. To overcome this obstacle, we develop tools to analyze MNI beyond the 1-sparse regime, namely by using the Woodbury inversion formula and carefully bounding error terms using Hanson-Wright. Under the subset ensemble, the explicit parameterizations for the survival and contamination then yield the stated conditions for weak-to-strong generalization.

## 5.4 Discussion

We now discuss further extensions which suggest interesting new weak-to-strong phenomena.

**Modeling assumptions.** In this paper, we assumed the features were Gaussian for the sake of theoretical tractability. One could relax this to vector-subgaussian, as this notion is preserved under rotation. On the other hand, the bi-level and subset ensembles were chosen as a minimal set of tractable theoretical conditions to study weak-to-strong generalization. In principle, our techniques could be extended to more complicated feature ensembles, but it would likely be quite involved.

As corroborated by our experiments in Figs. 5.2c and 5.2d, in practice one does not observe sharp transitions in test error from  $\frac{1}{2}$  to 0 using weak supervision — such a sharp transition corresponds to a Performance Gap Recovered (PGR) equal to 1 in [BIK<sup>+</sup>23] This

phenomenon can nevertheless be justified by our theory. First, the transition we prove is asymptotic; the rate of convergence matters to predict the PGR for finite  $n$ . Another factor is the 1-sparse assumption for the true labels. We argue in Section 5.11 that this is essentially necessary to get this sharp transition in asymptotic test error. In practice, the true directions are rarely aligned with the eigenvectors of the covariance, so the transition would instead be between different *constant* levels of test error.

**Weak imitation.** Our main result Theorem 5.3.2 uses the subset ensemble Assumption 5.2.5 to establish two distinct phases of weak-to-strong generalization. By adding a third intermediate eigenvalue level — thus relaxing the bi-level ensemble to a *tri-level* ensemble — and changing the relationship between the strong and weak basis, we expect a third distinct regime of weak imitation to occur in an overparameterized setting. At a high level, this can happen if the weak classifier puts most of its mass in its own disfavored direction  $v$ , but in the strong feature space  $v$  has an intermediate level of favoring. Given enough weakly labeled datapoints, the strong model will eventually learn to imitate the weak model. Nevertheless, it is still possible for weak-to-strong generalization to occur, with the regime for weak-to-strong generalization widening as the intermediate favoring level decreases.

**Multiclass classification.** By leveraging Theorem 5.3.1 and the techniques developed for binary and multilabel classification in the appendix, we expect our result to extend to the multiclass setting. The main technical challenge is controlling the expected weak-to-strong training behavior, which is significantly more complicated. However, multilabel weak supervision overcomes this challenge and is more tractable to analyze, as discussed after Theorem 5.3.5.

**Relevance to realistic settings.** One connection to more realistic settings comes through the NTK perspective, as touched upon in the introduction. To reiterate, if finetuning remains in the lazy training phase, then the supervised finetuning dynamics are captured by an appropriate generalized linear model defined by the NTK approximation. The strong and weak features are therefore learned from the pretraining phase. Another relevant setting is training linear probes on intermediate activations to interpret large models [AB16, MT23, NLW23, ZPC<sup>+</sup>23]. Hence, under appropriate overparameterization of the intermediate activations, our results also apply there. To go beyond classification problems, one could hope to use some recent implicit bias results for next token prediction [Thr24] to study language models.

## 5.5 Preliminaries

In this section, we lay out some preliminaries for the rigorous justification of our main results, and give a technical overview in Section 5.3 which motivates the study of the survival and contamination of the signal, defined formally in Definition 5.3.7. In Section 5.6, we will introduce tools from high dimensional probability which can rigorously control the survival and contamination. In Section 5.7, we will then use these tools to prove the main results of the paper for binary classification. Then, in Section 5.8, we will discuss how the same exact proofs extend to the multilabel setting, and in Section 5.9 we will prove a tight lower tail inequality for correlated Gaussians which yields Theorem 5.3.1. In Section 5.10 we include some simulations validating our theory, and finally in Section 5.11 we give a heuristic calculation which arrives at the same predictions for the weak-to-strong regimes which should lead to rigorous proofs even without the 1-sparse assumption.

**Notation.** For positive integers  $n$ , we use the shorthand  $[n] \triangleq \{1, \dots, n\}$ . For a vector  $v \in \mathbb{R}^n$ ,  $\|v\|_2$  always denotes the Euclidean norm. We index entries by using square brackets or subscripts, whichever is clearer, so  $v[j]$  and  $v_j$  both denote the  $j$ th entry of  $v$ . For any matrix  $M \in \mathbb{R}^{m \times n}$ , we denote its  $ij$ th entry by  $m_{ij}$ ,  $\|M\|_2$  denotes the spectral norm, and  $\|M\|_F = \text{Tr}(M^T M)$  denotes the Frobenius norm. If  $M \in \mathbb{R}^{n \times n}$  is symmetric, we write  $\mu_1(M) \geq \mu_2(M) \geq \dots \geq \mu_n(M)$  to denote the ordered eigenvalues of  $M$ .

We make extensive use of asymptotic notation. We say  $f(n) \asymp g(n)$  if  $f(n) = \Theta(g(n))$ ,  $f(n) \lesssim g(n)$  if  $f(n) = O(g(n))$ , and  $f(n) = \tilde{O}(g(n))$  if  $f(n) = O(g(n) \log(n)^c)$  for some constant  $c \geq 0$ . When we write  $f(n) \gg g(n)$ , we mean that  $\frac{f(n)}{g(n)} \geq n^\varepsilon$  for some constant  $\varepsilon > 0$ . An event  $\mathcal{E}$  is said to hold with high probability if it holds with probability  $1 - 1/n^c$  for some constant  $c > 0$ , and very high probability if it holds with probability  $1 - \exp(-n^c)$  for some constant  $c > 0$ . For us, high probability statements will be taken with respect to  $n$ , the number of datapoints we scale the weak and strong features with respect to, but under our terminology it makes no difference whether we use  $n$  or  $m$  because they are polynomially related.

In the weak-to-strong ensemble analysis, many of the expressions will depend on  $q + r$ , but the number of datapoints involved will change. In an effort to reduce confusion, we introduce the following notation.

**Definition 5.5.1** (Bi-level prefactor). If the covariance  $\Sigma = \Sigma(p, q, r)$  is drawn from the bi-level ensemble scaled with respect to  $n$ , then we define

$$\mu_n \triangleq \frac{an}{s} = n^{1-q-r}. \quad (5.9)$$



In particular, if  $q + r > 1$  then  $\mu_n \ll 1$ . During weak supervision, the strong learner has bi-level features with covariance  $\Sigma(p, q, r)$  scaled with respect to  $n$  and receives  $m = n^u$  weakly labeled examples. Hence, we define

$$\mu_m \triangleq \frac{am}{s} = n^{u-q-r}. \quad (5.10)$$

**Remark 5.5.2.** Note that once  $a, s$  are fixed,  $\mu_n$  is linear in  $n$ .

## 5.6 Tools for analyzing survival and contamination

In this section, we introduce the basic machinery which can justify the heuristic calculations carried out in Section 5.11. We will heavily rely on the analysis of [WS24], so in an attempt to not replicate effort we will comment on how the proofs change in the weak-to-strong setting.

**Basis change.** Since we have bi-level data and the true direction  $v_*$  is 1-sparse (Assumption 5.2.2), there are convenient basis changes which will greatly simplify the calculations. In particular, let  $S$  be the spiked subspace of  $\Sigma$ , and  $w$  be any unit vector. Then any vectors in  $S$  are eigenvectors of  $\Sigma$ , so we can pick an eigenbasis  $U_*$  so that  $x_{\text{strong}}$  has independent coordinates,  $v_*$  is 1-sparse, and  $w$  is 3-sparse. We state this more formally in the following definition.

**Definition 5.6.1** (Strong eigenbasis). Given  $\Sigma$ ,  $v_*$ , and  $w$  as above, let  $U_*$  be the distinguished eigenbasis of  $\Sigma$  such that after rotating to  $U_*$ , we have  $x_{\text{strong}} \sim N(0, \Lambda)$ ,  $v_* = e_1$ , and  $w = w_1 e_1 + w_2 e_2 + w_{s+1} e_{s+1}$ .

**Hanson Wright inequality.** Now, we recall a few versions of the Hanson-Wright inequality which handles bilinear forms between subgaussian vectors.

For the multiclass case, we also need to handle potentially sparse vectors; this version was proved in [WS24].

**Theorem 5.6.2** (Hanson-Wright for bilinear forms with soft sparsity). *Let  $x = (X_1, \dots, X_n) \in \mathbb{R}^n$  and  $y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$  be random vectors such that  $(X_i, Y_i)$  are independent pairs of (possibly correlated) centered random variables such that  $\|X_i\|_{\psi_2} \leq K$  and  $Y_i$  has soft sparsity at level  $\sigma$ , i.e.  $|Y_i| \leq 1$  almost surely, and  $\mathbb{E}[Y_i^2] \leq \sigma$ . Assume that conditioned on  $Y_j$ ,*

$\|X_j\|_{\psi_2} \leq K$ . Then there exists an absolute constant  $c > 0$  such that for all  $\mathbf{M} \in \mathbb{R}^{n \times n}$  and  $\epsilon \geq 0$  we have

$$\Pr[|\mathbf{x}^\top \mathbf{M} \mathbf{y} - \mathbb{E}[\mathbf{x}^\top \mathbf{M} \mathbf{y}]| > \epsilon] \leq 2 \exp\left(-c \min\left\{\frac{\epsilon^2}{K^2 \sigma \|\mathbf{M}\|_F^2}, \frac{\epsilon}{K \|\mathbf{M}\|_2}\right\}\right). \quad (5.11)$$

Because we are working with MNI, in a typical application we hope to set  $\mathbf{M} = \mathbf{A}^{-1}$ . However, in order to apply Hanson-Wright, one needs to ensure that the matrix  $\mathbf{M}$  is independent of  $\mathbf{x}$  and  $\mathbf{y}$ .

**Decorrelated Gram matrix.** As we are working with Gaussian covariates, one way to achieve this is to decorrelate  $\mathbf{A}$  with  $\mathbf{x}$  and  $\mathbf{y}$ .

**Definition 5.6.3** (Decorrelated Gram matrix). Let  $\mathbf{x}_1, \dots, \mathbf{x}_n \sim N(0, \Lambda)$  be iid samples from a Gaussian distribution in  $\mathbb{R}^d$ . Let  $R = \{\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(\ell)}\}$  be an arbitrary set of orthonormal vectors in  $\mathbb{R}^d$  and for each  $j \in [\ell]$ , let  $\mathbf{h}^{(j)} = (h_1, \dots, h_n) \in \mathbb{R}^n$  be the corresponding Gaussian observations, where  $h_i \sim \langle \Lambda^{1/2} \mathbf{g}_i, \mathbf{v}^{(j)} \rangle$ .

Define the projection matrix onto  $R$  by  $\Pi_R = \sum_{\mathbf{v} \in R} \mathbf{v} \mathbf{v}^\top$ . The  $R$ -projection of the data matrix  $\mathbf{X}_R \in \mathbb{R}^{n \times d}$  is defined by

$$\mathbf{X}_R \triangleq \mathbf{X} \Pi_R,$$

and the  $R$ -decorrelated data matrix  $\mathbf{X}_{-R} \in \mathbb{R}^{n \times d}$  is defined by

$$\mathbf{X}_{-R} \triangleq \mathbf{X}(\mathbf{I} - \Pi_R) = \mathbf{X} - \mathbf{X}_R.$$

The decorrelated Gram matrix is defined by

$$\mathbf{A}_{-R} \triangleq \mathbf{X}_{-R} \mathbf{X}_{-R}^\top = \mathbf{X}(\mathbf{I} - \Pi_R) \mathbf{X}^\top.$$

When  $R = \{\mathbf{v}\}$ , we will simply abbreviate  $\{\mathbf{v}\}$  with  $\mathbf{v}$ .

**Remark 5.6.4.** If  $R$  does not consist of orthonormal vectors, one can apply Gram-Schmidt first so that the above definition still makes sense.

It is not hard to show that as defined above,  $\mathbf{A}_{-R}$  is mutually independent of  $\{\mathbf{h}^{(j)}\}_{j \in [\ell]}$ .

**Fact 5.6.5.** In the same setting as Definition 5.6.3,  $\mathbf{A}_{-R}$  is mutually independent of  $\{\mathbf{h}^{(j)}\}_{j \in [\ell]}$ .

*Proof.* The fact that  $\mathbf{X}_{-R}$  is independent of the  $h_i$ 's is due to the fact that if  $g, h$  are jointly Gaussian with zero mean and unit variance, then  $g - \mathbf{E}[gh]h$  is independent of  $h$ . This immediately generalizes to each row of  $\mathbf{X}$ , since the entries are all independent for any row. Finally, as the rows are independent, the statement is proved.  $\square$

To relate a bilinear form against  $\mathbf{A}_{-R}^{-1}$  to the original bilinear form against  $\mathbf{A}^{-1}$ , we will also need the Woodbury inversion formula for inverting the Gram matrix after decorrelating it with various features.

Define the hat matrix for  $R$  by

$$\mathbf{H}_R \triangleq \mathbf{X}_R^\top \mathbf{A}_{-R}^{-1} \mathbf{X}_R. \quad (5.12)$$

These hat matrices appear in the Woodbury inversion formula. For the sake of notational compactness, define

$$\mathbf{M}_R \triangleq \mathbf{X}_R (\mathbf{I} + \mathbf{H}_R)^{-1} \mathbf{X}_R^\top. \quad (5.13)$$

**Fact 5.6.6** (Woodbury inversion formula). *We have*

$$\begin{aligned} \mathbf{A}^{-1} &= (\mathbf{X}_R \mathbf{X}_R^\top + \mathbf{A}_{-R})^{-1} \\ &= \mathbf{A}_{-R}^{-1} - \mathbf{A}_{-R}^{-1} \mathbf{X}_R (\mathbf{I} + \mathbf{H}_R)^{-1} \mathbf{X}_R^\top \mathbf{A}_{-R}^{-1} \end{aligned} \quad (5.14)$$

$$= \mathbf{A}_{-R}^{-1} - \mathbf{A}_{-R}^{-1} \mathbf{M}_R \mathbf{A}_{-R}^{-1}. \quad (5.15)$$

**Concentration of spectrum.** Finally, in view of Hanson-Wright, we will need to control  $\|\mathbf{A}_{-R}^{-1}\|_2$  and  $\text{Tr}(\mathbf{A}_{-R}^{-1})$ . To this end, we control the spectrum of  $\mathbf{A}_{-R}$ . For any PSD matrix  $\Sigma$  with eigenvalues  $\Lambda = \text{diag}(\lambda_i)_{i \in [d]}$  sorted in descending order and  $k \in [d]$ , define the effective rank  $r_k(\Sigma) \triangleq \frac{\sum_{i \geq k} \lambda_i}{\lambda_k}$ . We will use the following bounds adapted from [BLLT20], which show that the effective rank of the covariance matrix controls the spectrum of the Gram matrix (and hence the sample covariance).

**Lemma 5.6.7** (Eigenvalue bounds from [BLLT20]). *Suppose that  $\mathbf{x} \sim N(0, \Sigma)$ , where  $\Sigma = \mathbf{U} \Lambda \mathbf{U}^\top$ . Write  $\Lambda = \text{diag}(\lambda_i)$ , and let  $\mathbf{A} = \mathbf{X} \mathbf{X}^\top \in \mathbb{R}^{n \times n}$  denote the Gram matrix for  $n$  iid observations of  $\mathbf{x}$ . If  $\mathbf{x} = \mathbf{U}^\top \Lambda^{1/2} \mathbf{z}$ , where  $\mathbf{z}$  has independent, unit variance,  $O(1)$ -subgaussian coordinates, the following holds.*

*If  $r_1(\Sigma) = \omega(n)$ , then with very high probability*

$$\text{Tr}(\Lambda)(1 - o(1)) \leq \mu_n(\mathbf{A}) \leq \mu_1(\mathbf{A}) \leq \text{Tr}(\Lambda)(1 + o(1))$$

As a corollary, we show that in the regime  $q+r > 1$  in the bi-level ensemble, decorrelating the Gram matrix with any sublinear number of jointly Gaussian random variables will yield a flat matrix.

**Corollary 5.6.8.** *Suppose that  $\mathbf{x} \sim N(0, \Sigma)$ , where  $\Sigma \in \mathbb{R}^{d \times d}$  satisfies  $\mu_1(\Sigma) = o(\frac{d}{n})$  and for any  $\ell = o(d)$ , we have  $\mu_\ell(\Sigma)r_\ell(\Sigma) = d(1-o(1))$  (in particular, this holds if  $\Sigma$  follows the bi-level ensemble and  $q+r > 1$ ). Let  $R$  be an arbitrary set of orthonormal vectors in  $\mathbb{R}^d$  such that  $|R| = o(d)$ , and let  $\mathbf{A}_{-R}$  be the corresponding decorrelated Gram matrix for  $n$  iid observations of  $\mathbf{x}$  as in Definition 5.6.3. Then with very high probability,*

$$d(1 - o_n(1)) \leq \mu_n(\mathbf{A}_{-R}) \leq \mu_1(\mathbf{A}_{-R}) \leq d(1 + o_n(1))$$

*Proof.* First, note that the rows of  $\mathbf{X}_{-R}$  are iid samples from  $N(0, \Sigma_{-R})$ , where  $\Sigma_{-R} = (I - \Pi_R)\Sigma(I - \Pi_R)$ .

Let  $\ell = |R|$ ; by assumption  $\ell = o(d)$ . Since  $I - \Pi_R$  is a projection matrix, we conclude by Cauchy Interlacing that  $\mu_1(\Sigma_{-R}) \leq \mu_1(\Sigma)$  and for  $i \in \{\ell + 1, \dots, d\}$ ,  $\mu_i(\Sigma) \leq \mu_i(\Sigma_{-R})$ . It follows that

$$\begin{aligned} r_1(\Sigma_{-R}) &= \frac{\text{Tr}(\Sigma_{-R})}{\mu_1(\Sigma_{-R})} \geq \frac{\mu_\ell(\Sigma)r_\ell(\Sigma)}{\mu_1(\Sigma)} \\ &\geq \omega\left(\frac{d(1 - o(1))}{d/n}\right) \quad (\text{Assumption on } \Sigma) \\ &\geq \omega(n). \end{aligned}$$

Thus Lemma 5.6.7 implies that  $\mu_i(\mathbf{A}_{-R}) = d(1 \pm o(1))$  with high probability. □

This leads to the following bounds on the spectra of  $\mathbf{A}_{-R}^{-1}$  and the hat matrix  $\mathbf{H}_R$ ; this generalizes [WS24, Corollary B.5, Proposition B.6].

**Lemma 5.6.9.** *In the bi-level model, for any set of orthonormal vectors  $R$  with  $|R| = o(d)$ , if  $q+r > 1$ , then with very high probability we have*

$$\begin{aligned} \text{Tr}(\mathbf{A}_{-R}^{-1}) &= n^{1-p}(1 \pm o(1)) \\ \|\mathbf{A}_{-R}^{-1}\|_2 &\leq (1 + o(1))n^{-p} \\ \|\mathbf{X}_R \mathbf{X}_R^\top\|_2 &\leq o(n^p) \\ \mu_i(I + \mathbf{H}_R) &= 1 \pm o(1), \quad i \in [|R|] \end{aligned}$$

*Proof.* The first two statements are an immediate consequence of Corollary 5.6.8. For the third, let  $S, T \subseteq \mathbb{R}^d$  be the spiked and unspiked subspaces of  $\Sigma$ . Since  $\mathbf{X}_R = (\mathbf{X}_S + \mathbf{X}_T)\Pi_R$  and using  $(A + B)M(A + B) \leq 2AMA + 2BMB$ , we have

$$\mathbf{X}_R \mathbf{X}_R^\top \leq 2\mathbf{X}_S \mathbf{X}_S^\top + 2\mathbf{X}_T \Pi_R \mathbf{X}_T^\top.$$

[WS24, Lemma I.1] implies that  $\|\mathbf{X}_S \mathbf{X}_S^\top\|_2 \leq n^{1-q-r} \cdot n^p = o(n^p)$  since  $q + r > 1$ . On the other hand, since the law of  $\mathbf{X}_T$  is rotationally invariant, one can view  $\mathbf{X}_T \Pi_R \mathbf{X}_T^\top$  as a Wishart matrix  $\mathbf{Y} \mathbf{Y}^\top$  where  $\mathbf{Y} \in \mathbb{R}^{n \times |R|}$ . Hence the concentration of singular values for a matrix with independent subgaussian rows (see, e.g. [Ver10, Theorem 5.39]) implies that with probability  $1 - e^{-\sqrt{n}}$ ,

$$\|\mathbf{X}_T \Pi_R \mathbf{X}_T^\top\|_2 \leq (\sqrt{|R|} + \sqrt{n} + n^{1/4})^2 = O(R + n) = o(n^p).$$

Combining this with the bound on  $\|\mathbf{X}_S \mathbf{X}_S^\top\|_2$ , we conclude that  $\|\mathbf{X}_R \mathbf{X}_R^\top\|_2 = o(n^p)$ .

For the last statement, since  $I + \mathbf{H}_R = I + \mathbf{X}_R^\top \mathbf{A}_{-R}^{-1} \mathbf{X}_R$ , by applying the earlier bounds on  $\|\mathbf{A}_{-R}^{-1}\|_2$  and  $0 \leq \mathbf{X}_R^\top \mathbf{X}_R \leq o(n^p)$ , we conclude that  $\mu_i(I + \mathbf{H}_R) = 1 \pm o(1)$ .  $\square$

**Hanson-Wright calculations.** Combining the above results, we can establish a concentration inequality for the bilinear form  $\mathbf{z}_i^\top \mathbf{A}^{-1} \mathbf{y}$ .

**Proposition 5.6.10.** *Let  $\mathbf{y} = \text{sgn}(\langle \mathbf{g}, \mathbf{w} \rangle)$  for some unit vector  $\mathbf{w} \in \mathbb{R}^d$  and  $R$  be any set of orthonormal vectors such that  $\text{span}(R) \supseteq \{\mathbf{e}_i, \mathbf{w}\}$  and  $|R| = o(d)$ . Then conditioned on  $\mathbf{A}_{-R}^{-1}$ , with probability  $1 - \frac{1}{n}$ ,*

$$|\mathbf{z}_i^\top \mathbf{A}_{-R}^{-1} \mathbf{y} - \mathbf{E}[\mathbf{z}_i^\top \mathbf{A}_{-R}^{-1} \mathbf{y}]| \leq c_1 \|\mathbf{A}_{-R}^{-1}\|_F \sqrt{\log n},$$

where  $c_1$  is a positive constant. Furthermore, under the bi-level scaling, if  $q + r > 1$ , then with very high probability over  $\mathbf{A}_{-R}^{-1}$  we have

$$\begin{aligned} \mathbf{E}[\mathbf{z}_i^\top \mathbf{A}_{-R}^{-1} \mathbf{y} | \mathbf{A}_{-R}^{-1}] &= \left(\frac{2}{\pi}\right)^{3/2} (1 \pm o_n(1)) \arcsin(w_i) \cdot \frac{n}{d}, \\ \|\mathbf{A}_{-R}^{-1}\|_F &= (1 \pm o_n(1)) \frac{\sqrt{n}}{d}. \end{aligned}$$

*Proof.* Since  $\mathbf{A}_{-R}^{-1}$  is independent of  $(\mathbf{z}_i, \mathbf{y})$ , we can apply Theorem 5.6.2 to obtain the first statement.

For the other two expressions, we can apply Lemma 5.6.9 since we assumed  $q + r > 1$ . We now compute the expectation using the noise stability formula:

$$\mathbf{E}[\mathbf{z}_i^\top \mathbf{A}_{-R}^{-1} \mathbf{y} | \mathbf{A}_{-R}^{-1}] = \text{Tr}(\mathbf{A}_{-R}^{-1} \mathbf{E}[\mathbf{y} \mathbf{z}_i^\top])$$

$$\begin{aligned}
&= \left(\frac{2}{\pi}\right)^{3/2} \arcsin(w_i) \operatorname{Tr}(\mathbf{A}_{-R}^{-1}) && \text{(Noise stability)} \\
&= \left(\frac{2}{\pi}\right)^{3/2} (1 \pm o_n(1)) \arcsin(w_i) \cdot \frac{n}{d}, && \text{(Lemma 5.6.9)}
\end{aligned}$$

where the last line holds with very high probability over  $\mathbf{A}_{-R}^{-1}$ .

Finally, for the last statement we have

$$\begin{aligned}
\|\mathbf{A}_{-R}^{-1}\|_F &= \sqrt{\sum_{i \in [n]} \mu_i (\mathbf{A}_{-R}^{-1})^2} \\
&= \sqrt{\sum_{i \in [n]} d^{-2} (1 \pm o(1))} && \text{(Lemma 5.6.9)} \\
&= (1 \pm o(1)) \frac{\sqrt{n}}{d}.
\end{aligned}$$

□

## Controlling the error from Woodbury

In the previous subsection, we introduced some tools for obtaining concentration of the coefficients of the MNI classifier  $f_{\text{strong}}$ . The remaining survival and contamination reduce to understanding the typical behavior of bilinear forms  $\mathbf{z}_i^\top \mathbf{A}^{-1} \mathbf{y}$ . To do so, we work in the distinguished basis  $U_*$  introduced in Definition 5.6.1. We can thus pick  $R = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_{s+1}\}$  and apply Hanson-Wright to the bilinear form  $\mathbf{z}_i^\top \mathbf{A}_{-R}^{-1} \mathbf{y}$ . The next lemma bounds the error from replacing  $\mathbf{A}^{-1}$  with  $\mathbf{A}_{-R}^{-1}$ .

**Lemma 5.6.11.** *Suppose we are in the bi-level model,  $\mathbf{y} = \operatorname{sgn}(\langle \mathbf{g}, \mathbf{w} \rangle)$ , and  $q + r > 1$ . Suppose in the basis  $U_*$ ,  $\mathbf{w} = w_1 \mathbf{e}_1 + w_2 \mathbf{e}_2 + w_{s+1} \mathbf{e}_{s+1}$ . Let  $R = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_{s+1}\}$ . With probability  $1 - O(1/n)$ , we have uniformly over  $i \in [d]$  that*

$$|\mathbf{z}_i^\top \mathbf{A}^{-1} \mathbf{y} - \mathbf{z}_i^\top \mathbf{A}_{-R}^{-1} \mathbf{y}| \lesssim \|\mathbf{z}_i^\top \mathbf{A}_{-R}^{-1} \mathbf{X}_R\|_2 \|\mathbf{X}_R^\top \mathbf{A}_{-R}^{-1} \mathbf{y}\|_2.$$

Furthermore, in the bi-level scaling we have with probability  $1 - O(1/n)$  that

$$\|\mathbf{z}_i^\top \mathbf{A}_{-R}^{-1} \mathbf{X}_R\|_2 = \begin{cases} \tilde{O}(\sqrt{\mu_n} n^{\frac{1-p}{2}}) & i \in [2] \\ \tilde{O}(\sqrt{\mu_n} n^{-\frac{p}{2}} + n^{1-p}) & i = s+1 \\ \tilde{O}(\sqrt{\mu_n} n^{-\frac{p}{2}}) & \text{otherwise} \end{cases}$$

and

$$\|\mathbf{X}_R^\top \mathbf{A}_{-R}^{-1} \mathbf{y}\|_2 \lesssim \sqrt{\mu_n} (|w_1| + |w_2|) n^{\frac{1-p}{2}} + \tilde{O}(n^{-\frac{p}{2}}) + |w_{s+1}| n^{1-p} + \tilde{O}(n^{\frac{1}{2}-p}).$$

*Proof.* By relating these two bilinear forms using Woodbury, we have

$$\begin{aligned}
|z_i^\top A^{-1} \mathbf{y} - z_i^\top A_{-R}^{-1} \mathbf{y}| &= |z_i^\top A_{-R}^{-1} \mathbf{M}_R A_{-R}^{-1} \mathbf{y}| && \text{(Fact 5.6.6)} \\
&\leq \|(I + \mathbf{H}_R)^{-1}\|_2 \|z_i^\top A_{-R}^{-1} \mathbf{X}_R\|_2 \|\mathbf{X}_R^\top A_{-R}^{-1} \mathbf{y}\|_2 && \text{(Cauchy-Schwarz)} \\
&\leq (1 + o(1)) \|z_i^\top A_{-R}^{-1} \mathbf{X}_R\|_2 \|\mathbf{X}_R^\top A_{-R}^{-1} \mathbf{y}\|_2. && \text{(Lemma 5.6.9)}
\end{aligned}$$

It remains to control the above norms, which we achieve through another application of Hanson-Wright. Indeed, we can compute

$$\|z_i^\top A_{-R}^{-1} \mathbf{X}_R\|_2^2 = \sum_{j \in \{1, 2, s+1\}} \lambda_j (z_i^\top A_{-R}^{-1} z_j)^2$$

Now, by Hanson-Wright, we have with probability  $1 - O(\frac{1}{n})$  that

$$\lambda_U (z_i^\top A_{-R}^{-1} z_{s+1})^2 = \begin{cases} \tilde{O}(n^{2-2p}) & i = s+1 \\ \tilde{O}(n^{1-2p}) & \text{otherwise} \end{cases}$$

Similarly we have with probability  $1 - O(\frac{1}{n})$  that

$$\lambda_F \sum_{j \in [2]} (z_i^\top A_{-R}^{-1} z_j)^2 = \begin{cases} \tilde{O}(n^{p-q-r} \cdot n^{2-2p}) & i \in [2] \\ \tilde{O}(n^{p-q-r} \cdot n^{1-2p}) & \text{otherwise} \end{cases}$$

Since  $n^{p-q-r} \gg 1$  as  $p > q + r$  and  $n^{p-q-r} \cdot n^{1-p} = \mu_n$ , it follows that with probability  $1 - O(\frac{1}{n})$  we have simultaneously for  $i \in [d]$  that

$$\|z_i^\top A_{-R}^{-1} \mathbf{X}_R\|_2 = \begin{cases} \tilde{O}(\sqrt{\mu_n} n^{\frac{1-p}{2}}) & i \in [2] \\ \tilde{O}(\sqrt{\mu_n} n^{-\frac{p}{2}} + n^{1-p}) & i = s+1 \\ \tilde{O}(\sqrt{\mu_n} n^{-\frac{p}{2}}) & \text{otherwise} \end{cases}$$

A similar calculation, applying Proposition 5.6.10, yields that

$$(z_i^\top A_{-R}^{-1} \mathbf{y})^2 = \begin{cases} \frac{1}{d^2} \left[ \left( \frac{2}{\pi} \right)^{3/2} \arcsin(w_i) n \pm \tilde{O}(\sqrt{n}) \right]^2 & i \in \{1, 2, s+1\} \\ \tilde{O}(\frac{n}{d^2}) & \text{otherwise} \end{cases}$$

Hence,

$$\begin{aligned}
\|\mathbf{X}_R^\top A_{-R}^{-1} \mathbf{y}\|_2 &\lesssim \sum_{i \in \{1, 2, s+1\}} \frac{\sqrt{\lambda_i}}{d} (|w_i| n + \tilde{O}(\sqrt{n})) \\
&\lesssim \sqrt{\mu_n} (\|\mathbf{w}_S\|_1 n^{\frac{1-p}{2}} + \tilde{O}(n^{-\frac{p}{2}})) + |w_{s+1}| n^{1-p} + \tilde{O}(n^{\frac{1}{2}-p}).
\end{aligned}$$

This concludes the proof.  $\square$

We can simplify the above error bound greatly if  $\mathbf{w}$  satisfies the 1-sparse assumption.

**Corollary 5.6.12** (Woodbury error under 1-sparse assumption). *Suppose, in addition to the assumptions of Lemma 5.6.11, we have  $\mathbf{w} = \mathbf{v}_*$ , where  $\mathbf{v}_*$  satisfies the 1-sparse assumption. In other words, in the  $U_*$  basis we have  $\mathbf{v}_* = \mathbf{e}_1$ . Then we pick  $R = \{\mathbf{e}_1\}$ , and with probability  $1 - O(\frac{1}{n})$  we have*

$$|\mathbf{z}_i^\top \mathbf{A}^{-1} \mathbf{y} - \mathbf{z}_i^\top \mathbf{A}_{-R}^{-1} \mathbf{y}| \leq \begin{cases} \tilde{O}(\mu_n \cdot n^{1-p}) & i = 1 \\ \tilde{O}(\mu_n \cdot n^{\frac{1}{2}-p}) & \text{otherwise} \end{cases}$$

## Survival analysis

Let us first establish a general survival bound using Lemma 5.6.11. Afterward, we present special cases under additional assumptions.

**Proposition 5.6.13.** *Let  $\Sigma = \Sigma(p, q, r)$  be drawn from the bi-level ensemble scaled with respect to  $n$  with  $q + r > 1$ . Suppose in the distinguished basis  $U_*$ , we have  $\mathbf{w} = w_1 \mathbf{e}_1 + w_2 \mathbf{e}_2 + w_{s+1} \mathbf{e}_{s+1}$ , and set  $R = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_{s+1}\}$ . Then, given  $n$  labels generated by  $\mathbf{w}$ , we have*

$$\left| \mathbf{z}_1^\top \mathbf{A}^{-1} \mathbf{y} - \left( \frac{2}{\pi} \right)^{3/2} \cdot \frac{n}{d} (1 + o(1)) \arcsin w_1 \right| \lesssim \tilde{O}\left( \frac{\sqrt{n}}{d} \right) + \|\mathbf{z}_1^\top \mathbf{A}_{-R}^{-1} \mathbf{X}_R\|_2 \|\mathbf{X}_R^\top \mathbf{A}_{-R}^{-1} \mathbf{y}\|_2.$$

*Proof.* Recall our expressions for the survival:

$$\text{SU}_n(\mathbf{v}|\mathbf{w}) = \sum_{i \in [d]} \lambda_i \mathbf{z}_i^\top \mathbf{A}^{-1} \mathbf{y} \langle \mathbf{v}_i, \mathbf{v} \rangle, \quad (\text{MNI})$$

where each label is generated by  $y_i = \text{sgn}(\langle \mathbf{g}_i, \mathbf{w} \rangle)$ .

Setting  $\mathbf{v} = \mathbf{v}_*$ , in the basis  $U_*$  from Definition 5.6.1, the above expression simplifies to

$$\text{SU}_n(\mathbf{v}_*|\mathbf{w}) = \lambda_1 \mathbf{z}_1^\top \mathbf{A}^{-1} \mathbf{y}$$

Applying Woodbury (Fact 5.6.6) with  $R = \{\mathbf{v}, \mathbf{w}\}$ , we have

$$\mathbf{z}_1^\top \mathbf{A}^{-1} \mathbf{y} = \mathbf{z}_1^\top \mathbf{A}_{-R}^{-1} \mathbf{y} - \mathbf{z}_1^\top \mathbf{A}_{-R}^{-1} \mathbf{M}_R \mathbf{A}_{-R}^{-1} \mathbf{y}.$$

Now we can apply Proposition 5.6.10 to conclude that

$$\left| \mathbf{z}_1^\top \mathbf{A}_{-R}^{-1} \mathbf{y} - \left( \frac{2}{\pi} \right)^{3/2} \cdot \frac{n}{d} (1 + o(1)) \arcsin(w_1) \right| \leq \tilde{O}\left( \frac{\sqrt{n}}{d} \right)$$



with probability  $1 - \frac{1}{n}$ . Now, triangle inequality and Lemma 5.6.11 shows that

$$\left| \mathbf{z}_1^\top \mathbf{A}^{-1} \mathbf{y} - \left( \frac{2}{\pi} \right)^{3/2} \cdot \frac{n}{d} (1 + o(1)) \arcsin w_1 \right| \lesssim \tilde{O}\left(\frac{\sqrt{n}}{d}\right) + \|\mathbf{z}_1^\top \mathbf{A}_{-R}^{-1} \mathbf{X}_R\|_2 \|\mathbf{X}_R^\top \mathbf{A}_{-R}^{-1} \mathbf{y}\|_2.$$

Thus the statement follows.  $\square$

By way of Corollary 5.6.12, we get a clean expression for the survival if  $\mathbf{w} = \mathbf{v}_*$ .

**Corollary 5.6.14.** *If  $q + r > 1$ ,  $\mathbf{v}_*$  satisfies the 1-sparse assumption, and  $\mathbf{y} = \text{sgn}(\langle \mathbf{g}, \mathbf{v}_* \rangle)$ , then with probability  $1 - O(1/n)$*

$$\begin{aligned} \mathbf{z}_1^\top \mathbf{A}^{-1} \mathbf{y} &\asymp \frac{n}{d} \\ |z_i^\top \mathbf{A}^{-1} \mathbf{y}| &\leq \tilde{O}\left(\frac{\sqrt{n}}{d}\right) \quad \forall i > 1 \\ \text{SU}_n(\mathbf{v}_* | \mathbf{v}_*) &\asymp \mu_n \end{aligned}$$

*Proof.* If  $\mathbf{w} = \mathbf{v}_*$ , then by Corollary 5.6.12 the error bound simplifies to

$$|z_i^\top \mathbf{A}^{-1} \mathbf{y} - z_i^\top \mathbf{A}_{-R}^{-1} \mathbf{y}| \leq \begin{cases} \tilde{O}(\mu_n \cdot \frac{n}{d}) & i = 1 \\ \tilde{O}(\mu_n \cdot \frac{n}{d}) & \text{otherwise} \end{cases}$$

Since  $q + r > 1$ , these error terms are lower order, and the results follow.  $\square$

We can get a similarly clean bound under a few additional assumptions on  $\mathbf{w}$  and  $q + r$ .

**Corollary 5.6.15.** *If, in addition to the assumptions of Proposition 5.6.13, we have*

$$(15) \quad |w_i| \geq \omega(\sqrt{\frac{\log n}{n}}) \text{ for } i \in \{1, 2, s+1\}$$

$$(15) \quad |w_2| = o(|w_1|)$$

$$(15) \quad \sqrt{\mu_n} n^{\frac{1-p}{2}} \ll |w_1|$$

then with  $y_i = \text{sgn}(\langle \mathbf{g}_i, \mathbf{w} \rangle)$ , we have

$$\mathbf{z}_1^\top \mathbf{A}^{-1} \mathbf{y} \asymp \frac{n}{d} \arcsin w_1.$$

Furthermore, the survival of  $\mathbf{v}_*$  given  $n$  labels generated by  $\mathbf{w}$  is with high probability

$$\text{SU}_n(\mathbf{v}_* | \mathbf{w}) \asymp \mu_n w_1 \tag{5.16}$$

*Proof.* Under the additional assumptions, then from Lemma 5.6.11

$$\begin{aligned}\|X_R^\top A_{-R}^{-1} \mathbf{y}\|_2 &\lesssim \sqrt{\mu_n}(|w_1| + |w_2|)n^{\frac{1-p}{2}} + n^{1-p} & (|w_i| = \omega(\sqrt{\log n/n})) \\ &\lesssim \sqrt{\mu_n}|w_1|n^{\frac{1-p}{2}} + n^{1-p} & (|w_2| = o(|w_1|))\end{aligned}$$

Since  $i = 1$  here, we can apply Lemma 5.6.11 to see that  $\|z_1^\top A_{-R}^{-1} X_R\|_2 = \tilde{O}(\sqrt{\mu_n}n^{\frac{1-p}{2}})$ . Thus, the Woodbury error term is upper bounded by

$$\begin{aligned}\tilde{O}(\sqrt{\mu_n}n^{\frac{1-p}{2}})(\sqrt{\mu_n}|w_1|n^{\frac{1-p}{2}} + n^{1-p}) &\leq \tilde{O}(\mu_n|w_1|n^{1-p} + \sqrt{\mu_n}n^{\frac{3-3p}{2}}) \\ &\lesssim \tilde{O}(\mu_n|w_1|n^{1-p}) + o(n^{1-p}|w_1|) \\ &\leq o(n^{1-p}|w_1|)\end{aligned}$$

where the second line follows because we assumed  $\sqrt{\mu_n}n^{\frac{1-p}{2}} \ll |w_1|$ , and the last line follows because we assumed  $q + r > 1$  so  $\mu_n \ll 1$ . Therefore, Proposition 5.6.10 implies that

$$\begin{aligned}\left|z_1^\top A^{-1} \mathbf{y} - \left(\frac{2}{\pi}\right)^{3/2} n^{1-p}(1 + o(1)) \arcsin w_1\right| \\ \lesssim \tilde{O}(n^{\frac{1}{2}-p}) + o(n^{1-p} \arcsin(|w_1|)) \\ \leq o(n^{1-p} \arcsin w_1), & (|w_1| \geq \omega(\sqrt{\frac{\log n}{n}}))\end{aligned}$$

which recovers the stated result.  $\square$

## Contamination analysis

We now move onto the contamination.

**Proposition 5.6.16.** *Suppose  $v_*$  satisfies the 1-sparse assumption, and in the distinguished basis  $U_*$  we have  $v_* = e_1$  and  $w = w_1 e_1 + w_2 e_2 + w_{s+1} e_{s+1}$ . Under the same assumptions as Corollary 5.6.15, if there are  $n$  datapoints, and  $q + r > 1$ , then*

$$\begin{aligned}\text{CN}_n(v_*|w)^2 &\lesssim o(\mu_n^2|w_1|^2) + \mu_n^2 n^{r-1} \log(n)^2 + n^{1-p} \log(n) \\ \text{CN}_n(v_*|w)^2 &\gtrsim \mu_n^2 n^{r-1} + n^{1-p}.\end{aligned}$$

Furthermore, the lower bound holds even without the additional assumptions from Corollary 5.6.15.

*Proof.* Since  $v_*$  is 1-sparse, we have

$$\text{CN}(v_*|w)^2 = \sum_{i \in \{2, s+1\}} (\lambda_i z_i^\top A^{-1} \mathbf{y})^2 + \sum_{i \notin \{1, 2, s+1\}} (\lambda_i z_i^\top A^{-1} \mathbf{y})^2. \quad (5.17)$$

Since by definition  $w_i = 0$  for  $i \notin \{1, 2, s+1\}$ , the second term in Eq. (5.17) can be upper bounded up to constant factors using [WS24, Proposition A.2] by

$$\mu_n^2 \frac{s}{n} \log(n)^2 + \frac{n}{d} \log(n) = \mu_n^2 n^{r-1} \log(n)^2 + n^{1-p} \log(n).$$

Now, set  $R = \{e_1, e_2, e_{s+1}\}$ . For the first term, we know from Lemma 5.6.11 that

$$|z_i^\top A^{-1} \mathbf{y} - z_i^\top A_{-R}^{-1} \mathbf{y}| \lesssim \|z_i^\top A_{-R}^{-1} \mathbf{X}_R\|_2 \| \mathbf{X}_R^\top A_{-R}^{-1} \mathbf{y} \|_2,$$

where

$$\|z_i^\top A_{-R}^{-1} \mathbf{X}_R\|_2 = \begin{cases} \tilde{O}(\sqrt{\mu_n} n^{\frac{1-p}{2}}) & i = 2 \\ \tilde{O}(\sqrt{\mu_n} n^{-\frac{p}{2}} + n^{1-p}) & i = s+1 \end{cases}$$

and from our bounds in Proposition 5.6.13, we have

$$\|\mathbf{X}_R^\top A_{-R}^{-1} \mathbf{y}\|_2 \lesssim \sqrt{\mu_n} |w_1| n^{\frac{1-p}{2}} + n^{1-p}.$$

Hence we have

$$\|z_2^\top A_{-R}^{-1} \mathbf{X}_R\|_2 \| \mathbf{X}_R^\top A_{-R}^{-1} \mathbf{y} \|_2 \leq o(n^{1-p} |w_1|).$$

On the other hand, we have

$$\begin{aligned} \|z_{s+1}^\top A_{-R}^{-1} \mathbf{X}_R\|_2 \| \mathbf{X}_R^\top A_{-R}^{-1} \mathbf{y} \|_2 &\lesssim \mu_n |w_1| n^{\frac{1}{2}-p} + \sqrt{\mu_n} |w_1| n^{\frac{3-3p}{2}} + \sqrt{\mu_n} n^{1-\frac{3p}{2}} + n^{2-2p} \\ &\lesssim o(n^{\frac{1}{2}-p}) + o(n^{1-p} |w_1|) + n^{2-2p}, \end{aligned}$$

where the last line follows from the assumptions and  $p > 1$ . Putting these together with Hanson-Wright, we have

$$\begin{aligned} |z_2^\top A^{-1} \mathbf{y}| &\leq \tilde{O}(n^{\frac{1}{2}-p}) + o(n^{1-p} |w_1|) \\ |z_{s+1}^\top A^{-1} \mathbf{y}| &\leq \tilde{O}(n^{\frac{1}{2}-p}) + o(n^{1-p} |w_1|) + n^{2-2p} \end{aligned}$$

This yields an upper bound of

$$o(\mu_n^2 |w_1|^2) + \tilde{O}(n^{1-2p}) + o(n^{2-2p} |w_1|^2) + n^{4-4p} \leq o(\mu_n^2 |w_1|^2) + n^{1-p},$$

as  $p > 1$  and  $|w_1| \leq 1$ . This completes the proof of the upper bound.

For the lower bound, a careful inspection of the proof of [WS24, Proposition A.2] reveals that, for  $q + r > 1$ , the lower bound holds for any multiclass problem with  $t \geq 0$ . This recovers the desired bound by simply lower bounding the second term in the expansion Eq. (5.17), completing the proof.  $\square$

## 5.7 Analyzing the subset ensemble

In this section, we prove Theorem 5.3.2, which establishes weak-to-strong generalization in the simple weak-to-strong subset ensemble. We restate it below for convenience.

**Theorem 5.3.2** (Weak-to-strong generalization for subset ensemble). *Consider the setup in Procedure 5.2.6 where the weak model  $f_{\text{weak}}$  is MNI-trained on  $n$  correctly labeled examples and the strong model  $f_{\text{w2s}}$  is MNI-trained on  $m = n^u$  weakly pseudolabeled examples, where  $q + r > u$  and  $q_{\text{weak}} + r_{\text{weak}} > 1$ . In addition, we make the following data assumptions:*

- (363) *The true binary labels are 1-sparse (Assumption 5.2.2) for the strong covariance.*
- (363) *The weak and strong features follow the subset ensemble (Assumption 5.2.5) with bi-level eigenvalues  $\Lambda(p, q, r)$  and  $\Lambda(p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}})$ , respectively, scaled relative to  $n$ .*
- (363) *There are not too many weakly labeled examples:  $u < \frac{p+1+q+r-(q_{\text{weak}}+r_{\text{weak}})}{2}$ .*

Recall  $\tau_{\text{strong}} \triangleq p + 1 - 2(q + r)$ . Then, the resulting test error for  $f_{\text{w2s}}$  satisfies

$$\mathbf{E}[\ell(f_{\text{w2s}})] = \begin{cases} o_n(1), & \text{if } u > q_{\text{weak}} + r_{\text{weak}} - \min\{1 - r, \tau_{\text{strong}}\} \\ \frac{1}{2} - o_n(1), & \text{if } u < q_{\text{weak}} + r_{\text{weak}} - \min\{1 - r, \tau_{\text{strong}}\}. \end{cases} \quad (5.5)$$

*Proof.* The key idea of the proof is that, despite the error rate of the weak classifier being  $\frac{1}{2} - o(1)$ , the weak classifier still has a noticeable amount of mass on the true label defining direction. So long as the error rate is not  $\frac{1}{2} - O(\frac{1}{\sqrt{m}})$ , which random guessing can achieve, the strong learner can still pick up on the signal contained in the weak labels. To make this precise, we will control the coordinates of  $f_{\text{weak}}$  in the distinguished basis (Definition 5.6.1).

Quantitatively, for the weak learner  $f_{\text{weak}}$ , since we have clean labels and  $q_{\text{weak}} + r_{\text{weak}} > 1$ , we can apply Corollary 5.6.14. Let  $\mu_{n,\text{weak}} \triangleq \frac{a_{\text{weak}} n}{s_{\text{weak}}}$  denote the bi-level prefactor for the weak model. Then the corollary shows that in the basis where  $v_* = e_1$ , we have with high probability

$$\begin{aligned} f_{\text{weak}}[1] &= \sqrt{\lambda_{F,\text{weak}}} z_1^\top A^{-1} \mathbf{y} \\ &\asymp \sqrt{\frac{a_{\text{weak}} d_{\text{weak}}}{s_{\text{weak}}}} \frac{n}{d_{\text{weak}}} \\ &= \sqrt{\mu_{n,\text{weak}}} n^{\frac{1-p_{\text{weak}}}{2}}, \end{aligned}$$

whereas for all other coordinates we have with high probability

$$|f_{\text{weak}}[i]| \lesssim \begin{cases} \sqrt{\lambda_{F,\text{weak}}} \cdot \frac{\sqrt{n}}{d_{\text{weak}}} \sqrt{\log n} & i \in \{2, \dots, s_{\text{weak}}\} \\ \sqrt{\lambda_{U,\text{weak}}} \cdot \frac{\sqrt{n}}{d_{\text{weak}}} \sqrt{\log n} & i \in \{s_{\text{weak}} + 1, \dots, d_{\text{weak}}\} \end{cases}$$

$$\lesssim \begin{cases} \sqrt{\mu_{n,\text{weak}}} n^{-\frac{p_{\text{weak}}}{2}} \sqrt{\log n} & i \in \{2, \dots, s_{\text{weak}}\} \\ n^{\frac{1}{2}-p_{\text{weak}}} \sqrt{\log n} & i \in \{s_{\text{weak}} + 1, \dots, d_{\text{weak}}\} \end{cases}$$

Hence, to obtain  $w_{\text{weak}}$  in  $U_*$ , we first normalize  $f_{\text{weak}}$  and then do the basis change. Some calculation yields

$$\begin{aligned} \|f_{\text{weak}}\|^2 &\lesssim \mu_{n,\text{weak}} n^{1-p_{\text{weak}}} + n^{1-p_{\text{weak}}} \log n \cdot \frac{s_{\text{weak}} \lambda_{F,\text{weak}} + (d_{\text{weak}} - s_{\text{weak}}) \lambda_{U,\text{weak}}}{d_{\text{weak}}} \\ &= (\mu_{n,\text{weak}} + \log n) n^{1-p_{\text{weak}}} \\ &\lesssim n^{1-p_{\text{weak}}} \log n \end{aligned} \quad (q_{\text{weak}} + r_{\text{weak}} > 1)$$

where the second line follows because  $s_{\text{weak}} \lambda_{F,\text{weak}} + (d_{\text{weak}} - s_{\text{weak}}) \lambda_{U,\text{weak}} = d_{\text{weak}}$ . Similarly, we can compute a simple lower bound:

$$\begin{aligned} \|f_{\text{weak}}\|^2 &\geq \lambda_{U,\text{weak}} \sum_{i=s_{\text{weak}}+1}^{d_{\text{weak}}} (z_i^\top A^{-1} y)^2 = \lambda_{U,\text{weak}} y^\top A^{-1} \sum_{i=s_{\text{weak}}+1}^{d_{\text{weak}}} z_i z_i^\top A^{-1} y \\ &\gtrsim n^{1-p_{\text{weak}}}. \end{aligned} \quad (\text{Lemma 5.6.9})$$

We therefore obtain the sandwiching bounds

$$n^{\frac{1-p_{\text{weak}}}{2}} \lesssim \|f_{\text{weak}}\| \lesssim n^{\frac{1-p_{\text{weak}}}{2}} \sqrt{\log n}. \quad (5.18)$$

By rotating to the distinguished basis, since  $q_{\text{weak}} + r_{\text{weak}} > 1$ , this yields

$$w_{\text{weak}} = \frac{f_{\text{weak}}}{\|f_{\text{weak}}\|} = w_1 e_1 + w_2 e_2 + w_{s+1} e_{s+1},$$

where

$$\frac{1}{\sqrt{\log n}} \sqrt{\mu_{n,\text{weak}}} \lesssim w_1 \lesssim \sqrt{\mu_{n,\text{weak}}} \quad (5.19)$$

$$\frac{1}{\log n} n^{-\frac{q_{\text{weak}}}{2}} \lesssim |w_2| \lesssim n^{-\frac{q_{\text{weak}}}{2}} \sqrt{\log n} \quad (5.20)$$

$$1 \lesssim |w_{s+1}|. \quad (5.21)$$

Let us now check the conditions to apply Corollary 5.6.15, keeping in mind we need to scale with respect to  $m$  instead of  $n$ . In particular, we require

- $w_i = \omega(\sqrt{\frac{\log m}{m}})$  for all  $i \in \{1, 2, s+1\}$ .
- $|w_2| \ll w_1$

- $\sqrt{\mu_m} n^{\frac{u-p}{2}} \ll |w_1|$

Since we are going to check the second condition, and clearly the first condition is satisfied for  $i = s + 1$ , it suffices to check the first condition for  $i = 1$ . By plugging in the scaling for  $w_1$  in Eq. (5.19), we have  $w_1 = \omega(\sqrt{\frac{\log m}{m}})$  if

$$1 - q_{\text{weak}} - r_{\text{weak}} > -u \iff q_{\text{weak}} + r_{\text{weak}} < u + 1. \quad (5.22)$$

Let us now verify the second condition. It turns out that the second condition always holds under the bi-level parameterization. Indeed, from Eq. (5.19) and (5.20), we always have

$$|w_2| \leq n^{-\frac{q_{\text{weak}}}{2}} \sqrt{\log n} \ll \frac{1}{\sqrt{\log n}} n^{\frac{1-q_{\text{weak}}-r_{\text{weak}}}{2}} \leq w_1,$$

as  $r_{\text{weak}} < 1$ . Finally, for the third condition, we have

$$\begin{aligned} \sqrt{\mu_m} n^{\frac{u-p}{2}} &= n^{\frac{u-q-r}{2} + \frac{u-p}{2}} \ll n^{\frac{1-q_{\text{weak}}-r_{\text{weak}}}{2}} \iff 2u - (p + q + r) < 1 - (q_{\text{weak}} + r_{\text{weak}}) \\ &\iff q + r > 2u - (p + 1) + q_{\text{weak}} + r_{\text{weak}} \end{aligned} \quad (5.23)$$

With these preparations in hand, we now prove the positive side of the result.

**Sufficient condition for weak-to-strong generalization.** First we lower bound  $\text{SU}_m(v_*|w_{\text{weak}})$ .

Under Eq. (5.22) and (5.23), we conclude that

$$\text{SU}_{m,\text{strong}}(v_*|w_{\text{weak}}) \gtrsim \mu_m w_1 \geq \frac{1}{\sqrt{\log n}} \mu_m \cdot n^{\frac{1-q_{\text{weak}}-r_{\text{weak}}}{2}}$$

On the other hand, by Proposition 5.6.16 we have

$$\text{CN}_{m,\text{strong}}(v_*|w)^2 \leq o(\mu_m^2 |w_1|^2) + \mu_m^2 n^{r-u} \log n + n^{u-p} \log n.$$

We will now analyze the survival-to-contamination ratio for the strong learner. The first term is clearly at least  $\omega(1)$ . For the second term, we get

$$\frac{1}{\log n} n^{\frac{1-q_{\text{weak}}-r_{\text{weak}}}{2} - \frac{r-u}{2}},$$

which is  $\omega(1)$  if

$$q_{\text{weak}} + r_{\text{weak}} < u + 1 - r \quad (5.24)$$

Finally, for the third term, we get

$$\frac{1}{\log n} n^{u-q-r+\frac{1-q_{\text{weak}}-r_{\text{weak}}}{2}-\frac{u-p}{2}},$$

which is  $\omega(1)$  if

$$q_{\text{weak}} + r_{\text{weak}} < u + p + 1 - 2(q + r). \quad (5.25)$$

Collecting the conditions Eq. (5.22), (5.24) and (5.25) yields

$$\begin{aligned} u &> q_{\text{weak}} + r_{\text{weak}} - \min \{1, 1 - r, p + 1 - 2(q + r)\} \\ &= q_{\text{weak}} + r_{\text{weak}} - \min \{1 - r, p + 1 - 2(q + r)\}, \end{aligned}$$

which establishes the positive side of the result.

**Sufficient condition for failure of weak-to-strong generalization.** To get the negative result, we need to upper bound  $\text{SU}_{m,\text{strong}}(v_* | w_{\text{weak}})$  and lower bound  $\text{CN}_{m,\text{strong}}(v_* | w_{\text{weak}})$ . Our earlier calculations and Proposition 5.6.16 yield

$$|\text{SU}_{m,\text{strong}}(v_* | w_{\text{weak}})| \lesssim \mu_m \cdot \max \left\{ w_1, n^{\frac{u}{2}-p} \sqrt{\log n} \right\} \quad (5.26)$$

$$\text{CN}_{m,\text{strong}}(v_* | w_{\text{weak}}) \gtrsim \mu_m n^{\frac{r-u}{2}} + n^{\frac{u-p}{2}}. \quad (5.27)$$

The above bound only differ by log factors from the upper bounds. Hence, in this case, the sufficient condition merely flips the direction of the inequality, which recovers the stated result. Otherwise, if  $u < q_{\text{weak}} + r_{\text{weak}} - 1$ , then certainly  $u < q_{\text{weak}} + r_{\text{weak}} - 1 + r$ , and  $n^{\frac{u}{2}-p} \sqrt{\log n}$  dominates. The latter is in turn dominated by  $n^{\frac{u-p}{2}}$ , so the survival-to-contamination ratio is  $o(1)$ . This completes the proof.  $\square$

## 5.8 Multilabel classification

In this section, we analyze multilabel classification with  $k$  classes, which is a variant of multiclass classification where a single datapoint can be a positive example of multiple classes. Given  $n$  datapoints, we encode this label information using  $k$  label vectors  $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(k)}$ , where for  $i \in [k]$ , we encode the  $i$ th label vector as  $\mathbf{y}^{(i)} \in \{\pm 1\}^n$ , with 1 representing positive examples. We make the following 1-sparse assumption for multilabel data.

**Assumption 5.8.1** (1-sparse assumption for multilabel). The label defining directions  $v_*^{(1)}, \dots, v_*^{(k)}$  are aligned with a top  $k$  eigenbasis of the strong covariance  $\Sigma$ . In a strong eigenbasis where  $v_*^{(i)} = e_i$ , we have

$$y^{(j)} = \text{sgn}(x_j), \quad \forall j \in [k]$$

**Remark 5.8.2.** In fact, the analysis below will turn out to work even if the label defining directions are *not* orthogonal; they could all be the same! The analysis only relies on there existing some basis  $U_i$  such that  $v_*^{(i)} = e_1$ . However, the tight misclassification rate would be different depending on the relationship between the label defining directions.

The weak-to-strong setup is defined as follows.

- (8)  $f_{\text{weak}} \in \mathbb{R}^{d_{\text{weak}} \times k}$ : train on  $n$  datapoints using weak features and ground-truth labels.
- (8)  $f_{\text{w2s}} \in \mathbb{R}^{d \times k}$ : train on  $m \gg n$  datapoints using strong features and hard pseudolabels generated from  $f_{\text{weak}}$ .

As before, we will study weak-to-strong generalization using overparameterized linear classifiers. Hence,  $f_{\text{weak}}$  will now consist of  $k$  different linear classifiers  $f_{\text{weak}}^{(i)} \in \mathbb{R}^{d_{\text{weak}}}$  for  $i \in [k]$ , all trained by MNI on  $n$  clean multilabel datapoints. To train  $f_{\text{w2s}}^{(i)}$ , we generate hard pseudolabel vectors  $\hat{y}^{(i)} = \text{sgn}(\langle f_{\text{weak}}^{(i)}, x_{\text{weak}} \rangle)$ , and then perform MNI on these hard pseudolabel vectors. We deem that a multilabel classifier  $f$  generalizes if, for a fresh test sample  $x_{\text{test}}$  and for every class  $i \in [k]$ ,  $f$  correctly labels whether  $x_{\text{test}}$  is a positive example of class  $i$ . More formally, define for a collection of classifiers  $f = (f^{(1)}, \dots, f^{(k)})$ , the loss function  $\ell(f) = \Pr \left[ \exists i \in [k] : \text{sgn}(\langle f^{(i)}, x_{\text{test}} \rangle) \neq y_{\text{test}}^{(i)} \right]$ , where the probability is taken over a fresh test sample  $(x_{\text{test}}, y_{\text{test}})$ .

**Remark 5.8.3.** The weak model can also be trained on clean multiclass data rather than clean multilabel data. This will only affect the regimes for the weak model to satisfy Desiderata 1, based on Theorem 5.3.1.

The subset ensemble definition is essentially the same as before (Assumption 5.2.5), with the main difference being that we require all  $k$  of the label defining directions to be favored and axis-alignable in both the weak and strong favored feature subspace.

**Assumption 5.8.4** (Subset ensemble for multilabel classification). Let  $\Lambda = \Lambda(p, q, r) \in \mathbb{R}^{d \times d}$  denote the strong eigenvalues and  $\Lambda_{\text{weak}} = \Lambda(p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) \in \mathbb{R}^{d_{\text{weak}} \times d_{\text{weak}}}$  denote the weak eigenvalues, both drawn from the bi-level ensemble. Suppose the 1-sparse assumption (Assumption 5.2.2) holds for the strong covariance, with any distinguished eigenbasis  $U$  where  $v_*^{(i)} = e_i$  for  $i \in [k]$ . The following conditions relate the weak and strong features after rotating to  $U$ .

- (4)  $x_{\text{strong}} \sim N(0, \Lambda)$ , where  $\Lambda = \lambda_F I_{[s]} + \lambda_U I_{[d] \setminus [s]}$ .



- (4) There exists subsets of coordinates  $S \subseteq [s], T \subseteq [d] \setminus [s]$ , with  $[k] \subseteq S, |S| = s_{\text{weak}}$ , and  $|T| = d_{\text{weak}} - s_{\text{weak}}$ , such that

$$\mathbf{x}_{\text{weak}} = \left( \sqrt{\frac{\lambda_{F,\text{weak}}}{\lambda_F}} \Pi_S + \sqrt{\frac{\lambda_{U,\text{weak}}}{\lambda_U}} \Pi_T \right) \mathbf{x}_{\text{strong}} \stackrel{d}{=} N(0, \lambda_{F,\text{weak}} I_S + \lambda_{U,\text{weak}} I_T).$$

The crucial observation is that multilabel training boils down to  $k$  (nearly) independent binary classification problems. The main difference to establish successful generalization is that we now need to union bound over all  $k$  multilabel classifiers. The high probability bounds from the binary analysis come from two sources: (1) applying spectral bounds (Lemma 5.6.7), which holds with very high probability ( $\exp(-n^{1/2})$ ), and (2) Hanson-Wright calculations (Proposition 5.6.10), where bounds that hold with probability  $\delta$  have deviation  $\text{poly} \log(1/\delta)$ . Hence, for  $k$  a constant,  $\delta$  is only affected by a constant, and this will only change the bounds in our analysis by a constant, which does not shift the regimes. Furthermore, even if we allow  $k$  to scale with  $n$  as in Definition 5.2.4, since  $k$  is polynomial in  $n$ , the dependence for the high probability bounds on  $k$  is at most polylogarithmic in  $k$ , so again nothing changes with the analysis, which looks at polynomial regime shifts.

For the converse direction, one can use a crude bound of just analyzing the probability of one classifier failing, which gives an error rate bounded from below by  $\frac{1}{2} - o_n(1)$ . However, we expect that a more refined analysis would give the expected error rate of  $1 - O(2^{-k})$ ; we sketch an argument for how to get this improved error rate after the theorem statement. With these minor modifications, the formal details go through unchanged, and we arrive at our main theorem for weak-to-strong multilabel classification using hard pseudo-multilabels.

**Theorem 5.8.5** (Weak-to-strong generalization for multilabel subset ensemble). *Consider the setup where the weak model  $f_{\text{weak}}$  is trained on  $n$  correctly labeled examples and the strong model  $f_{\text{w2s}}$  is trained on  $m = n^u$  weakly labeled examples using MNI, where  $u < p, q + r > u$ , and  $q_{\text{weak}} + r_{\text{weak}} > 1$ . Assume the following:*

- (5) *The true multilabels satisfy the 1-sparse assumption (Assumption 5.8.1) for the strong covariance.*
- (5) *The weak and strong features follow the subset ensemble (Assumption 5.2.5) with bi-level eigenvalues  $\Lambda(p, q, r)$  and  $\Lambda(p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}})$ , respectively, scaled relative to  $n$ .*
- (5) *There are not too many weakly labeled examples:  $u < \frac{p+1+q+r-(q_{\text{weak}}+r_{\text{weak}})}{2}$ .*

Let  $\tau_{\text{strong}} \triangleq p + 1 - 2(q + r)$ . Then, the resulting test error for  $f_{\text{w2s}}$  satisfies

$$\mathbf{E}[\ell(f_{\text{w2s}})] = \begin{cases} o_n(1), & \text{if } u > q_{\text{weak}} + r_{\text{weak}} - \min\{1 - r, \tau_{\text{strong}}\} \\ \Omega(1), & \text{if } u < q_{\text{weak}} + r_{\text{weak}} - \min\{1 - r, \tau_{\text{strong}}\}. \end{cases} \quad (5.28)$$

We now sketch out an approach which should give the correct error rate for constant  $k$  (or even  $k$  growing slowly with  $n$ ). Note that the complement of the failure event is

$$\forall i \in [k] : \text{sgn}(\langle f_{w_{2s}}^{(i)}, x_{\text{test}} \rangle) = y_{\text{test}}^{(i)}.$$

From the Gram-Schmidt decomposition Eq. (5.8) and the noise stability formula, we can decompose this event as

$$\forall i \in [k] : \text{sgn}\left(\frac{\text{SU}(v_*^{(i)}|w)}{\text{CN}(v_*^{(i)}|w)} x_*^{(i)} + g^{(i)}\right) = \text{sgn}(x_*^{(i)}),$$

where  $x_*^{(i)} = \langle g_{\text{test}}, v_*^{(i)} \rangle$  and each  $g^{(i)}$  is a standard Gaussian residual. In the converse regime, the survival to contamination ratio is polynomially decaying, so for typical  $g^{(i)}$ , with probability  $\exp(-n^c)$  we have  $\text{sgn}\left(\frac{\text{SU}(v_*^{(i)}|w)}{\text{CN}(v_*^{(i)}|w)} x_*^{(i)} + g^{(i)}\right) = \text{sgn}(g^{(i)})$ . Furthermore, by unpacking the analysis of  $f_{w_{2s}}^{(i)}$  (specifically, see Eq. (5.20)) for  $i \neq j$ ,  $g^{(i)}$  and  $g^{(j)}$  only differ by a Gaussian with polynomially decaying variance, so up to a failure event with probability  $\exp(-n^c)$ , we can replace these with the same Gaussian  $\tilde{g}$  which is independent of both  $x_*^{(i)}$  and  $x_*^{(j)}$ . We can do this for all pairs  $(i, j)$  and union bound over  $k$  (this is where we use the slow growing condition on  $k$ ). Hence, up to these error terms, which are negligible, the complementary event occurs with probability  $2^{-k}$ , so the test error will indeed be  $1 - O(2^{-k})$ .

## Multilabel supervision for multiclass classification

In this section, we expand upon the arguments to use weak multilabel supervision to train a weak-to-strong multiclass classifier  $f_{w_{2s}}$ . The difficulty of the multiclass analysis is studying the MNI behavior for weak supervision. However, since the multilabel MNI behavior is well under control by the arguments above, it is tractable to study this setup instead.

The key insight is that, for multiclass classification to succeed, it suffices to look at pairwise comparisons between the score functions for the  $k$  different classes (see [WS24] for more justification). In particular, one studies the relative survival given different class label vectors:

$$\text{SU}(v|\hat{y}_{\text{weak}}^{(i)}) - \text{SU}(v|\hat{y}_{\text{weak}}^{(j)}),$$

with the signal component for class  $i$  being the above quantity with  $v = v_*^{(i)}$ . One can then define the relative contamination in analogous way. However, the path towards studying the survival comes from understanding the coefficients of  $f_{\text{weak}}$  and  $f_{w_{2s}}$ , which is feasible.

Thus, to determine whether  $f_{w2s}$  generalizes for multiclass classification, it reduces back down to the multilabel/binary behavior. As argued in [WS24], because of the margin between the features and the survivals, it suffices to get a polynomially increasing SU/CN ratio. Hence, in the successful regime for weak-to-strong multilabel generalization, this relative SU/CN ratio is polynomially increasing, which implies that multiclass classification succeeds.

## 5.9 Improving the bounds for misclassification rate

In this section we tighten the bounds on misclassification rate for the multiclass setting.

**Theorem 5.9.1** (Tightening of [WS24], Proposition A.7). *Assume we are in the bi-level ensemble model (Definition 5.2.3), the true data generating process is 1-sparse (Assumption 5.2.2), and the number of classes follows the scaling defined in Definition 5.2.4. Then, in the negative regime where the model does not achieve vanishing error, we have*

$$\mathbf{E}[\ell(f_{\text{strong}})] = 1 - \Theta\left(\frac{1}{k}\right), \quad (5.29)$$

where the expectation is taken over the randomness of the training data and the test point.

At a high level, the misclassification event  $\mathcal{E}_{\text{err}}$  is governed by the following inequality holding:

$$\frac{\text{SU}_n}{\text{CN}_n} \max_{j \in [k]} |x_j| \leq \max_{i \in [k]} g_i,$$

where  $(x_j)_{j \in [k]}$  are iid standard normals, and  $(g_i)_{i \in [k]}$  are jointly Gaussian with some correlation structure.

Now, [WS24, Proposition A.3] implies that  $\frac{\text{SU}_n}{\text{CN}_n} \leq n^{-u}$  for some constant  $u > 0$  with probability  $1 - O(1/n)$ . Also, we can upper bound  $\max_{j \in [k]} |x_j| \leq O(\sqrt{\log k})$  with probability at least  $1 - O(1/k)$ . Moreover, in the regime where classification fails, we know that from [WS24, Proposition F.1], we have that  $\mathbf{E}[g_i g_j] \leq \frac{1}{2} + n^{-\delta}$  for some constant  $\delta > 0$  with probability at least  $1 - O(1/n)$ . Since  $k = o(n)$ , when we union bound over the above events, they get absorbed by the  $O(\cdot)$  and  $\Omega(\cdot)$  terms.

To tighten the misclassification rate, we will improve upon [LY22, Theorem 2.1]. In particular, we show the following lower tail inequality for the maximum of correlated Gaussians. We will prove the following theorem, which captures the lower tail behavior of the maximum of correlated Gaussians very far from its expectation. This result complements the lower tail bound of [LY22], which holds for more moderate deviations.

**Theorem 5.3.6** (Lower tail for correlated Gaussians). *Let  $\rho_0 \in (0, 1)$  be a parameter bounded away from 0 and 1, and let  $(g_i)_{i \in [N]}$  be jointly Gaussian with zero mean and unit variance. Suppose  $\mathbf{E}[g_i g_j] \leq \rho_0$  for all distinct  $i, j \in [N]$ . For any  $0 \leq t_N = \delta_0 \sqrt{2(1 - \rho_0) \log N}$  where  $\delta_0 \in [0, 1)$  is bounded away from 1, there is a constant  $C > 0$  depending only on  $\rho_0$  such that*

$$\Pr \left[ \max_{i \in [N]} g_i \leq t_N \right] \leq C \cdot N^{(1-\delta_0)^2(1-\frac{1}{\rho_0})} (\log N)^{\frac{1-\rho_0(2-\delta_0)-\delta_0}{2\rho_0}}.$$

*In particular, one can take  $C = \sqrt{\frac{\rho_0}{1-\rho_0}}$ . If we further have  $\mathbf{E}[g_i g_j] = \rho_0$  for all distinct  $i, j$  and  $t_N = O\left(\frac{\log \log N}{\sqrt{\log N}}\right)$ , then  $\Pr[\max_{i \in [N]} g_i \leq t_N] = \Theta\left(N^{1-\frac{1}{\rho_0}} (\log N)^{\frac{1}{2\rho_0}-1}\right)$ .*

Before we prove the theorem, let us see how it tightens the misclassification rate.

*Proof of Theorem 5.9.1.* We will apply Theorem 5.3.6 with  $N = k - 1$ ,  $\rho_0 = \frac{1}{2} + n^{-\delta}$ , and  $\delta_0 = \frac{1}{\log k}$ .

Noting that  $\delta_0 \sqrt{2(1 - \rho_0) \log k} \geq n^{-u}$  if  $k = c_k n^t$  for any  $t \in [0, 1)$ , we see that

$$\begin{aligned} \Pr \left[ \max_{i \in [k]} g_i \leq n^{-u} \right] &\leq \Pr \left[ \max_{i \in [k]} g_i \leq \delta_0 \sqrt{2(1 - \rho_0) \log k} \right] && (\delta_0 \sqrt{2(1 - \rho_0) \log k} \geq n^{-u}) \\ &\leq O \left( k^{-(1-\delta_0)^2(1-O(n^{-\delta}))} (\log k)^{-\frac{\delta_0}{2}+O(n^{-\delta})} \right) && \text{(Theorem 5.3.6)} \\ &\leq O \left( k^{-(1-\delta_0)^2} (\log k)^{-\frac{\delta_0}{2}} \right) (1 + o_n(1)) && (k^{O(n^{-\delta})} = 1 + o_n(1)) \\ &\leq O(k^{-(1-\delta_0)^2}) && ((\log k)^{-\delta_0/2} = 1 - o_k(1)) \\ &\leq O\left(\frac{1}{k}\right). && (\delta_0 = 1/\log k) \end{aligned}$$

Inverting this bound, we see that

$$\Pr \left[ \max_{i \in [k]} g_i > n^{-u} \right] \geq 1 - O\left(\frac{1}{k}\right).$$

For the upper bound on the above probability, we can use Slepian's lemma and then compare to Gaussians  $\underline{g}_i$  which have correlation  $\frac{1}{2} - n^{-\delta}$ . Writing it all out, we have

$$\begin{aligned} \Pr \left[ \max_{i \in [k]} g_i \leq n^{-u} \right] &\geq \Pr \left[ \max_{i \in [k]} \underline{g}_i \leq n^{-u} \right] && \text{(Slepian's lemma)} \\ &\geq \Pr \left[ \max_{i \in [k]} \underline{g}_i \leq 0 \right] \end{aligned}$$

$$\begin{aligned}
&\geq \Omega\left(\frac{1}{k^{1+n^{-\delta}}}\right) && ([\text{PSZ21, Theorem 2.1}]) \\
&\geq \Omega\left(\frac{1}{k}\right). && (k^{n^{-\delta}} = 1 + o_n(1))
\end{aligned}$$

from which we get a nearly matching upper bound on the probability of  $1 - \Omega(\frac{1}{k})$ .  $\square$

We return to the proof of Theorem 5.3.6.

*Proof of Theorem 5.3.6.* The lower bound directly follows from the proof of [LY22], so we focus on proving the upper bound. We remark that the constant hidden by  $\Theta$  can be pinpointed to  $(1 + o_N(1))\sqrt{\frac{\rho_0}{1-\rho_0}}$ , but we will not discuss this further.

To reduce confusion, we will attempt to follow the notation and treatment from [LY22]. To prove the upper bound, we can use Slepian's lemma to reduce to the case where  $\mathbf{E}[g_i g_j] = \rho_0$  for all  $i \neq j$ . Then, we can explicitly decompose

$$g_i = \sqrt{\rho_0}x + \sqrt{1 - \rho_0}h_i,$$

where  $x, h_i$  are iid standard Gaussians. Via this decomposition, we can write an integral representation for the desired probability:

$$\Pr\left[\max_{i \in [N]} g_i \leq t_N\right] = \int_{-\infty}^{\infty} \psi(s) ds, \quad (5.30)$$

$$\psi(s) \triangleq \phi(s) \Phi^N\left(\frac{t_N - \sqrt{\rho_0}s}{\sqrt{1 - \rho_0}}\right), \quad (5.31)$$

where  $\phi(\cdot)$  and  $\Phi(\cdot)$  are the standard Gaussian density and CDF, respectively. We will estimate this integral by splitting it into a couple pieces. To this end, we will bound the Gaussian CDF using Mills' inequality. For any  $t > 0$ , we have

$$\frac{t}{1+t^2} \phi(t) \leq 1 - \Phi(t) = \Phi(-t) \leq \frac{1}{t} \phi(t).$$

We will need the following multiplicative estimate on  $\Phi^N(\cdot)$ , which holds for any  $\delta \in [0, 1]$ :

$$\begin{aligned}
\Phi^N(\sqrt{2 \log N}(1 - \delta)) &= (1 - \Phi(-\sqrt{2 \log N}(1 - \delta)))^N \\
&\leq \left(1 - \frac{1}{\sqrt{2\pi}N^{(1-\delta)^2}(\sqrt{2 \log N}(1-\delta)+1)}\right)^N && (\text{Mill's inequality}) \\
&\leq \exp\left(-\frac{N^{1-(1-\delta)^2}}{\sqrt{2\pi}(\sqrt{2 \log N}+1)}\right). && (5.32)
\end{aligned}$$

The above bound (5.32) yields nontrivial bounds only when  $\delta > \frac{c \log \log N}{\log N}$  for a constant  $c > \frac{1}{4}$ , and decays superpolynomially if  $c > \frac{3}{4}$ .

These bounds motivate splitting (5.31) into a few different pieces, based on which term dominates the behavior of the integral  $\psi$ . To this end, let  $c_N, d_N > 0$  be parameters we specify shortly. Then we split the integral into three pieces:

$$\int_{-\infty}^{-c_N} \psi(s) ds + \int_{-c_N}^{-d_N} \psi(s) ds + \int_{-d_N}^{\infty} \psi(s) ds.$$

As in [LY22], we define  $\alpha_0 \triangleq (\frac{1}{\rho_0} - 1)(1 - \delta_0)^2$  and  $\beta_0 \triangleq \frac{\alpha_0}{1 - \delta_0}$ . Then the ultimate bound we want to prove is  $\int_{-\infty}^{\infty} \psi(s) ds \leq O\left(N^{-\alpha_0}(\log N)^{\frac{\beta_0 - 1}{2}}\right)$ . We explain the choice of  $c_N, d_N > 0$  as follows. For succinctness, we introduce the following two functions on  $\mathbb{R}$ :

$$s_N(u) \triangleq -\sqrt{\frac{2(1 - \rho_0) \log N}{\rho_0}} u = -\frac{\sqrt{2\alpha_0 \log N}}{1 - \delta_0} u$$

$$f_N(s) \triangleq \frac{t_N - \sqrt{\rho_0} s}{\sqrt{1 - \rho_0}},$$

and reparameterize the interval  $[-c_N, -d_N]$  as another interval  $\mathcal{I}$  on  $u$  below.

1. We want  $c_N$  to satisfy  $\Phi(-c_N) = O\left(N^{-\alpha_0}(\log N)^{\frac{\beta_0 - 1}{2}}\right)$ . Given this, the first term can be bounded by

$$\int_{-\infty}^{-c_N} \psi(s) ds \leq \int_{-\infty}^{-c_N} \phi(s) ds = \Phi(-c_N) \leq O\left(\frac{1}{N^{\alpha_0}}(\log N)^{\frac{\beta_0 - 1}{2}}\right).$$

By inverting Mills' inequality, we see that it suffices to pick

$$c_N \triangleq \sqrt{\frac{2(1 - \rho_0) \log N}{\rho_0}} \left(1 - \delta_0 - \frac{\log \log N}{4 \log N}\right) \tag{5.33}$$

$$= -s_N \left(1 - \delta_0 - \frac{\log \log N}{4 \log N}\right).$$

Indeed, we have for sufficiently large  $N$  that

$$\begin{aligned} \Phi(-c_N) &\leq \frac{\phi(c_N)}{c_N} \leq \frac{\exp\left(-\alpha_0 \log N \left(1 - \frac{\log \log N}{4(1 - \delta_0) \log N}\right)^2\right)}{\sqrt{2\pi} \sqrt{2\left(\frac{1}{\rho_0} - 1\right) \log N (1 - o_N(1))}} \\ &\leq \sqrt{\frac{\rho_0}{1 - \rho_0}} N^{-\alpha_0} (\log N)^{\frac{\beta_0 - 1}{2}} \end{aligned} \tag{5.34}$$

$(\beta_0 = \frac{\alpha_0}{1 - \delta_0})$

2. We pick  $d_N$  such that  $\Phi^N(f_N(-d_N)) \ll \frac{1}{N^{\alpha_0}}(\log N)^{\frac{\beta_0-1}{2}}$ . Given this, by monotonicity we evidently have

$$\int_{-d_N}^{\infty} \psi(s) ds \leq \Phi^N(f_N(-d_N)) \ll \frac{1}{N^{\alpha_0}}(\log N)^{\frac{\beta_0-1}{2}}.$$

Owing to (5.32), to achieve the desired superpolynomial decay it suffices to pick  $d_N \geq 0$  such that  $\frac{t_N + \sqrt{\rho_0} d_N}{\sqrt{1-\rho_0}} \leq \sqrt{2 \log N} (1 - (\frac{3}{4} + \varepsilon) \frac{\log \log N}{\log N})$ , where  $\varepsilon > 0$  is a constant.

Recalling that  $t_N = \delta_0 \sqrt{2(1-\rho_0) \log N}$ , we see that  $f_N(-d_N) = \sqrt{2 \log N} \delta_0 + \frac{\sqrt{\rho_0}}{\sqrt{1-\rho_0}} d_N$ . Hence the desired inequality holds by picking

$$\begin{aligned} d_N &\triangleq \sqrt{\frac{2(1-\rho_0) \log N}{\rho_0}} \left( 1 - \delta_0 - \left( \frac{3}{4} + \varepsilon \right) \frac{\log \log N}{\log N} \right) \\ &= -s_N \left( 1 - \delta_0 - \left( \frac{3}{4} + \varepsilon \right) \frac{\log \log N}{\log N} \right). \end{aligned} \quad (5.34)$$

Given the above choices of  $c_N, d_N$  in Eq. (5.33) and (5.34), we see that the interval  $\mathcal{I}$  that  $u$  belongs to is  $\mathcal{I} \triangleq [1 - \delta_0 - (\frac{3}{4} + \varepsilon) \frac{\log \log N}{\log N}, 1 - \delta_0 - \frac{1}{4} \cdot \frac{\log \log N}{\log N}]$ . Hence, it is natural to reparameterize the integral in terms of  $\eta \in [0, \frac{1}{2} + \varepsilon]$  via the following change of variables

$$u_N(\eta) \triangleq 1 - \delta_0 - \frac{(\eta + \frac{1}{4}) \log \log N}{\log N},$$

and an easy computation yields

$$ds = s'_N(u_N(\eta)) u'_N(\eta) d\eta = \frac{\sqrt{2\alpha_0}}{1 - \delta_0} \cdot \frac{\log \log N}{\sqrt{\log N}} d\eta.$$

It is not hard to see that  $s_N(u_N([0, \frac{1}{2} + \varepsilon])) = [-c_N, -d_N]$ . Let us introduce the following abbreviations for the compositions of our changes of variable:

$$\begin{aligned} \tilde{s}_N(\eta) &\triangleq s_N(u_N(\eta)) \\ \tilde{f}_N(\eta) &\triangleq f_N(\tilde{s}_N(\eta)) = \sqrt{2 \log N} \left( 1 - \frac{(\eta + \frac{1}{4}) \log \log N}{\log N} \right). \end{aligned}$$

In this notation, we have

$$\int_{-c_N}^{-d_N} \psi(s) ds = \frac{\sqrt{2\alpha_0}}{1 - \delta_0} \cdot \frac{\log \log N}{\sqrt{\log N}} \int_0^{1/2+\varepsilon} \psi(\tilde{s}_N(\eta)) d\eta.$$

Now the argument as in [LY22, Eq. 4.22] shows that this integral is  $O(N^{-\alpha_0}(\log N)^{\frac{\beta_0-1}{2}})$ , which completes the proof of the first part of the theorem.<sup>5</sup> However, let us give an alternative proof which is a bit simpler. We will estimate the integral by performing a Riemann sum with subintervals of width  $\tau > 0$ , which we pick to satisfy  $\beta_0\tau < 1$  and such that  $\tau$  evenly divides  $\frac{1}{2} + \varepsilon$ . Then, we have

$$\int_0^{1/2+\varepsilon} \psi(\widetilde{s}_N(\eta)) d\eta \leq \sum_{k=0}^{\frac{1/2+\varepsilon}{\tau}} \int_{k\tau}^{(k+1)\tau} \psi(\widetilde{s}_N(\eta)) d\eta.$$

By monotonicity, for any  $k$  we have

$$\begin{aligned} \int_{k\tau}^{(k+1)\tau} \psi(\widetilde{s}_N(\eta)) d\eta &\leq \tau \cdot \phi(\widetilde{s}_N((k+1)\tau)) \Phi^N(\widetilde{f}_N(k\tau)) \\ &\leq C \cdot \tau \cdot N^{-\alpha_0} (\log N)^{\beta_0 \cdot (\frac{1}{2} + 2(k+1)\tau)} \exp\left(-(\log N)^{2k\tau}\right), \end{aligned}$$

where  $C$  is a universal constant. Here, the last line used  $\phi(\widetilde{s}_N(\eta)) = O(N^{-\alpha_0}(\log N)^{2\beta_0 \cdot (\frac{1}{4} + \eta)})$  and  $\Phi^N(\widetilde{f}_N(\eta)) = O(\exp(-\log N)^{2\eta})$ .

If  $2\beta_0(k+1)\tau < \frac{1}{2}$ , then  $\frac{\log \log N}{\log N} (\log N)^{2\beta_0(k+1)\tau} \ll (\log N)^{-\frac{1}{2}}$ , as desired. On the other hand, if  $2\beta_0(k+1)\tau \geq \frac{1}{2}$ , then as  $\beta_0\tau < \frac{1}{8}$ , we have  $2k\tau \geq \frac{1}{4\beta_0} > 0$ . Since  $\exp(-(\log N)^{2k\tau})$  dominates any polylog terms, so it is not hard to see that the total contribution here is  $\ll N^{-\alpha_0}(\log N)^{\frac{\beta_0-1}{2}}$ . Hence, we conclude that  $\int_{-c_N}^{-d_N} \psi(s) ds \leq \sqrt{\frac{\rho_0}{1-\rho_0}} N^{-\alpha_0} (\log N)^{\frac{\beta_0-1}{2}}$ , as desired.  $\square$

## 5.10 Experiments

In this section, we describe the simulations we conducted to validate the theory. We generated Gaussian data following the subset ensembles specified in the figures, and constructed two linear models from them: the MNI classifier and the simple averaging classifier. In the averaging classifier, we average over the positive examples of a label, which approximates the behavior of the first few iterations of gradient descent. In contrast, the MNI classifier governs the asymptotic behavior of gradient descent.

The weak-to-strong behavior for these two learning algorithms was compared to two other baselines: the weak accuracy for  $f_{\text{weak}}$  and the accuracy for the strong model trained

---

<sup>5</sup>To be explicit, their calculation requires us to verify that  $\sqrt{2 \log N}(\delta_0 + u_N(\eta)) = \omega(1)$ , which is certainly true.



on  $m$  clean labels:  $f_{\text{strong}}$ . The test accuracies were evaluated on  $n_{\text{test}} = 100$  fresh datapoints. We ran 8 independent trials to train  $f_{\text{weak}}$  with  $n = 50$  so that we could explore how the weak-to-strong behavior scales with  $p$  and  $u$ . For each  $f_{\text{weak}}$ , we conducted 16 independent trials to train  $f_{\text{w2s}}$ .

We swept out  $u$  using five equally spaced points in  $[1, 1.3]$ . In Figures 5.3 and 5.4, we show the results of the averaging and MNI experiments, respectively, for four different slices. In the top row, we show two slices where the theory predicts weak-to-strong generalization to occur for MNI, and in the bottom row we show two slices where the theory predicts failure of weak-to-strong generalization. The error bars show the estimated 95% CI over all sources of uncertainty in the inner and outer loop ( $f_{\text{w2s}}$  and  $f_{\text{weak}}$ ). In both the averaging and MNI plots, the theory successfully predicts whether weak-to-strong generalization occurs. Furthermore, in every plot the ground truth trained strong model  $f_{\text{strong}}$  trained on  $m$  clean labels has better test accuracy than  $f_{\text{weak}}$  and  $f_{\text{w2s}}$ , as expected. Another interesting experimental observation is that the averaging classifier does significantly better than MNI in non-asymptotic settings. This corroborates the view of practitioners of the benefits of early-stopping for gradient descent.

## 5.11 Heuristic calculations

Recall from the definition of MNI that

$$f_{\text{strong}} = \mathbf{X}^\top \mathbf{A}^{-1} \mathbf{y},$$

where  $\mathbf{y} \in \{\pm 1\}^n$  is a label vector generated by either the true feature  $x_*$  or a weak feature  $x_{\text{weak}}$  and  $\mathbf{A} = \mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{n \times n}$  is the Gram matrix. The key to our analysis is studying the survival and contamination of various features when the labels are possibly generated by another feature. Recall that we performed a basis change  $\mathbf{X} \mapsto \mathbf{X}\mathbf{U}$ , so that the strong features are drawn iid from  $N(0, \Lambda)$ .

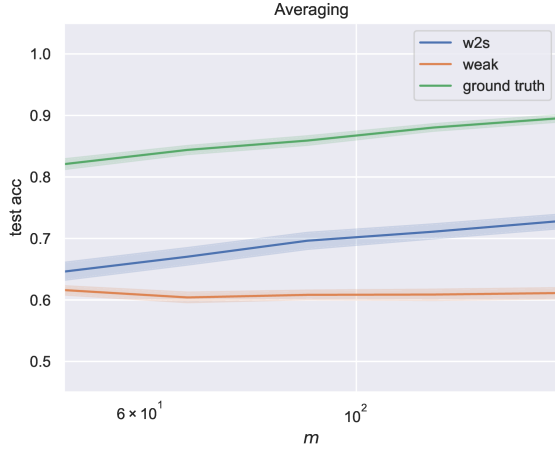
Writing the transformed data matrix now as

$$\mathbf{X} = [\sqrt{\lambda_1} \mathbf{z}_1 \quad \sqrt{\lambda_2} \mathbf{z}_2 \quad \dots \quad \sqrt{\lambda_d} \mathbf{z}_d],$$

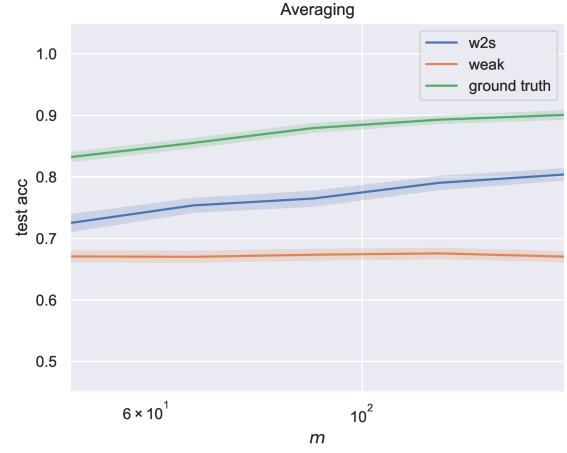
where each  $\mathbf{z}_i \sim N(0, I_n)$ , we obtain for any unit norm  $v \in \mathbb{R}^D$  that

$$\text{SU}(v) = \sum_{i \in [d]} \lambda_i \mathbf{z}_i^\top \mathbf{A}^{-1} \mathbf{y} \langle \mathbf{v}_i, v \rangle \tag{MNI}$$

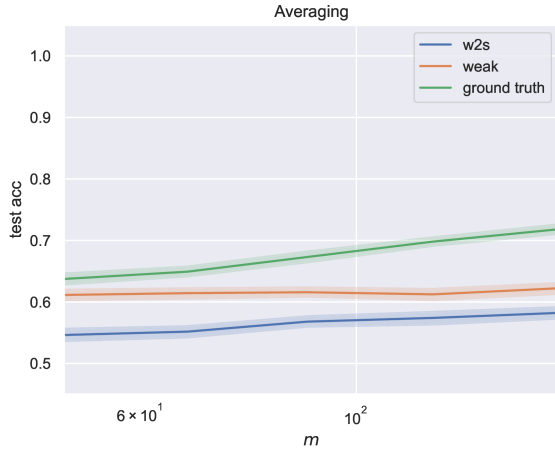
$$\text{CN}(v) = \sqrt{\sum_{i \in [d]} (\lambda_i \mathbf{z}_i^\top \mathbf{A}^{-1} \mathbf{y})^2 (1 - \langle \mathbf{v}_i, v \rangle^2)}. \tag{Orthonormality of } \mathbf{v}_i$$



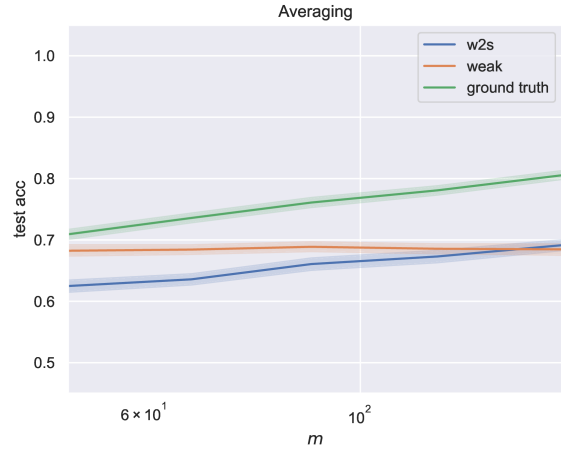
(a)  $(p, q, r) = (2, 0.6, 0.6)$ ,  $(p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) = (1.4, 0.9, 0.5)$ .



(b)  $(p, q, r) = (2, 0.9, 0.4)$ ,  $(p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) = (1.4, 0.9, 0.4)$

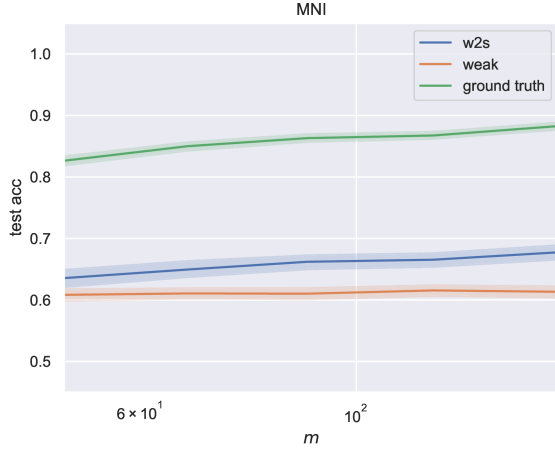


(c)  $(p, q, r) = (1.5, 0.6, 0.8)$ ,  $(p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) = (1.4, 0.9, 0.5)$ .

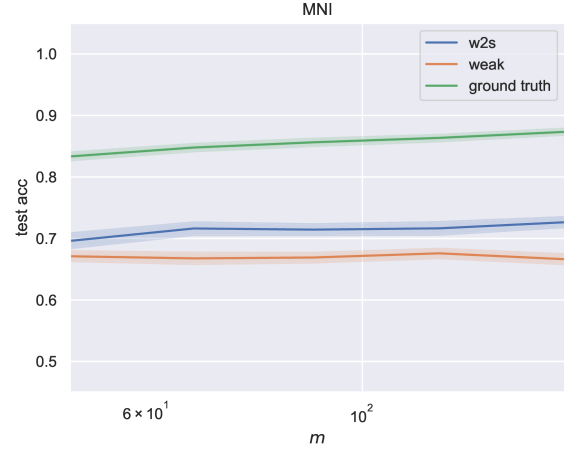


(d)  $(p, q, r) = (2, 0.9, 0.4)$ ,  $(p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) = (1.4, 0.9, 0.4)$

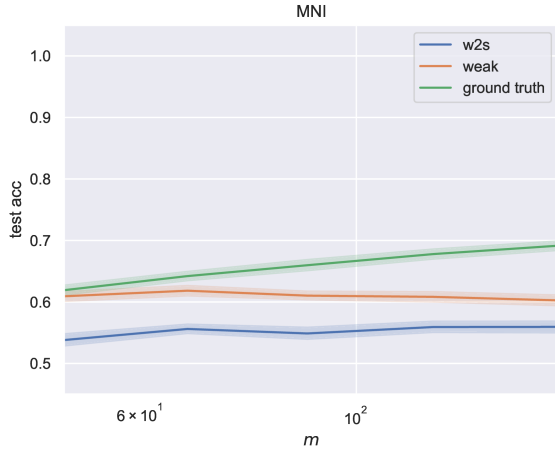
Figure 5.3: Comparison of test accuracies for four different models using averaging training. The  $x$ -axis plots  $m$ , the number of additional labeled datapoints. The models are trained using class averaging, which approximates the behavior of the initial few gradient descent iterations. Note how the weak model has low accuracy, whereas the weak-to-strong model and ground truth have higher accuracies that increase as  $m$  increases. The top row Figures 5.3a and 5.3b are in a regime where we predict MNI weak-to-strong generalization to succeed, whereas the bottom row Figures 5.3c and 5.3d depict regimes where we expect MNI weak-to-strong generalization to fail.



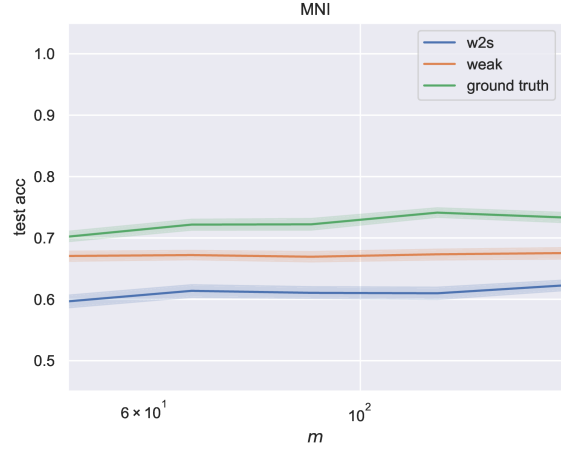
(a)  $(p, q, r) = (2, 0.6, 0.6), (p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) = (1.4, 0.9, 0.5)$ .



(b)  $(p, q, r) = (2, 0.9, 0.5), (p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) = (1.4, 0.9, 0.5)$



(c)  $(p, q, r) = (1.5, 0.6, 0.8), (p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) = (1.4, 0.9, 0.5)$ .



(d)  $(p, q, r) = (2, 0.9, 0.4), (p_{\text{weak}}, q_{\text{weak}}, r_{\text{weak}}) = (1.4, 0.9, 0.4)$

Figure 5.4: Comparison of MNI test accuracies for four different models. Observe how the weak-to-strong accuracy increases as  $m$  increases. Again, the top row Figures 5.4a and 5.4b are in a regime where we predict MNI weak-to-strong generalization to succeed, whereas the bottom row Figures 5.4c and 5.4d depict regimes where we expect MNI weak-to-strong generalization to fail. The plots corroborate these theoretical predictions.

Since  $A$  is close to  $dI_n$  in the isotropic case, and we are working in the regime where PCA fails to extract the bi-level structure ( $q + r > 1$ ), one could hope for the best and pretend that  $A^{-1} = \frac{1}{d}I_n$ . This step is not rigorous, but we will justify these approximations in Section 5.6.

Based on the decompositions in Definition 5.3.7, we study the case where  $\mathbf{y} = \text{sgn}(\langle \mathbf{g}, \mathbf{w} \rangle)$  for some  $\mathbf{w} \in \mathbb{R}^D$  but we want to recover the planted direction  $\mathbf{v} \in \mathbb{R}^D$ . For a subspace  $V \subseteq \mathbb{R}^D$  and a vector  $\mathbf{u} \in \mathbb{R}^D$ , let  $\mathbf{u}_V$  denote the projection of  $\mathbf{u}$  onto  $V$ . For axis-aligned subspaces  $V \subseteq [d]$ , this just corresponds to restricting to the coordinates in  $V$ . To simplify the heuristic calculation, we will make the following assumptions.

**Assumption 5.11.1.** Let  $S = [s]$  denote the spiked subspace after the basis change, and let  $\alpha, \rho > 0$  be parameters possibly depending on  $n$ . For any vector  $\mathbf{u} \in \mathbb{R}^d$ , let  $T_u \triangleq \left\{ i \in S : |u_i| = \omega\left(\frac{1}{\sqrt{n}}\right) \right\}$  denote the spiked coordinates where  $\mathbf{u}$  is large, and  $R_u = S \setminus T_u$  denote the spiked coordinates where  $\mathbf{u}$  is small.

We assume the following holds for  $T = T_w$  and  $R = R_w$ :

- (1) We have  $\langle \mathbf{v}_T, \mathbf{w}_T \rangle = \alpha \|\mathbf{v}_T\|_2 \|\mathbf{w}_T\|_2$ . In other words,  $\mathbf{v}$  and  $\mathbf{w}$  have correlation  $\alpha$  restricted to the heavy coordinates for  $\mathbf{w}$ .
- (1) We have  $\sum_{i \in T} v_i^2 w_i^2 = \rho^2 \|\mathbf{w}_T\|_2^2$ . Note that by  $L^p$  norm inequalities we always have  $\rho^2 \leq 1$ .
- (1) We have  $|R| = \Omega(s)$ , i.e. a constant fraction of  $\mathbf{w}$ 's spiked coordinates are small.

**Remark 5.11.2.** When  $\mathbf{v} = \mathbf{w}$ , Assumption 5.11.1 can be thought of as a relaxed notion of  $\mathbf{v}$  being 1-sparse; for a given  $\rho$  one should roughly think of  $\mathbf{v}$  as being  $\frac{1}{\rho}$ -sparse.

**Survival bound.** Recall that  $\mathbf{y} = \text{sgn}(\langle \mathbf{g}, \mathbf{w} \rangle)$ . By applying the noise stability formula again and the fact that  $z_{ij} \sim N(0, 1)$ , we deduce that

$$\begin{aligned}
\mathbf{E}[z_i^\top \mathbf{y}] &= n \mathbf{E}[z_{ij} y_j] \\
&= \Pr[\text{sgn}(z_{ij}) = y_j] \mathbf{E}[|z_{ij}|] - (1 - \Pr[\text{sgn}(z_{ij}) = y_j]) \mathbf{E}[|z_{ij}|] \\
&= \sqrt{\frac{2}{\pi}} (2 \Pr[\text{sgn}(z_{ij}) = y_j] - 1) \\
&= n \left(\frac{2}{\pi}\right)^{3/2} \arcsin w_i,
\end{aligned}$$

where in the second to last line we have used the fact that the expected magnitude of a standard Gaussian is  $\sqrt{2/\pi}$ , and in the last line we have used the noise stability formula. By standard concentration inequalities, the deviations will be of order  $O(\sqrt{n})$ . As the behavior of  $\frac{2}{\pi} \arcsin(x) \approx \frac{2}{\pi} x$  for small  $x$ , we deduce that the expectation will dominate whenever  $|w_i| \gg \frac{1}{\sqrt{n}}$ .

We will now plug in the bi-level scaling. Recall that  $\lambda_i = \lambda_F = \frac{ad}{s}$  for  $i \in [s]$  and  $\lambda_i = \lambda_U = \frac{(1-a)d}{d-s}$  for  $i > s$ .

Thus, ignoring constants, with high probability we should have

$$\text{SU}(v|w) = \frac{\lambda_F n}{d} \sum_{i \in T_w} v_i \arcsin w_i \pm \frac{\lambda_F \sqrt{n}}{d} \sum_{i \in R_w} v_i \pm \frac{\lambda_U \sqrt{n}}{d} \sum_{i > s} v_i.$$

Observe that

$$\frac{2}{\pi}|x| \leq \frac{2}{\pi}|\arcsin x| \leq |x|,$$

for all  $x \in [-1, 1]$ .

Consequently, for  $i \in T_w$ , each summand contributes  $v_i \arcsin(w_i) \approx \Theta(v_i w_i)$ .<sup>6</sup> Thus, by Assumption 5.11.1 and the definition of  $\lambda_F$ , the first term is  $\Theta(\alpha \cdot \frac{an}{s} \cdot \|v_{T_w}\|_2 \|w_{T_w}\|_2)$ .

For the second term, we will upper bound its magnitude by

$$\begin{aligned} \frac{\lambda_F \sqrt{n}}{d} \left| \sum_{i \in R_w} v_i \right| &\leq \frac{an}{s} \cdot \frac{\|v_{R_w}\|_1}{\sqrt{n}} \\ &\leq \frac{an}{s} \cdot \sqrt{\frac{s}{n}} \cdot \|v_{R_w}\|_2 \quad (R_w \subseteq S) \\ &\leq \frac{an}{s} \cdot \sqrt{\frac{s}{n}}. \quad (\|v\|_2 = 1) \end{aligned}$$

Finally, for the third term, since  $|\sum_{i > s} v_i| \leq \|v\|_1 \leq \sqrt{d}$ , after plugging in the definition of  $\lambda_U$ , we obtain the asymptotics

$$\text{SU}(v|w) \asymp \frac{an}{s} \cdot \left( \alpha \cdot \|v_{T_w}\|_2 \|w_{T_w}\|_2 \pm \sqrt{\frac{s}{n}} \right) \pm \sqrt{\frac{n}{d}}. \quad (5.35)$$

Note that for the first term to dominate, we must have  $\alpha = \Omega(\frac{s}{n})$ . Also, if we want to improve our estimate on the second term, we can further split it by  $T_v$ , we can gain and get deviations of order

$$\begin{aligned} \left| \sum_{i \in R_w \cap T_v} v_i + \sum_{i \in R_w \cap R_v} v_i \right| &\leq \|v_{R_w \cap T_v}\|_1 + \frac{|R_v|}{\sqrt{n}} \\ &\leq \|v_{R_w \cap T_v}\|_1 + \frac{s}{\sqrt{n}}, \end{aligned} \quad (\text{Definition of } R_v)$$

yielding an ultimate relative deviation of  $\frac{\|v_{R_w \cap T_v}\|_1}{\sqrt{n}} + \frac{s}{n}$ . This is significantly better if, say  $|T_v| = o(s)$ , as it allows us to beat the contamination bounds which have relative deviations  $\sqrt{\frac{s}{n}}$ .

---

<sup>6</sup>The reason this is not an equality is that there might be some heavy  $w'_i$ 's which disagree in sign with  $v'_i$ 's, but for the settings we consider this estimate will be true.

**Contamination bound.** For succinctness, introduce the shorthand  $h_i^2 = (1 - v_i^2) \langle \mathbf{z}_i, \mathbf{y} \rangle^2$ . Then the squared contamination is

$$\begin{aligned}
\text{CN}(\mathbf{v}|\mathbf{w})^2 &= \sum_{i \in [d]} \lambda_i^2 (1 - v_i^2) \cdot \mathbf{y}^\top \mathbf{A}^{-1} \mathbf{z}_i \mathbf{z}_i^\top \mathbf{A}^{-1} \mathbf{y} \\
&\approx \frac{1}{d^2} \sum_{i \in [d]} \lambda_i^2 h_i^2 \quad (\mathbf{A}^{-1} \approx \frac{1}{d} \mathbf{I}_d) \\
&= \frac{\lambda_F^2}{d^2} \sum_{i \in T_w} h_i^2 + \frac{\lambda_F^2}{d^2} \sum_{i \in R_w} h_i^2 + \frac{\lambda_U^2}{d^2} \sum_{i > s} h_i^2 \\
&\leq \frac{\lambda_F^2 n^2}{d^2} \sum_{i \in T_w} (1 - v_i^2) \left( \frac{2}{\pi} \arcsin w_i \right)^2 + \frac{\lambda_F^2 n}{d^2} \sum_{i \in R_w} (1 - v_i^2) + \frac{\lambda_U^2 n}{d^2} (d - s),
\end{aligned}$$

where in the last line we have used the observation that the expectation of  $h_i$  dominates if and only if  $i \in T_w$ .

The first term can be bounded (up to constants) as

$$\left( \frac{an}{s} \right)^2 \sum_{i \in T_w} (1 - v_i^2) w_i^2 = \left( \frac{an}{s} \right)^2 \|\mathbf{w}_{T_w}\|_2^2 (1 - \rho^2). \quad (\text{Assumption 5.11.1})$$

For the second term, we can bound up to constants as

$$\begin{aligned}
\left( \frac{an}{s} \right)^2 \cdot \frac{1}{n} \cdot \sum_{i \in R_w} (1 - v_i^2) &= \left( \frac{an}{s} \right)^2 \cdot \frac{|R_w| - \|\mathbf{v}_{R_w}\|_2^2}{n} \\
&\leq \left( \frac{an}{s} \right)^2 \cdot \frac{|R_w|}{n}. \quad (\|\mathbf{v}\|_2 = 1)
\end{aligned}$$

Finally, the third term can be bounded by  $\frac{n}{d}$ . Putting these together, we conclude that

$$\text{CN}(\mathbf{v}|\mathbf{w}) \asymp \left( \frac{an}{s} \right) \left( \sqrt{1 - \rho^2} \|\mathbf{w}_{T_w}\|_2 + \sqrt{\frac{|R_w|}{n}} \right) + \sqrt{\frac{n}{d}}. \quad (5.36)$$

Let  $\mu_n = \frac{an}{s}$ . In our regime,  $\mu_n \ll 1$  because  $q + r > 1$ . Combining Eq. (5.35) and (5.36) yields

$$\frac{\text{SU}(\mathbf{v}|\mathbf{w})}{\text{CN}(\mathbf{v}|\mathbf{w})} \asymp \frac{\mu_n \cdot \left( \alpha \|\mathbf{v}_{T_w}\|_2 \|\mathbf{w}_{T_w}\|_2 \pm \frac{\|\mathbf{v}_{R_w \cap T_v}\|_1}{\sqrt{n}} \pm \frac{s}{n} \right) \pm \sqrt{\frac{n}{d}}}{\mu_n \left( \sqrt{1 - \rho^2} \|\mathbf{w}_{T_w}\|_2 + \sqrt{\frac{|R_w|}{n}} \right) + \sqrt{\frac{n}{d}}} \quad (5.37)$$

Hence, for the survival to contamination ratio to grow with  $n$ , we need

$$\alpha \|v_{T_w}\|_2 \gg \sqrt{1 - \rho^2} \quad (\text{Weak supervision})$$

$$\alpha \|v_{T_w}\|_2 \|w_{T_w}\|_2 \gg \sqrt{\frac{|R_w|}{n}} \quad (\text{Favored contamination})$$

$$\mu_n \alpha \|v_{T_w}\|_2 \|w_{T_w}\|_2 \gg \sqrt{\frac{n}{d}}$$

Let us put these scalings together to predict the scaling regimes for weak-to-strong generalization. For strong generalization, we have  $v = v_*$  and  $y = \text{sgn}(\langle g, v_* \rangle)$ . From the discussion in Section 5.3, we know that the strong learner generalizes if

$$\frac{\text{SU}(v_*|v_*)}{\text{CN}(v_*|v_*)} = \omega_n(1),$$

and fails to generalize if the ratio is  $o_n(1)$ . Under Assumption 5.2.2, we have  $v_* = e_1$ , so  $T_v = T_w = \{1\}$ ,  $\alpha = 1$ , and  $\rho = 1$ , and the expression simplifies to

$$\frac{\text{SU}(v_*|v_*)}{\text{CN}(v_*|v_*)} \asymp \frac{\mu_n}{\mu_n \cdot \sqrt{\frac{s}{n}} + \sqrt{\frac{n}{d}}},$$

which under the bi-level parameter scaling verifies the conditions for ground truth supervision.

This completes the proof sketch; it remains to justify the above estimates rigorously. In the subsequent subsections, we will assume that the above scalings of the survival and contamination are correct and use them to deduce that the 1-sparse assumption is necessary to get a sharp transition in the test error. These calculations can be upgraded to rigorous proofs using the tools developed in Section 5.6.

## The necessity of 1-sparse assumption

Let's suppose we get clean labels from  $\text{sgn}(\langle g, v \rangle)$  and want to learn the unit vector  $v$ . We will abbreviate  $T = T_v = T_w$ . In this case, we have  $\alpha = 1$  and  $\rho^2 = \frac{\|v_T\|_4^4}{\|v_T\|_2^2}$  in Assumption 5.11.1.

**Lemma 5.11.3.** *Suppose we are given labels according to  $v$  and want to learn  $v$ . Then, the survival to contamination ratio is  $\omega_n(1)$  only if*

$$\frac{\|v_T\|_4^4}{\|v_T\|_2^2} = 1 - o(1).$$

*In particular, the above condition holds only if the following two conditions hold:*

$$(3) \quad \|v_T\|_2^2 = 1 - o(1).$$

$$(3) \quad \text{For each } i \in T, \text{ either } v_i = o(1) \text{ or } v_i = 1 - o(1).$$

The upshot is that having 1-sparse labels is necessary for obtaining asymptotically perfect generalization.

*Proof.* Hence, focusing only on the survival terms coming from  $T$ , which are the only relevant coordinates for learning, and lower bounding the contamination with just the  $T$  terms, we have

$$\frac{\text{SU}(v|v)}{\text{CN}(v|v)} \leq \frac{\frac{an}{s} \|v_T\|_2^2}{\frac{an}{s} \sqrt{\|v_T\|_2^2 - \|v_T\|_4^4}} = \frac{\|v_T\|_2}{\sqrt{1 - \frac{\|v_T\|_4^4}{\|v_T\|_2^2}}} \leq \frac{1}{\sqrt{1 - \frac{\|v_T\|_4^4}{\|v_T\|_2^2}}},$$

where the last inequality used the fact that  $v$  is unit norm. This proves the first necessary condition.

We show that if  $\|v_T\|_4^4 / \|v_T\|_2^2 = 1 - o(1)$ , then the second set of necessary conditions hold.

Indeed, for the first claim, suppose  $\|v_T\|_2^2 \leq 1 - \varepsilon$  for some constant  $\varepsilon > 0$ . The  $L^p$  norm inequalities imply that  $\|v_T\|_4^4 \leq \|v_T\|_2^4 \leq (1 - \varepsilon) \|v_T\|_2^2$ , so the ratio is at most  $1 - \varepsilon$ , a contradiction. For the second claim, suppose instead there is a coordinate  $i$  with  $v_i^2 = 1 - \varepsilon$  for some constant  $\varepsilon \in (0, 1)$ . Then, we have  $\|v_{T \setminus \{i\}}\|_4^4 \leq \|v_{T \setminus \{i\}}\|_2^4 \leq \varepsilon$ , but then  $\|v_T\|_4^4 \leq (1 - \varepsilon)^2 + \varepsilon \leq 1 - \Omega(\varepsilon)$ , a contradiction. □



# Chapter 6

## Synthetic Error Injection for Self Correction in Language Models

### 6.1 Introduction

The current paradigm for eliciting reasoning in large language models is reinforcement learning (RL) [GYZ<sup>+</sup>25]. These methods have driven substantial gains in mathematics and coding benchmarks across both closed and open-source models [JKL<sup>+</sup>24, CBS<sup>+</sup>25, BBC<sup>+</sup>25, LMP<sup>+</sup>24]. Such systems often generate an extended chain of thought (CoT) before producing a final answer; within these CoTs, we observe emergent self-reflection and error correction, as the model outlines a solution strategy and then identifies and revises mistakes.

Is it possible to obtain these self-reflection abilities without resorting to reinforcement learning, which is often computationally expensive? The robotics literature offers an illustrative analog. Goff et al. [GHH<sup>+</sup>25] trained a self-driving policy purely via imitation learning [SSHB18], where the ground-truth trajectory is always correct. When deployed on a real car, the policy gradually drifted until it straddled the lane boundary.

This behavior is a classic instance of distribution shift [RGB11]: as errors in real-world dynamics accumulate, a model trained only on “good” observations has little guidance on how to recover to the center of the lane. Goff et al. [GHH<sup>+</sup>25] mitigated this by artificially and randomly perturbing the camera pose during training using an explicit world model, then supervising the policy to steer back into the lane.

In this paper, we adopt a similar idea for reasoning language models (Figure 6.1). We use a purely supervised approach: insert synthetic errors into the CoT, mask them, and add a supervised correction step so the model learns to recognize and recover from its own mistakes.

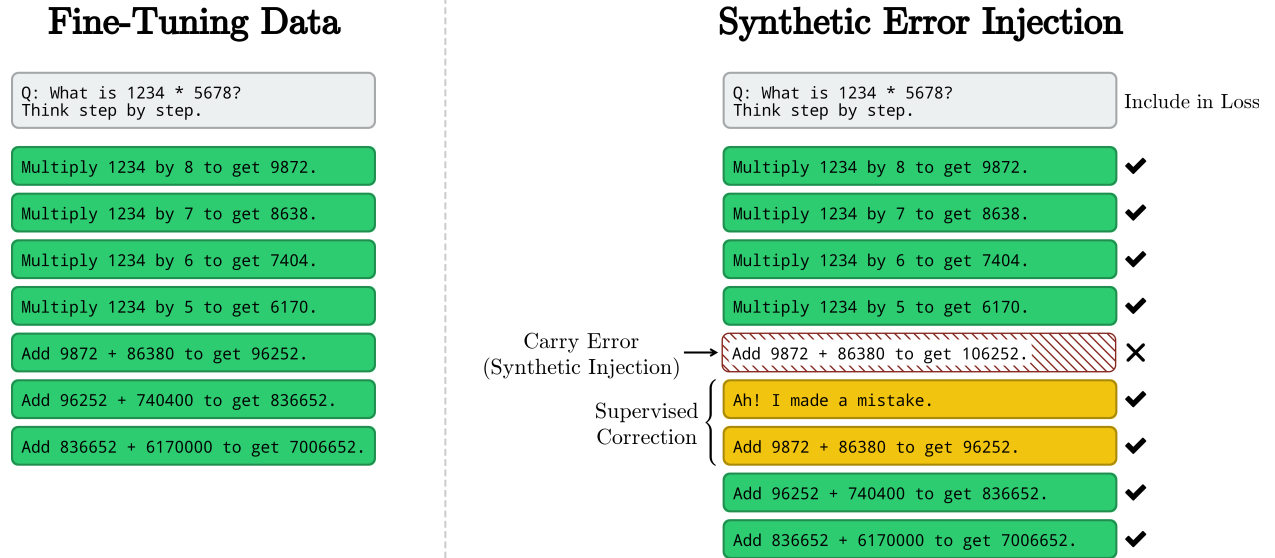


Figure 6.1: An illustration of the synthetic error injection process. The (Left) panel shows a correct Chain-of-Thought (CoT) trace used for standard fine-tuning (FT). The (Right) panel details the Error Injection Fine-Tuning (EIFT) methodology, where a correct step from a golden CoT is replaced by a three-part sequence: (1) a synthetically injected erroneous step (e.g., a “Carry Error”), (2) an explicit error recognition step (“Ah! I made a mistake.”), and (3) the original correct step as a supervised correction. As indicated by the “Include in Loss?” column, the loss is computed for the recognition and correction steps but masked for the synthetically injected error.

Surprisingly, even on toy tasks, the method fails to yield significant performance improvements across all tested models. We explore two potential hypotheses:

- (i) We investigate the “solver-verifier gap”—the difference in difficulty of generating and checking a step—but find that the method’s failure persists even when checking is significantly easier.
- (i) We propose that the true issue is a distribution mismatch. While our synthetic error distribution successfully covers the overall distribution of real errors, the model only learns to reliably correct synthetic errors. It fails to generalize this error-correction capability to the distribution of context-dependent errors that the base model itself is prone to making.

Our contributions in this paper are therefore twofold. First, we formally test the hypothesis that supervised error injection can elicit self-correction in LLMs, and we demonstrate its empirical failure. Second, our analysis pinpoints the reason for this failure: a critical

mismatch between synthetic error distributions and the model’s own innate error modes. This finding suggests that successful self-correction requires more precisely matching error distributions: it must target the specific, context-dependent failures a model is prone to, suggesting why on-policy methods like RL have proven uniquely effective.

## 6.2 Related work

Chain-of-thought (CoT) reasoning [WWS<sup>+</sup>22] is a popular approach for imparting reasoning abilities to language models. Recently, reinforcement learning (RL) [JKL<sup>+</sup>24, Ope25]—especially stabilization techniques such as GRPO [GYZ<sup>+</sup>25]—have emerged as the dominant paradigm for teaching models to reason and to self-correct. There is growing interest in this direction, including work such as [KZA<sup>+</sup>24], which uses RL to train models to recognize and correct their own errors.

There has also been substantial effort to improve reasoning and self-correction abilities *without* RL [SCG<sup>+</sup>23], given its computational cost. Approaches include using one LLM as a feedback model for another [MTG<sup>+</sup>23, SDYD<sup>+</sup>23], generate-then-rank strategies [HZR22, WZX<sup>+</sup>23], and feedback engines that guide autoregressive decoding [YDC22, XKZ<sup>+</sup>23]. For instance, Muennighoff et al. [MYS<sup>+</sup>25] propose a simple technique: append the token “Wait” multiple times when the model attempts to stop, thereby lengthening the reasoning chain and improving performance—without any reinforcement learning.

[KZA<sup>+</sup>24] also shows that offline–online distribution mismatch can yield poor on-policy self-correction. We examine this failure mode on even simpler tasks than MATH [HBK<sup>+</sup>21], specifically small multiplication and Sudoku problems. By constructing a covering set of plausible modeling errors, we show that even such coverage is insufficient, strengthening their result.

Synthetic error injection is widely used in robotics [GHH<sup>+</sup>25, SSHB18]. The work most closely related to ours is [YZW<sup>+</sup>25], which trains a language model to play Countdown and to backtrack in a tree-search harness using only supervised data and synthetic error injection. However, their method primarily distills the DFS procedure used to solve Countdown, rather than teaching the model to correct its own errors, and requires a costly sampling harness to generate solutions. In contrast, we teach the model to backtrack fully in a single linear CoT.

## 6.3 Experimental setup

We trained our models on a cluster of A100 and A6000 GPUs. For each model in {Qwen2.5-Math-1.5B-Instruct, gemma-3-1B-it, Llama-3.2-1B-Instruct} and task in

{mult, sudoku}, we trained two versions of the model which differed only in whether the data contained error-injection CoTs. To solidify the terminology, we define the model variants as follows.

**Definition 6.3.1** (FT model). We refer to the model trained only on clean CoTs as the *FT model* (Fine Tuned).

**Definition 6.3.2** (EIFT model). We refer to the model trained on error injected CoTs as the *EIFT model* (Error Injection Fine Tuned).

The details of the EIFT training methodology and data mix can be found in Section 6.3; see Section 6.6 for training hyperparameters.

## Error injection

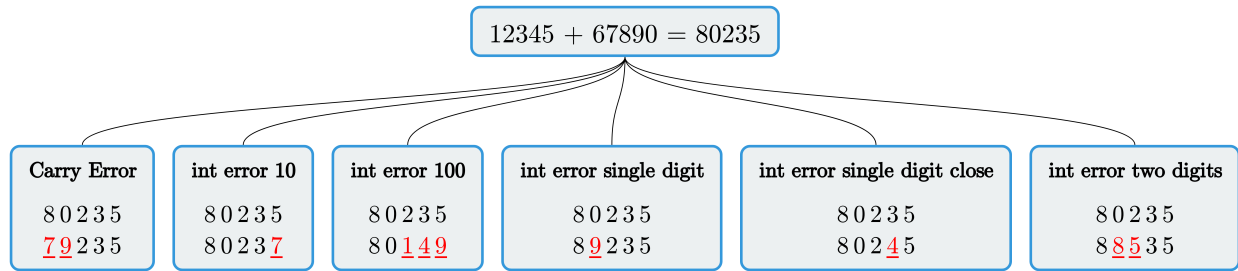


Figure 6.2: Types of errors we inject for the multiplication task. We take a correct step, depicted here as the addition step at the top of the figure. We then apply one of six types of errors with various probabilities; see Table 6.1 for more details. The underlined digits in each box depict the result of applying the corresponding error type.

For the EIFT model, we used a data mix where 20% of the CoTs were completely correct, and 80% had 1 to 4 synthetic errors injected, where the number and locations of errors were sampled uniformly at random (and without replacement for locations). We found that using more complex sampling schemes for the locations of errors did not improve the error correction abilities. Since the notion of a synthetic error is task-specific, each task is required to implement the distribution of synthetic errors that it attempts to model for the EIFT training. To determine this distribution, we trained FT models for various base models and examined the types of errors they tended to make; see the sections below for details. We validate this overall methodology in Section 6.3.

To inject a single error, we replace a golden step  $s_i$  at position  $i$  in the golden CoT with three steps according to the following procedure:

1. The erroneous step  $s'_i$ , which can easily be verified to be incorrect since we have access to the golden step  $s_i$ .
2. The error recognition step, consisting of the tokens AH! I MADE A MISTAKE.
3. Finally, the error correction step, which is just the original golden step  $s_i$ .

For EIFT training, we trained the model to predict the error recognition and error correction steps, and used loss masking to prevent the model from directly predicting the incorrect step  $s'_i$ . However, we did not modify the attention mask, as we wanted the model to learn an internal representation of what errors look like. We also ablated the token-level weight on the recognition and correction steps (Figure 6.8), but did not find any significant improvement in recognition or correction rates. Thus, for simplicity we did not use any special loss weighting beyond the loss masking.

## Multiplication

Our multiplication task consisted of uniformly random 4 digit multiplication problems. We found that 4 digit problems were the sweet spot for task difficulty: at this complexity, the base models had low accuracy, and the FT model was significantly more accurate than the base model. Figure 6.1 illustrates what the different types of CoTs look like for multiplication; see Section 6.7 for more examples.

**Error model.** To model the errors for multiplication, we trained the FT model for Qwen2.5 and a held out gemma2-2b-it. We then manually inspected the types of errors these models tended to make. Based on this study, we came up with the following broad error categories:

- `carry_error`: in an addition step, consider all locations where the column (including carry digits) sums to 9 or 10. Pick a random location and swap the column sum to the other value. Swapping 9 to 10 means incorrectly adding a carry, and swapping 10 to 9 means incorrectly omitting a carry.
- `_int_error_10`: add a random nonzero integer in  $[-10, 10]$  to the answer.
- `_int_error_100`: add a random nonzero integer in  $[-100, 100]$  to the answer.
- `_int_error_single_digit`: replace a randomly chosen single digit of the answer with a different random digit from 0 to 9.
- `_int_error_single_digit_close`: replace a randomly chosen single digit of the answer with a different random digit within 2 of it (respecting edge cases to keep the number syntactically valid after error injection).

- `_int_error_two_digits`: replace a random substring of two contiguous digits with two randomly chosen digits, resampling if the substring is unchanged.

For the specific frequencies and a visual depiction of these error types, see Table 6.1 and Fig. 6.2.

Table 6.1: Error probabilities for different step types. For addition steps, with probability 0.5 we inject an addition-carry error; otherwise we sample from the standard integer error injector (the `_int_error_*` distribution). For non-addition steps we always use the standard injector.

| Error type                                 | $\Pr[\text{type} \mid \text{addition}]$ | $\Pr[\text{type} \mid \text{non-addition}]$ |
|--------------------------------------------|-----------------------------------------|---------------------------------------------|
| <code>carry_error</code>                   | 0.5                                     | 0                                           |
| <code>_int_error_10</code>                 | 0.025                                   | 0.05                                        |
| <code>_int_error_100</code>                | 0.025                                   | 0.05                                        |
| <code>_int_error_single_digit</code>       | 0.125                                   | 0.25                                        |
| <code>_int_error_single_digit_close</code> | 0.125                                   | 0.25                                        |
| <code>_int_error_two_digits</code>         | 0.20                                    | 0.40                                        |

To pick the frequencies of the errors, we hill climbed on the final accuracy for the held out model, `gemma2-2b-it`.

## Sudoku

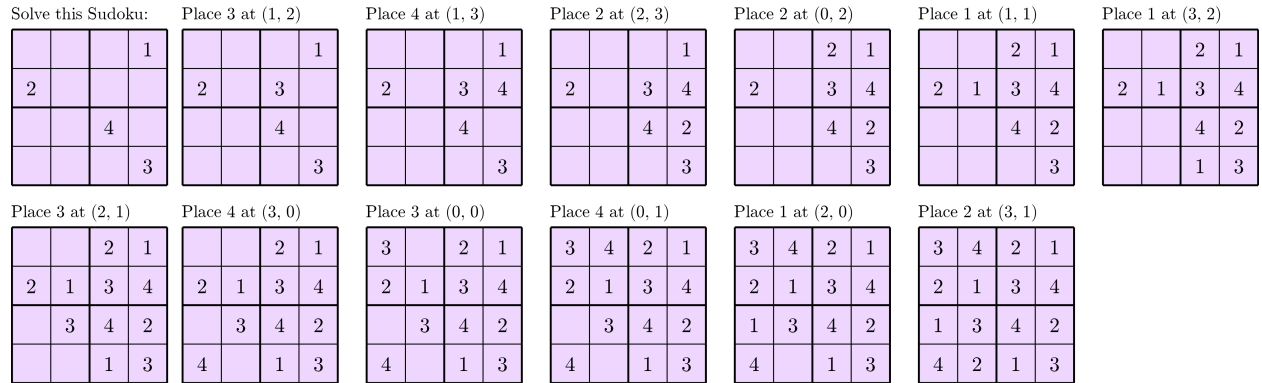


Figure 6.3: An example of a golden Chain-of-Thought (CoT) for the  $4 \times 4$  Sudoku task. The solution trajectory is restricted to “naked single” moves, defined as instances where a specific cell has exactly one possible candidate number remaining. The problem instances are generated by solving a random board and removing numbers such that a naked single solution path remains viable.

The Sudoku task consisted of  $4 \times 4$  Sudoku problems. To make the problem more tractable, we restricted to golden solutions that only make “naked single” moves.

**Definition 6.3.3** (Naked single move). For a given board state, a *naked single* is a cell that has exactly one possible candidate number remaining.

Correspondingly, we needed to sample Sudoku problems with naked singles solutions. To do so, we first generated problems by using a Sudoku solver with randomness to generate a random valid solution. We then randomly removed numbers one by one and rejected candidates unless a naked-single solver could complete the new board. See Figure 6.3 for an example of golden CoTs for sudoku.

**Error model.** To model errors, we used an equal split of invalid moves (i.e. moves with invalid coordinates or moves which violate a constraint) and moves that are not a naked single. We found that only injecting invalid moves did not impart error correction capabilities.

## Evaluation

For evaluation, we sampled 1000 problems for each task and evaluated the FT and EIFT models using greedy sampling. We also ablated the temperature setting (across  $T = 0.7, 1.0$ ), but did not find any significant difference in the accuracy or error correction capabilities; see Section 6.4 and Fig. 6.7 for more discussion.

To measure the accuracy of our models, we only graded their final answers. Even at modest dataset sizes, the models are very consistent with the exact answer format that we finetuned them on. Hence, we did not observe any false positives or negatives in our grading.

To measure the error correction capabilities, we follow a more involved procedure. Of course, one should not expect the model to correct from arbitrary error distributions. To narrow the scope, we compare the error correction capabilities where the errors are sampled from the following two error distributions.

**Definition 6.3.4** (Synthetic errors). A *synthetic* error is an erroneous step in a CoT generated by our synthetic error injector.

**Definition 6.3.5** (FT errors). An *FT* error is an erroneous step in a CoT generated by the FT model.

In other words, we generate synthetic errors by sampling synthetic CoTs and finding errors in them, and we generate FT errors by sampling FT CoTs and finding errors in them. See Figures 6.9 and 6.10 for synthetic and FT errors, respectively, for Qwen2.5 on multiplication.

With this in mind, we use the following process to measure error correction for the model.

1. Depending on which error source we are testing, we generate the appropriate type of CoT (synthetic or FT).
2. We truncate this CoT at the location of the first error, *including the error itself*. If the CoT does not have any errors, we discard this sample and repeat.
3. We prompt the model to complete the truncated CoT.
4. We grade the error recognition and correction for the guaranteed first error step produced by truncation. Namely, if first error step is at step  $i$ , we check for recognition on step  $i + 1$  and correction on step  $i + 2$ . Concretely, this entails checking for the tokens `AH! I MADE A MISTAKE` at step  $i + 1$  and checking that the step  $i + 2$  is correct *as defined by our task*.

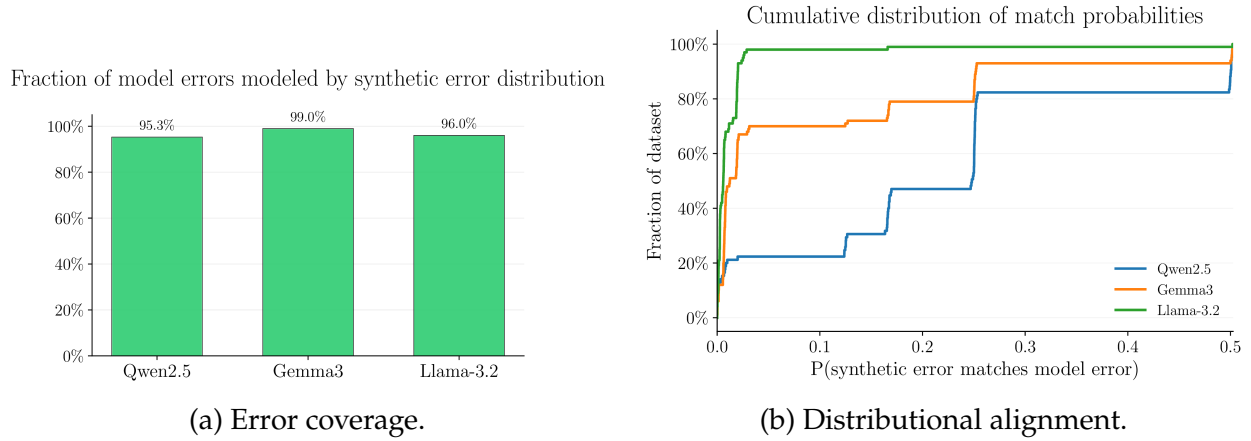
A couple points are in order about the above procedure. First, an important nuance to clarify is the exact notion of an incorrect step. To be conservative, we only check for errors that we have modeled. For multiplication, this means we only check steps where an atomic add or multiply is being performed incorrectly; for Sudoku, this means we check move placement steps and whether they are valid naked singles moves. In particular, we do not grade whether the model frivolously error corrects, as long as the proposed correction step is still correct.

Also, by convention, error corrections are a subset of error recognition—in order to correct an error, it must first declare that it has made a mistake. In practice, we never observed the model doing an error correction without first outputting an error recognition step.

## Error Distributions

While we qualitatively designed the injector to mimic observed error types, it is crucial to validate if the synthetic distribution accurately covers the on-policy error distribution. Hence, we sought to quantify the alignment between the synthetic errors used during training and the natural, on-policy errors made by the models.





**Figure 6.4: Validation of synthetic error distribution alignment with on-policy errors.** **(a) Error Coverage:** The percentage of on-policy model errors that appear at least once within  $n = 10,000$  samples from our synthetic error injector. We achieve near-perfect coverage ( $> 95\%$ ) across all models, indicating the injector correctly identifies the support of the error distribution. **(b) Distributional Alignment:** The cumulative distribution of the probability assigned by the synthetic injector to the *exact* on-policy error step (down and to the right is better). For a given probability threshold  $p$  on the  $x$ -axis, the  $y$ -axis represents what fraction of the 100 errors could be exactly matched with probability at most  $p$ . The probability mass placed on exact matches (particularly for Qwen2.5 and gemma-3) indicates that the synthetic distribution closely approximates the frequency of natural model errors.

We measured this alignment by estimating the probability of *exact* token-level match, i.e. how often the synthetic error injector exactly replicates an FT error. First, we collected 100 reasoning traces from the FT model containing calculation errors on the multiplication task. For each erroneous instance, we truncated the trace immediately prior to the first incorrect step and sampled  $n = 10,000$  candidate erroneous steps from our synthetic error injector. In comparison, during finetuning we trained the model on approximately 20,000 incorrect steps.

Figure 6.4a illustrates the error coverage: the percentage of on-policy errors that appeared at least once within the  $n$  synthetic samples. We observe near-perfect coverage across all models: 95.3% for Qwen2.5, 99.0% for gemma-3, and 96.0% for Llama-3.2. This indicates that the support of our synthetic distribution encompasses nearly all on-policy error modes.

Further, Figure 6.4b demonstrates that our synthetic injector achieves distributional alignment. We plot the empirical cumulative distribution function (CDF) of the probability assigned to the exact on-policy error step among the 100 reasoning traces. In generative tasks, the space of potential incorrect tokens is vast, making any significant probability

mass on an exact string match a strong indicator of alignment. We find that our injector frequently assigns high probabilities to the model’s specific natural errors. For instance, it assigns greater than 15% probability to the exact on-policy errors made by Qwen2.5 and gemma-3 in a large fraction of CoTs (70% for Qwen2.5 and 30% for gemma-3). Hence, during training, the models frequently encounter the *exact* errors they are prone to making, confirming the quality of the synthetic error distribution.

In Tables 6.2 and 6.3, we analyze the input FT error correction capabilities for Qwen2.5, bucketed by the synthetic error categories. Perhaps unsurprisingly, we do not exactly match the frequencies of error categories. More notably, the correction rates conditioned to specific error buckets were significantly lower for FT errors. Indeed, manual inspection of the FT errors revealed that they tended to be more subtle to catch and fix than synthetic errors: the relative error of FT errors was on average an order of magnitude smaller.

## 6.4 Results

For evaluation, we measured accuracy and error correction/recognition rates using a fixed set of 1000 randomly sampled problems. In all the figures below, we also report  $1.96 \cdot \text{SE}$  error bars.

We hypothesized that the EIFT model would have better performance than the corresponding FT model. In Figure 6.5 we see that this hypothesis is not strongly supported by the empirical results. For multiplication, we see modest to no performance gains from EIFT training. For Sudoku, we see more nontrivial gains of roughly 10% for Qwen2.5 and Llama-3.2. However, gemma-3 shows no performance boost from EIFT training. The baseline performance of the FT model varied drastically depending on the base model and task. For example, Qwen2.5 performed quite well (FT performance: 92%) on multiplication, but was not as strong on Sudoku. Conversely, Llama-3.2 had very poor performance on multiplication (FT performance: 2.0%), likely due to how Llama-3.2 tokenizes numbers, but had the best performance on Sudoku (FT performance: 52.2%). See Figure 6.11 for some examples of failure modes of the EIFT model. Apart from the issue of tokenization, we hypothesize that these fluctuations are due to the specific mid/post-training recipes for these models, e.g., since Qwen2.5 is a specialized math model.

One plausible explanation for why the performance is not significantly boosted from EIFT training is that the model does not actually reliably correct its own errors. Thus, we turn to investigating the correction and recognition capabilities of the EIFT model.

In Figure 6.6 we see the EIFT model, despite having high error recognition rates for the synthetic errors (Definition 6.3.4), suffers from a precipitous drop in correction and recognition rates when prompted with on-policy errors generated by the FT model (Defini-

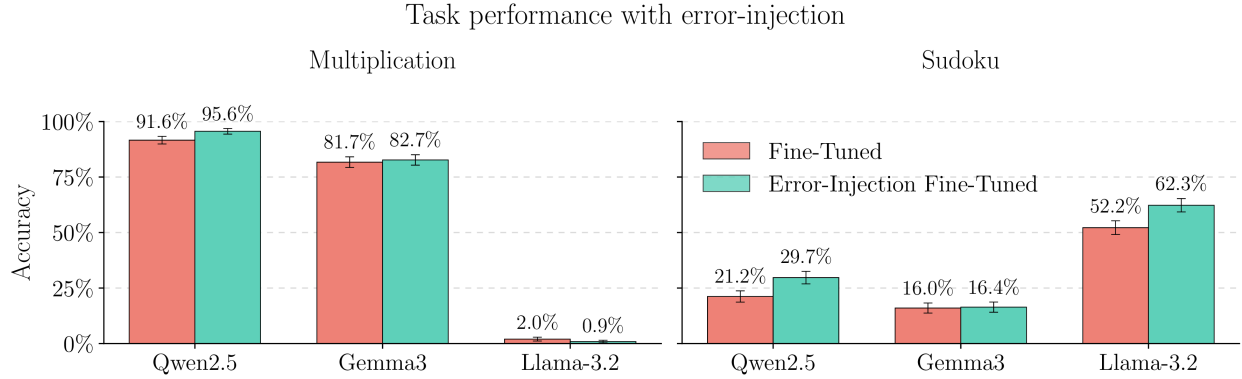


Figure 6.5: A comparison of model performance grouped by base model and task. We report the mean accuracy measured on a fixed set of 1000 randomly generated problems, and error bars are 1.96SE. The two labeled model types are (1) FT (Fine-Tuned) models which are trained solely on golden CoTs or (2) EIFT (Error-Injection Fine-Tuned) on error-injected CoTs. One would expect the EIFT models to have higher accuracy than the FT models. However, across the board, we see only modest ( $< 5\%$ ) performance gains for multiplication. For Sudoku tasks the performance is boosted by roughly 10% for Qwen and Llama, but none at all for Gemma.

tion 6.3.5). In all cases except for Qwen2.5 on multiplication, the FT error recognition rate drastically plummets—e.g., 94%  $\rightarrow$  8% for Qwen2.5 on Sudoku, 83%  $\rightarrow$  20% for gemma-3 on multiplication. Furthermore, there is a nontrivial gap in correction and recognition rates for FT errors. For Sudoku, this gap also exists for synthetic errors, suggesting that error correction is somewhat nontrivial for this task. For Qwen on multiplication, even though the model has better retention in error recognition (100%  $\rightarrow$  78%), it nevertheless significantly drops in error correction (99%  $\rightarrow$  40%). Interestingly, in this setting the EIFT model simply repeats the original error in 25% of the cases where it fails to correct from an FT error; see Figure 6.11 for some examples.

**Remark 6.4.1.** To be thorough, we also evaluated the FT models, and as expected the FT models did not possess error correction capabilities.

As mentioned in Section 6.3, we report our model evaluations using greedy sampling. We investigate the effect of positive temperature sampling on error correction and recognition rates; one might expect that error correction (and to a lesser extent error recognition) rates would be boosted by positive temperature sampling. In Figure 6.7 we test this hypothesis for the multiplication task using  $T \in \{0.0, 0.7, 1.0\}$ , and observe that in general the correction rates do not significantly change with positive temperature. In addition, we

observe no significant change in the recognition rate for Qwen2.5 and gemma-3. Interestingly, Llama-3.2 benefits the most from positive temperature sampling for error recognition, but remains unable to perform error correction. We conclude that our findings are relatively robust to the exact sampling temperature.

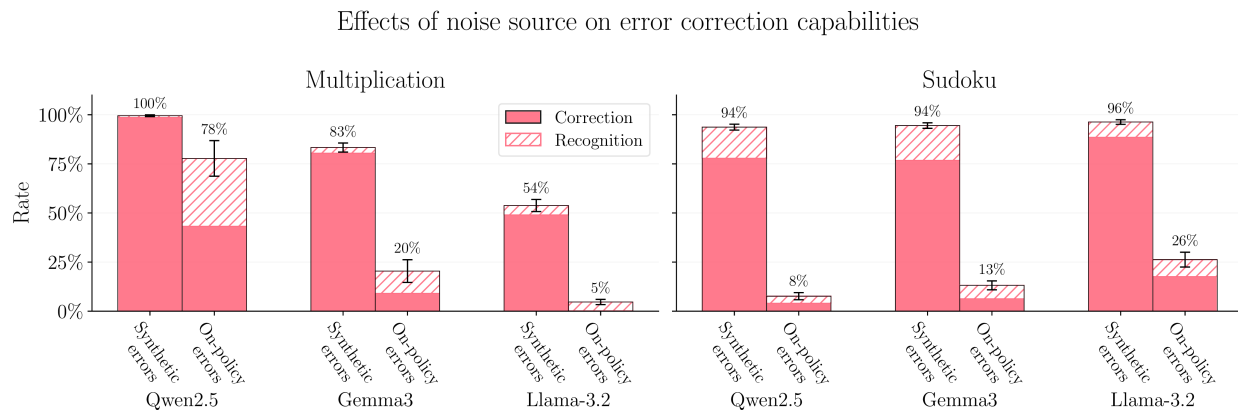


Figure 6.6: Comparison of error correction capabilities for EIFT models based on the error source and task. We report the mean correction and recognition rates measured on a fixed set of 1000 randomly generated problems, and error bars are 1.96SE. We compare errors that the EIFT model was trained to correct and recognize (Synthetic errors) versus errors the FT model naturally produces (On-policy errors). The recognition rates (hatched) are always larger than the correction rates (solid) as recognition is a superset of correction. The error correction capabilities would ideally generalize beyond the exact error distribution that the EIFT model is trained on. Aside from Qwen multiplication, we see that both correction and recognition rates drop drastically across all settings and base models. In general, there is a significant gap between error correction and error recognition in the on-policy error settings. Furthermore, for Sudoku such a gap is present even for synthetic errors.

## 6.5 Discussion

We have shown that, even in toy synthetic settings, synthetic error-injected SFT does not recover the performance of online RL, in spite of having access to golden solutions and perfect verifiers. Our results suggest that SFT-based approaches to error correction must essentially match the on-policy distribution, or else inject significantly richer information to condition the model to try different strategies.

Temperature effects on error recognition & correction

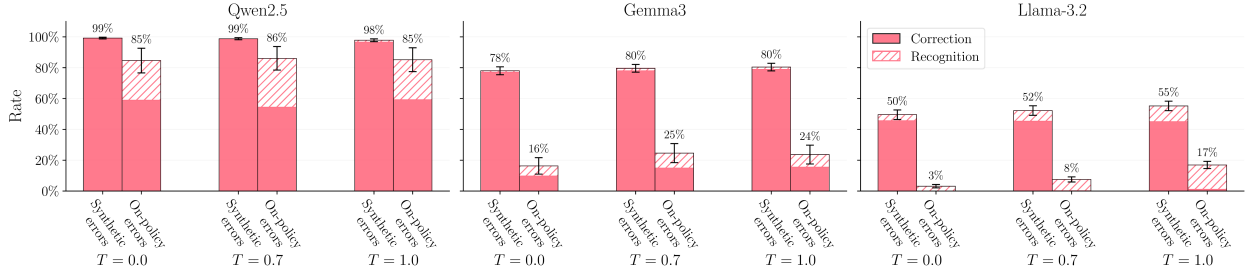


Figure 6.7: Temperature ablation for multiplication evaluation. We compare the correction and recognition rates for the exact same multiplication checkpoints, using different temperature settings:  $T \in \{0.0, 0.7, 1.0\}$ . We find no significant change in the capabilities across temperature settings for Qwen. For Gemma and Llama, we only see modest gains in performance with positive temperature, but the gap between synthetic and on-policy performance remains large ( $> 40\%$  absolute difference).

**Solver-Verifier gap.** When we conducted our initial experiments for multiplication, we were surprised that EIFT training did not significantly improve accuracy. One intuitive explanation was that for multiplication, the difficulty of a model verifying or checking an intermediate step is about as hard as doing it correctly in the first place.<sup>1</sup>

This led to designing the Sudoku task, which does have a more pronounced solver-verifier gap. For any particular move, verifying is trivial by just checking the constraints, whereas solving requires searching for the valid next step. Indeed, from Figure 6.5 we see that EIFT training boosts Sudoku performance more than multiplication performance for the Qwen2.5 and Llama-3.2 models. It would be interesting to study more systematically the benefits of error injection as a function of the solver-verifier gap. Some examples of more difficult solver-verifier gaps would be NP-hard search problems (such as general Sudoku where the solution need not be naked singles), but these would likely require training stronger base models than we could reasonably achieve given our compute limitations.

**Interpreting error correction.** We leave open the question of the precise mechanism behind the model’s error recognition and correction capabilities. For example, it was interesting that the model was able to recognize errors significantly more often than it could correct them. Most notably, we almost always observed that in CoTs where the model recognizes but doesn’t correct, it just parrots the original incorrect step; see Figure 6.11 for an

<sup>1</sup>Certain heuristics can approximately check intermediate work in arithmetic, such as rules based on modular arithmetic.

example. Moreover, this phenomenon was replicated across different temperature settings, suggesting that the model believes the most likely step is still the original (incorrect) step.

There are several different explanations for this type of behavior. One possibility is that the model learns two circuits, one for computing the answer to an intermediate step, and the other for verifying said answer. However, the model may not have learned how to use the verifier output to alter its circuit to get the correct (or even different) answer. Under this heuristic picture, it would be unsurprising that the model can rarely error correct. It would be interesting to investigate this hypothesis and to develop other training recipes that naturally “diversify” the strategies the model attempts towards solving the problem.

**Error injectors and RL.** One natural direction for further exploration is to generalize our results beyond synthetic error injectors. We now discuss some obstacles we faced when we tried to move in this direction.

The main issue one needs to overcome is the problem of efficiently generating EIFT datasets. Fundamentally, the usability of an EIFT method is bottlenecked by efficient verification, how accurate the error model is, and the availability of a golden solution. For example, the most natural error model is to take an on-policy approach and only use the model’s own errors. In this setting, the challenge then shifts to constructing an efficient and accurate enough verifier (e.g. a well-trained process reward model). Since our method is supposed to avoid doing RL in the first place, we did not pursue this direction further.

Another possibility is to prompt LLMs to scalably collect more on-policy errors. However, in early experiments, we found that it was somewhat difficult to steer the LLMs to produce errors in a reliable way that was useful for EIFT training. For instance, the model would often explicitly acknowledge its errors or make contrived errors.

These limitations could potentially be overcome by careful few-shot prompting. In practice, the process of constructing and iterating on few-shot prompts starts to look awfully similar to the construction of a synthetic error injector, especially if one wants to tune the frequency of certain types of errors appearing. On the other hand, it is possible that in more complicated reasoning scenarios, LLM based error injectors become the only scalable way to model errors.

## 6.6 Appendix: Training details

Each model was trained on a single GPU using bfloat16 mixed precision for 10000 steps with a batch size of 4 and a max sequence length of 1024 (the typical prompt length depended on the task). We used the AdamW optimizer with learning rate  $2e-5$ , weight decay  $1e-2$ , and default hyperparameters otherwise.

In Figure 6.8 we present our ablation over the token-level weight for the correction and backtrack steps.

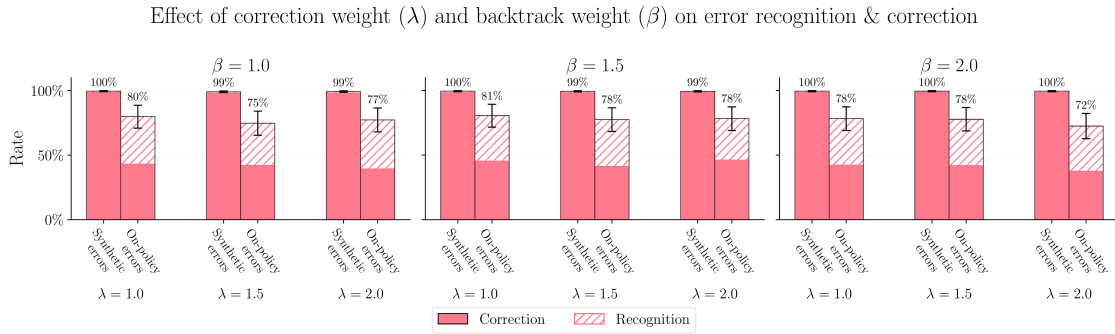


Figure 6.8: Sweep of correction weight  $\lambda$  and backtrack weight  $\beta$  across  $\{1.0, 1.5, 2.0\} \times \{1.0, 1.5, 2.0\}$  for the tokens in the error correction step for Qwen2.5 on multiplication. The error recognition and correction rates remain relatively consistent across the entire sweep.

## 6.7 Appendix: CoT examples

Table 6.2: Error recognition and correction breakdown for the Qwen2.5 synthetic EIFT model on 1000 synthetic errors. Note that there is some overlap in error categories; we greedily classified in the order `carry_error`  $\rightarrow$  `_int_error_single_digit_close`  $\rightarrow$  `_int_error_single_digit`  $\rightarrow$  `_int_error_two_digits`  $\rightarrow$  `_int_error_10`  $\rightarrow$  `_int_error_100`.

| Input Error Type                           | Frequency | Recognition Rate | Correction Rate | Correction given Recognition |
|--------------------------------------------|-----------|------------------|-----------------|------------------------------|
| <code>_int_error_10</code>                 | 0.002     | 1.000            | 1.000           | 1.000                        |
| <code>_int_error_100</code>                | 0.012     | 1.000            | 1.000           | 1.000                        |
| <code>_int_error_single_digit</code>       | 0.215     | 0.986            | 0.981           | 0.995                        |
| <code>_int_error_single_digit_close</code> | 0.287     | 0.997            | 0.990           | 0.993                        |
| <code>_int_error_two_digits</code>         | 0.348     | 1.000            | 0.986           | 0.986                        |
| <code>carry_error</code>                   | 0.136     | 0.993            | 0.985           | 0.993                        |

| Int Error Two Digits                                                                      | Int Error Single Digit                                                                    | Carry Error                                                                               |
|-------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|
| Q: What is 3057 * 7788? Think step by step.                                               | Q: What is 8498 * 4654? Think step by step.                                               | Q: What is 4670 * 9862? Think step by step.                                               |
| Multiply 3057 by 8 to get 24456.                                                          | Multiply 8498 by 4 to get 33993.                                                          | Multiply 4670 by 2 to get 9340.                                                           |
| Multiply 3057 by 8 to get 24456.                                                          | Ah! I made a mistake.                                                                     | Multiply 4670 by 6 to get 28020.                                                          |
| Add zeros to the right of 24456 to get 244560 because we are considering the 1th digit.   | Multiply 8498 by 4 to get 33992.                                                          | Add zeros to the right of 28020 to get 280200 because we are considering the 1th digit.   |
| Multiply 3057 by 7 to get 23499.                                                          | Multiply 8498 by 5 to get 42490.                                                          | Multiply 4670 by 8 to get 37360.                                                          |
| Ah! I made a mistake.                                                                     | Add zeros to the right of 42490 to get 424900 because we are considering the 1th digit.   | Add zeros to the right of 37360 to get 3736000 because we are considering the 2th digit.  |
| Multiply 3057 by 7 to get 21399.                                                          | Multiply 8498 by 6 to get 50988.                                                          | Multiply 4670 by 9 to get 42030.                                                          |
| Add zeros to the right of 21399 to get 2139900 because we are considering the 2th digit.  | Add zeros to the right of 50988 to get 5098800 because we are considering the 2th digit.  | Add zeros to the right of 42030 to get 42030000 because we are considering the 3th digit. |
| Multiply 3057 by 7 to get 21399.                                                          | Multiply 8498 by 4 to get 33992.                                                          | Add 9340 + 280200 to get 289540.                                                          |
| Add zeros to the right of 21399 to get 21399000 because we are considering the 3th digit. | Add zeros to the right of 33992 to get 33992000 because we are considering the 3th digit. | Add 289540 + 3736000 to get 3925540.                                                      |
| Add 24456 + 244560 to get 269016.                                                         | Add 33992 + 424900 to get 458892.                                                         | Ah! I made a mistake.                                                                     |
| Add 269016 + 2139900 to get 2408916.                                                      | Add 458892 + 5098800 to get 5557692.                                                      | Add 289540 + 3736000 to get 4025540.                                                      |
| Add 2408916 + 21399000 to get 23807916.                                                   | Add 5557692 + 33992000 to get 39549692.                                                   | Add 4025540 + 42030000 to get 46055540.                                                   |

Figure 6.9: Representative examples of synthetic errors for multiplication. The hatched steps are the incorrect steps, and the highlighted recognition/correction steps are separately labeled. See Section 6.3 for more details on what these types of errors entail.

| Carry Error                                                                               | Carry Error                                                                               | Int Error Two Digit                                                                       |
|-------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|
| Q: What is 2464 * 8669? Think step by step.                                               | Q: What is 2798 * 6894? Think step by step.                                               | Q: What is 6783 * 3006? Think step by step.                                               |
| Multiply 2464 by 9 to get 22176.                                                          | Multiply 2798 by 4 to get 11192.                                                          | Multiply 6783 by 6 to get 40708.                                                          |
| Multiply 2464 by 6 to get 14784.                                                          | Multiply 2798 by 9 to get 25182.                                                          | Multiply 6783 by 0 to get 0.                                                              |
| Add zeros to the right of 14784 to get 147840 because we are considering the 1th digit.   | Add zeros to the right of 25182 to get 251820 because we are considering the 1th digit.   | Add zeros to the right of 0 to get 0 because we are considering the 1th digit.            |
| Multiply 2464 by 6 to get 14784.                                                          | Multiply 2798 by 8 to get 22384.                                                          | Multiply 6783 by 0 to get 0.                                                              |
| Add zeros to the right of 14784 to get 1478400 because we are considering the 2th digit.  | Add zeros to the right of 22384 to get 2238400 because we are considering the 2th digit.  | Add zeros to the right of 0 to get 0 because we are considering the 2th digit.            |
| Multiply 2464 by 8 to get 19712.                                                          | Multiply 2798 by 6 to get 16788.                                                          | Multiply 6783 by 3 to get 20349.                                                          |
| Add zeros to the right of 19712 to get 19712000 because we are considering the 3th digit. | Add zeros to the right of 16788 to get 16788000 because we are considering the 3th digit. | Add zeros to the right of 20349 to get 20349000 because we are considering the 3th digit. |
| Add 22176 + 147840 to get 169016.                                                         | Add 11192 + 251820 to get 262012.                                                         | Add 40708 + 0 to get 40708.                                                               |
| Add 169016 + 1478400 to get 1647416.                                                      | Add 262012 + 2238400 to get 2490412.                                                      | Add 40708 + 0 to get 40708.                                                               |
| Add 1647416 + 19712000 to get 21359416.                                                   | Add 2490412 + 16788000 to get 19278412.                                                   | Add 40708 + 20349000 to get 20389708.                                                     |

Figure 6.10: Representative examples of FT errors for multiplication from Qwen2.5. The hatched steps are the incorrect steps. See Section 6.3 for more details on what these types of errors entail.



| Repeat Error                                                                              | Repeat Error                                                                              | Uncaught Error                                                                           |
|-------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|
| Q: What is 6223 * 9638? Think step by step.                                               | Q: What is 6783 * 3006? Think step by step.                                               | Q: What is 9567 * 1722? Think step by step.                                              |
| Multiply 6223 by 8 to get 49784.                                                          | Multiply 6783 by 6 to get 40708.                                                          | Multiply 9567 by 2 to get 19134.                                                         |
| Multiply 6223 by 3 to get 18669.                                                          | Ah! I made a mistake.                                                                     | Multiply 9567 by 2 to get 19134.                                                         |
| Add zeros to the right of 18669 to get 186690 because we are considering the 1th digit.   | Multiply 6783 by 6 to get 40708.                                                          | Add zeros to the right of 19134 to get 191340 because we are considering the 1th digit.  |
| Multiply 6223 by 6 to get 37338.                                                          | Multiply 6783 by 0 to get 0.                                                              | Multiply 9567 by 7 to get 67969.                                                         |
| Add zeros to the right of 37338 to get 3733800 because we are considering the 2th digit.  | Add zeros to the right of 0 to get 0 because we are considering the 1th digit.            | Add zeros to the right of 67969 to get 6796900 because we are considering the 2th digit. |
| Multiply 6223 by 9 to get 55997.                                                          | Multiply 6783 by 0 to get 0.                                                              | Multiply 9567 by 1 to get 9567.                                                          |
| Ah! I made a mistake.                                                                     | Add zeros to the right of 0 to get 0 because we are considering the 2th digit.            | Add zeros to the right of 9567 to get 9567000 because we are considering the 3th digit.  |
| Multiply 6223 by 9 to get 55997.                                                          | Multiply 6783 by 3 to get 20349.                                                          | Add 19134 + 191340 to get 210474.                                                        |
| Add zeros to the right of 55997 to get 55997000 because we are considering the 3th digit. | Add zeros to the right of 20349 to get 20349000 because we are considering the 3th digit. | Add 210474 + 6796900 to get 6997374.                                                     |
| Add 49784 + 186690 to get 236474.                                                         | Add 40708 + 0 to get 40708.                                                               | Ah! I made a mistake.                                                                    |
| Add 236474 + 3733800 to get 3970274.                                                      | Add 40708 + 0 to get 40708.                                                               | Add 210474 + 6796900 to get 7007374.                                                     |
| Add 3970274 + 55997000 to get 59967274.                                                   | Add 40708 + 20349000 to get 20389708.                                                     | Add 7007374 + 9567000 to get 16574374.                                                   |

Figure 6.11: Representative examples of EIFT self-correction failure modes from Qwen2.5. The hatched steps are incorrect, and the highlighted recognition steps are labeled. In the (Left) and (Center) panels we see instances where the EIFT model mimics its original mistake after recognizing it. In the (Right) panel we see a failure to recognize a mistake, along with a successful error correction near the end of the CoT.

Table 6.3: Error recognition and correction breakdown for the Qwen2.5 synthetic EIFT model on 1000 FT errors. The classification order is the same as in Table 6.2.

| Input Error Type              | Frequency | Recognition Rate | Correction Rate | Correction given Recognition |
|-------------------------------|-----------|------------------|-----------------|------------------------------|
| _int_error_10                 | 0.042     | 1.000            | 0.429           | 0.429                        |
| _int_error_100                | 0.054     | 1.000            | 0.574           | 0.574                        |
| _int_error_single_digit_close | 0.052     | 0.058            | 0.058           | 1.000                        |
| _int_error_two_digits         | 0.069     | 0.899            | 0.449           | 0.500                        |
| carry_error                   | 0.733     | 0.810            | 0.465           | 0.574                        |
| unknown                       | 0.050     | 0.780            | 0.420           | 0.538                        |

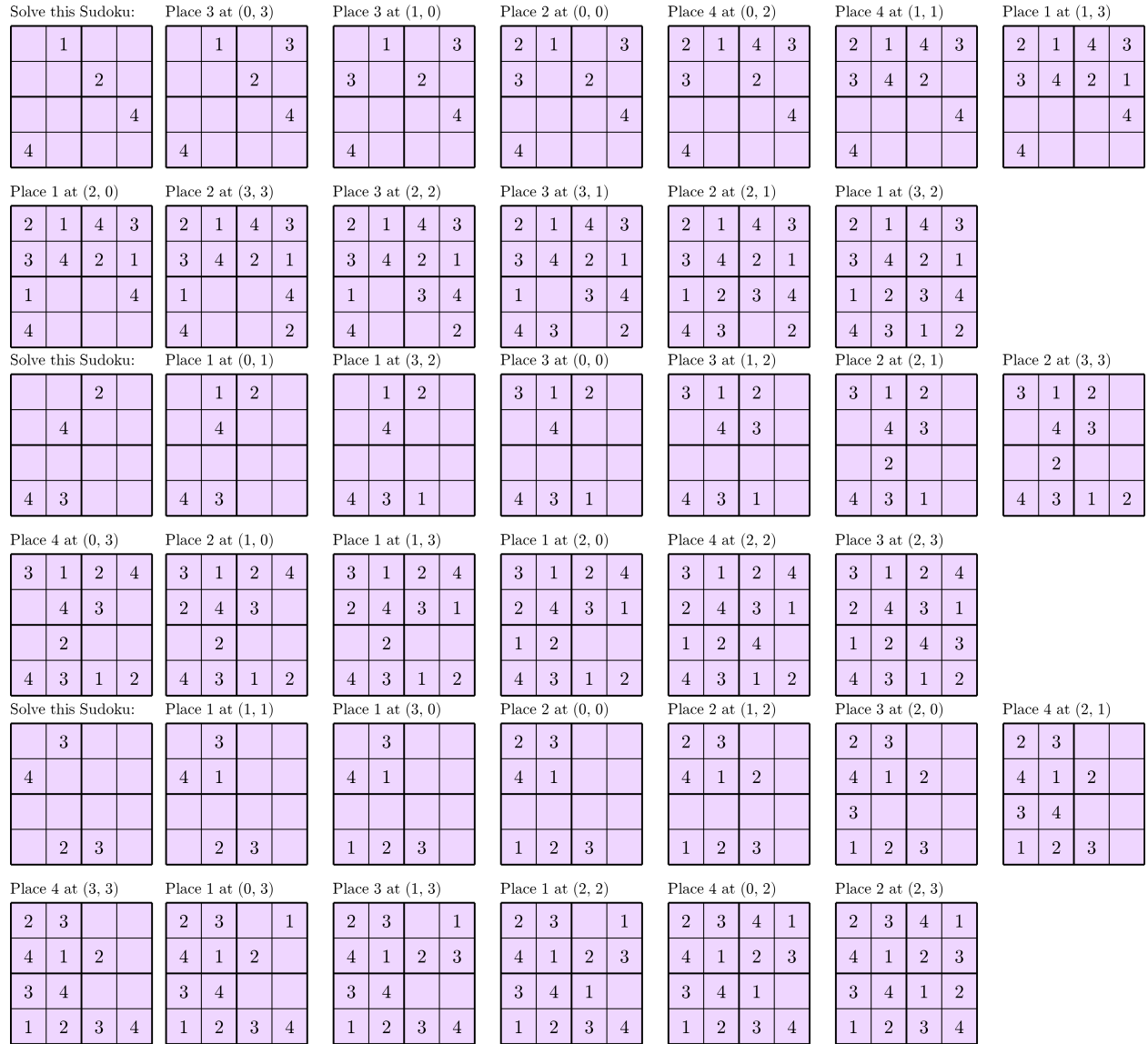


Figure 6.12: Representative examples of golden chain-of-thoughts (CoTs) from the sudoku task.

# Bibliography

- [A<sup>+</sup>20] N Aghanim et al. Planck 2018 results. vi. cosmological parameters. *Astron. Astrophys*, 641:A6, 2020.
- [AAA<sup>+</sup>16] Benjamin P Abbott, Richard Abbott, Thomas D Abbott, Matthew R Abernathy, Fausto Acernese, Kendall Ackley, Carl Adams, Thomas Adams, Paolo Addesso, Rana X Adhikari, et al. Observation of gravitational waves from a binary black hole merger. *Physical review letters*, 116(6):061102, 2016.
- [AB16] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.
- [Abb17] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.
- [ABXY24] Arka Adhikari, Christian Brennecke, Changji Xu, and Horng-Tzer Yau. Spectral gap estimates for mixed p-spin models at high temperature. *Probability Theory and Related Fields*, pages 1–29, 2024.
- [AC24] Norah Alballa and Marco Canini. Practical insights into knowledge distillation for pre-trained models, 2024.
- [ACO08] Dimitris Achlioptas and Amin Coja-Oghlan. Algorithmic barriers from phase transitions. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 793–802. IEEE, 2008.
- [AEGP23] Ahmed El Alaoui, Ronen Eldan, Reza Gheissari, and Arianna Piana. Fast relaxation of the random field Ising dynamics. *arXiv preprint arXiv:2311.06171*, 2023.
- [AG24] Ahmed El Alaoui and Jason Gaitonde. Bounds on the covariance matrix of the Sherrington–Kirkpatrick model. *Electronic Communications in Probability*, 29:1–13, 2024.

- [Aid98] Shigeki Aida. Uniform positivity improving property, Sobolev inequalities, and spectral gaps. *Journal of functional analysis*, 158(1):152–185, 1998.
- [AJK<sup>+</sup>21a] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence I: Modified log-Sobolev inequalities for fractionally log-concave distributions and high-temperature ising models. *arXiv preprint arXiv:2106.04105*, 2021.
- [AJK<sup>+</sup>21b] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence II: optimal sampling and concentration via restricted modified log-Sobolev inequalities. *arXiv preprint arXiv:2111.03247*, 2021.
- [AJK<sup>+</sup>22] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence: optimal mixing of down-up random walks. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1418–1430, 2022.
- [AJK<sup>+</sup>24] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Universality of spectral independence with applications to fast mixing in spin glasses. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5029–5056. SIAM, 2024.
- [AKS80] Miklós Ajtai, János Komlós, and Endre Szemerédi. A note on Ramsey numbers. *Journal of Combinatorial Theory, Series A*, 29(3):354–360, 1980.
- [AKUZ19] Fabrizio Antenucci, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Approximate Survey Propagation for Statistical Inference. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(2):023401, 2019.
- [AKV24] Nima Anari, Frederic Koehler, and Thuy-Duong Vuong. Trickle-Down in Localization Schemes and Applications. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1094–1105, 2024.
- [AL20] Vedat Levi Alev and Lap Chi Lau. Improved analysis of higher order random walks and applications. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1198–1211, 2020.
- [ALGV24] Nima Anari, Kuikui Liu, Shayan Oveis Gharan, and Cynthia Vinzant. Log-concave polynomials II: High-dimensional walks and an FPRAS for counting bases of a matroid. *Annals of Mathematics*, 199(1):259–299, 2024.

- [ALOG21] Nima Anari, Kuikui Liu, and Shayan Oveis Gharan. Spectral independence in high-dimensional expanders and applications to the hardcore model. *SIAM Journal on Computing*, (0):FOCS20–1, 2021.
- [AM22] Ahmed El Alaoui and Andrea Montanari. An information-theoretic view of stochastic localization. *IEEE Transactions on Information Theory*, 68(11):7423–7426, 2022.
- [AMS21] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Optimization of mean-field spin glasses. *The Annals of Probability*, 49(6):2922–2960, 2021.
- [AMS23] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from Mean-Field Gibbs Measures via Diffusion Processes. *arXiv preprint arXiv:2310.08912*, 2023.
- [AOA<sup>+</sup>20] Eric Arazo, Diego Ortego, Paul Albert, Noel E O’Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.
- [AOW15] Sarah R Allen, Ryan O’Donnell, and David Witmer. How to refute a random CSP. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 689–708. IEEE, 2015.
- [AS16] Noga Alon and Joel H Spencer. *The Probabilistic Method*. John Wiley & Sons, 2016.
- [BA23] Kenneth Borup and Lars Nørvang Andersen. Self-distillation for gaussian process regression and classification, 2023.
- [BAGJ20] Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Bounding flows for spherical spin glass dynamics. *Communications in Mathematical Physics*, 373:1011–1048, 2020.
- [BAP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643 – 1697, 2005.
- [BB19] Roland Bauerschmidt and Thierry Bodineau. A very simple proof of the LSI for high temperature spin systems. *Journal of Functional Analysis*, 276(8):2582–2588, 2019.

- [BBC<sup>+</sup>25] Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.
- [BCE<sup>+</sup>22] Krishna Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo. In *Conference on Learning Theory*, pages 2896–2923. PMLR, 2022.
- [BCKM98] Jean-Philippe Bouchaud, Leticia F Cugliandolo, Jorge Kurchan, and Marc Mézard. Out of equilibrium dynamics in spin-glasses and other glassy systems. *Spin glasses and random fields*, 12:161, 1998.
- [BCNM06] Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.
- [BCV25] Thiago Bergamaschi, Chi-Fang Chen, and Umesh Vazirani. A structural theory of quantum metastability: Markov properties and area laws. *arXiv preprint arXiv:2510.08538*, 2025.
- [BD00] Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The american mathematical monthly*, 107(4):353–357, 2000.
- [BDG06] Gérard Ben Arous, Amir Dembo, and Alice Guionnet. Cugliandolo-Kurchan equations for dynamics of spin-glasses. *Probability theory and related fields*, 136(4):619–660, 2006.
- [BDH16] Anton Bovier and Frank Den Hollander. *Metastability: a potential-theoretic approach*, volume 351. Springer, 2016.
- [BGJ20] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor PCA. *The Annals of Probability*, 48(4):2052–2087, 2020.
- [BGJM11] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press LLC, 2011.
- [BGL<sup>+</sup>14] Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.
- [BGP24] Gérard Ben Arous, Cédric Gerbelot, and Vanessa Piccolo. Langevin dynamics for high-dimensional optimization: the case of multi-spiked tensor PCA. *arXiv preprint arXiv:2408.06401*, 2024.

- [BGZ24] Antonio Blanca, Reza Gheissari, and Xusheng Zhang. Mean-field Potts and random-cluster dynamics from high-entropy initializations. *arXiv preprint arXiv:2404.13014*, 2024.
- [BHX20] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.
- [BIK<sup>+</sup>23] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.
- [BIL<sup>+</sup>16] Carlo Baldassi, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Local entropy as a measure for sampling solutions in constraint satisfaction problems. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(2):023301, 2016.
- [BJ24] Gérard Ben Arous and Aukosh Jagannath. Shattering versus metastability in spin glasses. *Communications on Pure and Applied Mathematics*, 77(1):139–176, 2024.
- [BKS23] Rares-Darius Buhai, Pravesh K Kothari, and David Steurer. Algorithms approaching the threshold for semi-random planted clique. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1918–1926, 2023.
- [BL16] Jinho Baik and Ji Oon Lee. Fluctuations of the free energy of the spherical Sherrington–Kirkpatrick model. *Journal of Statistical Physics*, 165:185–224, 2016.
- [BLLT20] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.
- [BM11] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.
- [BM19] Jean Barbier and Nicolas Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference. *Probability theory and related fields*, 174:1133–1185, 2019.

- [BNN24] Guy Bresler, Dheeraj Nagaraj, and Eshaan Nichani. Metastable mixing of Markov chains: Efficiently sampling low temperature exponential random graphs. *The Annals of Applied Probability*, 34(1A):517–554, 2024.
- [BRS11] Boaz Barak, Prasad Raghavendra, and David Steurer. Rounding semidefinite programming hierarchies via global correlation. In *2011 IEEE 52nd annual symposium on foundations of computer science*, pages 472–481. IEEE, 2011.
- [BSXY24] Christian Brennecke, Adrien Schertzer, Changji Xu, and Horng-Tzer Yau. The two point function of the SK model without external field at high temperature. *Probab. Math. Phys*, 5:131–175, 2024.
- [BT06] Sergey G Bobkov and Prasad Tetali. Modified logarithmic Sobolev inequalities in discrete settings. *Journal of Theoretical Probability*, 19(2):289–336, 2006.
- [BXY23] Christian Brennecke, Changji Xu, and Horng-Tzer Yau. Operator norm bounds on the correlation matrix of the SK model at high temperature. *arXiv preprint arXiv:2307.12535*, 2023.
- [BY05] ZD Bai and Jianfeng Yao. On the convergence of the spectral empirical process of Wigner matrices. *Bernoulli*, 11(6):1059–1092, 2005.
- [CBS<sup>+</sup>25] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- [CBTQO21] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6912–6920, 2021.
- [CCM21] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.
- [CDL<sup>+</sup>12] Paul Cuff, Jian Ding, Oren Louidor, Eyal Lubetzky, Yuval Peres, and Allan Sly. Glauber dynamics for the mean-field Potts model. *Journal of Statistical Physics*, 149:432–477, 2012.



- [CE22] Yuansi Chen and Ronen Eldan. Localization Schemes: A Framework for Proving Mixing Bounds for Markov Chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 110–122, Los Alamitos, CA, USA, Nov 2022. IEEE Computer Society.
- [Cel24] Michael Celentano. Sudakov–Fernique post-AMP, and a new proof of the local convexity of the TAP free energy. *The Annals of Probability*, 52(3):923–954, 2024.
- [CFM23] Michael Celentano, Zhou Fan, and Song Mei. Local convexity of the TAP free energy and AMP convergence for Z2-synchronization. *The Annals of Statistics*, 51(2):519–546, 2023.
- [CGG07] Patrick Cattiaux, Ivan Gentil, and Arnaud Guillin. Weak logarithmic Sobolev inequalities and entropic convergence. *Probability theory and related fields*, 139:563–603, 2007.
- [CGG<sup>+</sup>16] Jin-Yi Cai, Andreas Galanis, Leslie Ann Goldberg, Heng Guo, Mark Jerrum, Daniel Štefankovič, and Eric Vigoda. #BIS-hardness for 2-spin systems on bipartite bounded degree graphs in the tree non-uniqueness region. *Journal of Computer and System Sciences*, 82(5):690–711, 2016.
- [CGM12] Prasad Chebolu, Leslie Ann Goldberg, and Russell Martin. The complexity of approximately counting stable matchings. *Theoretical Computer Science*, 437:35–68, 2012.
- [Che21] Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture. *Geometric and Functional Analysis*, 31:34–61, 2021.
- [Che23] Sinho Chewi. Log-concave sampling. *Book draft available at <https://chewisinho.github.io>*, 2023.
- [CHKW23] Sergio Caracciolo, Alexander Hartmann, Scott Kirkpatrick, and Martin Weigel. Simulated annealing, optimization, searching for ground states. In *Spin Glass Theory and Far Beyond: Replica Symmetry Breaking After 40 Years*, pages 1–20. World Scientific, 2023.
- [CHS93] Andrea Crisanti, Heinz Horner, and H J Sommers. The spherical p-spin interaction spin-glass model: the dynamics. *Zeitschrift für Physik B Condensed Matter*, 92:257–271, 1993.

- [CK93] Leticia F Cugliandolo and Jorge Kurchan. Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Physical Review Letters*, 71(1):173, 1993.
- [CL21] Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021.
- [CLC<sup>+</sup>21] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.
- [CMV<sup>+</sup>23] Elisabetta Cornacchia, Francesca Mignacco, Rodrigo Veiga, Cédric Gerbelot, Bruno Loureiro, and Lenka Zdeborová. Learning curves for the multi-class teacher–student perceptron. *Machine Learning: Science and Technology*, 4(1):015019, 2023.
- [CMZ23] Zongchen Chen, Elchanan Mossel, and Ilias Zadik. Almost-Linear Planted Cliques Elude the Metropolis Process. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4504–4539. SIAM, 2023.
- [COKPZ17] Amin Coja-Oghlan, Florent Krzakala, Will Perkins, and Lenka Zdeborová. Information-theoretic thresholds from the cavity method. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 146–157, 2017.
- [COP19a] Amin Coja-Oghlan and Will Perkins. Bethe states of random factor graphs. *Communications in Mathematical Physics*, 366:173–201, 2019.
- [COP19b] Amin Coja-Oghlan and Will Perkins. Spin systems on Bethe lattices. *Communications in Mathematical Physics*, 372:441–523, 2019.
- [CPS24] Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in weak-to-strong generalization. *arXiv preprint arXiv:2405.15116*, 2024.
- [CT21] Wei-Kuo Chen and Si Tang. On convergence of the cavity and Bolthausen’s TAP iterations to the local magnetization. *Communications in Mathematical Physics*, 386(2):1209–1242, 2021.

- [CWZZ24] Xiang Cheng, Bohan Wang, Jingzhao Zhang, and Yusong Zhu. Fast Conditional Mixing of MCMC Algorithms for Non-log-concave Distributions. *Advances in Neural Information Processing Systems*, 36, 2024.
- [DAM16] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the binary stochastic block model. pages 185–189, 07 2016.
- [DAM17] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170, 2017.
- [dAT78] Jairo RL de Almeida and David J Thouless. Stability of the Sherrington-Kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General*, 11(5):983, 1978.
- [DF91] Martin Dyer and Alan Frieze. Computing the volume of convex bodies: a case where randomness provably helps. *Probabilistic combinatorics and its applications*, 44(123-170):0754–68052, 1991.
- [DFK91] Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991.
- [DGPZ25] Yatin Dandi, David Gamarnik, Francisco Pernice, and Lenka Zdeborová. Sequential dynamics in ising spin glasses. *arXiv preprint arXiv:2506.09877*, 2025.
- [Dia09] Persi Diaconis. The Markov Chain Monte Carlo Revolution. *Bull. Amer. Math. Soc.*, 46(2), 2009.
- [DJPR18] Ewan Davies, Matthew Jenssen, Will Perkins, and Barnaby Roberts. On the average size of independent sets in triangle-free graphs. *Proceedings of the American Mathematical Society*, 146(1):111–124, 2018.
- [DKMZ11] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, Dec 2011.
- [DM14] Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse PCA. In *2014 IEEE International Symposium on Information Theory*, pages 2197–2201. IEEE, 2014.

- [DMK<sup>+</sup>16] Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, Lenka Zdeborová, et al. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. *Advances in Neural Information Processing Systems*, 29, 2016.
- [DMM09] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.
- [DPHZ24] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.
- [DS23] Rudrajit Das and Sujay Sanghavi. Understanding self-distillation in the presence of label noise. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7102–7140. PMLR, 23–29 Jul 2023.
- [DSC96] Persi Diaconis and Laurent Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *The Annals of Applied Probability*, 6(3), 8 1996.
- [DSS15] Jian Ding, Allan Sly, and Nike Sun. Proof of the satisfiability conjecture for large k. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 59–68, 2015.
- [DW23] Partha S Dey and Qiang Wu. Mean field spin glass models under weak external field. *Communications in Mathematical Physics*, 402(2):1205–1258, 2023.
- [EAK18] Ahmed El Alaoui and Florent Krzakala. Estimation in the spiked wigner model: a short proof of the replica formula. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1874–1878. IEEE, 2018.
- [EAKJ20] Ahmed El Alaoui, Florent Krzakala, and Michael Jordan. Fundamental limits of detection in the spiked wigner model. 2020.
- [EAMS22] Ahmed El Alaoui, Andrea Montanari, and Mark 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.

- [EAMS23] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Shattering in pure spherical spin glasses. *arXiv preprint arXiv:2307.04659*, 2023.
- [EKZ22] Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. A spectral condition for spectral gap: fast mixing in high-temperature Ising models. *Probability theory and related fields*, 182(3):1035–1051, 2022.
- [Eld13] Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.
- [Eld20] Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3):737–755, 2020.
- [FFRT21] Giampaolo Folena, Silvio Franz, and Federico Ricci-Tersenghi. Gradient descent dynamics in the mixed p-spin spherical model: finite-size simulations and comparison with mean-field integration. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(3):033302, 2021.
- [FMM21] Zhou Fan, Song Mei, and Andrea Montanari. TAP free energy, spin glasses and variational inference. *The Annals of Probability*, 2021.
- [FV22] Alvaro Figueira and Bruno Vaz. Survey on synthetic data generation, evaluation methods and gans. *Mathematics*, 10(15):2733, 2022.
- [FW98] Mark I Freidlin and Alexander D Wentzell. Random perturbations. In *Random perturbations of dynamical systems*, pages 15–43. Springer, 1998.
- [FW15] Jianqing Fan and Weichen Wang. Asymptotics of empirical eigen-structure for ultra-high dimensional spiked covariance model. *arXiv preprint arXiv:1502.04733*, 2015.
- [GBP19] Junxian Geng, Anirban Bhattacharya, and Debdeep Pati. Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, 114(526):893–905, 2019.
- [GCW<sup>+</sup>24] Jianyuan Guo, Hanting Chen, Chengcheng Wang, Kai Han, Chang Xu, and Yunhe Wang. Vision superalignment: Weak-to-strong generalization for vision foundation models. *arXiv preprint arXiv:2402.03749*, 2024.
- [GHH<sup>+</sup>25] Mitchell Goff, Greg Hogan, George Hotz, Armand du Parc Locmaria, Kacper Raczy, Harald Schäfer, Adeeb Shihadeh, Weixing Zhang, and Yassine Yousfi.

- Learning to drive from a world model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1964–1973, 2025.
- [GHKM23] Venkatesan Guruswami, Jun-Ting Hsieh, Pravesh K Kothari, and Peter Manohar. Efficient Algorithms for Semirandom Planted CSPs at the Refutation Threshold. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 307–327. IEEE, 2023.
- [GJ12] Leslie Ann Goldberg and Mark Jerrum. Approximating the partition function of the ferromagnetic Potts model. *Journal of the ACM (JACM)*, 59(5):1–31, 2012.
- [GJ19] Reza Gheissari and Aukosh Jagannath. On the spectral gap of spherical spin glass dynamics. In *Annales de l’institut Henri Poincaré (B) Probability and Statistics*, volume 55, pages 756–776. Institute of Mathematical Statistics, 2019.
- [GJK23] David Gamarnik, Aukosh Jagannath, and Eren C Kızıldağ. Shattering in the Ising Pure  $p$ -Spin Model. *arXiv preprint arXiv:2307.07461*, 2023.
- [GJX23] Reza Gheissari, Aukosh Jagannath, and Yiming Xu. Finding planted cliques using Markov chain Monte Carlo. *arXiv preprint arXiv:2311.07540*, 2023.
- [GJZ17] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.
- [Gla63] Roy J Glauber. Time-dependent statistics of the ising model. *Journal of mathematical physics*, 4(2):294–307, 1963.
- [GLR18] Rong Ge, Holden Lee, and Andrej Risteski. Simulated tempering Langevin Monte Carlo ii: An improved proof using soft Markov chain decomposition. *arXiv preprint arXiv:1812.00793*, 2018.
- [GLZ<sup>+</sup>18] Ruiqi Gao, Yang Lu, Junpei Zhou, Song-Chun Zhu, and Ying Nian Wu. Learning Generative ConvNets via Multi-grid Modeling and Sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9155–9164, 2018.
- [GMT06] Sharad Goel, Ravi Montenegro, and Prasad Tetali. Mixing Time Bounds via the Spectral Profile. *Electronic Journal of Probability*, 11, 2006.
- [Goe04] Sharad Goel. Modified logarithmic Sobolev inequalities for some models of random walk. *Stochastic processes and their applications*, 114(1):51–79, 2004.

- [GS22] Reza Gheissari and Alistair Sinclair. Low-temperature ising dynamics with random initializations. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1445–1458, 2022.
- [GS24] Reza Gheissari and Alistair Sinclair. Spatial mixing and the random-cluster dynamics on lattices. *Random Structures & Algorithms*, 64(2):490–534, 2024.
- [GŠV15] Andreas Galanis, Daniel Štefankovič, and Eric Vigoda. Inapproximability for antiferromagnetic spin systems in the tree nonuniqueness region. *Journal of the ACM (JACM)*, 62(6):1–60, 2015.
- [GSVY16] Andreas Galanis, Daniel Stefankovic, Eric Vigoda, and Linji Yang. Ferromagnetic Potts model: Refined #BIS-hardness and related results. *SIAM Journal on Computing*, 45(6):2004–2065, 2016.
- [GTC24] Wei Guo, Molei Tao, and Yongxin Chen. Provable benefit of annealed langevin monte carlo for non-log-concave sampling. *arXiv preprint arXiv:2407.16936*, 2024.
- [Gue01] Francesco Guerra. Sum rules for the free energy in the mean field spin glass. *Mathematical Physics in Mathematics and Physics: Quantum and Operator Algebraic Aspects*, 30:161, 2001.
- [GV16] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck’s inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.
- [GW95] Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.
- [GYMT21] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.
- [GYZ<sup>+</sup>25] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [GZ00] A Guionnet and O Zeitouni. Concentration of the spectral measure for large matrices. *Electronic Communications in Probability*, 5:119–136, 2000.

- [Han07] Albert Hanen. A limit theorem for mean magnetisation in the sherrington-kirkpatrick model with an external field. In *Spin Glasses: Statics and Dynamics: Summer School, Paris 2007*, pages 177–201. Springer, 2007.
- [Har23] Hrayr Harutyunyan. On information captured by neural networks: connections with memorization and generalization, 2023.
- [HBCC24] Chi Hong, Robert Birke, Pin-Yu Chen, and Lydia Y. Chen. On dark knowledge for distilling generators. In De-Nian Yang, Xing Xie, Vincent S. Tseng, Jian Pei, Jen-Wei Huang, and Jerry Chun-Wei Lin, editors, *Advances in Knowledge Discovery and Data Mining*, pages 235–247, Singapore, 2024. Springer Nature Singapore.
- [HBCW24] Peter Hase, Mohit Bansal, Peter Clark, and Sarah Wiegrefe. The unreasonable effectiveness of easy training data for hard tasks. *arXiv preprint arXiv:2401.06751*, 2024.
- [HBK<sup>+</sup>21] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- [HBM<sup>+</sup>22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [HGJ<sup>+</sup>19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.
- [Hin10] Geoffrey Hinton. A Practical Guide to Training Restricted Boltzmann Machines. *Momentum*, 9(1):926, 2010.
- [HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [HK22] Jun-Ting Hsieh and Pravesh K Kothari. Approximating Max-Cut on Bounded Degree Graphs: Tighter Analysis of the FKL Algorithm. *arXiv preprint arXiv:2206.09204*, 2022.



- [HLW06] Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.
- [HLZ<sup>+</sup>24] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems*, 36, 2024.
- [HMP24] Brice Huang, Andrea Montanari, and Huy Tuan Pham. Sampling from spherical spin glasses in total variation via algorithmic stochastic localization. *arXiv preprint arXiv:2404.15651*, 2024.
- [HMRW25] Brice Huang, Sidhanth Mohanty, Amit Rajaraman, and David X. Wu. Weak Poincaré Inequalities, Simulated Annealing, and Sampling from Spherical Spin Glasses. *ACM Symposium on Theory of Computing*, 2025.
- [HPR19] Tyler Helmuth, Will Perkins, and Guus Regts. Algorithmic Pirogov-Sinai theory. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1009–1020, 2019.
- [HR04] David C Hoyle and Magnus Rattray. Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Physical Review E*, 69(2):026124, 2004.
- [HS17] Samuel B Hopkins and David Steurer. Efficient Bayesian estimation from few samples: community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–390. IEEE, 2017.
- [HS22] Brice Huang and Mark Sellke. Tight Lipschitz Hardness for optimizing Mean Field Spin Glasses. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 312–322. IEEE, 2022.
- [HS23] Brice Huang and Mark Sellke. A Constructive Proof of the Spherical Parisi Formula. *arXiv preprint arXiv:2311.15495*, 2023.
- [Hsu02] Elton P Hsu. *Stochastic analysis on manifolds*. Number 38. American Mathematical Soc., 2002.
- [HSW<sup>+</sup>21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

- [Hua24] Brice Huang. Capacity threshold for the Ising perceptron. *arXiv preprint arXiv:2404.18902*, 2024.
- [Hub59] J. Hubbard. Calculation of Partition Functions. *Phys. Rev. Lett.*, 3:77–78, Jul 1959.
- [HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [HW58] Wassily Hoeffding and J Wolfowitz. Distinguishability of sets of distributions. *The Annals of Mathematical Statistics*, 29(3):700–718, 1958.
- [HZR22] Hangfeng He, Hongming Zhang, and Dan Roth. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*, 2022.
- [JCL<sup>+</sup>24] Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, and Yaodong Yang. Aligner: Achieving efficient alignment through weak-to-strong correction. *arXiv preprint arXiv:2402.02416*, 2024.
- [JKL<sup>+</sup>24] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [JKP20] Matthew Jenssen, Peter Keevash, and Will Perkins. Algorithms for #BIS-hard problems on expander graphs. *SIAM Journal on Computing*, 49(4):681–710, 2020.
- [JNG<sup>+</sup>21] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.
- [JSV04] Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM (JACM)*, 51(4):671–697, 2004.
- [JT17] Aukosh Jagannath and Ian Tobasco. Some Properties of the Phase Diagram for Mixed  $p$ -Spin Glasses. *Probability Theory and Related Fields*, 167(3):615–672, 2017.
- [JT20] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

- [JT21] Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.
- [JZ20] Guangda Ji and Zhanxing Zhu. Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. *Advances in Neural Information Processing Systems*, 33:20823–20833, 2020.
- [KD25] Aayush Karan and Yilun Du. Reasoning with sampling: Your base model is smarter than you think. *arXiv preprint arXiv:2510.14901*, 2025.
- [KGJV83] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.
- [KKMO07] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O’Donnell. Optimal inapproximability results for MAX-CUT and other 2-variable CSPs? *SIAM Journal on Computing*, 37(1):319–357, 2007.
- [KL24] Bo’az Klartag and Joseph Lehec. Isoperimetric inequalities in high-dimensional convex sets. *arXiv preprint arXiv:2406.01324*, 2024.
- [KLR22] Frederic Koehler, Holden Lee, and Andrej Risteski. Sampling approximately low-rank Ising models: MCMC meets variational methods. In *Conference on Learning Theory*, pages 4945–4988. PMLR, 2022.
- [KLS97] Ravi Kannan, László Lovász, and Miklós Simonovits. Random walks and an  $O^*(n^5)$  volume algorithm for convex bodies. *Random Structures & Algorithms*, 11(1):1–50, 1997.
- [KLV24] Frederic Koehler, Holden Lee, and June Vuong. Efficiently learning and sampling multimodal distributions with data-based initialization. *Personal communication*, 2024.
- [KMH<sup>+</sup>20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [KO20] Tali Kaufman and Izhar Oppenheim. High order random walks: Beyond spectral gap. *Combinatorica*, 40(2):245–281, 2020.
- [Kra40] Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *physica*, 7(4):284–304, 1940.

- [KS66] Harry Kesten and Bernt P Stigum. Additional limit theorems for indecomposable multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, 37(6):1463–1481, 1966.
- [KS67] Harry Kesten and Bernt P Stigum. Limit theorems for decomposable multidimensional galton-watson processes. *Journal of Mathematical Analysis and Applications*, 17(2):309–338, 1967.
- [Kun23] Dmitriy Kunisky. Optimality of Glauber dynamics for general-purpose Ising model sampling and free energy approximation. *arXiv preprint arXiv:2307.12581*, 2023.
- [KV23] Frederic Koehler and Thuy-Duong Vuong. Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization. *arXiv preprint arXiv:2310.01762*, 2023.
- [KX25] Pravesh K Kothari and Jeff Xu. Smooth trade-off for tensor pca via sharp bounds for kikuchi matrices. *arXiv preprint arXiv:2510.03061*, 2025.
- [KZA<sup>+</sup>24] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.
- [L<sup>+</sup>13] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.
- [LA24] Yuejiang Liu and Alexandre Alahi. Co-supervised learning: Improving weak-to-strong generalization with hierarchical mixture of experts. *arXiv preprint arXiv:2402.15505*, 2024.
- [Lan08] Paul Langevin. Sur la théorie du mouvement brownien. *CR Acad. Sci. Paris*, 146(530-533):530, 1908.
- [LARC21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- [Liu23] Kuikui Liu. *Spectral Independence a New Tool to Analyze Markov Chains*. University of Washington, 2023.

- [LM17] Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. In *Conference on Learning Theory*, pages 1297–1301. PMLR, 2017.
- [LMP<sup>+</sup>24] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.
- [LMR<sup>+</sup>24] Kuikui Liu, Sidhanth Mohanty, Prasad Raghavendra, Amit Rajaraman, and David X. Wu. Locally stationary distributions: A framework for analyzing slow-mixing markov chains. *IEEE Annual Symposium on Foundations of Computer Science*, 2024.
- [LMRW24] Kuikui Liu, Sidhanth Mohanty, Amit Rajaraman, and David X. Wu. Fast mixing in sparse random ising models. *IEEE Annual Symposium on Foundations of Computer Science*, 2024.
- [LPW17] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2 edition, 2017.
- [LRG18] Holden Lee, Andrej Risteski, and Rong Ge. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. *Advances in neural information processing systems*, 31, 2018.
- [LS90] László Lovász and Miklós Simonovits. The mixing rate of Markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings [1990] 31st annual symposium on foundations of computer science*, pages 346–354. IEEE, 1990.
- [LS93] László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993.
- [LS16] Eyal Lubetzky and Allan Sly. Information percolation and cutoff for the stochastic Ising model. *Journal of the American Mathematical Society*, 29(3):729–774, 2016.
- [LS17] Eyal Lubetzky and Allan Sly. Universality of cutoff for the Ising model. 2017.

- [LSS22] Tengyuan Liang, Subhabrata Sen, and Pragma Sur. High-dimensional asymptotics of langevin dynamics in spiked matrix models. *arXiv preprint arXiv:2204.04476*, 2022.
- [LST21] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted Gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.
- [LSV24] Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. *arXiv preprint arXiv:2405.16043*, 2024.
- [LTM<sup>+</sup>22] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.
- [LV24] Yin Tat Lee and Santosh S Vempala. Eldan’s stochastic localization and the KLS conjecture: Isoperimetry, concentration and mixing. *Annals of Mathematics*, 199(3):1043–1092, 2024.
- [LW22] Gen Li and Yuting Wei. A non-asymptotic framework for approximate message passing in spiked models. *arXiv preprint arXiv:2208.03313*, 2022.
- [LWL<sup>+</sup>24] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*, 2024.
- [LY22] Miles E Lopes and Junwen Yao. A sharp lower-tail bound for gaussian maxima with application to bootstrap methods in high dimensions. *Electronic Journal of Statistics*, 16(1):58–83, 2022.
- [Mas14] Laurent Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703, 2014.
- [Mat06] P Mathieu. Quand l’inegalite log-Sobolev implique l’inegalite de trou spectral. In *Séminaire de Probabilités XXXII*, pages 30–35. Springer, 2006.
- [MFB20] Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361, 2020.

- [Mio17] Léo Miolane. Fundamental limits of low-rank matrix estimation: the non-symmetric case. *arXiv preprint arXiv:1702.00473*, 2017.
- [MM22] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.
- [MMZ06] Stephan Mertens, Marc Mézard, and Riccardo Zecchina. Threshold values of random K-SAT from the cavity method. *Random Structures & Algorithms*, 28(3):340–373, 2006.
- [MNS15] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.
- [MNS18] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.
- [MNS<sup>+</sup>21] Vidya Muthukumar, Adhyayan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in over-parameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.
- [Mon23] Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023.
- [Mon25] Andrea Montanari. Optimization of the sherrington–kirkpatrick hamiltonian. *SIAM Journal on Computing*, 54(4):FOCS19–1, 2025.
- [Moo17] Cristopher Moore. The computer science and physics of community detection: Landscapes, phase transitions, and hardness. *arXiv preprint arXiv:1702.00467*, 2017.
- [MPV87] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.
- [MPZ02] Marc Mézard, Giorgio Parisi, and Riccardo Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002.
- [MR14] Andrea Montanari and Emile Richard. A statistical model for tensor PCA. *Advances in neural information processing systems*, 27, 2014.

- [MR15] Andrea Montanari and Emile Richard. Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *IEEE Transactions on Information Theory*, 62(3):1458–1484, 2015.
- [MRT04] Andrea Montanari and Federico Ricci-Tersenghi. Cooling-schedule dependence of the dynamics of mean-field glasses. *Physical Review B—Condensed Matter and Materials Physics*, 70(13):134406, 2004.
- [MS12] Morten Mørup and Mikkel N Schmidt. Bayesian community detection. *Neural computation*, 24(9):2434–2456, 2012.
- [MT06] Ravi Montenegro and Prasad Tetali. Mathematical Aspects of Mixing Times in Markov Chains. *Foundations and Trends in Theoretical Computer Science*, 1(3):237–354, 2006.
- [MT23] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.
- [MTG<sup>+</sup>23] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.
- [MV21] Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1):321–345, 2021.
- [MVSS20] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.
- [MW23] Andrea Montanari and Yuchen Wu. Posterior sampling from the spiked models via diffusion processes. *arXiv preprint arXiv:2304.11449*, 2023.
- [MW24] Andrea Montanari and Yuchen Wu. Provably efficient posterior sampling for sparse linear regression via measure decomposition. *arXiv preprint arXiv:2406.19550*, 2024.
- [MWW07] Elchanan Mossel, Dror Weitz, and Nicholas C. Wormald. On the hardness of sampling independent sets beyond the tree threshold. *Probability Theory and Related Fields*, 143:401–439, 2007.



- [MYS<sup>+</sup>25] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hananeh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332, 2025.
- [Nes18] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.
- [NHH<sup>+</sup>20] Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the Anatomy of MCMC-Based Maximum Likelihood Learning of Energy-Based Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5272–5280, 2020.
- [Nik21] Sergey I Nikolenko. *Synthetic data for deep learning*, volume 174. Springer, 2021.
- [NLW23] Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.
- [NMB<sup>+</sup>23] Vaishnavh Nagarajan, Aditya K Menon, Srinadh Bhojanapalli, Hossein Mobahi, and Sanjiv Kumar. On student-teacher deviations in distillation: does it pay to disobey? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 5961–6000. Curran Associates, Inc., 2023.
- [OLSR<sup>+</sup>23] Utkarsh Ojha, Yuheng Li, Anirudh Sundara Rajan, Yingyu Liang, and Yong Jae Lee. What knowledge gets distilled in knowledge distillation? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 11037–11048. Curran Associates, Inc., 2023.
- [Ope25] OpenAI. OpenAI o3 and o4-mini system card. 2025.
- [Par80] Giorgio Parisi. A sequence of approximated solutions to the SK model for spin glasses. *Journal of Physics A: Mathematical and General*, 13(4):L115, 1980.
- [PL19] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International conference on machine learning*, pages 5142–5151. PMLR, 2019.

- [PSZ21] Damián Pinasco, Ezequiel Smucler, and Ignacio Zalduendo. Orthant probabilities and the attainment of maxima on a vertex of a simplex. *Linear Algebra and its Applications*, 610:785–803, 2021.
- [PWBM18] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of pca i: Spiked random matrix models. *The Annals of Statistics*, 46(5):2416–2451, 2018.
- [Rag08] Prasad Raghavendra. Optimal algorithms and inapproximability results for every CSP? In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 245–254, 2008.
- [RDRS21] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021.
- [RGB11] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.
- [RH03] Fredrik Ronquist and John P Huelsenbeck. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, 2003.
- [RSS<sup>+</sup>25] Dhruv Rohatgi, Abhishek Shetty, Donya Saless, Yuchen Li, Ankur Moitra, Andrej Risteski, and Dylan J Foster. Taming imperfect process verifiers: A sampling perspective on backtracking. *arXiv preprint arXiv:2510.03149*, 2025.
- [RV18] Cynthia Rush and Ramji Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286, 2018.
- [RW01] Michael Röckner and Feng-Yu Wang. Weak Poincaré inequalities and L2-convergence rates of Markov semigroups. *Journal of Functional Analysis*, 185(2):564–603, 2001.
- [RW25] Amit Rajaraman and David X. Wu. Markov Chains Approximate Message Passing. *arXiv preprint arXiv:2512.02384*. In submission, 2025.
- [RWC<sup>+</sup>19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

- [SAS22] Vignesh Subramanian, Rahul Arya, and Anant Sahai. Generalization for multiclass classification with overparameterized linear models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- [SBAH23] Felix Sarnthein, Gregor Bachmann, Sotiris Anagnostidis, and Thomas Hofmann. Random teachers are good teachers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30022–30041. PMLR, 23–29 Jul 2023.
- [SBG<sup>+</sup>21] Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Furong Huang, Uzi Vishkin, Micah Goldblum, and Tom Goldstein. Can you learn an algorithm? generalizing from easy to hard problems with recurrent networks. *Advances in Neural Information Processing Systems*, 34:6695–6706, 2021.
- [SCG<sup>+</sup>23] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.
- [SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [SDYD<sup>+</sup>23] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.
- [Sel24a] Mark Sellke. Optimizing mean field spin glasses with external field. *Electronic Journal of Probability*, 29:1–47, 2024.
- [Sel24b] Mark Sellke. The threshold energy of low temperature Langevin dynamics for pure spherical spin glasses. *Communications on Pure and Applied Mathematics*, 77(11):4065–4099, 2024.
- [She83] James B Shearer. A note on the independence number of triangle-free graphs. *Discrete Mathematics*, 46(1):83–87, 1983.

- [SIK<sup>+</sup>21] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? *Advances in Neural Information Processing Systems*, 34:6906–6919, 2021.
- [SK75] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.
- [SKC94] Bart Selman, Henry A Kautz, and Bram Cohen. Noise strategies for improving local search. In *AAAI*, volume 94, pages 337–343, 1994.
- [Sly10] Allan Sly. Computational Transition at the Uniqueness Threshold. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 287–296, 2010.
- [SPA23] Mher Safaryan, Alexandra Peste, and Dan Alistarh. Knowledge distillation performs partial variance reduction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 75229–75258. Curran Associates, Inc., 2023.
- [SPB<sup>+</sup>24] Seamus Somerstep, Felipe Maia Polo, Moulinath Banerjee, Ya’acov Ritov, Mikhail Yurochkin, and Yuekai Sun. A statistical framework for weak-to-strong generalization. *arXiv preprint arXiv:2405.16236*, 2024.
- [SS14] Allan Sly and Nike Sun. The Computational Hardness of Counting in Two-Spin Models on  $d$ -Regular Graphs. *The Annals of Probability*, 42(6):2383–2416, 2014.
- [SSDK<sup>+</sup>21] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021.
- [SSHB18] Harald Schafer, Eder Santana, Andrew Haden, and Riccardo Biasini. A commute in data: The comma2k19 dataset, 2018.
- [SSZM13] Dan Shen, Haipeng Shen, Hongtu Zhu, and JS Marron. Surprising asymptotic conical structure in critical sample eigen-directions. *arXiv preprint arXiv:1303.6171*, 2013.
- [STL20] Ruoqi Shen, Kevin Tian, and Yin Tat Lee. Composite Logconcave Sampling with a Restricted Gaussian Oracle. *arXiv preprint arXiv:2006.05976*, 2020.

- [Str57] RL Stratonovich. On a method of calculating quantum distribution functions. In *Soviet Physics Doklady*, volume 2, page 416, 1957.
- [Sub21] Eliran Subag. Following the ground states of full-RSB spherical spin glasses. *Communications on Pure and Applied Mathematics*, 74(5):1021–1044, 2021.
- [SV16] Igal Sason and Sergio Verdú.  $f$ -divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.
- [Tal06] Michel Talagrand. The Parisi formula. *Annals of mathematics*, pages 221–263, 2006.
- [Tal10] Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010.
- [Tal11] Michel Talagrand. *Mean Field Models for Spin Glasses: Volume II: Advanced replica-symmetry and low temperature*, volume 55. Springer Science & Business Media, 2011.
- [Thr24] Christos Thrampoulidis. Implicit bias of next-token prediction. *arXiv preprint arXiv:2402.18551*, 2024.
- [TL24] Leitian Tao and Yixuan Li. Your weak llm is secretly a strong teacher for alignment. *arXiv preprint arXiv:2409.08813*, 2024.
- [Ton02] Fabio Lucio Toninelli. About the Almeida-Thouless transition line in the Sherrington-Kirkpatrick mean-field spin glass model. *Europhysics letters*, 60(5):764, 2002.
- [Tre01] Luca Trevisan. Non-approximability results for optimization problems on bounded degree instances. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 453–461, 2001.
- [Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- [Wan06] Feng-Yu Wang. *Functional inequalities Markov semigroups and spectral theory*. Elsevier, 2006.
- [WBZ<sup>+</sup>21] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

- [WDY22] Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum  $\ell_1$ -norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics*, pages 10572–10602. PMLR, 2022.
- [WEM19] Alexander Wein, Ahmed El Alaoui, and Cristopher Moore. The Kikuchi hierarchy and tensor PCA. *Journal of the ACM*, 2019.
- [WHS22] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23549–23588. PMLR, 17–23 Jul 2022.
- [WKM<sup>+</sup>22] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.
- [WKS<sup>+</sup>25] David X. Wu, Shreyas Kapur, Anant Sahai, and Stuart Russell. Synthetic Error Injection Fails to Elicit Self-Correction In Language Models. *arXiv preprint arXiv:2512.02389*. In *submission.*, 2025.
- [WMT21] Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign Overfitting in Multiclass Classification: All Roads Lead to Interpolation. *arXiv e-prints*, page arXiv:2106.10865, June 2021.
- [WS24] David Wu and Anant Sahai. Precise asymptotic generalization for multiclass classification with overparameterized linear models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [WS25] David X. Wu and Anant Sahai. Provable weak-to-strong generalization via benign overfitting. *International Conference on Learning Representations (ICLR)*, 2025.
- [WT21] Ke Wang and Christos Thrampoulidis. Benign overfitting in binary classification of Gaussian mixtures. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4030–4034. IEEE, 2021.
- [WWS<sup>+</sup>22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning

- in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [WZX<sup>+</sup>23] Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, 2023.
- [XKZ<sup>+</sup>23] Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36:41618–41650, 2023.
- [XLZW16] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A Theory of Generative ConvNet. In *International conference on machine learning*, pages 2635–2644. PMLR, 2016.
- [XRH<sup>+</sup>24] Liuchi Xu, Jin Ren, Zhenhua Huang, Weishi Zheng, and Yunwen Chen. Improving knowledge distillation via head and tail categories. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5):3465–3480, 2024.
- [YDC22] Kaiyu Yang, Jia Deng, and Danqi Chen. Generating natural language proofs with verifier-guided search. *arXiv preprint arXiv:2205.12443*, 2022.
- [YSS<sup>+</sup>24] Wenkai Yang, Shiqi Shen, Guangyao Shen, Zhi Gong, and Yankai Lin. Super (ficial)-alignment: Strong models may deceive weak models in weak-to-strong generalization. *arXiv preprint arXiv:2406.11431*, 2024.
- [YSX<sup>+</sup>24] Hua Yuan, Yu Shi, Ning Xu, Xu Yang, Xin Geng, and Yong Rui. Learning from biased soft labels. *Advances in Neural Information Processing Systems*, 36, 2024.
- [YZW<sup>+</sup>25] Xiao-Wen Yang, Xuan-Yi Zhu, Wen-Da Wei, Ding-Chu Zhang, Jie-Jing Shao, Zhi Zhou, Lan-Zhe Guo, and Yu-Feng Li. Step back to leap forward: Self-backtracking for boosting reasoning of language models. *arXiv preprint arXiv:2502.04404*, 2025.
- [ZPC<sup>+</sup>23] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

- [ZWH<sup>+</sup>21] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021.
- [ZZ23] Qingyue Zhao and Banghua Zhu. Towards the fundamental limits of knowledge transfer over finite domains, 2023.
- [ZZS<sup>+</sup>24] Edwin Zhang, Vincent Zhu, Naomi Saphra, Anat Kleiman, Benjamin L Edelman, Milind Tambe, Sham M Kakade, and Eran Malach. Transcendence: Generative models can outperform the experts that train them. *arXiv preprint arXiv:2406.11741*, 2024.