UNIVERSITY OF CALIFORNIA

Los Angeles

# Unsupervised Methods on Structured Data

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Luciano Vinas

2025

ABSTRACT OF THE DISSERTATION

Unsupervised Methods on Structured Data

by

Luciano Vinas
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2025
Professor Arash A. Amini, Chair

Classical unsupervised algorithms, such as $k$-means and PCA, utilize a simple generative model where the sampling distribution is determined by a collection of unobserved, latent features. While this paradigm is powerful, it has the following consequence for applied settings: any structured trend in the data must be explained by the latent features and the assumptions therein. This requirement complicates the analysis of structured data sources, such as images, videos, and networks, especially when the latent features of interest do not govern every structured aspect of the data.

In this work, we consider scenarios where the latent features may be partially decoupled from the structure of the data. Under this new setting we develop new algorithmic improvements and insights for the following problems:

- Tissue intensity recovery for contaminated MRIs, where each pixel intensity is determined by an underlying tissue type and a spatially varying gain field.
- Semi-supervised node classification with graph aggregated features, where nodes are assumed to follow a community-based structure.

The dissertation of Luciano Vinas is approved.

Guido F. Montufar Cuartas

Mark S. Handcock

Qing Zhou

Oscar H. Madrid Padilla

Arash A. Amini, Committee Chair

University of California, Los Angeles

2025

*To my mom and dad,*
*who have supported me all throughout.*
*To Briana, who I cherish and care for*
*more than anything.*

TABLE OF CONTENTS

**1 Introduction** . . . . . **1**

1.1 Organization . . . . . 2

**I Level Recovery on Contaminated Images** **4**

**2 LapGM: A Rapid Multisequence MR Bias Correction and Normalization Model** . . . . . **5**

2.1 Introduction . . . . . 5

2.2 Methods . . . . . 7

2.2.1 Connection to Previous Works . . . . . 9

2.2.2 Gradient Weighting Heuristic . . . . . 11

2.3 Materials and Experiments . . . . . 12

2.4 Results and Discussion . . . . . 13

2.4.1 Simulated Data . . . . . 14

2.4.2 Patient Data . . . . . 17

2.4.3 Normalization Comparisons . . . . . 18

2.5 Conclusion . . . . . 20

A Expectation-Maximization for MAP . . . . . 21

B MAP Coordinate Updates . . . . . 23

**3 Step and Smooth Decompositions as Topological Clustering** . . . . . **25**

1 Introduction . . . . . 25

1.1 Applications . . . . . 27
1.2 Prior Work . . . . . 29
2 Identifiability Theory . . . . . 29
2.1 Topological Clustering . . . . . 30
2.2 Identifiability Results . . . . . 33
3 Methods and Optimization . . . . . 36
3.1 One-step Analysis . . . . . 37
4 Experiments . . . . . 43
4.1 Simulation Experiments . . . . . 45
4.2 MRI Decontamination . . . . . 46
5 Conclusion . . . . . 48
A Identifiability Proofs . . . . . 49
A.1 Proof of Proposition 1 . . . . . 52
B Supplement to Section 3.1 . . . . . 53
B.1 Proof of Proposition 2 . . . . . 53
B.2 Proof of Proposition 3 . . . . . 54
B.3 Sobolev Kernel Rates . . . . . 55
C Additional Experiments . . . . . 57
D AltMin Consistency Through Gamma-Convergence Techniques . . . . . 58
D.1 Convergence of Empirical Minimizers . . . . . 60
D.2 k-means Consistency . . . . . 62
D.3 Cluster Consistency for Contaminated Objectives . . . . . 63
D.4 Nearest Label Consistency from Cluster Convergence . . . . . 66

**II The Role of Neighbor Aggregation on Graphs 71**

**4 Simple GNNs with Low Rank Non-parametric Aggregators . . . . . 72**

1 Introduction . . . . . 72

2 GNN and SSNC Formalism . . . . . 74

2.1 Nonparametric Spectral Reshaping . . . . . 75

2.2 Motivating General Spectral Learners . . . . . 77

2.3 Complexity of Low Rank Spectral Learners . . . . . 78

3 Experiments . . . . . 78

3.1 SSNC Benchmarks . . . . . 80

3.2 CSBM Experiment . . . . . 81

3.3 Aggregation Ablation . . . . . 83

4 Changes in Evaluation Conventions . . . . . 85

4.1 Comparing Split Performances . . . . . 87

5 Conclusions . . . . . 88

**5 Sharp Bounds for Poly-GNNs and the Effect of Graph Noise . . . . . 90**

1 Introduction . . . . . 90

1.1 SSNC with GNNs . . . . . 91

1.2 CSBM and Noise Decompositions . . . . . 92

1.3 Paper Overview . . . . . 94

2 Main result . . . . . 96

2.1 Informal statement . . . . . 99

2.2 Formal statement . . . . . 100

2.3 Previous Work . . . 103
3 Signal Analysis . . . 105
3.1 Signal Proxy Growth . . . 106
3.2 Signal Proxy as Leading Order Approximation . . . 107
4 Noise Analysis . . . 109
4.1 Controlling Feature Noise . . . 110
4.2 Graph Noise and Walk Sequences . . . 114
4.3 Leading Order Walk Decomposition . . . 117
4.4 Proof of Theorem 11 . . . 119
4.5 Characterizing N* . . . 120
4.6 Proxy for Thi . . . 123
4.7 Controlling —Thi - Thit— . . . 125
4.8 Thit Lowerbound . . . 126
4.9 Proof of Theorem 12 . . . 128
5 Conclusion . . . 129
A Proofs For Signal Argument . . . 130
A.1 Proof of Lemma 5 . . . 132
B Concentration from Moments . . . 133
C Counting Lemmas . . . 135
C.1 Proof of Lemma 6 . . . 135
C.2 Proof of Lemma 7 . . . 135
C.3 Proof of Lemma 13 . . . 137
D Proofs for Noise Upperbound . . . 138

D.1 Proof of Lemma 10 . . . 138
D.2 Proof of Lemma 12 . . . 139
D.3 Proof of Lemma 14 . . . 139
E Proofs for the Noise Lower bound . . . 141
E.1 Proof of Lemma 15 . . . 141
E.2 Proof of Lemma 17 . . . 142
E.3 Proof of Lemma 19 . . . 143
E.4 Proof of Lemma 20 . . . 143
F Auxiliary Lemmas . . . 147

**6 A CLT for Polynomial GNNs on Community-Based Graphs . . . 149**

1 Introduction . . . 149
1.1 Overview of Our Contributions . . . 150
2 Preliminaries and Model Setup . . . 152
2.1 Poly-GNNs and Feature Definitions . . . 152
2.2 Community-Based Graph Model . . . 153
2.3 Assumptions . . . 154
2.4 Wasserstein Distance . . . 154
3 Asymptotic Distribution of Poly-GNN Embeddings . . . 155
3.1 Main Central Limit Theorems . . . 155
3.2 Proof Outline and Key Steps . . . 157
4 Implications for Classification and GNN Oversmoothing . . . 159
4.1 Convergence of Linear Classification on Poly-GNN Features . . . 159
4.2 A Precise Mechanism for GNN Oversmoothing . . . 162

5 Conclusion . . . . . 164

A Detailed Proofs of Main Theorems . . . . . 165

A.1 Proof of Theorem 15 (CLT for Centered and Scaled Features) . . . . 166

A.2 Proof of Theorem 14 (CLT for Degree-Normalized Features and Labels) 168

A.3 Supporting Lemmas for Moment Analysis . . . . . 171

A.4 Supporting Results for Specialization to Community-Based Graphs . 174

B Moment Characterization for Wp . . . . . 181

C Psir sub-Gaussians . . . . . 181

D Results on Triangular Arrays . . . . . 184

E Proof of Proposition 15 . . . . . 186

F Remaining proofs . . . . . 192

F.1 Proof of Lem . . . . . 192

F.2 Proof of Lemma 37 . . . . . 195

G Simulation Details for Figures . . . . . 196

G.1 Details for Figure 6.1 . . . . . 197

G.2 Details for Figure 6.2 . . . . . 198

G.3 Details for Figure 6.3 . . . . . 198

G.4 Details for Figure .4 . . . . . 199

G.5 Details for Figure .5 . . . . . 199

**References** . . . . . **200**

LIST OF FIGURES

2.1 Model diagram of LapGM and its use cases. Model assumes intensity distribution follows a Gaussian mixture with additional spatial information determined by a bias field. . . . . 8

2.2 Debias error results on Line32 with N4 and LapGM 3-seq. algorithm. Second column of this figure reveals LapGM's tendency to increase contrast for certain dominant tissue groups. . . . . 15

2.3 Probability density comparison between true tissue density and debiased tissue density functions. . . . . 16

2.4 Line profiles of the lower-left section of patient pelvic region. Line profile in red (top) with corresponding intensity shown in blue (bottom). Top dotted line is placed on dotted peak and the bottome dotted line is placed on the second peak next to the trough. . . . . 17

2.5 Debias and bias field comparison of single sequence LapGM for varying regularizations $\tau$ and gradient penalty weights $h_\alpha(r, z)$. . . . . 18

2.6 Visualization of the $\mu_*$ normalization technique with real patient data. After normalization, the second peak of each patient distribution clusters around the target intensity of 1000. . . . . 19

2.7 Combined tissue distributions of 10 patient scans using different normalization techniques. The distributions produced from each normalization technique is peak-aligned with the tissue distribution of a water mask ROI normalized distribution. Starting from the leftmost column the tissue TV errors relative to the water mask ROI are 9.52%, 14.0%, and 21.7%. . . . . 20

3.1 Example of a prominent bias field modifying the BrainWeb [CKK97] phantom. Leftmost image is the MRI of a synthetic brain which has been perturbed by a contaminant field. Middle image is a tissue categorization for the synthetic brain. Rightmost image is the contaminant field in ambient space (in absence of the synthetic brain). . . . . 28

3.2 Caption for LOF . . . . . 32

3.3 Sobolev-2 interpolants for step functions on $[0, 1]$. As sampled points $\{x_i\}_i$ approach a discontinuity of $f$, the corresponding interpolant $\bar{f}_n$ has Hilbert norm $\|\cdot\|_{\mathbb{H}}$ going to infinity. . . . . 40

3.4 Experiment results for the 2-state, $p$-probability Markov chain. 10000 chains were simulated for each $p \in \{k/10\}_{k=1}^{10}$. Shown in subfigures are median results with 95% probability intervals shaded in the corresponding colors. In the case of $p = 1$, there is no shading. . . . . 42

3.5 Signal, field and composite observation simulated from (3.27) for two and three classes. . . . . 44

3.6 AltMin recovery results for a noiseless simulated setting. Worst-case theory bounds are shown as dashed lines for each of the different settings. . . . . 45

3.7 AltMin recovery results for a noisy simulated setting. Bayes error rates for classification are shown as dashed lines for the various noise levels. . . . . 46

3.8 Cluster and level accuracy of the AltMin algorithm on the biased BrainWeb phantom. Final accuracies for single and multi-sequence settings are 91.07% and 98.91% respectively. Level deviations in the multi-sequence setting are calculated with respect to the vector 2-norm. . . . . 47

3.9 AltMin decomposition for the biased BrainWeb dataset. Class maps of the single sequence setting show anomalous tissue patches in areas where the field changes most rapidly. . . . . 48

3.10 exa use of the $\phi(r)$ function and $\{(u_q, v_q)\}_{q=1}^{Q}$ sequence, shown in blue above, for a 4-class path of length 7. The colors denote the estimated cluster labels. This example has $\{(u_q, v_q)\}_{q=1}^{Q} = \{(i_1, i_2), (i_4, i_5), (i_6, i_7)\}$ with $Q = 3$. . . . . 50

3.11 Residual norm decays $\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2$, based on the optimal $\tau_n$ selections, for different Sobolev-$\alpha$ kernels. Slight numerical inaccuracies are shown in the norm decay of the Sobolev-2 kernel. . . . . 56

4.1 Accuracy comparison of the KERNEL model for different graph representations $\boldsymbol{A}$ and $\boldsymbol{D} - \boldsymbol{A}$. Shown above is the signed accuracy difference between the adjacency and Laplacian representations. Best performing kernel was selected per dataset. 83

4.2 LR KERNEL performance relative to the full-rank KERNEL for different truncation factors $r$. Performance is seen to gradually decline on most datasets as the truncation factor $r$ decreases (that is truncation percentage increases). LR KERNEL performance can also be seen to periodically increase above full-rank KERNEL performance for the datasets Chameleon (red) and Squirrel (purple). . 85

4.3 Performance homogenization achieved by LR KERNEL model on directed networks. 86

4.4 Accuracy results and uncertainties on the citation datasets using different splits with linear models $\boldsymbol{XW}$ and $\boldsymbol{AXW}$. "Public" refers to the split introduced by [KW17]. Both "Sparse" and "Public" are single splits, so one cannot associate uncertainty to them. . . . . 86

4.5 Accuracy results on datasets introduced by [PWC20]. "Dense" refers to the original split while "Balanced" refers to the split introduced by [CPL21]. Test results and uncertainties are evaluated using models $\boldsymbol{XW}$ and $\boldsymbol{AXW}$. Results shown are for method with best validation. . . . . 87

5.1 Walks and their corresponding graphs for $n \geq 5$. . . . . 110

5.2 Walks and their unique undirected edges and nodes . . . . . 115

5.3 A walk sequence $\boldsymbol{w}$ with $r = 6$ components of length $k = 3$ each, and its corresponding undirected graph $G(\boldsymbol{w})$. This walk sequence belongs to $\mathcal{N}_{r,t,v}$ defined in (5.35) with $r = 6$, $t = |[\boldsymbol{w}]| = 15$ (number of unique edges) and $v = |[\![\boldsymbol{w}]\!]| = 16$ (number of unique vertices). . . . . 116

5.4 The unlabeled graph $G^*$ that all $G(\boldsymbol{w}), \boldsymbol{w} \in \mathcal{N}_*$ are isomorphic to, for $r = 6$ and $k = 4$. Its core star is colored red, while matched pairs of emanating branches are in black. . . . . 122

6.1 Ten gradients steps of cross-entropy optimization problem for $(A, X)$ drawn from a 3-class CSBM. Shown on the right are gradient paths for samples drawn from empirical and theoretical distributions for $\overline{\phi}^{(k)}$. . . . . 161

6.2 Classifier comparison for data which is 2-dimensional CSBM. On the left is the theoretical density of the 2-class CSBM. The two right plots are the estimated log-likelihood ratios for the QDA and CE estimator respectively. The slight bend in the data is correctly captured by the QDA estimator. . . . . 162

6.3 Estimated kernel density plots of the aggregated features $\overline{\phi}^{(k)}$ of a 2-class CSBM at different features depths $k$. A feature collapse in the mean vectors and the class covariances is visible by $k = 4$ and $k = 6$. . . . . 164

.4 Comparison of $\xi^{(k)}$ distribution for $k = 3$ and a fixed, expected degree Erdős–Réyni graph. As graph size increases, the overall histogram resolution may be increased but this does not qualitatively change the shape of the histogram. That is, growing degree $\nu_n \to \infty$, is a necessary condition for $\xi^{(k)}$ to be Gaussian. . . . . 197

.5 Empirical distribution of a two-class CSBM with exaggerated class proportions and edge probabilities. Both mixture components are centered at zero with a visible difference between the peak widths and heights of each component. . . . 198

# LIST OF TABLES

2.1 Debias test results on BrainWeb augmented data. . . . . 13

3.1 Clustering results for different debiasing methods for single and multi-sequence settings. . . . . 57

4.1 Performance: Mean test accuracy ± std. dev. over 10 data splits. Models include our own variations of "Linear" and "Aggregated Linear" GNNs, along with other state-of-the-art (SOTA) GNNs. Dashed entry in for LR Kernel signifies validated choice is the same as the full-rank Kernel. Performance is comparable between our simple GNNs and SOTA in some cases. Results for GPRGNN, SGC/ASGC. JacobiConv and ACMII-GCN are cited from [CPL21], [CM22], [WZ22], and [LHL22] respectively. Entries marked with '$*$' report 95% confidence intervals. . . . . 81

4.2 Simulation experiments on a three-class CSBM. Mean test accuracy and std. dev. of 10 runs are reported. $X$-only oracle is the accuracy associated with oracle classification on solely $X$. Maximum parameter counts for the two methods are also summarized. Relevant average degree $\Delta_n$ for the simulations are $\Delta_{300} = 1.83$, $\Delta_{600} = 3.68$, $\Delta_{1200} = 7.58$, and $\Delta_{1500} = 9.44$. . . . . 82

4.3 Impact of the kernel choice on the performance of the full-rank Kernel model. Underlined entries correspond to the model selected by validation. . . . . 84

# ACKNOWLEDGMENTS

I would like to thank Arash Amini whose guidance and tutelage have shaped me into the researcher I am today. Thank you for fostering my interests and for allowing me to explore a multitude of interesting math subjects. More than anything, I cherish our discussions and will take them with me far into my life.

I would like to thank the rest of my doctoral committee: Guido Montufar, Mark Handcock, Qing Zhou, and Oscar Madrid Padilla for their frequent communication and insightful questions during my doctoral journey.

Finally, I would like to thank Atchar Sudhyadhom and Jessica Scholey who ignited my passions for applied research. Thank you for your support and I appreciate all of your collaboration on our many projects.

# VITA

2015–2019 B.A. Physics and Applied Math, UC Berkeley

2020–2025 Graduate Student Researcher, UCLA.

# PUBLICATIONS

Vinas L, Amini A. A. (2025). A CLT for Polynomial GNNs on Community-Based Graphs. In preparation.

Vinas L, Amini A. A. (2024). Sharp Bounds for Poly-GNNs and the Effect of Network Noise. Submitted. eprint: https://arxiv.org/abs/2407.19567

Vinas L, Amini A. A. (2024). Simple GNNs with Low Rank Non-parametric Aggregators. *Learning on Graphs Conference (LOG 2024).*

Vinas L, Amini A. A. (2023). Step and Smooth Decompositions as Topological Clustering. Submitted. eprint: https://arxiv.org/abs/2311.05756

Vinas L., Amini A. A., Fischer J., Sudhyadhom A. (2022). LapGM: A Multisequence MR Bias Correction and Normalization Model. eprint: https://arxiv.org/abs/2209.13619

Sudhyadhom A, Scholey J, Descovich M, Kearney V, and Vinas L. (2022). Improved accuracy

of relative electron density and proton stopping power ratio through CycleGAN machine learning. *Physics in Medicine & Biology*, 67(10), 105001.

Edmonds A., Brown D., Vinas L., Pagan S. (2021). Using machine learning to select high-quality measurements. *Journal of Instrumentation*, 16(08), T08010.

Vinas L, Scholey J, Descovich M, Kearney V, Sudhyadhom A. (2021). Improved contrast and noise of megavoltage computed tomography (MVCT) through cycle-consistent generative machine learning. *Medical Physics*, 48(2), 676–690.

# CHAPTER 1

# Introduction

The advent of unsupervised algorithms has a rich history which begins in the era of bio-inspired designs for learning algorithms [Bar89]. As the field has progressed, these algorithms evolved to the familiar techniques we know today, such as $k$-means, PCA, and matrix factorization methods [HTF09]. In its most recent iteration, modern variations of unsupervised algorithms utilize deep neural networks such as the autoencoder family [Bal12] and other generative adversarial network (GAN) architectures [ZPI17].

Of particular interest to us is the original clustering algorithm, $k$-means [and58]. The $k$-means algorithm, is particularly attractive given its relatively few assumptions and its solid track record in empirical performance [AV07]. The classic $k$-means generative model assumes responses $\{y_1, y_2, \ldots, y_n\}$ which are drawn from a mixture means plus some additive noise. More formally, let $\{\mu_1, \mu_2, \ldots, \mu_L\}$ be a collection of means defined on $\mathbb{R}^d$. Then, given knowledge of the latent label $z_i \in [L]$, we obtain the following marginal distribution for $y_i$ given $z_i$,

$$y_i = \mu_{z_i} + \varepsilon_i, \tag{1.1}$$

where $\varepsilon_i$ is a zero-mean, additive noise with common variance $\sigma^2$.

With the conditional model (1.1), we abstract away specific forms for the underlying mixture distribution and describe $y_i$ in terms of realized quantities $z_i$. At the same time, (1.1) is a limited model as it constrains the responses to being elements of an unstructured point cloud. As such, any structure found in $\{y_i\}_{i=1}^n$ must be modeled by the labels $\{z_i\}_{i=1}^n$ or, alternatively, must be absorbed by the error terms $\{\varepsilon_i\}_{i=1}^n$. This has the drawback of limiting

the scope of analysis as well as violating the assumptions of (1.1).

To address this, we introduce an auxiliary structural term $S = (S_{ij})_{ij} \in \mathbb{R}^{n \times n}$ which captures pairwise dependencies between the samples $i, j \in [n]$. This auxiliary structural term $S$, which we refer to as the *structure* of the data, can be used to capture global-local dependencies of the observed samples. For example, for samples with positional coordinates $x = (x_i)_{i=1}^n$, the structure can be encoded as

$$S_{ij} = \begin{cases} s_i & \text{if } i = j, \\ s_i - s_j & \text{if } i \neq j. \end{cases} \tag{1.2}$$

In other settings we may consider the structure random, depending conditionally on the latent labels $z = (z_i)_{i=1}^n \in [L]^n$. For a community-based graph with community edge probabilities $B \in [0,1]^{L \times L}$ the structure can encoded as

$$S_{ij} = \text{Bern}(B_{z_i z_j}), \tag{1.3}$$

where $\text{Bern}(p)$ is the Bernoulli random variable parametrized by $p \in [0,1]$.

With this, the original generative formulation of (1.1) can be modified as

$$y_i = \psi_i(S, Z\mu) + \varepsilon_i \tag{1.4}$$

where $Z \in \{0,1\}^{n \times L}$ is a one-hot encoding of labels $z = (z_i)_{i=1}^L$ and $\psi : \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}$ is a problem specific mixing function between the structure $S \in \mathbb{R}^{n \times n}$ and the broadcasted means $Z\mu \in \mathbb{R}^{n \times d}$.

## 1.1 Organization

This paper is organized into two parts, with each part containing a self-contained problem highlighting the benefits of pairwise structure modeling for unsupervised clustering algorithms. Each part is motivated by an applied chapter of empirical results followed up by one or more theoretical chapters that provide insight into the earlier shown empirical results.

Part I visits the tissue recovery problem of contaminated MRIs and shows, how by modeling tissues independently from the source contamination, we can achieve both empirical improvements and novel recovery guarantees. Part II evaluates graph neural networks and their graph convolutions in popular semi-supervised node classification (SSNC) settings. By analyzing SSNC under the unsupervised lens of feature separation, we are able to obtain novel insights into the role of neighbor aggregation and how it leads to a limiting central limit theorem for the case of a fixed number of aggregations.

Material for each chapter is directly drawn from the following preprints and publications:

- Vinas L., Amini A. A., Fischer J., Sudhyadhom A. (2022). *LapGM: A Multisequence MR Bias Correction and Normalization Model.*
- Vinas L, Amini A. A. (2023). *Step and Smooth Decompositions as Topological Clustering.*
- Vinas L, Amini A. A. (2024). *Simple GNNs with Low Rank Non-parametric Aggregators.*
- Vinas L, Amini A. A. (2024). *Sharp Bounds for Poly-GNNs and the Effect of Network Noise.*

The notation of each chapter is self-contained and potentially unrelated to the notation of any subsequent chapter.

# Part I

# Level Recovery on Contaminated Images

# CHAPTER 2

# LapGM: A Rapid Multisequence MR Bias Correction and Normalization Model

## 2.1 Introduction

The maturation of machine learning methods has brought significant performance improvements to previously difficult image analyses tasks such as image segmentation [RFB15], anomaly detection [DKB13], and modality translation [YWB19]. That is not to say the gains realized by these methods are not without their own set of difficulties. In practice, large quantities of quality data are needed first before a performance ceiling is reached. Particularly in the case of medical research, it has been showed that training on poor-quality datasets leads to generalization issues such as out-of-distribution errors [JYJ20] and artifact generation [LCN19, VSD21]. A poor-quality dataset may also come about from combining multiple medical datasets into one. In these cases, the combination of images with different imaging protocols may bring about discernible, non-biological factors in the final dataset. These non-biological factors appear to degrade performance when left uncorrected for [DJN20].

To minimize such errors it is worthwhile to consider algorithmic alternatives which may be improve existing datasets through data-cleaning. Ideally, this data-cleaning routine should utilize domain-specific knowledge and be independent of any initial, training data composition. Following this principle, we show how the unsupervised method of gradient-regularized Gaussian mixtures can be used to correct for strong intensity spatial inhomogeneities as found in multi-coil parallel magnetic resonance imaging (MRI) reconstructions.

MRI intensity inhomogeneities artifacts, also known as bias fields, are non-anatomical intensity variations that vary slowly in the reconstructed image space [BMC06]. We will refer to the prominence of a bias field as a bias field's strength. The strength of a bias field depends both on patient geometry and radiofrequency (RF) receiver coils positioning in a magnetic resonance (MR) scanner [ABW10]. The greatest bias field strengths tend to be closer to the RF coils and decrease as distance away from the coil increases. Early coil technology preferred transmit and receive coils that provided more homogeneous signals at the expense of overall signal to noise (SNR). Current technology and preferences have moved towards the higher SNR coils which tend to high channel counts and higher SNR. As these types of coils will have high degrees of spatial signal heterogeneity, scanners will often include basic methods to estimate the signal sensitivity map for each coil/channel. Completely correcting for individual coil contributions at scan time can be difficult and any undercorrection may lead to the appearance of bias fields. These artifacts are made more prominent in multi-coil setups, where multiple coils interfere to produce stronger bias fields. It is common for scanner corrected images to still have significant bias field effects.

For MR images with strong bias fields, we found that state-of-the-art bias corrections [TAC10, DP14] would either under correct the bias field, leaving the image artifact intact, or over correct the bias field, reducing tissue contrast in the image. Following the Bayesian perspective of log-bias field generation suggested by W. Wells [WGK96], we highlight a class of covariance matrices that satisfy the low-pass property sought by W. Wells while remaining flexible enough to correct for strongly biased MR images. This class of matrices with sparse inverses are customizable under different weight transformations which allows for efficient and targeted correction on MR images.

Additionally, the fitted parameters of the Bayesian mixture model may be used to normalize the MR signal intensities of a scan across a set of scans within and across individuals. By combining debiasing and normalization in one step, our method decreases the possibility of chaining post-processing artifacts in the MR data cleaning pipeline. The end impact of which

is a streamlined process that may make MR data more consistent and quantitative and may have potential downstream advantages for machine learning as well.

## 2.2 Methods

Begin by considering an MR image that, due to incomplete coil sensitivity correction, has been corrupted by a spatial, multiplicative gain field. This slow-varying multiplicative field will be the bias field of our image. Our analysis will include the case that the MR scanner in question may be able to perform multiple imaging sequences at once. In the case of multiple sequences, we assume the bias field remains relatively constant between the different sequences and image reconstruction effects.

More clearly, for an $m$ sequence scan with $n$ voxels per scan let $X = [X_1, \ldots, X_n] \in \mathbb{R}^{m \times n}$ be the matrix of log-intensity measurements corrupted by some additive log-bias field $B \in \mathbb{R}^n$. Let $Z_i \in [K]$ be the tissue group label for voxel $i$. The multi-sequence, log-intensity vector $X_i$ will be conditionally characterized by the multivariate normal

$$X_i \,|\, (Z_i = k, B_i) \sim \mathcal{N}(\mu_k + B_i \mathbb{1}_m, \Sigma_k),$$

where $\mathbb{1}_m$ is the $m$-dimensional ones vector and $\mu_k \in \mathbb{R}^m$, $\Sigma_k \in \mathbb{R}^{m \times m}$ are unknown Gaussian parameters. For simplicity, a basic categorical prior is assumed on $Z_i \sim \text{Cat}(\pi)$ where tissue group $k$ has independent proability $\pi_k$ of appearing. Additionally we will assume all Gaussian parameters $\theta = \big(\pi, \{\mu_k\}_{k=1}^K, \{\Sigma_k\}_{k=1}^K\big)$ are pulled from an improper uniform prior distribution. Next incorporating knowledge that $B$ varies slowly in space, we consider the following Gaussian prior

$$B \sim \mathcal{N}\big(0, \tau L^\dagger\big),$$

where $L^\dagger$ is the pseudo-inverse of the graph Laplacian associated with spatial structure of our scan and parameter $\tau^{-1}$ is the regularization strength of the log-bias gradient penalty. The precision matrix $L$ will be explained in more detail at Section 2.2.2.

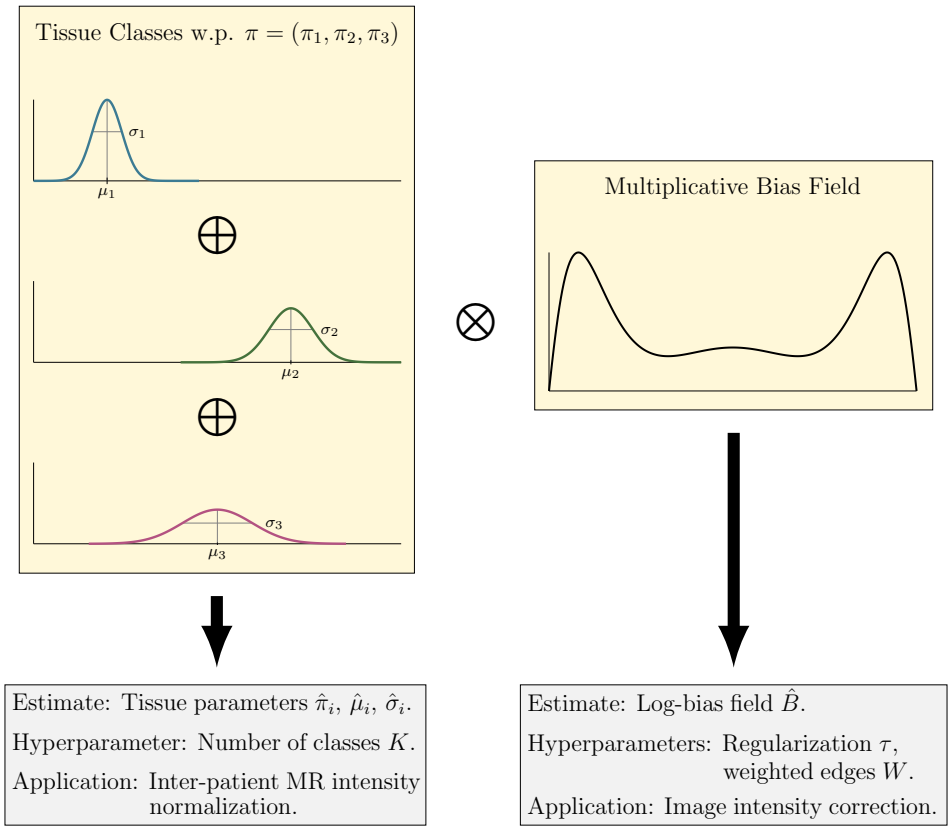



Figure 2.1: Model diagram of LapGM and its use cases. Model assumes intensity distribution follows a Gaussian mixture with additional spatial information determined by a bias field.

We will consider a posterior probability maximization for our model optimization

$$\max_{\theta, B} p_\theta\big(B \,\big|\, \{X_i\}_{i=1}^n\big). \tag{2.1}$$

In Appendix A we show how the minorization-maximization view of expectation-maximization (EM) [WL10] can be applied to (2.1) by iteratively optimizing

$$Q\big(\theta, B \,\big|\, \theta^{(t)}, B^{(t)}\big) := \mathbb{E}_{p_{\theta^{(t)}}(Z \,|\, X, B^{(t)})}[\log p_\theta(B, X, Z)], \tag{2.2}$$

with

$$\big(\theta^{(t+1)}, B^{(t+1)}\big) = \operatorname*{argmax}_{\theta, B} Q\big(\theta, B \,\big|\, \theta^{(t)}, B^{(t)}\big). \tag{2.3}$$

Note that in the optimization's current state, the parameters $\theta^{(t+1)}$ and $B^{(t+1)}$ cannot be

optimized independently of each other. Instead we decouple the expectation step from Gaussian and bias field estimation to produce the following couple closed-form updates:

- Expectation update

$$w_{ik}^{+} = \frac{\pi_k \, p_\theta\big(X_i \,|\, Z_i = k, B\big)}{\sum_{\ell=1}^{K} \pi_l \, p_\theta\big(X_i \,|\, Z_i = \ell, B\big)}. \tag{2.4}$$

- Gaussian update

$$\pi_k^{+} = \frac{1}{n} \sum_{i=1}^{n} w_{ik}, \tag{2.5}$$

$$\mu_k^{+} = \frac{1}{\sum_{i=1}^{n} w_{ik}} \sum_{i=1}^{n} w_{ik}(X_i - B_i \mathbb{1}_m), \tag{2.6}$$

$$\Sigma_k^{+} = \frac{1}{\sum_{i=1}^{n} w_{ik}} \sum_{i=1}^{n} w_{ik}(X_i - \mu_k^{+} - B_i \mathbb{1}_m) \cdot (X_i - \mu_k^{+} - B_i \mathbb{1}_m)^\top. \tag{2.7}$$

- Bias field update

$$B^{+} = \Big(\frac{1}{\tau} L + \sum_{k=1}^{K} \mathbb{1}_m^\top \Sigma_k^{-1} \mathbb{1}_m \operatorname{diag}(w_{\cdot k})\Big)^{-1} \cdot \Big(\sum_{k=1}^{K} \operatorname{diag}(w_{\cdot k})(X - \mu_k \mathbb{1}_n)^\top \Sigma_k^{-1} \mathbb{1}_m\Big). \tag{2.8}$$

Notation $(\cdot)^{+}$ indicates a subsequent iteration parameter estimate given the current parameter estimate $(\cdot)$. These block updates may be carried out in a randomly-permuted order [Nes12, SLY20] for increased parameter exploration. The updates are derived using the first order conditions of a concave objective. Derivations can be found in Appendix B.

### 2.2.1 Connection to Previous Works

The Laplacian-regularized, Gaussian mixture can be softly understood in the general framework of W. Wells [WGK96] where Gaussian prior

$$B \sim \mathcal{N}(0, \tau \, \psi_B)$$

uses a low-pass filter matrix $\psi_B$ for its covariance. To understand this frameworks connection to (2.4-2.8), note that the Laplacian matrix $L$ picks out high frequency content, such as edges, from areas of changing contrast. Under this frequency view, the inverse process $L^\dagger$ can be seen as a low-pass transformation which degrade high frequency content with extraneous low frequency content. As an added benefit, $L^\dagger$ can be understood as a proper inverse to signals $B$ with zero-mean.

This formulation, although exceedingly general, faced a combination of design and computational issues. The first difficulty was to construct a matrix $\psi_B$ which was positive-semidefinite and shared similarities to a low-pass filter. The second difficulty grappled with the computational cost of inverting

$$H = \frac{1}{\tau}\psi_B^{-1} + \sum_{k=1}^{K} \mathbb{1}_m^\top (\Sigma_k^-)^{-1} \mathbb{1}_m \operatorname{diag}\big(w_{\cdot k}^-\big),$$

which itself contained a matrix inverse $\psi_B^{-1}$.

The first issue could be addressed fairly generally by considering positive-definite kernel $K$ with fixed window $T$ and expanding $K$ to its Toeplitz matrix form $\psi_B$ for image dimensions $[n]^d$. This procedure would produce large but sparse covariance matrices with non-zero entries on the order $\mathcal{O}(nT^d)$. The downside to this approach is that there is no guaratee $\psi_B^{-1}$ itself will be sparse and in practice a dense inverse seems to be common. A dense $\psi_B^{-1}$ would make $H$-inversion computationally intractable in an iterative EM method. This problem may be partially solved by computing a convex program on $\psi_B$ which produces approximate sparse inverses $\hat{\psi}_B^{-1}$ depending on some known sparseness value $\alpha$ [FHT07]. However it is not clear how the forced sparseness of $\hat{\psi}_B^{-1}$ with parameter $\alpha$ will affect the low-pass properties of $\hat{\psi}_B$, potentially compromising the effectiveness and validity of the low-pass prior.

In the original paper of W. Wells, this was circumvented by assuming $H^{-1}$ could be approximated by a uniform filter matrix with kernel window of around 15 to 30 pixels. While computationally tractable, this heuristic is not directly connected to any known covariance matrix and, for its current setting, the large smoothing window could affect correction

effectiveness in the case of compact and prominent inhomogeneities.

By beginning with a penalty $B^\top LB$ and working backward to a generative model, we are able to avoid the inversion process $L^\dagger \to L$, enforce sparsity for $H$, and maintain the low-pass intuition for covariance matrix $L^\dagger$.

### 2.2.2 Gradient Weighting Heuristic

It is common in multi-coil setups for bias intensity to increase towards the boundary of the patient anatomy. In order to account for the spatial dependence in multi-coil bias fields, consider the graph Laplacian of an undirected graph $G = (V, E)$ with non-negative edge-weights $W$

$$L_{ij} = \begin{cases} -W_{ij}, & \text{if } i \neq j, \\ \sum_{k \in \mathrm{Nbr}(i)} W_{ik}, & \text{if } i = j, \end{cases}$$

where $\mathrm{Nbr}(i) \coloneqq \{j \in [n] : (i,j) \in E \text{ or } (j,i) \in E\}$. A reasonable heurestic would be to more sharply relax the gradient penalty as voxels approach the boundary of the patient's anatomy. One reasonable choice could be the inverse power function

$$h_\alpha(x, y, z) = (x^2 + y^2)^{-\alpha}, \quad \text{with } \alpha \geq 0,$$

with cylindrical symmetry about some $z$-axis. For efficiency, the Laplacian edge-weights are constructed through a set of vertex evaluations

$$W_{ij} = \mathbf{1}\{i > j\} \cdot h(V_j) + \mathbf{1}\{i < j\} \cdot h(V_i),$$

where $V_i$ is the spatial position of the $i$th voxel. For a $d$-dimension grid graph, distinct indices $i, j \in [n]^d$ are ordered as $i > j$ if and only if there exists some $k' \in [d]$ such that

$$\forall k \geq k',\ i_k > j_k \quad \text{and} \quad \forall k < k',\ i_k = j_k.$$

This inequality evaluation can be modified independently from the rest of the Bayesian model.

## 2.3 Materials and Experiments

We compare LapGM to the industry-standard N4ITK (N4) debiasing method. All calculations were run on a 64-core Linux machine equipped with a 3090 Nvidia GPU. The LapGM model was run using the author provided CUDA-accelerated Python package `lapgm`[1], while N4 was run using SimpleITK's [LCI13] multi-threaded CPU implementation.

LapGM and N4 were compared using simulated and real-world data. Performance on simualted data was evaluated using a 50-50 validation-testing split. During the validation phase a total of 120 hyperparameter combinations were evaluated for both LapGM and SimpleITK's N4. As suggested by the original N4 authors, all images were downsampled using a 2-factor downsample before running estimation. At inference all estimated bias fields were then upsample to their original image dimensions.

Simulated data was generated using bias simulation software `biasgen` [VS22] in conjuction with the noiseless, 0% RF non-uniformity BrainWeb dataset [CKK97]. A total of 10 bias fields were generated for the simulated dataset. The bias simulation settings include a sampling half-width of $L = 12$ per dimension and sampling rate of $(0.5, 1.75, 1.75)$ in the $(\omega_z, \omega_y, \omega_x)$ Fourier space. For the sampling grid $G$, plane $\omega_z = 0$ was omitted to produce more pronounced bias fields.

For real-world data, MR patient scans were acquired on a Siemens 3T Vida with thorax and pelvis body regions being scanned using a 32 channel posterior spine array. All scans were acquired using a gradient-echo based VIBE Dixon dual echo sequence with lowest possible repetition time (TR) and time to echo (TE) values. Image resolution was $2 \times 2 \times 2\,\text{mm}$ isotropic and patient images were retrospectively included in this Insitutional Review Board (IRB) approved study.

[1]Code available at https://github.com/lucianoAvinas/lapgm.

| Experiment | Method | Bias | Debias | Runtime [s] | |
|---|---|---|---|---|---|
| | | RMSE [1] | RMSE [1] | CPU | GPU |
| Line32 | N4 | 0.1684 | 2334 | 623 | — |
| Line32 | LapGM 1-seq. | 0.1889 | 2564 | 64.1 | 7.96 |
| Line32 | LapGM 3-seq. | 0.0614 | 934.7 | 86.2 | 13.6 |
| Rect3_Angn90 | N4 | 0.0738 | 2791 | 605 | — |
| Rect3_Angn90 | LapGM 1-seq. | 0.0720 | 2680 | 53.7 | 7.78 |
| Rect3_Angn90 | LapGM 3-seq. | 0.0236 | 930.5 | 86.2 | 14.7 |
| Rect4 | N4 | 0.0729 | 2798 | 481 | — |
| Rect4 | LapGM 1-seq. | 0.0745 | 2624 | 84.3 | 7.76 |
| Rect4 | LapGM 3-seq. | 0.0220 | 840.7 | 90.4 | 15.4 |
| Rect5 | N4 | 0.0699 | 2975 | 546 | — |
| Rect5 | LapGM 1-seq. | 0.0687 | 2586 | 87.4 | 6.53 |
| Rect5 | LapGM 3-seq. | 0.0607 | 1166 | 104 | 13.7 |
| Rect7 | N4 | 0.0566 | 2508 | 600 | — |
| Rect7 | LapGM 1-seq. | 0.0787 | 3223 | 71.1 | 7.99 |
| Rect7 | LapGM 3-seq. | 0.0245 | 1046 | 90.5 | 16.2 |

Table 2.1: Debias test results on BrainWeb augmented data.

## 2.4 Results and Discussion

Let $B^e, X^e$ be defined by the element-wise exponentiation

$$B_i^e \coloneqq \exp(B) \quad \text{and} \quad X_i^e \coloneqq \exp(X_i).$$

Simulated data were evaluated using the following metrics:

$$\text{Bias RMSE} = \left( \frac{1}{|\Omega|} \sum_{i \in \Omega} \left( B_i^e - \hat{B}_i^e \right)^2 \right)^{1/2},$$

$$\text{Debias RMSE} = \left( \frac{1}{|\Omega|} \sum_{i \in \Omega} \left( X_i^e / B_i^e - X_i^e / \hat{B}_i^e \right)^2 \right)^{1/2}.$$

For notation $\hat{B}$ is the estimated bias field are $\Omega$ is the set of indices contained within the tissue mask provided by BrainWeb. Simulated test results for N4 and LapGM can be seen in Table 2.1.

### 2.4.1 Simulated Data

Table 2.1 shows comparable performance between N4 and LapGM in the single-sequence setting and superior LapGM performance in the multi-sequence case. Large consistent improvements can be found in the debias RMSE for LapGM 3-sequence. The metric with the smallest improvement was tissue total variation. Fig. 2.2 and 2.3 have been provided to better understand performance differences in these metrics. Table 2.1 also shows that, for the given testing environment, LapGM runs significantly faster that N4 with a near 80-fold improvement in runtime for the GPU-accelerated case.

Fig. 2.2 shows the debiasing result for the Line32 biased experiment. N4 and LapGM show similar errors in the central tissue group of the BrainWeb phantom, with LapGM's errors closely following the contour of the central tissue group. At the boundary of the phantom's anatomy we see a clear difference between the errors of N4 and LapGM. Here N4 has trouble adapting to the spatial variations found in the simulated bias field. As mentioned before, multi-channel RF configurations have the tendency to sharply increase bias at the boundary of the patient's anatomy. This is something we are able to account for when setting up the graph Laplacian for LapGM.

Fig. 2.3 shows the recovered tissue distributions for N4 and LapGM. A few features we will focus on in the recovered tissue distributions are: number of peaks recovered, width/location

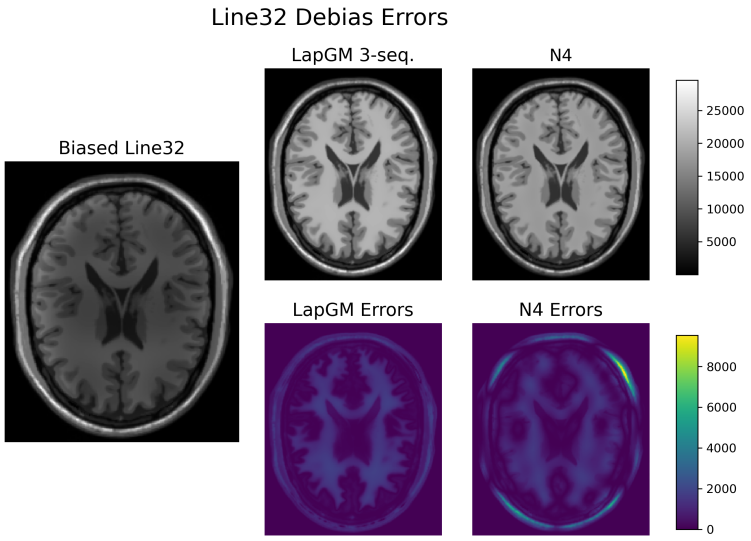



Figure 2.2: Debias error results on Line32 with N4 and LapGM 3-seq. algorithm. Second column of this figure reveals LapGM's tendency to increase contrast for certain dominant tissue groups.

of peaks, and recovery consistency. With these qualitative metrics we will be able to glean more insight into how the N4 and LapGM methods work.

First is the number of peaks recovered. In this regard LapGM 3-seq. is able to consistently identify major tissue groups between differently biased examples. Depending on the kind of bias field, a correctly calibrated N4 may sometimes recover the major tissue groups. In the 1-sequence setting, LapGM does not have as much data to contrast different tissue groups. As such, LapGM 1-seq. shows the tendency of overlapping certain tissue groups which are similar in intensity.

For the width and location of peaks, LapGM 3-seq. shows the sharp recovery with each peak being near its original location. N4's performance depends on the number of peaks recovered but in general shows a peaked recovery. Here LapGM 1-seq. is faced with the same issue from before with peak width and location being off for the last two tissue groups.

Next is the topic of consistent recovery. To LapGM 1-seq.'s benefit this is a category it performs fairly well in. LapGM 3-seq. performs similarly well and N4 has issues with consistent

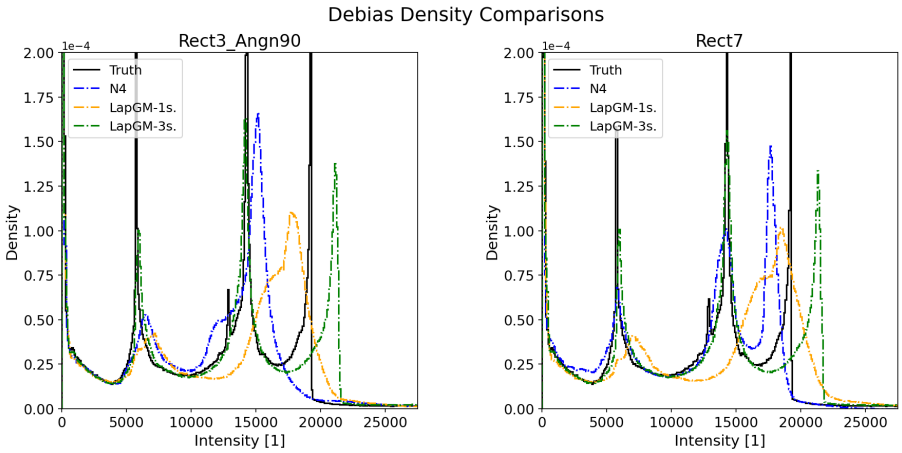



Figure 2.3: Probability density comparison between true tissue density and debiased tissue density functions.

distribution recovery. Inconsistent distribution recovery has negative downstream impacts to any supervised learning model. Modern machine learning methods can be remarkably good at accounting for missing or corrupted information, as long as this missingness or corruption is consistent between data samples. In the case of N4, inconsistent distribution recovery could lead to contradictory training signals for supervised methods which rely on N4 for data cleaning.

Lastly we identify the outlier on LapGM 3-seq. bias RMSE performance on testing data Rect5. Rect5 features five rectangular coils in a pentagon pattern arround the BrainWeb phantom. The deviation in LapGM 3-seq.'s perfomance may be an issue of incomplete convergence during the optimization process. In practice, incomplete optimization can be evaluated by analyzing whether the final bias field estimate of LapGM is too smooth or too rough. After identifying incomplete optimization can be corrected for by modifying the regularization strength $\tau^{-1}$ for the example in question.

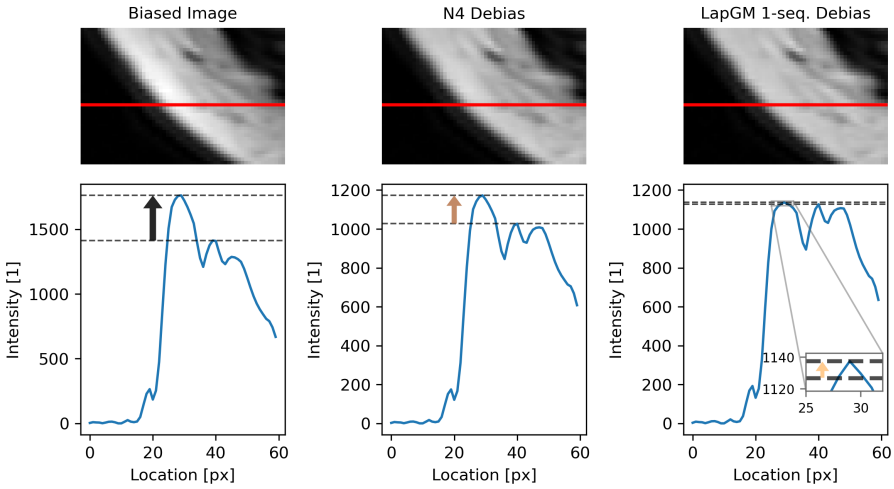



Figure 2.4: Line profiles of the lower-left section of patient pelvic region. Line profile in red (top) with corresponding intensity shown in blue (bottom). Top dotted line is placed on dotted peak and the bottome dotted line is placed on the second peak next to the trough.

### 2.4.2 Patient Data

For our patient scan analysis we will focus on the debiasing results of a patient's pelvic between N4 and the single sequence LapGM model, in particular we will be focussing on how image contrast can be balanced for improved bias correcting performance. Fig. 2.4 shows slice line profiles for the lower-left section of a patient scan. The full slice with its bias comparison can be found in Fig. 2.5. The leftmost column of Fig. 2.4 shows a compact but bright bias peak at around the 25 pixels mark. At this point, N4 shows a rough 60% reduction in bias while the LapGM single-sequence model shows an almost complete removal of bias. The degree of bias correction for LapGM can be modified through the regularization stength parameter $\tau^{-1}$. Although stronger debiasing capabilities come at the cost of image contrast, this setting of the LapGM single-sequence model is able to preserve meaningful contrast. This can be verified by noting both N4 and LapGM share similar trough amplitudes for the 35 pixel mark.

Fig. 2.5 gives a visual comparison on trade-off between bias removal and image contrast for

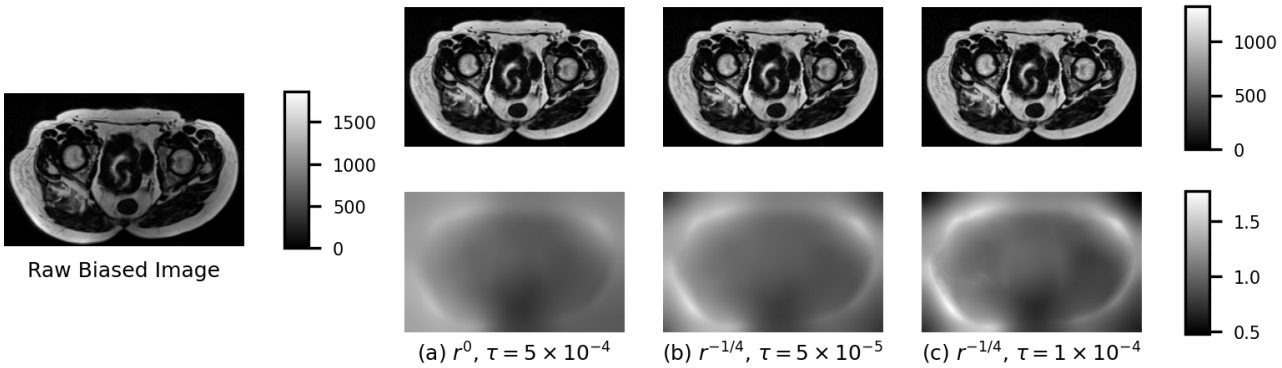



Figure 2.5: Debias and bias field comparison of single sequence LapGM for varying regularizations $\tau$ and gradient penalty weights $h_\alpha(r, z)$.

different $\tau$ regularizations and edge-weight settings. With decaying gradients weights, LapGM can fit for bias fields that are stonger along the extremity of the patient's anatomy. Fig. 2.5 also shows that the rate at which regularization $\tau$ trades contrast for bias correction may depend on the chosen edge-weight function $h_\alpha(r)$. As shown in sub-Fig. 2.5b, practitioners may choose to leave a weak underlying bias field for better image contast.

### 2.4.3 Normalization Comparisons

We will refer to any normalization scheme that utilizes the fitted Gaussian parameters of LapGM as a LapGM-based normalization scheme. In its full generality, an involved LapGM-based normalization scheme could incorporate both class posterior probabilities $w$ with Gaussian information $\theta$. That said, it is still possible to get good normalization results with simple LapGM-based normalization schemes. In this section we analyze a LapGM-based normalization called $\mu_*$ normalization which is done by taking the largest fitted mean value

$$\mu_* = \max_{k \in [K]} \mu_k$$

and applying the appropriate scaling factor $\beta$ which scales $\mu_*$ to some target intensity value of choice. A visualization of the $\mu_*$ normalization process is shown in Fig. 2.6.

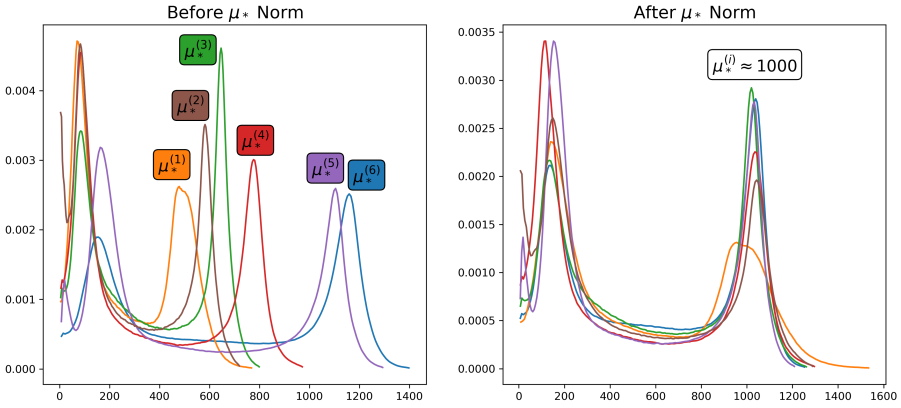



Figure 2.6: Visualization of the $\mu_*$ normalization technique with real patient data. After normalization, the second peak of each patient distribution clusters around the target intensity of 1000.

The LapGM $\mu_*$ normalization was compared to the max normalization and Z-score normalization techniques. These comparisons were done relative to a water-masked region-of-interest (ROI) normalization baseline. The intensity distribution produced from each normalization is shown in Fig. 2.7. A total of 10 LapGM-debiased patient scans were used for the normalization comparison. To allow for meaningful comparisons between the techniques, all three normalizations were peak-aligned to the baseline water mask ROI normalization.

Tissue total variations were computed for each of the peak-aligned distributions. The LapGM $\mu_*$ normalization had a TV error of 9.52%, the max normalization had a TV error of 14.0%, and the Z-score normalization had a TV error 21.7%. The Z-score normalization featured a significant left tail which was not show in Fig. 2.7. This left tail contributed Z-score's larger TV error calculation.

One important qualitative difference between the recovered distributions of $\mu_*$ normalization and the other normalization techniques, is the presence of bumps along the recovered tissue distributions. Small bumps, like the ones visible along the max and Z-score normalization distributions, are an indication of intensity scaling mismatch between different patients. As shown by Fig. 2.6, each of the 10 patient distributions feature a prominent second peak. Incorrect inter-patient scalings will cause these peaks to be misaligned and as a result form

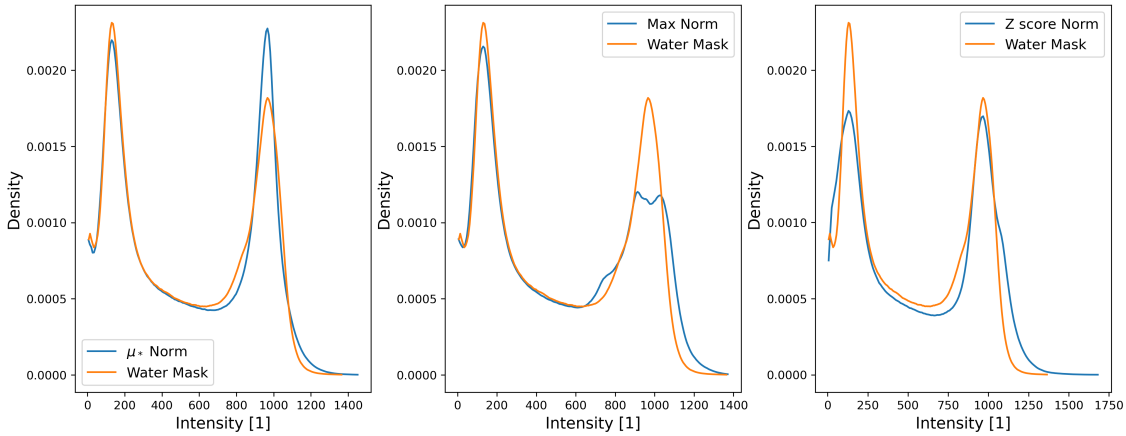



Figure 2.7: Combined tissue distributions of 10 patient scans using different normalization techniques. The distributions produced from each normalization technique is peak-aligned with the tissue distribution of a water mask ROI normalized distribution. Starting from the leftmost column the tissue TV errors relative to the water mask ROI are 9.52%, 14.0%, and 21.7%.

small bumps along the recovered tissue distribution. Note that this may even be the case for the small bump visible at the 800 intensity value for the water mask ROI distribution. As each water mask ROI is calculated by hand, it is possible for slight errors to accrue in the final aggregate tissue distribution.

This analysis was carried out on images where clear contours could be made for the different water ROIs, however we note that similar peak recovery can achieved by applying $\mu_*$ normalization to fat-dominant MR sequences as well.

## 2.5 Conclusion

In this paper, a multi-sequence debiasing algorithmic alternative to the common N4 debiasing algorithm is proposed. We have shown superior performance in various metrics for the multi-sequence case, as well as ease-of-use and interpretability in the single-sequence case. Both methods were tested on a variety simulated bias field configurations specified by [VS22] as well as real-world patient data. Implementations for LapGM and the bias generation can

be found in the Python packages `lapgm` and `biasgen` respectively.

In this paper a normalization technique using the fitted Gaussian parameters of LapGM is also proposed. This normalization is competitive with manual water mask ROI techniques and superior to max normalization and Z-score normalization techniques. Only one of the many possible LapGM-based normalization technique were analyzed. It may be possible to further improve normalization performance by utilizing more of LapGM's Gaussian parameters.

Future work could probe some theoretical properties of LapGM or attempt to explain some of the distributional tendencies of LapGM. As Fig. 2.3 showcases, N4 has a tendency to produce tissue peaks which vary based on bias field, while LapGM, both in the single and multi-sequence settings, seems to produce stable tissue distributions. Some other interesting phenomena to explore include why debias methods in general struggle to disentangle peak locations of high intensity tissue groups and under what conditions will multi-sequence LapGM increase tissue contrast rather than decrease it.

# Appendix

## A Expectation-Maximization for MAP

Let $X \coloneqq \{X_i\}_{i=1}^n$. We are interested in the maximum a posteriori probability (MAP) for

$$\operatorname*{argmax}_{\theta,B} p_\theta(B \mid X),$$

where $\theta$ and $B$ are conditionally independent from each other and an improper uniform prior $p(\theta) = 1$ is assumed on $\theta$. As the argmax is invariant to scalings and monotonic transforms, it is equivalent to consider the log joint probability

$$\operatorname*{argmax}_{\theta,B} \log p_\theta(X, B),$$

where normalization factor $p_\theta(X)$ has been dropped.

A function $f(\psi)$ is said to be minorized by $g(\psi \,|\, \psi^{(t)})$ at $\psi = \psi^{(t)}$ if

$$g\big(\psi \,|\, \psi^{(t)}\big) \leq f(\psi),\ \forall \psi \quad \text{and} \quad g\big(\psi^{(t)} \,|\, \psi^{(t)}\big) = f\big(\psi^{(t)}\big).$$

For shorthand let $X \coloneqq \{X_i\}_{i=1}^n$. The goal will be to show

$$\begin{aligned} g\big(\theta, B \,|\, \theta^{(t)}, B^{(t)}\big) \coloneqq\ & Q\big(\theta, B \,|\, \theta^{(t)}, B^{(t)}\big) \\ & + \log p_{\theta^{(t)}}\big(X, B^{(t)}\big) \\ & - Q\big(\theta^{(t)}, B^{(t)} \,|\, \theta^{(t)}, B^{(t)}\big) \end{aligned}$$

is a minorizing function of $(\theta, B) \mapsto \log p_\theta(X, B)$ at $(\theta^{(t)}, B^{(t)})$ where

$$Q\big(\theta, B \,|\, \theta^{(t)}, B^{(t)}\big) \coloneqq \mathbb{E}_{p_{\theta^{(t)}}(Z \,|\, X, B^{(t)})}[\log p_\theta(Z, X, B)].$$

This can be done by introducing an auxilliary distribution $q(Z)$ and decomposing the KL divergence of $q$ and $p_\theta$

$$D(q \,||\, p_\theta) \coloneqq \mathbb{E}_{Z \sim q}\Big[\log \frac{q(Z)}{p_\theta(Z \,|\, X, B)}\Big].$$

With some manipulation

$$\begin{aligned} D(q \,||\, p_\theta) &= \log p_\theta(X, B) + \mathbb{E}_q\big[\log q(Z) - \log p_\theta(Z, X, B)\big] \\ &= \log p_\theta(X, B) - \mathbb{E}_q[\log p_\theta(Z, X, B)] \\ &\quad + \mathbb{E}_q\big[\log\big(q(Z) p_{\theta^{(t)}}(X, B^{(t)})\big)\big] \\ &\quad - \log p_{\theta^{(t)}}(X, B^{(t)}). \end{aligned}$$

Now set $q(Z) = p_{\theta^{(t)}}(Z \,|\, X, B^{(t)})$ and rearrange

$$\begin{aligned} \log p_\theta(X, B) = & Q\big(\theta, B \,|\, \theta^{(t)}, B^{(t)}\big) + \log p_{\theta^{(t)}}\big(X, B^{(t)}\big) \\ & - Q\big(\theta^{(t)}, B^{(t)} \,|\, \theta^{(t)}, B^{(t)}\big) + D(p_{\theta^{(t)}} \,||\, p_\theta). \end{aligned}$$

The non-negativity and equality properties of KL divergence

$$D(q \,||\, p) \geq 0,\ \forall q \quad \text{and} \quad D(q \,||\, p) = 0,\ \text{if } p = q,$$

show that $g\big(\theta, B \,|\, \theta^{(t)}, B^{(t)}\big)$ is indeed a minorizing function of $(\theta, B) \mapsto \log p_\theta(X, B)$ at $(\theta^{(t)}, B^{(t)})$. Lastly since $g\big(\theta, B \,|\, \theta^{(t)}, B^{(t)}\big)$ only has one term which depends on $(\theta,\, B)$, the maximization step may be simplified to

$$\begin{aligned}\big(\theta^{(t+1)}, B^{(t+1)}\big) &= \operatorname*{argmax}_{\theta, B} g\big(\theta, B \,|\, \theta^{(t)}, B^{(t)}\big) \\ &= \operatorname*{argmax}_{\theta, B} Q\big(\theta, B \,|\, \theta^{(t)}, B^{(t)}\big).\end{aligned}$$

## B MAP Coordinate Updates

With the shorthands $a_{ik} \coloneqq X_i - \mu_k - B_i \mathbb{1}_m$ and $w_{ik} = p_{\theta^{(t)}}\big(Z_i = k \,|\, X_i, B_i^{(t)}\big)$, we expand the following objective

$$\begin{aligned}-Q\big(\theta, B \,|\, \theta^{(t)}, B^{(t)}\big) \propto \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \Big( &\frac{1}{2} a_{ik}^\top \Sigma_k^{-1} a_{ik} - \frac{1}{2} \log |\Sigma_k^{-1}| \\ &- \log \pi_k \Big) + \frac{1}{2\tau} B^\top L B.\end{aligned}$$

We will be interested in the optimization

$$\min_{\substack{\mu \in \mathbb{R}^m,\, B \in \mathbb{R}^n, \\ \Sigma_k \in \mathbb{R}^{m \times m} \colon \Sigma_k \succ 0, \\ \pi_k \in [0,1] \colon \sum_{k=1}^K \pi_k = 1}} -Q\big(\theta, B \,|\, \theta^{(t)}, B^{(t)}\big).$$

Parameter $\pi_k$ can be independently optimized as

$$\pi_k^* = \frac{1}{n} \sum_{i=1}^{n} w_{ik}.$$

The simplified objective

$$\mathcal{L} = \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{w_{ik}}{2} \Big( a_{ik}^\top \Sigma_k^{-1} a_{ik} - \log |\Sigma_k^{-1}| \Big) + \frac{1}{2\tau} B^\top L B$$

can be optimized as an unconstrained problem

$$\min_{\substack{\mu \in \mathbb{R}^m,\, B \in \mathbb{R}^n \\ \Sigma_k^{-1} \in \mathbb{R}^{m \times m} \colon \Sigma_k^{-1} \succ 0}} \mathcal{L}(\mu, \Sigma^{-1}, B),$$

where for $\Sigma_k \succ 0$ it is equivalent to optimize over the reparametrized $\Sigma_k^{-1}$. The local optimality conditions of $\mathcal{L}$ are

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu_k} &= \sum_{i=1}^{n} -w_{ik}\, \Sigma_k^{-1} a_{ik} = 0, \\
\frac{\partial \mathcal{L}}{\partial \Sigma_k^{-1}} &= \sum_{i=1}^{n} w_{ik}\, a_{ik} a_{ik}^{\top} - w_{ik} \Sigma_k = 0, \\
\frac{\partial \mathcal{L}}{\partial B_i} &= \sum_{k=1}^{K} -w_{ik} \mathbb{1}_m^{\top} \Sigma_k^{-1} a_{ik} + \frac{1}{\tau}(LB)_i = 0.
\end{aligned}$$

This last condition can be rewritten in vector notation

$$\frac{\partial \mathcal{L}}{\partial B} = \sum_{k=1}^{K} \operatorname{diag}(w_{\cdot k})\, (X - \mu_k \mathbb{1}_n - \mathbb{1}_m B^{\top})^{\top} \Sigma_k^{-1} \mathbb{1}_m - \frac{1}{\tau} LB,$$

where $\operatorname{diag}(w_{\cdot k})$ is a diagonal matrix with entries $w_{ik}$. Parameter updates can be done in a block-cooridinate fashion until convergence.

After sufficiently optimizing to a new set of parameters $(\theta', B')$, $w_{ik}$ can be updated as

$$w_{ik} = \frac{\pi_k'\, p_{\theta'}\big(X_i \,|\, Z_i = k, B'\big)}{\sum_{\ell=1}^{K} \pi_l'\, p_{\theta'}\big(X_i \,|\, Z_i = \ell, B'\big)}.$$

For practical applications, the authors have found that sequential updates of the form $(w \to \theta \to B \to w \to \ldots)$ yield quick and stable convergences.

As a final comment, we note that neither the original log-likelihood nor its expectation-step surrogate is necessarily convex for all $\Sigma_k^{-1} \succ 0$. For the case of a standard Gaussian mixture, it has been shown [JZB16] that overall performance is dependent on the quality of initial estimate parameter. In practice, a K-means initialization step with a potential data transform can help to avoid bad local minima.

# CHAPTER 3

# Step and Smooth Decompositions as Topological Clustering

## 1 Introduction

The prototypical recovery problem is nonparametric regression where we observe an unknown function corrupted by additive white noise: $y_i = f^*(x_i) + \varepsilon_i$, for $i = 1, \ldots, n$, where $f^*$ belongs to some function class $\mathcal{F}$ and $\varepsilon_i$ is the measurement noise. Important to the recovery is the structure of $\mathcal{F}$ and how it can be leveraged to differentiate observations from noise. Examples of previously explored structures in nonparametric regression include: smoothness [Tsy09], sparsity [Wai09, BRT09], homogeneity [KFW15], and piecewise simplicity [KKB09, Tib14]. In each of these problems, there is a particular interest in uncovering the structure-specific recovery conditions under which a finite-sample, data-estimate $\widehat{f}$ eventually recovers the optimal, data-generating $f^*$.

Another flavor of recovery problems include decompositions of the form

$$y_i = f^*(x_i) + g^*(x_i) + \varepsilon_i, \tag{3.1}$$

where the recovery quantities of interest include both $f^*$ and $g^*$. Naturally, this type of recovery problem, with its multiple recoverable quantities, is more difficult than basic nonparametric regression. Examples of such decompositions with provable recovery guarantees are rare but some notable examples include the case of sparse plus low-rank matrix recovery [CSP09, BR16, TV23] and compressed sensing in a pair of orthogonal bases [DK13].

In this paper, we consider a nonparametric decomposition of the form (3.1) where the signal is a combination of continuous and step functions. We provide identifiability conditions for the continuous and step functions $f^*$ and $g^*$ in terms of the modulus of continuity of $f^*$ and the height between steps in $g^*$. Analysis of $f^*$ and $g^*$ will be sufficiently general, where each function is considered to be a mapping from a metric space $(\mathcal{X}, d)$ to a normed vector space $(\mathcal{Y}, \|\cdot\|)$.

In its simplest formulation, we consider $f^*$ to be real-valued and continuous, lying in a Hilbert-norm $R$-ball of a reproducing kernel Hilbert space (RKHS). For this scenario, a practical estimation algorithm is proposed with consistency guarantees given in terms of spectral quantities related to the observed kernel matrix of the RKHS.

As in most regression analysis, we conduct our analysis under finite sampling constraints. For $g^*$ which attains at most $M$ unique values within a given sample, the composite observations will be re-expressed as

$$y_i = f^*(x_i) + \mu^*_{z^*_i} + \varepsilon_i, \quad \text{for } i = 1, \ldots, n \tag{3.2}$$

where $\boldsymbol{\mu}^* \in \mathbb{R}^M$ is a vector of values referred to as the levels of $g^*$, and $z^*_i \in [M]$ are labels to the corresponding levels of $g^*$. Our main goal is to recover the labels $z^*_i$ correctly, with a secondary goal of recovering the levels $\boldsymbol{\mu}^*$ and the continuous function $f^*$. For our finite sample setting, recovery of $f^* \in \mathcal{F}$ will be relaxed to finding an element of the equivalence class

$$[f^*]_n = \big\{ f \in \mathcal{F} : \ f(x_i) = f^*(x_i), \ \forall i \in [n] \big\}. \tag{3.3}$$

This recovery condition may be refined to instead selecting a representative solution from $[f^*]_n$, such as a minimum-norm solution. An approach of this sort will depend on the regularity available in the function space $\mathcal{F}$ and will not be a topic of focus in our forthcoming analysis.

## 1.1 Applications

To motivate the problem, let us give some concrete applications of the step and smooth decomposition model (3.2).

### Decompositions in Non-linear ICA

Non-linear independent component analysis (ICA) [HP99] provides a general framework to describe signal mixing problems. In non-linear ICA, the mixed observation $\boldsymbol{y} = \psi(\boldsymbol{s})$ is generated using independent, latent sources $\boldsymbol{s} \in \mathbb{R}^n$ and a non-linear, mixing function $\psi : \mathbb{R}^n \to \mathbb{R}^n$. In other ICA formulations [HST19], joint independence of $\boldsymbol{s}$ is relaxed to a conditional independence given some auxiliary information $\boldsymbol{u} \in \mathbb{R}^n$. That is,

$$\log p(\boldsymbol{s}|\boldsymbol{u}) = \sum_{i=1}^{n} q_i(s_i, \boldsymbol{u})$$

for appropriately defined densities $q_i$.

Decomposition (3.2) can be understood in terms of a self-mixing, non-linear ICA problem. In the simplest scenario, we may consider sources $s_i = (x_i, u_i)$ with auxiliary information $u_i \sim \text{Unif}[0, 1)$ and mixing defined by

$$\psi(s_i) = f^*(x_i) + \mu^*_{\phi(x_i,u_i)} \quad \text{where} \quad \phi(x_i, u_i) = \lfloor u_i \cdot M \rfloor + 1. \tag{3.4}$$

Generalizations to (3.4) may consider different cut-off functions $\phi(x_i, u_i)$ which also incorporate sample spatial information $x_i$ in their cut-offs.

In contrast to traditional ICA problems, the mixing function defined in (3.4) is not necessarily injective on $\mathbb{R}^n$ for all choices of $f^*$ and $\phi$. This a recovery setting not covered in recent non-linear ICA literature [HST19, KKM20, ZNZ22] and one we are interested in exploring in this paper. In particular, when given partial information $\{(x_i, y_i)\}_i$, which properties of the data, if any at all, can help overcome the non-injectivity of a general $f^*$ and $\mu^*_\phi$?

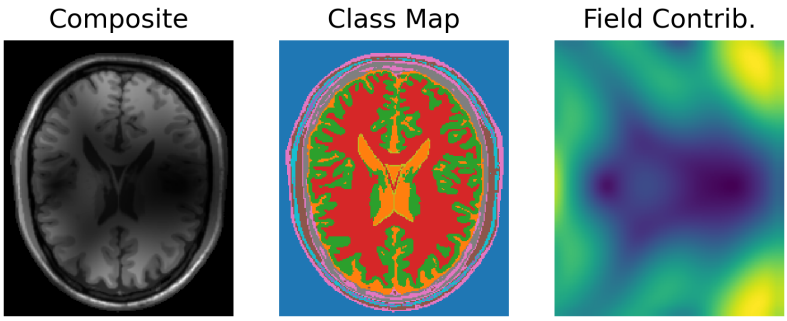



Figure 3.1: Example of a prominent bias field modifying the BrainWeb [CKK97] phantom. Leftmost image is the MRI of a synthetic brain which has been perturbed by a contaminant field. Middle image is a tissue categorization for the synthetic brain. Rightmost image is the contaminant field in ambient space (in absence of the synthetic brain).

### Decompositions in Medical Image Correction

In magnetic resonance imaging (MRI), image quality can be affected by factors ranging from radiofrequency coil setup to patient positioning and geometry [ABW10]. Dependent on these factors, MRI images may be contaminated with a spatially smooth, multiplicative field, known as the bias field. Figure 3.1 illustrates an example of a contaminated MRI image.

The MRI bias field problem admits the following multiplicative formulation [VPL07],

$$y(x) = f^*(x) \cdot \mu^*(x), \quad \text{for } x \in \mathcal{X} \tag{3.5}$$

where $f^*$ is a positive smooth field on $\mathcal{X}$, and $\mu^*(x)$ are, by convention, positive tissue values at locations $x \in \mathcal{X}$. Given a fixed number of tissues classes $M$, process (3.5) can be reformulated as (3.2) under a log-transformation.

In supervised learning tasks, the visual inconsistencies caused by MRI bias fields present significant challenges, as they prevent the acquisition of accurate ground truth signal information from patient scans. This issue parallels the earlier discussed problem of non-linear ICA, where, again, learning is hampered due to partial information and concerns regarding injectivity.

### 1.2 Prior Work

To the authors' best knowledge, the closest work on the theory of continuous and step decompositions is [KT14], where they provide a characterization of the set of viable functions given an observed composite signal $h^*$. The composite $h^* = f^* \cdot g^*$ is assumed to be the product of a positive continuous function $f^*$ and a positive step-wise function $g^*$. Assuming knowledge of the tissue ratios $\{\mu_k/\mu_{k+1}\}_{k=1}^{M-1}$, [KT14] have shown that one there are scalars $\{a_k\}_{k=1}^{M}$ such that the set

$$\widetilde{\mathcal{F}} = \left\{ f : \mathcal{X} \to \mathbb{R} : \forall x \in \mathcal{X},\, f(x) \in \{a_k h^*(x)\}_{k=1}^{M} \right\}$$

contains a unique scalar multiple of $f^*$. This result is then followed by a practical algorithm which optimizes over a soft-label surrogate of $\widetilde{\mathcal{F}}$.

The theoretical result of [KT14] is interesting since it dramatically reduces the search space for a viable $f$, esp. when $\mathcal{X}$ is finite. What this result does not tell us is how to identify $f^*$ in the set $\widetilde{\mathcal{F}}$, and whether $f^*$ is identifiable at all. This issue becomes readily apparent in finite sample scenarios, where there may be multiple ways to construct observations $h^*$ from different smooth-and-step pairs $(\tilde{f}, \tilde{g})$. In short, the work of [KT14] does not address the question of identifiability which is a focus of our work. Moreover, when no level information is available, $\widetilde{\mathcal{F}}$ itself is unknown. In this regime, attempts to approximate the set $\widetilde{\mathcal{F}}$ would ultimately be sensitive to initialization choice for scale parameters $\{a_k\}_k$.

## 2 Identifiability Theory

We consider the problem of identifying components $(f^*, \boldsymbol{\mu}^*, \boldsymbol{z}^*)$ from observations

$$y_i = f^*(x_i) + \mu^*_{z_i^*}, \quad \text{for } i = 1, \ldots, n, \tag{3.6}$$

where $f^*$ belongs to a class of smooth functions, from a metric space $(\mathcal{X}, d)$ to a normed space $(\mathcal{Y}, \|\cdot\|)$. Specifically, we assume that $f^* \in \mathcal{F}_\omega(\mathcal{X})$, the set of uniformly continuous

functions with *modulus of continuity* $\omega : [0, \infty) \to [0, \infty)$, that is,

$$\mathcal{F}_\omega(\mathcal{X}) := \big\{ f : \mathcal{X} \to \mathcal{Y} : \ \|f(x) - f(x')\| \leq \omega(d(x, x')), \ \forall x, x' \in \mathcal{X} \big\}. \tag{3.7}$$

For ease of presentation we will assume $\mathcal{Y} = (\mathbb{R}, |\cdot|)$. Proofs of the results for a general normed space $\mathcal{Y}$ can be found in Appendix A.

A model is identifiable if the ground-truth parameters $(f^*, \boldsymbol{\mu}^*, \boldsymbol{z}^*)$ can be unambiguously recovered from observed samples $\{y_i\}_i$ following (3.6). For our recovery procedure we consider solving the optimization

$$(\widehat{f}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{z}}) = \underset{\substack{f \in \mathcal{F}_\omega(\mathcal{X}), \\ \boldsymbol{\mu} \in \mathbb{R}^M, \, \boldsymbol{z} \in [M]^n}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_{z_i} - f(x_i))^2, \tag{3.8}$$

as well as *zero-mean* version where a constraint is added ensuring that $f$ is empirically zero-mean, i.e., $\sum_{i=1}^n f(x_i) = 0$. This zero-mean constraint addresses issues analogous to the scalar multiple problem described by [KT14]. Note that, in our case, correctly recovering $(\boldsymbol{\mu}^*, \boldsymbol{z}^*)$ will also recover $[f^*]_n$ since $f^*(x_i) = y_i - \mu^*_{z_i^*}$. Then, by providing conditions under which (3.8) unambiguously recovers the sampled clusters $\{\mu^*_{z_i^*}\}_i$, we will have shown identifiability for the step and smooth decomposition (3.6).

## 2.1 Topological Clustering

To motivate our forthcoming topological definitions, consider the following failure case for step and smooth identifiability:

**Example 1.** Consider the two cluster case ($M = 2$) on $\mathcal{X} = (\mathbb{R}^d, \|\cdot\|_2)$ and take $\omega(t) = t$, that is, $\mathcal{F}_\omega(\mathcal{X})$ contains all 1-Lipschitz functions on $\mathcal{X}$. Let $f^*(\boldsymbol{x}) = 0$ and $\mu_1^* = -\mu_2^* = 1$ with linearly separable clusters $\mathcal{C}_k = \{i \in [n] : \ z_i^* = k\}$; that is, there exists unit-norm $\boldsymbol{w} \in \mathbb{R}^d$ and $c_1, c_2 \in \mathbb{R}$ such that

$$\boldsymbol{w}^T \boldsymbol{x}_i \leq c_1 < c_2 \leq \boldsymbol{w}^T \boldsymbol{x}_j$$

for all $i \in \mathcal{C}_1$ and $j \in \mathcal{C}_2$.

Consider the piecewise cluster-interpolating function $\widetilde{f}(\boldsymbol{x}) = -\min\{\max\{\boldsymbol{w}^T\boldsymbol{x} - c_1, 0\}, 2\}$. Clearly, $\widetilde{f}$ is 1-Lipschitz, that is $\widetilde{f} \in \mathcal{F}_\omega(\mathcal{X})$. Now, if $\mathcal{C}_1$ and $\mathcal{C}_2$ are sufficiently separated so that $c_2 - c_1 \geq 2 = \mu_1^* - \mu_2^*$, then

$$\widetilde{f}(\boldsymbol{x}_i) + \mu_1^* = f^*(\boldsymbol{x}_i) + \mu_{z_i^*}^*$$

for every $i \in [n]$. That is, rather than two cluster, we can put all points in a single cluster (Cluster 1) and explain the variation in $y_i$ entirely by the smooth function $\widetilde{f}$, or we can put the points in two clusters and explain the residual variation (which turns out to be zero in this case) by the original $f^*$. In other words, in this case the observations $y_i = f^*(\boldsymbol{x}_i) + \mu_{z_i^*}^*$ do not allow for an unambiguous recovery of $(f^*, \boldsymbol{\mu}^*, \boldsymbol{z}^*)$.

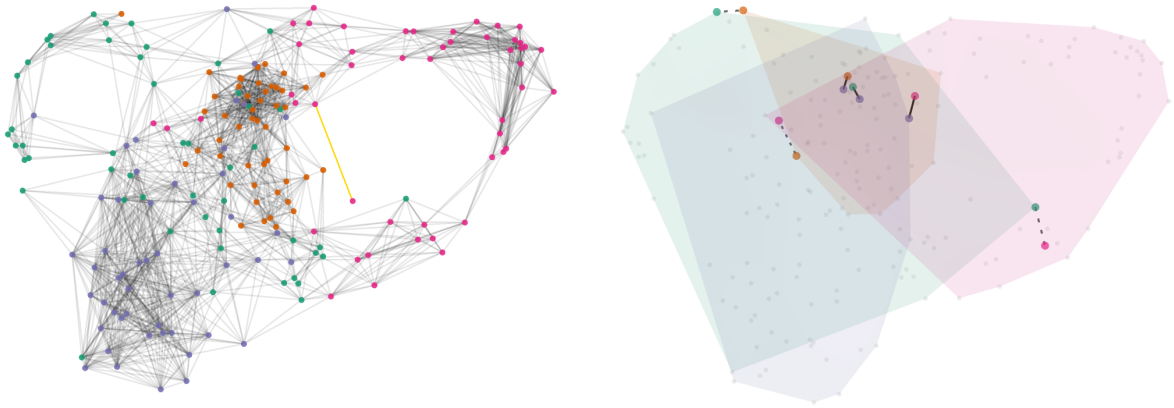
Example 1 shows that without a constraint on how far apart samples in the point cloud $\{x_i\}_{i=1}^n$ are placed, it is always possible to construct examples where one can interpolate between different clusters using a smooth $\widetilde{f}$. Stated differently, the distance between samples, compared to the difference in cluster levels $\mu_k^*$ for $k \in [M]$, must be within the scale allowed by the modulus of continuity $\omega$. It turns out the the proper way to measure the separation of samples is through the connectivity of the $\rho$-neighbor graph, which we recall next:

**Definition 1** (Neighbor Graph)**.** The $\rho$-neighbor graph $G_\rho(X)$ of point cloud $X = \{x_i\}$ is the graph with vertex set $[n]$ and edge set

$$\big\{(i,j) \in [n]^2 : i \neq j \ \text{ and } d(x_i, x_j) \leq \rho\big\}.$$

The $\rho$-neighbor graph captures some aspect of the topology of the point cloud. Paired with the modulus of continuity $\omega$, this graph allows us to quantify long-range variation of a particular $f^* \in \mathcal{F}_\omega(\mathcal{X})$ via its local variations along the edges.

In this sense, every point cloud $X$ has a minimum, necessary communication length $\rho_{\min}(X)$, such that all long-range variations in $\{d(x, x')\}_{x,x' \in X}$ can be bounded in terms of local ones $\{d(x_i, x_j)\}_{(i,j)\in E}$ where $G_{\rho_{\min}}(X) = ([n], E)$. More formally:



(a) $\rho$-neighbor connectivity graph

(b) Cluster connectivity graph

Figure 3.2: A $\rho$-neighbor and cluster connectivity graph on UMAP-reduced features for four topics from the "20 Newsgroups" classification dataset. Highlighted in gold in the left subfigure is an edge with length equal to connectivity $\rho_{\text{min}}$. Drawn in black in the right subfigure are the corresponding cluster distance edges $d(\mathcal{C}_k, \mathcal{C}_\ell)$. Cluster graph edges which are larger than $\delta_{\text{lbd}}$ are drawn in dashed. The final cluster graph $G_{\delta_{\text{lbl}}}(\mathcal{C})$ is a tree with a connecting hub at the blue colored cluster.[1]

**Definition 2** (Connectivity)**.** For a point cloud $X$, the connectivity is defined as

$$\rho_{\min}(X) := \inf\{\rho > 0 : \ G_\rho(X) \text{ is connected}\}.$$

So far, we have defined a connectivity parameter, $\rho_{\min}(X)$, such that deviations in $f^*$ can be translated into traversals between neighboring nodes in $X$. Let us define a similar concept for the deviations of the step component $\mu^*_{(\cdot)}$. Let $\mathcal{C} = \{\mathcal{C}_k\}_{k\in[M]}$ be the set of $M$ clusters with $\mathcal{C}_k = \{i \in [n] : z^*_i = k\}$. For each pair of clusters, there is a corresponding notion of *cluster distance* given by:

$$d(\mathcal{C}_k, \mathcal{C}_\ell) := \min_{i\in\mathcal{C}_k,\, j\in\mathcal{C}_\ell} d(x_i, x_j). \tag{3.9}$$

[1]For interactive 3D network representation: https://github.com/lucianoAvinas/topological-clustering-plots.

Then, for an associated tolerance parameter $\delta > 0$, let us construct the $\delta$-neighbor graph $G_\delta(\mathcal{C})$ with vertex set $\mathcal{C}$ and edge set

$$\big\{(\mathcal{C}_k, \mathcal{C}_\ell) : \ k \neq \ell \text{ and } \ d(\mathcal{C}_k, \mathcal{C}_\ell) \leq \delta\big\}.$$

Similar to the smooth case, there is a minimum necessary communication length $\delta_{\,\mathrm{lbl}}$, depending on both the point cloud $X$ and the set of labels $\boldsymbol{z}^*$, such that all deviations of $\mu^*_{(\cdot)}$ can be translated to traversals on $G_\delta(\mathcal{C})$.

**Definition 3** (Label distance)**.** The label distance for paired data $(X, \boldsymbol{z}^*)$ is

$$\delta_{\,\mathrm{lbl}}(X, \boldsymbol{z}^*) \coloneqq \inf\big\{\delta > 0 : \ G_\delta(\mathcal{C}) \text{ is connected}\big\}. \tag{3.10}$$

When it is clear from context, we omit dependence on sample $(X, \boldsymbol{z}^*)$ for the previously defined topological quantities. An example of these quantities in real-world data is given Figure 3.2.

Finally, we also need the following simple condition on the labels:

**Definition 4** (Label saturation)**.** A label vector $\boldsymbol{z}^* \in [M]^n$ *saturates* $[M]$ if every label in $[M]$ is present in $\boldsymbol{z}^* = (z_1^*, \ldots, z_n^*)$, that is, for every $\ell \in [M]$, there is $i \in [n]$, with $z_i^* = \ell$.

This condition is needed to avoid the trivial case where some of the levels $\{\mu_k^*\}$ are redundant (i.e. label $k$ does not appear until some $n > N$). It is clearly necessary for the identifiability of the levels and labels in the uncontaminated model. It is always possible to redefine $\boldsymbol{z}^*$ to be saturated by relabeling, and dropping redundant levels.

## 2.2 Identifiability Results

Our main result is the following cluster recovery guarantee:

**Theorem 1** (Cluster recovery)**.** *Let $X = \{x_i\}_{i=1}^n$ be a point cloud in a metric space $(\mathcal{X}, d)$ and let $\{y_i\}_{i=1}^n$ follow model* (3.6) *with $f^* \in \mathcal{F}_\omega(\mathcal{X})$ and $\boldsymbol{z}^* \in [M]^n$ that saturates $[M]$. If the*

*connectivity* $\rho_{\min}$ *of* $X$ *satisfies*

$$\omega(\rho_{\min}) < \frac{1}{2M} \min_{k \neq \ell} |\mu_k^* - \mu_\ell^*|, \tag{3.11}$$

*then, the labels* $\widehat{\boldsymbol{z}}$ *produced by* (3.8) *have zero misclassification error relative to* $\boldsymbol{z}^*$.

Our next result is an error bound on the recovered levels $\widehat{\boldsymbol{\mu}}$:

**Proposition 1** (Level recovery)**.** *Under the assumptions of Theorem 1, let* $(\widehat{f}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{z}})$ *be the solution of the zero-mean version of problem* (3.8)*. Then, we have*

$$\max_{k \in [M]} |\mu_k^* - \widehat{\mu}_k| \leq 2(M-1)\, \omega(\delta_{\,lbl}) + \Big| \frac{1}{n} \sum_{i=1}^{n} f^*(x_i) \Big|. \tag{3.12}$$

In essence, both Theorem 1 and Proposition 1 provide deviation bounds under specific connectivity constraints. The quantities $\rho_{\min}$ and $\delta_{\,\mathrm{lbl}}$ gauge the minimum jump distances at which the induced graphs of $\{x_i\}_{i=1}^n$ and $\{\mathcal{C}_k\}_{k=1}^M$ remain connected. The modulus $\omega(\cdot)$ then translates these jumps in distances into equivalent jumps in levels, observed indirectly through $\{y_i\}_i$.

Theorem 1 says that perfect cluster recovery $\widehat{\boldsymbol{z}} \equiv \boldsymbol{z}^*$ is attainable if this translated jump is roughly below the minimum resolution of the true levels $\{\mu_k^*\}$. Proposition 1 has a similar theme but now in the context of level recovery where, unlike $\boldsymbol{z}^* \in [M]^n$, the levels $\mu_k^* \in \mathbb{R}$ lie on a continuum. This leads to a gradual reduction in error as outlined in Proposition 1, contrasting with the sharp recovery of discrete labels $z_i^* \in [M]$ in Theorem 1.

The remainder term $|\frac{1}{n} \sum_i f^*(x_i)|$ in (3.12) highlights the scalar-shift ambiguity inherent in the components of model (3.6), where for any scalar $c \in \mathbb{R}$, it is possible to rewrite (3.6) as

$$y_i = (f^*(x_i) - c) + (\mu_{z_i^*}^* + c).$$

As such, the two components are only identifiable up to a scalar shift. More generally, problem (3.8) can be extended to include a constraint $\frac{1}{n} \sum_{i=1}^n f(x_i) = \overline{f^*}$ for some select mean value $\overline{f^*}$, in which case Proposition 1 holds with the remainder term $|\frac{1}{n} \sum_i f^*(x_i) - \overline{f^*}|$.

As an immediate corollary to Proposition 1, one can show that, under mild regularity on the sampling of $(X, \boldsymbol{z}^*)$, the zero-mean recovery problem (3.8) achieves asymptotic identifiability of $(\boldsymbol{\mu}^*, \boldsymbol{z}^*)$. As this corollary references multiple sets of samples, the notation $(\cdot)^{(n)}$ will be used to differentiate parameters belonging to different sets of observations $\{y_i\}_i$. We also allow the number of observed levels $M_n$ to grow with $n$. We say that a condition is *eventually satisfied* if it holds for all $n \geq N$ for some $N \in \mathbb{N}$.

**Corollary 1.** *Consider a sequence of point clouds $\{X^{(n)}\}$, with corresponding true labels $\{\boldsymbol{z}^{*(n)}\}$ and class levels $\boldsymbol{\mu}^{*(n)} \in \mathbb{R}^{M_n}$. Let $\delta_{lbl}^{(n)}$ be the label distance for $(X^{(n)}, \boldsymbol{z}^{*(n)})$. Assume that the connectivity condition* (3.11) *is eventually satisfied, and as $n \to \infty$,*

$$\omega(\delta_{lbl}^{(n)}) = o(M_n^{-1}), \quad \frac{1}{n} \sum_{x \in X^{(n)}} f^*(x) = o(1).$$

*Then for any solution $(\widehat{f}^{(n)}, \widehat{\boldsymbol{\mu}}^{(n)}, \widehat{\boldsymbol{z}}^{(n)})$ of the zero-mean version of problem* (3.8)*,*

$$\lim_{n \to \infty} \max_{k \in M_n} \left| \mu_k^{*(n)} - \widehat{\mu}_k^{(n)} \right| = 0.$$

According to Corollary 1, when $\{M_n\}$ is bounded, a set of sufficient conditions for recovery of both clusters and levels is:

$$\rho_{\min}^{(n)} = o(1), \quad \delta_{\text{lbl}}^{(n)} = o(1), \quad \text{and} \quad \Delta_n := \min_{k \neq \ell} |\mu_k^{*(n)} - \mu_\ell^{*(n)}| = \Omega(1),$$

i.e., minimum level gap is bounded below. When $\{M_n\}_n$ is unbounded, both the connectivity $\rho_{\min}$ and the label distance $\delta_{\text{lbl}}$ must decrease more rapidly. For example, when the smooth component is Lipschitz (i.e., $\omega(t) = Lt$), a set of sufficient conditions are

$$\rho_{\min}^{(n)} = o(\Delta_n / M_n), \quad \delta_{\text{lbl}}^{(n)} = o(1/M_n).$$

Note that, while Corollary 1is a deterministic result, it can be translated to a high probability version given appropriate assumptions on the sampling distribution of $(X, \boldsymbol{z})$.

The identifiability results of this section are intuitive and are described in terms of easily understood topological quantities. It is worth emphasizing that, prior to our analysis,

obtaining a perfect classification result similar to Theorem 1 is not immediately clear for a general context. That is, irrespective of the placements of labels $\boldsymbol{z}^*$ on the point cloud $X$, and regardless of the dimension of the space carrying $X$, we have shown that one can globally control $\widehat{\boldsymbol{z}}$ using only a scalar parameter of the point cloud, namely, the radius of connectivity of its associated neighbor graphs $G_\rho(X)$.

## 3 Methods and Optimization

For practical estimation, we consider estimating functions $f^* \in \mathbb{H}$ lying in the Hilbert-norm $R$-ball of an RKHS. The following example shows that this case can be treated as a special case of (3.7) with a linear modulus $\omega(t) = O(t)$.

**Example 2.** Consider the case where $f^*$ lies in RKHS $\mathbb{H}$. The natural metric to consider on $\mathcal{X}$ is the so-called *kernel metric*

$$d_{\mathcal{K}}(x, x') := \|\mathcal{K}(x, \cdot) - \mathcal{K}(x', \cdot)\|_{\mathbb{H}} = \sqrt{\mathcal{K}(x, x) - 2\mathcal{K}(x, x') + \mathcal{K}(x', x')}. \tag{3.13}$$

Using the Cauchy–Schwarz inequality, it is straightforward to show the following Lipschitz property: For any $f \in \mathbb{H}$, we have

$$|f(x) - f(x')| \leq \|f\|_{\mathbb{H}}\, d_{\mathcal{K}}(x, x')$$

for all $x, x' \in \mathcal{X}$. Letting $\omega_f$ denote a modulus of continuity of function $f$, the above shows that one can take $\omega_f(t) = \|f\|_{\mathbb{H}} \cdot t$ for all $f \in \mathbb{H}$. If we further assume $\|f^*\|_{\mathbb{H}} \leq R$ for some constant $R$, then $\omega(t) = O(t)$.

**AltMin Algorithm** For our estimation procedure, we propose a blockwise coordinate descent with alternating updates on $(\boldsymbol{\mu}, \boldsymbol{z})$ and $f$. More specfically, in each iteration, the

current estimates $(\widehat{f}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{z}})$ are updated to the new ones $(\widehat{f}^+, \widehat{\boldsymbol{\mu}}^+, \widehat{\boldsymbol{z}}^+)$ by

$$\widehat{f}^+ = \operatorname*{argmin}_{f \in \mathbb{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{\mu}_{\widehat{z}_i} - f(x_i))^2 + \tau \|f\|_{\mathbb{H}}^2, \tag{3.14}$$

$$(\widehat{\boldsymbol{\mu}}^+, \widehat{\boldsymbol{z}}^+) = \operatorname*{argmin}_{\boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{z} \in [M]^n} \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_{z_i} - \widehat{f}^+(x_i))^2, \tag{3.15}$$

with $\tau$ and $M$ being values to be determined through a cross-validation procedure.

For fixed $\widehat{f}$, optimization (3.15) can be solved through a $k$-means procedure. For RKHS $\mathbb{H}$ equipped with kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, optimization (3.14) has the following representer solution

$$\widehat{f}^+ = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\alpha}_i^+ \mathcal{K}(x_i, \cdot), \quad \widehat{\boldsymbol{\alpha}}^+ := (K + \tau I_n)^{-1} (\boldsymbol{y} - \widehat{Z} \widehat{\boldsymbol{\mu}}) / \sqrt{n} \tag{3.16}$$

where $K$ is the $n \times n$ kernel matrix with entries $K_{ij} = \mathcal{K}(x_i, x_j)/n$ and $\widehat{Z} \in \{0,1\}^{n \times L}$ is the one-hot encoding label matrix for previous label estimate $\widehat{\boldsymbol{z}}$.

### 3.1 One-step Analysis

In general, the interaction between updates (3.14) and (3.15) may be quite complicated. In this section we show a positive result: In the large sample limit, classification with AltMin simplifies to classification with regular $k$-means on the uncontaminated (step) signal.

We consider observations $\{y_i\}_i$ drawn from (3.1) with i.i.d. zero-mean noise $\varepsilon_i$ of variance $\sigma^2$. As before, $g^*$ will be assumed to be a step signal with $g^*(x_i) = \mu^*_{z_i^*}$, although the results of this section hold for any $g^*$ that is *sufficiently outside* the RKHS, as will be made precise in Theorem 2. For our analysis, we consider a half-step of the AltMin algorithm, evaluating performance after update (3.16). Our goal is to show the pointwise consistency of the KRR estimator $\widehat{\boldsymbol{f}} \coloneqq (\widehat{f}(x_i))$, that is

$$\lim_{n \to \infty} \mathbb{E}_{\boldsymbol{\varepsilon}} \operatorname{MSE}(\boldsymbol{f}^*, \widehat{\boldsymbol{f}}) = 0, \tag{3.17}$$

where $\operatorname{MSE}(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|_2^2 / n$.

Let $K = V\Lambda V^\mathsf{T}$ be the eigenvalue decomposition of the kernel matrix where $\Lambda = \mathrm{diag}(\lambda_i, i \in [n])$, and define

$$h(\lambda;\tau) := \frac{\lambda^2}{(\lambda+\tau)^2}, \quad \Gamma_\tau := \sqrt{h(\Lambda;\tau)} = \Lambda(\Lambda+\tau I)^{-1}$$

extending a scalar function to diagonal matrices in the natural way (i.e., by applying to each diagonal entry.) We assume the eigenvalues are ordered as follows: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Consider the Fourier expansion of $f^*$ and $g^*$ in the (empirical) eigen-basis of the kernel, that is, $\breve{\boldsymbol{f}} \coloneqq (\breve{f}_i) \coloneqq V^\mathsf{T}\boldsymbol{f}^*$ and $\breve{\boldsymbol{g}} \coloneqq (\breve{g}_i) \coloneqq V^\mathsf{T}\boldsymbol{g}^*$. Then

$$\begin{aligned}
\frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon}}\|\boldsymbol{f}^* - \widehat{\boldsymbol{f}}\|_2^2 &= \frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon}}\|(I_n-\Gamma_\tau)V^\mathsf{T}\boldsymbol{f}^* - \Gamma_\tau V^\mathsf{T}\boldsymbol{g}^* - V^\mathsf{T}\boldsymbol{\varepsilon}\|_2^2 \\
&\leq \frac{2}{n}\|(I_n-\Gamma_\tau)\breve{\boldsymbol{f}}\|_2^2 + \frac{2}{n}\|\,\Gamma_\tau\breve{\boldsymbol{g}}\,\|_2^2 + \frac{\sigma^2}{n}\,\mathrm{tr}(\Gamma_\tau^2) \\
&= \frac{2}{n}\sum_{i=1}^n \frac{\tau^2\breve{f}_i^2}{(\lambda_i+\tau)^2} + \frac{2}{n}\sum_{i=1}^n h(\lambda_i;\tau)\,\breve{g}_i^2 + \frac{\sigma^2}{n}\sum_{i=1}^n \frac{\lambda_i^2}{(\lambda_i+\tau)^2}
\end{aligned} \tag{3.18}$$

The first and the third terms are the bias and variance, respectively, for recovering $\boldsymbol{f}^*$ in classical kernel ridge regression (KRR). Both can be made to go to zero as $n \to \infty$ for a proper choice of $\tau = \tau_n = o(1)$. The middle term is new to our decomposition, and is the filtering effect of KRR on the step component $g^*(\cdot)$.

To expand on the filtering behavior of the middle term, we first establish a result on the Hilbert norm of minimum-norm interpolants for functions which are not in the continuous RKHS $\mathbb{H}$, the proof of which can be found in Appendix B.

**Proposition 2.** *Let $\mathcal{H}$ be an RKHS of real-valued functions on the metric space $\mathcal{X}$. Assume that the RKHS has a continuous kernel $K(\cdot,\cdot)$ and let $\{x_i\}_{i\geq 1} \subset \mathcal{X}$ be dense in $\mathcal{X}$. Consider a function $g^* : \mathcal{X} \to \mathbb{R}$ that is not in $\mathbb{H}$. Let $\bar{f}_n$ be the minimum $\mathbb{H}$-norm interpolation of $g^*$ based on $\{x_i\}_{i=1}^n$, that is,*

$$\bar{f}_n \in \underset{\substack{f\in\mathbb{H},\\ f(x_i)=g^*(x_i),\,\forall i}}{\mathrm{argmin}} \|f\|_{\mathbb{H}}. \tag{3.19}$$

*Then, $\|\bar{f}_n\|_{\mathbb{H}} \to \infty$ as $n \to \infty$.*

It is well-known [Wai19, Section 12.5] that the minimum-norm interpolant (3.19) can be written as $\bar{f}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \bar{\boldsymbol{\alpha}}_i \mathcal{K}(x_i, \cdot)$ where $\bar{\boldsymbol{\alpha}} = K^{-1}\boldsymbol{g}^*/\sqrt{n}$ and $\boldsymbol{g}^* = (g^*(x_1), \ldots, g^*(x_n))$. Here, we recall that the kernel matrix $K_{ij} = \mathcal{K}(x_i, x_j)/n$ is entrywise-normalized by $n$. It follows that

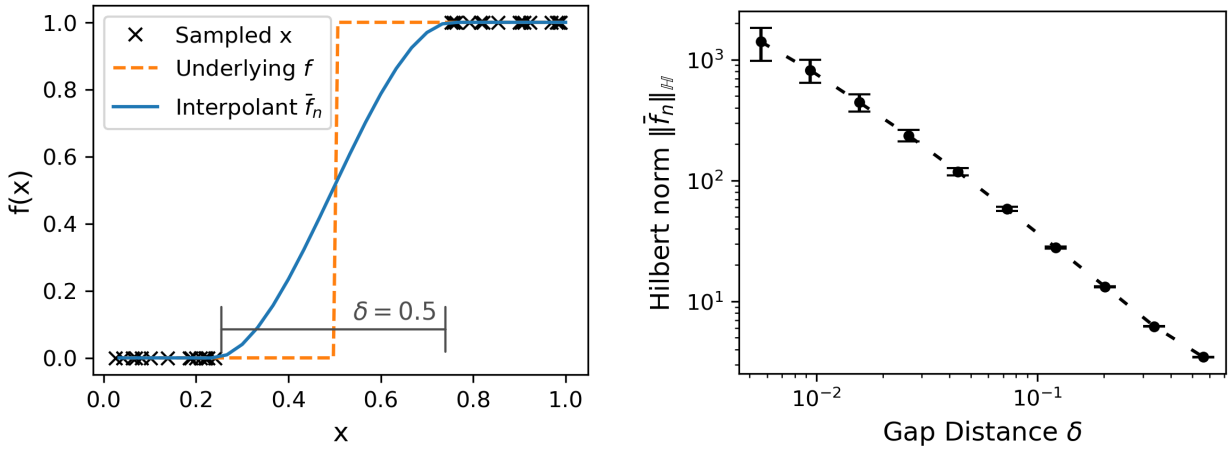
$$\|\bar{f}_n\|_{\mathbb{H}}^2 = \bar{\boldsymbol{\alpha}}^T K \bar{\boldsymbol{\alpha}} = (\boldsymbol{g}^*/\sqrt{n})^T K^{-1} (\boldsymbol{g}^*/\sqrt{n}) = \frac{1}{n} \sum_{i=1}^n \breve{g}_i^2/\lambda_i. \tag{3.20}$$

Proposition 2 then implies that if $g^* \notin \mathbb{H}$ and the sequence $\{x_i\}_{i=1}^n$ becomes dense in $\mathcal{X}$, the Fourier coefficients of $g^*$ with respect to $K$, that is $(\breve{g}_i)$, exhibit the following drift:

$$\frac{1}{n} \sum_{i=1}^n \breve{g}_i^2/\lambda_i \to \infty. \tag{3.21}$$

Figure 3.3 illustrates a pathological scenario in which the "density" requirement is violated. Here, $\mathcal{H}$ is a Sobolev-2 RKHS on $[0, 1]$, consisting of continuous functions, while $g^*$ is a step function with a discontinuity at 0.5; thus, $g^* \notin \mathcal{H}$. The sequence $\{x_i\}_{i \geq 1}$ is deliberately constructed to maintain a gap of size $\delta$ around the discontinuity, resulting in a bounded $\mathcal{H}$-norm for the interpolant $\bar{f}_n$. However, as $\delta \to 0$, the norm diverges, as demonstrated in Figure 3.3(b), consistent with Proposition 2. This example is adversarially constructed; in contrast, under i.i.d. sampling from a continuous distribution on $[0, 1]$, the sequence $\{x_i\}_i$ is almost surely dense.

The sample-level, spectral drift (3.21) of $g^*$ opens up the possibility of KRR effectively filtering out contributions of $g^*$ and making the middle term in (3.18) negligible. To see this, note that since $\lambda_i$ are decaying as a function of $i$, for the expression $\frac{1}{n} \sum_{i=1}^n \breve{g}_i^2/\lambda_i$ to grow without bound, most of the energy of $g^*$ (where "energy" is defined as $\frac{1}{n} \sum_{i=1}^n \breve{g}_i^2$) must be concentrated on the higher-index components, which correspond to smaller eigenvalues. Multiplication by $h(\lambda_i, \tau)$ filters out components of $g^*$ associated with small eigenvalues; equivalently it acts a a low-pass filter, filtering out higher index (i.e., higher frequency) components.




(a) RKHS interpolant with pathological sampling.

(b) $\mathbb{H}$-norm as distance to discontinuity decreases.

Figure 3.3: Sobolev-2 interpolants for step functions on $[0, 1]$. As sampled points $\{x_i\}_i$ approach a discontinuity of $f$, the corresponding interpolant $\bar{f}_n$ has Hilbert norm $\|\cdot\|_{\mathbb{H}}$ going to infinity.

To make the above intuition more precise, consider the *spectral survival function* of $g^*$:

$$S_{g^*}(t) := \sum_{i=1}^{n} \frac{\breve{g}_i^2}{n} \cdot 1\{\lambda_i > t\}. \tag{3.22}$$

As $t \to \infty$, $S(t)$ goes to zero, and the faster this decay, the more $g^*$ is concentrated on higher-index components. That is, the tail behavior of $S_{g^*}(t)$ is what determines how well $g^*$ is filtered by KRR. Let $r_n = \max\{i \in [n] : \breve{g}_i^2 > 0\}$ and let $\beta_n$ be the largest $\beta \geq 0$ that satisfies

$$S_{g^*}(t) \;\leq\; \|\boldsymbol{g}^*\|_\infty^2 \cdot \left(\frac{\lambda_{r_n}}{t}\right)^{\beta}, \qquad \text{for all } t > 0. \tag{3.23}$$

Such a tail bound always exists, since the trivial case $\beta = 0$ reduces to $\|\boldsymbol{g}^*/\sqrt{n}\|_2^2 \leq \|\boldsymbol{g}^*\|_\infty^2$. The parameters of the tail bound are influenced by how much the higher-index components of $\breve{\boldsymbol{g}}$ contribute to the total norm (or energy). The tail bound works together with the spectral filter $h(\lambda; \tau)$ to give the following control for the middle term of (3.18):

**Proposition 3.** *Consider KRR with regularization parameter $\tau_n$ and let $\xi_n := \lambda_{r_n}/\tau_n$. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} h(\lambda_i; \tau_n)\, \breve{g}_i^2 \;\lesssim\; \max\{\xi_n^2,\, \xi_n^{\beta_n}\}, \tag{3.24}$$

*where $\lesssim$ denotes inequality up to universal constants.*

We note that the best case scenario in Proposition 3 is obtained when $r_n = n$ and $\beta_n \geq 2$, leading to the quickest possible decay of $O(\xi_n^2) = O((\lambda_n/\tau_n)^2)$ for the residual norm.

Next we consider the case where $\mathcal{X}$ is compact and $\mathcal{K}$ is continuous, that is, kernel $\mathcal{K}$ is a Mercer kernel. Then, under the assumption that $\{x_i\}$ are i.i.d. draws, the sampling operator associated with $K$ converges compactly, almost surely, to an integral operator $T_{\mathcal{K}} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ [LBB08, Proposition 11-13]. This in turn implies that as long as $r_n \to \infty$, we will have $\lambda_{r_n} \to 0$. Combined with Proposition 3, this lead to the following consistency result for the one-step procedure:

**Theorem 2.** *Consider a Mercer kernel and i.i.d. sample $\{x_i\}_i$. Let the regularization parameter $\tau = \tau_n$ be chosen such that the first and third term in* (3.18) *go to zero and $\xi_n = o(1)$. Further suppose that $\liminf r_n/n > 0$ and $\liminf \beta_n > 0$. Then,*

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\varepsilon}}\, MSE(\boldsymbol{f}^*, \widehat{\boldsymbol{f}}) = 0.$$

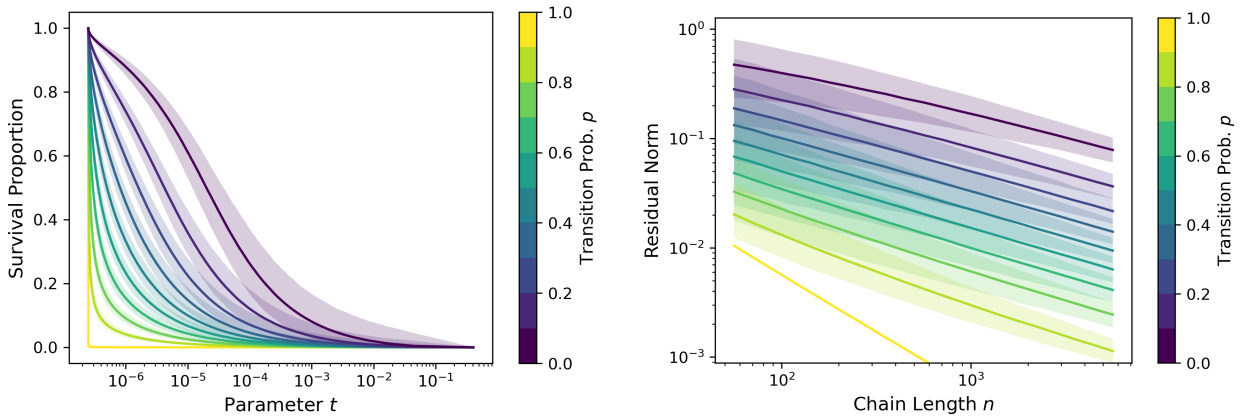
This MSE decay can be made explicit given the knowledge of the population eigenvalue decay. For example, consider the Sobolev-$\alpha$ RKHS, where $\lambda_i \asymp i^{-2\alpha}$. With some additional knowledge of $g^*$ and its spectral tail decay $\beta_n$, an upperbound on the MSE can be obtained for the general misspecified case of (3.18). Rate calculations for the case of a Sobolev-$\alpha$ RKHS are done in Appendix B, the result of which we now present as a corollary:

**Corollary 2.** *Suppose $\mathcal{K}(s,t)$ is the reproducing kernel for the Sobolev-$\alpha$ inner product on $[0,1]$, namely,*

$$\langle f, h\rangle_{\mathbb{H}} = \sum_{k=0}^{\alpha} \int_0^1 f^{(k)}(x) h^{(k)}(x)\, dx.$$

*If $g^*$ satisfies spectral tail decay $\beta_n$, then there is a selection of $\tau_n$ such that $\xi_n = n^{\frac{-2\alpha}{2\alpha+1}}$, and*

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\, MSE(\tau_n) \lesssim n^{\frac{-2\alpha}{2\alpha+1}(\beta_n \wedge 2)}.$$




(a) Survival functions of $\breve{\boldsymbol{g}}$ for length $n = 1000$.

(b) Residual norm $\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2$ for $\tau_n = \sqrt{\lambda_n}$.

Figure 3.4: Experiment results for the 2-state, $p$-probability Markov chain. 10000 chains were simulated for each $p \in \{k/10\}_{k=1}^{10}$. Shown in subfigures are median results with 95% probability intervals shaded in the corresponding colors. In the case of $p = 1$, there is no shading.

Next, to provide intuition for the spectral tail decay $\beta_n$, we provide an example analyzing the decay of the spectral survival function $S_{g^*}(t)$ in a general two-class signal.

**Example 3.** Consider step signal $\boldsymbol{g}^* \in \{-1, 1\}^n$ generated from an $n$-length, 2-state Markov chain with transition probability $p$. For estimation, we consider the following RKHS:

$$\mathbb{H}^1[0,1] = \big\{ f : [0,1] \to \mathbb{R} \mid \ f \text{ is abs. cts.}, \ \|\partial_x f\|_{L^2} < \infty, \ f(0) = 0 \big\}. \tag{3.25}$$

This RKHS has kernel $\mathcal{K}(x, x') = \min(x, x')$. In this example, we assume the data is sampled at regularly spaced intervals with $x_i = i/n$.

The RKHS $\mathbb{H}^1[0,1]$ organizes functions by roughness through the Hilbert-norm $\|f\|_{\mathbb{H}^1} = \|\partial_x f\|_{L^2}$. Hence, signals $\boldsymbol{g}^*$ produced by chains with high transition probabilities are expected to have a larger corresponding Hilbert-norm and, intuitively, a rapidly decaying spectral survival function. This intuition is corroborated in Figure 3.4, where the survival function $S_{g^*}(t)$ and residual norm $\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2 = \frac{1}{n}\sum_{i=1}^n h(\lambda_i; \tau_n)\, \breve{g}_i^2$ are plotted for various transition probabilities. One observes that as the transition probability increases, the tail decay of the survival function becomes sharper (Figure 3.4a) and the norm decay steeper (Figure 3.4b).

The kernel matrix $K$ in this case has minimum eigenvalue $\lambda_n \approx (4n)^{-1}$. For the regularization choice $\tau_n = \sqrt{\lambda_n}$ shown in Figure 3.4b, the quickest rate of decay guaranteed by Proposition 3—namely, $\mathcal{O}((\lambda_n/\tau_n)^2)$—will be on the order of $\mathcal{O}(n^{-1})$. This rate is attained in the log-log plot of Figure 3.4b where the curve associated with chain transition probability $p = 1$ shows a linear slope of $-1$.

Figure 3.4 also provides evidence that the conditions of Theorem 2 are met for this general signal class. The survival function plots in Figure 3.4a show natural tail decays for all probabilities $p$ at $n = 1000$ and the stable linear decays of Figure 3.4b show that the $\liminf$ conditions on $r_n/n$ and $\beta_n$ are attainable for a general signal model.

Lastly, we provide a complete consistency result for the AltMin algorithm. More precisely, given that the contamination error $\mathrm{MSE}(\boldsymbol{f}^*, \widehat{\boldsymbol{f}})$ goes to zero in $\varepsilon$-expectation, the AltMin returned clustering parameters $(\widehat{\boldsymbol{\mu}}^{(n)}, \widehat{\boldsymbol{z}}^{(n)})$ converge to population parameters of the uncontaminated $k$-means optimization. That is:

**Theorem 3.** *Under the same assumptions of Theorem 2, the minimizer sequence $\{\widehat{\boldsymbol{\mu}}^{(n)}\}_n$ converge, in probability, to the minimizers of the population objective*

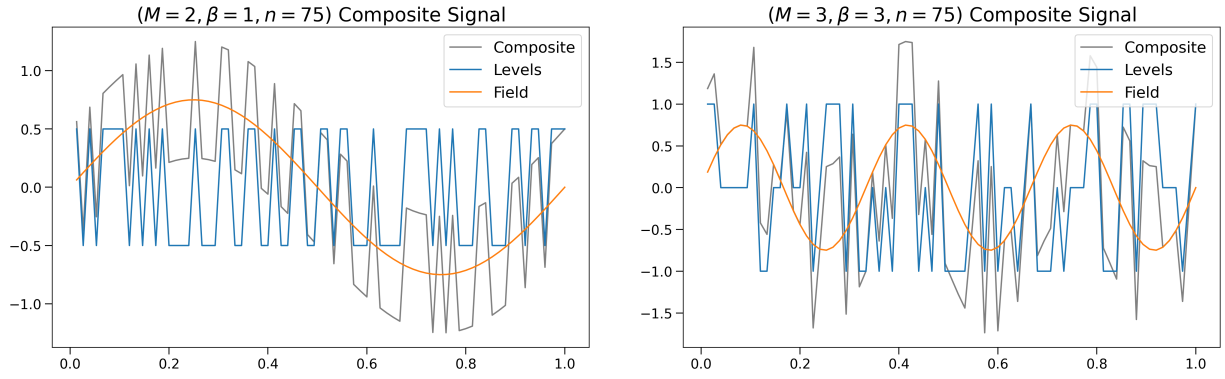
$$L_*(\boldsymbol{\mu}) = \int \min_{k\in[M]} |\mu_k - (g^*(x) + \varepsilon)| \, \mathbb{P}(dx \times d\varepsilon). \tag{3.26}$$

*Likewise, the misclassification rate between estimated labels $\widehat{\boldsymbol{z}}^{(n)}$ and the nearest label assignement of $\{x_i\}_{i=1}^n$ to the minimizers of (3.26) goes to zero in probability.*

A more precise version of this theorem can be found in Appendix D.

## 4 Experiments

We now provide experimental results on the performance of the AltMin algorithm. First, we consider simulated data from an $M$-class data generating process on $\mathcal{X} = [0, 1]$ where data $X^{(n)} = \{i/n\}_{i=1}^n$ is equispaced, cluster labels $z_i^* \sim$ Unif($[M]$) are uniformly distributed,




(a) Smoother field

(b) Rougher field

Figure 3.5: Signal, field and composite observation simulated from (3.27) for two and three classes.

and the step and smooth components follow

$$\mu_k^* = k - \frac{M+1}{2}, \qquad f_\beta^*(x) = \frac{3}{4}\sin(2\pi\beta x). \tag{3.27}$$

The min kernel from Example 3 was chosen for estimation due to its sinusoidal eigenfunctions. Given the equispaced data, the smallest radius $\rho$ that guarantees the connectivity condition of Theorem 1 is

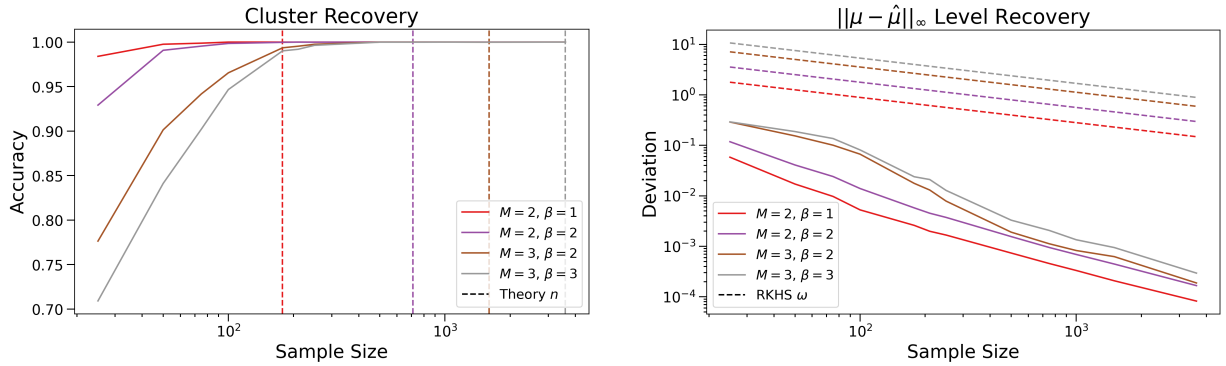
$$\min_{i\neq j} d_{\mathcal{K}}(x_i, x_j) = \sqrt{(i+1)/n - 2i/n + i/n} = n^{-1/2},$$

where the kernel-metric $d_{\mathcal{K}}(x, x')$ was defined in Example 2. The Hilbert-norm of $f_\beta^*$ can be computed using inner product $\langle f, g\rangle_{\mathbb{H}^1} = \int_0^1 \partial_x f(x)\,\partial_x g(x)\,dx$. Evaluating this norm gives the following worst-case bound on the modulus of continuity of $f_\beta^*$,

$$\omega(\rho_{\min}) \leq \|f_\beta^*\|_{\mathbb{H}^1} \cdot \rho_{\min} \leq \frac{3\sqrt{2}}{4}\pi\beta \cdot n^{-1/2}. \tag{3.28}$$

Finally, a noisy recovery setting will be considered where i.i.d. noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is added to mixed observations $f_\beta^*(x_i) + \mu_{z_i^*}^*$.

In both recovery settings, sample size is grown in roughly exponential manner starting from $n = 25$ to $n = 3600$. At each sample size $n$, a total of 100 datasets $(X^{(n)}, \boldsymbol{y})$ were simulated. Accuracy and deviation results at each $n$ were calculated using the mean score of the 100 datasets.




(a) (Noiseless) Classification curves

(b) (Noiseless) Deviation curves

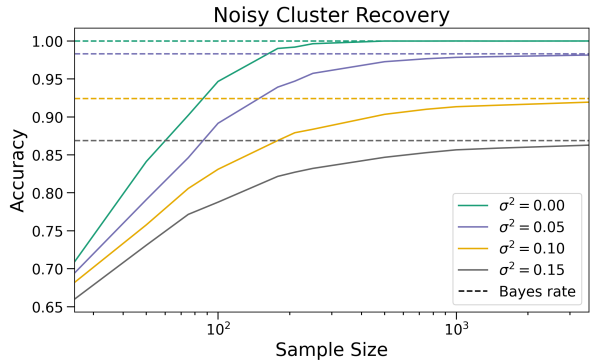
Figure 3.6: AltMin recovery results for a noiseless simulated setting. Worst-case theory bounds are shown as dashed lines for each of the different settings.

### 4.1 Simulation Experiments

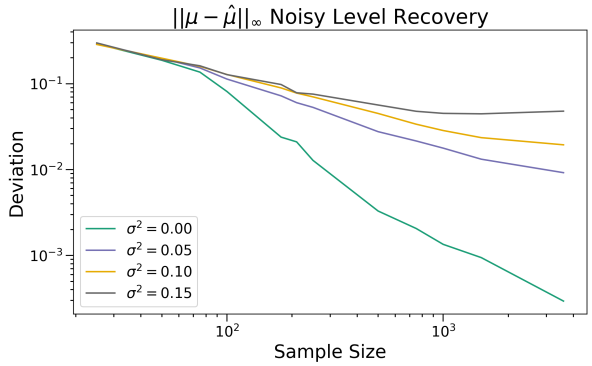
Four settings were considered for noiseless recovery: $(M, \beta) \in \{(2,1),(2,2),(3,2),(3,3)\}$. Cluster recovery and deviation results for the four settings can be found in Figure 3.6. For each setting of the optimization problem (3.8), worst-case recovery bounds, shown dashed in Figure 3.6, were calculated using Theorem 1 and Proposition 1. The AltMin algorithm stays well within these worst-case bounds, demonstrating the effectiveness of the simple blockwise updates for specific problem settings.

For noisy recovery, the setting with $M = \beta = 3$ was considered at noise levels $\sigma^2 \in \{0, 0.05, 0.1, 0.15\}$. Cluster recovery and deviation results for these four settings can be found in Figure 3.7. In each of the noisy settings, AltMin approaches the Bayes error of what is expected for a perfect classifier.

We note that the rate at which AltMin approaches Bayes error seems faster for the cases where $\sigma^2$ is low. This may suggest that the AltMin algorithm is well-suited for smooth field, cluster recovery problems which experience low amounts of background noise.




(a) (Noisy) Classification curves




(b) (Noisy) Deviation curves

Figure 3.7: AltMin recovery results for a noisy simulated setting. Bayes error rates for classification are shown as dashed lines for the various noise levels.

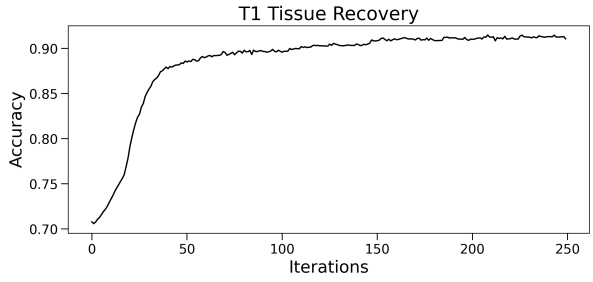
## 4.2 MRI Decontamination

For application, we return to the motivating MRI bias field problem. This is a real-world example where the magnitude of the inhomogeneity $f^*$ and the tissue intensity $g^*$ are much larger than the scale of the background noise $\sigma^2$ [ABW10]. As we have seen in Section 4.1, this is a type of problem which is a good candidate for the AltMin algorithm.

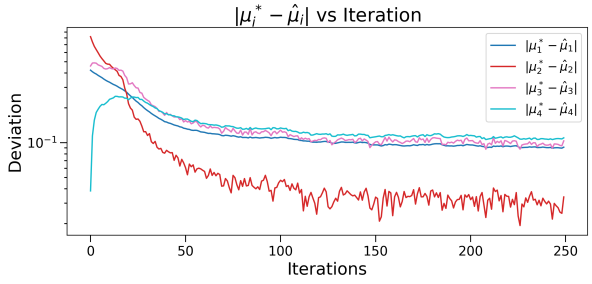
To make our experiment quantitative, we consider a 4-class, strongly-biased variant of the BrainWeb [CKK97] phantom. The field estimation step (3.14) is carried out using a Python spline routine `csaps` [Pri23]. This routine uses an RKHS tensor product of univariate smoothing splines to fit the multidimensional data. Relevant `csaps` smoothing parameters were selected using a post-fitting process. In practice, smoothing parameters would be selected using a validation set of data which corresponds to a specific coil cluster or MRI scanner.

For implementation, we consider modeling the bias field for both single sequence and multi-sequence scans. In a multi-sequence scan, it is understood that the bias field does not vary much between sequences [BMC06]. For this reason, we consider the following general $p$-sequence data model

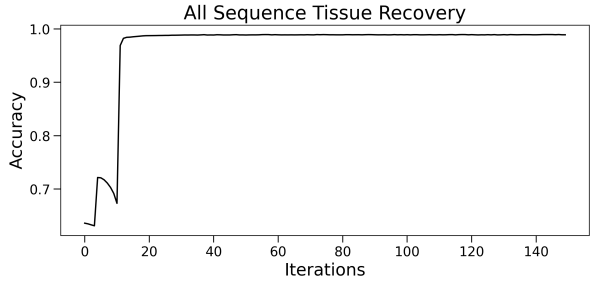
$$\boldsymbol{y}(x) = f^*(x) \cdot \boldsymbol{\mu}^*(x), \quad \text{for } x \in \mathcal{X}$$



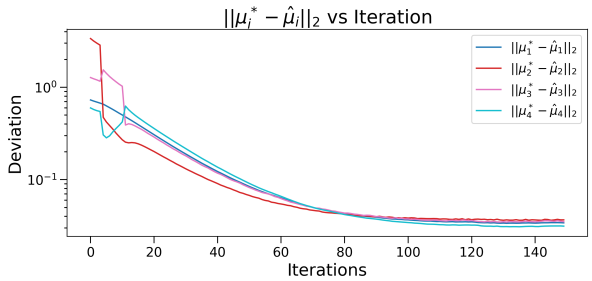

(a) (Single) Accuracy vs. iteration




(b) (Single) Deviation vs. iteration




(c) (Multi) Accuracy vs. iteration

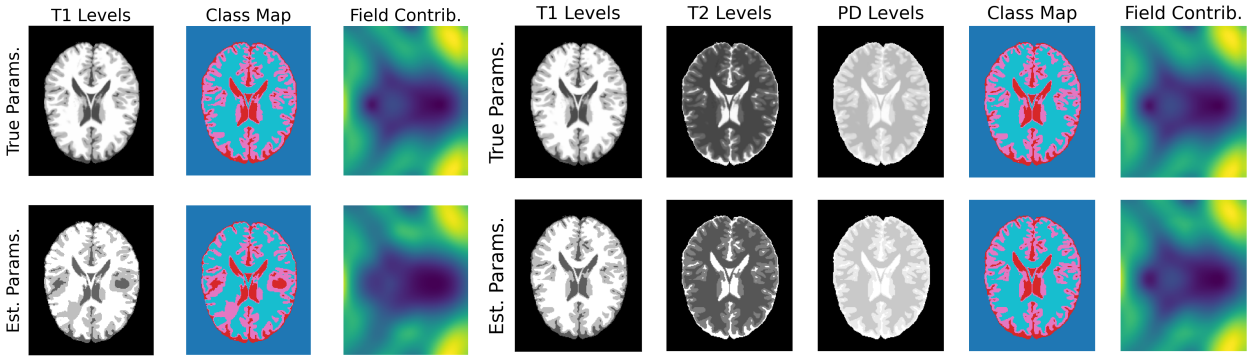



(d) (Multi) Deviation vs. iteration

Figure 3.8: Cluster and level accuracy of the AltMin algorithm on the biased BrainWeb phantom. Final accuracies for single and multi-sequence settings are 91.07% and 98.91% respectively. Level deviations in the multi-sequence setting are calculated with respect to the vector 2-norm.

where levels $\boldsymbol{\mu}^*(x)$ take value in $\mathbb{R}^p$ and bias $f^*(x)$ is still a scalar function.

Bias field and tissue decomposition results for the single and multi-sequence setting can be found in Figure 3.9 with the respective AltMin optimization results found in Figure 3.8. The presence of redundant sequencing data, albeit at different intensity scalings, seems to significantly improve AltMin convergence as shown in Figures 3.8c-3.8d. This also translate to an improved performance, as many of the anomalous tissue patches seen in Figure 3.9a no longer occur in Figure 3.9b. Additional experiments comparing AltMin to other medical debiasing methods can be found in Appendix C.




(a) Single sequence recovery

(b) Multi-sequence recovery

Figure 3.9: AltMin decomposition for the biased BrainWeb dataset. Class maps of the single sequence setting show anomalous tissue patches in areas where the field changes most rapidly.

## 5 Conclusion

In this paper, we defined the problem of composite signal decomposition for continuous contaminants and step-wise signals. We outlined recovery conditions that leverage the local and global topology of the data including: connectivity, minimum true level deviation, and the degree of oscillation of the contaminant. These quantities are natural, and their roles in recovery intuitively clear, allowing for a high-level understanding to be easily derived from our theoretical finding.

Besides identifiability, we developed a practical algorithm AltMin for handling contaminants that reside within an RKHS. This algorithm can be viewed as an extension of both kernel ridge regression (KRR) and $k$-means, with updates to each being performed alternately. MSE bounds for the algorithm were provided in terms of the spectral properties of the data, leading to a "one-step" consistency result in the large sample limit.

We evaluated AltMin empirically on both simulated and real-world data. In the case of simulated data, AltMin operated well within the worst-case theory bounds outlined in Section 2. When the data was further corrupted by noise, AltMin approached the best

possible classification rates for the given data generating process. In the real-world study, we conducted an MRI tissue recovery experiment, illustrating how tensor products of smoothing splines can be employed to estimate contaminant MRI bias fields. Given redundant data on the same bias field, AltMin significantly enhanced clustering performance and overall optimization stability.

These empirical studies, alongside the identifiability theory of Section 2, suggest that step and smooth decompositions are attainable within worst-case optimality guarantees. Regarding application, the alternating optimization of AltMin appears well-suited for data-dense tasks, especially when data is spatially uniform and low in noise. In this context, decomposition problems akin to MRI multi-sequence recovery could be promising avenues for further applications of the AltMin algorithm.

# Appendix

## A Identifiability Proofs

Any optimal candidate solution $(\widehat{f}, \widehat{\mu}, \widehat{z})$ to (3.8) which is fit to data $\{y_i\}$ generated from (3.6) must satisfy

$$f^*(x_i) + \mu^*_{z_i^*} = \widehat{f}(x_i) + \widehat{\mu}_{\widehat{z}_i}, \quad \text{for all } i \in [n]. \tag{3.29}$$

Since $\widehat{f} - f^* \in \mathcal{F}_{2\omega}(\mathcal{X})$, we may instead analyze the discrepancy

$$g(x_i) = \mu^*_{z_i^*} - \widehat{\mu}_{\widehat{z}_i}, \quad \text{for all } i = 1, \ldots, n,$$

for $g \in \mathcal{F}_{2\omega}(\mathcal{X})$. In addition, we will assume that function $g : \mathcal{X} \to \mathcal{Y}$ takes values on a normed vector space $(\mathcal{Y}, \|\cdot\|)$. As a result, the modulus of continuity $\omega$ will be related to the induced norm-metric as $\|g(x) - g(x')\| \leq \omega\big(d(x, x')\big)$.

The following result is the main ingredient in the proof of Theorem 1:

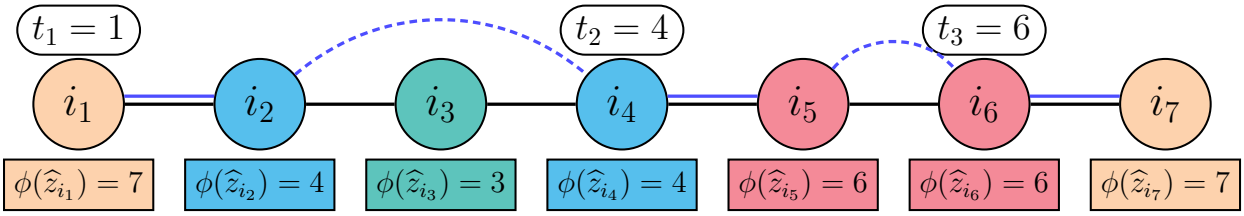



Figure 3.10: exa use of the $\phi(r)$ function and $\{(u_q, v_q)\}_{q=1}^{Q}$ sequence, shown in blue above, for a 4-class path of length 7. The colors denote the estimated cluster labels. This example has $\{(u_q, v_q)\}_{q=1}^{Q} = \{(i_1, i_2), (i_4, i_5), (i_6, i_7)\}$ with $Q = 3$.

**Theorem 4.** *Suppose for $g \in \mathcal{F}_{2\omega}(\mathcal{X})$ we have $g(x_i) = \mu^*_{z^*_i} - \widehat{\mu}_{\widehat{z}_i}$ for all $i \in [n]$ where $\boldsymbol{z}^* = (z^*_i)$ and $\widehat{\boldsymbol{z}} = (\widehat{z}_i)$ both belong to $[M]^n$. Assume the following holds:*

*(a) $\|\mu^*_k - \mu^*_\ell\| \geq \gamma$ for all $k \neq \ell$.*

*(b) $G_\rho(X)$ is connected for some $\rho$ with $2\,\omega(\rho) < \gamma/M$.*

*Then for all $i, j \in [n]$ we have*

$$\widehat{z}_i = \widehat{z}_j \implies z^*_i = z^*_j. \tag{3.30}$$

*Proof.* Start by considering the induction hypothesis that, for any path $\mathcal{P} \subseteq G_\rho(X)$ of length $T$, all element pairs $i, j \in \mathcal{P}$ satisfy (3.30). The base case of $T = 0$ holds trivially with $i = j$.

Throughout the proof, by the label of a node $i$, we mean its estimated label $\widehat{z}_i$. Consider a general path $\mathcal{P} = \{i_t\}_{t=1}^{T+1}$ of length $T + 1$ inside $G_\rho(X)$. As both $\{i_t\}_{t=1}^{T}$ and $\{i_t\}_{t=2}^{T+1}$ are paths of length $T$, we only need to verify (3.30) for $i_1$ and $i_{T+1}$. Therefore, for our induction step it is sufficient to show that $\widehat{z}_{i_1} = \widehat{z}_{i_{T+1}}$ and $z^*_{i_1} \neq z^*_{i_{T+1}}$ cannot simultaneously hold for the given assumptions (a) and (b).

For the sake of contradiction, assume $\widehat{z}_{i_1} = \widehat{z}_{i_{T+1}}$ and $z^*_{i_1} \neq z^*_{i_{T+1}}$. Under this assumption the induction hypothesis guarantees

$$\widehat{z}_{i_t} \neq \widehat{z}_{i_1} \quad \text{for } 1 < t < T + 1. \tag{3.31}$$

Note that if this was not the case with

$$\widehat{z}_{i_1} = \widehat{z}_{i_t} = \widehat{z}_{i_{T+1}} \quad \text{for some } 1 < t < T+1,$$

then the condition $z^*_{i_1} \neq z^*_{i_{T+1}}$ would have caused a contradiction at the earlier induction step $\max\{(T+1)-t, t-1\}$.

Next let $\mathcal{R}$ be the set of labels $\widehat{z}_{i_t}$ on path $\mathcal{P}$. Function $\phi(r)$ will be the index of the last node we see on the path from $i_1$ to $i_{T+1}$ that has label $r$, that is,

$$\phi(r) = \max_{t \in [T+1]} \{t : \ \widehat{z}_{i_t} = r\}.$$

We construct an edge sequence $\{(u_q, v_q)\}_{q=1}^{Q}$—where $Q$ is determined by the construction—recursively as follows: Let $(u_1, v_1) = (i_1, i_2)$ and for $q = 2, \ldots, Q$,

$$(u_q, v_q) = (i_{t_q}, i_{t_q+1}) \quad \text{where} \quad t_q = \phi(\widehat{z}_{v_{q-1}}).$$

The construction continues until $t_Q = T$, so that $(u_Q, v_Q) = (i_T, i_{T+1})$. See Figure 3.10 for a concrete example. By construction, the labels of $v_{q-1}$ and $u_q$ are the same, while the labels of $v_{q-1}$ and $v_q$ are necessarily different. By this latter property, the labels of $v_1, \ldots, v_{Q-1}$ are distinct elements of $\mathcal{R}$. The added uniqueness condition of (3.31) gives that the label of $v_Q$ is also distinct from $v_1, \ldots, v_{Q-1}$, hence $Q \leq |\mathcal{R}|$.

Using $\widehat{z}_{v_{q-1}} = \widehat{z}_{u_q}$, we obtain the decomposition

$$\widehat{\mu}_{\widehat{z}_{u_1}} - \widehat{\mu}_{\widehat{z}_{v_Q}} = \sum_{q=1}^{Q} (\widehat{\mu}_{\widehat{z}_{u_q}} - \widehat{\mu}_{\widehat{z}_{v_q}}). \tag{3.32}$$

From the induction hypothesis, $\widehat{z}_{v_{q-1}} = \widehat{z}_{u_q}$ implies $z^*_{v_{q-1}} = z^*_{u_q}$ for $2 \leq q \leq Q$. This gives the decomposition

$$\mu^*_{z^*_{u_1}} - \mu^*_{z^*_{v_Q}} = \sum_{q=1}^{Q} (\mu^*_{z^*_{u_q}} - \mu^*_{z^*_{v_q}}). \tag{3.33}$$

Moreover, since $u_q$ and $v_q$ are adjacent on the path, they satisfy $d(x_{u_q}, x_{v_q}) \leq \rho$, which by assumption (b) implies

$$\|(\mu^*_{z^*_{u_q}} - \mu^*_{z^*_{v_q}}) - (\widehat{\mu}_{\widehat{z}_{u_q}} - \widehat{\mu}_{\widehat{z}_{v_q}})\| = \|g(x_{u_q}) - g(x_{v_q})\| < \gamma/M. \tag{3.34}$$

By assumption, $\widehat{z}_{u_1} = \widehat{z}_{v_Q}$, hence the LHS of (3.32) is zero. Then, subtracting decomposition (3.32) from (3.33) and using the triangle inequality, we get

$$\|\mu^*_{z^*_{u_1}} - \mu^*_{z^*_{v_Q}}\| \leq \sum_{q=1}^{Q} \|(\mu^*_{z^*_{u_q}} - \mu^*_{z^*_{v_q}}) - (\widehat{\mu}_{\widehat{z}_{u_q}} - \widehat{\mu}_{\widehat{z}_{v_q}})\| < Q\,\gamma/M,$$

where the second inequality is by (3.34). If at the same time $z_{i_1} \neq z_{i_{T+1}}$ then $\mu^*_{z^*_{u_1}} \neq \mu^*_{z^*_{v_Q}}$, and by assumption (a), $\gamma \leq \|\mu^*_{z^*_{u_1}} - \mu^*_{z^*_{v_Q}}\|$. Hence,

$$\gamma < Q\,\gamma/M \leq |\mathcal{R}|\,\gamma/M.$$

Since $|\mathcal{R}| \leq M$, we arrive at a contradiction. This completes the induction step. Applying our induction claim to the connected $G_\rho(X)$ completes the proof. □

Theorem 4 shows that $\widehat{z}$ is a *refinement* of $z^*$. But since $\widehat{z}$ has at most $M$ classes and $z^*$ has exactly $M$ classes—due to being saturated by assumption—the classes of $\widehat{z}$ should, in fact, coincide with those of $z^*$. This proves Theorem 1.

Let us now prove Proposition 1. Under the assumptions of Theorem 1, we can relabel the classes of $(\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\mu}})$ so that $\widehat{\boldsymbol{z}} = \boldsymbol{z}^*$. Then, it follows from (3.29) that

$$\widehat{f}(x_i) - f^*(x_i) = \mu^*_{z^*_i} - \widehat{\mu}_{z^*_i} \quad \text{for all } i \in [n]. \tag{3.35}$$

### A.1 Proof of Proposition 1

For $\delta \geq \delta_{\,\mathrm{lbl}}$, the neighbor graph $G_\delta(\mathcal{C})$ is connected such that every $k, \ell \in [M]$ has a series of edges $\{(x_{i_t}, x_{j_t})\}_{t=1}^{T}$ with $d(x_{i_t}, x_{j_t}) \leq \delta$ such that $z^*_{i_1} = k$, $z^*_{j_T} = \ell$ and

$$z^*_{j_{t-1}} = z^*_{i_t} \neq z^*_{j_t}.$$

In particular, the condition $z^*_{i_t} \neq z^*_{j_t}$ ensures $T \leq M - 1$. Let $\delta = \delta_{\,\mathrm{lbl}}$ and with the shorthands $g = \widehat{f} - f^*$ and $\Delta_k = \mu^*_k - \widehat{\mu}_k$, we have $g(x_i) = \Delta_{z^*_i}$ for all $i \in [n]$. Then, the following

inequality holds for all $k, \ell \in [M]$,

$$\begin{aligned}\|\Delta_k - \Delta_\ell\| &\leq \sum_{t=1}^{T} \|g(x_{i_t}) - g(x_{j_t})\| \\ &\leq T \cdot 2\omega(\delta_{\text{ lbl}}) \leq 2(M-1) \cdot \omega(\delta_{\text{ lbl}}).\end{aligned}$$

Letting $\pi_k = |\mathcal{C}_k|/n$ be the proportion of class $k$, then

$$\begin{aligned}\left\|\Delta_k - \frac{1}{n}\sum_{i=1}^{n} g(x_i)\right\| &= \left\|\Delta_k - \sum_{\ell=1}^{M} \pi_\ell \Delta_\ell\right\| \\ &\leq \sum_{\ell=1}^{M} \pi_\ell \|\Delta_k - \Delta_\ell\|.\end{aligned}$$

Since $\widehat{f}$ is assumed zero-mean, $\frac{1}{n}\|\sum_{i=1}^{n} g(x_i)\| = \frac{1}{n}\|\sum_{i=1}^{n} f^*(x_i)\|$. Putting the pieces together, using the triangle inequality and noting that $\sum_\ell \pi_\ell = 1$ finishes the proof.

# B Supplement to Section 3.1

## B.1 Proof of Proposition 2

Since $K(\cdot, \cdot)$ is continuous, all the functions in the RKHS are also continuous with respect to the metric topology of $\mathcal{X}$. Moreover, by the definition of the RKHS, the evaluation functional $\delta_x$, given by $\delta_x f = f(x)$ for any $f \in \mathbb{H}$, is a continuous linear functional on $\mathbb{H}$ relative to $\|\cdot\|_{\mathbb{H}}$ for any $x \in \mathcal{X}$.

We prove the result by contradiction. Assume that $\|\bar{f}_n\|_{\mathbb{H}}$ does not converge to $\infty$. Then, there is a subsequence of $\bar{f}_n$ that is bounded in $\mathbb{H}$. Without loss of generality, let us pass to this subsequence for simplicity. Hence, we have $\|\bar{f}_n\|_{\mathbb{H}} \leq C$ for some $C > 0$ and all $n \geq 1$. Since $\mathbb{H}$ is a Hilbert space, the closed ball $\{f \in \mathbb{H} : \ \|f\|_{\mathbb{H}} \leq C\}$ is weakly compact. This follows from Kakutani's theorem: in a Banach space, the closed unit ball is weakly compact if and only if the Banach space is reflexive. Thus, there is a subsequence $\{\bar{f}_{n_k}\}_{k \geq 1}$ that weakly converges to some $f \in \mathbb{H}$. In particular, $\delta_x \bar{f}_{n_k} \to \delta_x f$, that is $\bar{f}_{n_k}(x) \to f(x)$ for all $x \in \mathbb{H}$.

This implies that for any $i \geq 1$, $\bar{f}_{n_k}(x_i) \to f(x_i)$, and since $g^*(x_i) = \bar{f}_{n_k}(x_i)$, by the definition of the interpolant, it follows that $g^*(x_i) = f(x_i)$ for all $i \geq 1$.

Consider the case where $g^*$ is continuous with respect to the metric topology of $\mathcal{X}$. Since $\{x_i\}_{i\geq 1}$ is a dense subset of the Hausdorff space $(\mathcal{X}, d)$ and since $f$ is continuous, it follows that $g^*(x) = f(x)$ for all $x \in \mathcal{X}$. But this is a contradiction since $g^* \notin \mathbb{H}$ and $f \in \mathbb{H}$.

On the other hand, if $g^*$ is discontinuous at a point $x_0 \in \mathcal{X}$, we can find two subsequences of $\{x_i\}$ converging to $x_0$, along which $g^*$ converges to two different values. But since $f$ matches $g^*$ on $\{x_i\}$, it means that $f$ converges to different values along those same subsequences. This contradict continuity of $f$ at $x_0$. The proof is complete.

### B.2 Proof of Proposition 3

Let $\boldsymbol{\lambda}$ be the discrete random variable defined by

$$\boldsymbol{\lambda} = \begin{cases} \lambda_i, & \text{w.p. } \breve{g}_i^2/(n\|\boldsymbol{g}^*\|_\infty^2) \\ 0, & \text{w.p. } 1 - \|\boldsymbol{g}^*\|_2^2/(n\|\boldsymbol{g}^*\|_\infty^2). \end{cases} \tag{3.36}$$

Further define $\psi(\lambda) = (1 + \tau/\lambda)^{-2}$, then

$$\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2 = \sum_{i=1}^n \frac{\breve{g}_i^2}{n}\psi(\lambda_i) \leq \|\boldsymbol{g}^*\|_\infty^2 \, \mathbb{E}[\psi(\boldsymbol{\lambda})].$$

Define $r$ and $\beta$ as before. Function $\psi(\cdot)$ is non-negative and monotone on $[0, \infty)$ so

$$\begin{aligned} \mathbb{E}[\psi(\boldsymbol{\lambda})] &= \int_0^\infty \Pr(\psi(\boldsymbol{\lambda}) > t)\, dt \\ &= \psi(\lambda_r) + \int_{\psi(\lambda_r)}^{\psi(\lambda_1)} \Pr(\boldsymbol{\lambda} > \psi^{-1}(t)) dt \\ &\leq \psi(\lambda_r) + \int_{\psi(\lambda_r)}^{\psi(\lambda_1)} \left(\frac{\lambda_r}{\psi^{-1}(t)}\right)^\beta dt \end{aligned} \tag{3.37}$$

Denote the last right-hand side integral as $\mathcal{I}(\beta)$. Integral $\mathcal{I}(\beta)$ is monotone decreasing with $\mathcal{I}(\beta) < \mathcal{I}(\beta')$ for $0 \leq \beta' < \beta$. Next, the inverse function $\psi^{-1}$ can be lower bounded as

$$\psi^{-1}(t) = \frac{\tau}{t^{-1/2} - 1} \geq \tau\, t^{1/2}(1-t)^{-1/2}. \tag{3.38}$$

Restricting focus to $\beta \in [0,2)$ and applying (3.38) to integral $\mathcal{I}(\beta)$ yields

$$\begin{aligned}\mathcal{I}(\beta) &\le (\lambda_r/\tau)^\beta \int_{\psi(\lambda_r)}^{\psi(\lambda_1)} t^{-\beta/2}(1-t)^{\beta/2}\,dt \\ &\le (\lambda_r/\tau)^\beta \int_0^1 t^{-\beta/2}(1-t)^{\beta/2}\,dt \\ &= (\lambda_r/\tau)^\beta \,\frac{\Gamma(1-\beta/2)\,\Gamma(1+\beta/2)}{\Gamma(2)}.\end{aligned}$$

Identity $\Gamma(z) = \Gamma(1+z)/z$ can be used with $\beta \in [0,2)$ to get

$$\frac{\Gamma(1-\beta/2)\,\Gamma(1+\beta/2)}{\Gamma(2)} \le \frac{2}{2-\beta}.$$

Lastly since $\psi(\lambda) \le (\lambda/\tau)^2$ and $\mathcal{I}(\beta)$ is monotone decreasing in $\beta$, we have

$$\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2 \le 2\|\boldsymbol{g}^*\|_\infty^2 \cdot \max\left\{(\lambda_r/\tau)^2,\ \inf_{\beta' \in [0,\beta)} \frac{2}{2-\beta'} \cdot (\lambda_r/\tau)^{\beta'}\right\}. \tag{3.39}$$

Let $\xi \coloneqq \lambda_r/\tau$. For $\xi < 1$, function $h(x) = \xi^x/(2-x)$ achieves global minimum at $x^* = 2 - (\log \frac{1}{\xi})^{-1}$. This minimum is non-negative for $\xi < e^{-1/2}$. That is, when $\beta = 2$, we have $\beta'$ approaching 2 as $\lambda_r/\tau$ approaches 0. More specifically, we obtain the following simplification to (3.39)

$$\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2 \le 4\|\boldsymbol{g}^*\|_\infty^2 \log(1/\xi)\,\xi^{2-(\log\frac{1}{\xi})^{-1}}. \tag{3.40}$$

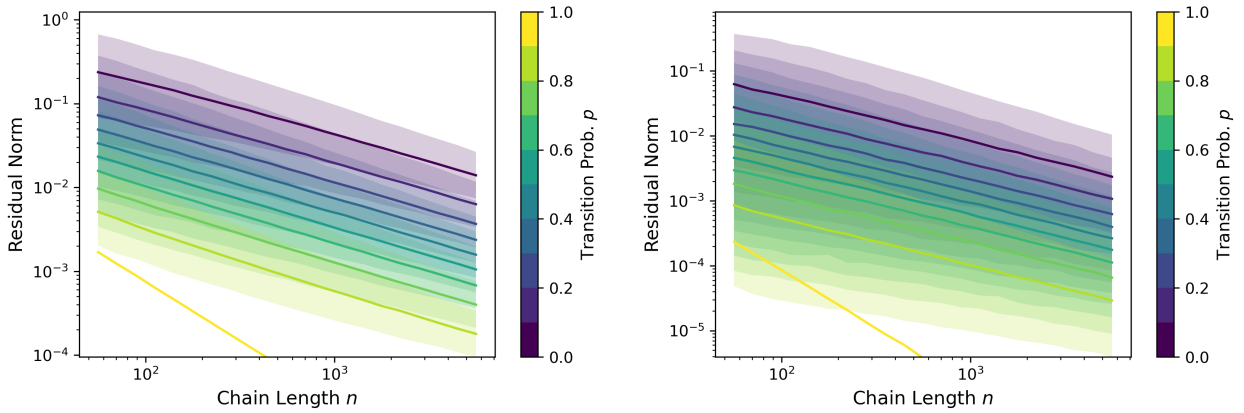
### B.3 Sobolev Kernel Rates

The Sobolev-$\alpha$ RKHS on $[0,1]$ has kernel defined by inner product

$$\langle f, g\rangle_{\mathbb{H}^\alpha} = \sum_{k=0}^{\alpha} \int_0^1 f^{(k)}(x)\, g^{(k)}(x)\,dx.$$

This Mercer kernel has eigenvalue decay $\lambda_i = i^{-2\alpha}$. For the standard KRR problem, a minimax optimal selection of the regularization parameter is given by $\tau_n \asymp n^{\frac{-2\alpha}{2\alpha+1}}$. When plugged into the MSE expression,

$$\overline{\mathrm{MSE}}(\tau) = \frac{1}{n}\sum_{i=1}^n \frac{\tau^2 \breve{f}_i^2}{(\lambda_i+\tau)^2} + \frac{\sigma^2}{n}\sum_{i=1}^n \frac{\lambda_i^2}{(\lambda_i+\tau)^2}, \tag{3.41}$$




(a) Sobolev-1 with $\tau = n^{-2/3}$.

(b) Sobolev-2 with $\tau = n^{-4/5}$.

Figure 3.11: Residual norm decays $\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2$, based on the optimal $\tau_n$ selections, for different Sobolev-$\alpha$ kernels. Slight numerical inaccuracies are shown in the norm decay of the Sobolev-2 kernel.

we obtain $\overline{\text{MSE}}(\tau_n) \asymp n^{\frac{-2\alpha}{2\alpha+1}}$ which decays to zero as $n \to \infty$. The resulting rate for $\xi_n = \lambda_n/\tau_n$ is

$$\xi_n \asymp n^{\frac{-4\alpha^2}{2\alpha+1}} = o(1)$$

which satisfies the condition of Corollary 2. Tying back to Example 3, Figure 3.11 shows norm decay plots for $\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2$, when using $\tau_n$ minimax selections on the different Sobolev-$\alpha$ kernels.

Similar to the case of the min-kernel in Figure 3.4, as the signal $\boldsymbol{g}^*$ becomes more rough, i.e. $p$ becomes larger, we see a quicker decay in filtered norm for the different Sobolev examples. Furthermore, these contributions are filtered at a faster rate for Sobolev kernels that are smoother, that is those with larger $\alpha$ values. This faster decay is not only intuitive but expected from our earlier derived $\xi_n$ decay rate.

| Method | # Seqs. | BrainWeb 4-class | | BrainWeb 10-class | |
|---|---|---|---|---|---|
| | | Acc. [%] | Max Dev. [1] | Acc. [%] | Max Dev. [1] |
| k-means | 1 | 73.62 | $8.21 \times 10^{-1}$ | 37.69 | $7.08 \times 10^{0}$ |
| | 3 | 74.38 | $3.41 \times 10^{0}$ | 44.21 | $1.19 \times 10^{1}$ |
| N4ITK + k-means | 1 | 74.10 | $1.17 \times 10^{0}$ | 40.14 | $6.01 \times 10^{0}$ |
| | 3 | 74.41 | $3.91 \times 10^{0}$ | 48.22 | $1.11 \times 10^{1}$ |
| LapGM | 1 | 76.14 | $2.87 \times 10^{0}$ | 50.16 | $\mathbf{2.47 \times 10^{0}}$ |
| | 3 | 87.28 | $4.08 \times 10^{0}$ | 78.38 | $\mathbf{4.19 \times 10^{0}}$ |
| AltMin | 1 | **91.07** | $\mathbf{1.10 \times 10^{-1}}$ | **56.27** | $4.49 \times 10^{0}$ |
| | 3 | **98.91** | $\mathbf{3.67 \times 10^{-2}}$ | **82.36** | $7.84 \times 10^{0}$ |

Table 3.1: Clustering results for different debiasing methods for single and multi-sequence settings.

## C Additional Experiments

We compare AltMin to other MRI debiasing techniques using the same biased phantom as Section 4.2. For comparison, we consider a standard debiasing technique N4ITK [TAC10] and a Bayesian modeling approach LapGM [VS22]. Hyperparameters for all methods, including AltMin, were selected using the same post-fitting process. Specific to N4ITK, bias estimates were calculated on the T1-sequence information and clusterings were calculated using an additional $k$-means estimation at the end of the debiasing procedure.

Performance of each method for the various recovery settings can be found in Table 3.1. In each recovery setting, AltMin either meets or exceeds the classification and level accuracies of the other tested methods. We highlight that, for all debias methods, recovery is significantly more difficult in the 10-class setting. Methods which eventually scored well in this setting were those which could effectively leverage multi-sequence information during debias and clustering. This emphasizes the importance of replicated information for practical step and

smooth recovery implementations.

# D AltMin Consistency Through Γ-Convergence Techniques

The AltMin algorithm is a blockwise optimization procedure that iteratively estimates parameter intermediaries through kernel ridge regression (KRR) and $k$-means update steps. Under the appropriate assumptions, each update step can be shown to be consistent with respect to their own parameter subset. However, a question remains of whether AltMin can achieve consistency in case of perturbed optimization steps.

In-sample consistency for the perturbed KRR step was already shown in Section 3.1 for samples with $x$-variates belonging to the set

$$\mathcal{B} \coloneqq \big\{(x_i)_{i=1}^{\infty} : \liminf_{n\to\infty} r_n/n > 0 \text{ and } \liminf_{n\to\infty} \beta_n > 0\big\}. \tag{3.42}$$

When it is clear from context we will let $\mathcal{B} \coloneqq \mathcal{B} \times \mathbb{R}^{\infty}$, since event $\mathcal{B}$ does not depend on noise $\varepsilon$. Let $\mathrm{MSE}_n : ((x_i)_{i=1}^{\infty}, (\varepsilon_i)_{i=1}^{\infty}) \mapsto \mathrm{MSE}(\boldsymbol{f}^*, \widehat{\boldsymbol{f}})$ denote the $n$-sample loss for finite sequence $(x_i, \varepsilon_i)_{i=1}^{n}$. Then, through Markov's inequality

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\big(\{\mathrm{MSE}_n(x,\varepsilon) \geq \delta\} \cap \mathcal{B}\big) &\leq \delta^{-1} \lim_{n\to\infty} \mathbb{E}_{(x,\varepsilon)}[\mathrm{MSE}_n(x,\varepsilon) \cdot 1_{\mathcal{B}}(x)] \\
&\leq \delta^{-1} \lim_{n\to\infty} \mathbb{E}_x[\mathbb{E}_{\boldsymbol{\varepsilon}}[\mathrm{MSE}(\boldsymbol{f}^*, \widehat{\boldsymbol{f}})] \cdot 1_{\mathcal{B}}(x)] \\
&\lesssim \delta^{-1} \lim_{n\to\infty} \mathbb{E}_x[e_n(x)]
\end{aligned}$$

where $e_n(x)$ is an error term which is almost-surely bounded and tending to zero. Then by dominated convergence theorem, for every $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(\{\mathrm{MSE}_n(x,\varepsilon) \geq \delta\} \cap \mathcal{B}\big) = 0 \tag{3.43}$$

Similarly, we will show that the AltMin clustering parameters, $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{z}})$, converge to population minima in probability. More rigorously, let $L_* : \mathbb{R}^M \to \mathbb{R}$ be the population $k$-means clustering objective defined by

$$L_*(\boldsymbol{\mu}) = \int \min_{k\in[M]} |\mu_k - (g^*(x) + \varepsilon)|^2 \, \mathbb{P}(dx \times d\varepsilon). \tag{3.44}$$

The minima of $L_*$, defined formally as the following set,

$$\mathcal{M} \coloneqq \{\overline{\boldsymbol{\mu}} \in \mathbb{R}^M : L_*(\overline{\boldsymbol{\mu}}) = \inf_{\boldsymbol{\mu} \in \mathbb{R}^M} L_*(\boldsymbol{\mu})\} \tag{3.45}$$

each admit a nearest label partition $\mathcal{V} = (\mathcal{V}_k)_{k=1}^M$, where

$$\mathcal{V}_k \coloneqq \{u \in \mathbb{R} : \operatorname*{argmin}_{\ell \in [M]} |\overline{\mu}_\ell - u| = k\} \tag{3.46}$$

and a nearest label sequence $\overline{z}_i = \operatorname{argmin}_{k \in [M]} |\overline{\mu}_k - (g^*(x_i) + \varepsilon_i)|$ for data pairs $(x_i, \varepsilon_i)$.

Individual deviations from the minima set $\mathcal{M} \subseteq (\mathbb{R}^M, \|\cdot\|)$ can be calculated by

$$\operatorname{Dist}(\boldsymbol{\mu}, \mathcal{M}) \coloneqq \inf_{\overline{\boldsymbol{\mu}} \in \mathcal{M}} \|\boldsymbol{\mu} - \overline{\boldsymbol{\mu}}\| \tag{3.47}$$

and sample misclassification relative to a sequence $\overline{z}$ can be calculated as

$$\operatorname{Miss}_n(\boldsymbol{z}, \overline{z}) = \frac{1}{n} \sum_{i=1}^n 1\{z_i = \overline{z}_i\} \tag{3.48}$$

This brings us to the following consistency result for the AltMin clustering step:

**Theorem 5.** *Let $\widehat{f}_n$ be the KRR estimate for $(y_i)_{i=1}^n$ and let $\widehat{\boldsymbol{\mu}}_n$ be a minimizer for*

$$\widetilde{L}_n(\boldsymbol{\mu}) = \frac{1}{n} \sum_{i=1}^n \min_{k \in [M]} |\mu_k + \widehat{f}_n(x_i) - y_i|^2$$

*with corresponding label estimates $\widehat{\boldsymbol{z}}_n \in [M]^n$. Suppose $(x_i, \varepsilon_i) \overset{i.i.d.}{\sim} \mathbb{P}$. If all $\mu_k^*$ are distinct and $(x_i)_{i=1}^\infty \in \mathcal{B}$, then the minimizer sequence $\{\widehat{\boldsymbol{\mu}}_n\}_n$ converges in probability in the sense*

$$\lim_{n \to \infty} \mathbb{P}(\{\ \mathit{Dist}(\widehat{\boldsymbol{\mu}}_n, \mathcal{M}) > \delta)\} \cap \mathcal{B}) = 0 \quad \textit{for every } \delta > 0.$$

*Furthermore, each converging subsequence $\widehat{\boldsymbol{\mu}}_m \to \overline{\boldsymbol{\mu}} \in \mathcal{M}$ has*

$$\lim_{m \to \infty} \ \mathit{Miss}_m(\widehat{\boldsymbol{z}}_m, \overline{z}) \to 0,$$

*where $\overline{z} = (\overline{z}_i)_{i=1}^\infty$ is the label sequence associated to $\overline{\boldsymbol{\mu}}$ and $(x_i, \varepsilon_i)_{i=1}^\infty$.*

*That is to say, under event $\mathcal{B}$,* AltMin *consistently recovers cluster and nearest label estimates for the uncontaminated $k$-means clustering problem $\{g(x_n) + \varepsilon_n\}_n$.*

Elaborating on the convergence shown in Theorem 5, a deviation quantity $\Delta_n \in \mathbb{R}$ converges in probability for $\mathcal{B}$ if

$$\lim_{n\to\infty} \mathbb{P}(\{\Delta_n > \delta\} \cap \mathcal{B}) = 0 \quad \text{for every } \delta > 0.$$

Equivalently stated, every subsequence of $\{\Delta_n\}_n$ has a sub-subsquence $\{\Delta_m\}_m$ such that $\Delta_m \to 0$ almost-surely in $\mathcal{B}$, that is,

$$\mathbb{P}\big(\big\{ \lim_{m\to\infty} \Delta_m = 0\big\} \cap \mathcal{B}\big) = \mathbb{P}(\mathcal{B}).$$

Armed with these definitions, we begin by fixing any countable, subsequence index set $\mathcal{I} \subseteq \mathbb{N}$. For brevity, we simply refer to $\mathcal{I}$ as an index set. Then, by (3.43), there exists a sub-index set $\mathcal{I}' \subseteq \mathcal{I}$ such that $\mathrm{MSE}_m(x, \varepsilon) \to 0$ for $m \in \mathcal{I}'$ and $\mathbb{P}$-almost all $(x_i, \varepsilon_i)_{i=1}^{\infty} \in \mathcal{B}$. Without loss of generality let $\mathcal{I}' = \mathcal{I} = [n]$, it suffices to show

$$\lim_{m\to\infty} \mathrm{Dist}(\widehat{\boldsymbol{\mu}}_m, \mathcal{M}) = 0 \quad \text{for } \mathbb{P}\text{-a.e. } (x_i, \varepsilon_i)_{i=1}^{\infty} \in \mathcal{B}. \tag{3.49}$$

for $m$ belonging to some sub-index set $\mathcal{I}'' \subseteq \mathcal{I}'$.

Our convergence result will be proven using techniques belonging to $\Gamma$-convergence. To this end, we begin with an overview $\Gamma$-convergence and how it pertains to the $k$-means objective and its minimizers. After this introduction, we show how the AltMin objective maintains key properties of the $k$-means objective plus a bracketing from the negligible $\mathrm{MSE}_n$ term. Lastly, we show how the empirical minimizers of the AltMin's $k$-means update must converge to population minimizers of the uncontaminated $k$-means objective.

### D.1 Convergence of Empirical Minimizers

In this section, we cover the basics of $\Gamma$-convergence and how it pertains to the convergence of empirical minimizers. We refer the reader to [Bra02] for an overview of $\Gamma$-convergence and its applications.

For functionals $F_n : \mathcal{U} \to \mathbb{R}$ defined on a common metric space $\mathcal{U}$, the $\Gamma$-convergence of a functional sequence $\{F_n\}_n$ admits the following characterization:

**Definition 5.** A sequence of functionals $\{F_n\}_n$ is said to $\Gamma$-converge to $\Gamma$-limit $F_*$ if

$$F_*(u) \leq \liminf_{n\to\infty} F_n(u_n) \qquad \text{for every } u_n \to u,$$

and there exists at least one $u_n \to u$ such that

$$\limsup_{n\to\infty} F_n(u_n) \leq F_*(u).$$

To properly characterize the behavior of minimizers for $\{F_n\}_n$, some additional regularity is needed regarding the optimization behavior of $F_n$. In particular, we will be interested in functional sequences that are *equi-mildly coercive*.

**Definition 6.** A sequence of functionals $\{F_n\}_n$ defined on a common metric space $\mathcal{U}$ is equi-mildly coercive if there exists a non-empty, compact subset $\mathcal{A} \subset \mathcal{U}$ such that

$$\inf_{u\in\mathcal{U}} F_n(u) = \inf_{u\in\mathcal{A}} F_n(u) \qquad \text{for all } n.$$

Equipped with these definitions, we can state the first result for convergence of minimizers.

**Theorem 6** ([Bra02], Theorem 1.21)**.** *Let $\{F_n\}_n$ be a sequence of equi-mildly coercive functionals on metric space $\mathcal{U}$ with $\Gamma$-limit $F_*$. Then $\min_{u\in\mathcal{U}} F_*(u)$ exists and equals $\lim_{n\to\infty} \min_{u\in\mathcal{U}} F_n(u)$. Furthermore, if the $F_n$-minimizer sequence $\{\widehat{u}_n\}_n$ is precompact, then all subsequence limits of $\{\widehat{u}_n\}_n$ are minimizers in $F_*$.*

Bounded sequences are necessarily precompact for any metric space $\mathcal{U}$ over the reals.

A stronger version of the coercivity condition for $F_n$ can be stated where, for some compact $\mathcal{A} \subset \mathcal{U}$, the sequence $\{F_n\}_n$ satisfies

$$\inf_{u\in\mathcal{A}} F_n(u) < \inf_{u\in\mathcal{U}\setminus\mathcal{A}} F_n(u) \qquad \text{for all } n.$$

Under this modified coercivity condition, the sequence of $F_n$-minimizers $\{\widehat{u}_n\}_n$ is bounded and, as such, must contain a convergent subsequence of minimizers for real-valued $\mathcal{U}$.

## D.2 $k$-means Consistency

In this section, we present the $k$-means consistency result for a specific normed space $\mathcal{U} = (\mathbb{R}, |\cdot|)$. Proofs for the following results and their subsequent generalizations can be found in [TTJ15].

The $k$-means algorithm takes a sample $(u_1, u_2, \ldots, u_n) \in \mathbb{R}^n$ and optimizes for the closest centers $\widehat{\boldsymbol{\mu}} \coloneqq (\widehat{\mu}_k)_{k=1}^M$ according to the distance based objective

$$L_n(\boldsymbol{\mu}) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \min_{k \in [M]} |\mu_k - u_i|_2^2. \tag{3.50}$$

The population analog of (3.50) is defined with respect to the probability law $\mathbb{P}$ under which marginal $u_1$ is sampled, that is,

$$L_*(\boldsymbol{\mu}) \coloneqq \int \min_{k \in [M]} |\mu_k - u|_2^2 \, \mathbb{P}(du). \tag{3.51}$$

The $k$-means optimization (3.50) is considered *consistent* if the sequence of empirical minimizers $\{\widehat{\boldsymbol{\mu}}_n\}_n$ of $\{L_n\}_n$ has a limit (or a subsequence limit) which minimizes (3.51).

Going forward, sample outcomes will be denoted as $\omega \coloneqq (u_i)_{i=1}^{\infty}$. Similarly, sample quantities like (3.50) which depend on $\omega$ will be made more specific using the notation $L_n^{(\omega)}$.

The first result of [TTJ15] is a $\mathbb{P}$-almost everywhere equivalence for the $\Gamma$-limit of $L_n$. For the specific normed space $\mathcal{U} = (\mathbb{R}, |\cdot|)$, no additional assumptions are needed.

**Theorem 7** ([TTJ15], Theorem 3.2)**.** *Let $L_n$ and $L_*$ be defined as they are in* (3.50) *and* (3.51)*. Further let each $u_i$ be independently drawn from $\mathbb{P}$. Then, for $\mathbb{P}$-almost every $\omega$, $L_*$ is the $\Gamma$-limit of $\{L_n^{(\omega)}\}_n$.*

The statement "$\mathbb{P}$-almost every $\omega$" refers to the law induced by $\mathbb{P}$ for $\omega$, that is $\mathbb{P}^{\infty} \coloneqq \prod_{i=1}^{\infty} \mathbb{P}$.

Equi-mild coercivity of objective for $\{L_n\}_n$ can be shown if optimizing with fewer than $M$ distinct centers is suboptimal for the population objective $L_*$. This can be guaranteed with the following support assumption for $\mathbb{P}$:

**Assumption 1.** *There exists $M$ distinct points $\mu_1, \mu_2, \ldots, \mu_M \in \mathcal{U}$ such that, for all $R > 0$,*

$$\min_{k \in [M]} \mathbb{P}(\{u : |\mu_k - u| < R\}) > 0.$$

In the case $\mathbb{P}$ is a mixture distribution, it suffices for $\mathbb{P}$ to have at least $M$ relative modes (e.g. for $M = 2$, a bimodal distributions with two distinct centers).

Next, provided by [TTJ15], is a coercivity result for the $k$-means algorithm. In fact, this claim is stronger than the usual equi-mild coercivity.

**Proposition 4** ([TTJ15] Proposition 3.3)**.** *Under Assumption 1 and there exists a $\delta > 0$ such that, for $\mathbb{P}$-almost every $\omega$, there is a $R > 0$ where*

$$\inf_{\|\boldsymbol{\mu}\| \leq R} L_n^{(\omega)}(\boldsymbol{\mu}) \leq \inf_{\|\boldsymbol{\mu}\| > R} L_n^{(\omega)}(\boldsymbol{\mu}) - \delta \tag{3.52}$$

*holds for all sufficiently large $n$.*

Proposition 4 can also be extended to all $\epsilon_n$-almost minimizers of $L_n^{(\omega)}$ [Lem03, Lemma 2.1].

Finally, by the $\Gamma$-convergence theorem of minimizers (Theorem 6), we achieve the following consistency result for the $k$-means algorithm:

**Corollary 3.** *Let $\{\widehat{\boldsymbol{\mu}}_n\}_n$ be a sequence of minimizers for $\{L_n\}_n$. Then, under Assumption 1, any limit or subsequence limit of $\{\widehat{\boldsymbol{\mu}}_n\}_n$ must be a minimizer for $L_*$, $\mathbb{P}$-almost surely.*

Under this result, it is valid for $\{\widehat{\boldsymbol{\mu}}_n\}_n$ to alternate between different empirical minimizers. Corollary 3 stipulates that each of these empirical minimizers is also a minimizer for $L_*$.

### D.3 Cluster Consistency for Contaminated Objectives

Recall that a set of samples $\{y_i\}_{i=1}^n \subset \mathbb{R}$ is step-and-smooth on $\mathcal{X}$ if

$$y_i = f(x_i) + u_i = f(x_i) + g(x_i) + \varepsilon_i$$

for continuous $f$ and bounded $g$. For the KRR estimate $\widehat{f}_n$ of $f$, the new perturbed optimization of interest is

$$\widetilde{L}_n(\boldsymbol{\mu}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \min_{k\in[M]} |\mu_k + \widehat{f}_n(x_i) - y_i|_2^2. \tag{3.53}$$

Define $\mathrm{MSE}_n^{(\omega)} \coloneqq \frac{1}{n}\sum_{i=1}^n |f(x_i) - \widehat{f}(x_i)|^2$ where $\omega_i = (x_i, \varepsilon_i)$. It can be shown that the perturbed objective (3.53) follows a triangle inequality. More generally, consider minimization over a general set $\mathcal{A}$, then

$$\begin{aligned}
\sum_{i=1}^{n} \min_{a\in\mathcal{A}} \|a + b_i + c_i\|^2 &\le \sum_{i=1}^{n} \min_{a\in\mathcal{A}} (\|a + b_i\| + \|c_i\|)^2 \\
&= \sum_{i=1}^{n} \big((\min_{a\in\mathcal{A}} \|a + b_i\|) + \|c_i\|\big)^2 \\
&= \sum_{i=1}^{n} (\min_{a\in\mathcal{A}} \|a + b_i\|)^2 + 2\sum_{i=1}^{n} \|c_i\| \min_{a\in\mathcal{A}} \|a + b_i\| + \sum_{i=1}^{n} \|c_i\|^2 \\
&\le \bigg( \Big( \sum_{i=1}^{n} \min_{a\in\mathcal{A}} \|a + b_i\|^2 \Big)^{1/2} + \Big( \sum_{i=1}^{n} \|c_i\|^2 \Big)^{1/2} \bigg)^2.
\end{aligned}$$

For the reverse-inequality, express $b_i \coloneqq b_i' + c_i'$ and $c_i \coloneqq -c_i'$ to obtain

$$\Big( \sum_{i=1}^{n} \min_{a\in\mathcal{A}} \|a + b_i'\|^2 \Big)^{1/2} \le \Big( \sum_{i=1}^{n} \min_{a\in\mathcal{A}} \|a + b_i' + c_i'\|^2 \Big)^{1/2} + \Big( \sum_{i=1}^{n} \|c_i'\|^2 \Big)^{1/2}.$$

Specified to the loss in (3.53), one can rewrite these inequalities in terms of the MSE and (3.50)

$$\big(L_n(\boldsymbol{\mu})\big)^{1/2} - \big(\mathrm{MSE}_n\big)^{1/2} \le \big(\widetilde{L}_n(\boldsymbol{\mu})\big)^{1/2} \le \big(L_n(\boldsymbol{\mu})\big)^{1/2} + \big(\mathrm{MSE}_n\big)^{1/2}, \quad \forall \mu \in \mathbb{R}^M. \tag{3.54}$$

This perturbation inequality is important as it is the key to inheriting both the $\Gamma$-convergence and the equi-mild coercivity properties from $L_n^{(\omega)}$.

Recall that for sequences $\{a_n\}_n$ and $\{b_n\}_n$ where $a_n$ is potentially divergent and $\lim_n b_n = b$, the following limits hold with equality

$$\liminf_{n\to\infty}(a_n + b_n) = \liminf_{n\to\infty} a_n + b \quad \text{and} \quad \limsup_{n\to\infty}(a_n + b_n) = \limsup_{n\to\infty} a_n + b. \tag{3.55}$$

Equalities (3.55) paired with Definition 5 yields the following proposition for $\Gamma$-convergence.

**Proposition 5.** *Let $Q$ be the probability distribution given by transformation $u = g(x) + \varepsilon$. Suppose $MSE_n^{(\omega)} \to 0$. Then, for almost every $\omega$, objective $\widetilde{L}_n^{(\omega)}$ $\Gamma$-converges to*

$$L_*(\boldsymbol{\mu}) = \int \min_{k\in[M]} |\mu_k - u|^2 \, Q(du). \tag{3.56}$$

*Proof.* It suffices to show that $\widetilde{L}_n^{(\omega)}$ and $L_n^{(\omega)}$ have the same $\Gamma$-limit whenever $\mathrm{MSE}_n^{(\omega)} \to 0$. The pushforward distribution $Q$ inherits coordinate-wise independence from $\mathbb{P}$ so, by Theorem 7, the $\Gamma$-limit of $L_n^{(\omega)}$ equals (3.56) for almost every $\omega$.

We square-root transform $\widetilde{L}_n^{(\omega)}$ and apply (3.54) to obtain

$$\begin{aligned}
\liminf_{n\to\infty} \widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}_n) &= \left( \liminf_{n\to\infty} \left(\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}_n)\right)^{1/2} \right)^2 \\
&\geq \left( \liminf_{n\to\infty} \left\{ \left(L_n^{(\omega)}(\boldsymbol{\mu}_n)\right)^{1/2} - \left(\mathrm{MSE}_n^{(\omega)}\right)^{1/2} \right\} \right)^2 \\
&= \left( \liminf_{n\to\infty} \left(L_n^{(\omega)}(\boldsymbol{\mu}_n)\right)^{1/2} - \lim_{n\to\infty} \left(\mathrm{MSE}_n^{(\omega)}\right)^{1/2} \right)^2 \\
&= \liminf_{n\to\infty} L_n^{(\omega)}(\boldsymbol{\mu}_n)
\end{aligned}$$

where on the last line (3.55) was used. Similarly for the recovery sequence inequality

$$\begin{aligned}
\limsup_{n\to\infty} \widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}_n) &= \left( \liminf_{n\to\infty} \left(\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}_n)\right)^{1/2} \right)^2 \\
&\leq \left( \limsup_{n\to\infty} \left\{ \left(L_n^{(\omega)}(\boldsymbol{\mu}_n)\right)^{1/2} + \left(\mathrm{MSE}_n^{(\omega)}\right)^{1/2} \right\} \right)^2 \\
&= \limsup_{n\to\infty} L_n^{(\omega)}(\boldsymbol{\mu}_n).
\end{aligned}$$

Reviewing Definition 5, we see that any $\Gamma$-limit of $L_n^{(\omega)}$ must also be a $\Gamma$-limit of $\widetilde{L}_n^{(\omega)}$. □

Next, we show equi-mild coercivity (Definition 6) for the perturbed objective $\widetilde{L}_n$.

**Proposition 6.** *Suppose $MSE_n^{(\omega)} \to 0$. If $L_n^{(\omega)}$ satisfies* (3.52)*, then $\{\widetilde{L}_n^{(\omega)}\}_n$ is a sequence of equi-mild coercive objectives.*

*Proof.* It suffices to show that the sequence of $\widetilde{L}_n^{(\omega)}$-minimizers, $\{\widetilde{\boldsymbol{\mu}}_n\}_n$, are bounded for $\omega$ satisfying (3.52). Suppose, for the sake of contradiction, $\|\widetilde{\boldsymbol{\mu}}_n\| \to \infty$. By the diverging nature

of $\widetilde{\boldsymbol{\mu}}_n$, every $R > 0$ can be associated with a $N \in \mathbb{N}$ such that, for all $n > N_1$,

$$\widetilde{L}_n^{(\omega)}(\widetilde{\boldsymbol{\mu}}_n) = \inf_{\|\boldsymbol{\mu}\|>R} \widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}).$$

Let $R$ be defined as in Proposition 4 where, when paired with $\delta > 0$,

$$\inf_{\boldsymbol{\mu}\in\mathcal{U}} \big(L_n^{(\omega)}(\boldsymbol{\mu})\big)^{1/2} \leq \big(\inf_{\|\boldsymbol{\mu}\|>R} L_n^{(\omega)}(\boldsymbol{\mu}) - \delta\big)^{1/2}$$

Additionally by (3.54),

$$\begin{aligned}\inf_{\boldsymbol{\mu}\in\mathcal{U}} \big(L_n^{(\omega)}(\boldsymbol{\mu})\big)^{1/2} &\geq \inf_{\boldsymbol{\mu}\in\mathcal{U}} (\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}))^{1/2} - \big(\mathrm{MSE}_n^{(\omega)}\big)^{1/2} \\ &= \inf_{\|\boldsymbol{\mu}\|>R} (\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}))^{1/2} - \big(\mathrm{MSE}_n^{(\omega)}\big)^{1/2}, \quad \text{for suff. large } n.\end{aligned}$$

Stringing both inequalities together and further upperbounding $L_n^{(\omega)}$ by (3.54) yields

$$\inf_{\|\boldsymbol{\mu}\|>R} (\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}))^{1/2} \leq \big(\inf_{\|\boldsymbol{\mu}\|>R} \widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}) + \big(\mathrm{MSE}_n^{(\omega)}\big)^{1/2} - \delta\big)^{1/2} + \big(\mathrm{MSE}_n^{(\omega)}\big)^{1/2}.$$

However, since $\mathrm{MSE}_n^{(\omega)} \to 0$, there exists $N_2 \in \mathbb{N}$ such that, for all $n > N_2$,

$$\inf_{\|\boldsymbol{\mu}\|>R} (\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}))^{1/2} \leq \big(\inf_{\|\boldsymbol{\mu}\|>R} \widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}) - \delta/2\big)^{1/2}$$

which is a contradiction for all $\delta > 0$. This completes the proof. □

Combining Propositions 5 and 6 to apply Theorem 6, we show that, for almost-every $\omega \in \mathcal{B}$, the sequence $\{\widehat{\boldsymbol{\mu}}_n\}_n$ minimizes $L_*$ in the limit. In particular, there exists a sub-index set $\mathcal{I}'' \subseteq \mathcal{I}'$ such that $\widehat{\boldsymbol{\mu}}_m \to \overline{\boldsymbol{\mu}} \in \mathcal{M}$ with $m \in \mathcal{I}''$. This shows the desired claim (3.49) for any initial choice of index set $\mathcal{I} \subset \mathbb{N}$.

### D.4 Nearest Label Consistency from Cluster Convergence

The convergence of cluster centers $\{\widehat{\boldsymbol{\mu}}_m\}_m \to \overline{\boldsymbol{\mu}}$ guarantees a convergence in the nearest label estimates $\{\widehat{\boldsymbol{z}}_m\}_m$ to limiting labels $\bar{z}$. These convergences are abundant in $\mathcal{B}$ in the sense that every subsequence of $\{\widehat{\boldsymbol{\mu}}_n\}_n$ has a further subsequence which converges to an element in $\mathcal{M}$.

Fix one such limiting center $\overline{\boldsymbol{\mu}} \in \mathcal{M}$ and let $\{\mathcal{V}_k\}_{k=1}^M$ be the Voronoi-partition on $\mathbb{R}$ according to $\{\overline{\mu}_k\}_{k=1}^M$. Refer to (3.46) for a definition of $\mathcal{V}_k$. In the case when centers $\{\overline{\mu}_k\}_{k=1}^M$ are distinct, each $\mathcal{V}_k$ can be expressed as the intersection of $M$ half-spaces $\mathcal{V}_k = \bigcap_{\ell=1}^M H_{k\ell}(\overline{\boldsymbol{\mu}})$ where

$$H_{k\ell}(\boldsymbol{\mu}) \coloneqq \big\{u \in \mathbb{R} : \big(u - (\mu_k + \mu_\ell)/2\big) \cdot (\mu_k - \mu_\ell) \geq 0\big\}. \tag{3.57}$$

By Assumption 1, any minimizer of $L_*$ must have $M$ distinct values. As such, the map $\{\overline{\mu}_k\}_k \mapsto \{\mathcal{V}_k\}_k$ is well-defined for all minimizers of (3.56).

Let $\mathring{A}$ denote the interior of a set $A$. To obtain nearest labels $\overline{z}$ according to $\overline{\boldsymbol{\mu}}$, we consider the following routine:

1. If a point $u_i$ lies in the interior of a cell $\mathring{\mathcal{V}}_k$ then $\overline{z}_i = k$.

2. Otherwise $\overline{z}_i$ is selected arbitrarily from $[M]$.

Note that the nearest labels $\overline{z}$ are not guaranteed to agree with the generating labels $z^*$. This is to be expected whenever $(x, \varepsilon) \mapsto g^*(x) + \varepsilon$ is not injective over $\operatorname{supp}(\mathbb{P})$. In the case of separable clusters, that is, when noise $\varepsilon$ is bounded and small relative to the levels of $g^*(x)$, one indeed has, up to a label permutation, $\overline{z}_i = z_i^*$ for all generated samples $(x_i, \varepsilon_i)_{i=1}^\infty$.

The nearest center labeling routine can also be extended to any estimated centers $\widehat{\boldsymbol{\mu}}_m \in \mathbb{R}^M$. Let $\{\mathcal{V}_{k,m}\}_k$ be the Voronoi partition of $\widehat{\boldsymbol{\mu}}$ and let $\widehat{\boldsymbol{z}}_n$ be the corresponding nearest labels. Similar to the population case, partition $\{\mathcal{V}_{k,m}\}_{k=1}^M$ is well-defined whenever $\widehat{\boldsymbol{\mu}}_m \to \overline{\boldsymbol{\mu}}$ and $m$ is sufficiently large.

When calculating misclassification, we will avoid ambiguity by measuring misclassification relative to the interior of different Voronoi cells. With this in mind, the misclassification (3.48) for nearest labels $\widehat{\boldsymbol{z}}_m$ and $\overline{z}$ can be expressed as

$$\operatorname{Miss}_m(\widehat{\boldsymbol{z}}_m, \overline{z}) = \sum_{k \neq \ell} C_{k\ell}, \tag{3.58}$$

where

$$C_{k\ell} \coloneqq \sum_{k\neq\ell} \frac{1}{n} \sum_{i=1}^{n} 1_{\mathring{\mathcal{V}}_{k,n}}(u_i) \cdot 1_{\mathring{\mathcal{V}}_{\ell}}(u_i) \tag{3.59}$$

and $u_i = g^*(x_i) + \varepsilon_i$ are sample coordinates with joint distribution $Q$.

Next, we present the following almost-sure misclassification convergence result for any sequence of centers which satisfies $\widehat{\boldsymbol{\mu}}_m \to \overline{\boldsymbol{\mu}}$.

**Theorem 8.** *Let $\{\widehat{\boldsymbol{z}}_m\}_m$ and $\bar{z}$ be defined as before. Suppose $g^*(x_i) + \varepsilon_i = u_i \overset{i.i.d.}{\sim} Q$ and $\widehat{\boldsymbol{\mu}}_m \to \overline{\boldsymbol{\mu}}$. Then, for almost every $(x_i, \varepsilon_i)_{i=1}^{\infty} \in \mathcal{B}$,*

$$\lim_{m\to\infty} \mathit{Miss}_m(\widehat{\boldsymbol{z}}_m, \bar{z}) = 0.$$

*Proof.* It suffices to show $C_{k\ell} \to 0$ for all $k \neq \ell$. Recall that $\operatorname{int}\big(\bigcap_i A_i\big) \subseteq \bigcap_i \mathring{A}_i$. Applying the interior intersection relation to the Voronoi cells $\mathcal{V}_{k,m}$ and $\mathcal{V}_{\ell}$ yields

$$\mathring{\mathcal{V}}_{k,m} \subseteq \bigcap_{\ell=1}^{M} \mathring{H}_{k\ell}(\widehat{\boldsymbol{\mu}}_m) \subseteq \mathring{H}_{k\ell}(\widehat{\boldsymbol{\mu}}_m) \qquad \text{and} \qquad \mathring{\mathcal{V}}_{\ell} \subseteq \bigcap_{k=1}^{M} \mathring{H}_{\ell k}(\overline{\boldsymbol{\mu}}) \subseteq \mathring{H}_{\ell k}(\overline{\boldsymbol{\mu}}).$$

And as a consequence,

$$C_{k\ell} \leq \frac{1}{m} \sum_{i=1}^{m} 1_{\mathring{H}_{k\ell}(\widehat{\boldsymbol{\mu}}_m)}(u_i) \cdot 1_{\mathring{H}_{\ell k}(\overline{\boldsymbol{\mu}})}(u_i).$$

To decouple from $\widehat{\mu}_n$ we define the following $\delta$-silhouette about $\bar{\mu}$ for half-spaces $H_{k\ell}(\cdot)$,

$$\mathring{H}^{\delta}_{k\ell}(\overline{\boldsymbol{\mu}}) \coloneqq \bigcup_{\|\boldsymbol{\mu}-\overline{\boldsymbol{\mu}}\|<\delta} \mathring{H}_{k\ell}(\boldsymbol{\mu}). \tag{3.60}$$

By the convergence of centers $\widehat{\boldsymbol{\mu}}_m \to \overline{\boldsymbol{\mu}}$, one eventually has $\mathring{H}_{k\ell}(\widehat{\boldsymbol{\mu}}_m) \subseteq \mathring{H}^{\delta}_{k\ell}(\overline{\boldsymbol{\mu}})$ for all $k \neq \ell$ and any fixed $\delta > 0$. Therefore,

$$\begin{aligned}
C_{k\ell} &\leq \frac{1}{m} \sum_{i=1}^{m} 1_{\mathring{H}^{\delta}_{k\ell}(\overline{\boldsymbol{\mu}})}(u_i) \cdot 1_{\mathring{H}_{\ell k}(\overline{\boldsymbol{\mu}})}(u_i) \\
&\leq \frac{1}{m} \sum_{i=1}^{m} |1_{\mathring{H}^{\delta}_{k\ell}(\overline{\boldsymbol{\mu}})}(u_i) - 1_{H_{k\ell}(\overline{\boldsymbol{\mu}})}(u_i)| \cdot 1 + \frac{1}{m} \sum_{i=1}^{m} |1_{H_{k\ell}(\overline{\boldsymbol{\mu}})}(u_i)| \cdot 1_{\mathring{H}_{\ell k}(\overline{\boldsymbol{\mu}})}(u_i) \\
&= \frac{1}{m} \sum_{i=1}^{m} |1_{\mathring{H}^{\delta}_{k\ell}(\overline{\boldsymbol{\mu}})}(u_i) - 1_{H_{k\ell}(\overline{\boldsymbol{\mu}})}(u_i)| + 0,
\end{aligned}$$

where the last line follows from the fact $H_{k\ell}(\overline{\boldsymbol{\mu}}) \cap \mathring{H}_{\ell k}(\overline{\boldsymbol{\mu}}) = \varnothing$. Going forward we will suppress all dependence on $\overline{\boldsymbol{\mu}}$ for half-space sets $H_{k\ell}$ and $\mathring{H}^{\delta}_{k\ell}$.

Next, note that $H_{k\ell} \subseteq \mathring{H}^{\delta}_{k\ell}$ for all $\delta > 0$. Clearly by construction $\mathring{H}_{k\ell} \subseteq \mathring{H}^{\delta}_{k\ell}$. So, for $u \in H_{k\ell} \setminus \mathring{H}_{k\ell}$, consider an arbitrarily small perturbation $\epsilon \propto \operatorname{sgn}(\mu_k - \mu_\ell)$. Since $u$ satisfies

$$\big(u - (\overline{\mu}_k + \overline{\mu}_\ell)/2\big) \cdot (\overline{\mu}_k - \overline{\mu}_\ell) = 0,$$

one can shift $\overline{\mu}$ as $\mu = \overline{\mu} + \epsilon$ to obtain

$$\big(u - (\overline{\mu}_k + \overline{\mu}_\ell + 2\epsilon)/2\big) \cdot (\overline{\mu}_k - \overline{\mu}_\ell) = \epsilon \cdot (\overline{\mu}_k - \overline{\mu}_\ell) \propto |\overline{\mu}_k - \overline{\mu}_\ell|,$$

where the RHS is strictly positive whenever $\overline{\mu}_k \neq \overline{\mu}_\ell$.

As such, we can combine indicators and have

$$\begin{aligned} \lim_{m\to\infty} C_{k\ell} &\leq \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{n} 1_{\mathring{H}^{\delta}_{k\ell} \setminus H_{k\ell}}(u_i) \\ &= Q(\mathring{H}^{\delta}_{k\ell} \setminus H_{k\ell}), \end{aligned} \tag{3.61}$$

where the last line holds by independence for almost every $(u_i)_{i=1}^{\infty}$. By similar reasoning, measure convergence (3.61) holds, simultaneously, the countable family of sets $\big\{\mathring{H}^{\delta_p}_{k\ell} \setminus H_{k\ell}\big\}_{p=1}^{\infty}$. Let $\delta_p$ be any positive real sequence with $\delta_p \to 0^+$, then, by continuity of measure,

$$\begin{aligned} \lim_{m\to\infty} C_{k\ell} &\leq \inf_{p\to\infty} Q\big(\mathring{H}^{\delta_p}_{k\ell} \setminus H_{k\ell}\big) \\ &= Q\Big(\bigcap_{p=1}^{\infty} \big(\mathring{H}^{\delta_p}_{k\ell} \setminus H_{k\ell}\big)\Big) \\ &= Q\Big(\Big(\bigcap_{p=1}^{\infty} \mathring{H}^{\delta_p}_{k\ell}\Big) \setminus H_{k\ell}\Big) \\ &= Q(\varnothing) = 0. \end{aligned}$$

To prove the empty set assertion, note the following contrapositive statement

$$u \notin H_{k\ell} \implies u \notin \bigcap_{p=1}^{\infty} \mathring{H}^{\delta_p}_{k\ell}.$$

Indeed, fix $u$ and define the continuous function $t_u(\mu) \coloneqq \big(u - (\mu_k + \mu_\ell)/2\big) \cdot (\mu_k - \mu_\ell)$. If $u \notin H_{k\ell}$ then $t_u(\bar{\mu}) < 0$ and, by continuity, there exists some $\epsilon_u > 0$ such that

$$\|\mu - \bar{\mu}\| < \epsilon_u \implies t_u(\mu) < 0.$$

Therefore, for every $m$ satisfying $\delta_m < \epsilon_u$, one has $u \notin \mathring{H}_{k\ell}^{\delta_p}$. □

As a last comment, we note that the proof of Theorem 8 can be straight-forwardly extended for samples $u_i$ belonging to a general Hilbert space $\mathbb{H}$. In the case where $\mathbb{H}$ is a Banach space, modifications must be made to the half-space representation of $\mathcal{V}_k$.

# Part II

# The Role of Neighbor Aggregation on Graphs

# CHAPTER 4

# Simple GNNs with Low Rank Non-parametric Aggregators

## 1 Introduction

The problem of semi-supervised node classification (SSNC) [See02, BNS06] has been a focal point in graph-based classification for roughly 20 years. At the task's inception, classical methods such as label propagation [ZGL03] and kernel learning [BNS06] had seen moderate success in predicting unobseved node labels. Now, in an era where computation is more plentiful, modern approaches to the classification problem on graphs make use of the multilayer Graph Neural Network (GNNs) [SGT09].

These networks, trained to predict node labels in SSNC, draw on both the individual node features ($\boldsymbol{X}$) and the broader network structure ($\boldsymbol{A}$) to inform their prediction.

The fundamental premise of SSNC is that the network structure allows us to borrow information from neighboring nodes for which we lack a response. This borrowing can enhance the prediction of the unobserved responses ($\boldsymbol{y}$) beyond what could be achieved with a traditional regression solely on node features. Recently, there has been a wide breadth of literature [VCC18, CPL21, LHL22] which attempts to better leverage the network structure of the graph using GNNs. This recent flurry of activity has led to the proposal of many competing, and often intricate, architectures to solve the SSNC problem.

Our study of the leading GNN architectures and the benchmarks used to prove their algorithmic effectiveness, has led us to believe that many of the design choices found in

modern GNNs may be drastically simplified, or even removed completely, at little-to-no cost to predictive performance. In our efforts to validate model performances, we revisit traditional estimation techniques like non-parametric regression. These techniques happen to be very effective for SSNC and highlight the importance of learnable feature aggregation in SSNC problems.

To this end, we devise a flexible non-parametric learner for feature aggregation. This learner generalizes the specific polynomial form used in spectral GNNs [DBV16, WZ22]. That is, given a singular value decomposition of the network graph $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, the non-parametric learner $f : \mathbb{R} \to \mathbb{R}$ transforms the spectrum of $\boldsymbol{A}$ to produce a new aggregation matrix

$$\boldsymbol{P}_f = \boldsymbol{U} f(\boldsymbol{\Sigma}) \boldsymbol{V}^T$$

where $f$ is applied entry-wise across the diagonal of $\boldsymbol{\Sigma}$. This singular value extension to the previous symmetric spectral approach of [WZ22] helps clear a directed graph hurdle faced by previous spectral GNN techniques.

Our contributions are as follows:

1. Propose a nonparametric approach to learn $f$, hence a GNN aggregation operator, by borrowing ideas from the theory of reproducing kernel Hilbert spaces (RKHS), thus generalizing polynomial aggregation to a much broader class of spectral functions. By controlling the underlying kernel, one can impose different regularity constraints on the spectral filters.

2. Highlight the importance and sensitivity of nonparametric spectral reshaping and show how it can be used to simpify model hyperparameters (e.g. dropout probabilities, model depth, parameter-specific optimizers) at near-no-cost to SOTA performance.

3. Classification improvements of +5% and +20% compared to competing spectral methods and other non-linear GNN baselines for the challenging benchmark datasets Chameleon and Squirrel [RAS21].

4. Outline common evaluation practices which have an outsized effect on model performance.

By standardizing evaluation practices and simplifying modeling considerations, we aim to disambiguate performance in the GNN model-space and hope to encourage more interpretable models and heuristics for future SSNC problems.

## 2 GNN and SSNC Formalism

In our observation framework, we consider observing a, potentially noisy realization, of the network with adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and node feature matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$. Specifically, each node in the network $i \in [n]$ is associated with a feature vector $\boldsymbol{x}_i$ and a label $y_i \in [C] := \{1, \ldots, C\}$.

In SSNC, it is assumed that for a subset of nodes $\mathcal{O} \subset [n]$ the labels $(y_i)_{i \in \mathcal{O}}$ are observed. In this setting, both the adjacency matrix $\boldsymbol{A}$ and the feature matrix $\boldsymbol{X}$ are assumed to be fully observed. The goal then is to correctly predict unobserved labels $(y_i)_{i \in \mathcal{O}^c}$ from the previously stated knowns.

GNNs are designed layerwise, with non-linearity $\phi^\ell : \mathbb{R} \to \mathbb{R}$, weight matrix $\boldsymbol{W}^\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ and aggregation matrix $\boldsymbol{P}^\ell \in \mathbb{R}^{n \times n}$ all depending on layer $\ell \in [L]$. Placed altogether, the intermediate features of the GNN can be expressed as

$$\boldsymbol{Z}^{\ell+1} = \phi^\ell(\boldsymbol{P}^\ell \boldsymbol{Z}^\ell \boldsymbol{W}^\ell) \tag{4.1}$$

with $\phi^\ell$ applied element-wise, $d_0 = d$ and $\boldsymbol{Z}^1 = \boldsymbol{X}$. In the case of a $C$-class classification problem, it is common to extract row-wise "argmax"s of the final features $\boldsymbol{Z}^L \in \mathbb{R}^{n \times C}$ using differentiable argmax surrogates such as softmax. Choice of the aggregation matrix $\boldsymbol{P}^\ell$ may vary dramatically depending on architecture, but common choices include the adjacency matrix $\boldsymbol{A}$, its transformed variants (e.g. normalized Laplacian), and other, learnable, attention-based mechanisms [VCC18].

## 2.1 Nonparametric Spectral Reshaping

In our proposed model, we consider the simplest variant of GNN: a one layer ($L = 1$), linear GNN, that is $\phi = \mathrm{id}$, where special attention is paid to the propagation structure $\boldsymbol{P}$. For ease of exposition, we first consider the undirected case where the adjacency matrix $\boldsymbol{A}$ is symmetric. Let $\boldsymbol{M}$ be a (symmetric) *network matrix* derived from $\boldsymbol{A}$. Examples include $\boldsymbol{M} \in \{\boldsymbol{A}, \boldsymbol{D}-\boldsymbol{A}, \hat{\boldsymbol{A}}, \boldsymbol{I}-\hat{\boldsymbol{A}}\}$ where $\hat{\boldsymbol{A}} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$. Our approach is to consider a general nonlinear deformation of $\boldsymbol{M}$, namely, $f(\boldsymbol{M})$ where $f : \mathbb{R} \to \mathbb{R}$ is a univariate function extended to the space of symmetric matrices by the so-called *functional calculus*. More precisely, given the eigendecomposition $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ of the $\boldsymbol{M}$ matrix, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_i, i \in [n])$, one has

$$f(\boldsymbol{M}) = \boldsymbol{U} f(\boldsymbol{\Lambda}) \boldsymbol{U}^T$$

where $f(\boldsymbol{\Lambda}) = \mathrm{diag}(f(\lambda_i), i \in [n])$ is the natural extension of $f$ to diagonal matrices. This way of extending univariate functions to self-adjoint operators has a long history in operator theory. Thus, our propagation operator is $\boldsymbol{P}_f = f(\boldsymbol{M})$ and we propose to optimize a loss over a general class of functions $\mathcal{F}$:

$$\hat{f} = \underset{f \in \mathcal{F},\, \boldsymbol{W} \in \mathbb{R}^{d \times C}}{\mathrm{argmin}} \sum_{i \in \mathcal{O}} \ell\big(y_i, (f(\boldsymbol{M})\boldsymbol{X}\boldsymbol{W})_i\big) + \mathrm{pen}(f) \tag{4.2}$$

where $\mathrm{pen}(f)$ is some regularization penalty on $f$. Our main claim is that rather than assuming a specific parametric form for $f$, one can allow $f$ to range in a potentially infinite-dimensional function space $\mathcal{F}$.

Of particular interest to us is when $\mathcal{F} = \mathbb{H}$, a reproducing kernel Hilbert space (RKHS) of functions, characterized by a kernel function $\mathcal{K} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. In such a space, the Hilbert norm $\|f\|_{\mathbb{H}}$ measures irregularity of $f$. Then, as long as $\mathrm{pen}(f)$ is a monotonic function of the Hilbert norm $\|f\|_{\mathbb{H}}$, by the so-called represented theorem [SHS01], problem (4.2) reduces to

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^n,\, \boldsymbol{W}}{\mathrm{argmin}} \sum_{i \in \mathcal{O}} \ell\big(y_i, (\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{\alpha})\boldsymbol{X}\boldsymbol{W})_i\big) + \widetilde{\mathrm{pen}}(\boldsymbol{\alpha}), \tag{4.3}$$

$$\text{where} \quad \boldsymbol{P}_{\mathcal{K}}(\boldsymbol{\alpha}) := \boldsymbol{U}(\mathrm{diag}(\boldsymbol{K}\boldsymbol{\alpha}))\boldsymbol{U}^T, \tag{4.4}$$

and $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ is the kernel matrix with entries $K_{ij} = \mathcal{K}(\lambda_i, \lambda_j)$. If $\text{pen}(f) = \omega(\|f\|_{\mathbb{H}})$ for monotonic function $\omega : \mathbb{R}_+ \to \mathbb{R}$, then $\widetilde{\text{pen}}(\boldsymbol{\alpha}) = \omega(\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha})$. Given $\hat{\boldsymbol{\alpha}}$ one can explicitly write down the solution $\hat{f}$ of the functional problem (4.2) as

$$\hat{f}(\lambda) := \sum_j \hat{\boldsymbol{\alpha}}_j \mathcal{K}(\lambda, \lambda_j)$$

which is the learned spectral filter.

**Practical considerations.** We found slight improvements in performance when regularizing with $\boldsymbol{\alpha}^T\boldsymbol{\alpha}$ rather than the Hilbert norm surrogate $(\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha})$. This amounts to using $\widetilde{\text{pen}}(\boldsymbol{\alpha}) = \rho\,\boldsymbol{\alpha}^T\boldsymbol{\alpha}$ for some $\rho > 0$. When minimizing with GD type methods, this is equivalent to introducing weight decay $\rho$, and is already built into SOTA solvers.

Additionally, we consider the possibility that edges in the network themselves have a component of randomness associated with them (Section 2.2).

This means that our initial spectral inputs $\lambda_1, \lambda_2, \ldots, \lambda_n$ are themselves noisy. It is then natural to truncate the spectral decomposition of $\boldsymbol{M}$ to the top $r$ eigenvalues (in absolute values). Thus, if the eigenvalues are ordered as $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$, we consider $\boldsymbol{M}^{(r)} = \boldsymbol{U}\boldsymbol{\Lambda}^{(r)}\boldsymbol{U}^T$ where $\boldsymbol{\Lambda}^{(r)} = (\lambda_i, i \in [r])$ and let the aggregation matrix be $f(\boldsymbol{M}^{(r)}) = \boldsymbol{U}f(\boldsymbol{\Lambda}^{(r)})\boldsymbol{U}^T$. Following through as before, the only changes to the algorithm is to replace $\boldsymbol{K}$ in (4.4) with $\boldsymbol{K}^{(r)} = (\mathcal{K}(\lambda_i, \lambda_j))_{i,j=1}^r$. We also note that $\boldsymbol{\alpha}$, the learnable spectral parameter, will be $r$-dimensional in this case. We treat the $r \in [n]$ as a hyperparameter and study its effect in simulations. We refer to the case $r < n$ as low-rank (LR) kernel model.

**Directed/asymmetric case.** All the above naturally extends to directed networks, where $\boldsymbol{M}$ is not necessarily symmetric, by replacing the eigenvalue decomposition with the SVD: $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ where $\boldsymbol{\Sigma} = \text{diag}(\sigma_i, i \in [n])$ collects the singular values of $\boldsymbol{M}$. The aggregation matrix in this case is $\boldsymbol{P}_f = \boldsymbol{U}f(\boldsymbol{\Sigma})\boldsymbol{V}^T$ and its finite-dimensional version is $\boldsymbol{P}_{\mathcal{K}}(\boldsymbol{\alpha}) = \boldsymbol{U}(\text{diag}(\boldsymbol{K}\boldsymbol{\alpha}))\boldsymbol{V}^T$ with the kernel matrix $\boldsymbol{K} = (\mathcal{K}(\sigma_i, \sigma_j))_{i,j=1}^n$ now based on singular values.

Everything else follows similarly, including rank truncation, where we use ordered singular values instead.

**Multiple layers.** We mainly focus on a single-layer model (with identity activation) and empirically show that a single layer of this model is enough to achieve near SOTA performance. However, it is straightforward to extend the model to multiple layers via the general blueprint (4.1) where each layer will have aggregation operator $\boldsymbol{P}^\ell = f_\ell(\boldsymbol{M})$ with $f_\ell$ belonging to $\mathbb{H}$.

### 2.2 Motivating General Spectral Learners

Implicit in all graph learning problems is the assumption that node features $\boldsymbol{X}$ are only partially informative towards predicting $\boldsymbol{y}$. To motivate why a spectral GNN of the form (4.2), with a general reshaping function $f$ can improve prediction, let us consider perhaps the simplest theoretical model of SSNC, the so-called Contextual Stochastic Block Model (CSBM) [DSM18]. The idea is that the labels $\boldsymbol{y}$ are latent variables generating both $\boldsymbol{A}$, via a $C$-class SBM: $\mathbb{P}(A_{ij} = 1 \,|\, \boldsymbol{y}) = B_{y_i, y_j}$, and the node features via a mixture model: $\boldsymbol{x}_i \,|\, y_i \sim N(\boldsymbol{\mu}_{y_i}, \sigma^2 I)$.

In this case, the idealized version of $\boldsymbol{A}$ is $\mathbb{E}[\boldsymbol{A}]$ which is a rank $C$ matrix with $C$ eigenvectors that are *indicator vectors* of each of the $C$ classes $\Gamma_1, \ldots, \Gamma_C \in \{0,1\}^n$. Consequently, if $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ is the EVD of $\boldsymbol{A}$, one expects $\boldsymbol{U}_{:j} \approx \mathbf{1}_{\Gamma_j}$ for $j \in [C]$, while $\boldsymbol{U}_{:j}$ for $j > C$ are expected to be mostly noise. So an ideal aggregation operator is close to $f(\boldsymbol{M}) = \boldsymbol{U} f(\boldsymbol{\Lambda}) \boldsymbol{U}^T$ where $f$ is a step function that passes the $\lambda_j, j \in [C]$ through and zero out the rest. As the experiment in Section 3.2 show, this is mostly what happens when we train (4.2) on CSBM, albeit with more nuance. In finite samples, $\boldsymbol{A}$ is not exactly low-rank and lower eigenvectors might still have information about $\boldsymbol{y}$. This is what we observe in practice where the learned $f$ is a *tapered* thresholding operator, that gradually downweights lower frequencies.

The general low-rank behavior of the $\mathbb{E}[\boldsymbol{A}]$ is not limited to SBMs and holds for more realistic network models such as random dot product graph (RDPG) [YS07] where depending

on the distribution of latent positions, more complex tapering might be optimal.

### 2.3 Complexity of Low Rank Spectral Learners

Scalability remains an issue for dense spectral methods. However in the case of low-rank non-parametric aggregators, this issue can be addressed through the use of a low-rank spectral approximation. By first selecting a rank parameter $r$ for the non-parametric aggregator, computation can be better budgeted ahead of time for a graph $G = (V, E)$ through the use of a low-rank SVD approximations. Specifically for PyTorch, a low-rank, random SVD routine based on [HMT11] is implemented in the function `torch.svd_lowrank`.

Computation and error complexity for this routine can be found in section 6.2 of [HMT11]. For a sparse adjacency matrix given by $G$ and a number of total iterations $q$, this routine has a time complexity of $\mathcal{O}(qr|E| + r^2|V|)$ and an error complexity, in operator norm, of $(r|V|)^{1/2(2q+1)}\sigma_{r+1}$. In total, we obtain a decomposition procedure which is: exponentially exact with respect to $q$, at most quadratic in time with respect to $r$, at most linear in time with respect to graph parameters $|V|$ and $|E|$.

In the forward pass of the non-parametric aggregator, a graph with $d$-dimensional node features and $c$ classes will contribute a computational complexity of $\mathcal{O}(|V|c(d+r))$. Additionally, since the non-parametric aggregator is linear with respect to weights $W$ and parameter $\alpha$, gradient computation in the backward pass can re-use intermediaries found in the forward pass, potentially saving computation.

## 3 Experiments

In an effort to show the power of feature aggregation for SSNC problems, our modeling effects will focus entirely on the aggregation matrix $\boldsymbol{P}$. No modifications are made to the original features $\boldsymbol{X}$ or the structure of the linear weight $\boldsymbol{W}$. As such, in our experiments we do not consider any model-specific augmentations such as dropout [SHK14], batchnorm [IS15], or

per-parameter optimizers (i.e. different learning rates for different layers). The design of $\boldsymbol{P}$ will have the following degrees of freedom:

- **Matrix representation of network ($\boldsymbol{M}$):** We refer to any matrix $\boldsymbol{M}$ derived from algebraic manipulations of the adjacency of a network $\boldsymbol{A}$, to be a *matrix representation* of the network. In particular we consider the following two representations:
  - *Adjacency*: This is simply an identity transformation on $\boldsymbol{A}$ with $\boldsymbol{M} = \boldsymbol{A}$.
  - *Laplacian*: This is $\boldsymbol{M} = \boldsymbol{D} - \boldsymbol{A}$ where $\boldsymbol{D}$ is the row-sum degree matrix of $\boldsymbol{A}$.
- **Spectral truncation factor ($r$):** Given a truncation factor $r$, the spectral system $(\boldsymbol{U}, \boldsymbol{\Lambda})$, resp. $(\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}^T)$, will be reduced to $(\boldsymbol{U}_{:r}, \boldsymbol{\Lambda}_{:r})$, resp. $(\boldsymbol{U}_{:r}, \boldsymbol{\Sigma}_{:r}, (\boldsymbol{V}_{:r})^T)$, where the eigenvectors associated with the bottom $n - r$ eigenvalue magnitudes are dropped. In our experiments, spectral truncations from 0 to 95% in 5% intervals are considered.
- **Choice of kernel ($\mathcal{K}$):** In ordering our RKHS we select among the following kernels:
  - *Identity:* $K_{ij} = 1\{i = j\}$
  - *Linear (Outer product):* $\mathcal{K}(\sigma_i, \sigma_j) = \sigma_i \sigma_j$
  - *Compact Sobolev*: $\mathcal{K}(\sigma_i, \sigma_j) = \min(\sigma_i, \sigma_j)$
  - *Unbounded Sobolev:* $\mathcal{K}(\sigma_i, \sigma_j) = \exp(\gamma|\sigma_i - \sigma_j|)$
  - *Gaussian Radial Basis:* $\mathcal{K}(\sigma_i, \sigma_j) = \exp(\gamma|\sigma_i - \sigma_j|^2)$

  Note, in the case of identity, the "kernel" does not generate a continuous RKHS. For the last two kernels, the bandwidth parameter $\gamma \in \mathbb{R}_+$ can be determined on validation.

Note that, the choice of matrix representation $\boldsymbol{M}$ matters here insofar that it determines the "modes" or partitions of the network with its left and right eigenvectors $(\boldsymbol{U}, \boldsymbol{V})$.

For our optimizer, we use the standard Adam optimizer [KB15] with weight decay. For simplicity, both parameter $\boldsymbol{\alpha}$ and weight matrix $\boldsymbol{W}$ share the same weight decay under Adam.

### 3.1 SSNC Benchmarks

Our methods are evaluated against common SSNC benchmarks. The Chameleon, Squirrel, and Actor benchmarks contain directed networks, while the other benchmarks contain undirected networks. More information on all benchmarks can be found in [PWC20]. All values are recorded using the *balanced splits* defined in [CPL21]. Section 4 provides a comprehensive analysis on the impact of splitting conventions. Although not covered in this paper, alternative benchmarks for simple spectral models can be found in Zhu and Koniusz [ZK21].

The following linear and kernel models are considered for evaluation: Linear ($\boldsymbol{XW}$), Aggregated Linear ($\boldsymbol{MXW}$), Kernel ($\boldsymbol{P}_{\mathcal{K}}\boldsymbol{XW}$), and LR Kernel ($\boldsymbol{P}_{\mathcal{K},r}\boldsymbol{XW}$). Model hyperparameters such as learning rate, weight decay, the specific aggregator $\boldsymbol{P}$ will be determined for each dataset using the mean accuracies of the validation splits. For completeness, we have also implemented a non-linear baseline which learns using only feature information $\boldsymbol{X}$. This model is a simple two-layer ReLU multi-layer perceptron MLP2 ($\phi(\boldsymbol{XW^1})\boldsymbol{W}^2$) with hidden layer size determined on validation.

Our models and their results compared to other current SOTA methods can be found in Table 4.1. We note that, for almost all of the larger graph benchmarks, our models perform within uncertainty or better compared to SOTA. In particular for directed graphs like Chameleon and Squirrel, we see gains in accuracy as high as 5% and 20% over other SOTA methods. A point of emphasis here is the relative simplicity of our models compared to the performance they attain. The absence of any post-model augmentations distinguishes our approach from the implementations of other competing SOTA spectral methods like JacobiConv [WZ22].

A point of difficulty where the performance gap persists, is where the node response $\boldsymbol{y}$ is overwhelming described by its node information $\boldsymbol{X}$. Graphs with this property (Actor, Cornell, Texas, and Wisconsin) can be identified by the negative performance gap between Linear and Aggregated Linear as well as the SOTA-like performance of MLP2. Note

| | Cora | CiteSeer | PubMed | Chameleon | Squirrel | Actor | Cornell | Texas | Wisconsin |
|---|---|---|---|---|---|---|---|---|---|
| MLP2 | $77.8_{\pm 1.6}$ | $77.2_{\pm 1.1}$ | $88.2_{\pm 0.5}$ | $48.5_{\pm 2.6}$ | $34.8_{\pm 1.4}$ | $40.3_{\pm 2.3}$ | $86.1_{\pm 3.0}$ | $91.7_{\pm 4.4}$ | $95.0_{\pm 2.6}$ |
| Linear | $78.9_{\pm 2.0}$ | $76.2_{\pm 1.2}$ | $85.8_{\pm 0.4}$ | $48.1_{\pm 3.2}$ | $34.9_{\pm 1.4}$ | $38.9_{\pm 1.2}$ | $84.9_{\pm 5.6}$ | $89.7_{\pm 3.8}$ | $95.0_{\pm 3.8}$ |
| Agg. Linear | $84.0_{\pm 2.0}$ | $73.9_{\pm 1.4}$ | $82.6_{\pm 0.5}$ | $79.0_{\pm 1.4}$ | $78.0_{\pm 1.1}$ | $32.4_{\pm 1.3}$ | $67.8_{\pm 8.7}$ | $86.8_{\pm 3.5}$ | $83.8_{\pm 3.2}$ |
| Kernel | $88.6_{\pm 1.0}$ | $81.1_{\pm 1.0}$ | $89.4_{\pm 0.8}$ | $78.7_{\pm 1.1}$ | $76.0_{\pm 1.2}$ | $32.2_{\pm 1.8}$ | $83.3_{\pm 5.9}$ | $88.2_{\pm 2.6}$ | $92.1_{\pm 3.4}$ |
| LR Kernel | — | — | — | $79.4_{\pm 1.4}$ | $76.8_{\pm 1.3}$ | $32.3_{\pm 1.7}$ | — | — | — |
| GPRGNN* | $79.5_{\pm 0.4}$ | $67.6_{\pm 0.4}$ | $85.1_{\pm 0.1}$ | $67.5_{\pm 0.4}$ | $49.9_{\pm 0.5}$ | $39.3_{\pm 0.3}$ | $91.4_{\pm 0.7}$ | $92.9_{\pm 0.6}$ | NA |
| SGC/ASGC* | $73.9_{\pm 2.5}$ | $70.2_{\pm 1.0}$ | $79.1_{\pm 1.0}$ | $72.3_{\pm 0.9}$ | $59.0_{\pm 1.0}$ | $36.5_{\pm 0.8}$ | $86.8_{\pm 3.6}$ | $86.2_{\pm 3.1}$ | NA |
| JacobiConv* | $89.0_{\pm 0.5}$ | $80.8_{\pm 0.8}$ | $89.6_{\pm 0.4}$ | $74.2_{\pm 1.0}$ | $55.8_{\pm 0.6}$ | $40.7_{\pm 1.0}$ | $92.3_{\pm 2.8}$ | $92.8_{\pm 2.0}$ | NA |
| ACMII-GCN | $89.0_{\pm 0.7}$ | $81.8_{\pm 1.0}$ | $90.7_{\pm 0.5}$ | $68.4_{\pm 1.4}$ | $54.5_{\pm 2.1}$ | $41.8_{\pm 1.2}$ | $95.9_{\pm 1.8}$ | $95.1_{\pm 2.0}$ | $96.6_{\pm 2.4}$ |

Table 4.1: Performance: Mean test accuracy ± std. dev. over 10 data splits. Models include our own variations of "Linear" and "Aggregated Linear" GNNs, along with other state-of-the-art (SOTA) GNNs. Dashed entry in for LR Kernel signifies validated choice is the same as the full-rank Kernel. Performance is comparable between our simple GNNs and SOTA in some cases. Results for GPRGNN, SGC/ASGC. JacobiConv and ACMII-GCN are cited from [CPL21], [CM22], [WZ22], and [LHL22] respectively. Entries marked with '$*$' report 95% confidence intervals.

that, even without using any graph information, MLP2 is able to achieve SOTA within uncertainty on almost all of the $X$-dominated, network datasets. Furthermore, for the cases of Cornell, Texas, and Wisconsin, there is a possibility of running into sample size issues for graph based methods. With the exception of Actor, these datasets are only 100-200 nodes large (less than 1/10 the size of the other network benchmarks).

### 3.2 CSBM Experiment

To illustrate the effectiveness of nonparametric spectral learners, we performed an experiment on simulated CSBM data. We consider $C = 3$ classes and node features in $\mathbb{R}^3$ ($\boldsymbol{X} \in \mathbb{R}^{n\times 3}$) generated using `make_blobs` function of `scikit-learn` package with cluster standard deviation of 10. This leads to a hard classification problem for an oracle that only knows

| | max params. | $n = 300$ | $n = 600$ | $n = 1200$ | $n = 1500$ |
|---|---|---|---|---|---|
| $X$-ONLY ORACLE | 0 | $64.3 \pm 3.9$ | $66.2 \pm 2.9$ | $63.3 \pm 2.7$ | $64.6 \pm 1.4$ |
| KERNEL | 1512 | $75.0 \pm 3.5$ | $86.6 \pm 3.3$ | $94.5 \pm 1.2$ | $97.3 \pm 0.8$ |
| ACMII-GCN | 102623 | $75.3 \pm 6.4$ | $89.5 \pm 3.1$ | $96.0 \pm 1.2$ | $97.7 \pm 0.8$ |

Table 4.2: Simulation experiments on a three-class CSBM. Mean test accuracy and std. dev. of 10 runs are reported. $X$-ONLY ORACLE is the accuracy associated with oracle classification on solely $X$. Maximum parameter counts for the two methods are also summarized. Relevant average degree $\Delta_n$ for the simulations are $\Delta_{300} = 1.83$, $\Delta_{600} = 3.68$, $\Delta_{1200} = 7.58$, and $\Delta_{1500} = 9.44$.

$X$, with optimal Bayes accuracy of roughly 0.63 (for large $n$). The SBM component has connection probabilities $B_{kk} = 0.015$ and $B_{k\ell} = 0.02$ for $k \neq \ell$. We vary the number of nodes $n$ over $300, 600$, 1200, and 1500.

Table 4.2 summarizes the results for our nonparametric learner (KERNEL) and ACMII-GCN as a competing SOTA. Also shown is the average degree of the resulting networks. As $n$ increases the CSBM model becomes more informative, which is reflected in increased prediction accuracy. At the two ends of the SNR spectrum ($n = 300$ and $n = 1500$) the performance of the KERNEL GNN and ACMII-GCN are very close, while there is a slight advantage for ACMII-GCN in the middle ($n = 600$ and $n = 1200$), though the two methods are still comparable due to the overlap of the wide uncertainty ranges.

What, however, is noteworthy is the significant effect of the spectral shaping in GNN performance: the KERNEL GNN significantly improves the performance beyond the $X$-only oracle with very few parameters and at very low graph SNRs; for example, at $n = 300$, where the parameter count is 312 and the average degree is barely 2 (a very weak graph signal). The simplicity of the KERNEL GNN allows us to exactly quantify the effect of nonparametric spectral learning since this is the only operation performed outside of applying the learned

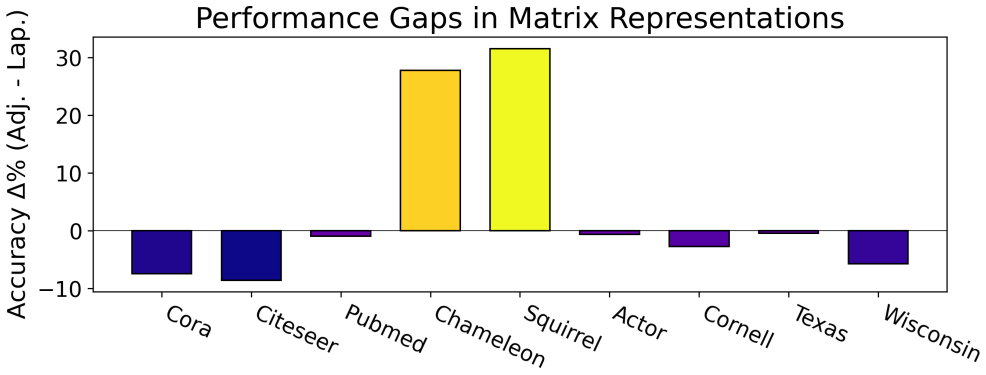



Figure 4.1: Accuracy comparison of the KERNEL model for different graph representations $\boldsymbol{A}$ and $\boldsymbol{D} - \boldsymbol{A}$. Shown above is the signed accuracy difference between the adjacency and Laplacian representations. Best performing kernel was selected per dataset.

linear weights $\boldsymbol{W}$.

### 3.3 Aggregation Ablation

To understand the impact of the degrees of freedom defined for the aggregation matrix in section 3, we conduct an ablation study on the three hyperparameters: matrix representation $\boldsymbol{M}$, truncation factor $r$, and the choice of kernel $\mathcal{K}$.

**Matrix Representation ($\boldsymbol{M}$):** For this experiment we keep spectral truncation fixed at 0% and choose the best kernel through validation splits. In other words, this experiment is conducted using the full-rank KERNEL model with a best validated kernel fit to each dataset. In the experiment, we explore affects of fixing either $\boldsymbol{M} = \boldsymbol{A}$ or $\boldsymbol{M} = \boldsymbol{D} - \boldsymbol{A}$.

Figure 4.1 shows the accuracy change across datasets when using a Laplacian matrix representation $\boldsymbol{D} - \boldsymbol{A}$ rather than an adjacency matrix representation $\boldsymbol{A}$. As shown by the figure, directed graphs such as Chameleon and Squirrel show large benefits when using the adjacency matrix representation. Otherwise there seems to be a slight but persistent benefit in using the Laplacian representation for undirected datasets.

| Kernel | Cora | CiteSeer | PubMed | Chameleon | Squirrel | Actor | Cornell | Texas | Wisconsin |
|---|---|---|---|---|---|---|---|---|---|
| Identity | $78.8 \pm 2.7$ | $72.6 \pm 2.0$ | $81.6 \pm 0.9$ | $69.7 \pm 2.7$ | $44.9 \pm 2.9$ | $28.6 \pm 3.0$ | $60.4 \pm 8.1$ | $76.2 \pm 4.3$ | $71.6 \pm 5.7$ |
| Sob. Cmpct. | $75.1 \pm 1.9$ | $73.0 \pm 1.4$ | $88.5 \pm 0.4$ | $41.4 \pm 2.2$ | $33.2 \pm 1.1$ | $\underline{32.2 \pm 1.8}$ | $\underline{83.3 \pm 5.9}$ | $88.6 \pm 4.0$ | $\underline{92.1 \pm 3.4}$ |
| Linear | $81.1 \pm 2.0$ | $72.1 \pm 1.8$ | $82.3 \pm 1.0$ | $\underline{78.7 \pm 1.2}$ | $\underline{76.0 \pm 1.2}$ | $31.6 \pm 0.9$ | $66.5 \pm 6.1$ | $77.2 \pm 8.0$ | $81.3 \pm 4.8$ |
| Sob. Unbnd. | $88.8 \pm 0.8$ | $\underline{81.1 \pm 1.0}$ | $89.2 \pm 2.0$ | $54.5 \pm 6.4$ | $68.8 \pm 8.2$ | $30.7 \pm 1.0$ | $80.6 \pm 6.4$ | $\underline{88.2 \pm 2.6}$ | $90.4 \pm 5.6$ |
| Gauss. RBF | $\underline{88.6 \pm 1.0}$ | $80.3 \pm 1.9$ | $\underline{89.4 \pm 0.8}$ | $60.4 \pm 8.4$ | $71.3 \pm 4.4$ | $30.4 \pm 1.3$ | $79.4 \pm 5.3$ | $84.0 \pm 4.5$ | $85.8 \pm 4.7$ |

Table 4.3: Impact of the kernel choice on the performance of the full-rank KERNEL model. Underlined entries correspond to the model selected by validation.

**Choice of Kernel ($\mathcal{K}$):** For this experiment we once again use the full-rank KERNEL model. This time the matrix representation $\boldsymbol{M}$ is chosen through validation and the choice of kernel is varied across datasets. Table 4.3 shows performance results for the various choices of kernels. In Table 4.3, we see a complicated dependence between kernel choice and the accuracy of node prediction. Although some results are within uncertainty, the dependence between kernel regularity and SSNC performance is not immediately clear. In the case of the Chameleon and Squirrel datasets, it is apparent that the wrong choice in kernel may lead to significant performance degradations (up to $\sim$30%).

**Spectral Truncation Factor ($r$):** For this experiment, both the matrix representation and the choice of kernel have been selected based on best validation with truncation factor $r$ fixed for the extent of each sub-experiment. Figure 4.2 demonstrates the effect of truncation on performance and how it gradually degrades with the truncation percentage. The rate at which performance degrades seems dependent on the dataset, but most benchmarks retain $\sim$90% performance even after a 40% spectral truncation. In special cases like Squirrel and Chameleon, performance can be seen to increase at larger truncation values.

**Alleviating Kernel Dependent Performance:** For this experiment, we explore the affects of kernel choice for the LR KERNEL model. In particular, we focus on the performance impact of kernel choice for the directed dataset benchmarks. Rather than reporting mean

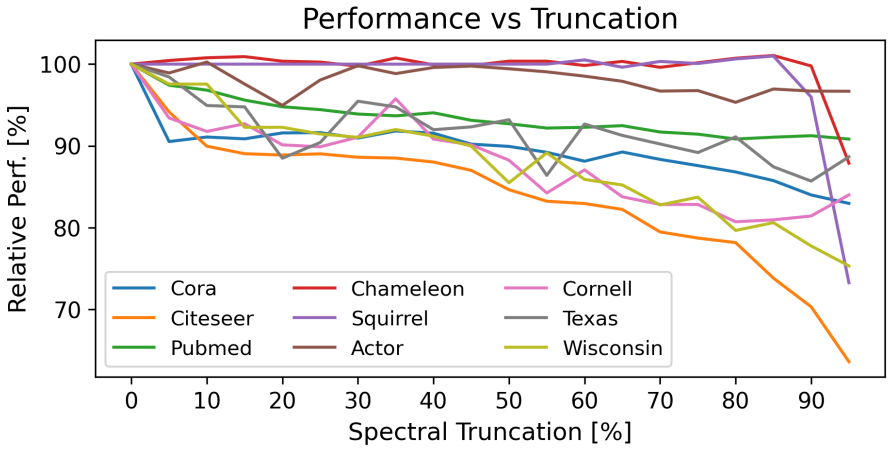



Figure 4.2: LR KERNEL performance relative to the full-rank KERNEL for different truncation factors $r$. Performance is seen to gradually decline on most datasets as the truncation factor $r$ decreases (that is truncation percentage increases). LR KERNEL performance can also be seen to periodically increase above full-rank KERNEL performance for the datasets Chameleon (red) and Squirrel (purple).

accuracies, Figure 4.3 shows the full violin plot of test split performances for each kernel-dataset combination. We notice a *homogenization* of results, where the choice of kernel is negligible to the overall SSNC performance. We stress however that this solution is partial, as the same order of homogenization is not observed for the other undirected datasets. Identifying relevant graph statistics which may describe this homogenization discrepancy is something which is left to future work.

## 4 Changes in Evaluation Conventions

The convention of using citation networks [SNB08] (Cora, Citeseer, Pubmed) in SSNC benchmarks was popularized by the graph embedding work of [YCS16]. [YCS16] defined the "sparse" train-test split of the citation datasets and their node masks were made publicly available. The sparse split fixed *20 nodes per class for training* and *1000 nodes total for testing*. These values were held constant across citation datasets, meaning larger networks

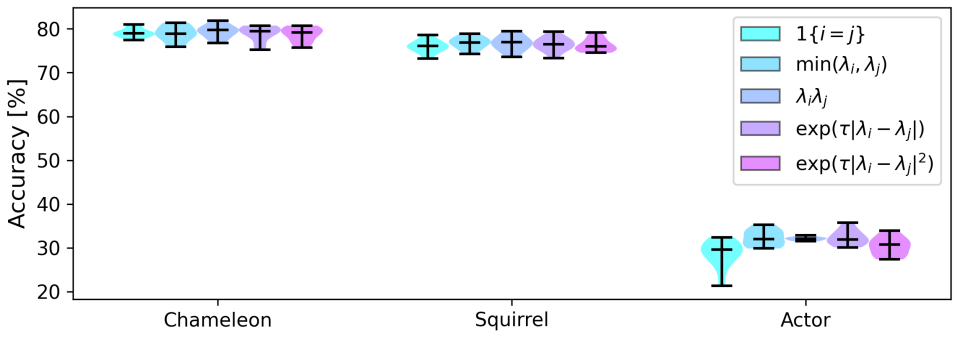



Figure 4.3: Performance homogenization achieved by LR KERNEL model on directed networks.

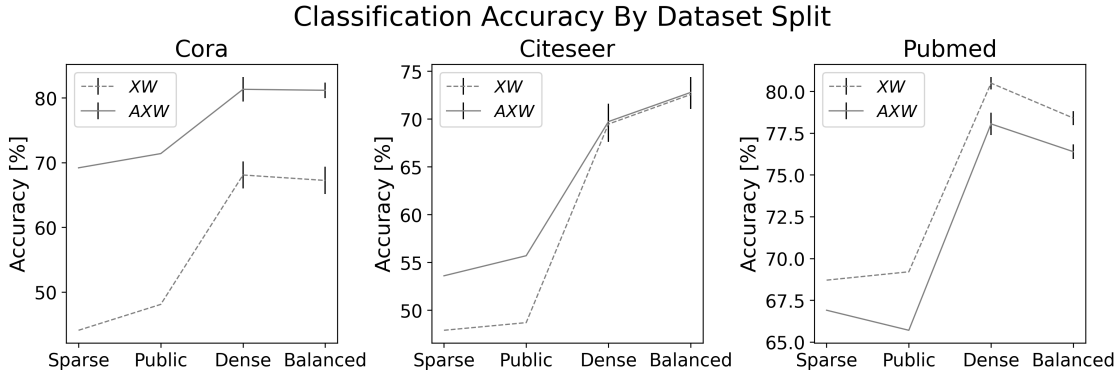



Figure 4.4: Accuracy results and uncertainties on the citation datasets using different splits with linear models $\boldsymbol{XW}$ and $\boldsymbol{AXW}$. "Public" refers to the split introduced by [KW17]. Both "Sparse" and "Public" are single splits, so one cannot associate uncertainty to them.

likes Pubmed were left with a relatively low label rate of ∼5%.

Quickly following was the semi-supervised work of [KW17] and [VCC18]. These follow-up papers defined a new "public" split where *500 previously unlabeled nodes* in the sparse split were now used for validation. In the respective code implementations of each paper, the additional labels were used for early stopping criteria and to determine the final model checkpoint.

Introduced later was the "dense" split by [PWC20], where train, validation, and test were now fractions of the whole graph, set to *60%-20%-20% respectively.* This paper also popularized two new benchmark datasets, the WebKB dataset [CDF98] (Cornell, Texas,

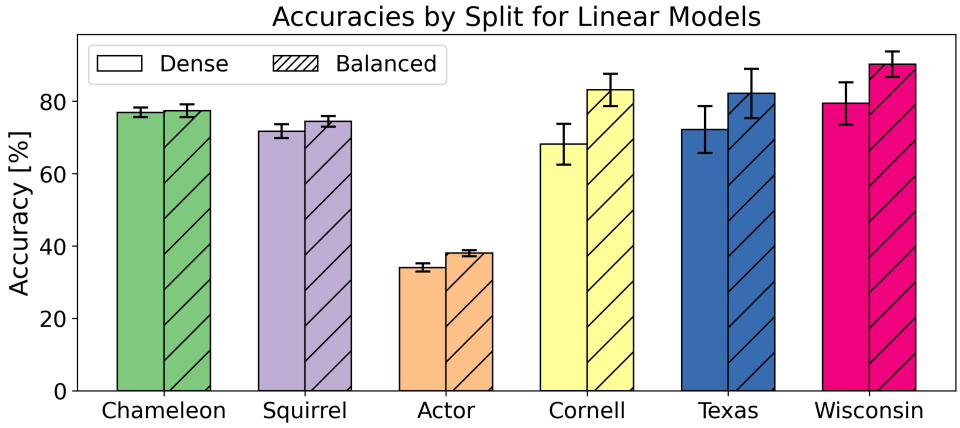



Figure 4.5: Accuracy results on datasets introduced by [PWC20]. "Dense" refers to the original split while "Balanced" refers to the split introduced by [CPL21]. Test results and uncertainties are evaluated using models $\boldsymbol{XW}$ and $\boldsymbol{AXW}$. Results shown are for method with best validation.

Wisconsin) and the Wikipedia animal page-page networks [RAS21] (Chameleon, Squirrel).

Most recently a "balanced" split was proposed by [CPL21]. This is a class-balanced split where, *for each class in a network*, a 60%-20%-20% mask is made with then each class mask being collected into a final, aggregate train-validation-test split. Both the balanced split and the datasets tested in Section 3 are commonplace benchmarking practices for current SSNC papers [LHL22, WZ22].

## 4.1 Comparing Split Performances

Provided in Figures 4.4-4.5 are visualizations on the impacts of different evaluation techniques on simple linear models ($\boldsymbol{XW}$ and $\boldsymbol{AXW}$). To keep things comparable to the sparse split, where no validation set exists, both the learning rate ($10^{-3}$) and the weight decay (0.0) were set to be fixed for the Adam optimizer. Despite this lack of tuning, the best of these models, per dataset, achieve roughly >85% relative performance when compared to SOTA SSNC methods. The high-end of this performance can be seen in the Squirrel column of Figure 4.5, where mean accuracy of the best linear model is 77.3%.

New GNN architectures which make use of the recent, balanced split may also experience an analogous performance bump relative to any older models tested before the split was introduced. In the worst case, this may lead to an overstatement in new modeling contributions and has the potential downside of muddying the signal of what makes for a successful and efficient GNN architecture in SSNC experiments. For this reason, we believe it is important to be clear on the impact of splitting conventions and how they contribute to recent performance upticks in SSNC benchmarking.

## 5 Conclusions

We have shown how classically-inspired, non-parametric techniques can be used to match, and sometimes exceed, previous spectral and non-linear GNN approaches. Our methods make no use of post-model augmentations, such as dropout [SHK14] or batchnorm [IS15], and allow for a clean theoretical analysis in future work.

Empirically, we explored and ablated pertinent hyperparameters to the spectral kernel model and have shown the various dependences between parameters across different datasets. On the aspect of low-rank kernel models, we have shown how spectral truncation can homogenize response outcomes for different kernel choices. Additionally for low-rank models, we have shown how performance decline is gradual with increases in spectral truncation, pointing to practical speed-ups for non-parametric kernel implementations.

On the aspect of testing conventions, we looked at how evaluation has changed for SSNC tasks since the first introduction of popular citation datasets [SNB08]. We have shown how the class-balanced split can produce improvements in performance outside of what is expected by uncertainty.

In summary, non-parametric kernel aggregators provide a simple yet effective means of recovering unobserved labels in SSNC tasks. As our implementations are free from post-model augmentations, we expect future theoretical insights obtained for low rank kernel

aggregators to be closely reproduced in experimental settings such as those seen in Section 3. Future work may further develop these insights, adding to the list of favorable properties for non-parametric kernel aggregators.

# CHAPTER 5

# Sharp Bounds for Poly-GNNs and the Effect of Graph Noise

## 1 Introduction

Graph neural networks (GNNs), like other deep learning models, have been shown to be best-in-class in empirical performance relative to standard kernel methods [YCS16, KW17]. Choices of architecture, non-linearity, and most importantly depth, are major determinants of GNN performance. However, outside of an empirically-based selection, there is little in the way of theoretical understanding as to why one choice of parameter may work better than another.

A parameter of particular interest is the depth of the network. A central dogma of deep learning is that deeper networks are better. They are easier to optimize and achieve better performance than their shallow counterparts [KSH12]. However, in the case of graph machine learning, this is not always the case. The phenomenon, known as GNN *oversmoothing*, is well-documented [RBM23] and represents a departure from commonly held deep learning beliefs.

In this paper, we study the theoretical implications of GNN depth in a common graph learning task, namely, the semi-supervised node classification (SSNC). We consider community-structured graphs where both the graph structure and the node features are allowed to be noisy. In our study, we derive exact rates of misclassification and add nuance to the discussion of GNN oversmoothing. Importantly, we show that there is a fundamental misclassification

rate which is available to all GNNs with polynomial features. Furthermore, this rate is sharp and invariant to network depth for sufficiently large graph inputs.

### 1.1 SSNC with GNNs

In the task of SSNC, one is given a graph, often in the form of an adjacency matrix $A \in \{0,1\}^{n\times n}$, and is asked to make predictions using a partially observed set of labels. More formally, each node $i$ has a feature vector $x_i \in \mathbb{R}^d$ and a label $y_i \in [L] := \{1,\dots,L\}$ determining its class. We observe the graph, $A$, all the node features, collected in a matrix $X \in \mathbb{R}^{n\times d}$ where the $i$th row is $x_i^T$, and a subset of the node labels $y_i$, $i \in \mathcal{O} \subset [n]$. The goal is to predict the unseen labels $y_i, i \in \mathcal{O}^c$.

The prototypical GNN design is defined layer-wise where, for $Z^{(0)} = X$, intermediate feature $Z^{(\ell+1)}$ are expressed as

$$Z^{(\ell+1)} = \varphi(AZ^{(\ell)}W^{(\ell)}). \tag{5.1}$$

Here, $\ell = 0, 1, \dots, k-1$ denotes the layer index, $\varphi : \mathbb{R} \to \mathbb{R}$ is a non-linear function applied elementwise, and $W^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ is the weight matrix for layer $\ell$. The rows of $Z^{(\ell)} \in \mathbb{R}^{n\times d_\ell}$ are the (latent) node representations produced by GNN at layer $\ell$. One hopes that, by adding more layers and repeatedly aggregating over the graph, the final representation $Z^{(k)}$ will be more informative compared to the initial node features $Z^{(0)} = X$.

In practice, one may replace $A$ with other graph aggregation operators $P$, such as the Laplacian matrix (in all its variants), and even change the aggregation operator per layer $P^{(\ell)}$, such as in the case of graph attention networks [VCC18]. Similarly, the nonlinearity can be varied layerwise $\varphi^{(\ell)}$; nonetheless we will focus on the form given in (5.1).

Critically, recent empirical work [WZ22] have suggested that one can drop the nonlinearity $\varphi$, essentially replacing it with the identity function without noticeable change in performance on various SSNC benchmarks. Taking $\varphi$ to be the identity map, we can recursively unravel the layers to obtain a simple form for $Z^{(k)} = A^k X W^{(0)} \cdots W^{(k-1)}$. Reparametrizing the

product of weight matrices into a single weight matrix $W$, we obtain

$$Z^{(k)} = A^k X W,$$

the basic polynomial (or rather monomial) GNN which is the object of study in this paper, which we call the *poly-GNN*. Training a classifier for the poly-GNN amounts to first forming the graph-aggregated features $A^k X$ and then training a linear classifier on the observed pairs $\big((A^k X)_{i*}, y_i\big), i \in \mathcal{O}$, where $(\cdot)_{i*}$ denotes the operator that extracts the $i$th row of a matrix.

Given this framing, our interest naturally lies in the predictive ability of graph-aggregated features $\phi^{(k)} \coloneqq A^k X \in \mathbb{R}^{n \times d}$. In this case, the pivotal quantity to explain performance is the feature signal-to-noise ratio (SNR) for $\phi^{(k)}$:

$$\frac{1}{\rho^{(k)}} \coloneqq \min_{i,j:\, y_i \neq y_j} \frac{\|\mathbb{E}[\phi_i^{(k)}] - \mathbb{E}[\phi_j^{(k)}]\|_2}{\left(\frac{1}{n}\sum_{i'} \|\phi_{i'}^{(k)} - \mathbb{E}[\phi_{i'}^{(k)}]\|_2^2\right)^{1/2}}. \tag{5.2}$$

where $\phi_i^{(k)}$ is the $i$th row of $\phi^{(k)}$ viewed as a column vector. Specific to misclassification, a feature SNR $(\rho^{(k)})^{-1}$ which increases with the sample size, $n$, is sufficient for the recovery of node labels $y$. Additionally, the feature SNR parameterizes the misclassification error, and is strictly more flexible than other related notions, such as linear separability.

### 1.2 CSBM and Noise Decompositions

A suitable theoretical model for SSNC and, by extension, for the aggregated features $\phi^{(k)}$, is the contextual stochastic block model (CSBM) [DSM18]. The CSBM is an extension of the stochastic block model (SBM) [HLL83] where latent labels $y = (y_i)_i$ determine both the distribution of the random edge $A_{ij}$ and that of the node feature $x_i$.

In particular, network data $(A, X)$ is said to be CSBM-generated if, for some cluster centers $\mu_1, \ldots, \mu_L \in \mathbb{R}^d$ and a connectivity matrix $B \in \mathbb{R}^{L \times L}$, the data follows

$$x_i \,|\, y_i \sim \mu_{y_i} + \varepsilon_i, \qquad \text{and} \qquad A_{ij} \,|\, y_i, y_j \sim \mathrm{Bern}(B_{y_i y_j}),$$

with zero-mean, sub-Gaussian noise vector $\varepsilon_i$. Originally, CSBM referred to the case where $B_{\ell\ell} = p$ and $B_{k\ell} = q$ for all $k \neq \ell$. We refer to this special case as $(p, q)$-CSBM.

The CSBM allows us to give a simple intuition for why graph aggregation in a GNN leads to better features for classification. Consider the matrix monomial $AX$, describing a first-order neighbor aggregation on the features $X$. For $X$ which is CSBM-generated (and more generally for any $X$ which contains additive noise), the first-order aggregation can be decomposed as

$$(AX)_{i*}^T = \sum_j A_{ij}\mu_{y_j} + \sum_j A_{ij}\varepsilon_j. \tag{5.3}$$

Consider an ideal CSBM where $B_{k\ell} = 0$ for $k \neq \ell$ and $B_{kk} > 0$ for all $k$. For simplicity, for this example, assume $L = 2$, $\mu_1 = 1$, $\mu_2 = -1$, and $\varepsilon_i \sim N(0, 1)$. Then, if node $i$ is in cluster 1, it will only be connected to nodes in cluster 1, hence the first term above is equal to $\deg(i)\,\mu_1 = \deg(i)$ where $\deg(i)$ is the degree of node $i$. On the other hand the second term is a sum of $\deg(i)$ independent zero mean variables, hence has standard deviation on the order of $\sqrt{\deg(i)}$. That is, the noise grows slower than the signal $\big(\sqrt{\deg(i)}$ versus $\deg(i)\big)$, improving the SNR after one round of aggregation.

The above observation is the principal idea familiar to statisticians that averaging reduces noise. It also suggests that the left and right terms in (5.3) are the signal and noise, respectively. However, this description is deceiving in a general CSBM where $B_{k\ell} > 0$ for $k \neq \ell$. In the general case, each $A_{ij}$ contains additional information about the generation process, both signal and noise, hence the first term in (5.3) carries noise as well.

**Noise Decomposition** Suppose that cluster centers $\{\mu_\ell\}$ were known. Then, there is an idealized aggregation operator $\mathbb{E}[A]$ that would maximally enhance the signal if it was available in place of $A$, that is, $\sum_j \mathbb{E}[A_{ij}]\mu_{y_j}$ rather than $\sum_j A_{ij}\mu_{y_j}$ should be considered the true signal. The price we pay in the signal for not knowing $\mathbb{E}[A]$ is $\sum_j (A_{ij} - \mathbb{E}[A_{ij}])\mu_{y_j}$ and can be referred to as the *graph noise* $\Delta$.

A mirror image of the above is obtained by examining feature $\mathbb{E}[A]x_j$, assuming we know $\mathbb{E}[A]$. Here the maximally enhanced version is obtained by replacing $x_j$ with the ideal center

$\mu_{y_j}$. The associated cost of not knowing $\mu_{y_j}$ is $\mathbb{E}[A]\varepsilon_j$ and will be referred to as *aggregated feature noise* $\Delta^{\varepsilon}$ or just feature noise for short.

These noise terms hold more generally for a $k$-hop aggregated feature where

$$\begin{aligned}\phi_i^{(k)} - \mathbb{E}[\phi_i^{(k)}] &= (A^k X)_{i*}^T - \mathbb{E}[(A^k X)_{i*}^T] \\ &= \sum_j \big((A^k)_{ij} - \mathbb{E}[A^k]_{ij}\big)\mu_{y_j} + \sum_j \mathbb{E}[A^k]_{ij}\varepsilon_j + \sum_j \big((A^k)_{ij} - \mathbb{E}[A^k]_{ij}\big)\varepsilon_j \\ &=: \sum_j \Delta_{ij}^{\mu} + \Delta_{ij}^{\varepsilon} + \widetilde{\Delta}_{ij}.\end{aligned}$$

For both $\mathbb{E}[A]$ and $\mu_{y_j}$ unknown, an additional noise interaction $\widetilde{\Delta}$ is introduced. This additional term can be absorbed as a graph noise, where $\Delta_{ij} \coloneqq \Delta_{ij}^{\mu} + \widetilde{\Delta}_{ij}$ now acts on the random centers $x_j = \mu_{y_j} + \varepsilon_j$.

**Walk Decomposition of Noise** Each $(A^k)_{ij}$ admits a linear decomposition $\sum_{w \in \mathcal{W}_k} A_w$ which pass on to similar decompositions for $\Delta$ and $\Delta^{\varepsilon}$. Here, $w = (i_\ell, i_{\ell+1})_\ell$ is a (directed) walk of length $k$ on the complete graph on $[n]$, and $\mathcal{W}_k$ is the set of such walks. The notation $A_w$ is shorthand for the product of edges along the walk, that is, $A_w = \prod_{\ell=1}^k A_{i_\ell i_{\ell+1}}$. The variance of walk product $A_w$ scales with the number of unique edges in $w = (i_\ell, i_{\ell+1})_\ell$. Since edge direction does not change the walk product, we can classify subgraphs of the complete graph on $[n]$ in terms of their contributions to noise terms $\Delta$ and $\Delta^{\varepsilon}$. In this sense, it is possible to collect walks $\mathcal{N}_\alpha \subset \mathcal{W}_k$ by the type $\alpha$ of their subgraph, and organize them according to their total noise contribution. We will show that for the problem at hand, there is a restricted set of walks, $\mathcal{N}_*$, which overwhelming contribute to the noise. This fact both streamlines our analysis and allows for exceedingly tight approximations of the noise.

### 1.3 Paper Overview

In this paper we provide a sharp analysis of the feature SNR $\left(\rho^{(k)}\right)^{-1}$. This requires providing upper and lowerbounds for the signal as well as bounding and disentangling contributions

from noise terms $\Delta$ and $\Delta^{\varepsilon}$.

**Technical Contributions** In our analysis, we make the following technical contributions:

- Introduce novel tools from matrix perturbation theory to generalize previous spectral structure arguments (i.e. circumventing Davis–Kahan).
- Define a higher-order notion of walks, *walk sequences*, which naturally arise from matrix moments.
- Provide a complete characterizations of dominant walk structures $\mathcal{N}_*$ for both graph and feature noise under general sparsity conditions.
- Provide general concentrations from moments for heavy-tailed sub-Weibull [VGN20] distributions.
- Provide a novel analysis connecting signal structure to a necessary lowerbound on graph noise for community structured graphs.

In Section 2, we provide an overview of the main result, along with its various components (separate signal and noise bounds), and a discussion of their implications. We also discuss connections to previous work at the end of this section. Section 3 gives a self-contained analysis of the signal component of the SNR, providing a streamlined treatment of the signal component using standard matrix concentration results and a matrix mean value theorem.

Section 4 provides a detailed walkthrough on the sources of the noise component of the SNR. Here, we bring tools from random matrix theory to bear on the analysis of GNNs. This section is by far the most technically involved, partly due to the use of general assumptions (see Section 2), and partly due to the natural difficulty of deriving high-probability results for collections of dependent quantities like $(\phi_i^{(k)})_i$. With that said, the payoff of our analysis is clear, as we can provide SNR rates that are tight in the sample size $n$ for any fixed or slow growing GNN depth $k$.

Together, Sections 3 and 4 provide the high-level proof of the main results. We have included the high-level argument in the main text since the proof provides more insights into the behavior of GNNs than what is reflected in the statement of the main results. An example is the subtle distinction between even and odd-layered GNNs in the effect of feature noise; see the discussion on dominant walks in Section 4.1. Another example is the characterization of the dominant walks for the graph noise in Section 4.5. We conclude the paper with discussion in Section 5.

**Notation** For a vector $x = (x_i) \in \mathbb{R}^d$ we write $\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$ for its $\ell_p$ norm. For a matrix $X \in \mathbb{R}^{n\times d}$, we write $\|A\|_{p\to q} = \max_{\|x\|_p\le 1} \|Ax\|_q$ for its $\ell_p/\ell_q$ operator norm. The $\ell_2/\ell_2$ operator norm is simply written as $\|A\|$. We write $\|A\|_{\max} = \max_{i,j} |A_{ij}|$. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there is a universal constant $C > 0$ such that $a_n \le Cb_n$. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. These notations will be used more generally, for any two expressions, to mean existence of the corresponding inequalities up to universal constants. We write $Z \sim \mathrm{SG}(\sigma)$ to denote a zero-mean sub-Gaussian random variable $Z$ with parameter $\sigma$, that is, $\mathbb{E}e^{\lambda Z} \le e^{\lambda^2\sigma^2/2}$ holds for all $\lambda \in \mathbb{R}$. The complete graph on nodes $[n] := \{1, \dots, n\}$ is denoted $K_n$.

## 2 Main result

Let $\mathcal{C}_\ell = \{j : y_j = \ell\}$ denote the set of indices corresponding to the $\ell$th class. Pick some $i \in \mathcal{C}_\ell$. Then, the ideal center of $\mathcal{C}_\ell$ can be defined as

$$\widetilde{\mu}_\ell^{(k)} := \mathbb{E}\phi_i^{(k)} = \sum_j \mathbb{E}[A^k]_{ij}\mu_{y_j} = \sum_{\ell'=1}^{L} \sum_{j\in\mathcal{C}_{\ell'}} \mathbb{E}[A^k]_{ij}\mu_{\ell'}. \tag{5.4}$$

By the community symmetry of CSBM, $\widetilde{\mu}_\ell^{(k)}$ is independent of which node $i \in \mathcal{C}_\ell$ was picked. As a consequence, the SNR in (5.2) simplifies to

$$\frac{1}{\rho^{(k)}} = \frac{\min_{\ell\neq\ell'} \|\widetilde{\mu}_\ell^{(k)} - \widetilde{\mu}_{\ell'}^{(k)}\|_2}{\left(\frac{1}{n}\sum_i \|\phi_i^{(k)} - \mathbb{E}[\phi_i^{(k)}]\|_2^2\right)^{1/2}} = \min_{\ell\neq\ell'} \frac{S(\ell,\ell')}{\overline{D}} \tag{5.5}$$

where

$$S(\ell,\ell') \coloneqq \|\widetilde{\mu}_\ell^{(k)} - \widetilde{\mu}_{\ell'}^{(k)}\|_2, \quad \overline{D} \coloneqq \big(\frac{1}{n}\sum_i \|\phi_i^{(k)} - \mathbb{E}[\phi_i^{(k)}]\|_2^2\big)^{1/2}. \tag{5.6}$$

Let us now summarize the assumptions we make. We start with the sparsity regime for CSBM. Let $p_{ij} = \mathbb{E}[A_{ij}]$, and

$$p_{\max} := \max_{i,j} p_{ij} = \max_{\ell,\ell'} B_{\ell\ell'}, \quad \text{and} \quad \nu_n := np_{\max}.$$

Parameter $\nu_n$ captures the sparsity of the graph. More precisely, we assume

(A1) For every $\ell \in [L]$, there is $\mathcal{I}_\ell \subseteq [L]$ with $|\mathcal{I}_\ell| \geq 1$ such that

$$nB_{\ell\ell'} \geq c_B\,\nu_n, \qquad \ell' \in \mathcal{I}_\ell, \tag{5.7}$$

$$nB_{\ell\ell'} \leq C_B\nu_n^{1-\delta}, \quad \ell' \notin \mathcal{I}_\ell, \tag{5.8}$$

for some constants $c_B, C_B > 0$ and $\delta \in (0,\infty]$.

On the first reading, one can take $\mathcal{I}_\ell = [L]$ for all $\ell$, so that (5.8) is vacuous (we can take $\delta = \infty$ in this case for subsequent results). This case corresponds to the most common setting in the literature where one assumes the entries of $B$ all grow at the same rate $\asymp \nu_n/n$, in which case $\nu_n$ roughly corresponds to the average degree and is a measure of graph sparsity. In particular, letting $\nu_n = o(n)$ leads to asymptotically sparse graphs. The general form of assumption (A1) significantly relaxes the standard setting above, by only requiring at least one entry of $B$ in each row to grow at the rate similar to $p_{\max} = \nu_n/n$ while the other entries are allowed to decay to zero faster.

Tacking on the following assumption prohibits degenerate cases where edge probabilities $p_{ij} \to 1$:

$$\text{(A2)} \quad \nu_n \leq (1-c_\nu)n \text{ for } c_\nu \in (0,1).$$

Let $\pi_\ell = |\mathcal{C}_\ell|/n$ and let $\pi = (\pi_1,\ldots,\pi_L)$ be the vector collecting the class proportions of the $L$-classes. We make the following assumption on class proportions:

$$\text{(A3)} \quad L\pi_\ell \geq c_\pi \text{ and } \sqrt{L}\|\pi\|_2 \leq C_\pi,$$

Assumption (A3) requires clusters $\mathcal{C}_\ell$ to be of similar order with $\pi_\ell \asymp 1/L$.

Next, consider the matrix $\mu$ whose $\ell$th column is the cluster mean for node features in class $\ell$:

$$\mu = [\mu_1, \mu_2, \ldots, \mu_L] \in \mathbb{R}^{d\times L} \tag{5.9}$$

where $\mu_\ell = \mathbb{E}[x_i]$ for $y_i = \ell$. We will need assumptions on the size of $\mu$ and its interaction with $B$. Consider the growth-normalized, $k$-aggregated population vectors

$$\bar{\xi}_\ell^{(k)} \coloneqq \mu \left(\Pi \cdot \frac{nB}{\nu_n}\right)^k e_\ell, \tag{5.10}$$

where $\Pi \coloneqq \operatorname{diag}(\pi) \in \mathbb{R}^{L\times L}$ is the diagonal matrix collecting class proportions. We assume

$$\text{(A4)} \quad \|\mu\| \leq C_\mu\sqrt{d}, \qquad \text{(A5)} \quad \|\bar{\xi}_\ell^{(k)} - \bar{\xi}_{\ell'}^{(k)}\|_2 \geq c_\xi\sqrt{d},$$

and, without loss of generality, take $c_\xi \leq 1$. We refer to $c_\xi$ as the separation factor of the graph. Note that $\bar{\xi}_\ell^{(k)}$ are growth-normalized, since $B_{\ell\ell'} \asymp \nu_n/n$ for $\ell, \ell' \in [L]$ by Assumption (A1). As a result, $c_\xi$ is free from scaling with $n$.

Assumption (A4) characterizes the growth of $\mu \in \mathbb{R}^{d\times L}$ to be $d$-dominant in operator norm. Alternatively, it can be considered the definition of constant $C_\mu$. Since we keep track of all the constants in the result, even if $C_\mu$ depends on $d$ and $L$, one can track its effect on the final bound. On the first pass, however, it is helpful to think of $C_\mu$ as constant, which is the case if, for example, the entries of $\mu$ are of order 1, the number of classes $L$ is kept fixed and the dimension $d$ grows.

Assumption (A5) is the key condition of the result. It is a constraint that prevents any two population means from becoming indistinguishable after $k$ smoothings. For a simple example where the condition is violated consider a balanced $(p, q)$-CSBM with $p = q$ (i.e., an Erdös-Rényi graph) with means $\mu = [1, -1]$. One should not be able to improve SNR by graph aggregation in this case.

## 2.1 Informal statement

Our main result shows that the SNR in $k$-hop aggregated features has strong invariance to the depth $k$ as $n$ grows:

**Theorem 9** (Informal)**.** *Let $(A, X)$ be generated from an $L$-class CSBM satisfying (A1)-(A5) with $\nu_n \gtrsim \log n$ and sufficiently large $n$. Then, for any $k \geq 1$, with high probability,*

$$\sqrt{\nu_n}\, \rho^{(k)} \leq C\, c_\xi^{-1} \tag{5.11}$$

*for a constant $C$ independent of $n$ and $k$. Furthermore, with probability bounded away from zero,*

$$\sqrt{\nu_n}\, \rho^{(k)} \geq c\, c_\xi \tag{5.12}$$

*for a constant $c > 0$ independent of $n$ and $k$.*

A more precise statement of the result can be found in Section 2.2 (Theorem 13). Theorem 9 states that there is a fundamental rate of separation, $\sqrt{\nu_n}$, which is the same for any $k$-hop feature, given that the sample size $n$ is sufficiently large. That is, increasing $k$ neither improves nor degrades the rate of separation. Furthermore, any $k$-dependence in the SNR must come from the composition of the separation factor $c_\xi$. In this way, we say that $\rho^{(k)}$ is *rate invariant* to the poly-GNN depth $k$.

Another consequence of Theorem 9 is that graph aggregation by GNN does indeed help, compared to classifying solely based on node features $X$, precisely at the relative rate $\sqrt{\nu_n}$, whenever $\nu_n$ grows with $n$.

Next, to illustrate the $k$-dependence found in $c_\xi$, consider a simple example where $\mu$ is full rank with $\sigma_\ell(\mu) \asymp \sqrt{d}$. More precisely, assume $c_\mu\sqrt{d} \leq \sigma_\ell(\mu) \leq C_\mu\sqrt{d}$. Set $\widetilde{B} \coloneqq \Pi(nB/\nu_n)$, and note that under our assumptions, $c_{\widetilde{B}} \leq \sigma_\ell(\widetilde{B}) \leq C_{\widetilde{B}}$ for constant $c_{\widetilde{B}}$ and $C_{\widetilde{B}}$ that potentially only depend on $L$. Then,

$$\|\bar{\xi}_\ell^{(k)} - \bar{\xi}_{\ell'}^{(k)}\|_2 \geq \sigma_{\min}(\mu)\big(\sigma_{\min}(\widetilde{B})\big)^k \geq c_\mu c_{\widetilde{B}}^k \sqrt{d}$$

and we can take $c_\xi = c_\mu c_{\widetilde{B}}^k$. Similarly for an upperbound one has

$$\|\bar{\xi}_\ell^{(k)} - \bar{\xi}_{\ell'}^{(k)}\|_2 \leq \sqrt{2}\,\|\mu\|\,\|\widetilde{B}^k\| \leq \sqrt{2d}\,C_\mu C_{\widetilde{B}}^k.$$

We see that $c_\xi \asymp c^k$ which, for pessimistic estimates $c < 1$, has a deflationary effect on the SNR. This potential deflationary effect between depth $k$ and classification accuracy is not surprising and matches the well-documented GNN oversmoothing found in the machine learning community. Perhaps more interesting is the following facts which can be gleaned from Theorem 9:

1. Oversmoothing is a scale effect that does not influence the SNR rate.

2. The rate optimal choice for SNR is obtained at $k = 1$.

These insights increase our understanding of GNNs and can inform future network architecture decisions for semi-supervised classification problems.

### 2.2 Formal statement

In order to state the precise version of Theorem 9, we first state a sequence of results on the upper and lower bounds of the two components of the SNR (signal and noise). Each bound requires a set of assumptions, with some more relaxed that the others. In fact, separating the assumptions reveals that the noise upper bound holds beyond CSBM, in the so-called general inhomogeneous Erdös-Rényi (IER) model. Given all the pieces, we then restate our main result in Theorem 13.

In our analysis of $\rho^{(k)} = \overline{D}/(\min_{\ell \neq \ell'} S(\ell, \ell'))$, we individually characterize the growth of all signal scales $S(\ell, \ell') = \|\widetilde{\mu}_\ell^{(k)} - \widetilde{\mu}_{\ell'}^{(k)}\|_2$ and the noise deviation $\overline{D} = \left(\frac{1}{n}\sum_i \|\phi_i^{(k)} - \mathbb{E}[\phi_i^{(k)}]\|_2^2\right)^{1/2}$. The high-level overview of the results are as follows:

1. The signal $S(\ell, \ell')$ grows precisely at the rate $\nu_n^k$ (Theorem 10).

2. The noise $\overline{D}$ grows precisely at the rate $\nu_n^{k-1/2}$ (Theorems 11 and 12).

Combining the two, the SNR grows at the precise rate $\nu_n^{1/2}$ independent of $k$.

**Signal bounds** To control the signal, consider the growth condition

$$\nu_n \;\geq\; \max\Big\{c'_\nu \log n,\, \frac{32LC_\mu^2 C_k^2}{c_\pi c_\xi^2}\Big\} \tag{5.13}$$

for some constant $c'_\nu > 0$, potentially different from $c_\nu$ in (A2), and $C_k = k2^k(C + \sqrt{(c/c_\nu)(k+1)})^k$ where $c, C > 0$ are some universal constants (see Lemma 5).

**Theorem 10** (Signal bounds)**.** *Assume (A3)–(A5) and growth condition* (5.13)*. Then, for* $\ell \neq \ell'$*,*

$$\frac{c_\xi}{2}\sqrt{d}\,\nu_n^k \;\leq\; S(\ell,\ell') \;\leq\; \sqrt{8d}\,C_\mu C_\pi^k\,\nu_n^k.$$

In addition to establishing a precise rate of $\nu_n^k$ for the signal growth, Theorem 10 shows that both the scale of $\mu$ and the cluster proportions $\pi_\ell$ affect the signal growth. These considerations, in addition to the cluster connectivity, are also captured in the lowerbound by the constant $c_\xi$. Note that Theorem 10 implies the following bound on $c_\xi \leq \sqrt{32}C_\mu C_\pi^k$. Similar bounds on $c_\xi$ with tighter universal constants can be obtained directly from the definition (A5). Theorem 10 is proved in Section 3.

**Noise bounds** Let $\kappa_{0,m} = 4\max\{\frac{C_1\sigma}{\|\mu_{m*}\|_\infty}, 1\}$, where $C_1$ is the universal constant in Lemma 9—controlling moment growth of sub-Gaussian variables. Set $\kappa_0 = \max_m \kappa_{0,m}$ and let

$$r_n(\epsilon) \coloneqq \max\big\{r \in 2\mathbb{N} :\, 3\big(\kappa_0 rke^k\big)^r \leq \nu_n^{1-\epsilon}\big\}, \tag{5.14}$$

for some $\epsilon \in (0,1)$. Let us also define constants

$$\kappa_1 = \frac{c_B c_\nu c_\pi c_\xi^2}{48L}, \quad \kappa_2 = 8\big(32\|\mu\|_{\max}^4 + (8\,C_1\sigma)^4\big), \quad \kappa_3 = \max\{8\,C_1\sigma, \|\mu\|_{\max}\}. \tag{5.15}$$

Consider the following growth conditions:

$$\min\Big\{\frac{n}{k \vee (4C_\mu/c_\xi)},\, C_B^{-1}\nu_n^\delta\Big\} \geq \frac{4C_\mu L}{c_\pi c_\xi}, \tag{5.16}$$

$$\min\Big\{(2k-1)^{-2}n,\, \nu_n^\epsilon\Big\} \geq \frac{\kappa_1}{2\|\mu\|_{\max}^2}. \tag{5.17}$$

Then we have the following control of the noise $\overline{D}$:

**Theorem 11** (Noise upper bound)**.** *Assume $\nu_n \geq ke^{2(k-1)}$ and $r_n(\epsilon) \geq 2$. Then for all real $r \in [2, r_n(\epsilon)]$,*

$$\mathbb{E}[|\overline{D}|^r] \leq \big(\kappa_3\sqrt{8dr}\,\nu_n^{k-1/2}\big)^r.$$

*Moreover for $u \geq 8de$,*

$$\mathbb{P}\big(\overline{D} \geq \kappa_3\nu_n^{k-1/2}\sqrt{u}\big) \leq \exp\Big(-\frac{1}{2}\min\Big\{\frac{u}{4de}, r_n(\epsilon)\Big\}\Big).$$

**Theorem 12** (Noise lower bound)**.** *Assume (A1)–(A3), (A5), the growth conditions* (5.16)–(5.17)*, and $r_n(\epsilon) \geq 4$ as defined in* (5.14)*. Then, for any $\eta \in (0,1)$,*

$$\mathbb{P}\Big(\,\overline{D} \geq \sqrt{\eta\,\kappa_1 d}\,\nu_n^{k-1/2}\,\Big) \geq (1-\eta)^2\frac{\kappa_1^2}{\kappa_2}.$$

Theorems 11 and 12 are proved in Section 4. Combining these bounds we can state our main result more precisely:

**Theorem 13** (Main result)**.** *Assume (A1)–(A5), growths conditions* (5.13), (5.16)*, and* (5.17)*, and $r_n(\epsilon) \geq 4$ as defined in* (5.14)*. Then, for any $\alpha \geq \sqrt{2}$, with probability at least $1 - \exp(-\frac{1}{2}\min\{\alpha^2, r_n(\epsilon)\})$, we have*

$$\sqrt{\nu_n}\,\rho^{(k)} \;\leq\; \sqrt{e}\alpha\Big(\frac{\kappa_3}{c_\xi}\Big) \tag{5.18}$$

*Moreover, for any $\eta \in (0,1)$, with probability at least $(1-\eta)^2\kappa_1^2/\kappa_2$, we have*

$$\sqrt{\nu_n}\,\rho^{(k)} \geq \sqrt{\frac{\eta}{8}} \cdot \frac{\sqrt{\kappa_1}}{C_\mu C_\pi^k}. \tag{5.19}$$

*Proof of Theorem 13.* Note that condition $\nu_n \geq ke^{2(k-1)}$ of Theorem 11 is automatically implied by $r_n(\epsilon) \geq 4$. Take $u = \alpha^2 4de$ for $\alpha^2 \geq 2$ in Theorem 11. Then, with probability at least $1 - \exp(-\frac{1}{2}\min\{\alpha^2, r_n\})$, we have $\overline{D} \leq \kappa_3\nu_n^{k-1/2}\sqrt{4de}\,\alpha$. Combined with the lower bound in Theorem 10, we obtain with the same probability

$$\rho^{(k)} \leq \frac{\kappa_3\nu_n^{k-1/2}\sqrt{4de}\,\alpha}{c_\xi\sqrt{d}\,\nu_n^k/2} = (\kappa_3/c_\xi)\sqrt{e}\alpha\,\nu_n^{-1/2}$$

which is the claimed upper bound. For the lower bound, it is enough to combine Theorem 12 with the lower bound in Theorem 10. □

A couple of comments are in order: For the upper bound (5.18) to truly hold with high probability, we must have $r_n(\epsilon) \to \infty$ as $n \to \infty$. This is the case when $\nu_n \to \infty$. In fact, one can show that, roughly $r_n(\epsilon) \gtrsim \log \nu_n / (\log \log \nu_n)$; see Lemma 30. The noise upper bound (Theorem 11) holds beyond CSBM, in a general IER model with $np_{ij} \leq \nu_n$. On the other hand, both the signal and noise lower bounds rely on the CSBM structure, as is evidenced by their dependence on parameter $c_\xi$ (via $\kappa_1$ in the case of the noise lower bound). One needs some form of structure for any lower bound to hold; this is clear in the case of the signal, but more subtle in the case of noise. The signal upper bound also relies on the CSBM structure.

As mentioned earlier, the binary nature of adjacency matrix $A$ allows noise $\overline{D}$ to be described in terms of walks on the complete graph $K_n$. As will be shown, walks which are tree-like, specifically star-like and path-like, have the largest contribution to the noise. For the aggregated feature noise ($\Delta^\varepsilon$ introduced in Section 1.2), the sparsity level $\nu_n$ influences the dominant walk type and the rate of growth, with a subtle distinction between the even and odd-layered GNNs; see Lemma 7 and the disscusion at the end of Section 4.1.

We suspect something similar may be true in the case of graph noise in general. As we show in Section 4.5, under structure guarantees like that of (A5), the dominant walk type for graph noise can be completely characterized. In the absence of such guarantees, the resulting dominant walk may change, and as a result, change $\nu_n^{k-1/2}$ noise growth rate.

### 2.3 Previous Work

The work of Baranwal et al. [BFJ21] is the first to our knowledge to explore the separation improvement in first-order aggregated features (i.e., $k = 1$) for CSBM data. Their results were obtained for a $(p, q)$-CSBM with a $\nu_n \gtrsim \log^2 n$ sparsity assumption. For the aggregation, a degree-normalized adjacency matrix with self-loops was used. In their paper, a $\sqrt{\nu_n}$

separation rate for the first-order aggregated features was recovered. This rate matches the fundamental separation rate shown in our main result. Our setting is more general, as it considers an $L$-class CSBM, relaxes the sparsity assumption to $\nu_n \gtrsim \log n$ and considers $k$-aggregated features for all $k \geq 1$. It is worth noting that the case $k \geq 2$ is technically much more challenging than $k = 1$ due to the dependence introduced by multi-hop aggregation. Furthermore, by focusing on the fundamental information content of $\phi^{(k)}$ and not necessarily just its linear separability, we are able to streamline the signal analysis by using simpler tools from matrix analysis [Bha97, BH16]

A similar work by Wu et al. [WCW23] explores the oversmoothing effect in features $\phi^{(k)} = A^k X$, assuming $X$ are normally distributed. A key claim in their work is that $\phi^{(k)}$ are exactly normal with distribution $\phi^{(k)} \sim \text{Gauss}(\mathbb{E}[\phi^{(k)}], \text{Var}(\phi^{(k)}))$. The authors claim this result follows from the linearity of the matrix $A^k$, however, this cannot be the case since even $\sum_j A_{ij} x_j$ is, by definition, a (scaled) mixture of Gaussians. Nevertheless, under the simplification that $\phi^{(k)} \approx \mathbb{E}[A]^k X$, the authors show that misclassification of GNNs can be described in terms of a Z-score of a standard normal. As we shall see, the approximation $\mathbb{E}[A^k] \approx \mathbb{E}[A]^k$ is not a bad one, especially when considering the overall size of $\|\mathbb{E}[A^k]\|_2$. However, the fact that $A^k$ is a matrix of dependent quantities complicates any high probability results for $\phi^{(k)}$.

Another work by Wei et al. [WYJ22] derives 1-hop MAP estimators, that is estimators which are locally optimal for a given neighborhood, for the case of a $(p, q)$-CSBM with normally distributed node covariates. The resulting estimator bears resemblance to a ReLU GNN utilizing a first-order aggregation scheme. In their main result, sparsity and mean separation are assumed to be $\nu_n \gtrsim \log^2 n$ and $\|\mu_1 - \mu_2\| \ll \log n / \sqrt{d}$ respectively. Additionally, this work has been recently extended by Baranwal et al. [BFJ23] to cover $\ell$-hop locally optimal MAP estimators for CSBMs satisfying sparsity $\nu_n \lesssim 1$. The $\mathcal{O}(1)$ sparsity constraint plays an important role in this analysis, as the networks generated from the CSBM become locally tree-like with high probability. This in turn makes the analysis more tractable for the fixed

hop case.

Our work in this paper was partly inspired by the empirical findings we report in [VA24a] where a simple single-layer GNN showed similar performance to more complicated state-of-the-art architectures on SSNC benchmarks. Within this same class of simple GNNs, the largest performance changes were observed when depth increased from $k = 0$ to $k = 1$.

On the topic of graph learning outside of GNNs, there is a locus of works which revolve around enforcing a Laplacian regularization to the traditional supervised learning context. This line of work traces its roots back to the manifold learning approach proposed by Belkin et al. [BNS06]. Recent works consider modifying the data fidelity term [LLZ19] or providing minimax rates for classes of non-parametric estimators with Laplacian regularization [GBT21, GBT23]. In the context of multi-graph regression, there are related works [ZM22] which consider regressing node-features with respect to multiple graph Laplacians.

## 3 Signal Analysis

In this section, we provide the analysis leading to the proof of Theorem 10 controlling the signal component of the SNR. The analysis is broken into several lemmas; the proofs are given in the text when short, otherwise deferred to the appendices.

We first introduce some notation. Let $Z \in \{0,1\}^{n \times L}$ be the cluster membership matrix for $y$, that is, $Z_{ij} = 1\{y_i = j\}$, and consider

$$P := ZBZ^T. \tag{5.20}$$

We note that $\mathbb{E}[A] = P - \mathrm{diag}(P)$ where $\mathrm{diag}(P)$ denotes the diagonal matrix with the same diagonal as $P$. We have subtracted $\mathrm{diag}(P)$, since we assume $A_{ii} = 0$ (no self-loops). Rewriting (5.4),

$$\widetilde{\mu}_\ell^{(k)} = \sum_{\ell'} \sum_j \mathbb{E}[A^k]_{ij} Z_{j\ell'} \mu_{\cdot \ell'} = \left(\mu Z^T \mathbb{E}[A^k]\right)_{*i} \tag{5.21}$$

for any $i \in \mathcal{C}_\ell$. Here, we have used the symmetry of $A$ and that $\mu$ is a matrix with columns $\mu_\ell = \mu_{*\ell}$. This implies that $\widetilde{\mu}_\ell^{(k)}$ is also the average, over $i \in \mathcal{C}_\ell$, of the RHS of (5.21), that is,

$$\widetilde{\mu}_\ell^{(k)} = \mu Z^T \mathbb{E}[A^k] \frac{\mathbb{1}_{\mathcal{C}_\ell}}{n_\ell} \tag{5.22}$$

where $\mathbb{1}_{\mathcal{C}_\ell} \in \{0,1\}^n$ denotes the indicator vector of cluster $\ell$, that is, $(\mathbb{1}_{\mathcal{C}_\ell})_i = 1\{i \in \mathcal{C}_\ell\}$.

### 3.1 Signal Proxy Growth

In showing $\|\widetilde{\mu}_\ell^{(k)} - \widetilde{\mu}_{\ell'}^{(k)}\|_2 \asymp \nu_n^k$, we first construct a proxy $\widetilde{S}(\ell,\ell')$ where $\widetilde{S}(\ell,\ell') \asymp \nu_n^k$. Motivated by the approximation $\mathbb{E}[A]^k \approx P^k$ in (5.22), let us write

$$\xi_\ell^{(k)} := \mu Z^T (ZBZ^T)^k \frac{\mathbb{1}_{\mathcal{C}_\ell}}{n_\ell} = MP^k \bar{\mathbb{1}}_{\mathcal{C}_\ell}. \tag{5.23}$$

where we have introduced

$$M \coloneqq \mu Z^T \in \mathbb{R}^{d\times n} \quad \text{and} \quad \bar{\mathbb{1}}_{\mathcal{C}_\ell} := \mathbb{1}_{\mathcal{C}_\ell}/n_\ell. \tag{5.24}$$

Next we need the following identity which can be proved by induction on $k$ (the proof is omitted):

**Lemma 1.** *$Z^T(ZBZ^T)^{k-1}ZB = (Z^TZB)^k$ for any $k \geq 1$.*

Using this lemma, each $\xi_\ell^{(k)}$ can be re-expressed into a simpler form

$$\begin{aligned}\xi_\ell^{(k)} &= \mu Z^T (ZBZ^T)^k Z e_\ell/n_\ell \\ &= \mu Z^T (ZBZ^T)^{k-1} ZB \cdot Z^T Z e_\ell/n_\ell = \mu (Z^TZB)^k e_\ell = \mu(\Pi \cdot nB)^k e_\ell\end{aligned}$$

where the third equality follows form $Z^TZe_\ell/n_\ell = e_\ell$ and Lemma 1, and the final equality from $\Pi = Z^TZ/n$. This allows for the following bracket bound for the growth of $\widetilde{S}(\ell,\ell')$:

**Lemma 2.** *Under assumptions (A3)–(A5), $c_\xi\sqrt{d}\,\nu_n^k \leq \widetilde{S}(\ell,\ell') \leq \sqrt{2d}\,C_\mu C_\pi^k \nu_n^k$.*

*Proof.* By the definition of $\nu_n$ and assumption (A3), we have

$$\begin{aligned}\|\Pi \cdot nB\| &\leq \|I_L\|_{2\to 1} \cdot \|\Pi \cdot nB\|_{1\to 2} \\ &\leq \sqrt{L} \cdot \max_{\ell'} \|(\Pi \cdot nB)_{\cdot \ell'}\|_2 \leq \sqrt{L} \cdot \nu_n \|\pi\|_2 \leq C_\pi \nu_n.\end{aligned}$$

Using assumptions (A3) and (A4), we obtain

$$\widetilde{S}(\ell, \ell') \;\leq\; \|\mu\| \, \|\Pi \cdot nB\|^k \, \|e_\ell - e_{\ell'}\|_2 \;\leq\; \sqrt{2d}\, C_\mu (C_\pi \nu_n)^k.$$

For a lowerbound, recalling definition (5.10), we note the identity

$$\xi_\ell^{(k)} = \nu_n^k \, \bar{\xi}_\ell^{(k)}, \tag{5.25}$$

which by assumption (A5) gives $\widetilde{S}(n,k) = \nu_n^k \, \|\bar{\xi}_\ell^{(k)} - \bar{\xi}_{\ell'}^{(k)}\|_2 \geq c_\xi \sqrt{d}\, \nu_n^k$. Altogether then, we have,

$$c_\xi \sqrt{d}\, \nu_n^k \;\leq\; \widetilde{S}(\ell, \ell') \;\leq\; \sqrt{2d}\, C_\mu C_\pi^k \, \nu_n^k.$$

which is the desired result. □

### 3.2 Signal Proxy as Leading Order Approximation

It remains to show that $\widetilde{S}(\ell, \ell')$ is indeed close to the signal deviation $\|\widetilde{\mu}_\ell^{(k)} - \widetilde{\mu}_{\ell'}^{(k)}\|_2$. We frequently use the following estimates:

**Lemma 3.** *Under (A3) and (A4),* $\|M\| \leq C_\mu \sqrt{nd}$ *and* $\|\bar{\mathbb{1}}_{\mathcal{C}_\ell}\| \leq (L/c_\pi)^{1/2} n^{-1/2}$.

*Proof.* We have $\|Z^T\| = \sqrt{\|Z^T Z\|} = \max_{\ell'} \sqrt{n_{\ell'}} \leq \sqrt{n}$, and the first claim follows from $\|M\| \leq \|\mu\| \|Z^T\|$ and (A4). Moreover, $\|\bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2 = n_\ell^{-1/2} = (\pi_\ell n)^{-1/2} \leq (L/c_\pi)^{1/2} n^{-1/2}$ by (A3). □

Using a Banach-valued variant to the mean-value theorem [Bha97], one has:

**Lemma 4.** $\|\mathbb{E}[A]^k - P^k\| \leq k \nu_n^k / n$.

*Proof.* Recall that $\mathbb{E}[A] = P - \operatorname{diag}(P)$ where $\|\mathbb{E}[A]\|$ and $\|P\|$ are upper-bounded by $np_{\max} = \nu_n$ and $\|\operatorname{diag}(P)\| \le p_{\max} = \nu_n/n$. Then, the second statement of Lemma 23 gives the desired bound. □

Using the same Banach-valued mean-value theorem and and sharp concentrations on $\|A - \mathbb{E}[A]\|$ [BH16], we get the following concentration inequality for $A^k$ which is proved in Appendix A:

**Lemma 5.** *Suppose that $\nu_n \ge c'_\nu \log n \ge 1$ for some constant $c'_\nu > 0$. Then, for any integer $k \ge 1$, the spectrum of $A^k$ concentrates as*

$$\mathbb{E}\|A^k - \mathbb{E}[A]^k\| \le C_k\, \nu_n^{k-1/2}.$$

*where $C_k = k2^k(C + \sqrt{(c/c'_\nu)(k+1)})^k$ for some universal constants $C > 1$ and $c > 0$.*

Next fix $\ell$ and $\ell'$, and let $w := \bar{\mathbb{1}}_{\mathcal{C}_\ell} - \bar{\mathbb{1}}_{\mathcal{C}_{\ell'}}$ Using (5.24) and (5.23),

$$\widetilde{\mu}_\ell^{(k)} - \widetilde{\mu}_{\ell'}^{(k)} = M\mathbb{E}[A^k]w \quad \text{and} \quad \xi_\ell^{(k)} - \xi_{\ell'}^{(k)} = MP^k w.$$

For $\ell \ne \ell'$, we have $\|w\|_2 = \sqrt{\|\bar{\mathbb{1}}_{\mathcal{C}_\ell}\|^2 + \|\bar{\mathbb{1}}_{\mathcal{C}_{\ell'}}\|^2} \le (2L/c_\pi)^{1/2} n^{-1/2}$ by Lemma 3. Moreover,

$$\begin{aligned}
\|\mathbb{E}[A^k] - P^k\| &\le \|\mathbb{E}[A^k] - \mathbb{E}[A]^k\| + \|\mathbb{E}[A]^k - P^k\| \\
&\le C_k\, \nu_n^{k-1/2} + (k\nu_n^k/n) \\
&\le 2C_k\, \nu_n^{k-1/2}
\end{aligned}$$

where the second line uses Lemmas 4 and 5 and the last line uses $C_k \nu_n^{-1/2} \ge k/n$ which is satisfied since $C_k \ge k$ and $\nu_n \le n$. We are now ready to prove Theorem 10.

*Proof of Theorem 10.* From our earlier results

$$\begin{aligned}
\left|\|\widetilde{\mu}_\ell^{(k)} - \widetilde{\mu}_{\ell'}^{(k)}\|_2 - \widetilde{S}(\ell,\ell')\right| &\le \|M(\mathbb{E}[A^k] - P^k)w\|_2 \\
&\le \|M\| \cdot \|\mathbb{E}[A^k] - P^k\| \cdot \|w\|_2 \\
&\le C_\mu \sqrt{nd} \cdot (2C_k\, \nu_n^{k-1/2}) \cdot (2L/c_\pi)^{1/2} n^{-1/2} \\
&\le \sqrt{8dL/c_\pi}\, C_\mu C_k\, \nu_n^{k-1/2}
\end{aligned}$$

using Lemmas 3 and 4 in the third line. Under the growth condition condition (5.13) we obtain $1 \geq c_\xi/2 \geq \sqrt{8L/c_\pi} C_\mu C_k \nu_n^{-1/2}$ and

$$\frac{c_\xi}{2}\sqrt{d}\,\nu_n^k \;\leq\; \|\widetilde{\mu}_\ell^{(k)} - \widetilde{\mu}_{\ell'}^{(k)}\|_2 \;\leq\; \sqrt{8d}\,C_\mu C_\pi^k\,\nu_n^k$$

which is the desired result. □

# 4 Noise Analysis

In this section, we develop probability bounds for the noise deviation

$$\overline{D} \coloneqq \Big(\frac{1}{n}\sum_i \|\phi_i^{(k)} - \mathbb{E}[\phi_i^{(k)}]\|_2^2\Big)^{1/2} = \Big(\frac{1}{n}\sum_{i,m} D_{im}^2\Big)^{1/2}$$

where $D_{im} \coloneqq \phi_{im}^{(k)} - \mathbb{E}\big[\phi_{im}^{(k)}\big]$, leading to the proofs of Theorems 11 and 12. These probability bounds will be obtained through a high-moment Markov bound, by analyzing the leading term of the moments of $\mathbb{E}(\overline{D})^r$ for $r \in 2\mathbb{N}$. For such $r$, the $r$th moment of the noise can be upperbound as

$$\mathbb{E}(\overline{D})^r \leq \frac{d^{r/2-1}}{n}\sum_{i,m}\mathbb{E}D_{im}^r, \tag{5.26}$$

where the right-hand side follows from Jensen inequality with expectation operator $\frac{1}{nd}\sum_{i,m}$. We further decompose the inner terms as

$$\begin{aligned} D_{im} &= \sum_j \big((A^k)_{ij} - \mathbb{E}[A^k]_{ij}\big)x_{jm} + \sum_j \mathbb{E}[A^k]_{ij}\varepsilon_{jm} \\ &=: \Delta_{im} + \Delta_{im}^\varepsilon \end{aligned}$$

We will control the moments $\mathbb{E}D_{im}^r$. For $r = 2$, we have $\mathbb{E}D_{im}^2 = \mathbb{E}\Delta_{im}^2 + \mathbb{E}(\Delta_{im}^\varepsilon)^2$ and, more generally for $r \in 2\mathbb{N}$, by the convexity of $x \mapsto x^r$,

$$D_{im}^r \leq 2^{r-1}\big(\Delta_{im}^r + (\Delta_{im}^\varepsilon)^r\big). \tag{5.27}$$

Let us first control $\mathbb{E}(\Delta_{im}^\varepsilon)^r$.

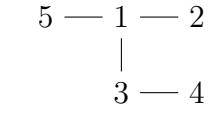
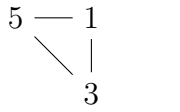
| | **Walk** $w$ | **Graph** $G(w)$ | $k$ | $t$ |
|---|---|---|---|---|
| $w_1$ | $5 \to 1 \to 2 \to 1 \to 3 \to 4 \to 3 \to 1 \to 5$ |  | 8 | 4 |
| $w_2$ | $5 \to 1 \to 3 \to 1 \to 5 \to 3 \to 5$ |  | 6 | 3 |

Figure 5.1: Walks and their corresponding graphs for $n \geq 5$.

### 4.1 Controlling Feature Noise

Recall that $\varepsilon_{im}$ are independent zero-mean sub-Gaussian random variables with parameter $\leq \sigma$; that is, $\varepsilon_{im} \sim \mathrm{SG}(\sigma)$. It follows that $\Delta^{\varepsilon}_{im} \sim \mathrm{SG}\big(\big(\sigma^2 \sum_j \mathbb{E}[A^k]^2_{ij}\big)^{1/2}\big)$. We can control $\mathbb{E}[A^k]_{ij}$ via a *walk analysis* which will be the common theme in Section 4. A more sophisticated version of such analysis appears in Section 4.2 where we control the graph noise.

Let us set up some notation and terminology. A $k$-walk on $[n]$ is a walk of length $k$ in the complete graph with nodes $[n]$. We represent a $k$-walk, $w$, as an ordered tuple of *directed* edges

$$w = ((i_1, i_2), (i_2, i_3), \ldots, (i_k, i_{k+1})) \tag{5.28}$$

We also denote the above walk as $i_1 \to i_2 \to \cdots \to i_k \to i_{k+1}$. For such a walk, we write $G(w)$ for the graph obtained by considering the nodes in $w$ and all the *undirected edges* present in $w$. We often denote the number of edges in $G(w)$ by $t$, which is the number of unique undirected edges in $w$. Figure 5.1 shows examples of walks and their corresponding graphs. The reason for considering the "undirected" graph of a walk is the symmetry of $A$. The undirected graph captures the truly independent entries of $A$ that appear in the walk.

Let $\mathcal{N}_t(i, j)$ be the set of $k$-walks going from $i$ to $j$ with $t$ unique undirected edges.

Representing $w \in \mathcal{N}_t(i,j)$ as in (5.28) with $i_1 = i$ and $i_{k+1} = j$, we have

$$\mathbb{E}[A^k]_{ij} = \sum_{t=1}^{k} \sum_{w \in \mathcal{N}_t(i,j)} \mathbb{E}\Big[\prod_{\ell=1}^{k} A_{i_\ell, i_{\ell+1}}\Big] \tag{5.29}$$

Necessary to our argument is the following counting lemmas on the number of $k$-walks:

**Lemma 6.** *$|\mathcal{N}_t(i,j)| \le \binom{n-2}{t-1} t^{k-1}$ for distinct $i, j \in [n]$.*

The case $i = j$ is more subtle. We partition $\mathcal{N}_t(i,i)$ into walks $w$ whose undirected graph $G(w)$ has loops, $\mathring{\mathcal{N}}_t(i,i)$, and walks for which $G(w)$ has no loops, $\breve{\mathcal{N}}_t(i,i)$. As example, consider $w_1$ and $w_2$ given in Figure 5.1 and note that $w_2 \in \mathring{\mathcal{N}}_t(i,i)$ while $w_1 \in \breve{\mathcal{N}}_t(i,i)$.

**Lemma 7.** *We have $|\mathring{\mathcal{N}}_t(i,i)| \le \binom{n-1}{t-1} t^{k-1}$ and*

$$|\breve{\mathcal{N}}_t(i,i)| \le C_t \binom{n-1}{t} \cdot t! \left\{ {k/2 \atop t} \right\} \cdot \mathbb{1}\big\{k \in 2\mathbb{N},\ t \le k/2\big\} \tag{5.30}$$

*where $C_t = \frac{1}{t+1}\binom{2t}{t}$ is the Catalan number and $\left\{ {m \atop t} \right\}$ is the Stirling number of the second kind. Bound* (5.30) *holds with equality when $k = 2t$. A further upper bound is*

$$|\breve{\mathcal{N}}_t(i,i)| \le (2e^2 n)^t \, t^{k/2-t-1}. \tag{5.31}$$

Note that these bounds imply that, for a given $t \le k/2$, the walks in $|\mathring{\mathcal{N}}_t(i,i)|$ have the fastest growth in $n$, of order $O(n^t)$, compared to walks in the other two categories whose growth is $O(n^{t-1})$. These lemmas allows us to bound adjacency moments $\mathbb{E}[A^k]$ elementwise:

**Lemma 8.** *Assume $\nu_n \ge k e^{2(k-1)}$, then*

$$\mathbb{E}[A^k]_{ij} \ \le \ 2\, p_{max} \nu_n^{k-1} + 2\nu_n^{k/2} \mathbb{1}\{i = j,\ k \in 2\mathbb{N}\}.$$

*Proof.* First, assume $i \ne j$. For $w = ((i_\ell, j_\ell))$ with $t$ unique edges, we have $\mathbb{E}[\prod_{\ell=1}^{k} A_{i_\ell j_\ell}] \le p_{\max}^t = (\nu_n/n)^t$ which gives

$$\mathbb{E}[A^k]_{ij} \ \le \ \sum_{t=1}^{k} |\mathcal{N}_t(i,j)| p_{\max}^t \ \le p_{\max} \sum_{t=1}^{k} \binom{n}{t-1} t^{k-1} p_{\max}^{t-1}$$

Using $\binom{n}{t-1} \leq (en/(t-1))^{t-1}$, and $(t/(t-1))^{t-1} \leq e$ for $t > 1$, we have $\binom{n}{t-1} \leq e \cdot (en/t)^{t-1}$. Plugging in and noting $np_{\max} = \nu_n$, we obtain $\mathbb{E}[A^k]_{ij} \leq ep_{\max} \sum_{t=1}^{k} (e\nu_n/t)^{t-1} t^{k-1}$. Further dividing both sides by $(e\nu_n)^{k-1}$ we have

$$\frac{\mathbb{E}[A^k]_{ij}}{(e\nu_n)^{k-1}} \leq ep_{\max} \sum_{t=1}^{k} (t/(e\nu_n))^{k-t}$$

Let $\rho = k/(e\nu_n)$. By assumption $ke^{k-1} \leq \nu_n$ so that $\rho \leq e^{-k} < 1/2$. Then, we have

$$\frac{\mathbb{E}[A^k]_{ij}}{p_{\max}\nu_n^{k-1}} \leq e^k \sum_{t=1}^{k} \rho^{k-t} \leq e^k \sum_{u=0}^{\infty} \rho^u \leq 2e^k \rho \leq 2.$$

which is the desired result.

Next for $\mathbb{E}[A^k]_{ii}$, we have the following bounds

$$\mathbb{E}[A^k]_{ii} \;\leq\; \sum_{t=1}^{k} |\mathcal{N}_t(i,i)| p_{\max}^t \;\leq\; \sum_{t=1}^{k} |\mathring{\mathcal{N}}_t(i,i)| p_{\max}^t + \sum_{t=1}^{k/2} |\breve{\mathcal{N}}_t(i,i)| p_{\max}^t.$$

The first sum bounds exactly as above. The second sum is zero unless $k \in 2\mathbb{N}$, which we assume for the rest of the argument. Let $c = 2e^2$ and note that by (5.31) of Lemma 7,

$$\begin{aligned}\sum_{t=1}^{k/2} |\breve{\mathcal{N}}_t(i,i)| p_{\max}^t \leq \sum_{t=1}^{k/2} (c\nu_n)^t \, t^{k/2-t} &= (c\nu_n)^{k/2} \sum_{t=1}^{k/2} (t/(c\nu_n))^{k/2-t} \\ &\leq (c\nu_n)^{k/2} \sum_{u=0}^{\infty} \rho^u \leq (c\nu_n)^{k/2} \cdot 2\rho \leq 2\nu_n^{k/2}\end{aligned}$$

where $\rho = (k/2)/(c\nu_n)$, $\sum_{u=0}^{\infty} \rho^u \leq 2\rho$ since $\rho < 1/2$, and

$$c^{k/2}\rho = (\sqrt{2}e)^k k/(4e^2\nu_n) \leq e^{2k}k/(e^2\nu_n) \leq 1$$

by assumption. The proof is complete. □

This style of walk argument will appear again and in more detail as we consider the network noise components $\Delta_{im}$. For the feature noise, we now need to translate the moment bound in Lemma 8 to a concentration bound. The following is well-known [Ver18, Proposition 2.5]:

**Lemma 9.** *If $Z$ is sub-Gaussian with parameter $\sigma$, then, $\mathbb{E}|Z|^r \leq (C_1\sigma r^{1/2})^r$ where $C_1$ is a numerical constant.*

The reverse is also true in the sense that a moment growth of the form above implies $Z$ is sub-Gaussian. This also follows from [Ver18, Proposition 2.5]. Alternatively, it follows from the high-probability bound provided in Appendix B of the Supplementary Material [VA24b] with $\eta = 1/2$. The bound in Lemma 8 gives (using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$):

$$\begin{aligned}\Big(\sum_j \mathbb{E}[A^k]_{ij}^2\Big)^{1/2} &\leq \mathbb{E}[A^k]_{ii} + \sqrt{n}\max_{j\neq i}\mathbb{E}[A^k]_{ij} \\ &\leq 2\nu_n^{k/2}\mathbb{1}\{k \in 2\mathbb{N}\} + 2p_{\max}^{1/2}\nu_n^{k-1/2} \\ &\leq 2\Big(\nu_n^{-k/2+1/2}\mathbb{1}\{k \in 2\mathbb{N}\} + p_{\max}^{1/2}\Big)\nu_n^{k-1/2} \leq 4\nu_n^{k-1/2} \end{aligned} \tag{5.32}$$

using $\nu_n \geq 1$ and $p_{\max} \leq 1$. Applying Lemma 9 gives the following

$$\mathbb{E}(\Delta_{im}^{\varepsilon})^r \leq \big(4C_1\sigma\nu_n^{k-1/2}r^{1/2}\big)^r \tag{5.33}$$

showing that $\Delta_{im}^{\varepsilon}$ is sub-Gaussian with parameter $\lesssim \sigma\nu_n^{k-1/2}$. Later in Section 4.4, we combine this with the bound on $\Delta_{im}$ to finish the proof of Theorem 11.

**On Dominant Walk Types** A careful review of Lemma 8 reveals that, when $\nu_n \gg 1$, there are two dominant walk types in the feature noise: the simple cycles / path graphs of length $k$ and the Dyck paths with $k/2$ edges. Out of all potential subgraphs constructed from a $k$-walk, these are the two subgraphs which aggregate, or "amplify," the feature noise the most.

For a walk type to contribute the most in expectation, its subgraph must have many configurations (for $k \ll n$ this roughly translates to maximizing the number of vertices in a graph) and it must limit the number of unique edges $t$ in its subgraph, otherwise any particular subgraph realization is less likely to appear. These two conditions lead to path and tree graphs to be the most natural contenders for subgraphs of a dominant walk type.

The feature noise dominant walk type has potentially alternate behavior in $k$, in that, for $k$ even, it is determined by the sparsity boundary $\nu_n \sim n^{1/k}$ (obtained by setting $\nu_n^{-k/2+1/2} \sim p_{\max}^{1/2}$ in (5.32)). This boundary is caused by the special nature of backtracking walks that can only appear when $k \in 2\mathbb{N}$. Backtracking walks limit the number of unique edges in their subgraphs at the cost of having fewer vertices. However, if the probability of making an edge is low enough, that is $\nu_n$ is small enough, then theses backtracking walks will be dominant with the largest category of the backtracking walks being the Dyck paths $\breve{\mathcal{N}}_{k/2}(i,i)$.

The previous argument, and much of our future analysis, hinges on the fact $\nu_n \gg 1$. In this case, subgraph multiplicity associated with increasing the number of edges $t$ can largely be discounted. In this regime, additional vertices add a factor of approximately $\nu_n/t$ to the noise where $t \leq k$ by the nature of our walks. For $\nu_n \sim 1$, it is no longer the case that adding a vertex uniformly increases the contribution across different walk types. By similar reasoning, one can see that the dominant walk type for $\nu_n \ll 1$ would simply the edge graph where $t = 1$.

### 4.2 Graph Noise and Walk Sequences

It remains to bound the graph noise, $\Delta_{im}$, for which we rely on a high-order notion of walks, *walk sequences* (or walk products), which, given the various independence properties of the adjacency matrix $A$ and the node features $X$, can be used to derive tight moment inequalities.

Let us revisit the $k$-walk $w$ as in (5.28). For such $w$, we write

$$A_w \coloneqq \prod_{\ell=1}^{k} A_{i_\ell, i_{\ell+1}} = \prod_{\{i_\ell, i_{\ell+1}\} \in [w]} A_{i_\ell, i_{\ell+1}}$$

where $[w] = \{\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_k, i_{k+1}\}\}$ is the set of unique *undirected* edges of $w$. The second equality follows since $A$ is a binary symmetric matrix, i.e., $A_{ij} \in \{0, 1\}$ and $A_{ij} = A_{ji}$. The number of unique undirected edges of $w$ is the cardinality of set $[w]$, denoted as $|[w]|$. Occasionally, we will need the set of unique vertices found in $w$ which we denote as

| | **Walk** $w$ | **Unique Edges** $[w]$ | **Nodes** $[\![w]\!]$ |
|---|---|---|---|
| $w_1$ | $5 \to 1 \to 2 \to 1 \to 3 \to 4 \to 3 \to 1 \to 5$ | $\{\{1,2\},\{1,3\},\{1,5\},\{3,4\}\}$ | $\{1,2,3,4,5\}$ |
| $w_2$ | $5 \to 1 \to 3 \to 1 \to 5 \to 3 \to 5$ | $\{\{1,3\},\{1,5\},\{3,5\}\}$ | $\{1,3,5\}$ |

Figure 5.2: Walks and their unique undirected edges and nodes

$[\![w]\!] = \{i_\ell\}_{\ell=1}^{k+1}$. Figure 5.2 illustrates $[w]$ and $[\![w]\!]$ for the two walks introudced in Figure 5.1. Note that, $[w]$ is the edge set of $G(w)$, while $[\![w]\!]$ is its vertex set.

Let $\mathcal{W} \coloneqq \mathcal{W}_k(i)$ be the set of $k$-walks which start at $i$, that is,

$$\mathcal{W} \coloneqq \mathcal{W}_k(i) = \{w \text{ as in (5.28) with } i_1 = i\}.$$

With the above notation, we have $\sum_j (A^k)_{ij} = \sum_{w\in\mathcal{W}} A_w$. Let $\mathfrak{p} : \mathcal{W} \to [n]$ be the projection giving the last vertex of a walk, that is, for $w$ as in (5.28), $\mathfrak{p}(w) = i_{k+1}$. Then

$$\Delta_{im} = \sum_{w\in\mathcal{W}} (A_w - \mathbb{E}[A_w])(x_{\mathfrak{p}(w)})_m.$$

Now, let $\boldsymbol{w} = (w_1, w_2, \ldots, w_r)$ be an ordered $r$-tuple of walks from $\mathcal{W}$. The set of such $r$-tuples is the $r$-fold Cartesian product $\mathcal{W}^r \coloneqq \bigotimes_{s=1}^r \mathcal{W}$. We refer to elements of $\mathcal{W}^r$ as *walk sequences*. Let us also write $(\cdot)^s$ for the coordinate projection of such $r$-tuples where $\boldsymbol{w}^s = w_s$ for $s \in [r]$. In the case of multiple coordinate projection with coordinates $S = \{s_1, s_2, \ldots, s_m\} \subseteq [r]$, we preserve the tuple ordering of $\boldsymbol{w}$ such that the corresponding projection $\boldsymbol{w}^S$ satisfies

$$s_1 < s_2 < \cdots < s_m \implies \boldsymbol{w}^S \coloneqq (w_{s_1}, w_{s_2}, \ldots, w_{s_m}).$$

The set of unique undirected edges and vertices in a walk sequence $\boldsymbol{w}$ can be computed as

$$[\boldsymbol{w}^S] = \bigcup_{s\in S} [w_s], \qquad [\![\boldsymbol{w}^S]\!] = \bigcup_{s\in S} [\![w_s]\!]$$

| | Walk Sequence $\boldsymbol{w}$ |
|---|---|
| $w_1$ | $1 \to 2 \to 3 \to 5$ |
| $w_2$ | $1 \to 2 \to 4 \to 6$ |
| $w_3$ | $1 \to 7 \to 9 \to 11$ |
| $w_4$ | $1 \to 7 \to 8 \to 10$ |
| $w_5$ | $1 \to 12 \to 13 \to 15$ |
| $w_6$ | $1 \to 12 \to 14 \to 16$ |

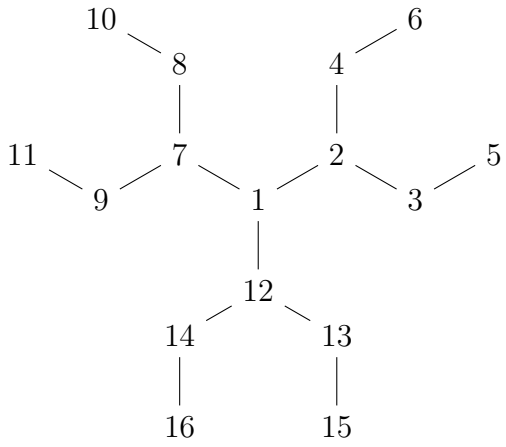



Figure 5.3: A walk sequence $\boldsymbol{w}$ with $r = 6$ components of length $k = 3$ each, and its corresponding undirected graph $G(\boldsymbol{w})$. This walk sequence belongs to $\mathcal{N}_{r,t,v}$ defined in (5.35) with $r = 6$, $t = |[\boldsymbol{w}]| = 15$ (number of unique edges) and $v = |[\![\boldsymbol{w}]\!]| = 16$ (number of unique vertices).

where, by convention, $[\boldsymbol{w}] \coloneqq [\boldsymbol{w}^{[r]}]$ and $[\![\boldsymbol{w}]\!] \coloneqq [\![\boldsymbol{w}^{[r]}]\!]$. Similar to the case of a single walk, we write $G(\boldsymbol{w})$ for the undirected graph associated with the walk sequence $\boldsymbol{w}$, that is, the graph with vertex set $[\![\boldsymbol{w}]\!]$ and edge set $[\boldsymbol{w}]$. Figure 5.3 shows an example of a walk sequence with $r = 6$ components each of length $k = 3$, together with its undirected graph $G(\boldsymbol{w})$.

With these notations, we have

$$\Delta_{im}^r = \sum_{\boldsymbol{w} \in \mathcal{W}^r} \prod_{s=1}^{r} (A_{\boldsymbol{w}^s} - \mathbb{E}[A_{\boldsymbol{w}^s}])(x_{\mathfrak{p}(\boldsymbol{w}^s)})_m.$$

Finally, we write

$$\varrho_1(\boldsymbol{w}) \coloneqq \mathbb{E}\Big[\prod_{s=1}^{r} (A_{\boldsymbol{w}^s} - \mathbb{E}[A_{\boldsymbol{w}^s}])\Big], \quad \varrho_2(\boldsymbol{w}) \coloneqq \mathbb{E}\Big[\prod_{s=1}^{r} (x_{\mathfrak{p}(\boldsymbol{w}^s)})_m\Big] \tag{5.34}$$

and $\varrho(\boldsymbol{w}) \coloneqq \varrho_1(\boldsymbol{w})\varrho_2(\boldsymbol{w})$. Note that we are suppressing the dependence of $\varrho(\boldsymbol{w})$ on $i$ and $m$ for simplicity. By independence of $A$ and $x$,

$$\mathbb{E}[\Delta_{im}^r] = \sum_{\boldsymbol{w} \in \mathcal{W}^r} \varrho(\boldsymbol{w}).$$

For a walk sequence $\boldsymbol{w} \in \mathcal{W}^r$, let us write $A_{\boldsymbol{w}} \coloneqq \prod_{s=1}^{r} A_{\boldsymbol{w}^s}$ and similarly $A_{\boldsymbol{w}^S} = \prod_{s \in S} A_{\boldsymbol{w}^s}$. We have the following control of $\varrho(\boldsymbol{w})$:

**Lemma 10.** $|\varrho_1(\boldsymbol{w})| \leq 2^r \mathbb{E}[A_{\boldsymbol{w}}] \leq 2^r p_{\max}^{|[\boldsymbol{w}]|}$ *and* $|\varrho_2(\boldsymbol{w})| \leq (2 \max\{C_1 \sigma r^{1/2}, \|\mu_{m*}\|_\infty\})^r$.

### 4.3 Leading Order Walk Decomposition

For the walk analysis, we are interested in walk types which contribute most to $\mathbb{E}[\Delta_{im}^r]$. To this end, we make use of the partial centering found in the walk products $\varrho_1(\boldsymbol{w})$. To accomplish this, we partition walks $\boldsymbol{w}^s$ within a walk sequence $\boldsymbol{w}$ according to their overlapped edges. This partition approach is similar to the edge partitions introduced in [EKY13], where it was necessary to control the moments of a related deviation term, $\frac{1}{n}\mathbb{E}\big[\mathbb{1}_{[n]}^T (A - \mathbb{E}[A])^r \mathbb{1}_{[n]}\big]$.

For every walk sequence $\boldsymbol{w} \in \mathcal{W}^r$, we define a partition $\Gamma(\boldsymbol{w})$ on $[r]$ by declaring $s, s' \in [r]$ to be equivalent if and only if $\boldsymbol{w}^s$ and $\boldsymbol{w}^{s'}$ share an undirected edge, that is, $[\boldsymbol{w}^s] \cap [\boldsymbol{w}^{s'}] \neq \varnothing$. Indexing the equivalences classes in $\Gamma(\boldsymbol{w})$ as $\Gamma_q(\boldsymbol{w}) \subseteq [r]$, we have $[r] = \bigsqcup_q \Gamma_q(\boldsymbol{w})$. We will refer to $\Gamma(\boldsymbol{w})$ as the $[r]$-partition of $\boldsymbol{w}$. We say $\boldsymbol{w} \in \mathcal{W}^r$ is *non-overlapping* if $|\Gamma_q(\boldsymbol{w})| = 1$ for some $q$, otherwise, it is *overlapping*. We occasionally treat $\Gamma(\boldsymbol{w})$ as an ordered tuple, by ordering the partition components $\Gamma_q(\boldsymbol{w})$ according to their smallest element, as in the example below:

**Example 4.** Walk sequence $\boldsymbol{w}$ in Figure 5.3 is overlapping with $[6]$-partition $\Gamma(\boldsymbol{w}) = (\{1,2\},\{3,4\},\{5,6\})$, meaning that walks, e.g., $\boldsymbol{w}^3$ and $\boldsymbol{w}^4$ share at least an edge (in this case $\{1,7\}$), while walks $\boldsymbol{w}^3$ and $\boldsymbol{w}^6$ share no edge, and so on. We can refer to the second component of the partition which is $\Gamma_2(\boldsymbol{w}) = \{3,4\}$, since $\{3,4\}$ has the second smallest element among the three components.

Given the independence of $A$ and $X$, we only need to consider overlapping $\boldsymbol{w}$, for which $|\Gamma_q(\boldsymbol{w})| \geq 2$ for all $q$:

**Lemma 11.** $\varrho(\boldsymbol{w}) = 0$ *if* $\boldsymbol{w}$ *is non-overlapping.*

Seen simply, Lemma 11 is stating that any walk $\boldsymbol{w}^s$ which does not overlap with any other walks in the walk sequence $\boldsymbol{w}$ will factor out and evaluate to zero in our moment calculation of $\mathbb{E}[\Delta_{im}]$. For overlapping walk sequences, we have the following control:

**Lemma 12.** *Let $\Gamma(\boldsymbol{w}) = \{\Gamma_q\}_{q=1}^{Q}$ for some overlapping $\boldsymbol{w}$. Then, $Q \le \lfloor r/2 \rfloor$ and*

$$|[\boldsymbol{w}^{\Gamma_q}]| \le |\Gamma_q|(k-1)+1, \qquad \forall q \in [Q]$$

*and hence $|[\boldsymbol{w}]| \le rk - \lceil r/2 \rceil$.*

To quantify the growth of $\mathbb{E}[\Delta_{im}^r]$, we partition overlapping walk sequences based on their number of unique edges, $t$, and unique vertices, $v$:

$$\mathcal{N}_{r,t,v} \coloneqq \big\{\ \boldsymbol{w} \in \mathcal{W}^r : |[\boldsymbol{w}]| = t,\ |[\![\boldsymbol{w}]\!]| = v,\ \boldsymbol{w} \text{ is overlapping } \big\}, \tag{5.35}$$

The following lemma bounds the size of $\mathcal{N}_{r,t,v}$:

**Lemma 13.** *We have $|\mathcal{N}_{r,t,v}| \le (v-1)^{rk}\binom{n-1}{v-1}$ and hence for $b \le t$, $\sum_{v=1}^{b+1} |\mathcal{N}_{r,t,v}| \le b^{rk-b}(en)^b$.*

The overlapping and rooted nature of the walks $\boldsymbol{w}^s$ give us a few consequences. Let

$$t_* \coloneqq r(k-1/2). \tag{5.36}$$

Then, by Lemma 12, for $r \in 2\mathbb{N}$, we have $\mathcal{N}_{r,t,v} = \varnothing$ for $t > t_*$, that is, $t_*$ is the maximum number of unique edges for an overlapping $r$-sequence of $k$-walks. The walk sequence in Figure 5.3, for example, achieves this maximum since $t = t_* = 6(3-1/2) = 15$ in this case. Additionally since each walk in the walk sequence $\boldsymbol{w}$ starts at $i$, the unique edges found in $\boldsymbol{w}$ correspond to a connected graph, meaning $\mathcal{N}_{r,t,v} = \varnothing$ for $v > t+1$.

Consequently, the $r$th moment of $\Delta_{im}$ can be decomposed as

$$\mathbb{E}[\Delta_{im}^r] = \sum_{t=1}^{t_*}\sum_{v=1}^{t+1}\sum_{\boldsymbol{w}\in\mathcal{N}_{r,t,v}} \varrho(\boldsymbol{w}) = T_{im}^{\mathrm{hi}}(r) + T_{im}^{\mathrm{lo}}(r)$$

where

$$T_{im}^{\mathrm{hi}}(r) \coloneqq \sum_{\boldsymbol{w}\in\mathcal{N}_{r,t_*,(t_*+1)}} \varrho(\boldsymbol{w}), \qquad T_{im}^{\mathrm{lo}}(r) \coloneqq \sum_{t=1}^{t_*}\sum_{v=1}^{t+1}\sum_{\boldsymbol{w}\in\mathcal{N}_{r,t,v}} \varrho(\boldsymbol{w})\, 1\{v \neq t^*+1\}. \tag{5.37}$$

When $i, m$ and $r$ are fixed, we often drop the dependence on them and simply write $T^{\mathrm{hi}}$ and $T^{\mathrm{lo}}$. For brevity, we also write

$$\mathcal{N}_* \coloneqq \mathcal{N}_{r,t_*,(t_*+1)}.$$

Terms $T^{\mathrm{hi}}$ and $T^{\mathrm{lo}}$ can be upper-bounded using the moment contributions $\varrho(\boldsymbol{w})$ of walk sequences $\boldsymbol{w}$ which have the largest number of unique vertices $v$. This leads to the following key result, by combining Lemmas 10 and 13:

**Lemma 14.** *Let $\kappa_{0,m} = 4\max\{\frac{C_1\sigma}{\|\mu_{m*}\|_\infty}, 1\}$ where $C_1$ is the constant in Lemma 9 and $\kappa_0 = \max_m \kappa_{0,m}$. Then for any even $r$ such that $\kappa_0^r\, t_*^r e^{t_*} \le \frac{1}{3}\nu_n^{1-\epsilon}$, we have*

$$|T_{im}^{hi}(r)| \;\le\; (\sqrt{r}\,\|\mu_{m*}\|_\infty)^r\, \nu_n^{t_*}, \tag{5.38}$$

$$|T_{im}^{lo}(r)| \;\le\; (\sqrt{r}\,\|\mu_{m*}\|_\infty)^r \cdot \nu_n^{t_*-\epsilon}. \tag{5.39}$$

Note that since $t_* < rk$, the condition of Lemma 14 is satisfied for all $r \le r_n(\epsilon)$ as defined in (5.14). We are now ready to prove the high-probability noise upperbound for $\overline{D}$.

### 4.4 Proof of Theorem 11

Let $r \in 2\mathbb{N}$. Combining (5.26) and (5.27), we have

$$\mathbb{E}(\overline{D})^r \le \frac{d^{r/2-1}}{n}\sum_{i,m} 2^{r-1}\big(\mathbb{E}[\Delta_{im}^r] + \mathbb{E}(\Delta_{im}^\varepsilon)^r\big).$$

Bounding $\|\mu_{m*}\|_\infty \le \|\mu\|_{\max}$, for the first term, we obtain

$$\mathbb{E}[\Delta_{im}^r] \;\le\; |T_{im}^{\mathrm{lo}}(r)| + |T_{im}^{\mathrm{hi}}(r)| \;\le\; 2(\sqrt{r}\,\|\mu\|_{\max})^r\, \nu_n^{t_*} \;=\; 2(\sqrt{r}\,\|\mu\|_{\max}\nu_n^{k-1/2})^r$$

using $t_* = rk - r/2$. For the second term, (5.33) gives

$$\mathbb{E}(\Delta_{im}^\varepsilon)^r \le \big(4C_1\sigma\sqrt{rp_{\max}}\nu_n^{k-1/2}\big)^r$$

Let $\kappa_3 = \max\{4C_1\sigma, \|\mu\|_{\max}\}$. Then, $\big(\mathbb{E}[\Delta_{im}^r] + \mathbb{E}(\Delta_{im}^{\varepsilon})^r\big) \le 3(\kappa_3\sqrt{r}\nu_n^{k-1/2})^r$. It follows that

$$\mathbb{E}(\overline{D})^r \le \frac{3}{2}d^{r/2}(\kappa_3\sqrt{r}\nu_n^{k-1/2})^r \le (\kappa_3\sqrt{2dr}\nu_n^{k-1/2})^r$$

using $3/2 \le (\sqrt{2})^r$ for all even $r \le r_n$. Applying Lemma 25 from Appendix B of the Supplementary Material with $K = \kappa_3\nu_n^{k-1/2}$, $\eta = 1/2$, $C = 4d$, the result follows.

### 4.5 Characterizing $\mathcal{N}_*$

The rest of Section 4 is devoted to proving the noise lowerbound (Theorem 12). To obtain a sharp lowerbound, we construct a sufficiently tight proxy $\widetilde{T}^{\text{hi}}$ to $T^{\text{hi}}$ (Section 4.6), one that satisfies (Section 4.7)

$$|T^{\text{hi}} - \widetilde{T}^{\text{hi}}| \lesssim \nu_n^{t_*-1}.$$

That is, the discrepancy between $\widetilde{T}^{\text{hi}}$ and $T^{\text{hi}}$ is of a lower order than $\nu_n^{t_*}$, which we know is the upper bound on $T^{\text{hi}}$ form (5.38). Then, we establish a lower bound on $\widetilde{T}^{\text{hi}}$ of the order $\nu_n^{t_*}$ (Section 4.8), which by the (reverse) triangle inequality implies the same lower bound on $T^{\text{hi}}$ up to constants, establishing that $T^{\text{hi}}$ is indeed tightly concentrated from above and below around $\nu_n^{t_*}$. This proxy approach is similar to the one used in Section 3.

To execute the above plan, we first investigate the structure of walk sequences in $\mathcal{N}_*$. We refer to the element of $\mathcal{N}_*$ as *maximal walk sequences*. Let $G(\boldsymbol{w}) = (\llbracket \boldsymbol{w} \rrbracket, [\boldsymbol{w}])$ be the undirected graph associated with the walk sequence $\boldsymbol{w}$. The following result provides a complete characterization of $G(\boldsymbol{w})$ for walks sequences in $\mathcal{N}_*$ as well as the associated $[r]$-partition $\Gamma(\boldsymbol{w})$. An $i$-rooted tree, is a rooted tree with root node $i$. Complete graph on $[r]$ is denoted $K_r$.

**Lemma 15** (Structure of $\mathcal{N}_*$)**.** *Let $\boldsymbol{w} \in \mathcal{N}_*$ and $\Gamma(\boldsymbol{w}) = \{\Gamma_q\}_{q=1}^Q$. Then, the following hold:*

*(a) $\Gamma(\boldsymbol{w})$ is a perfect matching on $[n]$. That is, $|\Gamma_q| = 2$ for all $q$, hence $Q = r/2$.*

*(b) $G(\boldsymbol{w})$ is an $i$-rooted tree.*

*(c)* $G(\boldsymbol{w}^{\Gamma_q}), q \in [Q]$ *are* $i$*-rooted subtrees of* $G(\boldsymbol{w})$*; they are vertex disjoint except at the root.*

*(d) For* $\Gamma_q = \{s, s'\}$*,* $G(\boldsymbol{w}^{\Gamma_q})$ *consists of two* $i$*-rooted subtrees* $G(\boldsymbol{w}^s)$ *and* $G(\boldsymbol{w}^{s'})$ *that share the same first edge* $(i, j_s)$*, but are otherwise disjoint.*

*Moreover,* $\Xi_r = \{\Gamma(\boldsymbol{w}) : \boldsymbol{w} \in \mathcal{N}_*\}$ *is the set of all perfect matchings on* $K_r$*, hence* $|\Xi_r| = (r-1)!!$.

Let us introduce the following terminology: For the two subtrees $G(\boldsymbol{w}^s)$ and $G(\boldsymbol{w}^{s'})$ in part (d) of Lemma 15, we refer to the disjoint parts of the $\boldsymbol{w}^s$ and $\boldsymbol{w}^{s'}$, after the initial edge, as a *matched pair of emanating branches*. Thus $G(\boldsymbol{w})$ can be described as follows: An $((r/2)+1)$-vertex star centered on $i$, which we refer to as the *core star*, to each of its $r/2$ leaves is attached a matched pair of emanating branches, each of length $k-1$, and mutually disjoint except at their root. The non-matched emanating branches (i.e., those attached to different leaves of the core star) are completely disjoint.

Fix $\boldsymbol{w} \in \mathcal{N}_*$ and let $\Gamma(\boldsymbol{w}) = \Gamma = (\Gamma_q)_{q=1}^{Q}$. Recall that, in general, $\Gamma$ forms a partition of $[r]$. By Lemma 15(a), each $\Gamma_q$ is of the form $\{s, s'\}$ for $s \neq s'$. Then, if $(i, j_s)$ and $(i, j_{s'})$ are the first edges of $\boldsymbol{w}^s$ and $\boldsymbol{w}^{s'}$, by Lemma 15(d), we will have $j_s = j_{s'}$ and that is the only overlap among the two walks. Let us write $j(\Gamma_q)$ for this common endpoint ($j_s = j_{s'}$) of the first edge of the two walks $\boldsymbol{w}^u, u \in \Gamma_q$. We also write $\boldsymbol{j}(\Gamma) \in [n]^Q$ for the vector whose $q$th coordinate is $j(\Gamma_q)$. We denote this $q$th coordinate with a superscript, that is, $\boldsymbol{j}^q(\Gamma) = j(\Gamma_q)$. In short, $\boldsymbol{j}(\Gamma)$ collects endpoints of the edges of the core star of $G(\boldsymbol{w})$, one for each matched pair in $\Gamma$. By Lemma 15, $Q = r/2$ hence $\boldsymbol{j}$ is $r/2$-dimensional.

**Example 5.** To illustrate, note that $\boldsymbol{w}$ in Figure 5.3 is a maximal walk sequence, belonging to $\mathcal{N}_* = \mathcal{N}_{6,15,16}$ for $r = 6$ and $k = 3$. The core store is the subgraph on nodes $\{1, 2, 7, 12\}$. The matched pair of emanating branches attached to, say 7 are $7-8-10$ and $7-9-11$, which do not overlap. Similarly, the matched pair of emanating branches attached to 2 are $2-4-6$ and $2-3-6$ and so on. As shown in Example 4, $\Gamma(\boldsymbol{w}) = \Gamma \coloneqq (\{1,2\},\{3,4\},\{5,6\})$

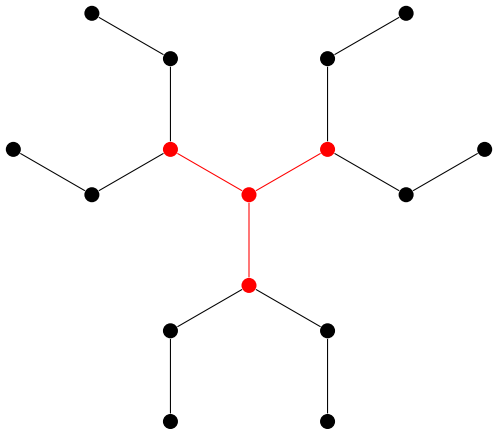


Figure 5.4: The unlabeled graph $G^*$ that all $G(\boldsymbol{w}), \boldsymbol{w} \in \mathcal{N}_*$ are isomorphic to, for $r = 6$ and $k = 4$. Its core star is colored red, while matched pairs of emanating branches are in black.

is indeed a perfect matching on $K_6$. We have $j(\Gamma_3) = j(\{5, 6\}) = 12$ and $\boldsymbol{j}(\Gamma) = (2, 7, 12)$. Similarly, $\boldsymbol{j}^3(\Gamma) = 12$, just the third coordinate of $\boldsymbol{j}(\Gamma)$.

As a consequence of Lemma 15, the graphs $G(\boldsymbol{w}), \boldsymbol{w} \in \mathcal{N}_*$ are all isomorphic to a single unlabeled graph, which we denote as $G^*$. This allows us to determine the size of $\mathcal{N}^*$ exactly (see the proof of Lemma 14). Figure 5.4 illustrates $G^*$ for the case $r = 6$ and $k = 4$, with the core star colored in red. This is essentially the unlabeled version of the graph shown in Figure 5.3. Every other graph $G(\boldsymbol{w}), \boldsymbol{w} \in \mathcal{N}_*$ can be obtained by assigning $t_* + 1 = 16$ elements from $[n]$ to the vertices of $G^*$ in Figure 5.4, and specifying a matching on $K_r$ to map the paired branches to constituent walks $\boldsymbol{w}^s$ (i.e., which of the $r$ walks in the sequence gets assigned to each branch). This matching can be equivalently seen as a branch coloring on the graph $G(\boldsymbol{w})$, where branches of $G$ are colored according to the order $(\boldsymbol{w}^1, \boldsymbol{w}^2, \ldots, \boldsymbol{w}^r)$ in the walk sequence $\boldsymbol{w}$. In this example, $n \geq 16$, but otherwise unspecified.

**Factorizing** $\rho(w)$ Lemma 15 can be used to completely factorize $\rho(\boldsymbol{w})$ for even $r$. Let

$$M \coloneqq (\mu_{y_i}, i \in [n]), \tag{5.40}$$

which we view as a $d \times n$ matrix and let $M_m$ be its $m$th row, viewed as a column vector, and $M_{mi} = (\mu_{y_i})_m$ be the $i$th coordinate of $M_m$. Note that this definition for $M$ agrees with definition (5.24) found in the signal analysis of Section 3.

**Corollary 4.** *Let $\boldsymbol{w} \in \mathcal{N}_*$, $\Gamma = \Gamma(\boldsymbol{w})$ and $\boldsymbol{j} = \boldsymbol{j}(\Gamma)$. Then,*

$$\varrho(\boldsymbol{w}) = \prod_{q=1}^{r/2} \varrho(\boldsymbol{w}^{\Gamma_q}) \tag{5.41}$$

*which for $\boldsymbol{w}^{\Gamma_q} = (\boldsymbol{v}^1, \boldsymbol{v}^2)$,*

$$\varrho(\boldsymbol{w}^{\Gamma_q}) = p_{i\boldsymbol{j}^q}(1 - p_{i\boldsymbol{j}^q}) \prod_{\alpha=1}^{2} \Big( M_{m\mathfrak{p}(\boldsymbol{v}^\alpha)} \prod_{\ell=2}^{k} p_{(\boldsymbol{v}^\alpha)_\ell} \Big), \tag{5.42}$$

*where $(\boldsymbol{v}^\alpha)_\ell$ is the $\ell$th edge in walk $\boldsymbol{v}^\alpha$ and $p_{(\boldsymbol{v}^\alpha)_\ell} = p_{i_\ell j_\ell}$ if $(\boldsymbol{v}^\alpha)_\ell = (i_\ell, j_\ell)$.*

Corollary 4 has no dependence on $\sigma$. Noise variance $\sigma$ appears in moments of $\mathbb{E}[\varepsilon_j^m]$ for $m > 1$, which is only possible for endpoint intersections on $\boldsymbol{w}$, that is $\mathfrak{p}(\boldsymbol{w}^s) = \mathfrak{p}(\boldsymbol{w}^{s'})$ for $s \neq s'$. The star shape of walks $\boldsymbol{w} \in \mathcal{N}_*$ guarantee $\mathfrak{p}(\boldsymbol{w}^s) \neq \mathfrak{p}(\boldsymbol{w}^{s'})$ for all $s \neq s'$.

When viewed as a function of $\mathcal{W}_k^2$, we express the RHS of (5.42) as $\varrho^*(\boldsymbol{v}) : \mathcal{W}_k^2 \to \mathbb{R}$ where

$$\varrho^*(\boldsymbol{v}) = p_{i\boldsymbol{j}^q}(1 - p_{i\boldsymbol{j}^q}) \prod_{\alpha=1}^{2} \Big( M_{m\mathfrak{p}(\boldsymbol{v}^\alpha)} \prod_{\ell=2}^{k} p_{(\boldsymbol{v}^\alpha)_\ell} \Big). \tag{5.43}$$

With some extra precaution, $\varrho^*$ can be extended according to (5.41) for inputs in $\mathcal{W}_k^r$. That is for $\boldsymbol{w} \in \mathcal{W}_k^r$ and for partition $\Gamma \in \Xi_r$, not necessarily equal to $\Gamma(\boldsymbol{w})$, we have

$$\varrho^*(\boldsymbol{w}; \Gamma) \coloneqq \prod_{q=1}^{r/2} \varrho^*(\boldsymbol{w}^{\Gamma_q}). \tag{5.44}$$

Note that we have suppressed the dependence of $\varrho^*(\boldsymbol{w}; \Gamma)$ on $i$ and $m$ for simplicity.

### 4.6 Proxy for $T^{\mathbf{hi}}$

From now on, let $\Gamma$ be an element of $\Xi_r$, that is, a perfect matching on $[r]$. Let

$$\mathcal{N}_*(\Gamma, \boldsymbol{j}) := \{\boldsymbol{w} \in \mathcal{N}_* : \ \Gamma(\boldsymbol{w}) = \Gamma, \ \boldsymbol{j}(\Gamma) = \boldsymbol{j}\},$$

the collection of walk sequences in $\mathcal{N}_*$ that have the same matching of walks, and the same elements of $[n] \setminus \{i\}$ as the leaves of the core star. Let $\mathcal{P}^{r/2}_{[n]\setminus\{i\}}$ be the set of all ordered $(r/2)$-tuples of distinct elements from $[n] \setminus \{i\}$. These represent all the possibilities for the leaves of the core star. We have

$$\mathcal{N}_* = \bigsqcup_{\Gamma \in \Xi_r} \bigsqcup_{\boldsymbol{j} \in \mathcal{P}^{r/2}_{[n]\setminus\{i\}}} \mathcal{N}_*(\Gamma, \boldsymbol{j}), \tag{5.45}$$

Consider the following collection of $k$-walk sequences of length $r$

$$\begin{aligned}\mathcal{W}^r_k(\Gamma, \boldsymbol{j}) \coloneqq \big\{(\boldsymbol{w}^1, \boldsymbol{w}^2, \ldots, \boldsymbol{w}^r) \text{ such that for all } q \in [r/2] \text{ and } s \in \Gamma_q \\ \boldsymbol{w}^s \text{ is a } k\text{-walk whose first edge is } (i, \boldsymbol{j}^q)\big\}.\end{aligned} \tag{5.46}$$

The walk sequences in $\mathcal{W}^r_k(\Gamma, \boldsymbol{j})$ are constructed almost similarly to those in $\mathcal{N}_*(\Gamma, \boldsymbol{j})$ except that the paired emanating branches attached to the core star are allowed to intersect. Now let

$$\widetilde{T}^{\mathrm{hi}}_{im}(r) \coloneqq \sum_{\Gamma \in \Xi_r} \sum_{\boldsymbol{j} \in \mathcal{P}^{r/2}_{[n]\setminus\{i\}}} \sum_{\boldsymbol{w} \in \mathcal{W}^r_k(\Gamma, \boldsymbol{j})} \varrho^*(\boldsymbol{w}; \Gamma). \tag{5.47}$$

Note that replacing $\mathcal{W}^r_k(\Gamma, \boldsymbol{j})$ in the inner sum above with $\mathcal{N}_*(\Gamma, \boldsymbol{j})$ would give us $T^{\mathrm{hi}}$ exactly. In fact, since $T^{\mathrm{hi}}$ and $\widetilde{T}^{\mathrm{hi}}$ utilize the same tree decomposition, their summands, $\varrho(\boldsymbol{w})$ and $\varrho^*(\boldsymbol{w}; \Gamma)$ respectively, can be upperbounded identical fashion:

**Lemma 16.** *We have* $|\varrho^*(\boldsymbol{w}; \Gamma)| \leq \|\mu_{m*}\|^r_\infty \, p^{t_*}_{\max}$ *for all* $\boldsymbol{w} \in \mathcal{W}^r_k(\Gamma, \boldsymbol{j})$. *The same bound holds for* $|\varrho(\boldsymbol{w})|$ *for any* $\boldsymbol{w} \in \mathcal{N}_*$.

*Proof.* From (5.44), we have

$$|\varrho^*(\boldsymbol{w}; \Gamma)| \leq \left(\|\mu_{m*}\|^2_\infty (p_{\max})^{2(k-1)+1}\right)^{r/2} = \|\mu_{m*}\|^r_\infty \, p^{rk-r/2}_{\max}$$

which is the desired result. The second assertion follows from the first by noting that $\rho(\boldsymbol{w}) = \varrho^*(\boldsymbol{w}; \Gamma(\boldsymbol{w}))$ for any $\boldsymbol{w} \in \mathcal{N}_*$. □

The expression for $\widetilde{T}^{\mathrm{hi}}$ can be further simplified. We have by definition (5.44)

$$\sum_{\boldsymbol{w}\in\mathcal{W}_k^r(\Gamma,\boldsymbol{j})} \varrho^*(\boldsymbol{w};\Gamma) = \sum_{\boldsymbol{w}} \prod_q \varrho^*(\boldsymbol{w}^{\Gamma_q}) = \prod_q \sum_{\boldsymbol{w}^{\Gamma_q}} \varrho^*(\boldsymbol{w}^{\Gamma_q}),$$

where the inner sum ranges over a pair of walks of the form $\boldsymbol{w}^{\Gamma_q} = (\boldsymbol{v}^1, \boldsymbol{v}^2)$ with $\boldsymbol{v}^\alpha = (i, \boldsymbol{j}^q) \,\square\, \widetilde{\boldsymbol{v}}^\alpha$ for $\alpha = 1, 2$ where $\square$ denotes a walk concatenation, and each $\widetilde{\boldsymbol{v}}^\alpha$ is a walk of length $k-1$ starting at $\boldsymbol{j}^q$, that is, $\widetilde{\boldsymbol{v}}^\alpha \in \mathcal{W}_{k-1}(\boldsymbol{j}^q)$. Using (5.43) we obtain

$$\sum_{\boldsymbol{w}^{\Gamma_q}} \varrho^*(\boldsymbol{w}^{\Gamma_q}) = p_{i\boldsymbol{j}^q}(1 - p_{i\boldsymbol{j}^q}) \sum_{\widetilde{\boldsymbol{v}}^1,\widetilde{\boldsymbol{v}}^2\in\mathcal{W}_{k-1}(\boldsymbol{j}^q)} \prod_{\alpha=1}^{2} \Big( M_{m\mathfrak{p}(\widetilde{\boldsymbol{v}}^\alpha)} \prod_{\ell=2}^{k} p_{(\widetilde{\boldsymbol{v}}^\alpha)_\ell} \Big),$$

since $\boldsymbol{v}^\alpha$ and $\widetilde{\boldsymbol{v}}^\alpha$ are identical after the first edge. The sum factorizes into two identical terms, hence equal to

$$\Big( \sum_{\widetilde{\boldsymbol{v}}\in\mathcal{W}_{k-1}(\boldsymbol{j}^q)} M_{m\mathfrak{p}(\widetilde{\boldsymbol{v}}^\alpha)} \prod_{\ell=2}^{k} p_{(\widetilde{\boldsymbol{v}}^\alpha)_\ell} \Big)^2 = \big(e_{\boldsymbol{j}^q}^T \mathbb{E}[A]^{k-1} M_m\big)^2.$$

Putting the pieces together, we have

$$\sum_{\boldsymbol{w}\,\in\,\mathcal{W}_k^r(\Gamma,\boldsymbol{j})} \varrho^*(\boldsymbol{w};\Gamma) = \prod_{q=1}^{r/2} p_{i\boldsymbol{j}^q}(1 - p_{i\boldsymbol{j}^q}) \big(e_{\boldsymbol{j}^q}^T \mathbb{E}[A]^{k-1} M_m\big)^2, \tag{5.48}$$

showing that the inner sum in (5.47) does not depend on $\Gamma$. This leads to the following simplified expression

$$\frac{\widetilde{T}^{\mathrm{hi}}}{(r-1)!!} = \sum_{\boldsymbol{j}\in\mathcal{P}_{[n]\setminus\{i\}}^{r/2}} \prod_{q=1}^{r/2} p_{i\boldsymbol{j}^q}(1 - p_{i\boldsymbol{j}^q}) \big(e_{\boldsymbol{j}^q}^T \mathbb{E}[A]^{k-1} M_m\big)^2. \tag{5.49}$$

## 4.7 Controlling $|T^{\mathbf{hi}} - \widetilde{T}^{\mathbf{hi}}|$

We will show that the difference $|T^{\mathrm{hi}} - \widetilde{T}^{\mathrm{hi}}|$ is of lower order, hence $\widetilde{T}^{\mathrm{hi}}$ will be a good surrogate for $T^{\mathrm{hi}}$ to analyze for the lower bound.

By our earlier walk construction we have $\mathcal{N}_*(\Gamma, \boldsymbol{j}) \subset \mathcal{W}_k^r(\Gamma, \boldsymbol{j})$. The difference is small relative to the larger set:

**Lemma 17.** *Let $\mathcal{N}_*(\Gamma, \boldsymbol{j})$ be defined as above. Then, for all $n \geq t_* = r(k-1/2)$*

$$\frac{|\mathcal{N}_*(\Gamma, \boldsymbol{j})|}{|\mathcal{W}_k^r(\Gamma, \boldsymbol{j})|} \geq 1 - \frac{t_*^2}{n},$$

*and consequently, $|\mathcal{W}_k^r(\Gamma, \boldsymbol{j}) \setminus \mathcal{N}_*(\Gamma, \boldsymbol{j})| \leq t_*^2\, n^{t_* - r/2 - 1}$.*

We are now ready to show that $|T^{\text{hi}} - \widetilde{T}^{\text{hi}}|$ is of lower order:

**Lemma 18.** *For any $\boldsymbol{j} \in \mathcal{P}^{r/2}_{[n]\setminus\{i\}}$ and $n \geq rk - r + 1$,*

$$|T^{hi}_{im}(r) - \widetilde{T}^{hi}_{im}(r)| \leq (r-1)!!\, t_*^2\, \|\mu_{m*}\|_\infty^r\, p_{max}\, \nu_n^{t_*-1}.$$

*Proof.* Let $\mathcal{R}(\Gamma, \boldsymbol{j}) = \mathcal{W}_k^r(\Gamma, \boldsymbol{j}) \setminus \mathcal{N}_*(\Gamma, \boldsymbol{j})$. Factorizing $T^{\text{hi}}$ into a similar form as $\widetilde{T}^{\text{hi}}$,

$$T^{\text{hi}} \;=\; \sum_{\boldsymbol{w} \in \mathcal{N}_*} \varrho(\boldsymbol{w}) \;=\; \sum_{\Gamma \in \Xi_r} \sum_{\boldsymbol{j} \in \mathcal{P}^{r/2}_{[n]\setminus\{i\}}} \sum_{\boldsymbol{w} \in \mathcal{N}_*(\Gamma, \boldsymbol{j})} \varrho^*(\boldsymbol{w}; \Gamma),$$

where $\varrho(\boldsymbol{w}) = \varrho^*(\boldsymbol{w}; \Gamma)$ for $\boldsymbol{w} \in \mathcal{N}^*(\Gamma, \boldsymbol{j})$ by Corollary 4. By the inclusion $\mathcal{N}_*(\Gamma, \boldsymbol{j}) \subset \mathcal{W}_k^r(\Gamma, \boldsymbol{j})$,

$$\widetilde{T}^{\text{hi}} - T^{\text{hi}} = \sum_{\Gamma \in \Xi_r} \sum_{\boldsymbol{j} \in \mathcal{P}^{r/2}_{[n]\setminus\{i\}}} \sum_{\boldsymbol{w} \in \mathcal{R}(\Gamma, \boldsymbol{j})} \varrho^*(\boldsymbol{w}; \Gamma).$$

Then, we have

$$\begin{aligned} |T^{\text{hi}} - \widetilde{T}^{\text{hi}}| &\leq |\Xi_r| \cdot |\mathcal{P}^{r/2}_{[n]\setminus\{i\}}| \cdot \max_{\Gamma, \boldsymbol{j}} \Big\{ |\mathcal{R}(\Gamma, \boldsymbol{j})| \cdot \max_{\boldsymbol{w} \in \mathcal{R}(\Gamma, \boldsymbol{j})} |\varrho^*(\boldsymbol{w}; \Gamma)| \Big\} \\ &\leq (r-1)!! \cdot \left(n^{r/2}\right) \cdot t_*^2\, n^{t_* - r/2 - 1} \cdot \|\mu_{m*}\|_\infty^r\, p_{\max}^{t_*} \end{aligned}$$

using Lemma 17 and Lemma 16. With $\nu_n = np_{\max}$, the last bound is the claimed bound. □

### 4.8 $\widetilde{T}^{\text{hi}}$ Lowerbound

We prove a $\widetilde{T}^{\text{hi}}$ lowerbound for the specific case of $r = 2$, for which $t_* = 2k - 1$. We note that this proof can be straightfowardly extended for the case of $r \in 2\mathbb{N}$.

Let $e_j$ be the $j$th standard basis vector of $\mathbb{R}^n$, and for $\ell \in [L]$, define

$$E_\ell \coloneqq \sum_{j:y_j \in \mathcal{I}_\ell} e_j e_j^T. \tag{5.50}$$

Note that for any matrix $H \in \mathbb{R}^{n\times n}$, we have $\|HE_\ell\|_F^2 = \sum_{j:y_j\in\mathcal{I}_\ell} \|He_j\|_2^2$.

**Lemma 19.** *Assume (A1)–(A2) and suppose $y_i = \ell$. Then we have*

$$\sum_m \widetilde{T}_{im}^{hi}(2) \;\geq\; c_B c_\nu \Big(\frac{\nu_n}{n}\|M\,\mathbb{E}[A]^{k-1}E_\ell\|_F^2 - dC_\mu^2\frac{\nu_n^{2k-1}}{n}\Big).$$

Continuing, we have the supplementary lemma which lowerbounds the average Frobenius-norm of $M\mathbb{E}[A]^{k-1}E_\ell$:

**Lemma 20.** *Assume (A3)–(A5) and* $\min\{n/k,\, \nu_n^\delta/C_B\} \geq 4C_\mu L/(c_\pi c_\xi)$. *Then,*

$$\frac{\nu_n}{n}\sum_\ell \pi_\ell \|M\,\mathbb{E}[A]^{k-1}E_\ell\|_F^2 \geq \Big(\frac{c_\pi c_\xi^2 d}{8L}\Big)\nu_n^{2k-1}.$$

Recall the definition of $\kappa_1 = (c_B c_\nu c_\pi c_\xi^2)/(48L)$ in (5.15).

**Proposition 7.** *Assume (A1)–(A5) and growth condition* (5.16) *on $n$. Then,*

$$\frac{1}{n}\sum_{i,m}\widetilde{T}_{im}^{hi}(2) \geq 3\kappa_1 d\,\nu_n^{2k-1}.$$

*Proof.* By Lemma 19, for all $i \in [n]$ and $\ell \in [L]$,

$$\mathbb{1}\{y_i=\ell\}\sum_m \widetilde{T}_{im}^{\mathrm{hi}}(2) \;\geq\; \mathbb{1}\{y_i=\ell\}\cdot c_B c_\nu\Big(\frac{\nu_n}{n}\|M\,\mathbb{E}[A]^{k-1}E_\ell\|_F^2 - dC_\mu^2\frac{\nu_n^{2k-1}}{n}\Big).$$

Summing both sides over $i \in [n]$ and $\ell \in [L]$, and dividing by $n$,

$$\frac{1}{n}\sum_{i,m}\widetilde{T}_{im}^{\mathrm{hi}}(2) \;\geq\; c_B c_\nu\Big(\frac{\nu_n}{n}\sum_\ell \pi_\ell\|M\,\mathbb{E}[A]^{k-1}E_\ell\|_F^2 - dC_\mu^2\frac{\nu_n^{2k-1}}{n}\Big).$$

Combining Lemmas 19 and 20, we note that if $\min\{(n/k),\,(\nu_n^\delta/C_B)\} \geq 4C_\mu L/(c_\pi c_\xi)$ and $c_\pi c_\xi^2/(16L) \;\geq\; C_\mu^2/n$, then the result follows. These conditions on $n$ can be combined into (5.16). □

### 4.9 Proof of Theorem 12

We are now ready to prove the noise lowerbound.

**Lemma 21.** *Assume growth conditions* (5.16) *and* (5.17)*, and* $r_n \geq 2$*. We have the following lower bound*

$$\mathbb{E}(\overline{D})^2 \geq \kappa_1 d \nu_n^{2k-1}.$$

*Proof.* Growth condition (5.17) can be written as

$$2\|\mu\|_{\max}^2 \max\{(2k-1)^2 n^{-1}, \nu_n^{-\epsilon}\} \leq \kappa_1, \tag{5.51}$$

We apply the results with $r = 2$, and let $t_* = 2k - 1$. We have

$$\begin{aligned}\mathbb{E}(\overline{D})^2 \geq \frac{1}{n}\sum_{i,m}\mathbb{E}[\Delta_{im}^2] &= \frac{1}{n}\sum_{i,m}(T_{im}^{\mathrm{hi}}(2) + T_{im}^{\mathrm{lo}}(2)) \\ &\geq \frac{1}{n}\sum_{i,m}\Big(\widetilde{T}_{im}^{\mathrm{hi}}(2) - |T_{im}^{\mathrm{hi}}(2) - \widetilde{T}_{im}^{\mathrm{hi}}(2)| - |T_{im}^{\mathrm{lo}}(2)|\Big)\end{aligned}$$

By Proposition 7, $\frac{1}{n}\sum_{i,m}\widetilde{T}_{im}^{\mathrm{hi}}(2) \geq 3\kappa_1 d\nu_n^{t_*}$. From Lemma 18,

$$\sum_m |T_{im}^{\mathrm{hi}}(2) - \widetilde{T}_{im}^{\mathrm{hi}}(2)| \leq 2dt_*^2 p_{\max}\|\mu\|_{\max}^2 \nu_n^{t_*-1}$$

and from Lemma 14, for $2 \leq r_n$, we have

$$\sum_m |T_{im}^{\mathrm{lo}}(r)| \leq d(\sqrt{2}\,\|\mu\|_{\max})^2 \cdot \nu_n^{t_*-\epsilon}.$$

By assumption, $2t_*^2 p_{\max}\|\mu\|_{\max}^2 \nu_n^{-1} \leq \kappa_1$ and $2\|\mu\|_{\max}^2 \nu_n^{-\epsilon} \leq \kappa_1$. Combining, we obtain $\mathbb{E}(\overline{D})^2 \geq \kappa_1 d\nu_n^{t_*}$. which is the claimed bound. □

Recall the definition of $\kappa_2 = 8\big(32\|\mu\|_{\max}^4 + (8C_1\sigma)^4\big)$ in (5.15).

**Lemma 22.** *Under the assumptions of Lemma 21, further assume* $r_n \geq 4$*. Then,*

$$\frac{\big(\mathbb{E}\overline{D}^2\big)^2}{\mathbb{E}\,\overline{D}^4} \geq \frac{\kappa_1^2}{\kappa_2}.$$

*Proof.* We have the upper bound

$$\begin{aligned}\mathbb{E}(\overline{D})^4 &\le \frac{d}{n}2^3\sum_{i,m}\big(\mathbb{E}(\Delta_{im}^4)+\mathbb{E}(\Delta_{im}^{\varepsilon})^4\big)\\ &\le 8d^2\max_{i,m}\big(\mathbb{E}(\Delta_{im}^4)+\mathbb{E}(\Delta_{im}^{\varepsilon})^4\big)\end{aligned}$$

By Lemma 14, we have for $4 \le r_n$,

$$\mathbb{E}(\Delta_{im}^4) \le |T_{im}^{\mathrm{hi}}(4)| + |T_{im}^{\mathrm{lo}}(4)| \le 2|T_{im}^{\mathrm{hi}}(4)| \le 2(\sqrt{4}\|\mu\|_{\max})^4\nu_n^{4k-2}$$

and from (5.33),

$$\mathbb{E}(\Delta_{im}^{\varepsilon})^4 \le \big(4C_1\sigma\nu_n^{k-1/2}\sqrt{4}\big)^4 = (8C_1\sigma)^4\nu_n^{4k-2}.$$

With $\kappa_2$ as defined above, we have $\mathbb{E}(\overline{D})^4 \le \kappa_2 d^2\nu_n^{4k-2}$. Combining with the lower bound in Lemma 21, we get the result. □

*Proof of Theorem 12.* Applying the Paley–Zygmund inequality to non-negative quantity $\overline{D}^2$ yields

$$\mathbb{P}\big(\overline{D}^2 \ge \eta\,\mathbb{E}\overline{D}^2\big) \ge (1-\eta)^2\frac{\big(\mathbb{E}\overline{D}^2\big)^2}{\mathbb{E}\,\overline{D}^4}.$$

Using Lemma 21 on the LHS and Lemma 22 on the RHS, we have the result. □

## 5 Conclusion

In this work, we provide sharp bounds on the signal-to-noise ratio for graph-aggregated features. We show that these features have a fundamental information rate which is invariant to the overall depth of the network $k$. These are features which underpin many GNN architectures and are the differentiator between GNNs and traditional feed-forward neural networks. As such, the knowledge that a feature information limit exists and is attainable for all networks with depth $k \ge 1$, is likely to influence future GNN architecture choices for empirical studies.

Our results are the first of their kind with respect to the generality of their assumptions. We work with the common CSBM but make no assumptions on the connectivity structure and allow for potentially disconnected clusters. Furthermore, our results bring to light how much the signal and noise are intertwined for GNNs, that a separation in the features necessarily comes with a scaling of the noise. Additionally, the upperbound portions of our noise bounds hold for the general class of inhomogeneous Erdös-Rényi models. This is a family of generative graph models which supersede CSBMs and include models like the random dot-product graph (RDPG).

Our technical contributions provide a mix of results from matrix analysis and random matrix theory. Results related to walks and their combinatorics, provide simple recipes to extend our anylysis to other generative frameworks. Other results bring to light the interplay between path-counting and edge probabilities, where certain subgraphs (mainly non-tree subgraphs) are less likely to occur given that they contain fewer vertices than unique edges. As we have shown, it is exactly this interplay which leads to the presence of dominant walk types in the noise contribution. Lastly, the presence of dominant walk types allows for a clean analysis since, to the leading order, this means the noise can be approximated by special polynomial forms (refer to Section 4.6 for more details).

# Appendix

## A Proofs For Signal Argument

We start with a lemma matrix monomial deviations.

**Lemma 23.** *Let $U, V \in \mathbb{R}^{n\times n}$ then for any $k \in \mathbb{N}$,*

$$\|U^k - V^k\| \leq k2^{k-2}\|U - V\| \left(\|U - V\|^{k-1} + \|V\|^{k-1}\right). \tag{5.52}$$

*Alternatively one can also derive*

$$\|U^k - V^k\| \leq k\|U - V\| \left(\max\{\|U\|, \|V\|\right)^{k-1}. \tag{5.53}$$

*Proof.* Consider the matrix valued function $f(U) = U^k$. By [Bha97, Theorem X.4.5] one has, for matrix inputs $U$ and $V$,

$$\|f(U) - f(V)\| \leq \|U - V\| \sup_{W \in \mathscr{L}(U,V)} \|\partial f_W\|$$

where $\mathscr{L}(U, V) = \{\eta U + (1 - \eta)V : \eta \in [0, 1]\}$ is the line segment between matrices $U$ and $V$, $\partial f_W$ is the Fréchet derivative of $f$ at $W$ and $\|\partial f_W\|$ is the operator norm of the derivative defined as

$$\|\partial f_W\| \coloneqq \sup_{\|H\|=1} \|\partial f_W(H)\|.$$

The Fréchet derivative of a matrix monomial can be straightforwardly computed as

$$\partial f_W(H) = \sum_{\ell=0}^{k-1} W^\ell H W^{k-1-\ell}.$$

By submultiplicativity of operator norm

$$\sup_{\|H\|=1} \|\partial f_W(H)\| \ \leq \ \sum_{\ell=0}^{k-1} \sup_{\|H\|=1} \|W^\ell H W^{k-1-\ell}\| \ \leq \ k\|W\|^{k-1}.$$

Next with $W = \eta U + (1 - \eta)V$

$$\begin{aligned} \|W\|^{k-1} &= \|\eta(U - V) + V\|^{k-1} \\ &\leq (\|U - V\| + \|V\|)^{k-1}, \end{aligned}$$

where we recall $\eta \in [0, 1]$. By convexity of $x \mapsto x^{k-1}$ on $[0, \infty)$,

$$(\|U - V\| + \|V\|)^{k-1} \leq 2^{k-2}(\|U - V\|^{k-1} + \|V\|^{k-1})$$

Stringing along the previous inequalities proves (5.52). To obtain (5.53) simply bound $\|W\|$

as

$$\begin{aligned}\|W\|^{k-1} &= \|\eta U + (1-\eta)V\|^{k-1} \\ &\leq (\eta\|U\| + (1-\eta)\|V\|)^{k-1} \\ &\leq \eta\|U\|^{k-1} + (1-\eta)\|V\|^{k-1} \\ &\leq (\max\{\|U\|, \|V\|\})^{k-1}.\end{aligned}$$

using the convexity of $x \mapsto x^{k-1}$ in the third line. □

The final piece needed to prove Theorem 10 is Lemma 5 (in Section 3) which is a concentration result for exponentiated random matrices with bounded sub-Gaussian entries. The proof of this result follows.

### A.1 Proof of Lemma 5

Recall that wlog $p_{\max} = \nu_n/n$. So $\|\mathbb{E}[A]\| \leq np_{\max} = \nu_n$. Apply (5.52) of Lemma 23 with $U = A$ and $V = \mathbb{E}[A]$ to obtain

$$\|A^k - \mathbb{E}[A]^k\| \leq C_{1,k}\Big(\|\Delta\|^k + \nu_n^{k-1}\|\Delta\|\Big). \tag{5.54}$$

where $C_{1,k} = k2^{k-2}$, and $\Delta := A - \mathbb{E}[A]$.

Next consider a spectral concentration result from [BH16, Corollary 3.12; Remark 3.13] where, for a random matrix $U$ with independent sub-Gaussian entries, one has

$$\mathbb{P}\big(\|U\| > C\widetilde{\sigma} + t\big) \leq ne^{-t^2/c\widetilde{\sigma}_*^2}$$

for universal constants $C > 1$ and $c > 0$, with $\widetilde{\sigma} := \max_i \sqrt{\sum_j \mathbb{E}[U_{ij}^2]}$ and $\widetilde{\sigma}_* := \max_{ij} \|U_{ij}\|_\infty$. For $U = A - \mathbb{E}[A]$ these parameters become $\widetilde{\sigma} \leq \sqrt{\nu_n}$ and $\widetilde{\sigma}_* \leq 1$. Consider the event

$$\mathcal{E} = \big\{\|\Delta\| \leq C_{2,k}\sqrt{\nu_n}\big\}$$

where $C_{2,k} = C + \sqrt{(c/c'_\nu)(k+1)}$. Then, taking $t = (C_{2,k} - C)\sqrt{\nu_n}$ in (5.54), we have

$$\mathbb{P}(\mathcal{E}^c) \leq ne^{-(k+1)\nu_n/c'_\nu} \leq ne^{-(k+1)\log n} \leq n^{-k}.$$

Then, on the one hand,

$$\mathbb{E}[\|\Delta\|^k \mathbb{1}_{\mathcal{E}^c}] \leq n^k \mathbb{P}(\mathcal{E}^c) \leq 1 \leq \nu_n^{k-1/2}.$$

On the other hand,

$$\mathbb{E}[\|\Delta\|^k \mathbb{1}_{\mathcal{E}}] \leq C_{2,k}^k \nu_n^{k/2} \leq C_{2,k}^k \nu_n^{k-1/2}.$$

Putting the pieces together we have

$$\mathbb{E}\|A^k - \mathbb{E}[A]^k\| \leq C_{1,k}\Big((1 + C_{2,k}^k)\nu_n^{k-1/2} + \nu_n^{k-1}(1 + C_{2,1})\nu_n^{1/2}\Big)$$

proving the result with constant, e.g., $C_k = 4C_{1,k}C_{2,k}^k$ using $C_{2,k} \geq 1$ and $C_{2,k} \geq C_{2,1}$.

# B Concentration from Moments

We have the following concentration inequality from moments:

**Lemma 24** (Sub-Weibull concentration)**.** *Let $\eta > 0$. Assume that for all $r \in 2\mathbb{N}$, we have*

$$\mathbb{E}[|\Delta|^r] \leq \Big(K(C\eta r)^\eta\Big)^r. \tag{5.55}$$

*Then,*

$$\mathbb{P}\Big(|\Delta| \geq Kx^\eta\Big) \leq \exp\Big(-\frac{x}{2Ce}\Big) \qquad \text{for } x \geq 4\eta Ce.$$

Lemma 24 follows from a more general statement for partial moment control:

**Lemma 25.** *Let $\eta > 0$ and $r_0 \in 2\mathbb{N} \cup \{\infty\}$. Assume that for all even integers $r \leq r_0$, we have*

$$\mathbb{E}|\Delta|^r \leq \Big(K(C\eta r)^\eta\Big)^r. \tag{5.56}$$

*Then* (5.56) *holds for all real $r \in [2, r_0]$ with $C$ replaced with $2C$. Moreover, if $x \geq 4C\eta e$*

$$\mathbb{P}\Big(|\Delta| \geq Kx^\eta\Big) \leq \begin{cases} \exp\big(-\frac{x}{2Ce}\big), & \text{for } x \leq 2C\eta e r_0 \\ \big(\frac{2\eta C r_0}{x}\big)^{\eta r_0}, & \text{for } x > 2C\eta e r_0 \end{cases} \leq \exp\Big(-\min\Big\{\frac{x}{2Ce}, \eta r_0\Big\}\Big).$$

*Proof.* Let us first establish the claim that (5.56) holds for all real $r \in [2, r_0]$ with $C$ replaced with $2C$. If $r_0 = 2$ this is trivial. Otherwise for $r_0 \geq 4$, we will use the log-convexity of $L_p$ norms, $(\mathbb{E}|\Delta|^p)^{1/p}$. Fix $r \leq r_0 - 2$ and $\theta \in (0, 1)$. Log-convexity implies

$$(\mathbb{E}|\Delta|^p)^{1/p} \leq (\mathbb{E}|\Delta|^r)^{(1-\theta)/r} (\mathbb{E}|\Delta|^{r+2})^{\theta/(r+2)}$$

where $p$ is given by $\frac{1}{p} = \frac{1-\theta}{r} + \frac{\theta}{r+2}$. Applying (5.56) with both $r$ and $r+2$, we obtain

$$\begin{aligned}(\mathbb{E}|\Delta|^p)^{1/p} &\leq [K(Cr\eta)^\eta]^{1-\theta} [K(C(r+2)\eta)^\eta]^\theta \\ &= K(Cr^{1-\theta}(r+2)^\theta \eta)^\eta.\end{aligned}$$

We have $r^{1-\theta}(r+2)^\theta = 2(r/2)^{1-\theta}(r/2+1)^\theta \leq 2(r/2)^{1-\theta} r^\theta \leq 2^\theta r \leq 2p$, since $r < p$. This gives $(\mathbb{E}|\Delta|^p)^{1/p} \leq K(2Cp\eta)^\eta$ which is the desired bound.

Next, consider the tail bound. Let us redefine $2C$ to be $C$ for simplicity. By Markov inequality, for all real $r \geq 2$,

$$\mathbb{P}(|\Delta| \geq u) \leq \frac{\mathbb{E}|\Delta|^r}{u^r} \leq \Big(\frac{K(Cr\eta)^\eta}{u}\Big)^r.$$

Using a change of variable $u = Kx^\eta$, we have

$$\mathbb{P}\Big(|\Delta| \geq Kx^\eta\Big) \leq \Big(\frac{(C\eta r)^\eta}{x^\eta}\Big) = \alpha^r r^{\eta r} = \exp(f(r))$$

where $\alpha = (C\eta/x)^\eta$ and $f(r) = r \log \alpha + \eta r \log r$. Let us now optimize over $r$. The function $f$ has first and second derivative $f'(r) = \log \alpha + \eta \log r + \eta$ and $f''(r) = \eta/r > 0$. Hence, $f$ is strictly convex on $(0, \infty)$ and achieves its global minimum at $r^* = e^{-1-(\log \alpha)/\eta} = (1/e)\alpha^{-1/\eta}$. If $r_0 \geq r^* \geq 2$, that is $x \leq C\eta e r_0$, then we achieve the first case tail bound

$$\mathbb{P}\Big(|\Delta| \geq Kx^\eta\Big) \leq \exp\Big(-\frac{x}{2Ce}\Big).$$

Otherwise if $x > C\eta e r_0$, one has $r_0 < r^*$. In this case we use the next best value of $r$ for our probability upperbound $\exp(f(r))$, that is $r = r_0$. Altogether

$$\mathbb{P}\Big(|\Delta| \geq Kx^\eta\Big) \leq \begin{cases} \exp\big(-\frac{x}{Ce}\big), & \text{for } x \leq C\eta e r_0 \\ \big(\frac{\eta C r_0}{x}\big)^{\eta r_0}, & \text{for } x > C\eta e r_0. \end{cases}$$

We note that the two bounds match at the boundary. We can summarize as

$$\begin{aligned}\mathbb{P}\Big(|\Delta| \geq Kx^{\eta}\Big) &\leq \max\Big\{e^{-x/(Ce)}, \Big(\frac{Cr_0\eta}{x}\Big)^{\eta r_0}\Big\} \\ &\leq \exp\big(-\min\{x/(Ce), \eta r_0\}\big)\end{aligned}$$

where we evaluated the second bound at $x = C\eta e r_0$. Replacing $C$ with $2C$ finishes the proof. □

# C Counting Lemmas

## C.1 Proof of Lemma 6

Let $i, j \in [n]$ with $i \neq j$. We will over-enumerate the walks in $\mathcal{N}_t(i,j)$, each of which contain, at most, $t+1$ unique vertices. We construct a potential vertex set for a walk $w$ in $\mathcal{N}_t(i,j)$ by first fixing $\{i,j\}$ and selecting the remaining $t-1$ vertices from $[n] \setminus \{i,j\}$. There are exactly $\binom{n-2}{t-1}$ ways to do this. Next, starting from node $i$, each outgoing vertex forms an additional edge in our walk. As there can be no self-loops and the last outgoing vertex $j$ is already determined, there are, at most, $t^{k-1}$ ways to select such edges. Altogether, there are, at most, $\binom{n-2}{t-1}t^{k-1}$ ways to construct a walk in $\mathcal{N}_t(i,j)$, which is the desired result.

## C.2 Proof of Lemma 7

The proof for the cardinality of $\mathring{\mathcal{N}}_t(i,i)$ follows similarly to the case $i \neq j$ (Lemma 6). The key realization is that any walk $w \in \mathring{\mathcal{N}}_t(i,i)$ can have at most $t$ unique vertices. This is by definition since the undirected graph $G(w) = (V, E)$ is, one, connected ($|V| \leq |E| + 1$) and, two, not a tree ($|V| \neq |E| + 1$). So walk $w \in \mathring{\mathcal{N}}_t(i,i)$ must have at most $t$ unique vertices.

The rest of the proof is devoted to bounding $|\breve{\mathcal{N}}_t(i,i)|$. For $w \in \breve{\mathcal{N}}_t(i,i)$, the edges of $G(w)$ must all be traversed an even number of times. We refer to a traversal that occurs on an odd (resp. even) visits of an edge as positive (resp. negative) traversals. Given

the backtracking nature of walks $w \in \breve{\mathcal{N}}_t(i,i)$, it is sufficient to keep track of positive edge traversals to reconstruct $w$, and so a recipe for bounding $|\breve{\mathcal{N}}_t(i,i)|$ is to list all viable $G(w)$ and count the possible positive edge traversals.

Recall that $G(w) = (V, E)$ is loop-less and connected, hence a tree. The vertices $V$ have a natural ordering on $\mathbb{N}$ that allows for an encoding of $G(w)$ using the so-called "Balanced parentheses sequence" or "Dyck word" representation of the tree. The unlabeled graphs of these encodings, which we will call $\mathcal{B}(w)$, are precisely the (rooted) plane trees with $t+1$ vertices and their count is exactly the Catalan number:

$$\mathfrak{B} \coloneqq \{\mathcal{B}(w) : w \in \breve{\mathcal{N}}_t(i,i)\}, \quad |\mathfrak{B}| = C_t.$$

Next, since $\mathcal{B}(w)$ has the same number of edges as $G(w)$, we can construct edge selections on $\mathcal{B}(w)$ that correspond to edge traversals on $G(w)$, modulo a choice of vertices $V$. Selecting for the vertices $V$, where the root $i$ is determined, gives an additional factor of $\binom{n-1}{t}$.

Lastly, we must perform an edge selection for each $\mathcal{B}(w)$, in particular we must select for $k/2$ edges which will later be our positive traversals. An arbitrary selection of edges gives a factor of $D_{k,t} = t!\left\{{k/2 \atop t}\right\}$ where $\left\{{m \atop t}\right\}$ is the Stirling number of second kind. To see this note that we arranging $t$ objects in $m = k/2$ slots where (a) each object must be used at least once (to cover the tree), (b) objects can be used multiple times and (c) the order matters. This is equivalent to the number of surjections from a set of $m$ elements to a set of $t$ elements which is given by $t!\left\{{m \atop t}\right\}$.

Certain edge selections on $\mathcal{B}(w)$ will translate to disconnected (i.e., invalid) walks on $G(w)$, yielding an overcount in the number of walks $w \in \breve{\mathcal{N}}_t(i,i)$ with undirected graph $G(w)$. An exception is when $G(w)$ is a star. Here, any edge selection on $\mathcal{B}(w)$, with any vertex set $V$, gives a valid sequence of positive edge traversals for the star-shaped $G(w)$ with vertices $V$. In other words, the two star graphs in $\mathfrak{B}$ (corresponding to whether the root is a hub or a spoke) are maximal in terms of the counts of valid walks they produce, hence the count in this case provides an upper bound on the counts for all elements of $\mathfrak{B}$.

Combining all factors (encoding trees, vertex choice, maximality of stars) gives the desired bound

$$|\breve{\mathcal{N}}_t(i,i)| \le C_t \cdot \binom{n-1}{t} \cdot D_{k,t}.$$

For (5.31), we note that $C_t \le \frac{1}{t}(2e)^t$, and $t! \cdot S(k/2,t) \le t^{k/2}$ which says that the number of surjections from $[k/2]$ to $[t]$ is at most the total number of functions. Moreover, $\binom{n-1}{t} \le (en/t)^t$. Combining, we get the second claimed upper bound, concluding the proof.

### C.3 Proof of Lemma 13

Consider collapsing the walk sequence $\boldsymbol{w} \in \mathcal{W}^r$ to a single walk $\widetilde{w}$. That is, for walk sequence $\boldsymbol{w} = (((i_\ell^s, j_\ell^s), \ell \in [k]), s \in [r])$ let $F(\cdot)$ be the in-place tuple flattening such that

$$\widetilde{w} \coloneqq F(\boldsymbol{w}) = ((i_\ell^s, j_\ell^s), \ell \in [k], s \in [r]).$$

The tuple flattening function $F$ is a reversible process for known $r$ and $k$, meaning there is an isomorphism defined by $F$ between $\widetilde{\mathcal{N}}_{r,t,v} \coloneqq \{F(\boldsymbol{w}) : \boldsymbol{w} \in \mathcal{N}_{r,t,v}\}$ and $\mathcal{N}_{r,t,v}$. As such any cardinality upperbound on $\widetilde{\mathcal{N}}_{r,t,v}$ will hold for $\mathcal{N}_{r,t,v}$.

One way to count the walks of $\widetilde{\mathcal{N}}_{r,t,v}$ is to first choose $v-1$ non-root nodes (for which there are $\binom{n-1}{v-1}$ options). After which rename the non-root nodes without loss of generality to $1,\dots,(v-1)$ and the rename the root node $i$ to $n$. The walk $\widetilde{w}$ takes $rk$ steps on $\{1,\dots,v-1\} \cup \{n\}$. Since there are, at most, $v-1$ options to choose from at each step, we may over-enumerate the walks $\widetilde{w} \in \widetilde{\mathcal{N}}_{r,t,v}$ and arrive at the simple upperbound

$$|\mathcal{N}_{r,t,v}| = |\widetilde{\mathcal{N}}_{r,t,v}| \le (v-1)^{rk}\binom{n-1}{v-1}.$$

For the second assertion, we have

$$\sum_{v=1}^{b+1} |\mathcal{N}_{r,t,v}| \le \sum_{v=1}^{b+1}(v-1)^{rk}\binom{n-1}{v-1} = \sum_{v=0}^{b} v^{rk}\binom{n-1}{v} \le b^{rk}\sum_{v=0}^{b}\binom{n}{v} \le b^{rk}\Big(\frac{en}{b}\Big)^b.$$

The proof is complete.

# D Proofs for Noise Upperbound

## D.1 Proof of Lemma 10

We have

$$\begin{aligned}\Big|\mathbb{E}\prod_{s=1}^{r}(A_{\boldsymbol{w}^s}-\mathbb{E}[A_{\boldsymbol{w}^s}])\Big| &\le \mathbb{E}\Big[\prod_{s=1}^{r}(|A_{\boldsymbol{w}^s}|+\mathbb{E}|A_{\boldsymbol{w}^s}|)\Big]\\ &= \mathbb{E}\Big[\sum_{S\subset[r]}\prod_{s\in S}|A_{\boldsymbol{w}^s}|\cdot\prod_{u\in S^c}\mathbb{E}|A_{\boldsymbol{w}^u}|\Big]\\ &= \sum_{S\subset[r]}\Big[\mathbb{E}[A_{\boldsymbol{w}^S}]\cdot\prod_{u\in S^c}\mathbb{E}|A_{\boldsymbol{w}^u}|\Big] \le 2^r\cdot\mathbb{E}[A_{\boldsymbol{w}}] \qquad (5.57)\end{aligned}$$

where the final inequality is by Lemma 28. Then, we have $\mathbb{E}[A_{\boldsymbol{w}}] \le p_{\max}^{|[w]|}$ by (5.62).

To control $\varrho_2(\boldsymbol{w})$, we note that for $J=\{\mathfrak{p}(\boldsymbol{w}^s): s\in[r]\}\subset[n]$ and some integers $a_j\ge 1$ with $\sum_j a_j = r$, we have

$$\Big|\mathbb{E}\prod_{s=1}^{r}(x_{\mathfrak{p}(\boldsymbol{w}^s)})_m\Big| \le \mathbb{E}\Big(\prod_{s=1}^{r}|(x_{\mathfrak{p}(\boldsymbol{w}^s)})_m|\Big) = \mathbb{E}\Big(\prod_{j\in J}|(x_j)_m|^{a_j}\Big) \le \max_{j\in J}\mathbb{E}|(x_j)_m|^r$$

where the equality is by independence of $x_j, j\in J$, and the last step follows from Lemma 29. Combining with

$$\begin{aligned}\mathbb{E}|(x_j)_m|^r &\le 2^{r-1}\big(\mathbb{E}|\varepsilon_{im}|^r+|(\mu_{y_i})_m|^r\big)\\ &\le 2^{r-1}\big((C_1\sigma r^{1/2})^r+\|\mu_{m*}\|_\infty^r\big),\end{aligned}$$

where the last line used Lemma 9, the proof is complete.

### D.2 Proof of Lemma 12

With $\boldsymbol{w}$ overlapping we have for all $q$, $|\Gamma_q| \geq 2$. Hence, $2Q \leq \sum_{q=1}^{Q} |\Gamma_q| = r$ and the upper bound on $Q$ follows. Next fix $q$,

$$\begin{aligned} |[\boldsymbol{w}]^{\Gamma_q}| &\leq |[\boldsymbol{w}^1]| + \sum_{s=2}^{|\Gamma_q|} |[\boldsymbol{w}^s] \setminus [\boldsymbol{w}^1]| \\ &\leq k + (|\Gamma_q| - 1)(k-1) = |\Gamma_q|(k-1) + 1. \end{aligned}$$

Then combining across partitions $\{\Gamma_q\}_q$,

$$\begin{aligned} |[\boldsymbol{w}]| = \sum_{q=1}^{Q} |[\boldsymbol{w}]^{\Gamma_q}| &\leq \Big(\sum_{q=1}^{Q} |\Gamma_q|\Big)(k-1) + Q \\ &= r(k-1) + Q \leq rk - \lceil r/2 \rceil \end{aligned}$$

using $Q \leq \lfloor r/2 \rfloor$.

### D.3 Proof of Lemma 14

Throughout, we fix $i, m$ and $r$. We start with the bound for $T^{\mathrm{lo}}$. To simplify the notation, let

$$\beta_r := (\|\mu_{m*}\|_\infty \sqrt{r})^r.$$

From Lemma 10, for any $\boldsymbol{w} \in \mathcal{N}_{r,t,v}$, we have

$$\begin{aligned} |\rho(\boldsymbol{w})| &= |\varrho_1(\boldsymbol{w})| \cdot |\varrho_2(\boldsymbol{w})| \\ &\leq (4 \max\{C_1 \sigma r^{1/2}, \|\mu_{m*}\|_\infty\})^r \, p_{\max}^t \\ &\leq (\kappa_{0,m} \|\mu_{m*}\|_\infty r^{1/2})^r p_{\max}^t \leq \beta_r \, \kappa_0^r \, p_{\max}^t \end{aligned} \tag{5.58}$$

where $\kappa_{0,m}\|\mu_{m*}\|_\infty = 4\max\{C_1\sigma, \|\mu_{m*}\|_\infty\}$ and $\kappa_0 = \max_m \kappa_{0,m}$ by definition.

Let us first bound $T^{\mathrm{lo}}$. We have

$$T^{\mathrm{lo}} = \sum_{t=1}^{t_*} \sum_{v=2}^{b_t+1} \sum_{\boldsymbol{w} \in \mathcal{N}_{r,t,v}} \varrho(\boldsymbol{w})$$

where $b_t = t \wedge (t_* - 1)$. That is, the inner sum goes from $v = 2$ to $v = t + 1$ unless $t = t^*$ in which goes it only goes to $v = t^*$. Using (5.58),

$$\frac{|T^{\mathrm{lo}}|}{\beta_r} \leq \kappa_0^r \sum_{t=1}^{t_*} p_{\max}^t \sum_{v=2}^{b_t+1} |\mathcal{N}_{r,t,v}|.$$

By Lemma 13, we have $\sum_{v=2}^{b_t+1} |\mathcal{N}_{r,t,v}| \leq b_t^{r/2} b_t^{t_*-b_t} (en)^{b_t}$ since $t_* = rk - r/2$. Using this bound and $p_{\max} = \nu_n/n$, we have

$$\frac{|T^{\mathrm{lo}}|}{\beta_r} \leq \kappa_0^r \, t_*^{r/2} \sum_{t=1}^{t_*} \nu_n^t n^{-t} b_t^{t_*-b_t} (en)^{b_t}.$$

Multiplying and dividing by $\nu_n^{t_*}$ and rearranging, then separating the term $t = t_*$ from $t < t_*$, we have

$$\begin{aligned}
\frac{|T^{\mathrm{lo}}|}{\beta_r \, \nu_n^{t_*}} &\leq \kappa_0^r \, t_*^{r/2} \sum_{t=1}^{t_*} \nu_n^{t-t_*} n^{b_t - t} b_t^{t_*-b_t} e^{b_t}, \\
&= \kappa_0^r \, t_*^{r/2} \Big[ n^{-1}(t_* - 1) e^{t^*-1} + \sum_{t=1}^{t_*-1} \nu_n^{t-t_*} t^{t_*-t} e^t \Big] \\
&\leq \kappa_0^r \, t_*^{r/2} e^{t^*-1} \Big[ n^{-1}(t_* - 1) + \sum_{t=1}^{t_*-1} \Big( \frac{t}{\nu_n} \Big)^{t_*-t} \Big].
\end{aligned}$$

By assumption $\kappa_0^r \, t_*^r e^{t_*} \leq \frac{1}{3} \nu_n^{1-\epsilon}$, which implies $\kappa_0^r \, t_*^{r/2+1} e^{t_*} \leq \frac{1}{3} \nu_n^{1-\epsilon}$. (The result holds under this slightly weaker form). Let $\rho = \kappa_0^{-r} t_*^{-r/2} e^{-t_*} \nu_n^{-\epsilon}$. Then, $t_* \leq \frac{1}{3} \rho \nu_n$, hence

$$\sum_{t=1}^{t_*-1} \Big( \frac{t}{\nu_n} \Big)^{t_*-t} \leq \sum_{t=1}^{t_*-1} \rho^{t_*-t} \leq \sum_{u=1}^{\infty} \rho^u \leq 2\rho.$$

using $\rho \leq 1/2$ since $\nu_n \geq 1$ and $t_* \geq 1$. We obtain

$$\frac{|T^{\mathrm{lo}}|}{\beta_r \, \nu_n^{t_*}} \leq \kappa_0^r \, t_*^{r/2} e^{t_*} \Big[ t_* n^{-1} + \frac{2}{3} \kappa_0^{-r} \, t_*^{-r/2} e^{-t_*} \nu_n^{-\epsilon} \Big] \leq \nu_n^{-\epsilon}$$

where we have used $\kappa_0^r \, t_*^{r/2+1} e^{t_*} n^{-1} \leq \frac{1}{3} \nu_n^{1-\epsilon} n^{-1} \leq \frac{1}{3} \nu_n^{-\epsilon}$. This proves the claim upper bound on $T^{\mathrm{lo}}$.

Next, we prove the bound for $T^{\mathrm{hi}}$. For all $\boldsymbol{w} \in \mathcal{N}_*$, the unlabeled graph associated with the tree $G(\boldsymbol{w})$ is the same, namely the $G^*$ graph described in Section 4.5 and depicted in

Figure 5.4 for a given example. In other words, as $\boldsymbol{w}$ ranges over $\mathcal{N}_*$, $G(\boldsymbol{w})$ ranges over all possible labelings of the vertices of $G^*$ with $t^*$ distinct elements from $[n] \setminus \{i\}$. There are $|\mathcal{P}^{t_*}_{[n]\setminus\{i\}}|$ such labelings, hence

$$|\mathcal{N}_*| \;=\; (r-1)!!\,|\mathcal{P}^{t_*}_{[n]\setminus\{i\}}| \;\leq\; (r-1)!!\,n^{t_*}.$$

By Lemma 16, for every $\boldsymbol{w} \in \mathcal{N}_{r,t_*,t_*+1}$

$$|\rho(\boldsymbol{w})| \leq \|\mu_{m*}\|^r_\infty \cdot p^{t_*}_{\max}.$$

Altogether, $|T^{\text{hi}}| \leq (r-1)!!\,\|\mu_{m*}\|^r_\infty (np_{\max})^{t_*}$, which is the first claim. The double factorial is connected to the Gamma function for odd valued inputs. In particular, $(r-1)!! = \frac{2^{r/2}}{\sqrt{\pi}}(r/2-1/2)! \leq r^{r/2}$.

# E Proofs for the Noise Lower bound

## E.1 Proof of Lemma 15

For (a), each $\Gamma \in \Xi_r$ is generated by some overlapping $\boldsymbol{w} \in \mathcal{N}_{r,t_*,t_*+1}$. The overlapping nature of each walk in the walk sequence $\boldsymbol{w}$ enforces that $|\Gamma_q| \geq 2$ for all $q$. The number of equivalences classes, $Q$, is maximized when each $|\Gamma_q|$ is minimized since, by construction, $\Gamma = \{\Gamma_q\}_{q=1}^Q$ must satisfy $\bigsqcup_{q\in[Q]} \Gamma_q = [r]$. In the case $r \in 2\mathbb{N}$, $Q = r/2$ so $\{\Gamma_q\}_{q=1}^Q$ must contain pairs of elements from $[r]$ such that $\Gamma_q \cap \Gamma_{q'} = \varnothing$ for $q \neq q'$ and $|\Gamma_q| = 2$ for $q \in [Q]$. This is a perfect matching on the vertex set $[r]$.

For (b), we note that since each $\boldsymbol{w}^s$ is a walk and the walks have common first element $i$, the graph $G(\boldsymbol{w})$ is connected. That is, $G(\boldsymbol{w})$ is a connected graph with $t_*+1$ nodes and $t_*$ edges, hence a tree. We designate $i$ as its root.

For (c), the partition $\Gamma(\boldsymbol{w})$ guarantees that subgraphs $G(\boldsymbol{w}^{\Gamma_q})$ and $G(\boldsymbol{w}^{\Gamma_{q'}})$ cannot share an edge for $q, q'$ distinct. Furthermore by (b), if $G(\boldsymbol{w}^{\Gamma_q})$ and $G(\boldsymbol{w}^{\Gamma_{q'}})$ share a vertex other than the root $i$, then an undirected cycle forms, contradicting the tree property of $G(\boldsymbol{w})$.

For (d), the proof is similar to (c). By the construction of partition $\Gamma(\boldsymbol{w})$, the walks $\boldsymbol{w}^{\Gamma_q}$ must overlap at some edge. However, any overlap must occur on an outgoing edge from root $i$. Otherwise, for any other edge overlap, an undirected cycle can be made to and from the root $i$, contradicting the inherited tree property of $G(\boldsymbol{w}^{\Gamma_q})$.

### E.2 Proof of Lemma 17

By Lemma 15, each $\boldsymbol{w} \in \mathcal{N}_*(\Gamma, \boldsymbol{j})$ has root and branching nodes determined, that is, out of $t_* + 1$ nodes unique nodes, $(r/2) + 1$ nodes are determined. What is left is to select $t_* - r/2 = rk - r$ nodes from $[n] \setminus \{i, \boldsymbol{j}^1, \boldsymbol{j}^2, \ldots, \boldsymbol{j}^{r/2}\}$. An ordered selection of nodes can be used to determine walks $\boldsymbol{w}^s$ giving a cardinality of

$$|\mathcal{N}_*(\Gamma, \boldsymbol{j})| = |\mathcal{P}^{rk-r}_{[n]\setminus\{i,\boldsymbol{j}^1,\boldsymbol{j}^2,\ldots,\boldsymbol{j}^{r/2}\}}| = \prod_{\ell=1}^{rk-r} \big((n-1-r/2) - \ell + 1\big)$$

For the cardinality of $\mathcal{W}^r_k(\Gamma, \boldsymbol{j})$, what is left to determine is each $\mathcal{W}^2_{k-1}(\boldsymbol{j}^q)$. Although these walks have no self-loops, they can overlap and repeat edges. So

$$|\mathcal{W}^r_k(\Gamma, \boldsymbol{j})| = \prod_{q=1}^{r/2} |\mathcal{W}^2_k(\{\Gamma_q\}, (\boldsymbol{j}^q))| = \prod_{q=1}^{r/2} |\mathcal{W}_{k-1}(\boldsymbol{j}^q)|^2 = \big((n-1)^{2(k-1)}\big)^{r/2}$$

As such, for $n \geq rk - r/2$,

$$\frac{|\mathcal{N}_*(\Gamma, \boldsymbol{j})|}{|\mathcal{W}^r_k(\Gamma, \boldsymbol{j})|} = \prod_{\ell=1}^{rk-r} \Big(1 - \frac{r/2 + \ell - 1}{n-1}\Big) \geq \Big(1 - \frac{rk - r/2 - 1}{n-1}\Big)^{rk-r} \geq 1 - (rk-r)\frac{rk - r/2 - 1}{n-1}$$

using Bernoulli's inequality. Using $\frac{a-1}{n-1} \leq \frac{a}{n}$ for $n \geq a$, we have

$$(rk-r)\frac{rk - r/2 - 1}{n-1} \leq \frac{(rk - r/2)^2}{n}$$

proving the first assertion. For the second claim, we have

$$\begin{aligned}
|\mathcal{W}^r_k(\Gamma, \boldsymbol{j}) \setminus \mathcal{N}_*(\Gamma, \boldsymbol{j})| &= |\mathcal{W}^r_k(\Gamma, \boldsymbol{j})| - |\mathcal{N}_*(\Gamma, \boldsymbol{j})| \\
&= |\mathcal{W}^r_k(\Gamma, \boldsymbol{j})|\Big(1 - \frac{|\mathcal{N}_*(\Gamma, \boldsymbol{j})|}{|\mathcal{W}^r_k(\Gamma, \boldsymbol{j})|}\Big) \\
&\leq n^{rk-r} \cdot \frac{t_*^2}{n} = t_*^2\, n^{rk-r-1}
\end{aligned}$$

which is the desired result since $rk - r = t_* - r/2$.

### E.3 Proof of Lemma 19

Applying definition (5.47) of $\widetilde{T}^{\mathrm{hi}}_{im}(2)$,

$$\begin{aligned}\widetilde{T}^{\mathrm{hi}}_{im}(2) &= \sum_{\Gamma\in\Xi_2}\sum_{\boldsymbol{j}\in\mathcal{P}^1_{[n]\setminus\{i\}}}\prod_{q=1}^{r/2} p_{i\boldsymbol{j}^q}(1-p_{i\boldsymbol{j}^q})\big(e_{\boldsymbol{j}^q}^T\mathbb{E}[A]^{k-1}M_m\big)^2\\ &= \sum_{j\neq i} p_{ij}(1-p_{ij})\big(e_j^T\mathbb{E}[A]^{k-1}M_m\big)^2\end{aligned}$$

using $\Xi_2=\{(1,2)\}$ and $\mathcal{P}^1_{[n]\setminus\{i\}}=[n]\setminus\{i\}$. Recall that $y_i=\ell$. Then for any $j$ with $y_j\in\mathcal{I}_\ell$, by assumptions (A1)–(A2), $p_{ij}\geq c_B\nu_n$ and $1-p_{\max}\geq c_\nu$, hence $p_{ij}(1-p_{ij})\geq c_Bc_\nu\,\nu_n/n$ for $j\neq i : j\in\mathcal{I}_\ell$. Letting $c_1=c_Bc_\nu$, we have

$$\sum_m \widetilde{T}^{\mathrm{hi}}_{im}(2)\geq c_1\frac{\nu_n}{n}\sum_{\substack{j\neq i:\\ y_j\in\mathcal{I}_\ell}}\sum_m\big(M_m^T\mathbb{E}[A]^{k-1}e_j\big)^2,$$

using the symmetry of $A$ and invariance of a scalar to transpose. The inner sum over $m$ is equal to $\|M\,\mathbb{E}[A]^{k-1}e_j\|_2^2$ since $M_m^T$ is the $m$th row of $M$. Next, we note that

$$\|M\,\mathbb{E}[A]^{k-1}e_i\|_2\ \leq\ \|M\|\cdot\|\mathbb{E}[A]\|^{k-1}\ \leq\ \sqrt{d}\,C_\mu\nu_n^{k-1}.$$

Thus including the term $j=i$ in the sum, we pay only a small price of $\|M\mathbb{E}[A]^{k-1}e_i\|_2^2$, which gives

$$\sum_m\widetilde{T}^{\mathrm{hi}}_{im}(2)\ \geq\ c_1\frac{\nu_n}{n}\Big(\sum_{j:\,y_j\in\mathcal{I}_\ell}\|M\,\mathbb{E}[A]^{k-1}e_j\|_2^2-dC_\mu^2\nu_n^{2k-2}\Big).$$

Noting that the sum over $j\in[n]$ is the squared Frobenius norm of $M\mathbb{E}[A]^{k-1}$ finishes the proof.

### E.4 Proof of Lemma 20

Let us write $\bar{\mathbb{1}}_{\mathcal{C}_\ell}=\mathbb{1}_{\mathcal{C}_\ell}/n_\ell$. Let

$$Y_\ell:=\sum_{j:\,y_j=\ell}e_je_j^T,\quad Y_{\mathcal{L}}:=\sum_{\ell\in\mathcal{L}}Y_\ell$$

for any $\mathcal{L} \subset [L]$. Recalling the definition of $E_\ell$ from 5.50, we have $E_\ell = Y_{\mathcal{I}_\ell}$. We note the following identities: For any matrix $H \in \mathbb{R}^{m\times n}$, and $\mathcal{L} \subset [L]$

$$\|HY_{\mathcal{L}}\|_F^2 = \sum_{\ell\in\mathcal{L}} \|HY_\ell\|_F^2, \qquad \|HY_\ell\|_F^2 = \sum_{j:y_j=\ell} \|He_j\|_F^2. \tag{5.59}$$

Next, by the symmetry of the SBM:

**Lemma 26.** *For any matrix $H \in \mathbb{R}^{d\times n}$, and any $k \in \mathbb{N}$, we have*

$$\frac{1}{n}\|H\,\mathbb{E}[A]^k Y_\ell\|_F^2 = \pi_\ell\,\|H\,\mathbb{E}[A]^k \bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2^2.$$

*As a consequence, for any $\mathcal{L} \subset [L]$,*

$$\frac{1}{n}\|H\,\mathbb{E}[A]^k Y_{\mathcal{L}}\|_F^2 = \sum_{\ell\in\mathcal{L}} \pi_\ell\,\|H\,\mathbb{E}[A]^k \bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2^2. \tag{5.60}$$

*Proof.* Expanding along the columns as above, the left-hand side is equal to

$$\frac{1}{n}\sum_{j:\,y_j=\ell} \|H\,\mathbb{E}[A]^k e_j\|_2^2 = \frac{1}{n}\,n_\ell \|H\,\mathbb{E}[A]^k \bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2^2$$

where the equality is by $H\,\mathbb{E}[A]^k e_j = H\,\mathbb{E}[A]^k \bar{\mathbb{1}}_{\mathcal{C}_\ell}$ for any $j \in \mathcal{C}_\ell$, a consequence of the symmetry of SBM. The second claim follows by combining the first with identity (5.59). The proof is complete. □

We also need the following intermediate lemma:

**Lemma 27.** *Under (A1), (A3) and (A4), we have*

$$\frac{\nu_n}{n}\|M\mathbb{E}[A]^{k-1}E_\ell\|_F^2 \;\geq\; \frac{\nu_n^{-1}}{2}\|M\mathbb{E}[A]^k \bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2^2 - dC_\mu^2(L/c_\pi)C_B^2\nu_n^{2k-1-2\delta}.$$

*Proof.* We have

$$\begin{aligned}\|M\,\mathbb{E}[A]^k\,Y_\ell\|_F \;&\leq\; \|M\,\mathbb{E}[A]^{k-1}\,E_\ell\,\mathbb{E}[A]\,Y_\ell\|_F + \|M\,\mathbb{E}[A]^{k-1}(I-E_\ell)\,\mathbb{E}[A]\,Y_\ell\|_F \\ &\leq\; \|\mathbb{E}[A]Y_\ell\|\cdot\|M\,\mathbb{E}[A]^{k-1}\,E_\ell\|_F + \|(I-E_\ell)\mathbb{E}[A]Y_\ell\|\cdot\|M\mathbb{E}[A]^{k-1}\|_F.\end{aligned}$$

For $R \in \mathbb{R}^{m\times n}$, we have $\|R\| \le \|R\|_F \le \sqrt{mn}\|R\|_{\max}$. Let $\mathcal{D}_\ell := \bigcup_{\ell' \in \mathcal{I}_\ell^c} \mathcal{C}_{\ell'}$. Then, $(I - E_\ell)\,\mathbb{E}[A]\,Y_\ell$ is equal to $\mathbb{E}[A]$ on the submatrix indexed by $\mathcal{D}_\ell \times \mathcal{C}_\ell$, and zero elsewhere. Hence by (A1),

$$\|(I - E_\ell)\,\mathbb{E}[A]\,Y_\ell\|_{\max} \;\le\; C_B \nu_n^{1-\delta}/n,$$

and consequently, $\|(I-E_\ell)\,\mathbb{E}[A]\,Y_\ell\| \le \sqrt{n \cdot n_\ell}\, C_B \nu_n^{1-\delta}/n = C_B\sqrt{\pi_\ell}\nu_n^{1-\delta}$. Similarly $\|\mathbb{E}[A]\,Y_\ell\| \le \sqrt{n \cdot n_\ell}\, p_{\max} = \sqrt{\pi_\ell}\,\nu_n$. It follows that

$$\|M\mathbb{E}[A]^k Y_\ell\|_F \;\le\; \sqrt{\pi_\ell}\nu_n \|M\mathbb{E}[A]^{k-1}E_\ell\|_F + \sqrt{\pi_\ell} C_B \nu_n^{1-\delta}\|M\mathbb{E}[A]^{k-1}\|_F.$$

Squaring both sides, using the inequality $(a+b)^2 \le 2(a^2+b^2)$, and multiplying by $\nu_n^{-1}/\pi_\ell n$, we have

$$\frac{\nu_n^{-1}}{\pi_\ell n}\|M\mathbb{E}[A]^k Y_\ell\|_F^2 \;\le\; \frac{2\nu_n}{n}\|M\mathbb{E}[A]^{k-1}E_\ell\|_F^2 + 2C_B^2\nu_n^{1-2\delta}\cdot\frac{1}{n}\|M\mathbb{E}[A]^{k-1}\|_F^2 \tag{5.61}$$

By Lemma 26—specifically, by (5.60) with $Y_{\mathcal{L}} = I$—we have

$$\frac{1}{n}\|M\mathbb{E}[A]^{k-1}\|_F^2 = \sum_\ell \pi_\ell \|M\mathbb{E}[A]^{k-1}\bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2^2 \;\ge\; dC_\mu^2(L/c_\pi)\nu_n^{2k-2}.$$

The inequality follows since for any $\ell \in [L]$,

$$\begin{aligned}\|M\mathbb{E}[A]^{k-1}\mathbb{1}_{\mathcal{C}_\ell}\|_2 &\le \|M\| \cdot \|\mathbb{E}[A]\|^{k-1} \cdot \|\bar{\mathbb{1}}_{\mathcal{C}_\ell}\| \\ &\le \sqrt{d}C_\mu(L/c_\pi)^{1/2}\nu_n^{k-1}\end{aligned}$$

using bounds $\|M\| \le C_\mu\sqrt{nd}$ and $\|\bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2 = (L/c_\pi)^{1/2}n^{-1/2}$ from Lemma 3, and recalling that by definition $\nu_n = np_{\max}$, hence $\|\mathbb{E}[A]\| \le \nu_n$. Combining with (5.61), and rearranging

$$\frac{\nu_n}{n}\|M\mathbb{E}[A]^{k-1}E_\ell\|_F^2 \;\ge\; \frac{\nu_n^{-1}}{2\pi_\ell n}\|M\mathbb{E}[A]^k Y_\ell\|_F^2 - dC_\mu^2(L/c_\pi)C_B^2\nu_n^{2k-1-2\delta}$$

Finally, by Lemma 26, $\frac{1}{n}\|M\mathbb{E}[A]^k Y_\ell\|_F^2 = \pi_\ell\|M\mathbb{E}[A]^k\bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2^2$ establishing the desired result. □

The proof now follows from Lemma 27 and assumption and (A5).

*Proof of Lemma 20.* By Lemma 27,

$$\frac{\nu_n}{n}\sum_\ell \pi_\ell \|M\,\mathbb{E}[A]^{k-1}E_\ell\|_F^2 \geq \Big(\frac{\nu_n^{-1}}{2}\sum_\ell \pi_\ell \|M\,\mathbb{E}[A]^k\bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2^2\Big) - dC_\mu^2(L/c_\pi)C_B^2\nu_n^{2k-1-2\delta}.$$

Next, we replace $\mathbb{E}[A]^k$ with $P^k$ on the LHS, since the price is negligible by Lemma 4. We obtain

$$\begin{aligned}\|M(\mathbb{E}[A]^k - P^k)\bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2 \;&\leq\; \|M\|\cdot\|\mathbb{E}[A]^k - P^k\|\cdot\|\bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2\\ &\leq\; \sqrt{d}\,C_\mu(L/c_\pi)^{1/2}\,k\nu_n^k/n.\end{aligned}$$

using $\|\mathbb{E}[A]^k - P^k\| \leq k\nu_n^k/n$ from Lemma 4. and 3. Using $\|a\|^2 \geq \frac{1}{2}\|b\|^2 - 2\|a-b\|^2$,

$$\|M\,\mathbb{E}[A]^k\bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2^2 \;\geq\; \frac{1}{2}\|M\,P^k\bar{\mathbb{1}}_{\mathcal{C}_\ell}\|_2^2 - 2\big(\sqrt{d}\,C_\mu(L/c_\pi)^{1/2}\,k\big)^2\nu_n^{2k}n^{-2}.$$

Let $C_2 := 2C_\mu^2(L/c_\pi)$ and recall the definition of $\xi_\ell^{(k)} = MP^k\bar{\mathbb{1}}_{\mathcal{C}_\ell}$ from (5.23). Then,

$$\frac{\nu_n}{n}\sum_\ell \pi_\ell\|M\mathbb{E}[A]^{k-1}E_\ell\|_F^2 \;\geq\; \nu_n^{-1}\Big(\frac{1}{2}\sum_\ell \pi_\ell\|\xi_\ell^{(k)}\|_2^2 - d\,C_2\,\nu_n^{2k}\big((n/k)\wedge(\nu_n^\delta/C_B)\big)^{-2}\Big).$$

W.l.o.g. assume $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_L$. Then

$$\sum_\ell \pi_\ell\|\xi_\ell^{(k)}\|_2^2 \;\geq\; \pi_2\sum_{\ell\in\{1,2\}}\|\xi_\ell^{(k)}\|_2^2 \;\geq\; \pi_2\,\frac{1}{2}\|\xi_1^{(k)} - \xi_2^{(k)}\|_2^2.$$

Recalling the identity (5.25), namely $\xi_\ell^{(k)} = \nu_n^k\,\bar{\xi}_\ell^{(k)}$, and invoking assumptions (A3) and (A5),

$$\sum_\ell \pi_\ell\|\xi_\ell^{(k)}\|_2^2 \geq \frac{c_\pi}{2L}c_\xi^2 d\cdot\nu_n^{2k}.$$

Putting the pieces together,

$$\frac{\nu_n}{n}\|M\,\mathbb{E}[A]^{k-1}\|_F^2 \;\geq\; \nu_n^{2k-1}\Big(\frac{c_\pi}{4L}c_\xi^2 d - C_2\,\big((n/k)\wedge(\nu_n^\delta/C_B)\big)^{-2}d\Big) \;\geq\; \nu_n^{2k-1}\frac{c_\pi}{8L}c_\xi^2 d$$

using $\frac{c_\pi}{8L}c_\xi^2 \geq C_2\big((n/k)\wedge(\nu_n^\delta/C_B)\big)^{-2}$ which is equivalent to $(n/k)\wedge(\nu_n^\delta/C_B) \geq 4C_\mu L/(c_\pi c_\xi)$. The proof is complete. □

## F Auxiliary Lemmas

**Lemma 28.** *For (symmetric) binary $A$, we have, for any $U, V \subset [r]$,*

$$\mathbb{E}[A_{\boldsymbol{w}^V}] \cdot \prod_{u \in U} \mathbb{E}[A_{\boldsymbol{w}^u}] \;\le\; \mathbb{E}[A_{\boldsymbol{w}^{U \cup V}}].$$

*Proof.* Since the random variables $A_{ij}$ are binary, for every edge $(i,j)$ and $a, b \in \mathbb{N}$ we have $\big(\mathbb{E}[A_{ij}^a]\big)^b \le \mathbb{E}[A_{ij}]$. We have $\mathbb{E}[A_{\boldsymbol{w}^u}] = \prod_{e \in [\boldsymbol{w}^u]} \mathbb{E}[A_e]$, using independence of $A_e, e \in [\boldsymbol{w}^u]$. Similarly,

$$\mathbb{E}[A_{\boldsymbol{w}^U}] = \mathbb{E}\Big[\prod_{u \in U} A_{\boldsymbol{w}^u}\Big] \overset{(a)}{=} \mathbb{E}\Big[\prod_{e \in [\boldsymbol{w}^U]} A_e\Big] \overset{(b)}{=} \prod_{e \in [\boldsymbol{w}^U]} \mathbb{E}[A_e] \tag{5.62}$$

where (a) is by $A_e^a = A_e$ and (b) is by independence. This in turn implies that for any $U \subset [r]$,

$$\prod_{u \in U} \mathbb{E}[A_{\boldsymbol{w}^u}] = \prod_{u \in U,\, e \in [\boldsymbol{w}^u]} \mathbb{E}[A_e] \le \prod_{e \in [\boldsymbol{w}^U]} \mathbb{E}[A_e] = \mathbb{E}[A_{\boldsymbol{w}^U}]$$

where the inequality is by $(\mathbb{E}[A_e])^b \le \mathbb{E}[A_e]$ and the final equality is by (5.62). We obtain

$$\begin{aligned}
\mathbb{E}[A_{\boldsymbol{w}^V}] \cdot \prod_{u \in U} \mathbb{E}[A_{\boldsymbol{w}^u}] &\le \mathbb{E}[A_{\boldsymbol{w}^V}] \cdot \mathbb{E}[A_{\boldsymbol{w}^U}] \\
&= \Big( \prod_{e \in [\boldsymbol{w}^V]} \mathbb{E}[A_e] \Big) \cdot \prod_{e' \in [\boldsymbol{w}^U]} \mathbb{E}[A_{e'}] \\
&\le \prod_{e \in [\boldsymbol{w}^V] \cup [\boldsymbol{w}^U]} \mathbb{E}[A_e]
\end{aligned}$$

where the final inequality again uses $(\mathbb{E}[A_e])^b \le \mathbb{E}[A_e]$. Using $[\boldsymbol{w}^U] \cup [\boldsymbol{w}^V] = [\boldsymbol{w}^{U \cup V}]$ and (5.62) finishes the proof. □

**Lemma 29.** *For any collection of random variable $\{X_i\}_{i \in I}$ and positive numbers $\{a_i\}_{i \in I}$ such that $\sum_i a_i = a$, we have*

$$\prod_i (\mathbb{E}|X_i|^{a_i}) \le \max_i \mathbb{E}|X_i|^a$$

*Proof.* Let $\lambda_i = a_i/a$ so that $\sum_i \lambda_i = 1$. By Jensen's inequality

$$\mathbb{E}|X_i|^{a_i} \leq (\mathbb{E}|X_i|^{a_i/\lambda_i})^{\lambda_i}.$$

Then,

$$\prod_i (\mathbb{E}|X_i|^{\alpha_i}) \leq \prod_i (\mathbb{E}|X_i|^{a_i/\lambda_i})^{\lambda_i} \leq \max_i \mathbb{E}|X_i|^{a_i/\lambda_i}$$

where the last step uses the elementary inequality $\prod_i |z_i|^{\lambda_i} \leq \max |z_i|$. □

**Lemma 30.** *Let* $r_n = \max\{r \in 2\mathbb{N} : c(ar)^r \leq \nu_n^b\}$ *for* $a, c \geq 1$ *and* $b > 0$. *Then,*

$$r_n \gtrsim \frac{b \log \nu_n}{\log(abc) + \log\log \nu_n}.$$

*Proof.* To lowerbound $r_n$, evaluate the inequality $(acr_n)^{r_n} \leq \nu_n^b$. Redefine $a \coloneqq ac$. Let $r \in \mathbb{R}$ such that $(ar)^{ar} = \nu_n^{ab}$. Then, $r_n \asymp r$. Recall that $ye^y = x$ has solution $y = W_0(x)$, the Lambert function. Let $y = \log(ar)$, so that $ye^y = \log((ar)^{ar}) = ab \log(\nu_n)$. Then,

$$ar = e^{W_0(ab \log \nu_n)} = \frac{ab \log \nu_n}{W_0(ab \log \nu_n)} \geq \frac{ab \log \nu_n}{\log(ab \log \nu_n)}$$

since $W_0(x) < \log(x)$. Simplifying and relating $r$ to $r_n$ completes the proof. □

# CHAPTER 6

# A CLT for Polynomial GNNs on Community-Based Graphs

## 1 Introduction

Graph Neural Networks (GNNs) are now a key tool for machine learning on graphs. Their success is largely due to the graph convolution operation—also known as message passing or neighbor aggregation—where node features are updated by gathering information from their graph neighbors [GSR17, KW17, YCS16]. This process helps GNNs learn powerful embeddings for tasks like node classification and regression. For graphs with community structure, theory shows that even one aggregation step can improve feature separation between classes by a factor of $\sqrt{\nu_n}$, where $\nu_n$ is the average node degree [BFJ21].

Analyzing deep GNNs with multiple aggregation layers ($k > 1$) is important but theoretically difficult. Unlike single aggregations, the resulting features, $\phi^{(k)}$, lose desirable properties such as entry-wise independence. To study these multi-aggregated features, researchers have used techniques like walk-based decompositions, which classify feature contributions by underlying graph walk patterns [CSD25, VA24b]. For community-based graphs, these methods suggest that while feature cluster centers can separate at a rate of $\nu_n^k$, their standard deviation often grows as $\nu_n^{k-1/2}$.

This paper focuses on Polynomial GNNs (Poly-GNNs). In these models, features $\phi^{(k)} = A^k X$ are created by applying the adjacency matrix $A$, $k$ times to initial node features $X \in \mathbb{R}^{n \times d}$, without any non-linear functions in between. These features $\phi^{(k)}$, when passed

through a final linear layer $W$, produce classification scores. Poly-GNNs, despite their simplicity, are not just theoretical ideas. They form the basis of, or are similar to, several practical and effective GNNs like APPNP [KBG19], GPR-GNN [CPL21], and models using Chebyshev or Jacobi polynomials [DBV16, WZ22]. Such models have achieved strong results, sometimes state-of-the-art, on standard benchmarks [WZ22]. Therefore, understanding Poly-GNN features offers valuable insights into multi-hop aggregation and the behavior of these common GNN types.

### 1.1 Overview of Our Contributions

In this paper, we undertake a detailed asymptotic analysis of the embeddings generated by $k$-layer Poly-GNNs on community-based graphs as the number of nodes $n$ grows. To stabilize these features, we consider two types of normalized embeddings: the degree-normalized features $\overline{\phi}_i^{(k)} := \phi_i^{(k)}/\nu_n^k$, and the centered and scaled features $\xi_i^{(k)} := \sqrt{\nu_n}(\overline{\phi}_i^{(k)} - \mathbb{E}[\overline{\phi}_i^{(k)}])$. Here, $\nu_n$ is the average degree parameter which we assume tends to infinity. One of our main results is a Central Limit Theorem (CLT) demonstrating that the empirical distribution of $\xi_i^{(k)}$ converges in 1-Wasserstein distance to a centered mixture of multivariate Gaussian distributions.

Building upon this, we further demonstrate that the joint empirical distribution of the uncentered, degree-normalized features $\overline{\phi}_i^{(k)}$ (which are directly used in downstream classifiers) and their corresponding true labels $z_i$ also converges in the 1-Wasserstein distance. Specifically, as $n \to \infty$ and $\nu_n \to \infty$, this distribution approaches that of a random pair $(Z, Y_n)$ where $Z \sim \pi$ (the limiting class proportions) and $Y_n$ conditioned on $Z = \ell$ follows a multivariate Gaussian distribution $N(\mu_\ell, \Sigma_\ell/\nu_n)$. A core contribution of our work is the precise analytical characterization of these limiting class means $\mu_\ell$ and class-conditional covariance matrices $\Sigma_\ell$, expressed in terms of the graph's community structure and initial feature means.

This characterization has profound implications for understanding the training dynamics of GNNs. We prove that training a linear classifier on these Poly-GNN features $\overline{\phi}_i^{(k)}$ using

the standard cross-entropy (CE) loss converges to the equivalent optimization problem on this limiting Gaussian mixture. This convergence holds uniformly for the loss function, the gradient path during optimization, and the final learned classifier weights (under mild conditions on weight norms), due to the Lipschitz nature of the CE loss and its gradients with respect to the features. This result provides a strong theoretical basis for the behavior observed when training linear classifiers on GNN embeddings.

Furthermore, our explicit forms for $\mu_\ell$ and $\Sigma_\ell$ reveal a clear and precise mechanism behind the well-known phenomenon of GNN oversmoothing. The mean vectors $\mu_\ell$ involve terms of the form $(J^k M)^T$, while the covariance matrices $\Sigma_\ell$ involve $(J^{k-1}M)^T$, where $J$ is a matrix derived from the graph's inter-community edge probabilities and class proportions, and $M$ represents the initial class feature means. As the GNN depth $k$ increases, the repeated matrix exponentiation $J^k$ (and $J^{k-1}$) acts like a power iteration. This causes both the class means and the dominant eigen-directions of the class covariances to align with a low-dimensional (often 1-D) subspace determined by the leading eigenvector(s) of $J$. Consequently, the feature distributions for different classes, initially potentially well-separated in $d$ dimensions, collapse onto this common, typically 1-D, subspace. This results in a degenerate, poorly separated Gaussian mixture, thereby degrading classification performance. Our analysis, thus, provides a nuanced, quantitative view of oversmoothing in the sparse, large-graph limit.

**Paper Structure** The remainder of this paper is organized as follows. Section 2 introduces the Poly-GNN framework, the community-based graph model, key notation, and the assumptions underlying our analysis. Section 3 presents our main Central Limit Theorems, formally stating the convergence of Poly-GNN embeddings to specific Gaussian mixtures and outlining the key steps of their proofs. Section 4 details the significant implications of these CLTs, discussing the convergence of linear classification tasks performed on these embeddings and providing a detailed analysis of how our results explain the GNN oversmoothing phenomenon. Finally, Section 5 concludes the paper and summarizes our findings. Detailed proofs and

auxiliary technical results are provided in the Appendices.

## 2 Preliminaries and Model Setup

In this section, we formally define the Polynomial GNN (Poly-GNN) architecture, introduce the normalized features central to our analysis, describe the community-based graph model, state our key assumptions, and briefly define the Wasserstein distance used to quantify distributional convergence.

### 2.1 Poly-GNNs and Feature Definitions

We consider a simple yet powerful class of Graph Neural Networks known as Polynomial GNNs (Poly-GNNs). Given an undirected graph with $n$ nodes, represented by its adjacency matrix $A \in \{0,1\}^{n\times n}$, and initial node features $X \in \mathbb{R}^{n\times d}$, a $k$-layer Poly-GNN computes node embeddings, or features, $\phi^{(k)} \in \mathbb{R}^{n\times d}$ through $k$ successive aggregations:

$$\phi^{(k)} = A^k X. \tag{6.1}$$

The $i$-th row of $\phi^{(k)}$, denoted $\phi_i^{(k)} \in \mathbb{R}^d$, represents the embedding for node $i$ after $k$ layers of aggregation.

For our asymptotic analysis, we work with normalized versions of these features. Let $\nu_n$ be the average degree parameter of the graph, which we assume grows with $n$ (see Assumption 2). We define the *degree-normalized features* as:

$$\overline{\phi}_i^{(k)} := \frac{\phi_i^{(k)}}{\nu_n^k}, \quad i = 1, \ldots, n. \tag{6.2}$$

These features $\overline{\phi}_i^{(k)}$ are often the direct input to a downstream classifier.

To establish a stable limiting distribution under a Central Limit Theorem, we further define the *centered and scaled features*:

$$\xi_i^{(k)} := \sqrt{\nu_n}\left(\overline{\phi}_i^{(k)} - \mathbb{E}[\overline{\phi}_i^{(k)}]\right), \quad i = 1, \ldots, n. \tag{6.3}$$

The empirical distribution of these features, $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i^{(k)}}$, where $\delta_x$ is a point mass at $x$, will be a primary object of study.

## 2.2 Community-Based Graph Model

We assume the graph and its node features are generated from a community-based model. Let $z = (z_i)_{i=1}^n \in [L]^n$ be a vector of latent node labels, assigning each node $i$ to one of $L$ communities or classes. The graph structure and initial feature distributions are conditional on these labels.

Specifically, we adopt the Contextual Stochastic Block Model (CSBM) [DSM18]. The adjacency matrix $A$ is generated such that edges are conditionally independent given $z$, with probabilities:

$$A_{ij} \sim \text{Bern}(\nu_n B_{z_i z_j}/n) \quad \text{for } i \neq j, \text{ and } A_{ii} = 0, \tag{6.4}$$

where $B \in [0,1]^{L\times L}$ is a symmetric matrix of inter-community edge probability scalings. The parameter $\nu_n/n$ represents the average edge density scale.

The initial node features $X_i \in \mathbb{R}^d$ are assumed to be conditionally independent given $z_i$. Their expectations are determined by their class membership:

$$\mathbb{E}[X_i \mid z_i = \ell] = M_{\ell,\cdot}, \tag{6.5}$$

where $M \in \mathbb{R}^{L\times d}$ is a matrix whose $\ell$-th row, $M_{\ell,\cdot}$, is the mean feature vector for class $\ell$. This can be written compactly as $\mathbb{E}[X] = ZM$, where $Z \in \{0,1\}^{n\times L}$ is the one-hot encoding matrix of the labels $z$, i.e., $Z_{i\ell} = \mathbb{1}\{z_i = \ell\}$.

We define $\pi = (\pi_1, \ldots, \pi_L)^T$ as the vector of limiting class proportions (see Assumption 4). Let $\Pi = \text{diag}(\pi_1, \ldots, \pi_L)$ be the diagonal matrix of these proportions. A key matrix in our analysis is $J \in \mathbb{R}^{L\times L}$, defined as:

$$J = B\Pi. \tag{6.6}$$

This matrix captures the interplay between inter-community connectivity $B$ and class sizes $\Pi$.

### 2.3 Assumptions

Our theoretical results rely on the following assumptions:

**Assumption 2** (Degree Growth)**.** *The average degree parameter $\nu_n \to \infty$ as $n \to \infty$. Moreover, the expected degree of any node $i$, $\sum_{j \neq i} \nu_n B_{z_i z_j}/n$, is $O(\nu_n)$, and $\nu_n B_{\ell\ell'}/n \leq \nu_n C/n$ for some constant $C$ (i.e., $p_{ij} := \mathbb{E}[A_{ij}] \lesssim \nu_n/n$).*

**Assumption 3** (Sparse Graph)**.** *The graph is sparse, meaning $\nu_n = o(n)$.*

This assumption simplifies the analysis by ensuring that noise from the graph's randomness dominates certain terms. While some results might extend to denser regimes, sparsity is maintained for clarity.

**Assumption 4** (Cluster Convergence)**.** *For each class $\ell \in [L]$, let $\mathcal{C}_\ell = \{i \in [n] : z_i = \ell\}$ be the set of nodes in class $\ell$. We assume there exist $\pi_\ell > 0$ such that $\sqrt{\nu_n}\big(\pi_\ell - |\mathcal{C}_\ell|/n\big) = o(1)$, and $\sum_{\ell=1}^{L} \pi_\ell = 1$.*

**Assumption 5** (Feature Bounds)**.** *The initial node features $X_i$ are sub-gaussian. Specifically, for any unit vector $u \in \mathbb{R}^d$, $(X_i - \mathbb{E}[X_i])u \sim SG(\sigma^2)$ for some $\sigma^2 > 0$ uniformly for all $i, n$. Furthermore, their expected norms are uniformly bounded: $\limsup_{n \geq 1} \sup_{i \in [n]} \mathbb{E}\|X_i\|_2 \leq x_*$ for some finite $x_* \geq 0$.*

### 2.4 Wasserstein Distance

To measure the distance between probability distributions, we use the 1-Wasserstein distance, denoted $W_1(\mathbb{P}, \mathbb{Q})$. For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathbb{R}^d$, the Kantorovich-Rubinstein duality provides a convenient definition:

$$W_1(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathrm{Lip}(1)} \left| \int f d\mathbb{P} - \int f d\mathbb{Q} \right|, \tag{6.7}$$

where Lip(1) is the class of all 1-Lipschitz functions $f : \mathbb{R}^d \to \mathbb{R}$, i.e., functions satisfying $|f(x) - f(y)| \leq \|x - y\|_2$ for all $x, y \in \mathbb{R}^d$. We also write $\mathbb{P}f = \int f d\mathbb{P}$ for the expectation of

$f$ under $\mathbb{P}$. Convergence in $W_1$ implies weak convergence and convergence of first moments. Its connection to Lipschitz functions makes it particularly relevant for analyzing learning algorithms with Lipschitz loss functions.

# 3 Asymptotic Distribution of Poly-GNN Embeddings

In this section, we present our main theoretical results concerning the asymptotic distribution of Poly-GNN embeddings. We establish Central Limit Theorems (CLTs) for both the degree-normalized features $\overline{\phi}_i^{(k)}$ (jointly with their labels) and the centered-and-scaled features $\xi_i^{(k)}$. We then outline the key steps involved in proving these theorems, highlighting the key intermediate lemmas and propositions.

## 3.1 Main Central Limit Theorems

Our first main result characterizes the joint limiting distribution of the true node labels $z_i$ and the degree-normalized Poly-GNN features $\overline{\phi}_i^{(k)}$. These features are typically used for downstream classification tasks.

**Theorem 14** (CLT for Degree-Normalized Features and Labels)**.** *Let $(A, X)$ be a community-based graph satisfying Assumptions 2–5. Let $\overline{\phi}_i^{(k)} = \phi_i^{(k)}/\nu_n^k$ be the degree-normalized $k$-layer Poly-GNN features. Define the limiting class means $\mu_\ell \in \mathbb{R}^d$ and class-conditional covariance matrices $\Sigma_\ell \in \mathbb{R}^{d\times d}$ as:*

$$\mu_\ell := (J^k M)^T e_\ell, \tag{6.8}$$

$$\Sigma_\ell := (J^{k-1}M)^T \operatorname{diag}(e_\ell^T J)(J^{k-1}M), \tag{6.9}$$

*where $e_\ell$ is the $\ell$-th canonical unit vector in $\mathbb{R}^L$, $J = B\Pi$, and $M$ contains the initial class feature means. Let $\widetilde{\mathbb{P}}_n^{joint}$ be the empirical distribution of pairs $(z_i, \overline{\phi}_i^{(k)})$: $\widetilde{\mathbb{P}}_n^{joint} = \frac{1}{n}\sum_{i=1}^n \delta_{(z_i, \overline{\phi}_i^{(k)})}$. Let $\mathbb{G}_n^{joint}$ be the probability distribution of a random pair $(Z, Y_n)$ where $Z \sim$ Categorical$(\pi_1, \ldots, \pi_L)$ and, conditioned on $Z = \ell$, $Y_n \sim N(\mu_\ell, \Sigma_\ell/\nu_n)$. Then, as*

$n \to \infty$:

$$\mathbb{E}\left[W_1\left(\widetilde{\mathbb{P}}_n^{joint}, \mathbb{G}_n^{joint}\right)\right] \to 0. \tag{6.10}$$

*Furthermore, this convergence holds in the stronger class-conditional sense: for any* $R > 0$,

$$\lim_{n\to\infty} \mathbb{E}\left\{\sup_{f_1,\ldots,f_L \in \mathrm{Lip}(R)} \left|\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in\mathcal{C}_\ell} f_\ell(\overline{\phi}_i^{(k)}) - \sum_{\ell=1}^{L} \pi_\ell \mathbb{E}_{Y\sim N(\mu_\ell, \Sigma_\ell/\nu_n)}[f_\ell(Y)]\right|\right\} = 0. \tag{6.11}$$

Theorem 14 shows that for large $n$ and $\nu_n$, the features $\overline{\phi}_i^{(k)}$ behave as if drawn from a Gaussian mixture where each component $\ell$ is centered at $\mu_\ell$ and has a covariance $\Sigma_\ell/\nu_n$ that vanishes as $\nu_n \to \infty$.

Our second main result provides a CLT for the centered and scaled features $\xi_i^{(k)}$, showing they converge to a stable (non-degenerate variance) Gaussian mixture.

**Theorem 15** (CLT for Centered and Scaled Features)**.** *Under the same conditions as Theorem 14, let* $\xi_i^{(k)} = \sqrt{\nu_n}(\overline{\phi}_i^{(k)} - \mathbb{E}[\overline{\phi}_i^{(k)}])$ *be the centered and scaled features. Let* $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i^{(k)}}$ *be their empirical distribution. Let* $\mathbb{G}$ *be the centered Gaussian mixture distribution:*

$$\mathbb{G} = \sum_{\ell=1}^{L} \pi_\ell \, N(0, \Sigma_\ell), \tag{6.12}$$

*where* $\Sigma_\ell$ *is defined in Eq.* (6.9)*. Then, as* $n \to \infty$*:*

$$\mathbb{E}\left[W_1(\mathbb{P}_n, \mathbb{G})\right] \to 0. \tag{6.13}$$

*Furthermore, this convergence also holds in the stronger class-conditional sense: for any* $R > 0$,

$$\lim_{n\to\infty} \mathbb{E}\left\{\sup_{f_1,\ldots,f_L \in \mathrm{Lip}(R)} \left|\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in\mathcal{C}_\ell} f_\ell(\xi_i^{(k)}) - \sum_{\ell=1}^{L} \pi_\ell \mathbb{E}_{Y\sim N(0, \Sigma_\ell)}[f_\ell(Y)]\right|\right\} = 0. \tag{6.14}$$

Theorem 15 establishes that after appropriate centering and scaling, the Poly-GNN features converge to a mixture of Gaussians, each component having a non-vanishing covariance $\Sigma_\ell$.

### 3.2 Proof Outline and Key Steps

The proofs of Theorems 14 and 15 share a common foundation and proceed in several steps. We outline the general strategy here, focusing on the convergence of $\mathbb{P}_n$ to $\mathbb{G}$ (Theorem 15). The argument for Theorem 14 builds on Theorem 15 with adjustments for the non-zero means and the $\nu_n^{-1}$ scaling in the covariance. The full proofs are provided in Appendix A.

The overall strategy involves two main parts for establishing $\mathbb{E}[W_1(\mathbb{P}_n, \mathbb{G})] \to 0$:

1. Show that the empirical measure $\mathbb{P}_n$ concentrates around its expectation $\overline{\mathbb{P}}_n := \mathbb{E}[\mathbb{P}_n]$, i.e., $\mathbb{E}[W_1(\mathbb{P}_n, \overline{\mathbb{P}}_n)] \to 0$.

2. Show that the expected empirical measure $\overline{\mathbb{P}}_n$ converges to the target Gaussian mixture $\mathbb{G}$ in $W_1$ distance, i.e., $W_1(\overline{\mathbb{P}}_n, \mathbb{G}) \to 0$.

The argument for class-conditional convergence (e.g., Eq. (6.11)) builds upon this by considering per-class empirical measures and leveraging the convergence of class proportions $|\mathcal{C}_\ell|/n \to \pi_\ell$.

The key technical steps involve analyzing the moments of the features:

**Step 1: Moment Analysis for General Graphs** This step characterizes the behavior of feature moments without yet imposing the full community structure, relying mainly on Assumptions 2, 3, and 5.

- The centered, un-normalized features $\phi_i^{(k)} - \mathbb{E}[\phi_i^{(k)}]$ are decomposed into two terms: $\mathring{\Delta}_i$ (due to graph randomness) and $\mathring{\Lambda}_i$ (due to initial feature randomness):

$$\phi_i^{(k)} - \mathbb{E}[\phi_i^{(k)}] = \mathring{\Delta}_i + \mathring{\Lambda}_i. \tag{6.15}$$

  Normalizing appropriately, $\xi_i^{(k)} = (\mathring{\Delta}_i + \mathring{\Lambda}_i)/\nu_n^{k-1/2}$.

- The term $\Lambda_i := \mathring{\Lambda}_i/\nu_n^{k-1/2}$ is shown to be asymptotically negligible under our sparsity assumption (see Proposition 9 in Appendix A.3). Thus, $\xi_i^{(k)}$ is asymptotically equivalent to $\Delta_i := \mathring{\Delta}_i/\nu_n^{k-1/2}$ in terms of its contribution to moments (see Lemma 35 in Appendix A.3).

- The moments of $\Delta_{i,\theta} := \langle \Delta_i, \theta \rangle$ for any unit vector $\theta \in \mathbb{R}^d$ are analyzed.
  - Odd moments: $\mathbb{E}[\Delta_{i,\theta}^r] \to 0$ for odd $r$ (this follows from the moment bounds in Proposition 10, specifically the term $\nu_n^{p/2-\lceil p/2 \rceil}$, which is $\nu_n^{-1/2}$ for odd $p = r$).
  - Even moments: $\mathbb{E}[\Delta_{i,\theta}^r] \to (r-1)!! \cdot \widetilde{\sigma}_{i,\theta}^r$ for even $r$, where $\widetilde{\sigma}_{i,\theta}^2 := \|V_i \mathbb{E}[A/\nu_n]^{k-1} \mathbb{E}[X]\theta\|_2^2$ (see Lemma 37 in Appendix A.3).
- The expected normalized mean $\mathbb{E}[\overline{\phi}_i^{(k)}]$ is shown to converge to a limit $\gamma_i = e_i^T (\mathbb{E}[A/\nu_n])^k \mathbb{E}[X]$ (see Lemma 33 in Appendix A.2).

**Step 2: Specialization to Community-Based Graphs** Here, the community structure (Assumptions 4 and the CSBM formulation) is used to refine the limiting moments.

- The limiting mean $\gamma_i$ for a node $i \in \mathcal{C}_\ell$ converges to $\mu_\ell = (J^k M)^T e_\ell$ (as detailed in the proof of Proposition 11 in Appendix A.4, building on Lemma 33).
- The average of the per-node variances $\widetilde{\sigma}_{i,\theta}^2$ over class $\ell$ converges to $\theta^T \Sigma_\ell \theta$, where $\Sigma_\ell$ is defined in Eq. (6.9) (this is part of the derivation in the proof of Proposition 11).
- Consequently, the $r$-th moment of the $\theta$-projection of $\overline{\mathbb{P}}_n$, $m_r(\overline{\mathbb{P}}_{n,\theta}) = \frac{1}{n} \sum_i \mathbb{E}[\langle \xi_i^{(k)}, \theta \rangle^r]$, converges to $m_r(\mathbb{G}_\theta) = \sum_{\ell=1}^L \pi_\ell \mathbb{E}_{Y \sim N(0, \theta^T \Sigma_\ell \theta)}[Y^r]$ (see Proposition 11 in Appendix A.4).

Since the Gaussian mixture $\mathbb{G}_\theta$ is determined by its moments, this establishes that $\overline{\mathbb{P}}_{n,\theta} \rightsquigarrow \mathbb{G}_\theta$. Uniform integrability of moments (derived from the $\Psi_r$ norm bounds in appendix C, specifically Lemma 39, applied to $\Delta_{i,\theta}$ via Proposition 10) then promotes this weak convergence to $W_1(\overline{\mathbb{P}}_{n,\theta}, \mathbb{G}_\theta) \to 0$. A discretization argument (Proposition 14 from Appendix B) and Proposition 16 (from Appendix D) extend this to $W_1(\overline{\mathbb{P}}_n, \mathbb{G}) \to 0$.

**Step 3: Concentration and Convergence of Empirical Measure.** To show $\mathbb{E}[W_1(\mathbb{P}_n, \overline{\mathbb{P}}_n)] \to 0$, we rely on:

- Control over the variance of empirical moments: $\mathrm{Var}(n^{-1}\sum_{i=1}^{n}\langle\Delta_{i,\theta}\rangle^r) \lesssim n^{-1}$ (see Lemma 36 in Appendix A.3, which implies similar behavior for $\xi_i^{(k)}$ via Lemma 35). This corresponds to condition (b) of Proposition 17 in Appendix D.

- Tail control for $\langle\xi_i^{(k)},\theta\rangle$: The features $\langle\xi_i^{(k)},\theta\rangle$ are shown to be uniformly $\Psi_{r_n}$ sub-Gaussian for a growing $r_n$ (see Lemma 31 in Appendix A.1). This corresponds to condition (a) of Proposition 17.

- Uniform integrability of moments of $\overline{\mathbb{P}}_n$ (the convergence shown in Proposition 11 implies that for any fixed $r$, $\sup_n m_r(\overline{\mathbb{P}}_{n,\theta})$ is finite, which by Proposition 14 implies $\sup_n M_r(\overline{\mathbb{P}}_n)$ is finite, e.g. $M_1(\overline{\mathbb{P}}_n)$ needed for condition (c) of Proposition 17).

These conditions allow the application of Proposition 17 (from Appendix D), which establishes the desired concentration $\mathbb{E}[W_1(\mathbb{P}_n,\overline{\mathbb{P}}_n)] \to 0$. The triangle inequality for $W_1$ then combines these two main parts to yield the final convergence result.

# 4 Implications for Classification and GNN Oversmoothing

The Central Limit Theorems presented in Section 3 not only provide a fundamental understanding of the distributional properties of Poly-GNN embeddings but also have significant practical implications. In this section, we explore two key consequences: first, how our results explain the convergence of linear classifiers trained on these embeddings, and second, how they offer a precise, quantitative mechanism for the GNN oversmoothing phenomenon.

## 4.1 Convergence of Linear Classification on Poly-GNN Features

In many node classification tasks, GNN embeddings are fed into a final linear layer (often followed by a softmax activation) that is trained using a cross-entropy (CE) loss. Our results provide a theoretical basis for understanding this training process in the asymptotic limit. We focus on the degree-normalized features $\overline{\phi}_i^{(k)}$, as these are the quantities typically used by

the classifier.

Recall from Theorem 14 that the joint empirical distribution of labels $z_i$ and features $\overline{\phi}_i^{(k)}$ converges to that of $(Z, Y_n)$, where $Z \sim \text{Categorical}(\pi)$ and $Y_n \mid Z = \ell \sim N(\mu_\ell, \Sigma_\ell/\nu_n)$. The class means $\mu_\ell$ and covariances $\Sigma_\ell$ are given by Eqs. (6.8) and (6.9), respectively.

Consider a linear classifier with weights $W = (w_1, \ldots, w_L)^T \in \mathbb{R}^{L\times d}$ and biases $b = (b_1, \ldots, b_L)^T \in \mathbb{R}^L$. The empirical cross-entropy loss for a dataset of $n$ nodes is:

$$\mathfrak{L}_{\text{emp}}(W, b) := -\frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{L} \mathbb{1}\{z_i = \ell\} \log \frac{\exp(w_\ell^T \overline{\phi}_i^{(k)} + b_\ell)}{\sum_{u=1}^{L} \exp(w_u^T \overline{\phi}_i^{(k)} + b_u)}. \tag{6.16}$$

The corresponding limiting loss, based on the Gaussian mixture (GM) characterization from Theorem 14, is:

$$\mathfrak{L}_{\text{GM}}(W, b) := -\sum_{\ell=1}^{L} \pi_\ell \mathbb{E}_{Y\sim N(\mu_\ell, \Sigma_\ell/\nu_n)} \left[\log \frac{\exp(w_\ell^T Y + b_\ell)}{\sum_{u=1}^{L} \exp(w_u^T Y + b_u)}\right]. \tag{6.17}$$

For any fixed set of weights $(W, b)$ (e.g., within a ball $\|(W, b)\|_F \le \mathcal{R}$ for some radius $\mathcal{R}$), the individual loss term for class $\ell$, $\text{CE}_\ell(x; W, b) = -\log \frac{\exp(w_\ell^T x + b_\ell)}{\sum_{u=1}^{L} \exp(w_u^T x + b_u)}$, is Lipschitz with respect to the feature $x$. This Lipschitz property, combined with the 1-Wasserstein convergence established in Theorem 14 (specifically, the class-conditional form Eq. (6.11), leads to the following key result:

**Proposition 8** (Convergence of CE Loss and Gradients)**.** *Under the conditions of Theorem 14, for any fixed radius $\mathcal{R} > 0$:*

*(a) The empirical CE loss converges uniformly to the limiting GM CE loss:*

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{\|(W,b)\|_F \le \mathcal{R}} |\mathfrak{L}_{emp}(W, b) - \mathfrak{L}_{GM}(W, b)|\right] = 0. \tag{6.18}$$

*(b) The gradients of the empirical CE loss converge uniformly to the gradients of the limiting GM CE loss:*

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{\|(W,b)\|_F \le \mathcal{R}} \left\|\nabla_{(W,b)}\mathfrak{L}_{emp}(W, b) - \nabla_{(W,b)}\mathfrak{L}_{GM}(W, b)\right\|_F\right] = 0. \tag{6.19}$$

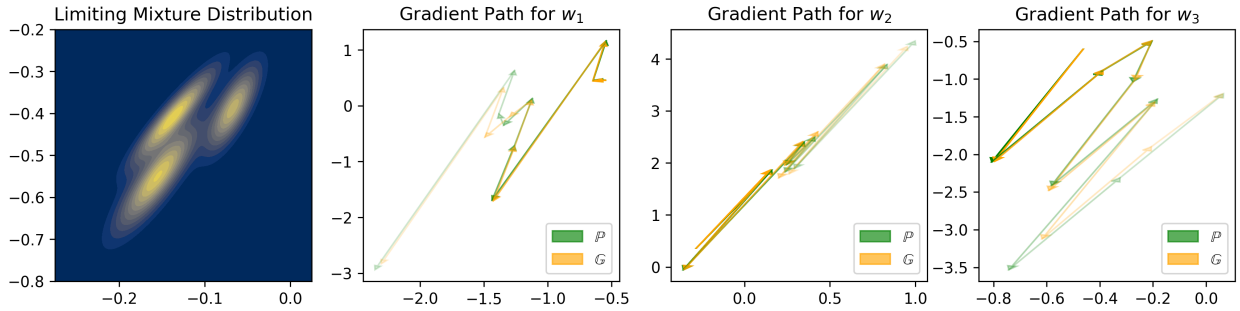



Figure 6.1: Ten gradients steps of cross-entropy optimization problem for $(A, X)$ drawn from a 3-class CSBM. Shown on the right are gradient paths for samples drawn from empirical and theoretical distributions for $\overline{\phi}^{(k)}$.

*Consequently, the sequence of parameters* $(W^*_{emp}, b^*_{emp})$ *minimizing* $\mathfrak{L}_{emp}(W, b)$ *within the ball converges in probability to the parameters* $(W^*_{GM}, b^*_{GM})$ *minimizing* $\mathfrak{L}_{GM}(W, b)$ *within the same ball, assuming uniqueness of the minimizer for the limiting problem.*

The proof of (b) relies on the fact that the gradients $\nabla_x \mathrm{CE}_\ell(x; W, b)$ are also Lipschitz in $x$ for bounded $(W, b)$. Proposition 8 formalizes the intuition that training a Poly-GNN with CE loss is asymptotically equivalent to performing CE optimization directly on the identified Gaussian mixture. This explains why gradient descent paths on the empirical loss track those on the limiting GM loss, as illustrated in Figure 6.1.

The stationarity conditions for the limiting optimization problem $\mathfrak{L}_{\mathrm{GM}}(W, b)$ reveal a moment-matching structure:

$$\pi_\ell \mu_\ell = \sum_{u=1}^{L} \pi_u \mathbb{E}_{Y \sim N(\mu_u, \Sigma_u/\nu_n)}[Y \cdot \widehat{p}_\ell(Y; W, b)], \tag{6.20}$$

$$\pi_\ell = \sum_{u=1}^{L} \pi_u \mathbb{E}_{Y \sim N(\mu_u, \Sigma_u/\nu_n)}[\widehat{p}_\ell(Y; W, b)], \tag{6.21}$$

for all $\ell \in [L]$, where $\widehat{p}_\ell(Y; W, b)$ is the softmax probability $\exp(w_\ell^T Y + b_\ell)/\sum_j \exp(w_j^T Y + b_j)$. It is important to note that while the GNN training process converges to this CE solution on the GM, this solution is not necessarily the Bayes optimal classifier for the Gaussian mixture

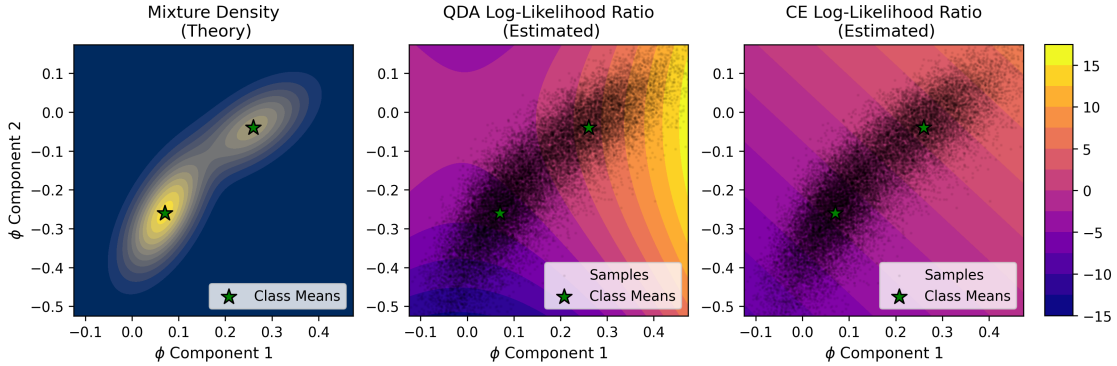



Figure 6.2: Classifier comparison for data which is 2-dimensional CSBM. On the left is the theoretical density of the 2-class CSBM. The two right plots are the estimated log-likelihood ratios for the QDA and CE estimator respectively. The slight bend in the data is correctly captured by the QDA estimator.

itself (which would be a Quadratic Discriminant Analysis, QDA, classifier). Figure 6.2 illustrates this, showing that even for large $n$ where $\overline{\phi}^{(k)}$ closely follows the GM, the linear CE boundary can differ from the optimal QDA boundary.

## 4.2 A Precise Mechanism for GNN Oversmoothing

The phenomenon of oversmoothing, where the performance of GNNs degrades as the number of layers $k$ increases, is a well-documented empirical observation. Our explicit characterization of the limiting class means $\mu_\ell$ and covariances $\Sigma_\ell$ provides a precise analytical explanation for this behavior in Poly-GNNs.

Recall the expressions from Eqs. (6.8) and (6.9):

$$\mu_\ell = (J^k M)^T e_\ell,$$
$$\Sigma_\ell = (J^{k-1} M)^T \operatorname{diag}(e_\ell^T J)(J^{k-1} M).$$

Consider the symmetric matrix $J_{\text{sym}} = \Pi^{1/2} B \Pi^{1/2}$, which is similar to $J$ (since $J = B\Pi = \Pi^{-1/2} J_{\text{sym}} \Pi^{1/2}$). Let $J_{\text{sym}} = Q \Lambda Q^T$ be its eigendecomposition, with $Q$ orthogonal and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_L)$ containing the eigenvalues, ordered by magnitude $|\lambda_1| \geq |\lambda_2| \geq \ldots$.

Then $J^k = \Pi^{-1/2} Q \Lambda^k Q^T \Pi^{1/2}$. If there is a dominant eigenvalue $\lambda_1$ (i.e., $|\lambda_1| > |\lambda_2|$), then for large $k$, the matrix $\Lambda^k \approx \text{diag}(\lambda_1^k, 0, \ldots, 0)$. This implies $J^k \approx \lambda_1^k (\Pi^{-1/2} q_1)(q_1^T \Pi^{1/2})$, where $q_1$ is the leading eigenvector of $J_{\text{sym}}$. Let $u_1 = \Pi^{-1/2} q_1$ (a right eigenvector of $J$) and $v_1^T = q_1^T \Pi^{1/2}$ (a left eigenvector of $J$). Then $J^k \approx \lambda_1^k u_1 v_1^T$.

Substituting this into the expressions for $\mu_\ell$ and $\Sigma_\ell$:

- **Class Means:** $\mu_\ell \approx \lambda_1^k (u_1 v_1^T M)^T e_\ell = \lambda_1^k (M^T v_1)(u_1^T e_\ell)$. This shows that for large $k$, all mean vectors $\mu_\ell$ become approximately proportional to the fixed vector $M^T v_1 = M^T \Pi^{1/2} q_1$. The specific proportionality constant $(u_1^T e_\ell)$ depends on the class $\ell$, but the direction is shared.

- **Class Covariances:** Similarly, $J^{k-1} \approx \lambda_1^{k-1} u_1 v_1^T$. Then $\Sigma_\ell \approx \lambda_1^{2(k-1)} (M^T v_1)(\text{scalar}_\ell)(v_1^T M)$, where $\text{scalar}_\ell = u_1^T \, \text{diag}(e_\ell^T J) u_1$. This indicates that $\Sigma_\ell$ (and thus $\Sigma_\ell / \nu_n$) becomes approximately rank-one, with its dominant direction also aligned with $M^T v_1$.

This power iteration effect driven by $J^k$ and $J^{k-1}$ is the core of the oversmoothing mechanism:

1. **Mean Collapse:** The mean vectors $\mu_\ell$ for different classes tend to align along a common direction $M^T \Pi^{1/2} q_1$. While their magnitudes might differ (scaled by $\lambda_1^k (u_1^T e_\ell)$), their angular separation diminishes. If the initial feature means $M$ projected onto $v_1$ do not maintain sufficient separation, or if $u_1^T e_\ell$ values are too similar across classes, the means become indistinguishable.

2. **Covariance Collapse and Alignment:** The covariance matrices $\Sigma_\ell$ also become rank-deficient and align their principal direction with the same direction as the means.

The net effect is that the $L$ Gaussian components $N(\mu_\ell, \Sigma_\ell / \nu_n)$ of the feature distribution $\overline{\phi}_i^{(k)}$ effectively collapse onto a 1-dimensional subspace. Within this subspace, they become a mixture of 1-D Gaussians. If the projected means are not well-separated relative to the

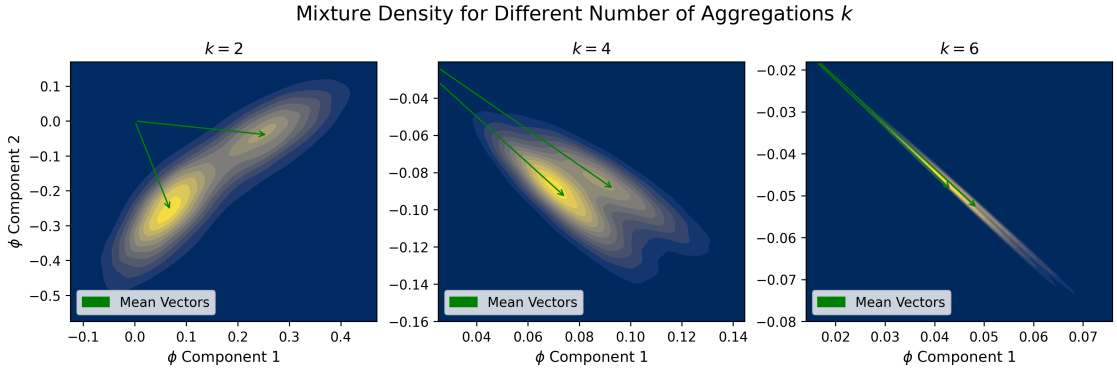



Figure 6.3: Estimated kernel density plots of the aggregated features $\overline{\phi}^{(k)}$ of a 2-class CSBM at different features depths $k$. A feature collapse in the mean vectors and the class covariances is visible by $k = 4$ and $k = 6$.

projected variances along this single dimension, classification becomes extremely difficult, regardless of the original dimensionality $d$ or the initial separability of $M$. This phenomenon is illustrated empirically in Figure 6.3, where increasing $k$ leads to feature distributions that are elongated along a common axis and overlap significantly. The parameter $\nu_n$ helps shrink the variances overall, but does not prevent this directional collapse induced by $k$.

This analysis provides a precise, quantifiable understanding of oversmoothing, directly linking it to the spectral properties of the community interaction matrix $J$ and the number of GNN layers $k$.

# 5 Conclusion

We conducted a rigorous asymptotic analysis of $k$-layer Polynomial GNN (Poly-GNN) embeddings on large, sparse, community-based graphs, establishing Central Limit Theorems that precisely characterize their limiting distributions. We showed that degree-normalized features $\overline{\phi}_i^{(k)}$, jointly with labels $z_i$, converge in $W_1$-distance to a Gaussian mixture $N(\mu_\ell, \Sigma_\ell/\nu_n)$ per class $\ell$. We provided exact forms for $\mu_\ell = (J^k M)^T e_\ell$ and $\Sigma_\ell = (J^{k-1}M)^T \operatorname{diag}(e_\ell^T J)(J^{k-1}M)$, determined by initial means $M$, layers $k$, and community interaction matrix $J$. Centered-

and-scaled features $\xi_i^{(k)}$ similarly converge to $\sum \pi_\ell N(0, \Sigma_\ell)$.

These findings have key implications. First, training linear classifiers on $\overline{\phi}_i^{(k)}$ with cross-entropy loss is asymptotically equivalent to optimizing on this limiting Gaussian mixture, with uniform convergence of the loss, gradient path, and optimal weights. This theoretically grounds the training behavior of GNN-based classifiers. Second, our explicit characterization of $\mu_\ell$ and $\Sigma_\ell$ offers a clear and nuanced understanding of the GNN oversmoothing phenomenon. The repeated multiplication by the matrix $J$ (to powers $k$ and $k-1$) acts as a power iteration, causing both the mean vectors and the principal directions of the covariance matrices to align with a low-dimensional (often 1-D) subspace dictated by the leading eigenvectors of $J$. As $k$ increases, this effect intensifies, leading to a collapse of the feature distributions of different classes onto this common subspace. This results in a degenerate, poorly separated Gaussian mixture, thereby diminishing the discriminative power of the GNN embeddings, irrespective of the initial feature dimensionality. Future directions include extensions to non-linear GNNs and different graph models.

# Appendix

## A Detailed Proofs of Main Theorems

This appendix provides the detailed proofs for Theorem 14 and Theorem 15 presented in Section 3.1. The general proof strategy follows the outline given in Section 3.2.

Before proceeding with the proofs, we establish some notation used throughout the appendices. For a probability measure $\mu$ on $\mathbb{R}^d$ and a vector $\theta \in \mathbb{R}^d$, we denote by $\mu_\theta$ the *$\theta$-projection* (or $\theta$-section) of $\mu$. This is the pushforward measure of $\mu$ under the map $x \mapsto \langle x, \theta \rangle = x^T\theta$. If $X \sim \mu$, then $\mu_\theta$ is the distribution of the real-valued random variable $\langle X, \theta \rangle$. For a measure $\nu$ on $\mathbb{R}$, its $r$-th moment is denoted by $m_r(\nu) = \int x^r d\nu(x)$. For a measure $\mu$ on $\mathbb{R}^d$, its $r$-th absolute moment is $M_r(\mu) = \int \|x\|_2^r d\mu(x)$. The empirical measure

of a set of $N$ points $\{Y_i\}_{i=1}^N$ is $\frac{1}{N}\sum_{i=1}^N \delta_{Y_i}$. We denote the expectation of a random empirical measure $\mathbb{P}_n$ as $\overline{\mathbb{P}}_n$, defined by its action on test functions $f$: $\overline{\mathbb{P}}_n[f] = \mathbb{E}[\mathbb{P}_n[f]]$. We use $\mathrm{Lip}(R)$ to denote the class of $R$-Lipschitz functions $f : \mathbb{R}^d \to \mathbb{R}$. Other notation, if not standard, will be defined as it appears.

The proofs presented herein for Theorem 14 and its supporting propositions (such as Proposition 13) ultimately rely on the weaker version of Assumption 4, i.e., requiring only $|\mathcal{C}_\ell|/n \to \pi_\ell$ rather than the faster $o(\nu_n^{-1/2})$ rate of convergence. While the faster rate was used in some intermediate rate calculations (e.g., for Lemma 34 to establish a rate for $\gamma_i \to \mu_\ell$), the final $W_1 \to 0$ convergence results for Theorem 14 hold under the weaker condition, as the dominant error terms vanish sufficiently fast or the vanishing variance of the target distribution $\widetilde{\mathbb{G}}_{n,\ell}$ accommodates slower convergence of the mean. Theorem 15 also holds under this weaker assumption as the involved moments converge appropriately.

### A.1 Proof of Theorem 15 (CLT for Centered and Scaled Features)

The proof of Theorem 15 largely follows the structure laid out in Section 3.2. First, we define the key components of the features. Recall the definition of the centered and scaled features from Eq. (6.3):

$$\xi_i^{(k)} = \sqrt{\nu_n}\Big(\frac{\phi_i^{(k)}}{\nu_n^k} - \mathbb{E}\Big[\frac{\phi_i^{(k)}}{\nu_n^k}\Big]\Big) = \frac{\phi_i^{(k)} - \mathbb{E}[\phi_i^{(k)}]}{\nu_n^{k-1/2}}.$$

The term $\phi_i^{(k)} - \mathbb{E}[\phi_i^{(k)}]$ represents the deviation of the $i$-th node's $k$-layer feature from its expectation. This deviation can be decomposed as follows. Let $\phi^{(k)}$ be the $n \times d$ matrix of all features.

$$\begin{aligned}
\phi^{(k)} - \mathbb{E}[\phi^{(k)}] &= A^k X - \mathbb{E}[A^k X] \\
&= A^k X - \mathbb{E}[A^k]\mathbb{E}[X] \quad \text{(since } A \text{ and } X \text{ are independent given } z\text{)} \\
&= (A^k - \mathbb{E}[A^k])X + \mathbb{E}[A^k](X - \mathbb{E}[X]) \\
&=: \mathring{\Delta} + \mathring{\Lambda}. \qquad (.22)
\end{aligned}$$

Here, $\mathring{\Delta} = (A^k - \mathbb{E}[A^k])X$ captures the randomness from the graph structure $A$, and $\mathring{\Lambda} = \mathbb{E}[A^k](X - \mathbb{E}[X])$ captures the randomness from the initial node features $X$ (around their means). Both $\mathring{\Delta}$ and $\mathring{\Lambda}$ are $n \times d$ matrices. Let $\mathring{\Delta}_i$ and $\mathring{\Lambda}_i$ denote their $i$-th rows (viewed as $d \times 1$ column vectors for consistency with $\xi_i^{(k)}$).

We define the normalized versions:

$$\Delta_i := \mathring{\Delta}_i / \nu_n^{k-1/2}, \quad \Lambda_i := \mathring{\Lambda}_i / \nu_n^{k-1/2}.$$

With this notation, the centered and scaled feature for node $i$ is:

$$\xi_i^{(k)} = \Delta_i + \Lambda_i.$$

For any projection vector $\theta \in S^{d-1}$, we denote $\Delta_{i,\theta} = \langle \Delta_i, \theta \rangle = \Delta_i^T \theta$ and $\Lambda_{i,\theta} = \langle \Lambda_i, \theta \rangle = \Lambda_i^T \theta$.

Now, we aim to show $\mathbb{E}[W_1(\mathbb{P}_n, \mathbb{G})] \to 0$ where $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i^{(k)}}$. This is achieved by showing:

1. $\mathbb{E}[W_1(\mathbb{P}_n, \overline{\mathbb{P}}_n)] \to 0$, where $\overline{\mathbb{P}}_n = \mathbb{E}[\mathbb{P}_n]$.
2. $W_1(\overline{\mathbb{P}}_n, \mathbb{G}) \to 0$.

**Part 1: Concentration of $\mathbb{P}_n$ around $\overline{\mathbb{P}}_n$.** This part relies on Proposition 17 (from Appendix D). To apply Proposition 17, we need to verify its conditions for $Y_{i,n} = \xi_i^{(k)}$:

(a) *Uniform $\Psi_{r_n}$ sub-Gaussianity of projections*: For any $\theta \in S^{d-1}$, $\{\langle \xi_i^{(k)}, \theta \rangle\}_{i=1}^n$ are uniformly $\Psi_{r_n}$ sub-Gaussian (see Appendix C for the definition):

**Lemma 31.** $\sup_{i \in [n]} \|\langle \xi_i^{(k)}, \theta \rangle\|_{\Psi_{r_n}} \lesssim C(\sigma, x_*)$ *for all* $\theta \in S^{d-1}$.

*Proof.* Combining Propositions 9 and 10 (from Appendix A.3), we have

$$\|\langle \xi_i^{(k)}, \theta \rangle\|_{\Psi_{r_n}} \lesssim \|\Delta_{i,\theta}\|_{\Psi_{r_n}} + \|\Lambda_{i,\theta}\|_{\Psi_{r_n}} \lesssim \kappa_0 + \sigma \beta_{k,n}$$

where $\kappa_0$ is a consntant only dependent on $\sigma$ and $x_*$, and $\beta_{k,n} = o(1)$. The result follows. □

(b) *Variance of empirical moments*: We have the following result:

**Lemma 32.** $\lim_{n\to\infty} Var\Big(n^{-1}\sum_i \langle \xi_i^{(k)}, \theta\rangle^r\Big) = 0$ *for all* $r \in \mathbb{N}$ *and* $\theta \in S^{d-1}$.

*Proof.* By Lemma 35, $\lim_{n\to\infty} \|\frac{1}{n}\sum_i \langle \xi_i^{(k)}, \theta\rangle^r - \frac{1}{n}\sum_{i=1}^n \Delta_{i,\theta}^r\|_{L^2} = 0$. Lemma 36 gives $\mathrm{Var}(n^{-1}\sum_i \Delta_{i,\theta}^r) \lesssim n^{-1}$. The result follows from the inequality $\mathrm{var}(A) \le 6\|A-B\|_{L^2}^2 + 3\,\mathrm{var}(B)$ for any random variables $A$ and $B$ in $L^2$. □

(c) *Uniformly bounded first moment of* $\overline{\mathbb{P}}_n$: $\sup_{n\ge 1} M_1(\overline{\mathbb{P}}_n) < \infty$. This follows since by Proposition 11 $m_1(\overline{\mathbb{P}}_{n,\theta})$ converge to $m_1(\mathbb{G}_\theta)$, which is finite, for all $\theta \in S^{d-1}$. The claims then follows from Proposition 14.

With these conditions met, Proposition 17 (from Appendix D) implies $\mathbb{E}[W_1(\mathbb{P}_n, \overline{\mathbb{P}}_n)] \to 0$.

**Part 2: Convergence of $\overline{\mathbb{P}}_n$ to $\mathbb{G}$.** Proposition 11 (from Appendix A.4) shows $m_r(\overline{\mathbb{P}}_{n,\theta}) \to m_r(\mathbb{G}_\theta)$ for all $\theta$ and $r$. Since $\mathbb{G}_\theta$ is a mixture of Gaussians, it is determined by its moments. This implies weak convergence $\overline{\mathbb{P}}_{n,\theta} \rightsquigarrow \mathbb{G}_\theta$. The convergence of moments also implies uniform integrability of all moments for $\{\overline{\mathbb{P}}_{n,\theta}\}_{n\ge 1}$. This, combined with weak convergence, yields $W_1(\overline{\mathbb{P}}_{n,\theta}, \mathbb{G}_\theta) \to 0$ for all $\theta \in S^{d-1}$ (e.g., by [AGS08, Proposition 7.1.5]). To lift this to $W_1(\overline{\mathbb{P}}_n, \mathbb{G}) \to 0$, we use Proposition 16 (from Appendix D). Condition (.30) for this proposition, $\sup_{n\ge 1}(M_1(\overline{\mathbb{P}}_n) + M_1(\mathbb{G})) < \infty$, is satisfied because $M_1(\overline{\mathbb{P}}_n)$ is uniformly bounded (as argued in Part 1c) and $M_1(\mathbb{G})$ is finite.

The class-conditional convergence statement in Theorem 15 follows from a similar argument by considering per-class empirical measures $\mathbb{P}_{n,\ell}$ and their expectations $\overline{\mathbb{P}}_{n,\ell}$, and showing their convergence to $N(0, \Sigma_\ell)$. See the proof of Proposition 12 (a restatement of the class-conditional convergence) for details.

### A.2 Proof of Theorem 14 (CLT for Degree-Normalized Features and Labels)

The proof of Theorem 14 closely mirrors that of Theorem 15, with adjustments for the non-zero means and the $\nu_n^{-1}$ scaling in the covariance. Let $\widetilde{\mathbb{P}}_n^{joint}$ be the empirical measure of

$(z_i, \overline{\phi}_i^{(k)})$ and $\mathbb{G}_n^{joint}$ be its target limit. The convergence in $W_1$ can be established by showing convergence of expectations of Lipschitz functions $f(z, x)$. The core argument involves showing that for $i \in \mathcal{C}_\ell$, $\overline{\phi}_i^{(k)}$ behaves like a draw from $N(\mu_\ell, \Sigma_\ell/\nu_n)$.

1. **Mean Convergence:** Lemma 33 establishes that $\mathbb{E}[\overline{\phi}_i^{(k)}]$ converges to a general limit $\gamma_i$.

   **Lemma 33.** *Define limiting mean $\gamma_i = e_i^T(\mathbb{E}[A/\nu_n])^k\mathbb{E}[X]$. Assume Assumption 5 and suppose $\nu_n \geq 1$. Then,*

   $$\max_{i\in[n]} \|\mathbb{E}[\overline{\phi}_i^{(k)}] - \gamma_i^T\|_2 \leq C(k)\, x_*\, \nu_n^{-1}.$$

   Under the specific community-based graph model, this general limit $\gamma_i$ further simplifies for nodes within a class $\mathcal{C}_\ell$ to the class-specific mean $\mu_\ell$, as stated in the following lemma.

   **Lemma 34.** *Under the conditions of Theorem 14 (which include the CSBM structure and Assumptions 2–5, with the strengthened Assumption 4), let $\gamma_i^T = e_i^T(\mathbb{E}[A/\nu_n])^k\mathbb{E}[X]$ be the $1 \times d$ row vector defined in Lemma 33. For any node $i \in \mathcal{C}_\ell$, its limiting mean $\gamma_i^T$ converges to $\mu_\ell^T = e_\ell^T J^k M$. More precisely,*

   $$\max_{\ell\in[L]} \sup_{i\in\mathcal{C}_\ell} \|\gamma_i^T - \mu_\ell^T\|_2 = o(\nu_n^{-1/2}).$$

   *Proof.* The expected adjacency matrix of an undirected, loop-less SBM is $\mathbb{E}[A] = (\nu_n/n)(P - \mathrm{diag}(P))$, where $P = ZBZ^T$ and $\mathrm{diag}(P)$ contains the diagonal entries of $P$. Thus, $\mathbb{E}[A/\nu_n] = (P/n) - (\mathrm{diag}(P)/n)$. The difference between $(\mathbb{E}[A/\nu_n])^k$ and $(P/n)^k$ can be bounded. Since matrix exponentiation $H \mapsto H^k$ is locally Lipschitz for matrices with bounded operator norm (which $(P/n)$ and $\mathbb{E}[A/\nu_n]$ are, as their entries are $O(1/n)$ and norms are $O(1)$), and

   $$\|\mathbb{E}[A/\nu_n] - P/n\|_{\mathrm{op}} = \|\mathrm{diag}(P)/n\|_{\mathrm{op}} = \max_{\ell'\in[L]} n^{-1}B_{\ell'\ell'} = O(n^{-1}), \tag{.23}$$

it follows that $\|(\mathbb{E}[A/\nu_n])^k - (P/n)^k\|_{\text{op}} = O(n^{-1})$. Given $\mathbb{E}[X] = ZM$ has bounded row norms (from Assumption 5), we can write:

$$\begin{aligned}\gamma_i^T &= e_i^T (P/n)^k \mathbb{E}[X] + e_i^T ((\mathbb{E}[A/\nu_n])^k - (P/n)^k)\mathbb{E}[X] \\ &= e_i^T (P/n)^k ZM + O(n^{-1}) \cdot \|e_i^T\|_{\text{op}} \|\mathbb{E}[X]\|_{\text{op}}.\end{aligned}$$

Since $\|e_i^T\|_{\text{op}} = 1$ and $\|\mathbb{E}[X]\|_{\text{op}}$ is bounded (e.g., by $\sqrt{L}\max_\ell \|M_{\ell,\cdot}\|_2 \leq \sqrt{L}x_*$), the error term is $O(n^{-1})$. So,

$$\gamma_i^T = e_i^T (P/n)^k ZM + O(n^{-1}).$$

Now consider the main term $e_i^T (P/n)^k ZM$. For a node $i \in \mathcal{C}_\ell$, we have $e_i^T Z = e_\ell^T$ (where $e_i \in \mathbb{R}^n, e_\ell \in \mathbb{R}^L$).

$$\begin{aligned}e_i^T (P/n)^k ZM &= e_i^T (ZBZ^T/n)^k ZM \\ &= e_i^T Z (B(Z^T Z/n))^k M \\ &= e_\ell^T (B\widetilde{\Pi})^k M \quad (\text{since } Z^T Z/n = \text{diag}(|\mathcal{C}_s|/n)_{s=1}^L = \widetilde{\Pi}) \\ &= e_\ell^T \widetilde{J}^k M,\end{aligned}$$

where $\widetilde{J} = B\widetilde{\Pi}$ and $\widetilde{\Pi} = \text{diag}(\widetilde{\pi}_1, \ldots, \widetilde{\pi}_L)$ with $\widetilde{\pi}_s = |\mathcal{C}_s|/n$. We are given $\mu_\ell^T = e_\ell^T J^k M$, where $J = B\Pi$. The difference is $e_\ell^T (\widetilde{J}^k - J^k)M$.

From the Assumption 4, $\widetilde{\pi}_s = \pi_s + o(\nu_n^{-1/2})$, which implies $\widetilde{\Pi} = \Pi + E_n$ where $E_n$ is a diagonal matrix with entries $o(\nu_n^{-1/2})$. Thus, $\|\widetilde{\Pi} - \Pi\|_{\text{op}} = o(\nu_n^{-1/2})$. Then, $\widetilde{J} - J = B(\widetilde{\Pi} - \Pi) = BE_n$. So, $\|\widetilde{J} - J\|_{\text{op}} \leq \|B\|_{\text{op}} \|E_n\|_{\text{op}} = O(1) \cdot o(\nu_n^{-1/2}) = o(\nu_n^{-1/2})$. Using the identity $A^k - B^k = \sum_{j=0}^{k-1} A^j (A - B) B^{k-1-j}$, and since $\|J\|_{\text{op}}$ and $\|\widetilde{J}\|_{\text{op}}$ are $O(1)$ (as $\|B\|_{\text{op}}$ and $\|\Pi\|_{\text{op}}$ are $O(1)$),

$$\|\widetilde{J}^k - J^k\|_{\text{op}} \leq k \cdot \max(\|J\|_{\text{op}}, \|\widetilde{J}\|_{\text{op}})^{k-1} \cdot \|\widetilde{J} - J\|_{\text{op}} = O(1) \cdot o(\nu_n^{-1/2}) = o(\nu_n^{-1/2}).$$

Therefore,

$$\|e_\ell^T (\widetilde{J}^k - J^k)M\|_2 \leq \|e_\ell^T\|_{\text{op}} \|\widetilde{J}^k - J^k\|_{\text{op}} \|M\|_{\text{op}} = 1 \cdot o(\nu_n^{-1/2}) \cdot O(1) = o(\nu_n^{-1/2}).$$

Combining the two error terms:

$$\gamma_i^T = e_\ell^T J^k M + o(\nu_n^{-1/2}) + O(n^{-1}).$$

Since $\nu_n = o(n)$ (Assumption 3), $n^{-1} = o(\nu_n^{-1})$ which is also $o(\nu_n^{-1/2})$. Thus, the dominant error term is $o(\nu_n^{-1/2})$. The bounds are uniform over $i \in \mathcal{C}_\ell$ and $\ell \in [L]$ because the operator norm bounds on $B, \Pi, M$ and the rate of convergence in Assumption 4 are uniform. □

2. **Covariance Characterization:** The deviation $\overline{\phi}_i^{(k)} - \mathbb{E}[\overline{\phi}_i^{(k)}] = \xi_i^{(k)}/\sqrt{\nu_n}$. The analysis for $\xi_i^{(k)}$ (specifically, the characterization of its moments leading to Proposition 11 in Appendix A.4) shows its asymptotic covariance, conditional on $z_i = \ell$, is $\Sigma_\ell$. Thus, the covariance of $\overline{\phi}_i^{(k)} - \mathbb{E}[\overline{\phi}_i^{(k)}]$ (and asymptotically, of $\overline{\phi}_i^{(k)} - \mu_\ell$ for $i \in \mathcal{C}_\ell$) is $\Sigma_\ell/\nu_n$.

3. **Moment Matching and Concentration:** Similar to Theorem 15, one shows that the moments of $(\overline{\phi}_i^{(k)} - \mu_\ell)$ (for $i \in \mathcal{C}_\ell$), when appropriately scaled, match those of $N(0, \Sigma_\ell/\nu_n)$. Concentration arguments analogous to Part 1 of Theorem 15's proof apply.

Steps 2 and 3 above are rigorously formalized during the proof of the class-conditional version of the statement (Eq. (6.11) which is stated and proved as Proposition 13 in Appendix A.4).

### A.3 Supporting Lemmas for Moment Analysis

We borrow the following two key results from [VA24b]:

**Proposition 9.** *Suppose $X_i - \mathbb{E}[X_i] \sim SG(\sigma^2)$ and $\nu_n \geq 1$. Then, for $\Lambda_{i,\theta} = \langle \mathring{\Lambda}_i/\nu_n^{k-1/2}, \theta \rangle$:*

$$\|\Lambda_{i,\theta}\|_{\Psi_{r_n}} \lesssim \sigma\beta_{k,n} \quad \textit{where} \quad \beta_{k,n} = (\nu_n^{-k+1})^{1/2} \cdot \mathbb{1}\{k \textit{ is even}\} + (\nu_n/n)^{1/2}.$$

*Proof.* A component-wise version of the above (i.e. with $\theta$ a coordinate basis vector) is proven in [VA24b, Section 4.1]. The general case follows by a similar argument. Broadly, this follows

from the fact that $\mathring{\Lambda}_i = \mathbb{E}[A^k](X - \mathbb{E}[X])/\nu_n^{k-1/2}$ is a linear transformation of sub-Gaussian random vectors $\{X_{i*} - \mathbb{E}[X_{i*}]\}_i$. The rate $\beta_{k,n}$ is obtained through path counting on $\mathbb{E}[A^k]_{ij}$ for each $j \in [n]$. □

**Proposition 10.** *Let $t_* = rk - \lceil r/2 \rceil$ and $\kappa_0 = 4\max\{C_1\sigma, x_*\}$ for a distribution-dependent constant $C_1$. Assume Assumption* (5) *and suppose $\nu_n \geq 1$. For $\epsilon \in [0,1]$, let*

$$r_n(\epsilon) := \max\{r \in 2\mathbb{N} : 3(\kappa_0 rke^k)^r \leq \nu_n^{1-\epsilon}\}. \tag{.24}$$

*Then, for $\mathring{\Delta}_{i,\theta} = \langle (A^k - \mathbb{E}[A^k])X, \theta\rangle$ and for all $r \leq r_n(0)$:*

$$\mathbb{E}|\mathring{\Delta}_{i,\theta}|^r \leq 2(\sqrt{r}\kappa_0)^r \nu_n^{t_*}. \tag{.25}$$

*As a consequence, for $\Delta_{i,\theta} = \mathring{\Delta}_{i,\theta}/\nu_n^{k-1/2}$, we have $\|\Delta_{i,\theta}\|_{\Psi_{r_n}} \lesssim \kappa_0$.*

*Proof.* This result is proven in [VA24b]. The power $\nu_n^{t_*}$ arises from counting dominant walk structures contributing to the $r$-th moment. □

**Lemma 35.** *Assume Assumptions 2, 3, and 5. For any $r \in \mathbb{N}$:*

$$\lim_{n\to\infty} \max_{i\in[n]} \left\| \langle \xi_i^{(k)}, \theta\rangle^r - \Delta_{i,\theta}^r \right\|_{L^2} = 0.$$

*Proof.* Using the decomposition $\langle \xi_i^{(k)}, \theta\rangle = \Delta_{i,\theta} + \Lambda_{i,\theta}$, we have

$$\langle \xi_i^{(k)}, \theta\rangle^r - \Delta_{i,\theta}^r = \sum_{s=1}^{r} \binom{r}{s} \Delta_{i,\theta}^{r-s} \Lambda_{i,\theta}^s.$$

By Minkowski inequality (for $L_2$ norm of sums):

$$\left\| \sum_{s=1}^{r} \binom{r}{s} \Delta_{i,\theta}^{r-s} \Lambda_{i,\theta}^s \right\|_{L^2} \leq \sum_{s=1}^{r} \binom{r}{s} \left\| \Delta_{i,\theta}^{r-s} \Lambda_{i,\theta}^s \right\|_{L^2}$$

By Hölder inequality, with $1/p = (r-s)/r$ and $1/q = s/r$,

$$\left\| \Delta_{i,\theta}^{r-s} \Lambda_{i,\theta}^s \right\|_{L^2}^2 = \mathbb{E}[\Delta_{i,\theta}^{2(r-s)} \Lambda_{i,\theta}^{2s}] \leq (\mathbb{E}\Delta_{i,\theta}^{2r})^{1-s/r} (\mathbb{E}\Lambda_{i,\theta}^{2r})^{s/r}.$$

This is $\|\Delta_{i,\theta}\|_{L^{2r}}^{2(r-s)} \cdot \|\Lambda_{i,\theta}\|_{L^{2r}}^{2s}$. For $n$ large enough so $2r \le r_n$, Proposition 10 (via Lemma 39) gives $\|\Delta_{i,\theta}\|_{L^{2r}} \lesssim \kappa_0\sqrt{2r}$. Proposition 9 (via Lemma 39) gives $\|\Lambda_{i,\theta}\|_{L^{2r}} \lesssim \sigma\beta_{k,n}\sqrt{2r}$. Take $n$ large enough so that $\beta_{k,n} \le 1$. Then, for $s \ge 1$, we have $\beta_{k,n}^{2s} \le \beta_{k,n}$, hence $\|\Lambda_{i,\theta}\|_{L^{2r}}^{2s} \lesssim \sigma^{2s}\beta_{k,n}(2r)^s$. Since $\beta_{k,n} \to 0$ from Proposition 9 (as $\nu_n \to \infty, \nu_n = o(n)$), and all other terms are bounded, the sum tends to 0. The convergence is uniform over $i$ as the bounds are uniform. □

**Lemma 36.** *Under Assumption 2 and 5. For every $i, i' \in [n]$, $r \in \mathbb{N}$ and $\theta \in S^{d-1}$,*

$$\mathrm{Cov}\big(\Delta_{i,\theta}^r, \Delta_{i',\theta}^r\big) \lesssim n^{-1\{i\neq i'\}}.$$

*In particular, $Var\big(n^{-1}\sum_{i=1}^n \Delta_{i,\theta}^r\big) \lesssim n^{-1}$ for $r \in \mathbb{N}$ and $\theta \in S^{d-1}$.*

The proof of Lemma 36 is quite involved, using combinatorics of walk sequences, and appears in Appendix F.

**Lemma 37.** *Let $r \in 2\mathbb{N}$ and $0 < \epsilon < 1$. Assume Assumption (5) and suppose $\nu_n \ge 1$. If $r \le r_n(\epsilon)$, then*

$$\max_{i\in[n]} \left|\mathbb{E}[\Delta_{i,\theta}^r] - (r-1)!! \cdot \widetilde{\sigma}_{i,\theta}^r\right| \;\le\; C(r)\, x_*^r \max\{n^{-1}, \nu_n^{-\epsilon}\}$$

*where $\widetilde{\sigma}_{i,\theta}^2 := \|V_i\mathbb{E}[A/\nu_n]^{k-1}\mathbb{E}[X]\theta\|_2^2$ and $V_i = \big[\mathrm{diag}\big((p_{i1}(1-p_{i1}), \ldots, p_{in}(1-p_{in})\big)/\nu_n\big]^{1/2}$.*

The proof of Lemma 37 appears in Appendix F and involves walk-based proxy term $\widetilde{T}_i^{\mathrm{hi}}(r)$ and careful counting of dominant vs non-dominant walk structures.

**Lemma 38** (Odd Moment Control for $\Delta_{i,\theta}$)**.** *Under Assumptions 2–5, for any odd integer $r \ge 1$ and any unit vector $\theta \in \mathbb{R}^d$,*

$$\lim_{n\to\infty} \max_{i\in[n]} |\mathbb{E}[\Delta_{i,\theta}^r]| = 0.$$

*More specifically, $\mathbb{E}[\Delta_{i,\theta}^r] = O(\nu_n^{-1/2})$.*

*Proof.* This follows from Proposition 10. For an odd $r$, the moment bound for $\Delta_{i,\theta}$ is $\mathbb{E}|\Delta_{i,\theta}|^r \lesssim (\sqrt{r}\kappa_0)^r \nu_n^{r/2-\lceil r/2\rceil} = (\sqrt{r}\kappa_0)^r \nu_n^{-1/2}$. □

### A.4 Supporting Results for Specialization to Community-Based Graphs

**Proposition 11.** *Under Assumptions 2–5,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\langle \xi_i^{(k)}, \theta\rangle^r = (r-1)!! \sum_{\ell=1}^{L} \pi_\ell \left((J^{k-1}M\theta)^T diag(e_\ell^T J)(J^{k-1}M\theta)\right)^{r/2} \cdot \mathbb{1}\{r \text{ is even}\}.$$

*Stated differently,* $m_r(\overline{\mathbb{P}}_{n,\theta}) \to m_r(\mathbb{G}_\theta)$ *where* $\mathbb{G}_\theta = \sum_{\ell=1}^{L} \pi_\ell N(0, \theta^T \Sigma_\ell \theta)$.

*Proof.* We proceed in steps:

*Step 1: Approximate with moments of* $\Delta_{i,\theta}$. By Lemma 35 (specifically, $\|\langle \xi_i^{(k)}, \theta\rangle^r - \Delta_{i,\theta}^r\|_{L^1} \to 0$ since $L_2$ convergence implies $L_1$), we have $\mathbb{E}\langle \xi_i^{(k)}, \theta\rangle^r = \mathbb{E}[\Delta_{i,\theta}^r] + o(1)$, where the $o(1)$ term is uniform over $i$. Thus,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\langle \xi_i^{(k)}, \theta\rangle^r = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\Delta_{i,\theta}^r] + o(1).$$

*Step 2: Handle odd moments.* If $r$ is an odd integer, by Lemma 38, $\mathbb{E}[\Delta_{i,\theta}^r] = o(1)$ uniformly in $i$. Therefore,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\Delta_{i,\theta}^r] = 0.$$

This matches the proposition statement, as $\mathbb{1}\{r \text{ is even}\} = 0$ for odd $r$.

*Step 3: Handle even moments using* $\widetilde{\sigma}_{i,\theta}^r$. If $r$ is an even integer, by Lemma 37,

$$\mathbb{E}[\Delta_{i,\theta}^r] = (r-1)!! \cdot \widetilde{\sigma}_{i,\theta}^r + o(1),$$

uniformly in $i$. Here, $\widetilde{\sigma}_{i,\theta}^2 = \|V_i \mathbb{E}[A/\nu_n]^{k-1}\mathbb{E}[X]\theta\|_2^2$. So we need to analyze the limit of $\frac{1}{n}\sum_{i=1}^{n}(r-1)!! \cdot \widetilde{\sigma}_{i,\theta}^r$.

*Step 4: Analyze* $\widetilde{\sigma}_{i,\theta}^2$ *under the CSBM structure.* We have

$$\widetilde{\sigma}_{i,\theta}^2 = (\mathbb{E}[A/\nu_n]^{k-1}\mathbb{E}[X]\theta)^T V_i^2 (\mathbb{E}[A/\nu_n]^{k-1}\mathbb{E}[X]\theta).$$

where $V_i^2 = \nu_n^{-1}\mathrm{diag}((p_{ij}(1-p_{ij}))_{j=1}^n) = \nu_n^{-1}\mathrm{diag}(e_i^T\mathbb{E}[A])(I_n - \mathrm{diag}(e_i^T\mathbb{E}[A]))$. Since by assumption $\nu_n = o(n)$, we have $e_i^T\mathbb{E}[A] = O(\nu_n/n) = o(1)$ uniformly in $i$. It follows that

$$V_i^2 = \nu_n^{-1}\mathrm{diag}(e_i^T\mathbb{E}[A]) + o(1).$$

Moreover, as shown in the proof of Lemma 34 (specifically Eq. (.23)), $\mathbb{E}[A/\nu_n] = P/n + O(n^{-1})$, where $P = ZBZ^T$. Substituting we get

$$\widetilde{\sigma}^2_{i,\theta} = ((P/n)^{k-1}\mathbb{E}[X]\theta)^T \mathrm{diag}(e_i^T P/n)((P/n)^{k-1}\mathbb{E}[X]\theta) + o(1).$$

Under the CSBM structure, we have $\mathbb{E}[X]\theta = ZM\theta$. Similar to the derivation in Lemma 34's proof:

$$(P/n)^{k-1} ZM\theta = Z(B\widetilde{\Pi})^{k-1} M\theta = Z\widetilde{J}_n^{k-1} M\theta.$$

If node $i \in \mathcal{C}_\ell$, then $e_i^T P/n = e_\ell^T (BZ^T/n)$. The term $Z^T \mathrm{diag}(e_i^T P/n) Z$ becomes a diagonal $L \times L$ matrix. For $i \in \mathcal{C}_\ell$:

$$\begin{aligned}
(Z^T \mathrm{diag}(e_i^T P/n) Z)_{s,s'} &= \sum_{j=1}^n Z_{js} (e_i^T P/n)_j Z_{js'} \\
&= \sum_{j \in \mathcal{C}_s, s=s'} (P_{ij}/n) = \sum_{j \in \mathcal{C}_s, s=s'} (B_{z_i z_j}/n) \\
&= \mathbb{1}\{s = s'\} \cdot (B_{\ell s} |\mathcal{C}_s|/n) = \mathbb{1}\{s = s'\} \cdot B_{\ell s} \widetilde{\pi}_s.
\end{aligned}$$

So, $Z^T \mathrm{diag}(e_i^T P/n) Z = \mathrm{diag}((B_{\ell s}\widetilde{\pi}_s)_{s=1}^L) = \mathrm{diag}(e_\ell^T \widetilde{J}_n)$. Therefore, for $i \in \mathcal{C}_\ell$:

$$\widetilde{\sigma}^2_{i,\theta} = (\widetilde{J}_n^{k-1} M\theta)^T \mathrm{diag}(e_\ell^T \widetilde{J}_n)(\widetilde{J}_n^{k-1} M\theta) + o(1).$$

Let $\sigma^2_{\ell,\theta}(\widetilde{J}_n) = (\widetilde{J}_n^{k-1} M\theta)^T \mathrm{diag}(e_\ell^T \widetilde{J}_n)(\widetilde{J}_n^{k-1} M\theta)$. This term is the same for all $i \in \mathcal{C}_\ell$ up to $o(1)$ errors.

*Step 5: Averaging over $i$ and taking limits.* For even $r$:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Delta^r_{i,\theta}] &= \frac{(r-1)!!}{n} \sum_{i=1}^n \widetilde{\sigma}^r_{i,\theta} + o(1) \\
&= (r-1)!! \sum_{\ell=1}^L \frac{|\mathcal{C}_\ell|}{n} \left( \frac{1}{|\mathcal{C}_\ell|} \sum_{i \in \mathcal{C}_\ell} \widetilde{\sigma}^r_{i,\theta} \right) + o(1) \\
&= (r-1)!! \sum_{\ell=1}^L \widetilde{\pi}_\ell \cdot (\sigma^2_{\ell,\theta}(\widetilde{J}_n))^{r/2} + o(1).
\end{aligned}$$

As $n \to \infty$, by Assumption 4, $\widetilde{\pi}_\ell \to \pi_\ell$. Also, $\|\widetilde{J}_n - J\|_{\mathrm{op}} \to 0$ (due to $\widetilde{\Pi} \to \Pi$). Since $\sigma^2_{\ell,\theta}(\cdot)$ is a continuous function of its matrix argument (in terms of matrix entries or operator norm for

fixed $M, \theta, B, e_\ell, k$), we have $\sigma^2_{\ell,\theta}(\widetilde{J}_n) \to \sigma^2_{\ell,\theta}(J)$. Let $\sigma^{*2}_{\ell,\theta} = (J^{k-1}M\theta)^T \mathrm{diag}(e_\ell^T J)(J^{k-1}M\theta)$. The limit becomes:

$$(r-1)!! \sum_{\ell=1}^{L} \pi_\ell (\sigma^{*2}_{\ell,\theta})^{r/2}.$$

This is precisely the $r$-th moment of $\mathbb{G}_\theta = \sum_{\ell=1}^{L} \pi_\ell N(0, \sigma^{*2}_{\ell,\theta})$. Note that $\sigma^{*2}_{\ell,\theta} = \theta^T \Sigma_\ell \theta$ where $\Sigma_\ell = (J^{k-1}M)^T \mathrm{diag}(e_\ell^T J)(J^{k-1}M)$. The proof is complete. □

**Proposition 12 (Part of Theorem 15).** *Consider the setting of Proposition 11. Let $\mathbb{G}_\ell = N(0, \Sigma_\ell)$ for $\ell \in [L]$. Then for any $R > 0$:*

$$\mathbb{E}\left\{ \sup_{f_1,\ldots,f_L \in \mathrm{Lip}(R)} \left| \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in \mathcal{C}_\ell} f_\ell(\xi_i^{(k)}) - \sum_{\ell=1}^{L} \pi_\ell \mathbb{E}_{Y \sim \mathbb{G}_\ell}[f_\ell(Y)] \right| \right\} \to 0.$$

*Proof.* Let $\mathbb{P}_{n,\ell} = \frac{1}{|\mathcal{C}_\ell|} \sum_{i \in \mathcal{C}_\ell} \delta_{\xi_i^{(k)}}$ be the class-conditional empirical measure for class $\ell$. Let $f_\ell \in \mathrm{Lip}(R)$. We can assume $f_\ell(0) = 0$ without loss of generality by considering $f_\ell(x) - f_\ell(0)$, as this does not change the difference of expectations for centered measures and preserves the Lipschitz constant. The term we want to show goes to zero is:

$$\Delta_n := \sup_{f_1,\ldots,f_L \in \mathrm{Lip}(R)} \left| \sum_{\ell=1}^{L} \frac{|\mathcal{C}_\ell|}{n} \mathbb{P}_{n,\ell}[f_\ell] - \sum_{\ell=1}^{L} \pi_\ell \mathbb{G}_\ell[f_\ell] \right|.$$

Using the triangle inequality:

$$\begin{aligned}
\Delta_n &\leq \sup_{f_1,\ldots,f_L \in \mathrm{Lip}(R)} \sum_{\ell=1}^{L} \left| \frac{|\mathcal{C}_\ell|}{n} \mathbb{P}_{n,\ell}[f_\ell] - \pi_\ell \mathbb{P}_{n,\ell}[f_\ell] \right| \\
&\quad + \sup_{f_1,\ldots,f_L \in \mathrm{Lip}(R)} \sum_{\ell=1}^{L} |\pi_\ell \mathbb{P}_{n,\ell}[f_\ell] - \pi_\ell \mathbb{G}_\ell[f_\ell]| \\
&\leq \sum_{\ell=1}^{L} \left| \frac{|\mathcal{C}_\ell|}{n} - \pi_\ell \right| \sup_{f_\ell \in \mathrm{Lip}(R)} |\mathbb{P}_{n,\ell}[f_\ell]| \\
&\quad + \sum_{\ell=1}^{L} \pi_\ell \sup_{f_\ell \in \mathrm{Lip}(R)} |\mathbb{P}_{n,\ell}[f_\ell] - \mathbb{G}_\ell[f_\ell]|.
\end{aligned}$$

The second term is $\sum_{\ell=1}^{L} \pi_\ell R \cdot W_1(\mathbb{P}_{n,\ell}, \mathbb{G}_\ell)$ by definition of $W_1$ (scaled by $R$). Let $T_{1,n}$ and $T_{2,n}$ be the two terms.

For $T_{1,n}$: Since $f_\ell(0) = 0$ and $f_\ell \in \text{Lip}(R)$, $|\mathbb{P}_{n,\ell}[f_\ell]| \leq \mathbb{P}_{n,\ell}[|f_\ell(x)|] \leq R \cdot \mathbb{P}_{n,\ell}[\|x\|]$. So, $\mathbb{E}[\sup_{f_\ell \in \text{Lip}(R)} |\mathbb{P}_{n,\ell}[f_\ell]|] \leq R \cdot \mathbb{E}[\mathbb{P}_{n,\ell}[\|x\|]] = R \cdot \overline{\mathbb{P}}_{n,\ell}[\|x\|]$. The term $\overline{\mathbb{P}}_{n,\ell}[\|x\|] = \frac{1}{|\mathcal{C}_\ell|} \sum_{i \in \mathcal{C}_\ell} \mathbb{E}[\|\xi_i^{(k)}\|]$. From the proof of Theorem 15 (specifically Part 1c, relying on uniform integrability of moments of $\overline{\mathbb{P}}_n$), $\sup_n \mathbb{E}[\|\xi_i^{(k)}\|]$ is bounded for all $i$. Thus, $\sup_n \overline{\mathbb{P}}_{n,\ell}[\|x\|]$ is bounded (as $|\mathcal{C}_\ell| \to \infty$). By Assumption 4, $\left|\frac{|\mathcal{C}_\ell|}{n} - \pi_\ell\right| \to 0$. Therefore, $\mathbb{E}[T_{1,n}] \to 0$.

For $T_{2,n}$: We need to show $\mathbb{E}[W_1(\mathbb{P}_{n,\ell}, \mathbb{G}_\ell)] \to 0$ for each $\ell$. By the triangle inequality, $W_1(\mathbb{P}_{n,\ell}, \mathbb{G}_\ell) \leq W_1(\mathbb{P}_{n,\ell}, \overline{\mathbb{P}}_{n,\ell}) + W_1(\overline{\mathbb{P}}_{n,\ell}, \mathbb{G}_\ell)$. For the two terms on we have:

(a) $\mathbb{E}[W_1(\mathbb{P}_{n,\ell}, \overline{\mathbb{P}}_{n,\ell})] \to 0$: $\mathbb{P}_{n,\ell}$ is an empirical measure of $N_\ell = |\mathcal{C}_\ell|$ variables $\{\xi_i^{(k)} : i \in \mathcal{C}_\ell\}$. Since $N_\ell \to \infty$ (as $\pi_\ell > 0$), we can apply Proposition 17 to this specific subset of variables. The conditions for Proposition 17 are: (i) Uniform $\Psi_{r_{N_\ell}}$ sub-Gaussianity of $\langle \xi_i^{(k)}, \theta \rangle$ for $i \in \mathcal{C}_\ell$: This holds from Lemma 31. (ii) Variance of their empirical moments $\text{Var}(N_\ell^{-1} \sum_{i \in \mathcal{C}_\ell} \langle \xi_i^{(k)}, \theta \rangle^r) \to 0$ holds from the more general formulation of Lemma 36 where $\text{Cov}(\Delta_{i,\theta}^r, \Delta_{\theta,i'}^r) \lesssim n^{-1\{i \neq i'\}}$. (iii) $\sup_n M_1(\overline{\mathbb{P}}_{n,\ell}) < \infty$: This holds as shown for $T_{1,n}$. Thus, $\mathbb{E}[W_1(\mathbb{P}_{n,\ell}, \overline{\mathbb{P}}_{n,\ell})] \to 0$.

(b) $W_1(\overline{\mathbb{P}}_{n,\ell}, \mathbb{G}_\ell) \to 0$: We analyze the moments of $\overline{\mathbb{P}}_{n,\ell}$ for a given $\theta \in S^{d-1}$. $m_r(\overline{\mathbb{P}}_{n,\ell,\theta}) = \frac{1}{|\mathcal{C}_\ell|} \sum_{i \in \mathcal{C}_\ell} \mathbb{E}[\langle \xi_i^{(k)}, \theta \rangle^r]$. From Steps 1, 2, 3 of the proof of Proposition 11, we know that $\mathbb{E}[\langle \xi_i^{(k)}, \theta \rangle^r] = \mathbb{E}[\Delta_{i,\theta}^r] + o(1)$. If $r$ is odd, $\mathbb{E}[\Delta_{i,\theta}^r] = o(1)$ by Lemma 38. So $m_r(\overline{\mathbb{P}}_{n,\ell,\theta}) \to 0 = m_r(N(0, \theta^T \Sigma_\ell \theta))$. If $r$ is even, $\mathbb{E}[\Delta_{i,\theta}^r] = (r-1)!! \cdot \widetilde{\sigma}_{i,\theta}^r + o(1)$, where the $o(1)$ is uniform in $i$. From Step 4 in the proof of Proposition 11, for any $i \in \mathcal{C}_\ell$, $\widetilde{\sigma}_{i,\theta}^2 \to \sigma_{\ell,\theta}^{*2} := \theta^T \Sigma_\ell \theta$. Thus, for $i \in \mathcal{C}_\ell$, $\mathbb{E}[\langle \xi_i^{(k)}, \theta \rangle^r] \to (r-1)!!(\sigma_{\ell,\theta}^{*2})^{r/2} \cdot \mathbb{1}\{r \text{ is even}\}$. This limit is uniform for all $i \in \mathcal{C}_\ell$. Therefore,

$$m_r(\overline{\mathbb{P}}_{n,\ell,\theta}) = \frac{1}{|\mathcal{C}_\ell|} \sum_{i \in \mathcal{C}_\ell} \left( (r-1)!!(\sigma_{\ell,\theta}^{*2})^{r/2} \cdot \mathbb{1}\{r \text{ is even}\} + o(1) \right)$$
$$\to (r-1)!!(\sigma_{\ell,\theta}^{*2})^{r/2} \cdot \mathbb{1}\{r \text{ is even}\}.$$

This is $m_r(N(0, \theta^T \Sigma_\ell \theta))$. Since $\mathbb{G}_{\ell,\theta} = N(0, \theta^T \Sigma_\ell \theta)$ is determined by its moments, and its moments are finite, $\overline{\mathbb{P}}_{n,\ell,\theta} \rightsquigarrow \mathbb{G}_{\ell,\theta}$. The uniform integrability of moments for $\overline{\mathbb{P}}_{n,\ell,\theta}$ is

inherited from the global case (as seen in $T_{1,n}$ argument, $M_p(\overline{\mathbb{P}}_{n,\ell})$ is bounded for any $p$). This promotes weak convergence to $W_1(\overline{\mathbb{P}}_{n,\ell,\theta}, \mathbb{G}_{\ell,\theta}) \to 0$. Then, by Proposition 16, $W_1(\overline{\mathbb{P}}_{n,\ell}, \mathbb{G}_\ell) \to 0$.

Since $\mathbb{E}[W_1(\mathbb{P}_{n,\ell}, \mathbb{G}_\ell)] \to 0$ for each $\ell$, and $\pi_\ell$ are constants, $\mathbb{E}[T_{2,n}] \to 0$. Combining $\mathbb{E}[T_{1,n}] \to 0$ and $\mathbb{E}[T_{2,n}] \to 0$ completes the proof. □

**Proposition 13** (Part of Theorem 14)**.** *Consider the settings of Proposition 11. Let* $\widetilde{\mathbb{G}}_{n,\ell} = N(\mu_\ell, \Sigma_\ell/\nu_n)$*. Then for any* $R > 0$*:*

$$\mathbb{E}\left\{ \sup_{f_1,\ldots,f_L \in \mathrm{Lip}(R)} \left| \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in \mathcal{C}_\ell} f_\ell(\overline{\phi}_i^{(k)}) - \sum_{\ell=1}^{L} \pi_\ell \mathbb{E}_{Y \sim \widetilde{\mathbb{G}}_{n,\ell}}[f_\ell(Y)] \right| \right\} \to 0.$$

*Proof.* Let $\widetilde{\mathbb{P}}_{n,\ell}$ be the class-conditional empirical measure for $\overline{\phi}_i^{(k)}$ for class $\ell$:

$$\widetilde{\mathbb{P}}_{n,\ell}[f] = \frac{1}{|\mathcal{C}_\ell|} \sum_{i \in \mathcal{C}_\ell} f(\overline{\phi}_i^{(k)}).$$

Let $\Delta'_n$ be the term inside the overall expectation:

$$\Delta'_n := \sup_{f_1,\ldots,f_L \in \mathrm{Lip}(R)} \left| \sum_{\ell=1}^{L} \frac{|\mathcal{C}_\ell|}{n} \widetilde{\mathbb{P}}_{n,\ell}[f_\ell] - \sum_{\ell=1}^{L} \pi_\ell \widetilde{\mathbb{G}}_{n,\ell}[f_\ell] \right|.$$

Similar to the proof of Proposition 12, using the triangle inequality:

$$\begin{aligned} \Delta'_n \le & \sum_{\ell=1}^{L} \left| \frac{|\mathcal{C}_\ell|}{n} - \pi_\ell \right| \sup_{f_\ell \in \mathrm{Lip}(R)} |\widetilde{\mathbb{P}}_{n,\ell}[f_\ell]| \quad (:= T'_{1,n}) \\ & + \sum_{\ell=1}^{L} \pi_\ell \sup_{f_\ell \in \mathrm{Lip}(R)} \left| \widetilde{\mathbb{P}}_{n,\ell}[f_\ell] - \widetilde{\mathbb{G}}_{n,\ell}[f_\ell] \right| \quad (:= T'_{2,n}). \end{aligned}$$

The second term is $\sum_{\ell=1}^{L} \pi_\ell R \cdot W_1(\widetilde{\mathbb{P}}_{n,\ell}, \widetilde{\mathbb{G}}_{n,\ell})$. We can assume $f_\ell(0) = 0$ by replacing $f_\ell(x)$ with $f_\ell(x) - f_\ell(0)$ and noting that $|\widetilde{\mathbb{P}}_{n,\ell}[f_\ell(0)] - \widetilde{\mathbb{G}}_{n,\ell}[f_\ell(0)]| = |f_\ell(0) - f_\ell(0)| = 0$.

Before bounding the two terms, we first show that

$$\mathbb{E}\|\overline{\phi}_i^{(k)} - \mu_\ell\|_2 \to 0 \quad \text{uniformly for } i \in \mathcal{C}_\ell. \tag{.26}$$

By Lemma 33 and Lemma 34, $\mathbb{E}[\overline{\phi}_i^{(k)}] \to \mu_\ell$ for $i \in \mathcal{C}_\ell$. Next,

$$\mathrm{Var}(\overline{\phi}_i^{(k)}) = \mathrm{Var}(\xi_i^{(k)}/\sqrt{\nu_n}) = \Sigma_\ell/\nu_n + o(\nu_n^{-1}),$$

uniformly over $i \in \mathcal{C}_\ell$, by noting that the convergence in the proof of Proposition 11 is, in fact, uniform over $i \in \mathcal{C}_\ell$ and $\theta \in \mathbb{S}^{d-1}$. Since $\mathbb{E}\|\overline{\phi}_i^{(k)} - \mu_\ell\|_2 \le \mathbb{E}\|\overline{\phi}_i^{(k)} - \mathbb{E}[\overline{\phi}_i^{(k)}]\|_2 + \|\mathbb{E}[\overline{\phi}_i^{(k)}] - \mu_\ell\|_2$, and the first term is bounded by $\left(\mathrm{tr}\,\mathrm{Var}(\overline{\phi}_i^{(k)})\right)^{1/2} = O(\nu_n^{-1/2}) = o(1)$, and the second terms is $o(1)$ for $i \in \mathcal{C}_\ell$, the claim follows.

For $T'_{1,n}$: $\mathbb{E}[\sup_{f_\ell \in \mathrm{Lip}(R)} |\widetilde{\mathbb{P}}_{n,\ell}[f_\ell]|] \le R \cdot \mathbb{E}[\widetilde{\mathbb{P}}_{n,\ell}[\|x\|]] = R \cdot \frac{1}{|\mathcal{C}_\ell|}\sum_{i\in\mathcal{C}_\ell} \mathbb{E}[\|\overline{\phi}_i^{(k)}\|]$. By eq. (.26), $\mathbb{E}[\|\overline{\phi}_i^{(k)}\|]$ converges to $\|\mu_\ell\|$ which is bounded. Thus, $\sup_n \mathbb{E}[\sup_{f_\ell} |\widetilde{\mathbb{P}}_{n,\ell}[f_\ell]|]$ is bounded. Since $\left|\frac{|\mathcal{C}_\ell|}{n} - \pi_\ell\right| \to 0$ by Assumption 4, $\mathbb{E}[T'_{1,n}] \to 0$.

For $T'_{2,n}$: We need to show $\mathbb{E}[W_1(\widetilde{\mathbb{P}}_{n,\ell}, \widetilde{\mathbb{G}}_{n,\ell})] \to 0$ for each $\ell$. Let $f \in \mathrm{Lip}(R)$ with $f(0) = 0$. Let $\overline{\widetilde{\mathbb{P}}}_{n,\ell} = \mathbb{E}[\widetilde{\mathbb{P}}_{n,\ell}]$. We first analyze

$$|\overline{\widetilde{\mathbb{P}}}_{n,\ell}[f] - \widetilde{\mathbb{G}}_{n,\ell}[f]| \le \frac{1}{|\mathcal{C}_\ell|}\sum_{i\in\mathcal{C}_\ell} \mathbb{E}|f(\overline{\phi}_i^{(k)}) - \widetilde{\mathbb{G}}_{n,\ell}[f]|,$$

where $\overline{\widetilde{\mathbb{P}}}_{n,\ell}[f] = \frac{1}{|\mathcal{C}_\ell|}\sum_{i\in\mathcal{C}_\ell} \mathbb{E}[f(\overline{\phi}_i^{(k)})]$. Using the decomposition for a single $i \in \mathcal{C}_\ell$:

$$\mathbb{E}|f(\overline{\phi}_i^{(k)}) - \widetilde{\mathbb{G}}_{n,\ell}[f]| \le \mathbb{E}|f(\overline{\phi}_i^{(k)}) - f(\mu_\ell)| + \mathbb{E}|f(\mu_\ell) - \widetilde{\mathbb{G}}_{n,\ell}[f]|.$$

Let these two terms be $A_i, B_i$. (Note $B_i$ is actually independent of $i$ for $i \in \mathcal{C}_\ell$). Since $f \in \mathrm{Lip}(R)$:

- $A_i \le R \cdot \mathbb{E}\|\overline{\phi}_i^{(k)} - \mu_\ell\|_2 = O(\nu_n^{-1/2})$ uniformly for $i \in \mathcal{C}_\ell$, by eq. (.26).
- For $B_i$, we have

$$B_i = |\mathbb{E}_{Y\sim N(0,\Sigma_\ell/\nu_n)}[f(\mu_\ell) - f(\mu_\ell + Y)]| \le R \cdot \mathbb{E}_{Y\sim N(0,\Sigma_\ell/\nu_n)}[\|Y\|_2]$$

  and $\mathbb{E}_{Y\sim N(0,\Sigma_\ell/\nu_n)}[\|Y\|_2] \le \sqrt{\mathrm{tr}(\Sigma_\ell/\nu_n)} = O(\nu_n^{-1/2})$. So $B_i = O(\nu_n^{-1/2})$.

Thus, uniformly over $i \in \mathcal{C}_\ell$ and $f \in \mathrm{Lip}(R)$, we have $\mathbb{E}|f(\overline{\phi}_i^{(k)}) - \widetilde{\mathbb{G}}_{n,\ell}[f]| = O(\nu_n^{-1/2})$. This establishes $W_1(\overline{\widetilde{\mathbb{P}}}_{n,\ell}, \widetilde{\mathbb{G}}_{n,\ell}) \to 0$.

Now, for the concentration part $\mathbb{E}[W_1(\widetilde{\mathbb{P}}_{n,\ell}, \overline{\widetilde{\mathbb{P}}}_{n,\ell})] \to 0$: We will verify the conditions of Proposition 17 for the variables $X_{i,n} = \overline{\phi}_i^{(k)}$ for $i \in \mathcal{C}_\ell$:

(i) Uniform $\Psi_{r_n}$ sub-Gaussianity of $\langle \overline{\phi}_i^{(k)}, \theta \rangle$: Since $\overline{\phi}_i^{(k)} = \mathbb{E}[\overline{\phi}_i^{(k)}] + \xi_i^{(k)}/\sqrt{\nu_n}$, we have

$$\begin{aligned}\|\langle \overline{\phi}_i^{(k)}, \theta \rangle\|_{\Psi_{r_n}} &\le \|\langle \mathbb{E}[\overline{\phi}_i^{(k)}], \theta \rangle\|_{\Psi_{r_n}} + \|\langle \xi_i^{(k)}/\sqrt{\nu_n}, \theta \rangle\|_{\Psi_{r_n}} \\ &= |\langle \mathbb{E}[\overline{\phi}_i^{(k)}], \theta \rangle| \cdot \|1\|_{\Psi_{r_n}} + \|\langle \xi_i^{(k)}/\sqrt{\nu_n}, \theta \rangle\|_{\Psi_{r_n}}\end{aligned}$$

the first term is bounded in the limit by $C\langle \mu_\ell, \theta \rangle$ where $C = \limsup_{n\to\infty} \|1\|_{\Psi_{r_n}}$ is a universal constant, and the second term is $O(\nu_n^{-1/2})$ by Lemma 31, both uniformly over $i \in \mathcal{C}_\ell$ and $\theta \in \mathbb{S}^{d-1}$.

and $\mu_\ell$ is bounded, and $\xi_i^{(k)}/\sqrt{\nu_n}$ has vanishing $\Psi$ norm (as $\xi_i^{(k)}$ has bounded $\Psi$ norm), $\langle \overline{\phi}_i^{(k)}, \theta \rangle$ will have bounded $\Psi_{r_{N_\ell}}$ norm (dominated by $\langle \mu_\ell, \theta \rangle$ plus a small term).

(ii) Variance of empirical moments: $\mathrm{Var}(N_\ell^{-1} \sum_{i \in \mathcal{C}_\ell} \langle \overline{\phi}_i^{(k)}, \theta \rangle^r)$. Again, we use $\overline{\phi}_i^{(k)} = \mathbb{E}[\overline{\phi}_i^{(k)}] + \xi_i^{(k)}/\sqrt{\nu_n}$. By an argument similar to Lemma 35, we obtain

$$\|\langle \overline{\phi}_i^{(k)}, \theta \rangle^r - \langle \mathbb{E}\overline{\phi}_i^{(k)}, \theta \rangle^r\|_{L^2} \le \sum_{s=1}^{r} \binom{r}{s} \|\langle \mathbb{E}\overline{\phi}_i^{(k)}, \theta \rangle\|_{L^{2r}}^{r-s} \cdot \|\langle \xi_i^{(k)}/\sqrt{\nu_n}, \theta \rangle\|_{L^{2r}}^{s} \qquad (.27)$$

We have $\|\langle \mathbb{E}\overline{\phi}_i^{(k)}, \theta \rangle\|_{L^{2r}}^{r-s} = |\langle \mathbb{E}\overline{\phi}_i^{(k)}, \theta \rangle|^{r-s}$ since the quantity is deterministic. This is uniformly bounded over $i \in \mathcal{C}_\ell$ and $\theta \in \mathbb{S}^{d-1}$, by eq. (.26). Similarly, $\|\langle \xi_i^{(k)}, \theta \rangle\|_{L^{2r}}$ is uniformly bounded over $i \in \mathcal{C}_\ell$ and $\theta \in \mathbb{S}^{d-1}$, by the argument in the proof of Proposition 11 (the convergence of the moments is uniform over $i \in \mathcal{C}_\ell$). It follows that $\|\langle \xi_i^{(k)}/\sqrt{\nu_n}, \theta \rangle\|_{L^{2r}}^{s} = O(\nu_n^{-s/2}) = O(\nu_n^{-1/2})$ for $s \ge 1$, uniformly over $i$ and $\theta$. The same then applies to LHS of eq. (.27). This in turn implies $\|N_\ell^{-1} \sum_{i \in \mathcal{C}_\ell} \langle \overline{\phi}_i^{(k)}, \theta \rangle^r - N_\ell^{-1} \sum_{i \in \mathcal{C}_\ell} \langle \mathbb{E}\overline{\phi}_i^{(k)}, \theta \rangle^r\|_{L^2} = o(1)$. Now, $\mathrm{Var}(N_\ell^{-1} \sum_{i \in \mathcal{C}_\ell} \langle \mathbb{E}\overline{\phi}_i^{(k)}, \theta \rangle^r) = 0$ since this quantity is deterministic. This implies (see the inequality in the proof of Lemma 32) $\mathrm{Var}(N_\ell^{-1} \sum_{i \in \mathcal{C}_\ell} \langle \overline{\phi}_i^{(k)}, \theta \rangle^r) = o(1)$ which is the desired result.

(iii) $\sup_n M_1(\overline{\widetilde{\mathbb{P}}}_{n,\ell}) < \infty$: This was shown for $T'_{1,n}$.

Thus, by Proposition 17, $\mathbb{E}[W_1(\widetilde{\mathbb{P}}_{n,\ell}, \overline{\widetilde{\mathbb{P}}}_{n,\ell})] \to 0$.

Since $\mathbb{E}[W_1(\widetilde{\mathbb{P}}_{n,\ell}, \widetilde{\mathbb{G}}_{n,\ell})] \to 0$ for each $\ell$, it follows that $\mathbb{E}[T'_{2,n}] \to 0$. Combining $\mathbb{E}[T'_{1,n}] \to 0$ and $\mathbb{E}[T'_{2,n}] \to 0$ completes the proof. □

# B Moment Characterization in $W_p$

In the following $\{\mathbb{H}_n\}_{n\geq 1}$ and $\mathbb{H}$ are all (Borel) probability measures on $\mathbb{R}^d$.

**Proposition 14.** *Assume that $\{\mathbb{H}_n\}_{n\geq 1}$ is a sequence of (Borel) measures on $\mathbb{R}^d$ such that*

$$\sup_{n\geq 1} \int |\theta^T x|^r d\mathbb{H}_n(x) < \infty, \quad \textit{for all } \theta \in \mathbb{R}^d.$$

*Then,* $\sup_{n\geq 1} \int \|x\|^r d\mathbb{H}_n(x) < \infty$.

*Proof.* Let $\{\theta_1, \ldots, \theta_m\}$ be a $\frac{1}{2}$-net of the unit sphere $S^{d-1} = \{\theta \in \mathbb{R}^d : \ \|\theta\| = 1\}$. We have $\|x\| = \sup_{\theta \in S^{d-1}} |\theta^T x| \leq 2 \max_{i\in[m]} |\theta_i^T x|$. It follows that $\|x\|^r \leq 2^r \max_{i\in[m]} |\theta_i^T x|^r \leq 2^r \sum_{i=1}^m |\theta_i^T x|^r$, hence

$$\sup_{n\geq 1} \int \|x\|^r d\mathbb{H}_n(x) \leq 2^r \sum_{i=1}^m \sup_{n\geq 1} \int |\theta_i^T x|^r \, d\mathbb{H}_n(x) < \infty$$

proving the result. □

# C $\Psi_r$ sub-Gaussians

**Definition 7** ($\Psi_r$ sub-Gaussian)**.** Let $r \geq 2$ be a real number, and $\Psi_r : [0, \infty) \to [0, \infty)$ be defined by

$$\Psi_r(x) = \sum_{j=1}^{\lfloor r/2 \rfloor} \frac{x^{2j}}{j!}. \tag{.28}$$

The corresponding Orlicz (or Luxembourg) norm for a random variable $X$ is:

$$\|X\|_{\Psi_r} = \inf\{K > 0 : \mathbb{E}[\Psi_r(|X|/K)] \leq 1\}. \tag{.29}$$

**Lemma 39** (Norm equivalence)**.** *Let $X$ be a random variable and $r \geq 2$. The following holds:*

*(a) Norm implies moments: If $\|X\|_{\Psi_r} \leq K$ for some $K > 0$, then*

$$(\mathbb{E}|X|^p)^{1/p} \leq C_1 K\sqrt{p} \quad \textit{for all } p \in [2, 2\lfloor r/2 \rfloor]$$

*where $C_1 > 0$ is a universal constant.*

*(b) Moments imply norm: If $(\mathbb{E}|X|^p)^{1/p} \leq C\sqrt{p}$ for some $C > 0$ and for all $p \in [2, r]$, then*

$$\|X\|_{\Psi_r} \leq C_2 C$$

*where $C_2 = 2\sqrt{e}$.*

*Proof.* **Part (a)** Assume $\|X\|_{\Psi_r} \leq K$. By definition, $\mathbb{E}[\Psi_r(|X|/K)] \leq 1$.

$$\mathbb{E}\left[\sum_{j=1}^{\lfloor r/2 \rfloor} \frac{(|X|/K)^{2j}}{j!}\right] \leq 1$$

For any integer $j_0 \in [1, \lfloor r/2 \rfloor]$, let $p = 2j_0$. Since all terms in the sum are non-negative:

$$\mathbb{E}\left[\frac{|X|^p}{K^p j_0!}\right] \leq \mathbb{E}[\Psi_r(|X|/K)] \leq 1$$

So, $\mathbb{E}|X|^p \leq K^p j_0! = K^p (p/2)!$. Taking the $p$-th root: $(\mathbb{E}|X|^p)^{1/p} \leq K((p/2)!)^{1/p}$. Using the inequality $m! \leq e\sqrt{m}(m/e)^m$ for $m = p/2 \geq 1$:

$$((p/2)!)^{1/p} \leq (e\sqrt{p/2}(p/2e)^{p/2})^{1/p} = (e\sqrt{p/2})^{1/p}(p/2e)^{1/2} = (e\sqrt{p/2})^{1/p}\sqrt{\frac{p}{2e}}$$

The term $(e\sqrt{p/2})^{1/p}$ is bounded by a universal constant $c'$ for $p \geq 2$. (It tends to 1 as $p \to \infty$). Thus, $(\mathbb{E}|X|^p)^{1/p} \leq Kc'\sqrt{1/(2e)}\sqrt{p}$ for even integers $p \in [2, 2\lfloor r/2 \rfloor]$. Now, let $p \in [2, 2\lfloor r/2 \rfloor]$ be any real number. Let $q = 2\lceil p/2 \rceil$. Then $q$ is an even integer, $p \leq q \leq p + 1 < p + 2$, and $q \leq 2\lceil (2\lfloor r/2 \rfloor)/2 \rceil = 2\lfloor r/2 \rfloor$. By Lyapunov's inequality:

$$(\mathbb{E}|X|^p)^{1/p} \leq (\mathbb{E}|X|^q)^{1/q} \leq Kc'\sqrt{1/(2e)}\sqrt{q}$$

Since $q \le p+2$ and $p \ge 2$, we have $q \le p+p = 2p$. So $\sqrt{q} \le \sqrt{2p} = \sqrt{2}\sqrt{p}$. Therefore, $(\mathbb{E}|X|^p)^{1/p} \le Kc'\sqrt{1/(2e)}\sqrt{2}\sqrt{p} = (c'\sqrt{1/e})K\sqrt{p}$. Setting $C_1 = c'\sqrt{1/e}$ (a universal constant) proves the first part.

**Part (b)** Assume $(\mathbb{E}|X|^p)^{1/p} \le C\sqrt{p}$ for $p \in [2, r]$. We want to find $k$ such that $\mathbb{E}[\Psi_r(|X|/k)] \le 1$.

$$\mathbb{E}[\Psi_r(|X|/k)] = \sum_{j=1}^{\lfloor r/2 \rfloor} \frac{\mathbb{E}[|X|^{2j}]}{k^{2j} j!}$$

Let $p = 2j$. Since $j \in [1, \lfloor r/2 \rfloor]$, $p \in [2, 2\lfloor r/2 \rfloor]$. This range is contained in $[2, r]$. So we can use the moment bound: $\mathbb{E}|X|^p \le (C\sqrt{p})^p = C^p p^{p/2}$.

$$\mathbb{E}[\Psi_r(|X|/k)] \le \sum_{j=1}^{\lfloor r/2 \rfloor} \frac{C^{2j}(2j)^j}{k^{2j} j!}$$

Using the bound $(2j)^j/j! \le (2e)^j$:

$$\mathbb{E}[\Psi_r(|X|/k)] \le \sum_{j=1}^{\lfloor r/2 \rfloor} \frac{C^{2j}(2e)^j}{k^{2j}} = \sum_{j=1}^{\lfloor r/2 \rfloor} \left(\frac{2eC^2}{k^2}\right)^j$$

This is a geometric series with ratio $R = 2eC^2/k^2$. If we choose $k$ such that $R \le 1/2$, the sum is bounded by $\sum_{j=1}^{\infty}(1/2)^j = 1$. We need $2eC^2/k^2 \le 1/2$, which means $k^2 \ge 4eC^2$. Let $k = \sqrt{4e}C = 2\sqrt{e}C$. With this choice of $k$, we have $\mathbb{E}[\Psi_r(|X|/k)] \le 1$. By the definition of the norm, $\|X\|_{\Psi_r} \le k = 2\sqrt{e}C$. Setting $C_2 = 2\sqrt{e}$ proves the second part. □

**Lemma 40** (Tail bound)**.** *Let $Y$ be a random variable and $r \ge 2$. Suppose $\|Y\|_{\Psi_r} \le K$ for some $K > 0$. Then there exists a universal constant $c_0 > 0$ such that for all $t \ge c_0 K$:*

$$\mathbb{P}(|Y| \ge t) \le \exp\Big(-c_1 \min\Big\{\frac{t^2}{K^2}, \lfloor r/2 \rfloor\Big\}\Big)$$

*where $c_1 = 1/(4C_1^2 e)$ and $C_1$ is the universal constant from Lemma 39(a). The threshold constant is $c_0 = 2C_1\sqrt{e}$.*

*Proof.* The assumption $\|Y\|_{\Psi_r} \le K$ implies $(\mathbb{E}|Y|^p)^{1/p} \le C_1 K\sqrt{p}$ for all $p \in [2, 2\lfloor r/2 \rfloor]$ by Lemma 39(a). Let $r_0' = 2\lfloor r/2 \rfloor$. This matches the condition (56) of [VA24b, Lemma 25]

with $\Delta = Y$, $\eta = 1/2$, $K_{lem} = K$, $C_{lem} = 2C_1^2$, and $r_0$ replaced by $r_0'$. Lemma 25 applies for $x \geq 4C_{lem}\eta e = 4(2C_1^2)(1/2)e = 4C_1^2 e$. It gives the tail bound:

$$\mathbb{P}(|Y| \geq Kx^{1/2}) \leq \exp\Big( - \min\Big\{ \frac{x}{2C_{lem}e},\, \eta r_0' \Big\}\Big) = \exp\Big( - \min\Big\{ \frac{x}{4C_1^2 e},\, \lfloor r/2 \rfloor \Big\}\Big)$$

Let $t = Kx^{1/2}$, so $x = (t/K)^2$. The condition on $x$ becomes $t \geq K\sqrt{4C_1^2 e} = 2C_1\sqrt{e}K$. Substituting $x$ in the bound yields:

$$\mathbb{P}(|Y| \geq t) \leq \exp\Big( - \min\Big\{ \frac{(t/K)^2}{4C_1^2 e},\, \lfloor r/2 \rfloor \Big\}\Big)$$

Setting $c_1 = 1/(4C_1^2 e)$ and $c_0 = 2C_1\sqrt{e}$ gives the desired result. □

# D Results on Triangular Arrays

**Proposition 15.** *Let $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_{i,n}}$ be the empirical measure of real-valued random variables $Y_{i,n}$ for $i \in [n]$, and let $\bar{\mu}_n = \mathbb{E}\mu_n$. Assume that for some sequence $r_n = \omega(1)$, we have*

*(a) $\{Y_{i,n}\}_{i=1}^n$ is uniformly $\Psi_{r_n}$ sub-Gaussian, that is, there exists $\zeta > 0$ such that $\sup_{i\in[n]} \|Y_{i,n}\|_{\Psi_{r_n}} \leq \zeta$.*

*(b) $\mathrm{var}\big(n^{-1}\sum_{i=1}^n Y_{i,n}^r\big) \to 0$ as $n \to \infty$ for all $r \in \mathbb{N}$.*

*Then, $\mathbb{E}\big[W_1(\mu_n, \bar{\mu}_n)\big] \to 0$ as $n \to \infty$.*

The full proof of this proposition will be deferred for a next section.

**Lemma 41.** *$W_1(\eta_{\theta_1}, \eta_{\theta_2}) \leq \|\theta_1 - \theta_2\| M_1(\eta)$, for any probability measure $\eta$ on $\mathbb{R}^d$ and $\theta_1, \theta_2 \in \mathbb{R}^d$.*

*Proof.* Let $X \sim \eta$. Using the dual formulation of $W_1$ for measures on $\mathbb{R}$:

$$\begin{aligned} W_1(\eta_{\theta_1}, \eta_{\theta_2}) &= \sup_{f\in\mathcal{L}} \big|\mathbb{E}f(\theta_1^T X) - \mathbb{E}f(\theta_2^T X)\big| \\ &\leq \sup_{f\in\mathcal{L}} \mathbb{E}\big|f(\theta_1^T X) - f(\theta_2^T X)\big| \leq \mathbb{E}|\theta_1^T X - \theta_2^T X| \leq \|\theta_1 - \theta_2\| M_1(\eta). \end{aligned}$$

This completes the proof. □

**Proposition 16.** *Let $\{\mu_n\}_{n\geq 1}$ and $\{\eta_n\}_{n\geq 1}$ be random probability measures on $\mathbb{R}^d$. Let $\bar{\mu}_n = \mathbb{E}\mu_n$ and $\bar{\eta}_n = \mathbb{E}\eta_n$. Assume that*

$$\sup_{n\geq 1}\big(M_1(\bar{\mu}_n) + M_1(\bar{\eta}_n)\big) < \infty, \tag{.30}$$

*and $\mathbb{E}\big[W_1(\mu_{n,\theta}, \eta_{n,\theta})\big] \to 0$ as $n \to \infty$ for every $\theta \in S^{d-1}$. Then, $\mathbb{E}\big[W_1(\mu_n, \eta_n)\big] \to 0$ as $n \to \infty$.*

*Proof.* The map $\theta \mapsto W_1(\mu_{n,\theta}, \eta_{n,\theta})$ is Lipschitz with constant $L_n := M_1(\mu_n) + M_1(\eta_n)$. This is shown in [BG], and we reproduce the argument here for completeness. The triangle inequality for $W_1$ gives,

$$W_1(\mu_{n,\theta_1}, \eta_{n,\theta_1}) \leq W_1(\mu_{n,\theta_1}, \mu_{n,\theta_2}) + W_1(\mu_{n,\theta_2}, \eta_{n,\theta_2}) + W_1(\eta_{n,\theta_2}, \eta_{n,\theta_1}).$$

Rearranging yields

$$\begin{aligned} W_1(\mu_{n,\theta_1}, \eta_{n,\theta_1}) - W_1(\mu_{n,\theta_2}, \eta_{n,\theta_2}) &\leq W_1(\mu_{n,\theta_1}, \mu_{n,\theta_2}) + W_1(\eta_{n,\theta_1}, \eta_{n,\theta_2}) \\ &\leq \|\theta_1 - \theta_2\| M_1(\mu_n) + \|\theta_1 - \theta_2\| M_1(\eta_n) \\ &= L_n \|\theta_1 - \theta_2\|. \end{aligned}$$

where the second inequality follows from Lemma 41. Switching $\theta_1$ and $\theta_2$ shows that the inequality holds with the LHS replaced with its absolute value, proving Lipschitz continuity.

Let $F_n(\theta) = W_1(\mu_{n,\theta}, \eta_{n,\theta})$. By the result of [BG], there is a constant $C(d)$ such that

$$W_1(\mu_n, \eta_n) \leq C(d) \max_{\theta \in S^{d-1}} F_n(\theta).$$

Let $\theta_1, \theta_2, \ldots, \theta_N$ be a $\varepsilon$-net of $S^{d-1}$, with $N = N(\varepsilon)$ finite. For every $\theta \in S^{d-1}$, there is a $\theta_j$ such that $F_n(\theta) \leq L_n\varepsilon + F_n(\theta_j) \leq L_n\varepsilon + \sum_{i=1}^N F_n(\theta_i)$. It follows that

$$\mathbb{E} \max_{\theta \in S^{d-1}} F_n(\theta) \leq \mathbb{E}[L_n] \cdot \varepsilon + \sum_{i=1}^N \mathbb{E}[F_n(\theta_i)]$$

Bounding $\mathbb{E}[L_n]$ further by $\sup_{n\geq 1}\mathbb{E}[L_n]$ and noting that $\mathbb{E}[L_n] = M_1(\bar{\mu}_n) + M_1(\bar{\eta}_n)$, we have

$$\mathbb{E}[W_1(\mu_n,\eta_n)] \;\leq\; C(d)\Big\{\varepsilon \sup_{m\geq 1}\big(M_1(\bar{\mu}_m) + M_1(\bar{\eta}_m)\big) + \sum_{i=1}^{N}\mathbb{E}W_1(\mu_{n,\theta_i},\eta_{n,\theta_i})\Big\}.$$

The sum goes to zero by assumption as $n\to\infty$, and the first term goes to zero taking $\varepsilon\downarrow 0$ and using (.30). □

**Proposition 17.** *Let $\mu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{Y_{i,n}}$ be the empirical measure of vector-valued random variables $Y_{i,n}\in\mathbb{R}^d$ for $i\in[n]$, and let $\bar{\mu}_n = \mathbb{E}\mu_n$. Assume that for some sequence $r_n = \omega(1)$ and for any $\theta\in S^{d-1}$,*

*(a) $\{\langle\theta, Y_{i,n}\rangle\}_{i=1}^{n}$ is uniformly $\Psi_{r_n}$ sub-Gaussian, that is, there exists $\zeta(\theta)>0$, such that $\sup_{i\in[n]}\|\langle\theta,Y_{i,n}\rangle\|_{\Psi_{r_n}}\leq\zeta(\theta)$.*

*(b) $\operatorname{var}\big(n^{-1}\sum_{i=1}^{n}\langle\theta,Y_{i,n}\rangle^r\big)\to 0$ as $n\to\infty$ for all $r\in\mathbb{N}$.*

*(c) $\sup_{n\geq 1} M_1(\bar{\mu}_n) < \infty$.*

*Then, $\mathbb{E}\big[W_1(\mu_n,\bar{\mu}_n)\big]\to 0$ as $n\to\infty$.*

*Proof.* First, note that $\mathbb{E}\mu_{n,\theta} = \bar{\mu}_{n,\theta}$. By Proposition 15, and assumptions (a) and (b), we have that $\mathbb{E}\big[W_1(\mu_{n,\theta},\bar{\mu}_{n,\theta})\big]\to 0$ as $n\to\infty$ for every $\theta\in S^{d-1}$. Next, applying Proposition 16 with $\eta_n = \bar{\mu}_n$ and noting that $\bar{\nu}_n = \bar{\mu}_n$, the result follows. □

## E Proof of Proposition 15

*Proof.* Let us write $\operatorname{Lip}(f) = \sup_{x\neq y}\frac{|f(x)-f(y)|}{|x-y|}$ for the Lipschitz constant of $f$. Consider the set of functions

$$\mathcal{L} = \{f:\mathbb{R}\to\mathbb{R} \mid \operatorname{Lip}(f)\leq 1, f(0)=0\},\quad \mathcal{L}_B = \{f1_{|x|\leq B} \mid f\in\mathcal{L},\ B>0\}.$$

Let $\varpi_n := \mu_n - \bar{\mu}_n$. By the dual characterization of $W_1$, we have

$$W_1(\mu_n,\bar{\mu}_n)\leq \sup_{f\in\mathcal{L}}|\varpi_n f|.$$

By breaking $f = f1_{|x|\le B} + f1_{|x|>B}$, we have

$$W_1(\mu_n, \bar{\mu}_n) \le \sup_{f\in\mathcal{L}_B} |\varpi_n f| + \sup_{f\in\mathcal{L}} |\varpi_n(f1_{|x|>B})|. \tag{.31}$$

Fix $\varepsilon \in (0,1)$ and consider the second term first. For any integrable $f$, we have

$$\begin{aligned} |\varpi_n(f1_{|x|>B})| &\le |\mu_n(f1_{|x|>B})| + |\bar{\mu}_n(f1_{|x|>B})| \\ &\le \mu_n(|f|1_{|x|>B}) + \bar{\mu}_n(|f|1_{|x|>B}). \end{aligned} \tag{.32}$$

For $f \in \mathcal{L}$, we have $|f(x)| = |f(x) - f(0)| \le |x - 0|$. Then, we have

$$|\varpi_n(f1_{|x|>B})| \le \mu_n(|x|1_{|x|>B}) + \bar{\mu}_n(|x|1_{|x|>B})$$

Taking the supremum over $f \in \mathcal{L}$ and then expectation, we have

$$\mathbb{E}\sup_{f\in\mathcal{L}} |\varpi_n(f1_{|x|>B})| \;\le\; 2\bar{\mu}_n(|x|1_{|x|>B}) = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}\big(|Y_{i,n}|1\{|Y_{i,n}| > B\}\big).$$

Take $n$ large enough so that

$$r_n \ge 2\Big(\frac{B^2}{\zeta^2} + 1\Big) \tag{.33}$$

which we will verify at the end. Also, take $B \ge B_0(\zeta) := c_0\zeta$ where $c_0$ is the constant in Lemma 40. Then, by this lemma, we have $\mathbb{P}(|Y_{i,n}| > B) \le \exp(-c_1B^2/\zeta^2)$, and by Lemma 40, we have $\mathbb{E}[Y_{i,n}^2] \le 2C_1^2\zeta^2$. Then, by Cauchy-Schwarz, we have

$$\mathbb{E}\big(|Y_{i,n}|1\{|Y_{i,n}| > B\}\big) \le \sqrt{\mathbb{E}[|Y_{i,n}|^2] \cdot \mathbb{P}(|Y_{i,n}| > B)} \le \sqrt{2}C_1\zeta \cdot \exp(-cB^2/2\zeta^2).$$

Taking $B \ge B_1(\zeta)$ for $B_1(\zeta)$ large enough, the RHS can be made $\le \varepsilon$, which gives

$$\mathbb{E}\sup_{f\in\mathcal{L}} |\varpi_n(f1_{|x|>B})| \le 2\varepsilon.$$

Consider now the first term in (.31). Viewing $\mathcal{L}_B$ as a subspace of $(C_b([-B,B]), \|\cdot\|_\infty)$, by restricting to $[-B,B]$, $\mathcal{L}_B$ is uniformly bounded and equicontinuous, hence by Arzelà–Ascoli, it is relatively compact in the sup-norm topology. This, in turn, implies $\mathcal{L}_B$ is totally bounded. Then, there exists $f_1, \ldots, f_M \in \mathcal{L}_B$ that form an $\varepsilon$-net for $\mathcal{L}_B$ in sup-norm, for

some $M = M(\varepsilon, B) < \infty$. That is, for any $f \in \mathcal{L}_B$, there is $f_\ell$ such that $\|f - f_\ell\|_\infty \le \varepsilon$, hence

$$\begin{aligned} |\varpi_n f| \;&\le\; |\varpi_n(f - f_\ell)| + |\varpi_n f_\ell| \\ &\le\; \|\varpi_n\|_{\mathrm{TV}} \cdot \|f - f_\ell\|_\infty + |\varpi_n f_\ell| \;\le\; 2\varepsilon + |\varpi_n f_\ell|. \end{aligned}$$

Taking supremum over $f \in \mathcal{L}_B$, we have

$$\sup_{f \in \mathcal{L}_B} |\varpi_n f| \;\le\; 2\varepsilon + \sup_{\ell \in [M]} |\varpi_n f_\ell|.$$

Take $B \ge 3$. By Lemma 43, each $f_\ell$ admits a (truncated) polynomial $Q_\ell(x) = 1\{|x| \le B\} \cdot \sum_{j=0}^m c_{j\ell} x^j$, with $m = 4\lceil C_2 B/\varepsilon \rceil \in 4\mathbb{N}$ (can take $C_2 = 18$) such that

$$\|f_\ell - Q_\ell\|_\infty \le \varepsilon,$$

and $|c_{j\ell}| \le 6B \cdot 3^{m-j} =: a_j$ for all $j \ge 0$ and $\ell \in [M]$. We have

$$|\varpi_n f_\ell| \le \|\varpi_n\|_{\mathrm{TV}} \cdot \|f_\ell - Q_\ell\|_\infty + |\varpi_n Q_\ell|.$$

It follows that

$$\sup_{\ell \in [M]} |\varpi_n f_\ell| \le 2\varepsilon + \sup_{\ell \in [M]} |\varpi_n Q_\ell|$$

and we have

$$\begin{aligned} \sup_{\ell \in [M]} |\varpi_n Q_\ell| &\le \sup_{\ell \in [M]} \Big| \sum_{j=0}^m c_{j\ell}\, \varpi_n(x^j 1_{|x| \le B}) \Big| \\ &\le \sum_{j=0}^m \big( \sup_{\ell \in [M]} |c_{j\ell}| \big) \cdot |\varpi_n(x^j 1_{|x| \le B})| \le \sum_{j=0}^m a_j\, |\varpi_n(x^j 1_{|x| \le B})| \end{aligned}$$

We have

$$|\varpi_n(x^j 1_{|x| \le B})| \le |\varpi_n(x^j)| + |\varpi_n(x^j 1_{|x| > B})|.$$

Then, for the second term, using (.32), we have, for all $j \in [m]$,

$$\begin{aligned} |\varpi_n(x^j 1_{|x| > B})| &\le \mu_n(|x^j| 1_{|x| > B}) + \bar{\mu}_n(|x^j| 1_{|x| > B}) \\ &\le \mu_n(|x^m| 1_{|x| > B}) + \bar{\mu}_n(|x^m| 1_{|x| > B}). \end{aligned}$$

Taking maximum over $j \in [m]$, followed by expectation, we have

$$\mathbb{E} \sup_{j\in[m]} |\varpi_n(x^j 1_{|x|>B})| \le 2\bar{\mu}_n(|x|^m 1_{|x|>B}) = \frac{2}{n}\sum_{i=1}^n \mathbb{E}\big(|Y_{i,n}|^m 1\{|Y_{i,n}| > B\}\big).$$

Take $n$ large enough so that

$$r_n \ge 2m = 8\lceil C_2 B/\varepsilon \rceil, \tag{.34}$$

which we will verify at the end. Then, by Lemma 40 we have $\mathbb{E}[|Y_{i,n}|^{2m}] \le (C_1\zeta)^{2m}(2m)^m = (2C_1^2\zeta^2 m)^m$. Then, by Cauchy-Schwarz, we have

$$\begin{aligned}\mathbb{E}\big(|Y_{i,n}|^m 1\{|Y_{i,n}| > B\}\big) &\le \sqrt{\mathbb{E}[|Y_{i,n}|^{2m}] \cdot \mathbb{P}(|Y_{i,n}| > B)} \\ &\le (2C_1^2\zeta^2 m)^m \cdot \exp(-cB^2/2\zeta^2)\end{aligned}$$

Using $a_j = 6B \cdot 3^{m-j}$, we have $\sum_{j=0}^m a_j \le 9B \cdot 3^m$. It follows that

$$\begin{aligned}\mathbb{E}\Big[\sum_{j=0}^m a_j |\varpi_n(x^j 1_{|x|>B})|\Big] &\le \Big(\sum_{j=0}^m a_j\Big) \cdot \mathbb{E} \sup_{j\in[m]} |\varpi_n(x^j 1_{|x|>B})| \\ &\le 9B \cdot 3^m \cdot 2(2C_1^2\zeta^2 m)^m \cdot \exp(-cB^2/2\zeta^2) \\ &\le 18 \exp\Big(\log B + m\log(6C_1^2\zeta^2 m) - cB^2/2\zeta^2\Big) \\ &\le 18 \exp\bigg(\log B + 4\lceil C_2 B/\varepsilon\rceil \log\Big(24C_1^2\zeta^2\lceil C_2 B/\varepsilon\rceil\Big) - cB^2/2\zeta^2\bigg).\end{aligned}$$

Since $B^2$ grows faster than $B\log B$, the RHS can be made $\le \varepsilon$ for $B \ge B_2(\zeta,\varepsilon)$ for some $B_2(\zeta,\varepsilon)$ large enough. For this choice of $B$, we have

$$\begin{aligned}\mathbb{E} \sup_{\ell\in[M]} |\varpi_n Q_\ell| &\le \sum_{j=0}^m a_j \mathbb{E}|\varpi_n(x^j)| + \varepsilon \\ &\le \sum_{j=0}^m a_j \operatorname{var}\Big(\frac{1}{n}\sum_{i=1}^n Y_{i,n}^j\Big) + \varepsilon.\end{aligned}$$

By assumption

$$\max_{0\le j\le m} \operatorname{var}\Big(\frac{1}{n}\sum_{i=1}^n Y_{i,n}^j\Big) \le \varepsilon/(\sum_{j=0}^m a_j) \tag{.35}$$

for sufficiently large $n$. This gives $\mathbb{E}\sup_{\ell\in[M]}|\varpi_n Q_\ell| \le 2\varepsilon$. Putting the pieces together, we have

$$\mathbb{E}\sup_{f\in\mathrm{Lip}_B}|\varpi_n f| \le 2\varepsilon + 2\varepsilon + 2\varepsilon = 6\varepsilon.$$

All in all, taking $B = \max\{3, B_0(\zeta), B_1(\zeta), B_2(\zeta,\varepsilon)\}$, and $n$ large enough so that (.33) and (.34) are satisfied for the chosen $B$, and (.35) holds, we obtain $\mathbb{E}W_1(\mu_n, \bar{\mu}_n) \le 8\varepsilon$. The proof is complete. □

**Lemma 42.** *Let $T_k$ be the $k$th Chebyshev polynomial, and let $[T_k]_j$ be the coefficient of $x^j$ in $T_k(x)$. Then, $|[T_k]_0| \le 1$ and*

$$\max_{1\le j\le k}|[T_k]_j| \le (1+\sqrt{2})^k \le 3^k.$$

*Proof.* The first part is clear, since $[T_k]_0 \in \{0,1\}$. For the second part, from the recurrence relation $T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$, we have

$$|[T_{k+1}]_j| \le 2|[T_k]_{j-1}| + |[T_{k-1}]_j|.$$

Assuming the result holds as $\max_{1\le j\le k}|[T_k]_j| \le c^k$ for some constant $c$ and for all $T_r, r \le k$, we have $|[T_{k+1}]_j| \le 2\cdot c^k + c^{k-1}$. Then, if $2c^k + c^{k-1} \le c^{k+1}$, the result follows by induction. But this holds for $c \ge 1+\sqrt{2}$. The proof is complete. □

**Lemma 43** (Chebyshev–Jackson approximation)**.** *Let $B \ge 3$. Then, for any $f : [-B, B] \to \mathbb{R}$ 1-Lipschitz with $f(0) = 0$, there exists a polynomial $P(x) = \sum_{j=0}^m c_j x^j$, with $m \in 4\mathbb{N}$, such that*

$$\sup_{x\,\in\,[-B,B]}|f(x) - P(x)| \le \frac{18B}{m}, \quad |c_j| \le 6B\cdot 3^{m-j}, \quad \textit{for all } j \ge 0.$$

*Proof.* Consider an $L$-Lipschitz function $g$ on $[-1,1]$ with $g(0) = 0$. Then, for each $m \in 4\mathbb{N}$, there is a polynomial of the form

$$Q_m(x) = \sum_{k=0}^{m} \lambda_{k,m} a_k(g) T_k(x)$$

where $\lambda_{k,m}$ are derived from a Jackson kernel, satisfying $0 \le \lambda_{k,m} \le 1$ and $a_k(g)$ are the Chebyshev coefficients of $g$, such that

$$\sup_{x\in[-1,1]} |g(x) - Q_m(x)| \le \frac{18L}{m}, \quad |a_k(g)| \le \frac{\sqrt{8/\pi}L}{k}, \quad k \ge 1.$$

See Facts 3.2 and 3.3 in [BKM]. The Chebyshev coefficients are given by

$$a_k(g) = \frac{2}{\pi}\int_{-1}^{1} \frac{g(x)T_k(x)}{\sqrt{1-x^2}}dx, \quad k \ge 1,$$

and for $k = 0$, the same fomrula holds with $2/\pi$ replaced with $1/\pi$. For $k = 0$, using $g(0) = 0$ so that $|g(x)| \le L|x|$ for all $x \in [-1, 1]$, and $T_0(x) = 1$, we have

$$|a_0(g)| \le \frac{1}{\pi}\int_{-1}^{1} \frac{L|x|}{\sqrt{1-x^2}}dx = \frac{2L}{\pi}.$$

Thus a crude upper bound that works for all $k \ge 0$ is $|a_k(g)| \le 2L$.

Let $a_{k,m} = \lambda_{k,m}a_k(g)$ and note that $|a_{k,m}| \le 2L$ for all $k \ge 0$, by the above discussion. Rewriting $Q_m(x) = \sum_{j=0}^m b_j x^j$, one has $b_j = \sum_{k=j}^m a_{k,m}[T_k]_j$ where $[T_k]_j$ is the coefficient of $x^j$ in $T_k(x)$. It follows that

$$|b_j| \le \sum_{k=j}^{m} 2L \cdot 3^k \le 2L \cdot 3^m \sum_{k=j}^{m} 3^{k-m} \le 2L \cdot 3^m \frac{1}{1-3^{-1}} \le 6L \cdot 3^m$$

for all $j \ge 0$.

If $f$ is 1-Lipschitz on $[-B, B]$ with $f(0) = 0$, then $g(x) = f(Bx)$ is $B$-Lipschitz on $[-1, 1]$ with $g(0) = 0$. Let $Q_m$ be the above polynomial for $g$, and let $P(x) = Q_m(x/B) = \sum_{j=0}^m (b_j/B^j)x^j =: \sum_{j=0}^m c_j x^j$. Then,

$$|c_j| \le 6B\frac{3^m}{B^j} \le 6B \cdot 3^{m-j}$$

assuming $B \ge 3$. We also have $\sup_{x\in[-B,B]} |f(x) - P(x)| = \sup_{x\in[-1,1]} |g(x) - Q_m(x)| \le \frac{18B}{m}$. The proof is complete. □

# F Remaining proofs

## F.1 Proof of Lemma 36

Let $\mathcal{W}_k(i)$ be the set of directed, length $k$ walks starting at node $i \in [n]$. We consider $r$-tuples of walks called *walk sequences* where $\boldsymbol{w} \in \mathcal{W}_k^r(i)$ gives $\boldsymbol{w} = (\boldsymbol{w}^s)_{s=1}^r$ with $\boldsymbol{w}^s \in \mathcal{W}_k(i)$. We define the last vertex projection $\mathfrak{p} : \mathcal{W}_k(i) \to [n]$ and walk products $A_{\boldsymbol{w}^s} \coloneqq \prod_{\ell=1}^k A_{i_\ell j_\ell}$ with $\boldsymbol{w}^s = ((i_\ell, j_\ell))_{\ell=1}^k$.

Relating back to $\Delta_{i,\theta}$, let

$$\varrho(\boldsymbol{w}) = \mathbb{E}\Big[\prod_{s=1}^r (A_{\boldsymbol{w}^s} - \mathbb{E}[A_{\boldsymbol{w}^s}]) x_{\mathfrak{p}(\boldsymbol{w}^s)}\Big]$$

with $x \coloneqq X\theta$. Then

$$\mathbb{E}[\mathring{\Delta}_{i,\theta}^r] = \sum_{\boldsymbol{w} \in \mathcal{W}_k^r(i)} \varrho(\boldsymbol{w}).$$

Further let $[w]$ and $[\![w]\!]$ be the set of unique edges and vertices, respectively, found on a walk $w$. A walk sequence $\boldsymbol{w}$ is said to be *overlapping* if for every $s \in [r]$ there exists a distinct $s' \in [r]$ such that $[\boldsymbol{w}^s] \cap [\boldsymbol{w}^{s'}] \neq \varnothing$. Walk sequence which are not overlapping have $\varrho(\boldsymbol{w}) = 0$. For this reason we define the following walk sets

$$\mathcal{N}_{r,t,v}(i) \coloneqq \{\boldsymbol{w} \in \mathcal{W}_k^r(i) : \boldsymbol{w} \text{ overlapping}, \ |[\boldsymbol{w}]| = t, \ |[\![\boldsymbol{w}]\!]| = v\} \tag{.36}$$

where $[\boldsymbol{w}] \coloneqq \bigcup_{s=1}^r [\boldsymbol{w}^s]$ and $[\![\boldsymbol{w}]\!] \coloneqq \bigcup_{s=1}^r [\![\boldsymbol{w}^s]\!]$.

The walk sets $\{\mathcal{N}_{r,t,v}(i)\}_{t,v}$ form a partition for $\mathcal{W}_k^r(i)$ with $2 \leq v \leq t+1$ and $1 \leq t \leq t_*$ where $t_* \leq rk - \lceil r/2 \rceil$. This gives the sum equivalence

$$\sum_{\boldsymbol{w} \in \mathcal{W}_k^r(i)} \varrho(\boldsymbol{w}) = \sum_{t=1}^{t_*} \sum_{v=2}^{t+1} \sum_{\boldsymbol{w} \in \mathcal{N}_{r,t,v}(i)} \varrho(\boldsymbol{w}),$$

which gives fine-grained control of $\varrho(\boldsymbol{w})$ for the specific walk sets $\mathcal{N}_{r,t,v}(i)$.

To prove the result, start by expanding the variance of the $r$-empirical moment of $\gamma$,

$$\mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n}\Delta_{i,\theta}^{r}\Big)=\frac{1}{n^2}\sum_{i,i'}\mathbb{E}[\Delta_{i,\theta}^{r}\Delta_{i',\theta}^{r}]-\mathbb{E}[\Delta_{i,\theta}^{r}]\mathbb{E}[\Delta_{i',\theta}^{r}]. \tag{.37}$$

By the $n^{-2}$ scaling over $i,i'\in[n]$, it suffices to show

$$\mathrm{Cov}\big(\Delta_{i,\theta}^{r},\Delta_{\theta,i'}^{r}\big)=\mathbb{E}[\Delta_{i,\theta}^{r}\Delta_{i',\theta}^{r}]-\mathbb{E}[\Delta_{i,\theta}^{r}]\mathbb{E}[\Delta_{\theta,i'}^{r}]\lesssim n^{-1\{i\neq i'\}},$$

for every $i,i'\in[n]$.

Introduce the new notation for walk-sequence pairs $(\boldsymbol{w},\widetilde{\boldsymbol{w}})$

$$\varrho(\boldsymbol{w},\widetilde{\boldsymbol{w}})=\mathbb{E}\Big\{\Big(\prod_{s=1}^{r}(A_{\boldsymbol{w}^s}-\mathbb{E}[A_{\boldsymbol{w}^s}])x_{\mathfrak{p}(\boldsymbol{w}^s)}\Big)\Big(\prod_{s=1}^{r}(A_{\widetilde{\boldsymbol{w}}^s}-\mathbb{E}[A_{\widetilde{\boldsymbol{w}}^s}])x_{\mathfrak{p}(\widetilde{\boldsymbol{w}}^s)}\Big)\Big\}.$$

Then, the walk-linearized covariance expansion is

$$\mathrm{Cov}\big(\Delta_{i,\theta}^{r},\Delta_{\theta,i'}^{r}\big)=\frac{1}{\nu_n^{r(2k-1)}}\sum_{(\boldsymbol{w},\widetilde{\boldsymbol{w}})\in\mathcal{W}_k^r(i)\times\mathcal{W}_k^r(i')}\varrho(\boldsymbol{w},\widetilde{\boldsymbol{w}})-\varrho(\boldsymbol{w})\varrho(\widetilde{\boldsymbol{w}}). \tag{.38}$$

We are interested in the case $\varrho(\boldsymbol{w},\widetilde{\boldsymbol{w}})$ does not factorize as $\varrho(\boldsymbol{w},\widetilde{\boldsymbol{w}})=\varrho(\boldsymbol{w})\varrho(\widetilde{\boldsymbol{w}})$. Collect walk pairs under the concatenation notation $\boldsymbol{w}|\widetilde{\boldsymbol{w}}=(\boldsymbol{w}^1,\dots,\boldsymbol{w}^r,\widetilde{\boldsymbol{w}}^1,\dots,\widetilde{\boldsymbol{w}}^r)$ and define the walk set

$$\mathcal{M}_{r,t,v}\coloneqq\{(\boldsymbol{w},\widetilde{\boldsymbol{w}})\in\mathcal{W}_k^r(i)\times\mathcal{W}_k^r(i'):\ \boldsymbol{w}|\widetilde{\boldsymbol{w}}\text{ overlapping},\ |[\boldsymbol{w}|\widetilde{\boldsymbol{w}}]|=t,\ |[\![\boldsymbol{w}|\widetilde{\boldsymbol{w}}]\!]|=v,\ |[\boldsymbol{w}]\cap[\widetilde{\boldsymbol{w}}]|>0\}. \tag{.39}$$

The last condition of (.39) filters out walk pairs $(\boldsymbol{w},\widetilde{\boldsymbol{w}})$ which factorize as $\varrho(\boldsymbol{w},\widetilde{\boldsymbol{w}})=\varrho(\boldsymbol{w})\varrho(\widetilde{\boldsymbol{w}})$. Similarly, if $\boldsymbol{w}|\widetilde{\boldsymbol{w}}$ is not overlapping $\varrho(\boldsymbol{w},\widetilde{\boldsymbol{w}})=0$ and, consequently, $\varrho(\boldsymbol{w})\varrho(\widetilde{\boldsymbol{w}})=0$.

Let's start with the case $i=i'$. By the set construction in (.39), $\mathcal{M}_{r,t,v}(i,i)\subseteq\mathcal{N}_{2r,t,v}(i)$. So $|\mathcal{M}_{r,t,v}(i,i)|\leq|\mathcal{N}_{2r,t,v}(i)|$ and by the counting result [VA24b, Lemma 13]

$$|\mathcal{M}_{r,t,v}(i,i)|\leq(v-1)^{2rk}\binom{n-1}{v-1}. \tag{.40}$$

A similar argument can be made when $i$ and $i'$ are distinct. By fixing $i$ and $i'$, we are left selecting $\binom{n-2}{v-2}$ unique vertices with a walk selection factor of $(v-1)^{2rk}$. Altogether,

$$|\mathcal{M}_{r,t,v}(i,i')|\leq(v-1)^{2rk}\binom{n-2}{v-2}. \tag{.41}$$

For bounds on $v$ and $t$, we note that $\boldsymbol{u} \coloneqq \boldsymbol{w}|\widetilde{\boldsymbol{w}}$ is an overlapping walk sequence, which by the partition result [VA24b, Lemma 12], means it must have, at most, $|[\boldsymbol{u}]| \leq 2rk - r$ unique edges. Similarly, the number of unique vertices bounds as $|[\![\boldsymbol{u}]\!]| \leq |[\boldsymbol{u}]| + 1$ since the discrete graph $([\![\boldsymbol{u}]\!], [\boldsymbol{u}])$ associated with $\boldsymbol{u}$ is necessarily connected by the rooted nature of the walks in the sequence $\boldsymbol{u}$ (walks must start at $i$ or $i'$) and the last condition of (.39).

Next, we consider the bound $|\varrho(\boldsymbol{w}, \widetilde{\boldsymbol{w}})| \leq 2\max\{|\varrho(\boldsymbol{w}, \widetilde{\boldsymbol{w}})|, |\varrho(\boldsymbol{w})\varrho(\widetilde{\boldsymbol{w}})|\}$. Introduce the notation, $\varrho_1(\boldsymbol{w}) = \mathbb{E}\big[\prod_{s=1}^{r}(A_{\boldsymbol{w}^s} - \mathbb{E}[A_{\boldsymbol{w}^s}])\big]$ and $\varrho_2(\boldsymbol{w}) = \mathbb{E}\big[\prod_{s=1}^{r} x_{\mathfrak{p}(\boldsymbol{w}^s)}\big]$. We analogously define, $\varrho_1(\boldsymbol{w}, \widetilde{\boldsymbol{w}}) \coloneqq \varrho_1(\boldsymbol{w}|\widetilde{\boldsymbol{w}})$ and $\varrho_2(\boldsymbol{w}, \widetilde{\boldsymbol{w}}) \coloneqq \varrho_2(\boldsymbol{w}|\widetilde{\boldsymbol{w}})$. From [VA24b, Lemma 10],

$$|\varrho_1(\boldsymbol{w})\varrho_1(\widetilde{\boldsymbol{w}})| \leq 2^{2r}(\nu_n/n)^{|[\boldsymbol{w}]|+|[\widetilde{\boldsymbol{w}}]|} \leq 2^{2r}(\nu_n/n)^{|[\boldsymbol{w}|\widetilde{\boldsymbol{w}}]|} \quad \text{and} \quad |\varrho_1(\boldsymbol{w}, \widetilde{\boldsymbol{w}})| \leq 2^{2r}(\nu_n/n)^{|[\boldsymbol{w}|\widetilde{\boldsymbol{w}}]|}$$

and

$$|\varrho_2(\boldsymbol{w})\varrho_2(\widetilde{\boldsymbol{w}})| \leq (2\sqrt{r}\kappa_0)^{2r} \quad \text{and} \quad |\varrho_2(\boldsymbol{w}, \widetilde{\boldsymbol{w}})| \leq (2\sqrt{r}\kappa_0)^{2r}$$

where $\kappa_0$ is defined as in Proposition 10. Let $t_* = r(2k-1)$ then

$$\mathrm{Cov}\big(\Delta^r_{i,\theta}, \Delta^r_{\theta,i'}\big) \leq \frac{1}{\nu_n^{r(2k-1)}} \sum_{t=1}^{t_*} \sum_{v=2}^{t+1} (4\sqrt{r}\kappa_0)^{2r} \cdot |\mathcal{M}_{r,t,v}(i,i')|(\nu_n/n)^t. \tag{.42}$$

For the case $i = i'$, cardinality and $|\mathcal{M}_{r,t,v}(i,i)| \lesssim n^{v-1} \leq n^t$ by (.40).

$$\begin{aligned}\mathrm{Cov}\big(\Delta^r_{i,\theta}, \Delta^r_{\theta,i'}\big) &\lesssim \frac{1}{\nu_n^{r(2k-1)}} \sum_{t=1}^{t_*} \sum_{v=2}^{t+1} (4\sqrt{r}\kappa_0)^{2r} \cdot \nu_n^t \\ &\lesssim \frac{\nu_n^{t_*}}{\nu_n^{r(2k-1)}},\end{aligned}$$

where the last line follows from the fact $r$ and $k$ are fixed relative to $n$. Similarly for the off-diagonal case of $i \neq i'$, $|\mathcal{M}_{r,t,v}(i,i')| \lesssim n^{v-2} \leq n^{t-1}$ by (.41) and

$$\frac{1}{\nu_n^{r(2k-1)}} \sum_{t=1}^{t_*} \sum_{v=2}^{t+1} (4\sqrt{r}\kappa_0)^{2r} \cdot |\mathcal{M}_{r,t,v}(i,i')|(\nu_n/n)^t \lesssim \frac{1}{\nu_n^{r(2k-1)}} \cdot \frac{\nu_n^{t_*}}{n}.$$

Noting $t_* = r(2k-1)$, this proofs the claim that $\mathrm{Cov}\big(\Delta^r_{i,\theta}, \Delta^r_{\theta,i'}\big) \lesssim n^{-1\{i \neq i'\}}$.

### F.2 Proof of Lemma 37

Shown in [VA24b] the dominant term in a walk-based for $\mathring{\Delta}_{i,\theta}$ is given by the proxy term

$$\widetilde{T}_i^{\mathrm{hi}}(r) = (r-1)!! \sum_{(j_\ell)_\ell \in \mathcal{P}^{r/2}_{[n]\setminus\{i\}}} \prod_{q=1}^{r/2} p_{ij_\ell}(1-p_{ij_\ell})(e_{j_\ell}^T \mathbb{E}[A]^{k-1}\mathbb{E}[X]\theta)^2$$

where $\mathcal{P}^{r/2}_{[n]\setminus\{i\}}$ is the set of coordinate distinct $(r/2)$-tuples on $[n]\setminus\{i\}$. Specifically, it was shown for $r \in 2\mathbb{N}$ and $\nu_n$ sufficiently large

$$\left|\mathbb{E}[\Delta_{i,\theta}^r] - \nu_n^{-(rk-r/2)}\, \widetilde{T}_i^{\mathrm{hi}}(r)\right| \lesssim n^{-1} + \nu_n^{-\epsilon}$$

where $\epsilon$ can be used to parameterize the separation of higher- and lower-order terms $\mathring{\Delta}_{i,\theta}$ [VA24b, Lemma 14 and Lemma 18].

To obtain the limiting closed form, we utilize $|[n]^{r/2} \setminus \mathcal{P}^{r/2}_{[n]\setminus\{i\}}| \le C(r) n^{r/2-1}$ and

$$\begin{aligned}\sum_{(j_\ell)_\ell \in [n]^{r/2}} \prod_{q=1}^{r/2} p_{ij_\ell}(1-p_{ij_\ell})(e_{j_\ell}^T \mathbb{E}[A]^{k-1}\mathbb{E}[X]\theta)^2 &= \Big(\sum_{j\in[n]} p_{ij}(1-p_{ij})(e_j^T \mathbb{E}[A]^{k-1}\mathbb{E}[X]\theta)^2\Big)^{r/2} \\ &= \big((\mathbb{E}[A]^{k-1}\mathbb{E}[X]\theta)^T(\nu_n V_i^2)(\mathbb{E}[A]^{k-1}\mathbb{E}[X]\theta)\big)^{r/2}.\end{aligned}$$

For brevity, let $f_i(j) \coloneqq (p_{ij}/\nu_n)(1-p_{ij})(e_j^T \mathbb{E}[A/\nu_n]^{k-1}\mathbb{E}[X]\theta)^2$. Then, noting $\mathcal{P}^{r/2}_{[n]\setminus\{i\}} \subseteq [n]^{r/2}$,

$$\begin{aligned}&|\nu_n^{-(rk-r/2)}\, \widetilde{T}_i^{\mathrm{hi}}(r) - (r-1)!!\, \|V_i \mathbb{E}[A]^{k-1}\mathbb{E}[X]\theta\|_2^r| \\ &\qquad = (r-1)!!\Big|\sum_{(j_\ell)_\ell \in \mathcal{P}^{r/2}_{[n]\setminus\{i\}}} \prod_{q=1}^{r/2} f_i(j_\ell) - \sum_{(j_\ell)_\ell \in [n]^{r/2}} \prod_{q=1}^{r/2} f_i(j_\ell)\Big| \\ &\qquad \le (r-1)!!\, |[n]^{r/2} \setminus \mathcal{P}^{r/2}_{[n]\setminus\{i\}}|\, (\max_{j\in[n]} f_i(j))^{r/2}.\end{aligned}$$

Let $\mathcal{W}_{k-1}(j)$ be the set of $k-1$ walks on $[n]$ starting at $j$. Then, with $\mathcal{W}^2_{k-1}(j) \coloneqq \mathcal{W}_{k-1}(j) \times \mathcal{W}_{k-1}(j)$

$$f_i(j) = (p_{ij}/\nu_n)(1-p_{ij}) \sum_{\boldsymbol{w}\in\mathcal{W}^2_{k-1}(j)} \prod_{s=1}^{2} \Big(\mathbb{E}[(X\theta)_{\mathfrak{p}(\boldsymbol{w}^s)}] \prod_{\ell=1}^{k-1} (p_{(\boldsymbol{w}^s)_\ell}/\nu_n)\Big)$$

Recall that $\mathbb{E}|(X\theta)_i| < x_*$ by assumption. Since $|\mathcal{W}_{k-1}(j)| \le |[n]^{k-1}| = n^{k-1}$ and $p_{ij}/\nu_n \le 1/n$ we have

$$f_i(j) \le x_*^2/n \qquad \text{for every } i, j \in [n].$$

Altogether, this yields the inequality

$$|\nu_n^{-(rk-r/2)}\widetilde{T}_i^{\mathrm{hi}}(r) - (r-1)!!\,\|V_i\mathbb{E}[A]^{k-1}\mathbb{E}[X]\theta\|_2^r| \le C(r)x_*^r n^{-1}$$

that, when pieced with a triangle inequality, produces the desired bound.

# G Simulation Details for Figures

This appendix details the experimental setups for the figures presented in the main text and introduces two supplementary figures for further illustration.

1. **Figure .4 (Erdős–Rényi Graph Behavior):** This figure illustrates the convergence to normality on an Erdős–Rényi graph. It demonstrates that the Central Limit Theorem (CLT) becomes evident even for a modest number of nodes (e.g., $n = 300$) once the average degree $\nu_n$ is moderately large (e.g., $\nu_n = 16$). While larger $n$ values would increase histogram resolution, the characteristic normal shape is already apparent at $n = 300$.

2. **Figure .5 (2-Class CSBM Validation):** This figure showcases the strong agreement between the theoretically predicted distribution and the empirical distribution of features on a 2-class Contextual Stochastic Block Model (CSBM).

Specific parameters for these and all other figures are provided in the subsequent subsections.

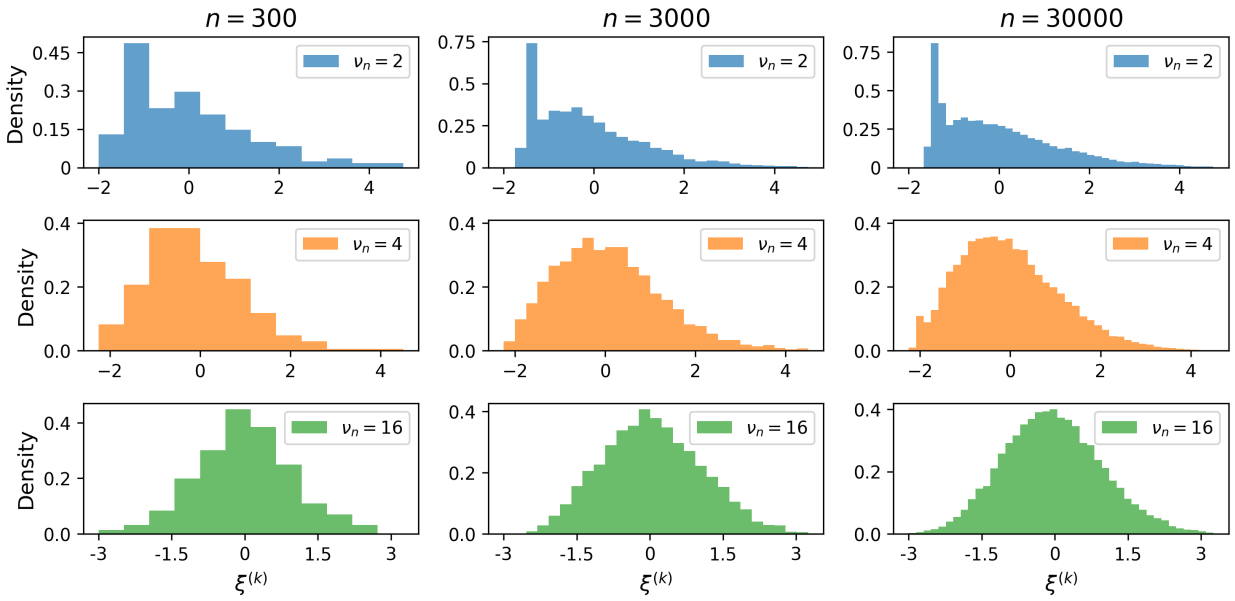



Figure .4: Comparison of $\xi^{(k)}$ distribution for $k = 3$ and a fixed, expected degree Erdős–Réyni graph. As graph size increases, the overall histogram resolution may be increased but this does not qualitatively change the shape of the histogram. That is, growing degree $\nu_n \to \infty$, is a necessary condition for $\xi^{(k)}$ to be Gaussian.

### G.1 Details for Figure 6.1

The plots in Figure 6.1 were generated using a 3-class CSBM with $n = 8192$ nodes. Class proportions were $\pi_1 = 0.25, \pi_2 = 0.45, \pi_3 = 0.30$, average degree parameter was $\nu_n = \sqrt{8192}$, and the inter-community probability scaling matrix was $B = (\nu_n/n) \cdot \begin{pmatrix} 0.4 & 1 & 1 \\ 1 & 0.4 & 1 \\ 1 & 1 & 0.4 \end{pmatrix}$. Initial features $X_i$ where $d = 2$ dimensional and generated as $X_i \sim N(M_{z_i,*}, \sigma^2 I_2)$ with $\sigma^2 = 0.25$ and $M_{1,*} = [2, 2]^T$, $M_{2,*} = [-1, -3]^T$, and $M_{3,*} = [-1, 0]^T$.

Cross entropy training was run for a single linear classifier layer for 10 epochs with learning rate 10 on the SGD optimization. Although small differences are expected at later time steps, Figure 6.1 still shows good agreement between the empirical and theoretical gradient average.

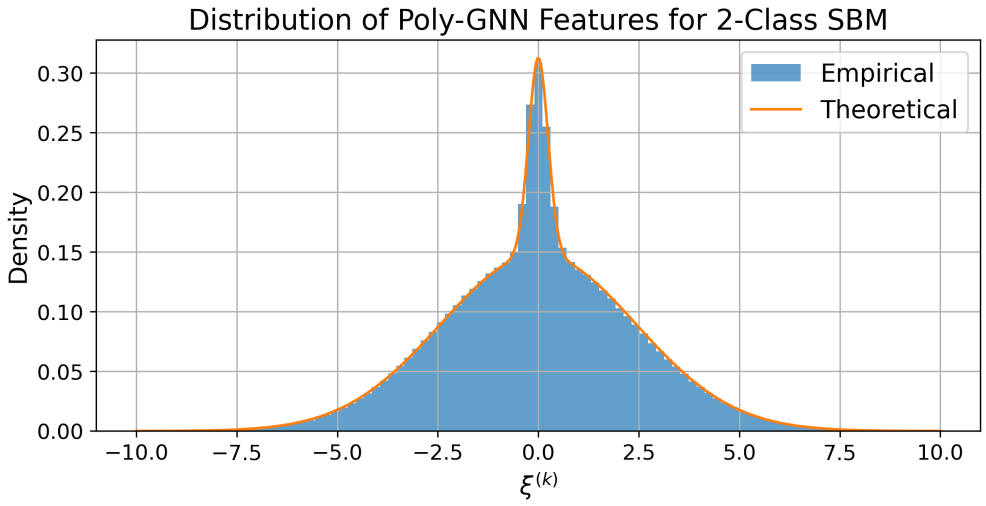



Figure .5: Empirical distribution of a two-class CSBM with exaggerated class proportions and edge probabilities. Both mixture components are centered at zero with a visible difference between the peak widths and heights of each component.

### G.2 Details for Figure 6.2

The plots in Figure 6.2 were generated using a 2-class SBM with $n = 32000$ nodes. Class proportions were $\pi_1 = 0.4, \pi_2 = 0.6$, average degree parameter was $\nu_n = 30$, and the inter-community probability scaling matrix was $B = (\nu_n/n) \cdot \begin{pmatrix} 0.5 & 1 \\ 1 & 0.5 \end{pmatrix}$. Initial features $X_i$ were $d = 2$ dimensional drawn from mean vectors $M_{1,*} = [2, 2]^T$ and $M_{2,*} = [-1, -2]^T$. Quadratic discriminant analysis was performed using the sample statistics of $\overline{\phi}_i^{(k)}$ with $k = 2$. Cross-entropy training consisted of single linear layer trained for 5000 epochs at learning rate 0.5 with a SGD optimizer

### G.3 Details for Figure 6.3

The plots in Figure 6.3 were generated in the same setting as Section G.2 with the exception of a higher average degree $\nu_n = 35$. The plots show Kernel Density Estimates (KDEs) of the $\overline{\phi}_i^{(k)}$ features for $k \in \{2, 4, 6\}$. The KDEs were computed using Gaussian kernels with

bandwidth selected by Scott's rule.

### G.4 Details for Figure .4

The plots in Figure .4 were simulated from a 1-class SBM, commonly referred to as an Erdős–Réyni graph, with probability parameter $p = \nu_n/n$. Depth $k = 3$ was used with unit, univariate features $X_i = 1$ for all $i \in [n]$. A grid search was performed on graph sizes $n \in \{300, 3000, 30000\}$ with expected degrees $\nu_n \in \{2, 4, 16\}$. These graph are very sparse, yet they approach Gaussianity fairly quickly. Particularly, the plot associated with $\nu_n = 16$ has nearly symmetrical tails and a bell curve shape.

### G.5 Details for Figure .5

The plot of Figure .5 was generated from a 2-class SBM with 32000 nodes. Class proportions were $\pi_1 = 0.9, \pi_2 = 0.1$, average degree parameter was $\nu_n = \sqrt{32000}$, and the inter-community probability scaling matrix was $B = (\nu_n/n) \cdot \begin{pmatrix} 10 & 0.1 \\ 0.1 & 10 \end{pmatrix}$. Initial features $X_i$ were $d = 1$ dimensional and generated as $X_i \sim N(M_{z_i}, \sigma^2)$ for $M_1 = 10^{-2}$, $M_2 = -10^{-2}$ and $\sigma^2 = 10^{-4}$.

For the plot of Figure .5 we simulate 100 CSBM graphs each at 32000 nodes. From these 100 replicates, we obtain an estimate for $\mathbb{E}[\xi^{(k)}]$ with $k = 3$. The final figure is a 100 bin histogram of the 3200000 empirical elements with a theoretical density given by our theory drawn on top.

REFERENCES


[ABW10] Kambiz A. Asher, Neal K. Bangerter, Ronald D. Watkins, and Garry E. Gold. "Radiofrequency Coils for Musculoskeletal Magnetic Resonance Imaging." *Topics in Magnetic Resonance Imaging*, **21**(5), 2010.

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Springer Science & Business Media, 2nd edition, 2008.

[and58] Walter D. Fisher and. "On Grouping for Maximum Homogeneity." *Journal of the American Statistical Association*, **53**(284):789–798, 1958.

[AV07] David Arthur and Sergei Vassilvitskii. "k-means++: the advantages of careful seeding." In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, p. 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics.

[Bal12] Pierre Baldi. "Autoencoders, Unsupervised Learning, and Deep Architectures." In Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver, editors, *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pp. 37–49, Bellevue, Washington, USA, 02 Jul 2012. PMLR.

[Bar89] H.B. Barlow. "Unsupervised Learning." *Neural Computation*, **1**(3):295–311, 09 1989.

[BFJ21] Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. "Graph Convolution for Semi-Supervised Classification: Improved Linear Separability and Out-of-Distribution Generalization." In *International Conference on Machine Learning*, 2021.

[BFJ23] Aseem Baranwal, Kimon Fountoulakis, and Aukosh Jagannath. "Optimality of Message-Passing Architectures for Sparse Graphs." In *Advances in Neural Information Processing Systems*, 2023.

[BG] Erhan Bayraktar and Gaoyue Guo. "Strong Equivalence between Metrics of Wasserstein Type." **26**:1–13.

[BH16] Afonso S. Bandeira and Ramon van Handel. "Sharp nonasymptotic bounds on the norm of random matrices with independent entries." *The Annals of Probability*, **44**(4):2479 – 2506, 2016.

[Bha97] Rajendra Bhatia. *Perturbation of Matrix Functions*, pp. 289–323. Springer New York, New York, NY, 1997.

[BKM] Vladimir Braverman, Aditya Krishnan, and Christopher Musco. "Sublinear Time Spectral Density Estimation." In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1144–1157.

[BMC06] Boubakeur Belaroussi, Julien Milles, Sabin Carme, Yue Min Zhu, and Hugues Benoit-Cattin. "Intensity non-uniformity correction in MRI: Existing methods and their validation." *Medical Image Analysis*, **10**(2):234–246, 2006.

[BNS06] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples." *Journal of Machine Learning Research*, **7**(85):2399–2434, 2006.

[BR16] Sohail Bahmani and Justin Romberg. "Near-optimal estimation of simultaneously sparse and low-rank matrices from nested linear measurements." *Information and Inference: A Journal of the IMA*, **5**(3):331–351, 05 2016.

[Bra02] Andrea Braides. *Gamma-Convergence for Beginners*. Oxford University Press, 07 2002.

[BRT09] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector." *The Annals of statistics*, **37**(4):1705–1732, 2009.

[CDF98] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Seán Slattery. "Learning to Extract Symbolic Knowledge from the World Wide Web." AAAI '98/IAAI '98, p. 509–516, USA, 1998.

[CKK97] Chris A. Cocosco, Vasken Kollokian, Remi K.-S. Kwan, G. Bruce Pike, and Alan C. Evans. "BrainWeb: Online Interface to a 3D MRI Simulated Brain Database." *NeuroImage*, **5**:425, 1997.

[CM22] Sudhanshu Chanpuriya and Cameron N Musco. "Simplified Graph Convolution with Heterophily." In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[CPL21] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. "Adaptive Universal Generalized PageRank Graph Neural Network." In *International Conference on Learning Representations*, 2021.

[CSD25] Juntong Chen, Johannes Schmidt-Hieber, Claire Donnat, and Olga Klopp. "Understanding the Effect of GCN Convolutions in Regression Tasks.", 2025.

[CSP09] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. "Sparse and low-rank matrix decompositions." In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 962–967, 2009.

[DBV16] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. "Convolutional neural networks on graphs with fast localized spectral filtering." In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, p. 3844–3852, Red Hook, NY, USA, 2016. Curran Associates Inc.

[DJN20] Nicola K Dinsdale, Mark Jenkinson, and Ana I L Namburete. "Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal." *Neuroimage*, **228**:117689, 12 2020.

[DK13] David Donoho and Gitta Kutyniok. "Microlocal Analysis of the Geometric Separation Problem." *Communications on Pure and Applied Mathematics*, **66**(1):1–47, 2013.

[DKB13] P. Dvořák, W. G. Kropatsch, and K. Bartušek. "Automatic Brain Tumor Detection in T2-weighted Magnetic Resonance Images." *Measurement Science Review*, **13**(5):223–230, 10 2013.

[DP14] Fangfang Dong and Jialin Peng. "Brain MR image segmentation based on local Gaussian mixture model and nonlocal spatial regularization." *Journal of Visual Communication and Image Representation*, **25**(5):827–839, 2014.

[DSM18] Yash Deshpande, Subhabrata Sen, Andrea Montanari, and Elchanan Mossel. "Contextual Stochastic Block Models." In *Advances in Neural Information Processing Systems*, 2018.

[EKY13] László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin. "Spectral statistics of Erdős–Rényi graphs I: Local semicircle law." *The Annals of Probability*, **41**(3B):2279 – 2375, 2013.

[FHT07] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics*, **9**(3):432–441, 12 2007.

[GBT21] Alden Green, Sivaraman Balakrishnan, and Ryan Tibshirani. "Minimax Optimal Regression over Sobolev Spaces via Laplacian Regularization on Neighborhood Graphs." In *International Conference on Artificial Intelligence and Statistics*, 2021.

[GBT23] Alden Green, Sivaraman Balakrishnan, and Ryan J Tibshirani. "Minimax optimal regression over Sobolev spaces via Laplacian Eigenmaps on neighbourhood graphs." *Information and Inference*, **12**(3):2423–2502, 2023.

[GSR17] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. "Neural Message Passing for Quantum Chemistry." In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW,*

*Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, 2017.

[HLL83] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. "Stochastic blockmodels: First steps." *Social Networks*, **5**(2):109–137, 1983.

[HMT11] N. Halko, P. G. Martinsson, and J. A. Tropp. "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions." *SIAM Review*, **53**(2):217–288, 2011.

[HP99] Aapo Hyvärinen and Petteri Pajunen. "Nonlinear independent component analysis: Existence and uniqueness results." *Neural Networks*, **12**(3):429–439, 1999.

[HST19] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. "Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning." In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 16–18 Apr 2019.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Unsupervised Learning*, pp. 485–585. Springer New York, New York, NY, 2009.

[IS15] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015.

[JYJ20] Woojin Jung, Jaeyeon Yoon, Sooyeon Ji, Joon Yul Choi, Jae Myung Kim, Yoonho Nam, Eung Yeop Kim, and Jongho Lee. "Exploring linearity of deep neural network trained QSM: QSMnet+." *NeuroImage*, **211**:116619, 2020.

[JZB16] Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, and Michael I Jordan. "Local Maxima in the Likelihood of Gaussian Mixture Models: Structural Results and Algorithmic Consequences." In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[KB15] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[KBG19] Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. "Predict then Propagate: Graph Neural Networks meet Personalized PageRank." In *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

[KFW15] Zheng Tracy Ke, Jianqing Fan, and Yichao Wu. "Homogeneity Pursuit." *Journal of the American Statistical Association*, **110**(509):175–194, 2015.

[KKB09] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. "$\ell_1$ Trend Filtering." *SIAM Review*, **51**(2):339–360, 2009.

[KKM20] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. "Variational Autoencoders and Nonlinear ICA: A Unifying Framework." In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 26–28 Aug 2020.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*, 2012.

[KT14] Yunho Kim and Hemant D. Tagare. "Intensity Nonuniformity Correction for Brain MR Images with Known Voxel Classes." *SIAM Journal on Imaging Sciences*, **7**(1):528–557, 2014.

[KW17] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks." In *International Conference on Learning Representations*, 2017.

[Lan95] Ken Lang. "Newsweeder: Learning to filter netnews." In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 331–339, 1995.

[LBB08] Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. "Consistency of spectral clustering." *The Annals of Statistics*, **36**(2):555 – 586, 2008.

[LCI13] Bradley Lowekamp, David Chen, Luis Ibanez, and Daniel Blezek. "The Design of SimpleITK." *Frontiers in Neuroinformatics*, **7**, 2013.

[LCN19] Xiao Liang, Liyuan Chen, Dan Nguyen, Zhiguo Zhou, Xuejun Gu, Ming Yang, Jing Wang, and Steve Jiang. "Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy." *Physics in Medicine & Biology*, **64**(12):125002, 6 2019.

[Lem03] Jüri Lember. "On minimizing sequences for k-centres." *Journal of Approximation Theory*, **120**(1):20–35, 2003.

[LHL22] Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and optdoina Precup. "Revisiting Heterophily For Graph Neural Networks." In *Advances in Neural Information Processing Systems*, volume 35, pp. 1362–1375, 2022.

[LLZ19] Tianxi Li, Elizaveta Levina, and Ji Zhu. "Prediction models for network-linked data." *The Annals of Applied Statistics*, **13**(1):132 – 164, 2019.

[MHS18] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. "UMAP: Uniform Manifold Approximation and Projection." *Journal of Open Source Software*, **3**(29):861, 2018.

[Nes12] Yu. Nesterov. "Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems." *SIAM Journal on Optimization*, **22**(2):341–362, 2012.

[Pri23] Eugene Prilepin. "csaps." `https://github.com/espdev/csaps`, 2023.

[PWC20] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. "Geom-GCN: Geometric Graph Convolutional Networks." In *International Conference on Learning Representations*, 2020.

[RAS21] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. "Multi-Scale attributed node embedding." *Journal of Complex Networks*, **9**(2):cnab014, 05 2021.

[RBM23] T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. "A Survey on Oversmoothing in Graph Neural Networks.", 3 2023. arXiv:2303.10993.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing.

[See02] Matthias Seeger. "Learning With Labeled and Unlabeled Data." Technical report, Institute for Adaptive and Neural Computation, University of Edinburgh, 2002.

[SGT09] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. "The Graph Neural Network Model." *IEEE Transactions on Neural Networks*, **20**(1):61–80, 2009.

[SHK14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research*, **15**(56):1929–1958, 2014.

[SHS01] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. "A Generalized Representer Theorem." In *Computational Learning Theory*, pp. 416–426, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[SLY20] Ruoyu Sun, Zhi-Quan Luo, and Yinyu Ye. "On the Efficiency of Random Permutation for ADMM and Coordinate Descent." *Mathematics of Operations Research*, **45**(1):233–271, 2020.

[SNB08] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. "Collective Classification in Network Data." *AI Magazine*, **29**(3):93, Sep. 2008.

[TAC10] Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, Yuanjie Zheng, Alexander Egan, Paul A. Yushkevich, and James C. Gee. "N4ITK: Improved N3 Bias Correction." *IEEE Transactions on Medical Imaging*, **29**(6):1310–1320, 2010.

[Tib14] Ryan J. Tibshirani. "Adaptive piecewise polynomial estimation via trend filtering." *The Annals of Statistics*, **42**(1):285 – 323, 2014.

[Tsy09] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[TTJ15] Matthew Thorpe, Florian Theil, Adam M. Johansen, and Neil Cade. "Convergence of the $k$-Means Minimization Problem using $\Gamma$-Convergence." *SIAM Journal on Applied Mathematics*, **75**(6):2444–2474, 2015.

[TV23] Jared Tanner and Simon Vary. "Compressed sensing of low-rank plus sparse matrices." *Applied and Computational Harmonic Analysis*, **64**:254–293, 2023.

[VA24a] Luciano Vinas and Arash A. Amini. "Simple GNNs with Low Rank Non-parametric Aggregators." In *The Third Learning on Graphs Conference*, 2024.

[VA24b] Luciano Vinas and Arash A. Amini. "Supplementary Material for: Sharp Bounds for Poly-GNNs and the Effect of Graph Noise.", 8 2024.

[VCC18] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. "Graph Attention Networks." In *International Conference on Learning Representations*, 2018.

[Ver18] Roman Vershynin. *Concentration of Sums of Independent Random Variables*, p. 11–37. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[VGN20] Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. "Sub-Weibull distributions: Generalizing sub-Gaussian and sub-Exponential properties to heavier tailed distributions." *Stat*, **9**(1):e318, 2020.

[VPL07] Uro Vovk, Franjo Pernus, and Botjan Likar. "A Review of Methods for Correction of Intensity Inhomogeneity in MRI." *IEEE Transactions on Medical Imaging*, **26**(3):405–421, 2007.

[VS22] Luciano Vinas and Atchar Sudyadhom. "Sinusoidal Sensitivity Calculation for Line Segment Geometries." *arXiv:2208.03059*, 2022.

[VSD21] Luciano Vinas, Jessica Scholey, Martina Descovich, Vasant Kearney, and Atchar Sudhyadhom. "Improved contrast and noise of megavoltage computed tomography (MVCT) through cycle-consistent generative machine learning." *Medical Physics*, **48**(2):676–690, 2021.

[Wai09] Martin J. Wainwright. "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using $\ell_1$-Constrained Quadratic Programming (Lasso)." *IEEE Transactions on Information Theory*, **55**(5):2183–2202, 2009.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[WCW23] Xinyi Wu, Zhengdao Chen, William Wei Wang, and Ali Jadbabaie. "A Non-Asymptotic Analysis of Oversmoothing in Graph Neural Networks." In *International Conference on Learning Representations*, 2023.

[WGK96] W.M. Wells, W.E.L. Grimson, R. Kikinis, and F.A. Jolesz. "Adaptive segmentation of MRI data." *IEEE Transactions on Medical Imaging*, **15**(4):429–442, 1996.

[WL10] Tong Tong Wu and Kenneth Lange. "The MM Alternative to EM." *Statistical Science*, **25**(4):492–505, 2010.

[WYJ22] Rongzhe Wei, Haoteng Yin, Junteng Jia, Austin R. Benson, and Pan Li. "Understanding Non-linearity in Graph Neural Networks from the Bayesian-Inference Perspective." In *Advances in Neural Information Processing Systems*, 2022.

[WZ22] Xiyuan Wang and Muhan Zhang. "How Powerful are Spectral Graph Neural Networks." In *International Conference on Machine Learning*, 2022.

[YCS16] Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. "Revisiting semi-supervised learning with graph embeddings." In *International Conference on Machine Learning*, 2016.

[YS07] Stephen J. Young and Edward R. Scheinerman. "Random Dot Product Graph Models for Social Networks." In *Algorithms and Models for the Web-Graph*, pp. 138–149, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[YWB19] Xin Yi, Ekta Walia, and Paul Babyn. "Generative adversarial network in medical imaging: A review." *Medical Image Analysis*, **58**:101552, 2019.

[ZGL03] Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions." In *International Conference on Machine Learning*, 2003.

[ZK21] Hao Zhu and Piotr Koniusz. "Simple Spectral Graph Convolution." In *International Conference on Learning Representations*, 2021.

[ZM22] Yidong Zhou and Hans-Georg Müller. "Network Regression with Graph Laplacians." *Journal of Machine Learning Research*, **23**(320):1–41, 2022.

[ZNZ22] Yujia Zheng, Ignavier Ng, and Kun Zhang. "On the Identifiability of Nonlinear ICA: Sparsity and Beyond." In *Advances in Neural Information Processing Systems*, 2022.

[ZPI17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks." In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.