

Abstract

Density, Function, and Parameter Estimation with High-Dimensional Data

Jason M. Klusowski

2018

This thesis seeks to describe the tradeoff between computational complexity and statistical estimation in a variety of high-dimensional settings (where the ambient dimension d is large or even possibly much larger than the available sample size n). Specifically, we focus on four representative problems that broadly fall under the guise of high-dimensional statistical modeling: (1) nonparametric, nonlinear regression, (2) parametric mixture models, (3) nonparametric density models, and (4) statistical network analysis.

1. In nonparametric, nonlinear regression setting, we impose conditions on a high-dimensional, multivariate regression function so that small predictive mean squared error can be achieved when $d \gg n$. Due to the non-convexity of the loss or likelihood-based surfaces, provably good, computationally feasible algorithms are also needed to overcome the associated (and challenging) optimization tasks. A complementary task to investigate what is computationally or theoretically achievable is to analyze the fundamental limits of statistical inference in terms of minimax rates, which are also investigated here.
2. The Expectation-Maximization (EM) algorithm is a widely used technique for parameter estimation. When the log-likelihood is not concave, it is well known that EM can converge to a non-global optimum. However, recent work has side-stepped the question of whether EM reaches the likelihood maximizer, instead by directly working out statistical guarantees on its loss. For a large enough sample size, the difference between the sample EM operator M and the population EM operator M_n can be bounded such that the empirical EM estimate approaches the true parameter with high probability. These explorations have identified regions of initialization for

which the empirical EM iterates $\theta^{t+1} \leftarrow M_n(\theta^t)$ approaches the true parameter in probability. Modern literature has focused on a few specific toy models that showcase this approach. We focus on a representative problem – the symmetric mixture of two regressions model $Y = R(\theta^* \cdot X) + \varepsilon$, where R is a Rademacher random variable, X is a d -dimensional Gaussian covariate, and ε is a univariate Gaussian error. In [1], it was shown that if the EM algorithm is initialized in a ball around θ^* with radius proportional $\|\theta^*\|$, the EM algorithm for the mixture of two regressions converges with high probability. We relax these conditions and show that as long as the cosine angle between θ^* and the initializer θ^0 is not too small (regardless of the size of $\|\theta^0\|$), the EM algorithm also converges. Furthermore, we also show that the population EM operator is not globally contractive for some initializers satisfying $\theta^0 \cdot \theta^* > 0$. In contrast, it is known that the population EM operator for a symmetric mixture of two Gaussians is globally contractive [2], provided $\theta^0 \cdot \theta^* > 0$.

3. A popular class of problem in statistics deals with estimating the support of a density from n observations drawn at random from a d -dimensional distribution. The one-dimensional case reduces to estimating the end points of a univariate density. In practice, an experimenter may only have access to a noisy version of the original data. Therefore, a more realistic model allows for the observations to be contaminated with additive noise.

We consider estimation of convex bodies when the additive noise is distributed according to a multivariate Gaussian distribution, even though our techniques could easily be adapted to other noise distributions. Unlike standard methods in deconvolution that are implemented by thresholding a kernel density estimate, our method avoids tuning parameters and Fourier transforms altogether. We show that our estimator, computable in $(O(\ln n))^{(d-1)/2}$ time, converges at a rate of $O_d(\log \log n / \sqrt{\log n})$ in Hausdorff distance, in accordance with the polylogarithmic rates encountered in Gaussian deconvolution problems. Part of our analysis also involves the optimality of the proposed estimator. We provide a lower bound for the minimax rate of estimation in Hausdorff distance that is $\Omega_d(1/\log^2 n)$.

4. Counting the number of features in a graph – ranging from basic local structures like motifs or graphlets (e.g., edges, triangles, cycles, cliques), or other more global features like the number of connected components – is an important statistical and computational problem. For instance, applied researchers seek to capture from such features the interactions and relationships between groups and individuals. In doing so, they typically collect data from a random sample of nodes in order to infer global properties of the parent population network from the sampled version. This setting is largely due to cost and time constraints (e.g., in-person interviews that are in remote locations) or an inability to gain access the full population (e.g., historical data). We consider two graph sampling models. The first is based on the subgraph sampling model, where we sample each vertex independently with probability p and observe the subgraph induced by these sampled vertices. The second is based on the neighborhood sampling model, where we sample each vertex independently with probability p , and additionally observe the edges between the sampled vertices and their neighbors. We obtain optimal sample complexity bounds for several classes of graphs (i.e. bounded degree, chordal, and planar). The methodology relies on topological identities of graph homomorphism numbers. They, in turn, also play a key role in proving minimax lower bounds based on construction of random instances of graphs with matching structures of small subgraphs.

Density, Function, and Parameter Estimation with High-Dimensional Data

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Jason M. Klusowski

Dissertation Director: Andrew R. Barron

May, 2018

Copyright © 2018 by Jason M. Klusowski
All rights reserved.

Dedication

This thesis is dedicated to my wife, Joowon Kim. Her love, patience, and support throughout the years have enabled me to achieve what I alone could only dream of.

Contents

Dedication	iii
Acknowledgements	xi
1 Introduction	1
2 Risk bounds for high-dimensional ridge function combinations including neural networks	8
2.1 Introduction	8
2.2 How far from optimal?	17
2.3 Computational aspects	19
2.4 Greedy algorithm	20
2.5 Risk bounds	23
2.5.1 Penalized estimators over the entire parameter space	23
2.6 Risk bounds in high dimensions	35
2.6.1 Penalty under Assumption 1	35
2.6.2 Penalty under Assumption 2	36
2.7 Risk bounds with improved exponents for moderate dimensions	37
2.7.1 Penalized estimators over a discretization of the parameter space . .	38
2.8 Proofs of the lemmata	39
3 Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls	52
3.1 Introduction	52

3.2	L^∞ approximation with bounded ℓ^1 norm	55
3.2.1	Positive results	55
3.2.2	Lower bounds	65
3.3	L^2 approximation with bounded ℓ^0 and ℓ^1 norm	69
4	Minimax lower bounds for ridge combinations including neural nets	73
4.1	Introduction	73
4.2	Results for sinusoidal nets	77
4.3	Implications for neural nets	81
4.4	Implications for polynomial nets	82
4.5	Discussion	84
5	Estimating the coefficients of a mixture of two linear regressions by expectation maximization	85
5.1	Introduction	85
5.2	The population EM operator	87
5.3	The sample EM operator	91
5.4	Without assuming symmetry	92
5.5	Proofs of main theorems	94
5.6	Extensions to other models	100
5.7	Discussion	101
5.8	Additional proofs	102
6	Recovering the endpoint of a density from noisy data with application to convex body estimation	112
6.1	Preliminaries	112
6.1.1	Introduction	112
6.1.2	Notation	114
6.1.3	Notation	114
6.1.4	Model and outline	115
6.2	Estimation when $d = 1$	118

6.3	Dominating bias in endpoint estimation	120
6.4	Application to convex support estimation from noisy data	122
6.4.1	Definition of the estimator	122
6.4.2	Lower bound for the minimax risk	124
6.5	Proofs	128
6.5.1	Proof of Theorem 17	128
6.5.2	Proof of Theorem 18	130
6.5.3	Proof of Theorem 19	130
6.5.4	Proof of Corollary 3	131
6.5.5	Proofs of the lemmas and corollaries	134
7	Estimating the number of connected components in a graph via subgraph sampling	144
7.1	Introduction	144
7.1.1	Organization	148
7.1.2	Notations	149
7.2	Model	150
7.2.1	Subgraph sampling model	150
7.2.2	Classes of graphs	152
7.3	Main results	154
7.4	Algorithms and performance guarantees	157
7.4.1	Combinatorial properties of chordal graphs	158
7.4.2	Estimators for chordal graphs	160
7.4.3	Unions of cliques	169
7.4.4	Extensions to uniform sampling model	171
7.4.5	Non-chordal graphs	171
7.5	Lower bounds	172
7.5.1	General strategy	172
7.5.2	Bounding total variations between sampled graphs	175
7.5.3	Lower bound for graphs with long induced cycles	179

7.5.4	Lower bound for chordal graphs	180
7.5.5	Lower bounds for forests	185
7.6	Numerical experiments	186
7.7	Additional proofs	189
8	Counting motifs with graph sampling	196
8.1	Introduction	196
8.1.1	Sampling model	199
8.1.2	Main results	201
8.1.3	Notations	203
8.1.4	Organization	204
8.2	Methodologies and performance guarantees	205
8.2.1	Subgraph sampling	205
8.2.2	Neighborhood sampling	207
8.3	Lower bounds	216
8.3.1	Auxiliary results	216
8.3.2	Subgraph sampling	222
8.3.3	Neighborhood sampling	224
8.4	Graphs with additional structures	225
8.5	Numerical experiments	228
8.6	Discussion	231
8.7	Auxiliary lemmas	232
8.8	Additional proofs	243
8.9	Neighborhood sampling without colors	248
8.10	Lower bounds for other motifs	250

List of Figures

5.1	The population EM operator $M(\theta)$ lies in the space spanned by θ and θ^* . The unit vector θ_0^\perp lies in the space spanned by θ and θ^* and is perpendicular to θ . The vector θ forms an angle α with θ^*	90
7.1	Subgraph sampling.	150
7.2	Examples of G (resp. G') consisting multiple copies of H (resp. H') with $r = 3$. Both graphs have 6 vertices and 6 edges.	153
7.3	Examples of chordal and non-chordal graphs both with three connected com- ponents.	154
7.4	A chordal graph G with PEO labelled. In this example, $\text{cc}(G) = 3 = 16 -$ $19 + 6 = \mathfrak{s}(K_1, G) - \mathfrak{s}(K_2, G) + \mathfrak{s}(K_3, G)$	158
7.5	Each connected subgraph with $k \leq 4$ vertices appears exactly $9 - k$ times in each graph.	178
7.6	Example where $U = \{u_1, u_2\}$ is an edge. If any of these vertices are not sampled and all incident edges are removed, the resulting graphs are isomorphic. 179	
7.7	Example for $\omega = 4$ and $m = 3$, where $U = \{u_1, u_2, u_3\}$ form a triangle. If any one or two (as shown in the figure) of these vertices are not sampled and all incident edges are removed, the graphs are isomorphic.	183
7.8	Illustration for the construction in (7.47) for $\omega = 3$. Each graph contains a matching number of cliques of size up to 2.	184
7.9	Illustration for the construction in (7.48) for $\omega = 4$. Each graph contains a matching number of cliques of size up to 3.	184

7.10	The two graphs are isomorphic if the center vertex is not sampled and all incident edges are removed. Thus, $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = p(1 - q^6)$	185
7.11	The relative error of $\hat{\text{cc}}$ with moderate values of d and ω	188
7.12	A comparison of the relative error of the unbiased estimator $\hat{\text{cc}}$ in (7.10) and its smoothed version $\hat{\text{cc}}_L$ in (7.25). The parent graph is a triangulated realization of $\mathcal{G}(1000, 0.0015)$ with $d = 88$, $\omega = 15$, and $\text{cc}(G) = 325$	188
7.13	The estimator $\hat{\text{cc}}(\text{TRI}(\tilde{G}))$ applied to non-chordal graphs.	189
8.1	A comparison of subgraph and neighborhood sampling: Five vertices are sampled in the parent graph, and the observed graph is shown in Fig. 8.1b and Fig. 8.1c for the subgraph and neighborhood sampling, respectively.	201
8.2	Relative error of estimating the edge count. In Fig. 8.2a and Fig. 8.2b, the parent graph G is the Facebook network with $d = 77$, $\text{v}(G) = 333$, $\text{e}(G) = 2519$. In Fig. 8.2c and Fig. 8.2d, G is a realization of the Erdős-Rényi graph $\mathcal{G}(1000, 0.05)$ with $d = 12$, and $\text{e}(G) = 2536$	229
8.3	Relative error of counting triangles. In Fig. 8.3a and Fig. 8.3b, the parent graph is the Facebook network with $d = 77$, $\text{v}(G) = 168$, $\text{t}(G) = 7945$. In Fig. 8.3c and Fig. 8.3d, the parent graph is a realization of the Erdős-Rényi graph $\mathcal{G}(1000, 0.02)$ with $d = 35$, and $\text{t}(G) = 1319$	230
8.4	Relative error of counting wedges. In Fig. 8.4a and Fig. 8.4b, the parent graph is a Facebook network with $d = 29$, $\text{v}(G) = 61$, $\text{w}(G) = 1039$. In Fig. 8.4c and Fig. 8.4d, the parent graph is a realization of the Erdős-Rényi graph $\mathcal{G}(1000, 0.001)$ with $d = 7$, and $\text{w}(G) = 514$	231

List of Tables

2.1	Main contributions to penalties for Theorem 2 over continuum of candidate fits	16
2.2	Main contributions to penalties for Theorem 2 over discretization of candidate fits	16
7.1	Sample complexity for various classes of graphs	157
8.1	Probability mass function of $\mathcal{K}_A \mathcal{K}_{A'}$ for two distinct intersecting edges (excluding zero values).	209
8.2	The graph H with $\ell = 5$ and $d(H) = \ell + 1 = 6$	242
8.3	The graph H' with $\ell = 5$ and $d(H') = \ell + 1 = 6$	242
8.4	The graph H with $\ell = 5$	242
8.5	The graph H' with $\ell = 5$	242

Acknowledgements

I am deeply grateful to my advisor, Prof. Andrew R. Barron, for his careful guidance and training during my career at Yale. I have had the honor of working on various problems with him and, in doing so, benefited from his profound and broad knowledge of information theory and statistics. I have also witnessed his incredible originality and uncanny intuition play out in multiple ways and his perspective has shaped how I think about the field. I can only hope to follow in his footsteps as a future academician. I have also had the opportunity to work with and be mentored by Prof. Yihong Wu, whose talent and energy is endlessly inspiring. His advice and example have made me a better author and researcher. I would also like to thank other members of the Yale Statistics & Data Science faculty, in particular, Profs. Harrison Zhou, Jessi Cisewski, Jay Emerson, Joseph Chang, and David Pollard for their advice and mentorship. Furthermore, Prof. Marina Niessner at the Yale School of Management opened my eyes to the world of finance and advised me on many aspects of being a young researcher. My graduate student colleagues, Dana Yang, David Brinda, Natalie Doss, Tal Sarig, Elena Khusainova, Derek Feng, Sören Kunzel, and Yu Lu have also provided an intellectually stimulating environment for research and debate. I am very fortunate to have had their friendship over the years.

The Yale administrative staff, Joann DelVecchio, JoAnn Falato, Karen Kavanaugh, and Elizavette Torres are world class. Their smiles and cheer brightened even the toughest of days.

I would also like to thank my undergraduate classmates at the University of Manitoba, Creagh Briercliffe, Kevin Mather, and Toban and Michael Wiebe, for encouraging me to do a PhD. My undergraduate supervisors, Nina Zorboska, David S. Gunderson, and

Brad Johnson, are also directly responsible for encouraging me to study mathematics and statistics and my admission to Yale.

Finally, I would like to thank my parents, Rose and Glen Klusowski, for their love and nurture. They never sacrificed the quality of my education or their support of my interests and curiosities.

Chapter 1

Introduction

This thesis focuses on describing the trade-off between computational complexity and statistical estimation in a variety of settings – mainly high-dimensional non-linear regression, mixture models, density support recovery, and network analysis. Below we provide a brief description of each chapter and the contents therein.

Chapter 2

Let f^\star be a function on \mathbb{R}^d with an assumption of a spectral norm v_{f^\star} . For various noise settings, we show that $\mathbb{E}\|\hat{f} - f^\star\|^2 \leq \left(v_{f^\star}^4 \frac{\log d}{n}\right)^{1/3}$, where n is the sample size and \hat{f} is either a penalized least squares estimator or a greedily obtained version of such using linear combinations of sinusoidal, sigmoidal, ramp, ramp-squared or other smooth ridge functions. The candidate fits may be chosen from a continuum of functions, thus avoiding the rigidity of discretizations of the parameter space. On the other hand, if the candidate fits are chosen from a discretization, we show that $\mathbb{E}\|\hat{f} - f^\star\|^2 \leq \left(v_{f^\star}^3 \frac{\log d}{n}\right)^{2/5}$.

This work bridges non-linear and non-parametric function estimation and includes single-hidden layer nets. Unlike past theory for such settings, our bound shows that the risk is small even when the input dimension d of an infinite-dimensional parameterized dictionary is much larger than the available sample size. When the dimension is larger than the cube root of the sample size, this quantity is seen to improve the more familiar risk bound of $v_{f^\star} \left(\frac{d \log(n/d)}{n}\right)^{1/2}$, also investigated here. The heart of the analysis relies on showing that one can restrict the ℓ_1 and ℓ_0 norms of the inner and outer parameters, without sacrificing

the flexibility and richness of these ridge combinations.

Chapter 3

In Chapter 2, it is shown that small mean squared prediction error is achieved by ℓ^1 penalized least squares estimators over the class of ridge combinations. These statistical error bounds are obtained by optimizing the tradeoff between *approximation error* and *descriptive complexity relative to sample size*, when the model consists of sparse ridge combinations. In this chapter, we establish L^∞ and L^2 approximation error bounds for functions of many variables that are approximated by linear combinations of ReLU (rectified linear unit) and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls on their inner and outer parameters. With the squared ReLU ridge function, we show that the L^2 approximation error is inversely proportional to the inner layer ℓ^0 sparsity and it need only be sublinear in the outer layer ℓ^0 sparsity. Our constructions are obtained using a variant of the Maurey-Jones-Barron probabilistic method, which can be interpreted as either stratified sampling with proportionate allocation or two-stage cluster sampling. We also provide companion error lower bounds that reveal near optimality of our constructions. Despite the sparsity assumptions, we showcase the richness and flexibility of these ridge combinations by defining a large family of functions, in terms of certain spectral conditions, that are particularly well approximated by them.

Chapter 4

In this chapter, we investigate the optimality of the risk bounds from Chapter 2. More specifically, estimation of functions of d variables is considered using ridge combinations of the form $\sum_{k=1}^m c_{1,k} \phi(\sum_{j=1}^d c_{0,j,k} x_j - b_k)$ where the activation function ϕ is a function with bounded value and derivative. These include single-hidden layer neural networks, polynomials, and sinusoidal models. From a sample of size n of possibly noisy values at random sites $X \in B = [-1, 1]^d$, the minimax mean square error is examined for functions in the closure of the ℓ_1 hull of ridge functions with activation ϕ . It is shown to be of order d/n to a fractional power (when d is of smaller order than n), and to be of order $(\log d)/n$ to a fractional power (when d is of larger order than n). In particular, we show that in the regimes

$n \gg d$ and $n \ll d$, the aforementioned risk upper bounds from Chapter 2, $\left(\frac{d \log(n/d)}{n}\right)^{1/2}$ and $\left(\frac{\log d}{n}\right)^{2/5}$, have accompanying lower bounds $\left(\frac{d \log(n/d)}{n}\right)^{1/2}$ and $\left(\frac{\log d}{n}\right)^{1/2}$, respectively, for analogously restricted parameter spaces (i.e. bounded ℓ_1 norm of inner and outer layer coefficients). Dependence on constraints v_0 and v_1 on the ℓ_1 norms of inner parameter c_0 and outer parameter c_1 , respectively, is also examined. The heart of the analysis is development of information-theoretic packing numbers for these classes of functions.

Chapter 5

We give convergence guarantees for estimating the coefficients of a symmetric mixture of two linear regressions by expectation maximization (EM). In particular, we show that convergence of the empirical iterates is guaranteed provided the algorithm is initialized in an unbounded cone. That is, if the initializer has a large cosine angle with the population coefficient vector and the signal to noise ratio (SNR) is large, a sample-splitting version of the EM algorithm converges to the true coefficient vector with high probability. Here “large” means that each quantity is required to be at least a universal constant. Finally, we show that the population EM operator is not globally contractive by characterizing a region where it fails. We give empirical evidence that suggests that the sample based EM performs poorly when initializers are drawn from this set. Interestingly, our analysis borrows from tools used in the problem of estimating the centers of a symmetric mixture of two Gaussians by EM [6]. We also discuss some extensions to mixtures of nonlinear regression models, such as ramp or step activation functions.

This chapter is based on joint work with W. D. Brinda and Dana Yang; see [7] for the manuscript in its original form.

Chapter 6

A popular class of problem in statistics deals with estimating the support G of a density μ from observations X_1, \dots, X_n drawn at random from a d -dimensional distribution with density μ . The one-dimensional case reduces to estimating the end point of a univariate density; a problem that has been extensively studied in the literature [8]. When the support

is assumed to have a convex shape-constraint and μ is the uniform density on G , the convex polytope $\text{Conv}(X_1, \dots, X_n)$ is a minimax optimal estimator and its statistical properties have a long history in stochastic geometry.

A natural question to ask is how the problem changes when the observations X are contaminated with some additive noise ε via $Y = X + \varepsilon$. Note that one can also view this problem as estimating the support of a mixing measure p_X under an infinite mixture model, e.g., $p_Y(y) = \mathbb{E}_{X \sim p_X}[p_\varepsilon(y - X)]$. Here we can no longer use the convex polytope estimator since there is a probability that at least one observation will land outside G and these outliers enlarge the boundary of $\text{Conv}(Y_1, \dots, Y_n)$ so that it overestimates G .

Such a problem falls under the guise of the so-called inverse or deconvolution problems and it is usually considered in the context of density estimation or regression. The analog of this model in the aforementioned univariate setting is to estimate the endpoint of a density when the observations have been contaminated by some additive noise. This scenario has only more recently been considered in [9, 10], where it was assumed that the density of μ is *exactly* equal to a polynomial in a neighborhood of the endpoint of the support. Ideally, one would like to relax this so that the density only behaves *approximately* like a polynomial near its boundary.

In the multidimensional case, techniques from deconvolution in density and function estimation can be applied. These are usually implemented as plug-in estimators, where the density is first estimated using Fourier transforms and kernel density estimators and the support estimator is then obtained by thresholding the density estimator. One major pitfall of these estimators is that there is a bandwidth parameter that must be selected a priori and it is not always clear how to do this in practice.

When ε is distributed according to a multivariate normal distribution, we consider estimation of compact convex supports under the deconvolution model that avoids tuning parameters and, as a byproduct, extends the results of [9] when the distribution function behaves *approximately* like a polynomial in the vicinity of the endpoint. The estimator we propose takes particular advantage of the spherical symmetry of the Gaussian density and the convexity of the support. The strategy is to estimate the support function of G , by $\hat{h}_n(u) = \max_{1 \leq i \leq n} Y_i \cdot u - b_n$ (where b_n is an explicit sequence) and then estimate G by

$\widehat{G}_n = \{x \in \mathbb{R}^d : \langle u, x \rangle \leq \widehat{h}_n(u) \text{ for all } u \in \mathbb{S}^{d-1}\}$. We show that \widehat{G}_n is a suitable estimator and that it converges to G at a rate of $O_d(\log \log n / \sqrt{\log n})$ in Hausdorff distance. This logarithmic rate of convergence is considerably worse than in the noiseless case and is consistent with the sort of slow rates encountered in Gaussian deconvolution problems [11]. Part of the analysis also involves the optimality of the proposed estimator. We provide a minimax lower bound for this estimation problem by selecting two sets G_1 and G_2 with equal Lebesgue measure for which the Fourier transform of their difference $|\mathcal{F}[\mathbb{1}_{G_1} - \mathbb{1}_{G_2}]|$ is small in some ball around the origin, akin to a lower bound construction used by [11] for deconvolution in manifold estimation under Hausdorff loss. Using these sets, we show that the minimax rate of estimating G in Hausdorff distance is $\Omega_d(1/\log^2 n)$. The lower bound is different than other lower bounds in deconvolution problems. For example, in standard density or regression deconvolution [12], the classes are rich enough to ensure the existence of a function whose Fourier transform vanishes on a compact interval. The uncertainty principle for Fourier transforms makes that impossible in this setting, since the function class consists of compactly supported functions.

This chapter is based on joint work with Victor-Emmanuel Brunel and Dana Yang; see [13] for the original manuscript in its full form. Although we will not include it here, we have also extended our theory for noise distributions other than Gaussian (i.e. Cauchy). For more details, see [14].

Chapter 7

Learning properties of large graphs from samples has been an important problem in statistical network analysis since the early work of Goodman [15] and Frank [16]. We revisit a problem formulated by Frank [16] of estimating the number of connected components in a large graph based on the subgraph sampling model, in which we randomly sample a subset of the vertices and observe the induced subgraph. The key question is whether accurate estimation is achievable in the *sublinear* regime where only a vanishing fraction of the vertices are sampled. We show that it is impossible if the parent graph is allowed to contain high-degree vertices or long induced cycles. For the class of chordal graphs, where induced cycles of length four or above are forbidden, we characterize the optimal sample complexity

within constant factors and construct linear-time estimators that provably achieve these bounds. This significantly expands the scope of previous results which have focused on unbiased estimators and special classes of graphs such as forests or cliques.

Both the construction and the analysis of the proposed methodology rely on combinatorial properties of chordal graphs and identities of induced subgraph counts. They, in turn, also play a key role in proving minimax lower bounds based on construction of random instances of graphs with matching structures of small subgraphs.

Let $\text{cc}(G)$ denote the number of connected components of a graph G . If $\mathcal{G}(N, d, \omega)$ denotes the collection of all chordal graphs on N vertices with clique number ω and maximum degree at most d , we show the minimax rate $\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{G}(N, d, \omega)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 = \Theta_\omega \left(\left(\frac{N}{p^\omega} \vee \frac{Nd}{p^{\omega-1}} \right) \wedge N^2 \right)$. In the large ω setting, We also show that $\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{G}(N, d, \omega)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 \leq N^2 \left(\frac{d\omega}{N} \right)^{\frac{p}{2q-p}}$ via a truncated estimator that achieves the optimal bias-variance tradeoff. Thus, even when $d = o(\sqrt{N})$, accurate estimating is still possible. Importantly, all estimators that achieve these rates are adaptive to both d and ω .

This chapter is based on joint work with Yihong Wu; see [17] for the manuscript in its original form.

Chapter 8

Applied researchers often construct a network from data that has been collected from a random sample of nodes, with the goal to infer properties of the parent network from the sampled version. Two of the most widely used sampling schemes are *subgraph sampling*, where we sample each vertex independently with probability p and observe the subgraph induced by the sampled vertices, and *neighborhood sampling*, where we additionally observe the edges between the sampled vertices and their neighbors.

In this chapter, we study the problem of estimating the number of motifs as induced subgraphs under both models from a statistical perspective. We show that: for parent graph G with maximal degree d , for any connected motif h on k vertices, to estimate the number of copies of h in G , denoted by $s = \mathbf{s}(h, G)$, with a multiplicative error of ϵ ,

- For subgraph sampling, the optimal sampling ratio p is $\Theta_k(\max\{(s\epsilon^2)^{-\frac{1}{k}}, \frac{d^{k-1}}{s\epsilon^2}\})$,

which only depends on the size of the motif but *not* its actual topology. Furthermore, we show that Horvitz-Thompson type estimators are universally optimal for any connected motifs.

- For neighborhood sampling, we propose a family of estimators, encompassing and outperforming the Horvitz-Thompson estimator and achieving the sampling ratio $O_k(\min\{(\frac{d}{s\epsilon^2})^{\frac{1}{k-1}}, \sqrt{\frac{d^{k-2}}{s\epsilon^2}}\})$, which again only depends on the size of h . This is shown to be optimal for all motifs with at most 4 vertices and cliques of all sizes.

For example, if $\mathcal{G}(m, d)$ is the collection of all graphs with at most m edges and maximum degree at most d , then under neighborhood sampling, $\inf_{\hat{e}} \sup_{G \in \mathcal{G}(m, d)} \mathbb{E}_G |\hat{e} - \mathbf{e}(G)|^2 \asymp \frac{m}{p^2} \wedge \frac{md}{p} \wedge m^2$, whereas under vertex sampling we have the worse rate $\inf_{\hat{e}} \sup_{G \in \mathcal{G}(m, d)} \mathbb{E}_G |\hat{e} - \mathbf{e}(G)|^2 \asymp \frac{m}{p^2} \vee \frac{md}{p} \wedge m^2$, in accordance with the more limited sampling model. The matching minimax lower bounds are established using certain algebraic properties of subgraph counts. These results allow us to quantify how much more informative neighborhood sampling is than subgraph sampling, as empirically verified by experiments on synthetic and real-world data. We also address the issue of adaptation to the unknown maximum degree, and study specific problems for parent graphs with additional structures, e.g., trees or planar graphs.

This chapter is based on joint work with Yihong Wu; see [18] for the manuscript in its original form.

Other work

In addition to conducting research Prof. Andrew R. Barron, and other collaborators in statistics, the author has worked closely with Prof. Marina Niessner in finance at the Yale School of Management on various applied projects involving statistical natural language processing and network analysis (see [19] for the outcome of this work).

The author has also completed work with W. D. Brinda in [20]; see his thesis for details of this work.

Chapter 2

Risk bounds for high-dimensional ridge function combinations including neural networks

2.1 Introduction

Functions f^\star in \mathbb{R}^d are approximated using linear combinations of ridge functions with one layer of nonlinearities. These approximations are employed via functions of the form

$$f_m(x) = f_m(x, \zeta) = \sum_{k=1}^m c_k \phi(a_k \cdot x + b_k), \quad (2.1)$$

which is parameterized by the vector ζ , consisting of a_k in \mathbb{R}^d , and b_k, c_k in \mathbb{R} for $k = 1, \dots, m$, where $m \geq 1$ is the number of nonlinear terms. Models of this type arise with considerable freedom in the choice of the activation function ϕ , ranging from general smooth functions of projection pursuit regression [21] to the unit step sigmoid and ramp functions of single-hidden layer neural nets [3, 5, 22–24].

Our focus in this chapter is on the case that ϕ is a fixed Lipschitz function (such as a sigmoid or ramp or sinusoidal function), though some of our conclusions apply more generally. For these activation functions, we will obtain statistical risk bounds using a penalized least squares criterion. We obtain generalization error bounds for these by balancing the

approximation error and descriptive complexity. The most general form of our bounds hold for quite general non-linear infinite dictionaries. A hallmark of our conclusions is to lay bare how favorable risk behavior can be obtained as long as the logarithm of the number of parameters relative to sample size is small. This entails a slower rate of convergence through a rate that is smaller than what is cemented in traditional cases, but leads to better results than these earlier bounds would permit in certain very high-dimensional situations. From an applied perspective, good empirical performance of neural net (and neural net like) models has been reported as in [25] even when d is much larger than n , though theoretical understanding has been lacking. Returning to the case of a single layer of nonlinearly parameterized function, it is useful to view the representation (2.1) as

$$\sum_h \beta_h h(x), \quad (2.2)$$

where the h are a selection of functions from the infinite library $\mathcal{H} = \mathcal{H}_\phi$ of functions of the form $\pm\phi(\theta \cdot x)$ for real vector θ and the β_h are coefficients of linear combination of $\pm\phi$ in the library. These representations are single hidden-layer networks. Deep network approximations are not very well understood. Nevertheless our results generalize provided some of our arguments are slightly modified.

We can reduce (2.1) to (2.2) as follows. Suppose the library is symmetric $\mathcal{H} = -\mathcal{H}$ and contains the zero function. Without loss of generality, we may assume that the c_k or β_h are non-negative by replacing the associated ϕ with $\phi \operatorname{sgn} c_k$, that by assumption also belongs to \mathcal{H} . One can assume the internal parameterization $a \cdot x + b$ take the form $\theta \cdot x$ by appending a coordinate of constant value 1 to x and a coordinate of value b to the vector a . Note that now x and θ are $(d+1)$ -dimensional.

We will take advantage of smoothness of the activation function (assumption that either ϕ is Lipschitz or that its first derivative ϕ' is Lipschitz). Suppose P is an arbitrary probability measure on $[-1, 1]^d$. Let $\|\cdot\|$ be the $L^2(P)$ norm induced by the inner product $\langle \cdot, \cdot \rangle$. For a symmetric collection of dictionary elements $\mathcal{H} = -\mathcal{H}$ containing the zero function, we let $\mathcal{F} = \mathcal{F}_\mathcal{H}$ be the linear span of \mathcal{H} .

The variation $v_f = \|f\|_\mathcal{H}$ of f with respect to \mathcal{H} (or the atomic norm of f with respect

to \mathcal{H}) is defined by

$$\lim_{\delta \downarrow 0} \inf_{f_\delta \in \mathcal{F}} \left\{ \|\beta\|_1 : f_\delta = \sum_{h \in \mathcal{H}} \beta_h h \text{ and } \|f_\delta - f\| \leq \delta, \beta_h \in \mathbb{R}^+ \right\},$$

where $\|\beta\|_1 = \sum_{h \in \mathcal{H}} \beta_h$. For functions in $\mathcal{F}_{\mathcal{H}}$, this variation picks out the smallest $\|\beta\|_1$ among representations $f = \sum_{h \in \mathcal{H}} \beta_h h$. In the particular case that $f = \sum_{h \in \mathcal{H}} \beta_h h$, we have $v_f = \|\beta\|_1$. For functions in the $L^2(P)$ closure of the linear span of \mathcal{H} , the variation is the smallest limit of such ℓ_1 norms among functions approaching the target. The subspace of functions with $\|f\|_{\mathcal{H}}$ finite is denoted $L_{1,\mathcal{H}}$. Such variation control provides for approximation (opportunity) for dimension independent rates of order $1/\sqrt{m}$ with an m term approximation.

It is fruitful to discuss spectral conditions for finite variation for various choices of ϕ . To this end, define $v_{f^*,s} = \int_{\mathbb{R}^d} \|\omega\|_1^s \tilde{f}(\omega) d\omega$, for $s \geq 0$. If f^* has a bounded domain in $[-1,1]^d$ and a Fourier representation $f^*(x) = \int_{\mathbb{R}^d} e^{i\omega \cdot x} \mathcal{F}(f)(\omega) d\omega$ with $v_{f^*,1} < +\infty$, it is possible to use approximating functions of the form (2.1) with a single activation function ϕ . Such activation functions ϕ can be general bounded monotone functions. We use x for vectors in \mathbb{R}^d and z for scalars such as $z = \theta \cdot x$. As we have said, to obtain risk bounds in later sections, we will assume that either ϕ is bounded Lipschitz or that, additionally, its derivative ϕ' is Lipschitz. These two assumptions are made precise in the following statements.

Assumption 1. *The activation function ϕ has L_∞ norm at most one and satisfies*

$$|\phi(z) - \phi(\tilde{z})| \leq L_1 |z - \tilde{z}|,$$

for all z, \tilde{z} in \mathbb{R} and for some positive constant $L_1 > 0$.

Assumption 2. *The activation function ϕ has L_∞ norm at most one and satisfies*

$$|\phi(z) - \phi(\tilde{z})| \leq L_1 |z - \tilde{z}|,$$

and

$$|\phi'(z) - \phi'(\tilde{z})| \leq L_2|z - \tilde{z}|,$$

for all z, \tilde{z} in \mathbb{R} and for some positive constants $L_1 > 0$ and $L_2 > 0$.

In particular, Assumption 2 implies that

$$|\phi(z) - \phi(\tilde{z}) - (z - \tilde{z})\phi'(\tilde{z})| \leq \frac{1}{2}(z - \tilde{z})^2 L_2,$$

for all z, \tilde{z} in \mathbb{R} .

A result from [23] provides a useful starting point for approximating general functions f^* by linear combinations of such objects. Suppose $v_{f^*,1}$ is finite. Then by [23] the function f^* has finite variation with respect to step functions and, consequently, there exists an artificial neural network of the form (2.1) with $\phi(x) = \text{sgn}(x)$, $\|a_k\|_1 = 1$, and $|b_k| \leq 1$ such that, if a suitable constant correction is subtracted from f^* , then

$$\|f^* - f_m\|^2 \leq \frac{v_{f^*,1}^2}{m}.$$

In particular, f^* minus a constant correction has variation less than $v_{f^*,1}$.

If ϕ has right at left limits -1 and $+1$, respectively, the fact that $\phi(\tau x) \rightarrow \text{sgn}(x)$ as $\tau \rightarrow +\infty$ allows one to use somewhat arbitrary activation functions as basis elements. For our results, it is undesirable to have unbounded weights. Accordingly, it is natural to impose a restriction on the size of the internal parameters and to also enjoy a certain degree of smoothness not offered by step functions. Although, it should be mentioned that classical empirical process theory allows one to obtain covering numbers for indicators of half-spaces (which are scale invariant in the size of the weights) by taking advantage of their combinatorial structure [26]. Nevertheless, we adopt the more modern approach of working with smoothly parameterized dictionaries. In this direction, we consider the result in [5], which allows one to approximate f^* by linear combinations of ramp ridge functions (also known as first order ridge splines or hinging hyper-planes) $(x \cdot \alpha - t)_+ = \max\{0, x \cdot \alpha - t\}$, with $\|\alpha\|_1 = 1$, $|t| \leq 1$.

The ramp activation function $\phi(x) = (x)_+$ (also called a lower-rectified linear unit or

ReLU) is currently one of the most popular form of artificial neural network activation functions, particularly because it is continuous and Lipschitz. In particular, it satisfies the conditions of Assumption 1 with $L_1 = 1$ depending on the size of its domain. In [27], we refine a result from [5]. An arbitrary target function f^\star with $v_{f^\star,2}$ finite has finite variation with respect to the ramp functions and, consequently, there exists an approximation of the form (2.1) activated by ridge ramp functions with $\|a_k\| = 1$ and $|b_k| \leq 1$ such that if a suitable linear correction is subtracted from f^\star , then

$$\|f^\star - f_m\|^2 \leq cv_{f^\star,2}^2 m^{-1/2-1/d}, \quad (2.3)$$

for some universal positive constant c . In particular, f^\star minus a linear correction has variation less than $v_{f^\star,2}$. The linear correction may be regarded as included in the approximation (2.1).

The second order spline $\phi(x) = (x)_+^2$, which may also be called ramp-squared, satisfies the conditions of Assumption 2 with constants $L_1 = 2$ and $L_2 = 2$ depending on the size of its domain. Likewise, in [27], we show that for an arbitrary target function f^\star with $v_{f^\star,3}$ finite, a quadratically corrected f^\star has finite variation with respect to second order splines, and consequently, there exists an approximation of the form (2.1) activated by second order ridge splines with $\|a_k\| = 1$ and $|b_k| \leq 1$ such that, if a suitable quadratic correction is subtracted from f^\star , then

$$\|f^\star - f_m\|^2 \leq cv_{f^\star,3}^2 m^{-1/2-1/d}, \quad (2.4)$$

for some universal positive constant c . In particular, f^\star minus a quadratic correction has variation less than $v_{f^\star,3}$.

For integer $s \geq 1$, we define the infinite dictionary

$$\mathcal{H}_s = \{x \mapsto \pm(\alpha \cdot x - t)_+^{s-1} : \|\alpha\|_1 = 1, |t| \leq 1\}.$$

We then set \mathcal{F}_s to be the linear span of \mathcal{H}_s . With this notation, $\mathcal{F}_{\text{ramp}} = \mathcal{F}_2$.

The condition $\int_{\mathbb{R}^d} \|\omega\|_1^s |\tilde{f}(\omega)| d\omega < +\infty$ ensures that f^\star (corrected by a $(s-1)$ -th degree ridge polynomial) belongs to L_{1,\mathcal{H}_s} and $\|f^\star\|_{\mathcal{H}_s} \leq v_{f^\star,s}$. Functions with moderate variation

are particularly closely approximated. Nevertheless, even when $\|f^\star\|_{\mathcal{H}}$ is infinite, we express the trade-offs in approximation accuracy for consistently estimating functions in the closure of the linear span of \mathcal{H} .

In what follows, we assume that the internal parameters have ℓ_1 norm at most v_0 . Likewise, we assume that $x \in [-1, 1]^d$ so that $|\theta \cdot x| \leq \|\theta\|_1 \leq v_0$. This control on the size of the internal parameters will be featured prominently throughout. In the case of spline activation functions, we are content with the assumption $v_0 = 1$. Note that if one restricts the size of the domain and internal parameters (say, to handle polynomials), the functions h are still bounded and Lipschitz but with possibly considerably worse constants.

Suppose data $\{(X_i, Y_i)\}_{i=1}^n$ are independently drawn from the distribution of (X, Y) . To produce predictions of the real-valued response Y from its input X , the target regression function $f^\star(x) = \mathbb{E}[Y|X = x]$ is to be estimated. The function f^\star is assumed to be bounded in magnitude by a positive constant B . We assume the noise $\varepsilon = Y - f^\star(X)$ has moments (conditioned on X) that satisfy a Bernstein condition with parameter $\eta > 0$. That is, we assume

$$\mathbb{E}(|\varepsilon|^k|X) \leq \frac{1}{2}k!\eta^{k-2}\mathbb{V}(\varepsilon|X), \quad k = 3, 4, \dots,$$

where $\mathbb{V}(\varepsilon|X) \leq \sigma^2$. This assumption is equivalent to requiring that $\mathbb{E}(e^{|\varepsilon|/\nu}|X)$ is uniformly bounded in X for some $\nu > 0$, i.e., X is subexponential. A stricter assumption is that $\mathbb{E}(e^{|\varepsilon|^2/\nu}|X)$ is uniformly bounded in X , which corresponds to an error distribution with sub-Gaussian tails. These two noise settings will give rise to different risk bounds, as we will see.

Because f^\star is bounded in magnitude by B , it is useful to truncate an estimator \hat{f} at a level B_n at least B . Depending on the nature of the noise ε , we will see that B_n will need to be at least B plus a term of order $\sqrt{\log n}$ or $\log n$. We define the truncation operator T that acts on function f in \mathcal{F} by $Tf = \min\{|f|, B_n\}\text{sgn}f$. This Tf is a fully rectified linear ramp with maximum value B_n . Associated with the truncation operator is a tail quantity $T_n = 2\sum_{i=1}^n(|Y_i|^2 - B_n^2)\mathbb{I}\{|Y_i| > B_n\}$ that appears in the following analysis and our risk bounds have a $\mathbb{E}[T_n/n]$ term, but this will be seen to be negligible when compared to the main terms. The behavior of $\mathbb{E}T_n$ is studied in Lemma 10.

The empirical mean squared error of a function f as a candidate fit to the observed data is $(1/n) \sum_{i=1}^n (Y_i - f(X_i))^2$. Given the collection of functions \mathcal{F} , a penalty $\text{pen}_n(f)$, $f \in \mathcal{F}$, and data, a penalized least squares estimator \hat{f} arises by optimizing or approximately optimizing

$$(1/n) \sum_{i=1}^n (Y_i - f(X_i))^2 + \text{pen}_n(f)/n. \quad (2.5)$$

Our method of risk analysis proceeds as follows. Given a collection \mathcal{F} of candidate functions, we show that there is a countable approximating set $\tilde{\mathcal{F}}$ of representations \tilde{f} , variable-distortion, variable-complexity cover of \mathcal{F} , and a complexity function $L_n(\tilde{f})$, with the property that for each f in \mathcal{F} , there is an \tilde{f} in $\tilde{\mathcal{F}}$ such that $\text{pen}_n(f)$ is not less than a constant multiple of $\gamma_n L_n(\tilde{f}) + \Delta_n(f, \tilde{f})$, where γ_n is a constant (depending on B , σ^2 , and η) and $\Delta_n(f, \tilde{f})$ is given as a suitable empirical measure of distortion (based on sums of squared errors). The variable-distortion, variable-complexity terminology has its origins in [28–30]. The task is to determine penalties such that an estimator \hat{f} approximately achieving the minimum of $\|Y - f\|_n^2 + \text{pen}_n(f)/n$ satisfies

$$\mathbb{E}\|T\hat{f} - f^\star\|^2 \leq c \inf_{f \in \mathcal{F}} \{\|f - f^\star\|^2 + \mathbb{E}\text{pen}_n(f)/n\}, \quad (2.6)$$

for some universal $c > 1$. Valid penalties take different forms depending on the size of the effective dimension d relative to the sample size n and smoothness assumption of ϕ .

- When d is large compared to n and if ϕ satisfies Assumption 1, a valid penalty divided by sample size $\text{pen}_n(f)/n$ is at least

$$16v_f \left(\frac{\gamma_n B_n^2 v_0^2 \log(d+1)}{n} \right)^{1/4} + 8 \left(\frac{\gamma_n B_n^2 v_0^2 \log(d+1)}{n} \right)^{1/2} + \frac{T_n}{n}. \quad (2.7)$$

- When the noise ε is zero and d is large compared to n and if ϕ satisfies Assumption 1, a valid penalty divided by sample size $\text{pen}_n(f)/n$ is at least

$$16v_f^{4/3} \left(\frac{\gamma_n v_0^2 \log(d+1)}{n} \right)^{1/3} + 4(v_f^{4/3} + 1) \left(\frac{\gamma_n v_0^2 \log(d+1)}{n} \right)^{2/3} + \frac{T_n}{n}. \quad (2.8)$$

- When d is large compared to n and if ϕ satisfies Assumption 2, a valid penalty divided

by sample size $\text{pen}_n(f)/n$ is at least of order

$$v_f^{4/3} \left(\frac{\gamma_n v_0^2 \log(d+1)}{n} \right)^{1/3} + v_f \left(\frac{1}{n} \sum_{i=1}^n |Y_i| \right) \left(\frac{\gamma_n v_0^2 \log(d+1)}{n} \right)^{1/3} + \frac{T_n}{n}. \quad (2.9)$$

- When d is small compared to n and if ϕ satisfies Assumption 1, a valid penalty divided by sample size $\text{pen}_n(f)/n$ is at least

$$\begin{aligned} 60v_f v_0 \left(\frac{d\gamma_n \log(n/d+1)}{n} \right)^{1/2+1/(2(d+3))} &+ \frac{1}{v_0^2} \left(\frac{d\gamma_n \log(n/d+1)}{n} \right)^{1/2+1/(2(d+3))} \\ &+ \left(\frac{d\gamma_n \log(n/d+1)}{n} \right)^{1/2+3/(2(d+3))} + \frac{d\gamma_n \log(n/d+1)}{n} + \frac{T_n}{n}. \end{aligned} \quad (2.10)$$

Here $\gamma_n = (2\tau)^{-1}(1 + \delta_1/2)(1 + 2/\delta_1)(B + B_n)^2 + 2(1 + 1/\delta_2)\sigma^2 + 2(B + B_n)\eta$ and $\tau = (1 + \delta_1)(1 + \delta_2)$ for some $\delta_1 > 0$ and $\delta_2 > 0$.

Accordingly, if f^* belongs to $L_{1,\mathcal{H}}$, then $\mathbb{E}\|T\hat{f} - f^*\|^2$ is not more than a constant multiple of the above penalties with v_f replaced by $\|f^*\|_{\mathcal{H}}$.

In the single-hidden layer case, we have the previously indicated quantification of the error of approximation $\|f - f^*\|^2$. Nevertheless, the general result (2.6) allows us to likewise say that the risk for multilayer networks will be at least as good as the deep network approximation capability will permit. The quantity

$$\inf_{f \in \mathcal{F}} \{\|f - f^*\|^2 + \mathbb{E}\text{pen}_n(f)/n\}.$$

is an index of resolvability of f^* by functions \mathcal{F} with sample size n . We shall take particular advantage of such risk bounds in the case that $\text{pen}_n(f)$ does not depend on \underline{X} . Our restriction of X to $[-1, 1]^d$ is one way to allow the construction of such penalties.

The following table expresses the heart of our results in the case of penalty based on the ℓ_1 norm of the outer layer coefficients of one-hidden layer networks expressible through v_f (subject to constraints on the inner layer coefficients). These penalties also provide risk bounds for moderate and high-dimensional situations.

Table 2.1: Main contributions to penalties for Theorem 2 over continuum of candidate fits

	Activation ϕ	$\text{pen}_n(f)/n \gtrsim$	$\lambda_n \gtrsim$
I	Assumption 1	$v_f \lambda_n$	$\left(\frac{\gamma_n^2 v_0^2 \log(d+1)}{n}\right)^{1/4}$
II	Assumption 2	$(v_f)^{4/3} \lambda_n$	$\left(\frac{\gamma_n^2 v_0^2 \log(d+1)}{n}\right)^{1/3}$
III	Assumption 1	$v_f \lambda_n$	$v_0 \left(\frac{d \gamma_n \log(n/d+1)}{n}\right)^{1/2+1/(2(d+3))}$

Table 2.2: Main contributions to penalties for Theorem 2 over discretization of candidate fits

	Activation ϕ	$\text{pen}_n(f)/n \gtrsim$	$\lambda_n \gtrsim$
A	Assumption 1	$(v_f)^{4/3} \lambda_n$	$\left(\frac{\gamma_n^2 v_0^2 \log(d+1)}{n}\right)^{1/3}$
B	Assumption 2	$(v_f)^{6/5} \lambda_n$	$\left(\frac{\gamma_n^2 v_0^2 \log(d+1)}{n}\right)^{2/5}$
C	Assumption 1	$v_f \lambda_n$	$v_0 \left(\frac{d \gamma_n \log(n/d+1)}{n}\right)^{1/2+1/(d+1)}$

The results we wish to highlight are contained in the first two rows of Table 2.1. The penalties as stated are valid up to modest universal constants and negligible terms that do not depend on the candidate fit. The quantity γ_n is of order $\log^2 n$ in the sub-exponential noise case, order $\log n$ in the sub-Gaussian noise case and of constant order in the zero noise case. This γ_n (as defined in Lemma 10) depends on the variance bound σ^2 , Bernstein parameter η , the upper bound B of $\|f^*\|_{\mathcal{H}}$, and the noise tail level B_n of the indicated order.

When f^* belongs to $L_{1,\mathcal{H}}$, a resulting valid risk bound is a constant multiple of $\|f^*\|_{\mathcal{H}} \lambda_n$ or $\|f^*\|_{\mathcal{H}}^{4/3} \lambda_n$, according to the indicated cases. In this way the λ_n expression provides a rate of convergence. Thus the columns of Table 2.1 provide valid risk bounds for these settings. The statistical rates for penalized estimation over a discretization of the parameter space are derived in Section 2.7.1.

The classical risk bounds for mean squared error, involving d/n to some power, are only useful when the sample size is much larger than the dimension. Here, in contrast, in the first two lines of Table 2.1, we see the dependence on dimension is logarithmic, permitting much smaller sample sizes. These results are akin to those obtained in [31] (where the role

of the dimension there is the size of the dictionary) for high-dimensional linear regression. However, there is an important difference. Our dictionary of non-linear parameterized functions is infinite dimensional. For us, the role of d is the input dimension, not the size of the dictionary. The richness of $L_{1,\mathcal{H}}$ is largely determined by the sizes of v_0 and v_f and $L_{1,\mathcal{H}}$ more flexibly represents a larger class of functions.

The price we pay for the smaller dependence on input dimension is a deteriorated rate with exponent $1/4$ in general and $1/3$ under slightly stronger smoothness assumptions on ϕ , rather than the familiar exponents of $1/2$.

The rate in the last row improves upon the familiar exponent of $1/2$ to $1/2 + 1/((2(d+3)))$. Note that when d is large, this enhancement in the exponent is negligible. The rate in the first row is better than the third approximately for $d > \sqrt{n}$, the second is better than the third row approximately for $d > n^{1/3}$, and both of these first two rows have risk tending to zero as long as $d < e^{o(n)}$.

For functions in $L_{1,\mathcal{H}_{\text{ramp}}}$, an upper bound of $((d/n) \log(n/d))^{1/2}$ for the squared error loss is obtained in [22]. The L^2 squared error minimax rates for functions in $L_{1,\mathcal{H}_1} = L_{1,\mathcal{H}_{\text{step}}}$ [32], was determined to be between

$$(1/n)^{1/2+1/(2(d+1))}(\log n)^{-(1+1/d)(1+2/d)(1+2/d)(2+1/d)}5$$

and

$$(\log n/n)^{1/2+1/(2(2d+1))}.$$

Using the truncated penalized ℓ_1 least squares estimator (2.6), we obtain an improved rate of order $((d\gamma_n/n) \log(n/d))^{1/2+1/(2(d+3))}$, where γ_n is logarithmic in n , using techniques that originate in [33] and [34], with some corrections here.

2.2 How far from optimal?

For positive v_0 , let

$$\mathcal{D}_{v_0} = \mathcal{D}_{v_0,\phi} = \{\phi(\theta \cdot x - t), x \in B : \|\theta\|_1 \leq v_0, t \in \mathbb{R}\} \quad (2.11)$$

be the dictionary of all such inner layer ridge functions $\phi(\theta \cdot x - t)$ with parameter restricted to the ℓ_1 ball of size v_0 and variables x restricted to the cube $[-1, 1]^d$. The choice of the ℓ_1 norm on the inner parameters is natural as it corresponds to $\|\theta\|_B = \sup_{x \in B} |\theta \cdot x|$ for $B = [-1, 1]^d$. Let $\mathcal{F}_{v_0, v_1} = \mathcal{F}_{v_0, v_1, \phi} = \ell_1(v_1, \mathcal{D}_{v_0})$ be the closure of the set of all linear combinations of functions in \mathcal{D}_{v_0} with ℓ_1 norm of outer coefficients not more than v_1 . For any class of functions \mathcal{F} on $[-1, 1]^d$, the minimax risk is

$$R_{n,d}(\mathcal{F}) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E} \|f - \hat{f}\|^2, \quad (2.12)$$

Consider the model $Y = f(X) + \varepsilon$ for $f \in \mathcal{F}_{v_0, v_1, \text{sine}}$, where $\varepsilon \sim N(0, 1)$ and $X \sim \text{Uniform}([-1, 1]^d)$. It was determined in [35], that for $\frac{d}{v_0} + 1 > \left(c \frac{v_1^2 n}{v_0 \log(1+d/v_0)}\right)^{1/v_0}$, roughly corresponding to $d \gg n$,

$$R_{n,d}(\mathcal{F}_{v_0, v_1, \text{sine}}) \geq C \left(\frac{v_0 v_1^2 \log(1+d/v_0)}{n} \right)^{1/2}, \quad (2.13)$$

and for $\frac{v_0}{d} + 1 > \left(c \frac{v_1^2 n}{d \log(1+v_0/d)}\right)^{1/d}$,

$$R_{n,d}(\mathcal{F}_{v_0, v_1, \text{sine}}) \geq C \left(\frac{d v_1^2 \log(1+v_0/d)}{n} \right)^{1/2}. \quad (2.14)$$

These lower bounds are similar in form to the risk upper bounds that are implied from the penalties in Table 2.2. These quantities have the attractive feature that the rate (the power of $1/n$) remains at least as good as $1/2$ or $2/5$ even as the dimension grows. However, rates determined by (2.14) and the last line in Table 2.2 are only useful provided d/n is small. In high dimensional settings, the available sample size might not be large enough to ensure this condition.

These results are all based on obtaining covering numbers for the library $\{x \mapsto \phi(\theta \cdot x) : \|\theta\|_1 \leq v_0\}$. If ϕ satisfies a Lipschitz condition, these numbers are equivalent to ℓ_1 covering numbers of the internal parameters or of the Euclidean inner product of the data and the internal parameters. The factor of d multiplying the reciprocal of the sample size is produced from the order $d \log(v_0/\epsilon)$ log cardinality of the standard covering of the library

$\{\theta : \|\theta\|_1 \leq v_0\}$. What enables us to circumvent this polynomial dependence on d is to use an alternative cover of $\{x \mapsto x \cdot \theta : \|\theta\|_1 \leq v_0\}$ that has log cardinality of order $(v_0/\epsilon)^2 \log(d+1)$. Misclassification errors for neural networks with bounded internal parameters have been analyzed in [24, 26, 36] (via Vapnik-Chervonenkis dimension and its implications for covering numbers). Unlike the setup considered here, past work [22, 24, 29, 32, 33, 37–42] has not investigated the role of such restricted parameterized classes in the determination of suitable penalized least squares criterion for non-parametric function estimation. After submission of the original form of this work, our results have been put to use in [43] to give risk statements about multi-layer (deep) networks activated by ramp functions.

2.3 Computational aspects

From a computational point of view, the empirical risk minimization problem (2.5) is highly non-convex, and it is unclear why existing algorithms like gradient descent or back propagation are empirically successful at learning the representation (2.1). There are relatively few rigorous results that guarantee learning for regression models with latent variables, while keeping both the sampling and computational complexities polynomial in n and d . Here we catalogue some papers that make progress toward developing a provably good, computationally feasible estimation procedure. Most of them deal with parameter recovery and assume that f^* has exactly the form (2.1). Using a theory of tensor decompositions from [44], the authors of [45] apply the method of moments via tensor factorization techniques to learn mixtures of sigmoids, but they require a special non-degeneracy condition on the activation function. It is assumed that the input distribution P is known apriori. In [46], the authors use tensor initialization and resampling to learn the parameters in a representation of the form (2.1) with smooth ϕ that has sample complexity $O(d)$ and computation complexity $O(dn)$.

In [47], the authors estimate the gradient of the regression function (where X is Gaussian and ϕ is the logistic sigmoid) at a set of random points, and then cluster the estimated gradients. They prove that the estimated gradients concentrate around the internal parameter vectors. However, unless the weights of the outer layer are positive and sum to 1,

the complexity is exponential in d . In [48], it was shown that for a randomly initialized neural network with sufficiently many hidden units, the generic gradient descent algorithm learns any low degree polynomial. Learning non-linear networks through multiple rounds of random initialization followed by arbitrary optimization steps was proposed in [49]. In [50], an efficiently learned kernel based estimator was shown to perform just as well as a class of deep neural networks. However, its ability to well-approximate general conditional mean regression functions is unclear.

The next section discusses an iterative procedure that reduces the complexity of finding the penalized least squares estimator (2.5).

2.4 Greedy algorithm

The main difficulty with constructing an estimator that satisfies (2.6) is that it involves a dm -dimensional optimization. Here, we outline a greedy approach that reduces the problem to performing m d -dimensional optimizations. This construction is based on the ℓ_1 -penalized greedy pursuit (LPGP) in [33], with the modification that the penalty can be a convex function of the candidate function complexity. Greedy strategies for approximating functions in the closure of the linear span of a subset of a Hilbert space has its origins in [51] and many of its statistical implications were studied in [38] and [33].

Let f^\star be a function, not necessarily in \mathcal{F} . Initialize $f_0 = 0$. For $m = 1, 2, \dots$, iteratively, given the terms of f_{m-1} as h_1, \dots, h_{m-1} and the coefficients of it as $\beta_{1,m-1}, \dots, \beta_{m-1,m-1}$, we proceed as follows. Let $f_m(x) = \sum_{j=1}^m \beta_{j,m} h_j(x) = \sum_{j=1}^m \beta_{j,m} \phi(\theta_{h_j} \cdot x)$, with the term h_m in \mathcal{H} chosen to come within a constant factor $c \geq 1$ of the maximum inner product with the residual $f^\star - f_{m-1}$; that is

$$\langle h_m, f^\star - f_{m-1} \rangle \geq \frac{1}{c} \sup_{h \in \mathcal{H}} \langle h, f^\star - f_{m-1} \rangle.$$

Define $f_m(x) = (1 - \alpha_m) f_{m-1}(x) + \beta_{m,m} h_m(x)$. Associated with this representation of f_m is the ℓ_1 norm of its coefficients $v_m = \sum_{j=1}^m |\beta_{j,m}| = (1 - \alpha_m) v_{m-1} + \beta_{m,m}$. The coefficients α_m and $\beta_{m,m}$ are chosen to minimize $\|f^\star - (1 - \alpha_m) f_{m-1} - \beta_{m,m} h_m\|^2 + \omega((1 - \alpha_m) v_{m-1} + \beta_{m,m})$.

In the empirical setting, with $R_i = Y_i - f_{m-1}(X_i)$, the high-dimensional optimization task is to find θ_m such that

$$\frac{1}{n} \sum_{i=1}^n R_i \phi(\theta_m \cdot X_i) \geq \frac{1}{c} \sup_{\theta} \frac{1}{n} \sum_{i=1}^n R_i \phi(\theta \cdot X_i)$$

The fact that one does not need to find the exact maximizer of the above empirical inner product, but only come to within a constant multiple of it, has important consequences. For example, in adaptive annealing, one begins by sampling from an initial distribution p_0 and then iteratively samples from a distribution proportional to $e^{t(\frac{1}{n} \sum_{i=1}^n R_i \phi(\theta \cdot X_i))} p_0(\theta)$, evolving according to $\theta_{t+h} = \theta_t - h G_t(\theta_t)$, where $G_t(\theta)$ satisfies $\nabla^T [G_t(\theta) p_t(\theta)] = \partial_t p_t(\theta)$. The mean of p_t is at least $\frac{1}{c} \sup_{\|\theta\|_1 \leq \Lambda} \frac{1}{n} \sum_{i=1}^n R_i \phi(\theta \cdot X_i)$ for sufficiently large t .

Theorem 1. *Suppose $w : \mathbb{R} \rightarrow \mathbb{R}$ is a real-valued non-negative convex function. If f_m is chosen according to the greedy scheme described previously, then*

$$\|f^* - f_m\|^2 + w(v_m) \leq \inf_{f \in \mathcal{F}} \left\{ \|f^* - f\|^2 + w(cv_f) + \frac{4b_f}{m} \right\}, \quad (2.15)$$

where $b_f = c^2 v_f^2 + 2v_f \|f^*\|(c+1) - \|f\|^2$. Furthermore, for all $\delta > 0$,

$$\begin{aligned} & \|f^* - f_m\|^2 + w(v_m) \\ & \leq \inf_{f \in \mathcal{F}} \inf_{\delta > 0} \left\{ (1+\delta) \|f^* - f\|^2 + w(cv_f) + \frac{4(1+\delta)\delta^{-1}(c+1)^2 v_f^2}{m} \right\}, \end{aligned} \quad (2.16)$$

and hence with $\delta = \frac{2(c+1)v_f}{\|f^* - f\|\sqrt{m}}$,

$$\|f^* - f_m\|^2 + w(v_m) \leq \inf_{f \in \mathcal{F}} \left\{ \left(\|f^* - f\| + \frac{2(c+1)v_f}{\sqrt{m}} \right)^2 + w(cv_f) \right\}.$$

Proof. Fix any f in the linear span \mathcal{F} , with the form $\sum_{h \in \mathcal{H}} \beta_h h$, with non-negative β_h and set

$$e_m = \|f^* - f_m\|^2 - \|f^* - f\|^2 + w(v_m).$$

From the definition of α_m and $\beta_{m,m}$ as minimizers of e_m for each h_m , and the convexity of

w ,

$$\begin{aligned}
e_m &= \|f^\star - (1 - \alpha_m)f_{m-1} - \beta_{m,m}h_m\|^2 - \|f^\star - f\|^2 + \\
&\quad w((1 - \alpha_m)v_{m-1} + \beta_{m,m}) \\
&\leq \|f^\star - (1 - \alpha_m)f_{m-1} - \alpha_m cv_f h_m\|^2 - \|f^\star - f\|^2 + \\
&\quad w((1 - \alpha_m)v_{m-1} + \alpha_m cv_f) \\
&\leq \|f^\star - (1 - \alpha_m)f_{m-1} - \alpha_m cv_f h_m\|^2 - \|f^\star - f\|^2 + \\
&\quad (1 - \alpha_m)w(v_{m-1}) + \alpha_m w(cv_f).
\end{aligned}$$

Now $\|f^\star - (1 - \alpha_m)f_{m-1} - \alpha_m cv_f h_m\|^2$ is equal to $\|(1 - \alpha_m)(f^\star - f_{m-1}) + \alpha_m(f^\star - ch_m v_f)\|^2$.

Expanding this quantity leads to

$$\begin{aligned}
\|f^\star - (1 - \alpha_m)f_{m-1} - \alpha_m cv_f h_m\|^2 &= (1 - \alpha_m)^2 \|f^\star - f_{m-1}\|^2 \\
&\quad - 2\alpha_m(1 - \alpha_m)\langle f^\star - f_{m-1}, ch_m v_f - f^\star \rangle \\
&\quad + \alpha_m^2 \|f^\star - ch_m v_f\|^2.
\end{aligned}$$

Next we add $(1 - \alpha_m)w(v_{m-1}) + \alpha_m w(cv_f) - \|f^\star - f\|^2$ to this expression to obtain

$$\begin{aligned}
e_m &\leq (1 - \alpha_m)e_{m-1} + \alpha_m^2 [\|f^\star - ch_m v_f\|^2 - \|f^\star - f\|^2] + \alpha_m w(cv_f) \\
&\quad - 2\alpha_m(1 - \alpha_m)\langle f^\star - f_{m-1}, ch_m v_f - f \rangle \\
&\quad + \alpha_m(1 - \alpha_m)[2\langle f^\star - f_{m-1}, f^\star - f \rangle - \|f^\star - f_{m-1}\|^2 - \|f^\star - f\|^2]. \quad (2.17)
\end{aligned}$$

The expression in brackets in (2.17) is equal to $- \|f - f_{m-1}\|^2$ and hence the entire quantity is further upper bounded by

$$\begin{aligned}
e_m &\leq (1 - \alpha_m)e_{m-1} + \alpha_m^2 [\|f^\star - ch_m v_f\|^2 - \|f^\star - f\|^2] + \alpha_m w(cv_f) \\
&\quad - 2\alpha_m(1 - \alpha_m)\langle f^\star - f_{m-1}, ch_m v_f - f \rangle.
\end{aligned}$$

Consider a random variable that equals h with probability β_h/v_f having mean f . Since a maximum is at least an average, the choice of h_m implies that $\langle f^\star - f_{m-1}, ch_m v_f \rangle$ is at least

$\langle f^\star - f_{m-1}, f \rangle$. This shows that e_m is no less than $(1 - \alpha_m)e_{m-1} + \alpha_m^2[\|f^\star - ch_mv_f\|^2 - \|f^\star - f\|^2] + \alpha_m w(cv_f)$. Expanding the squares in $\|f^\star - ch_mv_f\|^2 - \|f^\star - f\|^2$ and using the Cauchy-Schwarz inequality yields the bound $\|ch_mv_f\|^2 + 2\|f^\star\|(\|f - ch_mv_f\|) - \|f\|^2$. Since $\|h_m\| \leq \|h_m\|_\infty \leq 1$ and $\|f\| \leq v_f$, we find that $\|f^\star - ch_mv_f\|^2 - \|f^\star - f\|^2$ is at most $b_f = c^2 v_f^2 + 2v_f\|f^\star\|(c + 1) - \|f\|^2$. Hence we have shown that

$$e_1 \leq b_f + w(cv_f)$$

and

$$e_m \leq (1 - \alpha_m)e_{m-1} + \alpha_m^2 b_f + \alpha_m w(cv_f). \quad (2.18)$$

Because α is a minimizer of e_m , it can replace it by any value in $[0, 1]$ and the bound (2.18) holds verbatim. In particular, we can choose $\alpha_m = 2/(m + 1)$, $m \geq 2$ and use an inductive argument to establish (2.15). The second statement (2.16) follows from similar arguments upon consideration of

$$e_m = \|f^\star - f_m\|^2 - (1 + \delta)\|f^\star - f\|^2 + w(v_m),$$

together with the inequality $a^2 - (1 + \delta)b^2 \leq (1 + \delta)\delta^{-1}(a - b)^2$. \square

2.5 Risk bounds

2.5.1 Penalized estimators over the entire parameter space

Here we state our main theorem.

Theorem 2. *Let f^\star be a real-valued function on $[-1, 1]^d$ with finite variation v_{f^\star} with respect to the library $\mathcal{H} = \{h(x) = \phi(\theta \cdot x) : \|\theta\|_1 \leq v_0\}$. If \hat{f} is chosen to satisfy*

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}(X_i))^2 + \text{pen}_n(\hat{f})/n \leq \inf_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 + \text{pen}_n(f)/n \right\},$$

then for the truncated estimator $T\hat{f}$ and for $\text{pen}_n(f)$ depending on v_f as specified below, the

risk has the resolvability bound

$$\mathbb{E}\|T\hat{f} - f^\star\|^2 \leq (\tau + 1) \inf_{f \in \mathcal{F}} \{\|f - f^\star\|^2 + \mathbb{E}pen_n(f)/n\},$$

with penalties as described in (2.7), (2.8), (2.9), and (2.10). If \hat{f}_m is the LPGP estimator from the previous section, then by Theorem 1,

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}_m(X_i))^2 + w(v_{\hat{f}_m}) \leq \inf_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 + w(cv_f) + \frac{4b_f}{m} \right\},$$

where b_f is the empirical version of the same quantity in Theorem 1 and hence the risk has the resolvability bound

$$\mathbb{E}\|T\hat{f} - f^\star\|^2 \leq (\tau + 1) \inf_{f \in \mathcal{F}} \{\|f - f^\star\|^2 + \mathbb{E}pen_n(cf)/n + 4\mathbb{E}b_f/m\},$$

for a penalty, convex in v_f , $pen_n(f) = nw(v_f)$ as before. If m is chosen to be of order between \sqrt{n} and n so as to make the computational effects negligible, the previously described $L^2(P)$ rates for estimating f^\star in $L_{1,\mathcal{H}}$ via the truncated estimator $T\hat{f}_m$ are attainable under the appropriate penalties.

One can also extend these results to include penalties that depend on the number of terms m in an m -term greedy approximation \hat{f}_m to f^\star . We take \hat{f}_m to be an m term fit from an LPGP algorithm and choose \hat{m} among all $m \in \mathcal{M}$ (i.e. $\mathcal{M} = \{1, \dots, n\}$) to minimize

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}_m(X_i))^2 + pen_n(\hat{f}_m, m)/n.$$

This approach enables the use of a data-based stopping criterion for the greedy algorithm. For more details on these adaptive methods, we refer the reader to [33]. The resolvability risk bound allows also for interpolation rates between L_2 and $L_{1,\mathcal{H}}$ refining the results of [38] and in accordance with the best balance between error of approximation and penalty.

The target f^\star is not necessarily in \mathcal{F} . To each f in \mathcal{F} , there corresponds a function ρ ,

which assigns to (X, Y) the relative loss

$$\rho(X, Y) = \rho_f(X, Y) = (Y - f(X))^2 - (Y - f^*(X))^2.$$

Let $(\underline{X}', \underline{Y}')$ be an independent copy of the training data $(\underline{X}, \underline{Y})$ used for testing the efficacy of a fit \hat{f} based on $(\underline{X}, \underline{Y})$. The relative empirical loss with respect to the training data is denoted by $P_n(f||f^*) = \frac{1}{n} \sum_{i=1}^n \rho(X_i, Y_i)$ and that with respect to the independent copy is $P'_n(f||f^*) = \frac{1}{n} \sum_{i=1}^n \rho(X'_i, Y'_i)$. We define the empirical squared error on the training and test data by $D_n(f, \tilde{f}) = \frac{1}{n} \sum_{i=1}^n (f(X_i) - \tilde{f}(X_i))^2$ and $D'_n(f, \tilde{f}) = \frac{1}{n} \sum_{i=1}^n (f(X'_i) - \tilde{f}(X'_i))^2$ for all f, \tilde{f} in \mathcal{F} . Using the relationship $Y = f^*(X) + \varepsilon$, we note that $\rho(X, Y)$ can also be written as $(f(X) - f^*(X))^2 - 2\varepsilon(f(X) - f^*(X)) = g^2(X) - 2\varepsilon g(X)$, where $g(x) = f(x) - f^*(x)$. Hence we have the relationship $P_n(f||f^*) = D_n(f, f^*) - \frac{2}{n} \sum_{i=1}^n \varepsilon_i g(X_i)$.

The relative empirical loss $P'_n(\hat{f}||f^*)$ is an unbiased estimate of the risk $\mathbb{E}\|\hat{f} - f^*\|^2$. Since ε'_i has mean zero conditioned on X'_i , the mean of $P'_n(\hat{f}||f^*)$ with respect to $(\underline{X}', \underline{Y}')$ is $\|\hat{f} - f^*\|^2$. This quantity captures how well the fit \hat{f} based on the training data generalizes to a new set of observations. The goal is to control the empirical discrepancy $P'_n(f||f^*) - \tau P_n(f||f^*)$ between the loss on the future data and the loss on the training data for a constant $\tau > 1$. Toward this end, we seek a positive quantity $\text{pen}_n(f)$ to satisfy

$$\mathbb{E} \sup_{f \in \mathcal{F}} \{P'_n(f||f^*) - \tau P_n(f||f^*) - \tau \text{pen}_n(f)/n\} \leq 0,$$

Once such an inequality holds, the data-based choice \hat{f} in \mathcal{F} yields

$$\mathbb{E} P'_n(\hat{f}||f^*) \leq \tau \mathbb{E}[P_n(\hat{f}||f^*) + \text{pen}_n(f)/n].$$

If \hat{f} satisfies

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}(X_i))^2 + \frac{\text{pen}_n(\hat{f})}{n} \leq \inf_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 + \frac{\text{pen}_n(f)}{n} + A_f \right\}, \quad (2.19)$$

for some positive quantity A_f that decays to zero as the sample size grows, we see that

$$\mathbb{E}P'_n(\hat{f}||f^\star) \leq \tau \inf_{f \in \mathcal{F}} \mathbb{E}[P_n(f||f^\star) + \text{pen}_n(f)/n + A_f].$$

In our application, A_f is of the form $4b_f/m$ (as in (2.15)) and is made small with the number of greedy step m . Using $\mathbb{E}P'_n(\hat{f}||f^\star) = \mathbb{E}\|\hat{f} - f^\star\|^2$ and $\mathbb{E}P_n(f||f^\star) = \|f - f^\star\|^2$, the above expression is seen to be

$$\mathbb{E}\|\hat{f} - f^\star\|^2 \leq \tau \inf_{f \in \mathcal{F}} \{\|f - f^\star\|^2 + \mathbb{E}\text{pen}_n(f)/n + \mathbb{E}A_f\}. \quad (2.20)$$

For the purposes of proving results in the case when \mathcal{F} is uncountable, it is useful to consider complexities $L_n(\tilde{f})$ for \tilde{f} in a countable subset $\tilde{\mathcal{F}}$ of \mathcal{F} satisfying $\sum_{\tilde{f} \in \tilde{\mathcal{F}}} e^{-\gamma_n L_n(\tilde{f})} \leq 1$ for some $\gamma_n > 0$ and such that

$$\begin{aligned} & \sup_{f \in \mathcal{F}} \{P'_n(f||f^\star) - \tau P_n(f||f^\star) - \tau \text{pen}_n(f)/n\} \\ & \leq \sup_{\tilde{f} \in \tilde{\mathcal{F}}} \{P'_n(\tilde{f}||f^\star) - \tau P_n(\tilde{f}||f^\star) - \tau \gamma_n L_n(\tilde{f})/n\}, \end{aligned} \quad (2.21)$$

with

$$\mathbb{E} \sup_{\tilde{f} \in \tilde{\mathcal{F}}} \{P'_n(\tilde{f}||f^\star) - \tau P_n(\tilde{f}||f^\star) - \tau \gamma_n L_n(\tilde{f})/n\} \leq 0.$$

The condition in (2.21) is equivalent to requiring that

$$\sup_{f \in \mathcal{F}} \inf_{\tilde{f} \in \tilde{\mathcal{F}}} \{\Delta_n(f, \tilde{f}) + \gamma_n L_n(\tilde{f}) - \text{pen}_n(f)\} \leq 0,$$

where

$$\Delta_n(f, \tilde{f}) = n[P_n(\tilde{f}||f^\star) - P_n(f||f^\star)] - (n/\tau)[P'_n(\tilde{f}||f^\star) - P'_n(f||f^\star)].$$

If we truncate the penalized least squares estimator \hat{f} at a certain level B_n , for $\mathbb{E}\|T\hat{f} - f^\star\|^2$ to maintain the resolvability bound $\tau \inf_{f \in \mathcal{F}} \{\|f - f^\star\|^2 + \mathbb{E}\text{pen}_n(f)/n + \mathbb{E}A_f\}$, we require

that

$$\sup_{f \in \mathcal{F}} \inf_{\tilde{f} \in \tilde{\mathcal{F}}} \{ \Delta_n(f, \tilde{f}) + \gamma_n L_n(\tilde{f}) - \text{pen}_n(f) \} \leq 0,$$

where

$$\Delta_n(f, \tilde{f}) = n[P_n(T\tilde{f}||f^\star) - P_n(f||f^\star)] - (n/\tau)[P'_n(T\tilde{f}||f^\star) - P'_n(Tf||f^\star)].$$

Rather than working with the relative empirical loss $P'_n(Tf||f^\star)$, we prefer to work with $D'_n(Tf, f^\star)$. These two quantities are related to each other, provided $\frac{1}{n} \sum_{i=1}^n \varepsilon_i g(X'_i)$ is small and they are exactly equal in the no noise case. Hence we would like to determine penalties that ensure

$$\mathbb{E} \sup_{f \in \mathcal{F}} \{ D'_n(Tf, f^\star) - \tau P_n(f||f^\star) - \tau \text{pen}_n(f)/n \} \leq 0.$$

Suppose we require that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \{ \tau_1^{-1} D'_n(Tf, f^\star) - \tau P_n(f||f^\star) - \tau \text{pen}_n(f)/n \} \leq 0,$$

for some $\tau_1 \geq 1$. This further inflates the resulting risk bound by τ_1 so that the factor τ is replaced with $\tau\tau_1$ in (2.20). However, it enables us to create countable covers $\tilde{\mathcal{F}}$ with smaller errors in approximating functions from \mathcal{F} . To see this, suppose the countable cover $\tilde{\mathcal{F}}$ satisfies

$$\begin{aligned} & \sup_{f \in \mathcal{F}} \{ \tau_1^{-1} D'_n(Tf, f^\star) - \tau P_n(f||f^\star) - \tau \text{pen}_n(f)/n \} \\ & \leq \sup_{\tilde{f} \in \tilde{\mathcal{F}}} \left\{ D'_n(T\tilde{f}, f^\star) - \tau P_n(T\tilde{f}||f^\star) - \tau \gamma_n L_n(\tilde{f})/n \right\}, \end{aligned}$$

or equivalently that

$$\sup_{f \in \mathcal{F}} \inf_{\tilde{f} \in \tilde{\mathcal{F}}} \left\{ \Delta_n(f, \tilde{f}) + \gamma_n L_n(\tilde{f}) - \text{pen}_n(f) \right\} \leq 0,$$

where

$$\begin{aligned}\Delta_n(f, \tilde{f}) &= n[P_n(T\tilde{f}||f^\star) - P_n(f||f^\star)] + \\ &\quad n\tau^{-1}[\tau_1^{-1}D'_n(Tf, f^\star) - D'_n(T\tilde{f}, f^\star)].\end{aligned}$$

We set $\tau_1 = 1/\tau + 1$. Using the inequality, $\tau^{-1}a^2 - b^2 \leq \frac{1}{\tau-1}(b-a)^2$ that can be derived from $(a/\sqrt{\tau} - b\sqrt{\tau})^2 \geq 0$, we can upper bound the difference $\tau_1^{-1}D'_n(Tf, f^\star) - D'_n(T\tilde{f}, f^\star)$ by

$$(\tau_1 - 1)^{-1}D'_n(Tf, T\tilde{f}).$$

This quantity does not involve f^\star , which is desirable for the proceeding analysis. Hence $\Delta_n(f, \tilde{f})$ is not greater than

$$n[P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) + D'_n(Tf, T\tilde{f})].$$

and thus we seek a penalty $\text{pen}_n(f)$ that is at least

$$\inf_{\tilde{f} \in \tilde{\mathcal{F}}} \{\gamma_n L_n(\tilde{f}) + n[P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) + D'_n(Tf, T\tilde{f})]\}.$$

An estimator \hat{f} satisfying (2.19) with penalty $\text{pen}_n(f)$ that is at least

$$\inf_{\tilde{f} \in \tilde{\mathcal{F}}} \{\gamma_n L_n(\tilde{f}) + n[P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) + D'_n(Tf, T\tilde{f})]\}$$

satisfies the risk bound

$$\mathbb{E}\|T\hat{f} - f^\star\|^2 \leq (\tau + 1) \inf_{f \in \mathcal{F}} \{\|f - f^\star\|^2 + \mathbb{E}\text{pen}_n(f)/n + \mathbb{E}A_f\}.$$

By bounding the distortion in this way, we eliminate some error in approximating f by \tilde{f} that arises from analyzing $P_n(T\tilde{f}||f^\star) - P_n(f||f^\star)$ and $D'_n(T\tilde{f}, f^\star) - D'_n(Tf, f^\star)$. The next result in Theorem 3 summarizes what we have found so far.

Theorem 3. Suppose $\tilde{\mathcal{F}}$ is a countable collection of functions that satisfies

$$\mathbb{E} \sup_{\tilde{f} \in \tilde{\mathcal{F}}} \left\{ D'_n(T\tilde{f}, f^\star) - \tau P_n(\tilde{f}||f^\star) - \tau \gamma_n L_n(\tilde{f}) \right\} \leq 0.$$

If $\text{pen}_n(f)$ is at least

$$\inf_{\tilde{f} \in \tilde{\mathcal{F}}} \{ \gamma_n L_n(\tilde{f}) + n[P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) + D'_n(Tf, T\tilde{f})] \},$$

then the truncated estimator $T\hat{f}$ with \hat{f} satisfying (2.19) has the resolvability bound

$$\mathbb{E} \|T\hat{f} - f^\star\|^2 \leq (\tau + 1) \inf_{f \in \mathcal{F}} \{ \|f - f^\star\|^2 + \mathbb{E} \text{pen}_n(f)/n + \mathbb{E} A_f \}.$$

Furthermore, if $\tilde{\mathcal{F}} = \mathcal{F}$, and $\text{pen}_n(f)$ is at least $\gamma_n L_n(f) + P_n(Tf||f^\star) - P_n(f||f^\star)$, the truncated estimator $T\hat{f}$ with \hat{f} satisfying (2.19) has the resolvability bound

$$\mathbb{E} \|T\hat{f} - f^\star\|^2 \leq \tau \inf_{f \in \mathcal{F}} \{ \|f - f^\star\|^2 + \mathbb{E} \text{pen}_n(f)/n + \mathbb{E} A_f \}.$$

The main task is to construct the countable collection $\tilde{\mathcal{F}}$ and find a suitable upper bound on

$$\inf_{\tilde{f} \in \tilde{\mathcal{F}}} \{ \gamma_n L_n(\tilde{f}) + n[P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) + D'_n(Tf, T\tilde{f})] \}. \quad (2.22)$$

Here we outline a general strategy to obtain countable covers $\tilde{\mathcal{F}}$ of a given collection \mathcal{F} :

1. Given a function $f = \sum_h \beta_h h$ in \mathcal{F} , use the Jones-Barron probabilistic method to obtain an equally weighted, sparse linear combination of dictionary elements from \mathcal{H} , $\tilde{g} = \frac{v}{m} \sum_{i=1}^m h_i$, such that $P_n(\tilde{g}||f^\star) - P_n(f||f^\star) + D'_n(\tilde{g}, f)$ is small.
2. Construct a finite cover of \mathcal{H} , say $\tilde{\mathcal{H}}$, replace each h_i by an approximant \tilde{h}_i , and obtain $\tilde{f} = \frac{v}{m} \sum_{i=1}^m \tilde{h}_i$ such that $D(\tilde{g}, \tilde{f})$ and $D'_n(\tilde{g}, \tilde{f})$ are small. Finally, take \mathcal{F} to be all functions of the form $\frac{v}{m} \sum_{i=1}^m \tilde{h}_i$, for which there are finitely many.

Remark 1. Importantly, covers obtained from the above strategy do not depend on the empirical probability measure (i.e., depend on the data). Indeed, the individual representers

$\frac{v}{m} \sum_{i=1}^m \tilde{h}_i$ may be data-dependent, but they belong to a (data-independent) collection that is essentially an enumeration of all possible types.

This next lemma tells us how to use these approximants to bound (2.22).

Lemma 1. *For every \tilde{g} , \tilde{f} , and f ,*

$$\begin{aligned} & P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) + D'_n(Tf, T\tilde{f}) \\ & \leq P_n(\tilde{g}||f^\star) - P_n(f||f^\star) + D'_n(\tilde{g}, f) + 4B_n \left[\sqrt{D(\tilde{g}, \tilde{f})} + \sqrt{D'_n(\tilde{g}, \tilde{f})} \right] + \frac{T_n}{n}. \end{aligned}$$

Proof. By Lemma 9 (I) and (II),

$$\begin{aligned} (y - T\tilde{f}(x))^2 - (y - f(x))^2 &= [(y - \tilde{g}(x))^2 - (y - f(x))^2] + \\ & \quad [(y - T\tilde{f}(x))^2 - (y - T\tilde{g}(x))^2] + \\ & \quad [(y - T\tilde{g}(x))^2 - (y - \tilde{g}(x))^2] \\ &\leq [(y - \tilde{g}(x))^2 - (y - f(x))^2] + \\ & \quad 4B_n |\tilde{g}(x) - \tilde{f}(x)| + \\ & \quad 4B_n (|y| - B_n) \mathbb{I}\{|y| > B_n\} + \\ & \quad 2(|y| - B_n)^2 \mathbb{I}\{|y| > B_n\} \\ &= [(y - \tilde{g}(x))^2 - (y - f(x))^2] + \\ & \quad 4B_n |\tilde{g}(x) - \tilde{f}(x)| + \\ & \quad 2(|y|^2 - B_n^2) \mathbb{I}\{|y| > B_n\}. \end{aligned}$$

Summing over this inequality at the data points, we have

$$P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) \leq P_n(\tilde{g}||f^\star) - P_n(f||f^\star) + 4B_n \sqrt{D(\tilde{g}, \tilde{f})} + \frac{T_n}{n}.$$

By Lemma 9 (III),

$$(T\tilde{f}(x') - Tf(x'))^2 \leq (f(x') - \tilde{g}(x'))^2 + 4B_n |\tilde{f}(x') - \tilde{g}(x')|.$$

Again, summing over this inequality at the data points, we have

$$D'_n(Tf, T\tilde{f}) \leq D'_n(\tilde{g}, f) + 4B_n \sqrt{D'_n(\tilde{g}, \tilde{f})}.$$

□

Recall that g is equal to $f - f^*$. In this way, there is a one to one correspondence between f and g . To simplify notation, we sometimes write $D_n(f, f^*)$ as $D_n(g)$ and $D'_n(f, f^*)$ as $D'_n(g)$. Moreover, assume an analogous notation holds for the relative loss functions $P_n(f||f^*)$ and $P'(f||f^*)$ and complexities $L_n(f)$.

Theorem 4. *If \mathcal{F} is a countable collection of functions bounded in magnitude by B_n and $L_n(f)$ satisfies the Kraft inequality $\sum_{f \in \mathcal{F}} e^{-L_n(f)} \leq 1$, then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \{D'_n(f||f^*) - \tau P_n(f||f^*) - \tau \gamma_n L_n(f)/n\} \leq 0,$$

where $\tau = (1 + \delta_1)(1 + \delta_2)$ and $\gamma_n = (2\tau)^{-1}(1 + \delta_1/2)(1 + 2/\delta_1)(B + B_n)^2 + 2(1 + 1/\delta_2)\sigma^2 + 2(B + B_n)\eta$.

Proof. Let $s^2(g)$ be as in Lemma 2. Since g^2 is non-negative, $s^2(g) \leq D'_n(g^2) + D_n(g^2)$. Moreover, since $|f| \leq B_n$ and $|f^*| \leq B$, it follows that $s^2(g) \leq (B + B_n)^2(D'_n(g) + D_n(g))$. Let $\gamma_1 = A_1(B + B_n)^2/2$ with A_1 to be specified later. By Lemma 2, we have

$$\mathbb{E} \sup_{g \in \mathcal{G}} \left\{ (1 - 1/A_1)D'_n(g) - (1 + 1/A_1)D_n(g) - \frac{\gamma_1}{n}L(g) \right\} \quad (2.23)$$

$$\leq \mathbb{E} \sup_{g \in \mathcal{G}} \left\{ D'_n(g) - D_n(g) - \frac{\gamma_1}{n}L(g) - \frac{1}{2\gamma_1}s^2(g) \right\} \leq 0 \quad (2.24)$$

By Lemma 3, we also know that

$$\mathbb{E} \sup_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(X_i) - \frac{\gamma_2}{n}L(g) - \frac{1}{A_2 n}D_n(g) \right\} \leq 0, \quad (2.25)$$

where $\gamma_2 = A_2\sigma^2/2 + (B + B_n)\eta$. Adding the expression in (2.42) to $2a > 0$ times the expression in (2.43) and collecting terms, we find that $1 + 1/A_1 + 2a/A_2$ should be equal to a in order for $D_n(g)$ and $\frac{1}{n} \sum_{i=1}^n \varepsilon_i g(X_i)$ to be added together to produce $P_n(g)$. Thus we

find that

$$\mathbb{E} \sup_{g \in \mathcal{G}} \left\{ (1 - 1/A_1) D'_n(g) - a(P_n(g) + \frac{\gamma_n}{n} L(g)) \right\} \leq 0,$$

where $\gamma_n = \gamma_1/a + 2\gamma_2$. Choosing $A_1 = 1 + 2/\delta_1$, $A_2 = 2(1 + 1/\delta_2)$, and $\tau = (1 + \delta_1)(1 + \delta_2)$, we find that $a = \tau(1 - 1/A_1)$. Dividing the resulting expression by $1 - 1/A_1$ produces

$$\mathbb{E} \sup_{g \in \mathcal{G}} \left\{ D'_n(g) - \tau P_n(g) - \tau \gamma_n L(g)/n \right\} \leq 0.$$

□

In general, the penalty should not depend on the unknown test data \underline{X}' . However if one seeks to describe the error of a fit \hat{f} trained with the data $(\underline{X}, \underline{Y})$ at new data points \underline{X}' , a penalty that depends on \underline{X}' is natural. Also it is analogous to the trans-inductive setting in machine learning [52].

In deriving our variable complexity covers, we use empirical L^2 covers of certain sizes of the dictionary \mathcal{H} developed in lemmas in Section 2.8. Under the conditions on the class \mathcal{H} , these covers will not depend on the data. Here we show how these covers can be used to build covers of the class of function $f = \sum_h \beta_h h$.

Theorem 5. *Let $f = \sum_h \beta_h h$. Let $\tilde{\mathcal{H}}_1$ be an empirical L^2 ϵ_1 -net for \mathcal{H} of cardinality M_1 . Let $\tilde{\mathcal{H}}_2$ be an empirical L^2 ϵ_2 -net for \mathcal{H} of cardinality M_2 . Suppose these empirical covers do not depend on the underlying data. For every integer $m_0 \geq 1$, there exists a subset $\tilde{\mathcal{F}}$ of \mathcal{F} with cardinality at most $\binom{M_2 + M_1 + m_0}{M_1 + m_0}$ such that for $v \geq v_f$ and $\tilde{v} = v(1 + M_1/m_0)$, if ϕ satisfies Assumption 1,*

$$P_n(T\tilde{f}||f^*) - P_n(f||f^*) + D'_n(Tf, T\tilde{f}) \leq \frac{2\tilde{v}^2 \epsilon_1^2}{m_0} + \frac{\tilde{v}^2 M_1}{2m_0^2} + 8B_n \tilde{v} \epsilon_2 + \frac{T_n}{n}, \quad (2.26)$$

for some \tilde{f} in $\tilde{\mathcal{F}}$.

If ϕ satisfies Assumption 1, there exists a subset $\tilde{\mathcal{F}}$ of \mathcal{F} with cardinality at most $\binom{M_2 + m_0}{m_0}$ such that

$$P_n(T\tilde{f}||f^*) - P_n(f||f^*) + D'_n(Tf, T\tilde{f}) \leq \frac{2vv_f}{m_0} + 8B_n v \epsilon_2 + \frac{T_n}{n}, \quad (2.27)$$

for some \tilde{f} in $\tilde{\mathcal{F}}$.

If ϕ satisfies Assumption 2, then there exists a subset $\tilde{\mathcal{F}}$ of \mathcal{F} with cardinality at most $\binom{2^{(2d+m_0)}+m_1}{m_0 \atop m_1}$ such that

$$\begin{aligned} P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) + D'_n(Tf, T\tilde{f}) &\leq \frac{2vv_f}{m_1} + \frac{L_2v_f^2v_0^2}{m_0} + \frac{L_2^2v_f^2v_0^4}{4m_0} \\ &\quad + \left(\frac{1}{n} \sum_{i=1}^n |Y_i| \right) \frac{v_f L_2 v_0^2}{m_0} + \frac{T_n}{n}. \end{aligned} \quad (2.28)$$

for some \tilde{f} in $\tilde{\mathcal{F}}$.

Proof. We first prove (2.26) and (2.27). Let $\tilde{g} = f_m = (v/m_0) \sum_{k=1}^m h_k$ be as in (2.35) of Lemma 5. Then, using the empirical L^2 norm, we have that

$$P_n(\tilde{g}||f^\star) - P_n(f||f^\star) \leq \frac{\tilde{v}^2 \epsilon_1^2}{m_0} + \frac{\tilde{v}^2 M_1}{2m_0^2},$$

and

$$D'_n(\tilde{g}, f) \leq \frac{\tilde{v}^2 \epsilon_1^2}{m_0}.$$

Since $\tilde{\mathcal{H}}_2$ is an empirical L^2 ϵ_2 -net for \mathcal{H} , for each h_k there is an \tilde{h}_k in $\tilde{\mathcal{H}}_2$ such that $\frac{1}{n} \sum_{i=1}^n |h_k(x_i) - \tilde{h}_k(x_i)|^2$ and $\frac{1}{n} \sum_{i=1}^n |h_k(x'_i) - \tilde{h}_k(x'_i)|^2$ are less than ϵ_2^2 . Set $\tilde{f} = (v/m_0) \sum_{k=1}^m \tilde{h}_k$ and define $\tilde{\mathcal{F}}$ to be the collection of all such functions. Thus, it follows from Jensen's inequality that $D(\tilde{g}, \tilde{f})$ and $D'(\tilde{g}, \tilde{f})$ are less than $v^2 \epsilon_2^2$. Putting all these together, we have that from Lemma 1,

$$\begin{aligned} &P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) + D'_n(Tf, T\tilde{f}) \\ &\leq P_n(\tilde{g}||f^\star) - P_n(f||f^\star) + D'_n(\tilde{g}, f) + 4B_n \left[\sqrt{D(\tilde{g}, \tilde{f})} + \sqrt{D'_n(\tilde{g}, \tilde{f})} \right] + \frac{T_n}{n} \\ &\leq \left(\frac{\tilde{v}^2 \epsilon_1^2}{m_0} + \frac{\tilde{v}^2 M_1}{2m_0^2} \right) + \frac{\tilde{v}^2 \epsilon_1^2}{m_0} + 4B_n(v\epsilon_2 + v\epsilon_2) + \frac{T_n}{n}. \end{aligned}$$

The conclusion about the cardinality of $\tilde{\mathcal{F}}$ follows from Lemma 11. The bound in (2.27) is

obtained in a similar way, but this time we use Lemma 4, which yields

$$\begin{aligned}
& P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) + D'_n(Tf, T\tilde{f}) \\
& \leq P_n(\tilde{g}||f^\star) - P_n(f||f^\star) + D'_n(\tilde{g}, f) + 4B_n \left[\sqrt{D(\tilde{g}, \tilde{f})} + \sqrt{D'_n(\tilde{g}, \tilde{f})} \right] + \frac{T_n}{n} \\
& \leq \frac{vv_f}{m_0} + \frac{vv_f}{m_0} + 4B_n(v\epsilon_2 + v\epsilon_2) + \frac{T_n}{n}.
\end{aligned}$$

To prove (2.28), we use Lemma 6 and take $\tilde{g} = \tilde{f}$ so that

$$\begin{aligned}
& P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) + D'_n(Tf, T\tilde{f}) \\
& \leq P_n(\tilde{g}||f^\star) - P_n(f||f^\star) + D'_n(\tilde{g}, f) + 4B_n \left[\sqrt{D(\tilde{g}, \tilde{f})} + \sqrt{D'_n(\tilde{g}, \tilde{f})} \right] + \frac{T_n}{n} \\
& = P_n(\tilde{g}||f^\star) - P_n(f||f^\star) + D'_n(\tilde{g}, f) + \frac{T_n}{n} \\
& \leq \left(\frac{vv_f}{m_1} + \frac{L_2 v_f (\frac{1}{n} \sum_{i=1}^n |Y_i| + v_f) v_0^2}{m_0} \right) + \left(\frac{vv_f}{m_1} + \frac{L_2^2 v_f^2 v_0^4}{4m_0^2} \right).
\end{aligned}$$

Let $\tilde{\mathcal{F}}$ be the collection of all such functions \tilde{f} . The bound on the cardinality of $\tilde{\mathcal{F}}$ follows also from Lemma 6. \square

According to Theorem 3 and Theorem 4, a valid penalty is at least

$$\gamma_n L_n(\tilde{f}) + n[P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) + D'_n(Tf, T\tilde{f})],$$

where \tilde{f} belongs to a countable set $\tilde{\mathcal{F}}$ satisfying $\sum_{\tilde{f} \in \tilde{\mathcal{F}}} e^{-L_n(\tilde{f})} \leq 1$. The constant γ_n is as prescribed in Theorem 4. By Theorem 5, there is a set $\tilde{\mathcal{F}}$ with cardinality at most $\binom{M_2+M_1+m_0}{M_1+m_0}$ such that for all f with $v_f \leq v$, there is a \tilde{f} in $\tilde{\mathcal{F}}$ such that $P_n(T\tilde{f}||f^\star) - P_n(f||f^\star) + D'_n(Tf, T\tilde{f})$ is bounded by

$$\frac{2\tilde{v}^2 \epsilon_1^2}{m_0} + \frac{\tilde{v}^2 M_1}{2m_0^2} + 8B_n \tilde{v} \epsilon_2 + \frac{T_n}{n}.$$

Using the fact that the logarithm of $\binom{M_2+M_1+m_0}{M_1+m_0}$ is bounded by $(M_1 + m_0) \log(e(M_2/M_1 +$

1)), a valid penalty divided by sample size is at least

$$\frac{\gamma_n}{n}(M_1 + m_0) \log(e(M_2/M_1 + 1)) + \frac{2\tilde{v}^2\epsilon_1^2}{m_0} + \frac{\tilde{v}^2 M_1}{2m_0^2} + 8B_n \tilde{v}\epsilon_2 + \frac{T_n}{n}. \quad (2.29)$$

Alternatively, there is a set $\tilde{\mathcal{F}}$ with cardinality at most $\binom{M_2+m_0}{m_0}$ such that for all f with $v_f \leq v$, there is a \tilde{f} in $\tilde{\mathcal{F}}$ such that $P_n(T\tilde{f}||f^*) - P_n(f||f^*) + D'_n(Tf, T\tilde{f})$ is bounded by

$$\frac{2vv_f}{m_0} + 8B_n v\epsilon_2 + \frac{T_n}{n}$$

and hence a valid penalty divided by sample size is at least

$$\frac{\gamma_n m_0 \log M_2}{n} + \frac{2vv_f}{m_0} + 8B_n v\epsilon_2 + \frac{T_n}{n}. \quad (2.30)$$

Analogously, if ϕ satisfies Assumption 2, a valid penalty divided by sample size is at least

$$\frac{5m_0 m_1 \log(d+1)}{n} + \frac{2vv_f}{m_1} + \frac{L_2 v_f^2 v_0^2}{m_0} + \frac{L_2^2 v_f^2 v_0^4}{4m_0^2} + \left(\frac{1}{n} \sum_{i=1}^n |Y_i| \right) \frac{v_f L_2 v_0^2}{m_0} + \frac{T_n}{n}. \quad (2.31)$$

for some \tilde{f} in $\tilde{\mathcal{F}}$.

We now discuss how m_0 , m_1 , ϵ_1 , and ϵ_2 should be chosen to produce penalties that yield optimal risk properties for $T\hat{f}$.

2.6 Risk bounds in high dimensions

2.6.1 Penalty under Assumption 1

By Lemma 7, an empirical L^2 ϵ_2 -cover of \mathcal{H} has cardinality less than $\binom{2d + \lceil (v_0/\epsilon_2)^2 \rceil}{\lceil (v_0/\epsilon_2)^2 \rceil}$. The logarithm of $\binom{2d + \lceil (v_0/\epsilon_2)^2 \rceil}{\lceil (v_0/\epsilon_2)^2 \rceil}$ is bounded by $4(v_0/\epsilon_2)^2 \log(d+1)$.

Continuing from the expression (2.30), we find that $\text{pen}_n(f)/n$ is at least

$$\frac{4\gamma_n m_0 (v_0/\epsilon_2)^2 \log(d+1)}{n} + \frac{2vv_f}{m_0} + 8B_n v\epsilon_2 + \frac{T_n}{n}.$$

Choosing m_0 to be the ceiling of $\left(\frac{vv_f n \epsilon_2^2}{2\gamma_n v_0^2 \log(d+1)}\right)^{1/2}$, we see that $\text{pen}_n(f)/n$ must be at least

$$\frac{8\gamma_n v_0^2 \log(d+1)}{n \epsilon_2^2} + 8 \left(\frac{vv_f \gamma_n v_0^2 \log(d+1)}{n \epsilon_2^2} \right)^{1/2} + 8B_n v \epsilon_2 + \frac{T_n}{n}.$$

Finally, we set $v = v_f$ and $\epsilon_2 = \left(\frac{\gamma_n v_0^2 \log(d+1)}{n B_n^2}\right)^{1/4}$ so that $\text{pen}_n(f)/n$ must be at least

$$16v_f \left(\frac{\gamma_n B_n^2 v_0^2 \log(d+1)}{n} \right)^{1/4} + 8 \left(\frac{\gamma_n B_n^2 v_0^2 \log(d+1)}{n} \right)^{1/2} + \frac{T_n}{n}.$$

We see that the main term in the penalty divided by sample size is

$$16v_f \left(\frac{\gamma_n B_n^2 v_0^2 \log(d+1)}{n} \right)^{1/4}.$$

2.6.2 Penalty under Assumption 2

Looking at (2.31) suggests that we choose m_0 to be the floor of $v_0^2 m_1$ which results in a penalty divided by sample size of at least

$$\frac{5\gamma_n v_0^2 m_1^2 \log(d+1)}{n} + \frac{2v_f^2}{m_1} + \frac{L_2 v_f^2}{m_1} + \frac{L_2^2 v_f^2}{4m_1^2} + \left(\frac{1}{n} \sum_{i=1}^n |Y_i| \right) \frac{v_f L_2}{m_1} + \frac{T_n}{n},$$

with leading terms of order

$$\frac{\gamma_n v_0^2 m_1^2 \log(d+1)}{n} + \frac{v_f^2}{m_1}.$$

Choosing m_1 to be the floor of $\left(\frac{v_f^2 n}{\gamma_n v_0^2 \log(d+1)}\right)^{1/3}$ yields the conclusion that a valid penalty divided by sample size is at least of order

$$v_f^{4/3} \left(\frac{\gamma_n v_0^2 \log(d+1)}{n} \right)^{1/3} + v_f \left(\frac{1}{n} \sum_{i=1}^n |Y_i| \right) \left(\frac{\gamma_n v_0^2 \log(d+1)}{n} \right)^{1/3} + \frac{T_n}{n}.$$

2.7 Risk bounds with improved exponents for moderate dimensions

Continuing from the expression (2.29), we find that $\text{pen}_n(f)/n$ is at least

$$\frac{\gamma_n}{n}(M_1 + m_0) \log(e(M_2/M_1 + 1)) + \frac{2\tilde{v}^2\epsilon_1^2}{m_0} + \frac{\tilde{v}^2 M_1}{2m_0^2} + 8B_n\tilde{v}\epsilon_2 + \frac{T_n}{n}.$$

Note that we can bound B_n^2 by γ_n by choosing δ_1 and δ_2 appropriately. For the precise definition of γ_n , see Theorem 4. The strategy for optimization is to first consider the terms

$$\frac{\gamma_n}{n}m_0 \log(e(M_2/M_1 + 1)) + \frac{2\tilde{v}^2\epsilon_1^2}{m_0} + 8\sqrt{\gamma_n}\tilde{v}\epsilon_2. \quad (2.32)$$

After m_0 , M_1 , and M_2 have been selected, we then check that

$$\frac{\gamma_n}{n}M_1 \log(e(M_2/M_1 + 1)) + \frac{\tilde{v}^2 M_1}{2m_0^2} \quad (2.33)$$

is relatively negligible. Choosing m_0 to be the ceiling of $\left(\frac{2\tilde{v}^2 n \epsilon_1^2}{\gamma_n \log(e(M_2/M_1 + 1))}\right)^{1/2}$, we see that (2.32) is at most

$$\frac{\gamma_n}{n} \log(e(M_2/M_1 + 1)) + 4 \left(\frac{\tilde{v}^2 \gamma_n \epsilon_1^2 \log(e(M_2/M_1 + 1))}{n} \right)^{1/2} + 8\sqrt{\gamma_n}\tilde{v}\epsilon_2.$$

Note that an empirical L^2 ϵ -cover of \mathcal{H} has cardinality between $(v_0/\epsilon)^d$ and $(2v_0/\epsilon + 1)^d \leq (3v_0/\epsilon)^d$ whenever $\epsilon \leq v_0$. Thus $M_2/M_1 \leq (3\epsilon_1/\epsilon_2)^d$ whenever $\epsilon_2 \leq v_0$ and hence

$$\log(e(M_2/M_1 + 1)) \leq 1 + (d/2) \log(9\epsilon_1^2/\epsilon_2^2 + 1) \leq d \log(9\epsilon_1^2/\epsilon_2^2 + 1),$$

whenever $\epsilon_1^2 \geq \epsilon_2^2(e-1)/9$. These inequalities imply that (2.32) is at most

$$\frac{d\gamma_n \log(9\epsilon_1^2/\epsilon_2^2 + 1)}{n} + 4 \left(\frac{\tilde{v}^2 \epsilon_1^2 d \gamma_n \log(9\epsilon_1^2/\epsilon_2^2 + 1)}{n} \right)^{1/2} + 8\sqrt{\gamma_n}\tilde{v}\epsilon_2.$$

Next, set

$$\epsilon_2^2 = \frac{9d\epsilon_1^2}{n}.$$

This means that the assumption $\epsilon_1^2 \geq \epsilon_2^2(e-1)/9$ is valid provided $d \leq n/(e-1)$. Thus (2.32) is at most

$$\frac{d\gamma_n \log(n/d+1)}{n} + 20\epsilon_1 \tilde{v} \sqrt{\frac{d\gamma_n \log(n/d+1)}{n}}.$$

Next, we add in the terms from (2.33). The selections of m_0 and ϵ_1 make (2.33) at most

$$\frac{M_1 d\gamma_n \log(n/d+1)}{n} + \frac{M_1 d\gamma_n \log(n/d+1)}{n\epsilon_1^2}$$

Since $M_1 \leq (3v_0/\epsilon_1)^d$ whenever $\epsilon_1 \leq v_0$, we find that (2.33) is at most

$$\frac{(3v_0)^d d\gamma_n \log(n/d+1)}{n\epsilon_1^d} + \frac{(3v_0)^d d\gamma_n \log(n/d+1)}{n\epsilon_1^{d+2}}$$

Let $\epsilon_1 = 3v_0 \left(\frac{d\gamma_n \log(n/d+1)}{n} \right)^{1/(2(d+3))}$. Choosing $\tilde{v} = v_f$, we see that a valid penalty divided by sample size is at least

$$\begin{aligned} 60v_f v_0 \left(\frac{d\gamma_n \log(n/d+1)}{n} \right)^{1/2+1/(2(d+3))} &+ \frac{1}{v_0^2} \left(\frac{d\gamma_n \log(n/d+1)}{n} \right)^{1/2+1/(2(d+3))} \\ &+ \left(\frac{d\gamma_n \log(n/d+1)}{n} \right)^{1/2+3/(2(d+3))} + \frac{d\gamma_n \log(n/d+1)}{n} + \frac{T_n}{n}. \end{aligned}$$

Note that for the form of the above penalty to be valid, we need $\frac{d\gamma_n \log(n/d)}{n}$ to be small enough to ensure that ϵ_1 and ϵ_2 are both less than v_0 .

2.7.1 Penalized estimators over a discretization of the parameter space

In the case that $\mathcal{F} = \tilde{\mathcal{F}}$, it follows from Theorem 3 that a valid penalty is at least $\gamma_n L_n(f) + P_n(Tf||f^*) - P_n(f||f^*)$. By Lemma 9 (I), we have that $P_n(Tf||f^*) - P_n(f||f^*) \leq T_n$. Hence a valid penalty is at least $\gamma_n L_n(f) + T_n$, where γ_n is as prescribed in Theorem 4. Suppose $\tilde{\mathcal{F}} = \tilde{\mathcal{F}}(\epsilon, v)$ is an $L^2(P)$ ϵ -net of $L_{1,\mathcal{H}}$ for functions f with variation v_f at most v . We

choose $L_n(f) = \log |\mathcal{F}_\epsilon|$. Then

$$\begin{aligned}\mathbb{E}\|T\hat{f} - f^\star\|^2 &\leq \tau \inf_{f \in \mathcal{F}_\epsilon} \left\{ \|f - f^\star\|^2 + \frac{\gamma_n \log |\tilde{\mathcal{F}}(\epsilon, v_f)|}{n} + \mathbb{E} \left[\frac{T_n}{n} \right] \right\} \\ &\leq \tau \inf_{\epsilon > 0} \left\{ \epsilon^2 + \frac{\gamma_n \log |\tilde{\mathcal{F}}(\epsilon, v_{f^\star})|}{n} + \mathbb{E} \left[\frac{T_n}{n} \right] \right\}.\end{aligned}$$

By Theorem in [53], there exists a universal constant $C > 0$ such that $\log |\tilde{\mathcal{F}}(\epsilon, v)| \leq Cd(vv_0)^{\frac{2d}{d+2}} \epsilon^{-\frac{2d}{d+2}}$. Hence,

$$\begin{aligned}\mathbb{E}\|T\hat{f} - f^\star\|^2 &\leq \tau \inf_{f \in \mathcal{F}_\epsilon} \left\{ \|f - f^\star\|^2 + \frac{\gamma_n \log |\tilde{\mathcal{F}}(\epsilon, v_f)|}{n} + \mathbb{E} \left[\frac{T_n}{n} \right] \right\} \\ &\leq \tau \inf_{\epsilon > 0} \left\{ \epsilon^2 + \frac{\gamma_n Cd(v_{f^\star} v_0)^{\frac{2d}{d+2}} \epsilon^{-\frac{2d}{d+2}}}{n} + \mathbb{E} \left[\frac{T_n}{n} \right] \right\} \\ &\leq 2\tau(v_{f^\star} v_0)^{\frac{d}{d+1}} \left(\frac{C\gamma_n d}{n} \right)^{\frac{d+2}{2(d+1)}} + \mathbb{E} \left[\frac{T_n}{n} \right].\end{aligned}$$

This result is similar to [54], which also improved on the more familiar rate of $(\frac{d}{n})^{1/2}$ are obtained.

On the other hand, if $h = \phi(x \cdot \theta_h)$ and $\|\theta_h\|_1 \leq v_0$, we can use an alternative argument via Lemma 6 to produce $\log |\tilde{\mathcal{F}}(\epsilon, v)| \leq C\epsilon^{-2}v^4v_0^2 \log(d+1)$ and $\log |\tilde{\mathcal{F}}(\epsilon, v)| \leq C\epsilon^{-3}v^3v_0^2 \log(d+1)$ if ϕ satisfies Assumption 1 and Assumption 2, respectively. Hence, $\mathbb{E}\|T\hat{f} - f^\star\|^2$ is bounded by a multiple of $\left(\frac{\gamma_n v_{f^\star}^4 v_0^2 \log(d+1)}{n} \right)^{1/3}$ and $\left(\frac{\gamma_n v_{f^\star}^3 v_0^2 \log(d+1)}{n} \right)^{2/5}$ if ϕ satisfies Assumption 1 and Assumption 2, respectively.

Compare this result with the minimax risk lower bound (2.13) of order $(\frac{\log(d+1)}{n})^{1/2}$. The exponents of these rates should also be compared with the extension to optimize over the continuum in Section 2.5.1, where obtained the 1/3 power rate only under the stronger Assumption 2 and a 1/4 rate for the general bounded Lipschitz case Assumption 1.

2.8 Proofs of the lemmata

An important aspect of the above covers $\tilde{\mathcal{F}}$ is that they only depend on the data $(\underline{X}, \underline{X}')$ through $\|\underline{X}\|_\infty^2 + \|\underline{X}'\|_\infty^2$, where $\|\underline{X}\|_\infty^2 = \frac{1}{n} \sum_{i=1}^n \|X_i\|_\infty^2$. Since the coordinates of \underline{X} and \underline{X}' are restricted to belong to $[-1, 1]^d$, the penalties and quantities satisfying Kraft's inequality

do not depend on \underline{X} and \underline{X}' . This is an important implication for the following empirical process theory.

Lemma 2. *Let $(\underline{X}, \underline{X}') = (X_1, \dots, X_n, X'_1, \dots, X'_n)$, where \underline{X}' is an independent copy of the data \underline{X} and where (X_1, \dots, X_n) are component-wise independent but not necessarily identically distributed. A countable function class \mathcal{G} and complexities $L(g)$ satisfying $\sum_{g \in \mathcal{G}} e^{-L(g)} \leq 1$ are given. Then for arbitrary positive γ ,*

$$\mathbb{E} \sup_{g \in \mathcal{G}} \left\{ D'_n(g) - D_n(g) - \frac{\gamma}{n} L(g) - \frac{1}{2\gamma} s^2(g) \right\} \leq 0, \quad (2.34)$$

where $s^2(g) = \frac{1}{n} \sum_{i=1}^n (g^2(X_i) - g^2(X'_i))^2$.

Proof. Let $\underline{Z} = (Z_1, \dots, Z_n)$ be a sequence of independent centered Bernoulli random variables with success probability $1/2$. Since X_i and X'_i are identically distributed, $g^2(X_i) - g^2(X'_i)$ is a symmetric random variable and hence sign changes do not affect the expectation in (2.34). Thus the right hand side of the inequality in (2.34) is equal to

$$\mathbb{E}_{\underline{Z}, \underline{X}, \underline{X}'} \sup_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^n Z_i (g^2(X_i) - g^2(X'_i)) - \frac{\gamma}{n} L(g) - \frac{1}{2\gamma} s^2(g) \right\}.$$

Using the identity $x = \lambda \log(x/\lambda)$ with $\lambda = \gamma/n$, conditioning on \underline{X} and \underline{X}' , and applying Jensen's inequality to move $\mathbb{E}_{\underline{Z}}$ inside the logarithm, we have that

$$\begin{aligned} & \mathbb{E}_{\underline{Z}} \sup_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^n Z_i (g^2(X_i) - g^2(X'_i)) - \frac{\gamma}{n} L(g) - \frac{1}{2\gamma} s^2(g) \right\} \\ & \leq \frac{\gamma}{n} \log \mathbb{E}_{\underline{Z}} \sup_{g \in \mathcal{G}} \exp \left\{ \frac{1}{\gamma} \sum_{i=1}^n Z_i (g^2(X_i) - g^2(X'_i)) - L(g) - \frac{n}{2\gamma^2} s^2(g) \right\}. \end{aligned}$$

Replacing the supremum with the sum and using the linearity of expectation, the above expression is not more than

$$\begin{aligned} & \frac{\gamma}{n} \log \sum_{g \in \mathcal{G}} \mathbb{E}_{\underline{Z}} \exp \left\{ \frac{1}{\gamma} \sum_{i=1}^n Z_i (g^2(X_i) - g^2(X'_i)) - L(g) - \frac{n}{2\gamma^2} s^2(g) \right\} \\ & = \frac{\gamma}{n} \log \sum_{g \in \mathcal{G}} \exp \left\{ -L(g) - \frac{n}{2\gamma^2} s^2(g) \right\} \mathbb{E}_{\underline{Z}} \exp \left\{ \frac{1}{\gamma} \sum_{i=1}^n Z_i (g^2(X_i) - g^2(X'_i)) \right\}. \end{aligned}$$

Next, note that by the independence of Z_1, \dots, Z_n ,

$$\mathbb{E}_{\underline{Z}} \exp \left\{ \frac{1}{\gamma} \sum_{i=1}^n Z_i (g^2(X_i) - g^2(X'_i)) \right\} = \prod_{i=1}^n \mathbb{E}_{Z_i} \exp \left\{ \frac{1}{\gamma} Z_i (g^2(X_i) - g^2(X'_i)) \right\}.$$

Using the inequality $e^x + e^{-x} \leq 2e^{x^2/2}$, each $\mathbb{E}_{Z_i} \exp \left\{ \frac{1}{\gamma} Z_i (g^2(X_i) - g^2(X'_i)) \right\}$ is not more than $\exp \left\{ \frac{1}{2\gamma^2} (g^2(X_i) - g^2(X'_i))^2 \right\}$. Whence

$$\mathbb{E}_{\underline{Z}} \exp \left\{ \frac{1}{\gamma} \sum_{i=1}^n Z_i (g^2(X_i) - g^2(X'_i)) \right\} \leq \exp \left\{ \frac{n}{2\gamma^2} s^2(g) \right\}.$$

The claim follows from the fact that $\frac{\gamma}{n} \log \sum_{g \in \mathcal{G}} e^{-L(g)} \leq 0$. \square

Lemma 3. *Let $\underline{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$ be conditionally independent random variables given $\{X_i\}_{i=1}^n$, with conditional mean zero, satisfying Bernstein's moment condition with parameter $\eta > 0$. A countable class \mathcal{G} and complexities $L(g)$ satisfying*

$$\sum_{g \in \mathcal{G}} e^{-L(g)} \leq 1$$

are given. Assume a bound K , such that $|g(x)| \leq K$ for all g in \mathcal{G} . Then

$$\mathbb{E} \sup_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(X_i) - \frac{\gamma}{n} L(g) - \frac{1}{An} \sum_{i=1}^n g^2(X_i) \right\} \leq 0.$$

where A is an arbitrary constant and $\gamma = A\sigma^2/2 + Kh$.

Proof. Using the identity $x = \lambda \log(x/\lambda)$ with $\lambda = \gamma/n$, conditioning on \underline{X} , and applying Jensen's inequality to move $\mathbb{E}_{\underline{\varepsilon}}$ inside the logarithm, we have that

$$\begin{aligned} & \mathbb{E}_{\underline{\varepsilon}|\underline{X}} \sup_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(X_i) - \frac{\gamma}{n} L(g) - \frac{1}{An} \sum_{i=1}^n g^2(X_i) \right\} \\ & \leq \frac{\gamma}{n} \log \mathbb{E}_{\underline{\varepsilon}|\underline{X}} \sup_{g \in \mathcal{G}} \exp \left\{ \frac{1}{\gamma} \sum_{i=1}^n \varepsilon_i g(X_i) - L(g) - \frac{1}{\gamma A} \sum_{i=1}^n g^2(X_i) \right\}. \end{aligned}$$

Replacing the supremum with the sum and using the linearity of expectation, the above

expression is not more than

$$\begin{aligned} & \frac{\gamma}{n} \log \sum_{g \in \mathcal{G}} \mathbb{E}_{\underline{\varepsilon} | \underline{X}} \exp \left\{ \frac{1}{\gamma} \sum_{i=1}^n \varepsilon_i g(X_i) - L(g) - \frac{1}{\gamma A} \sum_{i=1}^n g^2(X_i) \right\} \\ &= \frac{\gamma}{n} \log \sum_{g \in \mathcal{G}} \exp \left\{ -L(g) - \frac{1}{\gamma A} \sum_{i=1}^n g^2(X_i) \right\} \mathbb{E}_{\underline{\varepsilon} | \underline{X}} \exp \left\{ \frac{1}{\gamma} \sum_{i=1}^n \varepsilon_i g(X_i) \right\}. \end{aligned}$$

Next, note that by the independence of $\varepsilon_1, \dots, \varepsilon_n$ conditional on \underline{X} ,

$$\mathbb{E}_{\underline{\varepsilon} | \underline{X}} \exp \left\{ \frac{1}{\gamma} \sum_{i=1}^n \varepsilon_i g(X_i) \right\} = \prod_{i=1}^n \mathbb{E}_{\varepsilon_i | X_i} \exp \left\{ \frac{1}{\gamma} \varepsilon_i g(X_i) \right\}.$$

By Lemma 8, each $\mathbb{E}_{\varepsilon_i | X_i} \exp \left\{ \frac{1}{\gamma} \varepsilon_i g(X_i) \right\}$ is not more than $\exp \left\{ \frac{\sigma^2 g^2(X_i)}{2\gamma^2(1-\eta K/\gamma)} \right\}$. Whence

$$\begin{aligned} \mathbb{E}_{\underline{\varepsilon} | \underline{X}} \exp \left\{ \frac{1}{\gamma} \sum_{i=1}^n \varepsilon_i g(X_i) \right\} &\leq \exp \left\{ \frac{\sigma^2 \sum_{i=1}^n g^2(X_i)}{2\gamma^2(1-\eta K/\gamma)} \right\} \\ &= \exp \left\{ \frac{1}{\gamma A} \sum_{i=1}^n g^2(X_i) \right\}, \end{aligned}$$

where the last line follows from the definition of γ . The proof is finished after observing that $\frac{\gamma}{n} \log \sum_{g \in \mathcal{G}} e^{-L(g)} \leq 0$. \square

Lemma 4. For $f = \sum_h \beta_h h$ and f_0 in \mathcal{F} , there is a choice of h_1, \dots, h_m in \mathcal{H} with $f_m = (v/m) \sum_{k=1}^m h_k$, $v \geq v_f$ such that

$$\|f_m - f_0\|^2 - \|f_0 - f\|^2 \leq \frac{vv_f}{m}.$$

Moreover, the same bound holds for any convex combination of $\|f_m - f_0\|^2 - \|f_0 - f\|^2$ and $\rho^2(f_m, f)$, where ρ is a possibly different Hilbert space norm.

Proof. Let H be a random variable that equals $h v$ with probability β_h/v and zero with probability $1 - v_f/v$. Let H_1, \dots, H_m be a random sample from the distribution defining H . Then $\overline{H} = \frac{1}{m} \sum_{j=1}^m H_j$ has mean f and furthermore the mean of $\|f_m - f_0\|^2 - \|f_0 - f\|^2$ is the mean is $\|f - \overline{H}\|^2$. This quantity is seen to be bounded by vv_f/m . As a consequence of the bound holding on average, there exists a realization of f_m of \overline{H} (having form $(v/m) \sum_{k=1}^m h_k$)

such that $\|f_m - f_0\|^2 - \|f_0 - f\|^2$ is also bounded by Vv_f/m . \square

The next lemma is an extension of a technique used in [55] to improve the L^2 error of an m -term approximation of a function in $L_{1,\mathcal{H}}$. The idea is essentially stratified sampling with proportional allocation [56] used in survey sampling as a means of variance reduction. In the following, we use the notation $\|\cdot\|$ to denote a generic Hilbert space norm.

Lemma 5. *Let $\tilde{\mathcal{H}}$ be an L^2 ϵ_1 -net of \mathcal{H} with cardinality M_1 . For $f = \sum_h \beta_h h$ and f_0 in \mathcal{F} , there is a choice of h_1, \dots, h_m in \mathcal{H} with $f_m = (1/m_0) \sum_{k=1}^m b_k h_k$, $m \leq m_0 + M_1$ and $\|b\|_1 \geq v_f$ such that*

$$\|f_0 - f_m\|^2 - \|f_0 - f\|^2 \leq \frac{vv_f \epsilon_1^2}{m_0}.$$

Moreover, there is an equally weighted linear combination $f_m = (v/m_0) \sum_{k=1}^m h_k$, $v \geq v_f$, $m \leq m_0 + M_1$ such that

$$\|f_0 - f_m\|^2 - \|f_0 - f\|^2 \leq \frac{v^2 \epsilon_1^2 (1 + M_1/m_0)}{m_0} + \frac{v^2 M_1}{4m_0^2}. \quad (2.35)$$

The same bound holds for any convex combination of $\|f_m - f_0\|^2 - \|f_0 - f\|^2$ and $\rho^2(f_m, f)$, where ρ is a possibly different Hilbert space norm.

Proof. Suppose the elements of $\tilde{\mathcal{H}}$ are $\tilde{h}_1, \dots, \tilde{h}_{M_1}$. Consider the M_1 sets (or “strata”)

$$\tilde{\mathcal{H}}_j = \{h \in \mathcal{H} : \|h - \tilde{h}_j\|^2 \leq \epsilon_1^2\},$$

$j = 1, \dots, M_1$. By working instead with disjoint sets $\tilde{\mathcal{H}}_j \setminus \bigcup_{1 \leq i \leq j-1} \tilde{\mathcal{H}}_i$, $\tilde{\mathcal{H}}_0 = \emptyset$, that are contained in $\tilde{\mathcal{H}}_j$ and whose union is \mathcal{H} , we may assume that the $\tilde{\mathcal{H}}_j$ form a partition of \mathcal{H} . Let $M = m_0 + M_1$ and $v_j = \sum_{h \in \tilde{\mathcal{H}}_j} \beta_h$. To obtain the first conclusion, define a random variable H_j to equal $h v_j$ with probability β_h / v_j for all $h \in \tilde{\mathcal{H}}_j$. let $H_{1,j}, \dots, H_{N_j,j}$ be a random sample of size $N_j = \left\lceil \frac{v_j M}{V} \right\rceil$, where $V = \frac{vM}{m_0}$ and $v \geq v_f$, from the distribution defining H_j . Note that the N_j sum to at most M . Define $g_j = \sum_{h \in \tilde{\mathcal{H}}_j} \beta_h h$ and $\bar{f} = \sum_{j=1}^{M_1} \frac{1}{N_j} \sum_{k=1}^{N_j} H_{k,j}$. Note that the mean of \bar{f} is f . This means the expectation of $\|f_0 - \bar{f}\|^2 - \|f_0 - f\|^2$ is the expectation of $\|f - \bar{f}\|^2$, which is equal to $\sum_{j=1}^{M_1} \mathbb{E} \|H_j - g_j\|^2 / N_j$. Now $\mathbb{E} \|H_j - g_j\|^2 / N_j$ is

further bounded by

$$(V/M) \sum_{h \in \tilde{\mathcal{H}}_j} \beta_h \inf_{h_j} \|h - h_j\|^2 \leq (V/M) \sum_{h \in \tilde{\mathcal{H}}_j} \beta_h \|h - \tilde{h}_j\|^2 \leq \frac{v_j v \epsilon_1^2}{m_0}.$$

The above fact was established by noting that the mean of a real-valued random variable minimizes its average squared distance from any point h_j . Summing over $1 \leq j \leq M_1$ produces the claim. Since this bound holds on average, there exists a realization f_m of \bar{f} (having form $(1/m_0) \sum_{k=1}^m b_k h_k$ with $\|b\|_1 \geq v_f$) such that $\|f_0 - f_m\|^2 - \|f_0 - f\|^2$ is also bounded by $\frac{v v_f \epsilon_1^2}{m_0}$.

For the second conclusion, we proceed in a similar fashion. Suppose n_j is a random variable that equals $\left\lceil \frac{v_j M}{V} \right\rceil$ and $\left\lfloor \frac{v_j M}{V} \right\rfloor$ with respective probabilities chosen to make its average equal to $\frac{v_j M}{V}$. Furthermore, assume n_1, \dots, n_{M_1} are independent. Define $V_j = \frac{V}{M} n_j$. Since $V_j \leq v_j + \frac{V}{M}$, the V_j sum to at most V . Let H_j be a random variable that equals $h v_j$ with probability β_h / v_j for all $h \in \tilde{\mathcal{H}}_j$. For each j and conditional on n_j , let $H_{1,j}, \dots, H_{n_j,j}$ be a random sample of size $N_j = n_j + \mathbb{I}\{n_j = 0\}$ from the distribution defining H_j . Note that the N_j sum to at most M . Define $g_j = \sum_{h \in \tilde{\mathcal{H}}_j} \beta_h h$ and $\bar{f} = \sum_{j=1}^{M_1} \frac{1}{N_j} \sum_{k=1}^{N_j} H_{k,j}$. Note that the conditional mean of \bar{H} given N_1, \dots, N_{M_1} is $g = \sum_{j=1}^{M_1} (V_j / v_j) g_j$ and hence the mean of \bar{f} is f . This means the expectation of $\|f_0 - \bar{f}\|^2 - \|f_0 - f\|^2$ is the expectation of $\|f - \bar{f}\|^2$, which is equal to $\sum_{j=1}^{M_1} \mathbb{E} \|H_j - (V_j / v_j) g_j\|^2 / N_j + \mathbb{E} \|f - g\|^2$ by the law of total variance. Now $\mathbb{E} \|H_j - (V_j / v_j) g_j\|^2 / N_j$ is further bounded by

$$(V/M)^2 (n_j / v_j) \sum_{h \in \tilde{\mathcal{H}}_j} \beta_h \inf_{h_j} \|h - h_j\|^2 \leq \frac{v^2 M \epsilon_1^2}{m_0^2}.$$

The above fact was established by noting that the mean of a real-valued random variable minimizes its average squared distance from any point h_j . Next, note that by the independence of the coordinates of v_1, \dots, v_{M_1} and the fact that V_j has mean v_j ,

$$\mathbb{E} \|f - g\|^2 = \mathbb{E} \left\| \sum_{j=1}^{M_1} (V_j / v_j - 1) g_j \right\|^2 = (V/M)^2 \sum_{j=1}^{M_1} (\|g_j\|^2 / v_j^2) \mathbb{V}(n_j).$$

Finally, observe that $\|g_j\|^2 \leq v_j^2$ and $\mathbb{V}(n_j) \leq 1/4$ (a random variable whose range is contained in an interval of length one has variance bounded by $1/4$). This shows that $\mathbb{E}\|f - g\|^2 \leq \frac{v^2 M_1}{4m_0^2}$. Since this bound holds on average, there exists a realization f_m of \bar{f} (having form $(v/m_0) \sum_{k=1}^m h_k$) such that $\|f_0 - f_m\|^2 - \|f_0 - f\|^2$ is also bounded by $\frac{v^2 \epsilon_1^2 (1+M_1/m_0)}{m_0} + \frac{v^2 M_1}{4m_0^2}$. \square

Lemma 6. *There is a collection of functions $\tilde{\mathcal{F}}$ with cardinality at most $2^{\binom{2d+m_0}{m_1}+m_1} \lesssim d^{m_0 m_1}$ such that for each $f(x) = \sum_h \beta_h h(x) = \sum_h \beta_h \phi(\theta_h \cdot x)$, there exists \tilde{f} in $\tilde{\mathcal{F}}$ such that for any $v \geq v_f$,*

$$\|\tilde{f} - f\|^2 \leq \frac{vv_f}{m_1} + \frac{L_2^2 v_f^2 v_0^4}{4m_0^2}. \quad (2.36)$$

and

$$\|g - \tilde{f}\|^2 - \|g - f\|^2 \leq \frac{vv_f}{m_1} + \frac{L_2 v_f (\|g\|_1 + v_f) v_0^2}{m_0}, \quad (2.37)$$

provided ϕ satisfies Assumption 2. If ϕ satisfies Assumption 1, then

$$\|\tilde{f} - f\|^2 \leq \frac{vv_f}{m_1} + \frac{L_1^2 v_f^2 v_0^2}{m_0}. \quad (2.38)$$

Proof. Define a joint probability distribution $(\tilde{\theta}_H, H)$ as follows. Let $\mathbb{P}[\tilde{\theta}_h = e_i \text{sgn}(\theta_h(i)) | H = h] = \frac{|\theta_h(i)|}{v_0}$, where e_i denotes the i -th standard basis vector for \mathbb{R}^d , and $\mathbb{P}[\tilde{\theta}_h = 0 | H = h] = 1 - \frac{\|\theta_h\|_1}{v_0}$ for $i = 1, 2, \dots, d$ and $\mathbb{P}[H = h] = \frac{|\beta_h|}{v}$ and $\mathbb{P}[H = 0] = 1 - \frac{v_f}{v}$ for all $h \in \mathcal{H}$ and $v \geq v_f$.

Take a random sample $\underline{H} = \{H_j\}_{1 \leq j \leq m_1}$ from the distribution defining H . Given \underline{H} take a random sample $\underline{\tilde{\theta}} = \{\tilde{\theta}_{k,H_j}\}_{1 \leq k \leq m_0, 1 \leq j \leq m_1}$, where $\tilde{\theta}_{k,H_j}$ is distributed according to $\tilde{\theta}_{H_j}$. Define

$$\tilde{f}_{m_0, m_1}(x) = \frac{v}{m_1} \sum_{j=1}^{m_1} \text{sgn}(\beta_{H_j}) \phi \left(\frac{v_0}{m_0} \sum_{k=1}^{m_0} \tilde{\theta}_{k, H_j} \cdot x \right). \quad (2.39)$$

By a similar argument to Lemma 4, there exists a realization of \tilde{f}_{m_0, m_1} such that

$$\|\tilde{f}_{m_0, m_1} - \mathbb{E} \tilde{f}_{m_0, m_1}\|^2 \leq \frac{vv_f}{m_1}. \quad (2.40)$$

By the bias-variance decomposition,

$$\mathbb{E}\|\tilde{f}_{m_0, m_1} - f\|^2 = \|\tilde{f}_{m_0, m_1} - \mathbb{E}\tilde{f}_{m_0, m_1}\|^2 + \|f - \mathbb{E}\tilde{f}_{m_0, m_1}\|^2. \quad (2.41)$$

Using a similar argument to Lemma 4, there exists a realization of \tilde{f}_{m_0, m_1} such that

$$\|\tilde{f}_{m_0, m_1} - \mathbb{E}\tilde{f}_{m_0, m_1}\|^2 \leq \frac{vv_f}{m_1}. \quad (2.42)$$

The second term of (2.41) may be bounded as follows. First, note that

$$\mathbb{E}\tilde{f}_{m_0, m_1}(x) = \sum_h \beta_h \mathbb{E}\phi\left(\frac{v_0}{m_0} \sum_{k=1}^{m_0} \tilde{\theta}_{k, H_j} \cdot x\right).$$

By Assumption 2, we have the pointwise bound

$$\begin{aligned} |f(x) - \mathbb{E}\tilde{f}_{m_0, m_1}(x)| &= \left| \sum_h \beta_h \phi(\theta_h \cdot x) - \sum_h \beta_h \mathbb{E}\phi\left(\frac{v_0}{m_0} \sum_{k=1}^{m_0} \tilde{\theta}_{k, h} \cdot x\right) \right| \\ &\leq \frac{L_2}{2} \sum_h |\beta_h| \mathbb{E} \left| \frac{v_0}{m_0} \sum_{k=1}^{m_0} \tilde{\theta}_{k, h} \cdot x - \theta_h \cdot x \right|^2 \leq \frac{L_2}{2} \sum_h |\beta_h| \frac{v_0^2 \|x\|_\infty}{m_0} \\ &\leq \frac{L_2 v_f v_0^2 \|x\|_\infty^2}{2m_0} \leq \frac{L_2 v_f v_0^2}{2m_0}. \end{aligned} \quad (2.43)$$

Here we used the fact that

$$\begin{aligned} \mathbb{E} \left| \frac{v_0}{m_0} \sum_{k=1}^{m_0} \tilde{\theta}_{k, h} \cdot x - \theta_h \cdot x \right|^2 &\leq \frac{v_0^2}{m_0} \mathbb{E} |\tilde{\theta}_{k, h} \cdot x|^2 \\ &= \frac{v_0}{m_0} \sum_{i=1}^d |\theta_h(i)| |x(i)| \\ &\leq \frac{v_0^2}{m_0}. \end{aligned}$$

Combining (2.42) and (2.43), we have shown that there exists a realization \tilde{f} of \tilde{f}_{m_0, m_1} such that

$$\|\tilde{f} - f\|^2 \leq \frac{vv_f}{m_1} + \frac{L_2^2 v_f^2 v_0^4}{4m_0^2}.$$

To show (2.37), we also use the bias-variance decomposition to write

$$\begin{aligned}\mathbb{E}\|g - \tilde{f}_{m_0, m_1}\|^2 - \|g - f\|^2 &= \mathbb{E}\|\tilde{f}_{m_0, m_1} - \mathbb{E}\tilde{f}_{m_0, m_1}\|^2 + \|g - \mathbb{E}\tilde{f}_{m_0, m_1}\|^2 - \|g - f\|^2 \\ &= \mathbb{E}\|\tilde{f}_{m_0, m_1} - \mathbb{E}\tilde{f}_{m_0, m_1}\|^2 + \langle f - \mathbb{E}\tilde{f}_{m_0, m_1}, 2g - f - \mathbb{E}\tilde{f}_{m_0, m_1} \rangle.\end{aligned}$$

As before, $\mathbb{E}\|\tilde{f}_{m_0, m_1} - \mathbb{E}\tilde{f}_{m_0, m_1}\|^2$ is less than $\frac{vv_f}{m_1}$. By (2.43), $|f(x) - \mathbb{E}\tilde{f}_{m_0, m_1}(x)| \leq \frac{L_2 v_f v_0^2}{2m_0}$, and combining this with the pointwise bounds $|f| \leq v_f$ and $|\mathbb{E}\tilde{f}_{m_0, m_1}| \leq v_f$, we have

$$|\langle f - \mathbb{E}\tilde{f}_{m_0, m_1}, 2g - f - \mathbb{E}\tilde{f}_{m_0, m_1} \rangle| \leq \frac{L_2 v_f (\|g\|_1 + v_f) v_0^2}{m_0}.$$

If ϕ satisfies Assumption 1, we use (2.41) together with the pointwise bound

$$\begin{aligned}|f(x) - \mathbb{E}\tilde{f}_{m_0, m_1}(x)| &= \left| \sum_h \beta_h \phi(\theta_h \cdot x) - \sum_h \beta_h \mathbb{E} \phi \left(\frac{v_0}{m_0} \sum_{k=1}^{m_0} \tilde{\theta}_{k, h} \cdot x \right) \right| \\ &\leq L_1 \sum_h |\beta_h| \mathbb{E} \left| \frac{v_0}{m_0} \sum_{k=1}^{m_0} \tilde{\theta}_{k, h} \cdot x - \theta_h \cdot x \right| \\ &\leq L_1 \sum_h |\beta_h| \sqrt{\mathbb{E} \left| \frac{v_0}{m_0} \sum_{k=1}^{m_0} \tilde{\theta}_{k, h} \cdot x - \theta_h \cdot x \right|^2} \leq L_1 \sum_h |\beta_h| \frac{v_0 \|x\|_\infty}{\sqrt{m_0}} \\ &\leq \frac{L_1 v_f v_0 \|x\|_\infty}{\sqrt{m_0}} \leq \frac{L_1 v_f v_0}{\sqrt{m_0}},\end{aligned}$$

which yields

$$\mathbb{E}\|\tilde{f}_{m_0, m_1} - f\|^2 \leq \frac{vv_f}{m_1} + \frac{L_1^2 v_f^2 v_0^2}{m_0}. \quad (2.44)$$

Thus there exists a realization \tilde{f} of \tilde{f}_{m_0, m_1} such that (2.44) holds.

By two applications of Lemma 11 with $m = m_1$ and $M = \binom{2d+m_0}{m_0}$, the number of functions having the form (2.39) is at most $2^{\binom{2d+m_0}{m_0} + m_1}$. \square

Lemma 7. *There is a subset $\tilde{\mathcal{H}}$ of \mathcal{H} with cardinality at most $\binom{2d+m_0}{m_0}$ such that for each $h(x) = \phi(x \cdot \theta)$ in \mathcal{H} with $\|\theta\|_1 \leq v_0$, there is $\tilde{h}(x) = \phi(x \cdot \tilde{\theta})$ in $\tilde{\mathcal{H}}$ such that $\|h - \tilde{h}\|^2 \leq L_1 v_0 \|\theta\|_1 / m_0$.*

Proof. Let $\tilde{\theta}$ be a random vector that equals $e_i \text{sgn}(\theta(i))$ with probability $|\theta(i)|/v_0$, $i = 1, 2, \dots, d$ and equals the zero vector with probability $1 - \|\theta\|_1/v_0$. Let $\{\tilde{\theta}_j\}_{1 \leq j \leq m_0}$ be a

random sample from the distribution defining $\tilde{\theta}$. Note that the average of $\bar{\theta} = \frac{v_0}{m_0} \sum_{j=1}^{m_0} \tilde{\theta}_j$ is θ and hence the average of $|\bar{\theta} \cdot x - \theta \cdot x|^2$ is the variance of $\tilde{\theta} \cdot x$ divided by m_0 . Taking the expectation of the desired quantity and using Assumption 1, we have the pointwise inequality

$$\begin{aligned} \mathbb{E} |\phi(\bar{\theta} \cdot x) - \phi(\theta \cdot x)|^2 &= L_1 \mathbb{E} |\bar{\theta} \cdot x - \theta \cdot x|^2 \\ &= \frac{L_1 v_0^2}{m_0} \mathbb{E} |\tilde{\theta} \cdot x - \theta \cdot x|^2 \\ &\leq \frac{L_1 v_0^2}{m_0} \mathbb{E} |\tilde{\theta} \cdot x|^2 \\ &= \frac{L_1 v_0}{m_0} \sum_{i=1}^d |\theta(i)| |x(i)| \\ &\leq \frac{L_1 v_0 \|\theta\|_1}{m_0}. \end{aligned}$$

Since this bound holds on average, there must exist a realization of $\bar{\theta}$ for which the inequality is also satisfied. Consider the collection of all vectors of the form $\frac{v_0}{m_0} \sum_{j=1}^{m_0} u_j$, where u_j is any of the $2d + 1$ signed standard basis vectors including the zero vector. By Lemma 11, this collection has cardinality bounded by $\binom{2d+m_0}{m_0}$ with its logarithm is bounded by the minimum of $m_0 \log(e(2d/m_0 + 1))$ and $2m \log(d + 1)$. \square

Lemma 8. *Let Z have mean zero and variance σ^2 . Moreover, suppose Z satisfies Bernstein's moment condition with parameter $\eta > 0$. Then*

$$\mathbb{E}(e^{tZ}) \leq \exp \left\{ \frac{t^2 \sigma^2 / 2}{1 - \eta |t|} \right\}, \quad |t| < 1/\eta. \quad (2.45)$$

Lemma 9. *Define $Tf = \min\{B_n, |f|\} \text{sgn} f$. Then*

$$(I) \quad (y - Tf)^2 \leq (y - f)^2 + 2(|y| - B_n)^2 \mathbb{I}\{|y| > B_n\},$$

$$(II) \quad (y - Tf)^2 \leq (y - T\tilde{f})^2 + 4B_n |f - \tilde{f}| + 4B_n (|y| - B_n) \mathbb{I}\{|y| > B_n\}, \text{ and}$$

$$(III) \quad (T\tilde{f} - Tf)^2 \leq (f - f_1)^2 + 4B_n |f_1 - \tilde{f}|.$$

Proof. (I) Since $(y - Tf)^2 = (y - f)^2 + 2(f - Tf)(2y - f - Tf)$, the proof will be complete

if we can show that

$$(f - Tf)(2y - f - Tf) \leq (|y| - B_n)^2 \mathbb{I}\{|y| > B_n\}.$$

Note that if $|f| \leq B_n$, the left hand side of the above expression is zero. Thus we may assume that $|f| > B_n$, in which case $f - Tf = \operatorname{sgn} f(|f| - B_n)$. Thus

$$\begin{aligned} (f - Tf)(2y - f - Tf) &= 2y \operatorname{sgn} f(|f| - B_n) - (|f| - B_n)(|f| + B_n) \\ &\leq 2|y|(|f| - B_n) - (|f| - B_n)(|f| + B_n). \end{aligned}$$

If $|y| \leq B_n$, the above expression is less than $-(|f| - B_n)^2 \leq 0$. Otherwise, it is a quadratic in $|f|$ that attains its global maximum at $|f| = |y|$. This yields a maximum value of $(|y| - B_n)^2$.

(II) For the second claim, note that

$$(y - Tf)^2 = (y - T\tilde{f})^2 + (T\tilde{f} - Tf)(2y - T\tilde{f} - Tf).$$

Hence, we are done if we can show that

$$(T\tilde{f} - Tf)(2y - T\tilde{f} - Tf) \leq 4B_n|f - \tilde{f}| + 4B_n(|y| - B_n) \mathbb{I}\{|y| > B_n\}.$$

If $|y| \leq B_n$, then

$$\begin{aligned} (T\tilde{f} - Tf)(2y - T\tilde{f} - Tf) &\leq 4B_n|T\tilde{f} - Tf| \\ &\leq 4B_n|\tilde{f} - f|. \end{aligned}$$

If $|y| > B_n$, then

$$\begin{aligned} (T\tilde{f} - Tf)(2y - T\tilde{f} - Tf) &\leq 2|T\tilde{f} - Tf||y| + 2B_n|T\tilde{f} - Tf| \\ &= 2|T\tilde{f} - Tf|(|y| - B_n) + 4B_n|T\tilde{f} - Tf| \\ &\leq 4B_n(|y| - B_n) + 4B_n|\tilde{f} - f|. \end{aligned}$$

(III) For the last claim, note that

$$\begin{aligned}
(T\tilde{f} - Tf)^2 &= (T\tilde{f} - Tf_1)^2 + [2T\tilde{f} - Tf_1 - Tf](Tf_1 - Tf) \\
&\leq (T\tilde{f} - Tf_1)^2 + 4B_n|Tf_1 - Tf| \\
&\leq (\tilde{f} - f_1)^2 + 4B_n|f_1 - f|
\end{aligned}$$

□

Lemma 10. *Let $Y = f^*(X) + \varepsilon$ with $|f^*(X)| \leq B$. Suppose*

$$(I) \quad \mathbb{E}e^{|\varepsilon|/\nu} < +\infty \text{ or}$$

$$(II) \quad \mathbb{E}e^{|\varepsilon|^2/\nu} < +\infty$$

for some $\nu > 0$. Then $\mathbb{E}[(Y^2 - B_n^2)\mathbb{I}\{|Y| > B_n\}]$ is at most

$$(I) \quad (4\nu^2/n)\mathbb{E}e^{|\varepsilon|/\nu} \text{ provided } B_n > \sqrt{2}(B + \nu \log n) \text{ or}$$

$$(II) \quad (2\nu/n)\mathbb{E}e^{|\varepsilon|^2/\nu} \text{ provided } B_n > \sqrt{2}(B + \sqrt{\nu \log n}).$$

Proof. Under assumption (I),

$$\begin{aligned}
\mathbb{P}(Y^2 - B_n^2 > t) &= \mathbb{P}(|Y| > \sqrt{t + B_n^2}) \\
&\leq \mathbb{P}(|\varepsilon| > \sqrt{t + B_n^2} - B) \\
&\leq \mathbb{P}(|\varepsilon| > (1/\sqrt{2})(\sqrt{t} + B_n) - B) \\
&\leq e^{-\frac{1}{\nu}\sqrt{\frac{t}{2}}} e^{-\frac{1}{\nu}(\frac{B_n}{\sqrt{2}} - B)} \mathbb{E}e^{|\varepsilon|/\nu}.
\end{aligned}$$

The last inequality follows from a simple application of Markov's inequality after exponentiation. Integrating the previous expression from $t = 0$ to $t = +\infty$ ($\int_0^\infty e^{-\frac{1}{\nu}\sqrt{\frac{t}{2}}} dt = 4\nu^2$) yields an upper bound on $\mathbb{E}[(Y^2 - B_n^2)\mathbb{I}\{|Y| > B_n\}]$ that is at most $(4\nu^2/n)\mathbb{E}e^{|\varepsilon|/\nu}$ provided $B_n > \sqrt{2}(B + \nu \log n)$.

Under assumption (II),

$$\begin{aligned}
\mathbb{P}(Y^2 - B_n^2 > t) &= \mathbb{P}(|Y|^2 > t + B_n^2) \\
&\leq \mathbb{P}(|\varepsilon|^2 > (1/2)(t + B_n^2) - B^2) \\
&\leq e^{-\frac{t}{2\nu}} e^{-\frac{1}{\nu}(\frac{B_n^2}{2} - B^2)} \mathbb{E}e^{|\varepsilon|^2/\nu}.
\end{aligned}$$

The last inequality follows from a simple application of Markov's inequality after exponentiation. Integrating the previous expression from $t = 0$ to $t = +\infty$ ($\int_0^\infty e^{-\frac{t}{2\nu}} dt = 2\nu$) yields an upper bound on $\mathbb{E}[(Y^2 - B_n^2)\mathbb{I}\{|Y| > B_n\}]$ that is at most $(2\nu/n)\mathbb{E}e^{|\varepsilon|^2/\nu}$ provided $B_n > \sqrt{2}(B + \sqrt{\nu \log n}) \geq \sqrt{2}(B^2 + \nu \log n)$. \square

Lemma 11. *The number of functions having the form $\sum_{k=1}^m f_k$, where f_k belong to a library of size M is at most $\binom{M-1+m}{m} \leq \binom{M+m}{m}$ and its logarithm bounded by $m \log(e(M/m + 1))$.*

Proof. Suppose the elements in the library are indexed by $1, 2, \dots, M$. Let w_i be the number of terms in $\sum_{k=1}^m f_k$ of type i . Hence the number of function of the form $\sum_{k=1}^m f_k$ is at most the number of non-negative integer solutions w_1, w_2, \dots, w_M to $w_1 + w_2 + \dots + w_M = m$. This number is $\binom{M-1+m}{m}$ with its logarithm bounded by the minimum of $m \log(e((M-1)/m + 1))$ and $m \log M$. \square

Chapter 3

Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls

3.1 Introduction

Functions of many variables are approximated using linear combinations of ridge functions with one layer of nonlinearities, viz.,

$$f_m(x) = \sum_{k=1}^m b_k \phi(a_k \cdot x - t_k), \quad (3.1)$$

where $b_k \in \mathbb{R}$ are the outer layer parameters and $a_k \in \mathbb{R}^d$ are the vectors of inner parameters for the single-hidden layer of functions $\phi(a_k \cdot x - t_k)$. The activation function ϕ is allowed to be quite general. For example, it can be bounded and Lipschitz, polynomials with certain controls on their degrees, or bounded with jump discontinuities. When the ridge activation function is a sigmoid, (3.1) is single-hidden layer artificial neural network.

One goal in a statistical setting is to estimate a regression function, i.e., conditional mean response, $f(x) = \mathbb{E}[Y \mid X = x]$ with domain $D \triangleq [-1, 1]^d$ from noisy observations $\{(X_i, Y_i)\}_{i=1}^n$, where $Y = f(X) + \varepsilon$. In classical literature [22], $L^2(P)$ mean squared pre-

diction error of order $(d/n)^{1/2}$, achieved by ℓ^1 penalized least squares estimators¹ over the class of models (3.1), are obtained by optimizing the tradeoff between *approximation error* and *descriptive complexity relative to sample size*. Bounds on the approximation error are obtained by first showing how models of the form (3.1) with $\phi(z) = \mathbb{1}\{z > 0\}$ can be used to approximate f satisfying $\int_{\mathbb{R}^d} \|\omega\|_1 |\mathcal{F}(f)(\omega)| d\omega < +\infty$, provided f admits a Fourier representation $f(x) = \int_{\mathbb{R}^d} e^{ix \cdot \omega} \mathcal{F}(f)(\omega) d\omega$ on $[-1, 1]^d$. Because it is often difficult to work with discontinuous ϕ (i.e., vanishing or exploding gradient issues), these step functions are replaced with smooth ϕ such that $\phi(\tau z) \wedge 1 \rightarrow \mathbb{1}\{z > 0\}$ as $\tau \rightarrow +\infty$. Thus, this setup allows one to work with approximants of the form (3.1) with smooth ϕ , but at the expense of *unbounded* ℓ^1 norm $\|a_k\|_1$.

Like high-dimensional linear regression [31], many applications of statistical inference and estimation require a setting where $d \gg n$. In contrast to the aforementioned mean square prediction error of $(d/n)^{1/2}$, it has been shown [57] how models of the form (3.1) with Lipschitz² ϕ (reps. Lipschitz derivative ϕ') and *bounded* inner parameters $\|a_k\|_0$ and $\|a_k\|_1$ can be used to give desirable $L^2(D)$ mean squared prediction error of order $((\log d)/n)^{1/3}$ (resp. $((\log d)/n)^{2/5}$), also achieved by penalized estimators.³ In fact, [35] shows that these rates are nearly optimal. A few natural questions arise from restricting the ℓ^0 and ℓ^1 norms of the inner parameters in the model:

- To what degree do the sparsity assumptions limit the flexibility of the model (3.1)?
- What condition can be imposed on f so that it can be approximated by f_m with Lipschitz ϕ (or Lipschitz derivative ϕ') and bounded $\|a_k\|_0$ and / or $\|a_k\|_1$?
- How well can f be approximated by f_m , given these sparsity constraints?

According to classic approximation results [3, 23], if the domain of f is contained in $[-1, 1]^d$ and f admits a Fourier representation $f(x) = \int_{\mathbb{R}^d} e^{ix \cdot \omega} \mathcal{F}(f)(\omega) d\omega$, then the spectral

1. That is, the fit minimizes $(1/n) \sum_{i=1}^n (f_m(X_i) - Y_i)^2 + \lambda \sum_{k=1}^m |b_k|$ for some appropriately chosen $\lambda > 0$.

2. Henceforth, when we say a function is Lipschitz, we assume it has bounded Lipschitz parameter.

3. With additional ℓ^0 inner sparsity, we might also consider an estimator that minimizes $(1/n) \sum_{i=1}^n (f_m(X_i) - Y_i)^2 + \lambda_0 \psi(\sum_{k=1}^m |b_k| \|a_k\|_0)$ for some convex function ψ and appropriately chosen $\lambda_0 > 0$.

condition $v_{f,1} < +\infty$, where $v_{f,s} \triangleq \int_{\mathbb{R}^d} \|\omega\|_1^s |\mathcal{F}(f)(\omega)| d\omega$, is enough to ensure that $f - f(0)$ can be approximated in $L^\infty(D)$ by equally weighted, i.e., $|b_1| = \dots = |b_m|$, linear combinations of functions of the form (3.1) with $\phi(z) = \mathbb{1}\{z > 0\}$. Typical L^∞ error rates $\|f - f_m\|_\infty$ of an m -term approximation (3.1) are at most $cv_{f,1}\sqrt{d} m^{-1/2}$, where c is a universal constant [3, 4, 58]. A rate of $c(p)v_{f,1}m^{-1/2-1/(pd)}$ was given in [55, Theorem 3] for $L^p(D)$ for nonnegative even integer p . Again, all these bounds are valid when the step activation function is replaced by a smooth approximant ϕ (in particular, *any* sigmoid satisfying $\lim_{z \rightarrow \pm\infty} \phi(z) = \pm 1$), but at the expense of unbounded $\|a_k\|_1$.

Towards giving partial answers to the questions we posed, in Section 7.4.4, we show how functions of the form (3.1) with ReLU (also known as a ramp or first order spline) $\phi(z) = (z)_+ = 0 \vee z$ (which is Lipschitz)⁴ or squared ReLU $\phi(z) = (z)_+^2$ (which has Lipschitz derivative) activation function can be used to give desirable $L^\infty(D)$ approximation error bounds, even when $\|a_k\|_1 = 1$, $0 \leq t_k \leq 1$, and $|b_1| = \dots = |b_m|$. Because of the widespread popularity of the ReLU activation function and its variants, these simpler forms may also be of independent interest for computational and algorithmic reasons as in [36, 45, 47, 50, 59], to name a few.

Unlike the case with step activation functions, our analysis makes no use of the combinatorial properties of half-spaces as in Vapnik-Chervonenkis theory [60, 61]. The $L^2(D)$ case for ReLU ridge functions (also known as hinging hyperplanes) with ℓ^1 -bounded inner parameters was considered in [5, Theorem 3] and our $L^\infty(D)$ bounds improve upon that line of work and, in addition, increase the exponent from $1/2$ to $1/2 + O(1/d)$. Our proof techniques are substantively different than [5] and, importantly, are more amenable to empirical process theory, which is the key to showing our error bounds.

These tighter rates of approximation, with ReLU and squared ReLU activation functions, are possible under two different conditions – finite $v_{f,2}$ or $v_{f,3}$, respectively. The main idea we use originates from [55] and [62] and can be seen as stratified sampling with proportionate allocation. This technique is widely applied in survey sampling as a means of variance reduction [56].

4. It is perhaps more conventional to write $(z)^+$ for $0 \vee z$, however, to avoid clutter in the exponent, we use the current notation.

At the end of Section 7.4.4, we will also discuss the degree to which these bounds can be improved by providing companion lower bounds on the minimax rates of approximation.

Section 3.3 will focus on how accurate estimation can be achieved even when $\|a_k\|_0$ is also bounded. In particular, we show how an m -term linear combination (3.1) with $\|a_k\|_0 \leq \sqrt{m}$ and $\|a_k\|_1 = 1$ can approximate f satisfying $v_{f,3} < +\infty$ in $L^2(D)$ with error at most $\sqrt{2}v_{f,3}m^{-1/2}$. In other words, the $L^2(D)$ approximation error is inversely proportional to the inner layer sparsity and it need only be sublinear in the outer layer sparsity. The constructions that achieve these error bounds are obtained using a variant of the Jones-Barron probabilistic method, which can be interpreted as two-stage cluster sampling.

Throughout this chapter, we will state explicitly how our bounds depend on d so that the reader can fully appreciate the complexity of approximation. If a is a vector in Euclidean space, we use the notation $a(k)$ to denote its k -th component.

3.2 L^∞ approximation with bounded ℓ^1 norm

3.2.1 Positive results

In this section, we provide the statements and proofs of the existence results for f_m with bounded ℓ^1 norm of inner parameters. We would like to point out that the results of Theorem 6 hold when all occurrences of the ReLU or squared ReLU activation functions are replaced by general ϕ which is Lipschitz or has Lipschitz derivative ϕ' , respectively.

Theorem 6. *Suppose f admits an integral representation*

$$f(x) = v \int_{[0,1] \times \{a: \|a\|_1=1\}} \eta(t, a) (a \cdot x - t)_+^{s-1} dP(t, a), \quad (3.2)$$

for x in $D = [-1, 1]^d$ and $s \in \{2, 3\}$, where P is a probability measure on $[0, 1] \times \{a \in \mathbb{R}^d : \|a\|_1 = 1\}$ and $\eta(t, a)$ is either -1 or $+1$. There exists a linear combination of ridge functions of the form

$$f_m(x) = \frac{v}{m} \sum_{k=1}^m b_k (a_k \cdot x - t_k)_+^{s-1}, \quad (3.3)$$

with $b_k \in [-1, 1]$, $\|a_k\|_1 = 1$, $0 \leq t_k \leq 1$ such that

$$\sup_{x \in D} |f(x) - f_m(x)| \leq c\sqrt{d + \log m} m^{-1/2-1/d}, \quad s = 2,$$

and

$$\sup_{x \in D} |f(x) - f_m(x)| \leq c\sqrt{d} m^{-1/2-1/d}, \quad s = 3,$$

for some universal constant $c > 0$. Furthermore, if the b_k are restricted to $\{-1, 1\}$, the upper bound is of order

$$\sqrt{d + \log m} m^{-1/2-1/(d+2)}, \quad s = 2$$

and

$$\sqrt{d} m^{-1/2-1/(d+2)}, \quad s = 3.$$

Theorem 7. Let $D = [-1, 1]^d$. Suppose f admits a Fourier representation $f(x) = \int_{\mathbb{R}^d} e^{ix \cdot \omega} \mathcal{F}(f)(\omega) d\omega$ and

$$v_{f,2} = \int_{\mathbb{R}^d} \|\omega\|_1^2 |\mathcal{F}(f)(\omega)| d\omega < +\infty.$$

There exists a linear combination of ReLU ridge functions of the form

$$f_m(x) = b_0 + a_0 \cdot x + \frac{v}{m} \sum_{k=1}^m b_k (a_k \cdot x - t_k)_+ \quad (3.4)$$

with $b_k \in [-1, 1]$, $\|a_k\|_1 = 1$, $0 \leq t_k \leq 1$, $b_0 = f(0)$, $a_0 = \nabla f(0)$, and $v \leq 2v_{f,2}$ such that

$$\sup_{x \in D} |f(x) - f_m(x)| \leq cv_{f,2} \sqrt{d + \log m} m^{-1/2-1/d},$$

for some universal constant $c > 0$. Furthermore, if the b_k are restricted to $\{-1, 1\}$, the upper bound is of order

$$v_{f,2} \sqrt{d + \log m} m^{-1/2-1/(d+2)}.$$

Theorem 8. *Under the setup of Theorem 7, suppose*

$$v_{f,3} = \int_{\mathbb{R}^d} \|\omega\|_1^3 |\mathcal{F}(f)(\omega)| d\omega < +\infty.$$

There exists a linear combination of squared ReLU ridge functions of the form

$$f_m(x) = b_0 + a_0 \cdot x + x^T A_0 x + \frac{v}{2m} \sum_{k=1}^m b_k (a_k \cdot x - t_k)_+^2 \quad (3.5)$$

with $b_k \in [-1, 1]$, $\|a_k\|_1 = 1$, $0 \leq t_k \leq 1$, $b_0 = f(0)$, $a_0 = \nabla f(0)$, $A_0 = \nabla \nabla^T f(0)$, and $v \leq 2v_{f,3}$ such that

$$\sup_{x \in D} |f(x) - f_m(x)| \leq cv_{f,3} \sqrt{d} m^{-1/2-1/d},$$

for some universal constant $c > 0$. Furthermore, if the b_k are restricted to $\{-1, 1\}$, the upper bound is of order

$$v_{f,3} \sqrt{d} m^{-1/2-1/(d+2)}.$$

The key observation for proving Theorem 7 and Theorem 8 is that f modulo linear or quadratic terms with finite $v_{f,s}$ can be written in the integral form (3.2). Unlike in [5, Theorem 3] where an interpolation argument is used, our technique of writing f as the mean of a random variable allows for more straightforward use of empirical process theory to bound the expected sup-error of the empirical average of m independent draws from its population mean. Our argument is also more flexible than [5] and can be readily adapted to the case of squared ReLU activation function. We should also point out that our $L^\infty(D)$ error bounds immediately imply $L^p(D)$ error bounds for all $p \geq 1$. In fact, using nearly exactly the same techniques, it can be shown that the results in Theorem 6, Theorem 7, and Theorem 8 hold verbatim in $L^2(D)$, sans the $\sqrt{d + \log m}$ or \sqrt{d} factors, corresponding to the ReLU or squared ReLU cases, respectively.

Remark 2. *In [62], it was shown that the standard order $m^{-1/2}$ $L^\infty(D)$ error bound alluded to earlier could be improved to be of order $\sqrt{\log m} m^{-1/2-1/(2d)}$ under an alternate condition of finite $v_{f,1}^* \triangleq \sup_{u \in \mathbb{S}^{d-1}} \int_0^\infty r^d |\mathcal{F}(f)(ru)| dr$, but with the requirement that $\|a_k\|_1$ be unbounded. In general, our assumptions are neither stronger nor weaker than this since*

the function f with Fourier transform $\mathcal{F}(f)(\omega) = e^{-\|\omega - \omega_0\|}/\|\omega - \omega_0\|$ for $\omega_0 \neq 0$ and $d \geq 2$ has infinite $v_{f,1}^*$ but finite $v_{f,s}$ for $s \geq 0$, while the function f with Fourier transform $\mathcal{F}(f)(\omega) = 1/(1 + \|\omega\|)^{d+2}$ has finite $v_{f,1}^*$ but infinite $v_{f,s}$ for $s \geq 2$.

Proof of Theorem 6. Case I: $s = 2$. Let $\mathcal{B}_1, \dots, \mathcal{B}_M$ be a partition of the space $\Omega = \{(\eta, t, a)' : \eta \in \{-1, +1\}, 0 \leq t \leq 1, \|a\|_1 = 1\}$ such that

$$\inf_{(\tilde{\eta}, \tilde{t}, \tilde{a})' \in \mathcal{B}_k, k=1, \dots, M} \sup_{(\eta, t, a)' \in \Omega} \|h(\tilde{\eta}, \tilde{t}, \tilde{a}) - h(\eta, t, a)\|_\infty < \epsilon, \quad (3.6)$$

where $h(\eta, t, a)(x) = h(x) = \eta(a \cdot x - t)_+^{s-1}$. It is not hard to show that $M \asymp \epsilon^{-d}$. For $k = 1, \dots, M$ define

$$dP_k(t, a) = dP(t, a) \mathbb{1}\{(\eta(t, a), t, a)' \in \mathcal{B}_k\} / L_k,$$

where L_k is chosen to make P_k a probability measure. A very important property we will use is that $\text{Var}_{P_k}[h] \leq \epsilon$, which follows from (3.6). Let m be a positive integer and define a sequence of M independent random variables $\{m_k\}_{1 \leq k \leq M}$ as follows: let m_k equal $\lfloor mL_k \rfloor$ and $\lceil mL_k \rceil$ with probabilities chosen to make its mean equal to mL_k . Given, $\underline{m} = \{m_k\}_{1 \leq k \leq M}$, take a random sample $\underline{a} = \{(t_{j,k}, a_{j,k})'\}_{1 \leq j \leq n_k, 1 \leq k \leq M}$ of size $n_k = m_k + \mathbb{1}\{m_k = 0\}$ from P_k . Thus, we split the population Ω into M “strata” $\mathcal{B}_1, \dots, \mathcal{B}_M$ and allocate the number of within-stratum samples to be proportional to the “size” of the stratum m_1, \dots, m_M (i.e., proportionate allocation). The within-stratum variability of h (i.e., $\text{Var}_{P_k}[h]$) is now smaller than the population level variability (i.e., $\text{Var}_P[h]$) by a factor of ϵ as evidenced by (3.6).

Note that the n_k sum to be at most $m + M$ because

$$\begin{aligned}
\sum_{k=1}^M n_k &= \sum_{k=1}^M m_k \mathbb{1}\{m_k > 0\} + \sum_{k=1}^M \mathbb{1}\{m_k = 0\} \\
&\leq \sum_{k=1}^M (mL_k + 1) \mathbb{1}\{m_k > 0\} + \sum_{k=1}^M \mathbb{1}\{m_k = 0\} \\
&= m \sum_{k=1}^M L_k \mathbb{1}\{m_k > 0\} + M \\
&\leq m + M,
\end{aligned} \tag{3.7}$$

where the last inequality follows from $\sum_{k=1}^M L_k \leq 1$. For $j = 1, \dots, m_k$, let $h_{j,k} = h(\eta(t_{j,k}, a_{j,k}), t_{j,k}, a_{j,k})$ and $f_k = \frac{vm_k}{mn_k} \sum_{j=1}^{n_k} h_{j,k}$. Also, let $\bar{f}_m = \sum_{k=1}^M f_k$. A simple calculation shows that the mean of \bar{f}_m is f . Write $\sum_{k=1}^M (f_k(x) - \mathbb{E}f_k(x)) = \frac{v}{m} \left(\sum_{k=1}^M (m_k - L_k m) \mathbb{E}_{P_k} h(x) \right) + \frac{v}{m} \left(\sum_{k=1}^M \sum_{j=1}^{n_k} \frac{m_k}{n_k} (h_{j,k}(x) - \mathbb{E}_{P_k} h(x)) \right)$. By the triangle inequality, we upper bound

$$\mathbb{E} \sup_{x \in D} |\bar{f}_m(x) - f(x)| = \mathbb{E} \sup_{x \in D} \left| \sum_{k=1}^M (f_k(x) - \mathbb{E}f_k(x)) \right|$$

by

$$\begin{aligned}
&\frac{v}{m} \mathbb{E}_m \sup_{x \in D} \left| \sum_{k=1}^M (m_k - L_k m) \mathbb{E}_{P_k} h(x) \right| + \\
&\frac{v}{m} \mathbb{E}_m \mathbb{E}_{a|m} \sup_{x \in D} \left| \sum_{k=1}^M \sum_{j=1}^{n_k} \frac{m_k}{n_k} (h_{j,k}(x) - \mathbb{E}_{P_k} h(x)) \right|.
\end{aligned} \tag{3.8}$$

Now

$$\begin{aligned}
&\mathbb{E}_{a|m} \sup_{x \in D} \left| \sum_{k=1}^M \sum_{j=1}^{n_k} \frac{m_k}{n_k} (h_{j,k}(x) - \mathbb{E}_{P_k} h(x)) \right| \leq \\
&2 \mathbb{E}_{a|m} \sup_{x \in D} \left| \sum_{k=1}^M \sum_{j=1}^{n_k} \sigma_{j,k} \frac{m_k}{n_k} [h_{j,k}(x) - \mu_{j,k}(x)] \right|,
\end{aligned} \tag{3.9}$$

where $\{\sigma_{j,k}\}$ is a sequence of independent identically distributed Rademacher variables and $\{x \mapsto \mu_{j,k}(x)\}$ is any sequence of functions defined on D [see for example Lemma 2.3.6

in [63]]. For notational brevity, we define $\tilde{h}_{j,k}(x) = \frac{m_k}{n_k}[h_{j,k}(x) - \mu_{j,k}(x)]$. By Dudley's entropy integral method [see Corollary 13.2 in [64]], the quantity in (3.9) can be bounded by

$$24 \int_0^{\delta/2} \sqrt{N(u, D)} du, \quad (3.10)$$

where $N(u, D)$ is the u -metric entropy of D with respect to the norm $\kappa(x, x')$ (i.e., the logarithm of the smallest u -net that covers D with respect to κ) defined by

$$\begin{aligned} \kappa^2(x, x') &\triangleq \sum_{k=1}^M \sum_{j=1}^{n_k} (\tilde{h}_{j,k}(x) - \tilde{h}_{j,k}(x'))^2 \\ &\leq (m + M) \|x - x'\|_\infty^2, \end{aligned} \quad (3.11)$$

and $\delta^2 = \sup_{x \in D} \sum_{k=1}^M \sum_{j=1}^{n_k} |\tilde{h}_{j,k}(x)|^2$. If we set $\mu_{j,k}$ to equal $\frac{m_k}{n_k} h(\eta(t_k, a_k), t_k, a_k)$, where $(\eta_k, t_k, a_k)'$ is any fixed point in \mathcal{B}_k , it follows from (3.6) and (3.7) that $\delta \leq \sqrt{m + M} \epsilon$ and from (3.11) that $N(u, D) \leq d \log(3\sqrt{m + M}/u)$. By evaluating the integral in (3.10), we can bound the second term in (3.8) by

$$24v\sqrt{d} m^{-1/2} \epsilon \sqrt{-\log \epsilon + 1} \sqrt{1 + M/m}. \quad (3.12)$$

For the first expectation in (3.8), we follow a similar approach. As before,

$$\begin{aligned} &\mathbb{E}_{\underline{m}} \sup_{x \in D} \left| \sum_{k=1}^M (m_k - L_k m) \mathbb{E}_{P_k} h(x) \right| \\ &\leq 2\mathbb{E}_{\underline{m}} \sup_{x \in D} \left| \sum_{k=1}^M \sigma_k (m_k - L_k m) \mathbb{E}_{P_k} h(x) \right|, \end{aligned} \quad (3.13)$$

where $\{\sigma_k\}$ is a sequence of independent identically distributed Rademacher variables. For notational brevity, we write $\tilde{h}_k(x) = (m_k - L_k m) \mathbb{E}_{P_k} h(x)$. We can also bound (3.13) by (3.10), except this time $N(u, D)$ is the u -metric entropy of D with respect to the norm

$\rho(x, x')$ defined by

$$\begin{aligned}\rho^2(x, x') &\triangleq \sum_{k=1}^M (\tilde{h}_k(x) - \tilde{h}_k(x'))^2 \\ &\leq M \|x - x'\|_\infty^2,\end{aligned}\tag{3.14}$$

where the last line follows from $|m_k - L_k m| \leq 1$ and $|\mathbb{E}_{P_k} h(x) - \mathbb{E}_{P_k} h(x')| \leq \|x - x'\|_\infty$. The quantity δ is also less than \sqrt{M} , since $\sup_{x \in D} |\tilde{h}_k(x)| \leq 1$ and moreover $N(u, D) \leq d \log(3\sqrt{M}/u)$. Evaluating the integral in (3.10) with these specifications yields a bound on the first term in (3.8) of

$$\frac{48v\sqrt{d}\sqrt{M}}{m}.\tag{3.15}$$

Adding (3.15) and (3.12) together yields a bound on $\mathbb{E} \sup_{x \in D} |\bar{f}_m(x) - f(x)|$ of

$$48v\sqrt{d}m^{-1/2}(\sqrt{M/m} + \epsilon\sqrt{1 + M/m}\sqrt{-\log \epsilon + 1}).\tag{3.16}$$

Choose

$$M = m \frac{\epsilon^2(-\log \epsilon + 1)}{1 - \epsilon^2(-\log \epsilon + 1)}.\tag{3.17}$$

Consequently, $\mathbb{E} \sup_{x \in D} |\bar{f}_m(x) - f(x)|$ is at most

$$96v\sqrt{d}m^{-1/2} \frac{\epsilon\sqrt{-\log \epsilon + 1}}{\sqrt{1 - \epsilon^2(-\log \epsilon + 1)}}.\tag{3.18}$$

We stated earlier that $M \asymp \epsilon^{-d}$. Thus (3.17) determines ϵ to be at most of order $m^{-1/(d+2)}$. Since the inequality (3.17) holds on average, there is a realization of \bar{f}_m for which $\sup_{x \in D} |\bar{f}_m(x) - f(x)|$ has the same bound. Note that \bar{f}_m has the desired equally weighted form.

For the second conclusion, we set $m_k = mL_k$ and $n_k = \lceil m_k \rceil$. In this case, the first term in (3.8) is zero and hence $\mathbb{E} \sup_{x \in D} |\bar{f}_m(x) - f(x)|$ is not greater than (3.12). The conclusion follows with $M = m$ and ϵ of order $m^{-1/d}$.

Case II: $s = 3$. The metric $\kappa(x, x')$ is in fact bounded by a constant multiple of

$\sqrt{m+M}\epsilon\|x-x'\|_\infty$. To see this, we note that the function $\tilde{h}_{j,k}(x)$ has the form

$$\pm \frac{m_k}{n_k} [(a \cdot x - t)_+^2 - (a_k \cdot x - t_k)_+^2],$$

with $\|a - a_k\|_1 + |t - t_k| < \epsilon$. Thus, the gradient of $\tilde{h}_{j,k}(x)$ with respect to x has the form

$$\nabla \tilde{h}_{j,k}(x) = \pm \frac{2m_k}{n_k} [(a(a \cdot x - t)_+ - a_k(a_k \cdot x - t_k)_+)].$$

Adding and subtracting $\frac{2m_k}{n_k} a(a_k \cdot x - t_k)_+$ to the above expression yields the bound of order ϵ for $\sup_{x \in D} \|\nabla \tilde{h}_{j,k}(x)\|_1$. Taylor's theorem yields the desired bound on $\kappa(x, x')$. Again using Dudley's entropy integral, we can bound $\mathbb{E} \sup_{x \in D} |\bar{f}_m(x) - f(x)|$ by a universal constant multiple of either $v\sqrt{d}m^{-1/2}(\sqrt{M/m} + \epsilon\sqrt{1+M/m})$ or $v\sqrt{d}m^{-1/2}\epsilon\sqrt{1+M/m}$ corresponding to the equally weighted or non-equally weighted cases, respectively. The corresponding results follow with $M = m\epsilon^2/(1 - \epsilon^2)$ and ϵ of order $m^{-1/(d+2)}$ or $M = m$ and ϵ of order $m^{-1/d}$. Note that here the additional smoothness afforded by the stronger assumption $v_{f,3} < +\infty$ allows one to remove the $\sqrt{-\log \epsilon + 1}$ factor that appeared in the final bound in the proof of Theorem 7. This rate is the same as what was achieved in Theorem 7, without a $\sqrt{(\log m)/d + 1}$ factor. \square

Proof of Theorem 7. If $|z| \leq c$, we note the identity

$$-\int_0^c [(z-u)_+ e^{iu} + (-z-u)_+ e^{-iu}] du = e^{iz} - iz - 1. \quad (3.19)$$

If $c = \|\omega\|_1$, $z = \omega \cdot x$, $a = a(\omega) = \omega/\|\omega\|_1$, and $u = \|\omega\|_1 t$, $0 \leq t \leq 1$, we find that

$$\begin{aligned} -\|\omega\|_1^2 \int_0^1 [(a \cdot x - t)_+ e^{i\|\omega\|_1 t} + (-a \cdot x - t)_+ e^{-i\|\omega\|_1 t}] dt = \\ e^{i\omega \cdot x} - i\omega \cdot x - 1. \end{aligned}$$

Multiplying the above by $\mathcal{F}(f)(\omega) = e^{ib(\omega)}|\mathcal{F}(f)(\omega)|$, integrating over \mathbb{R}^d , and applying

Fubini's theorem yields

$$f(x) - x \cdot \nabla f(0) - f(0) = \int_{\mathbb{R}^d} \int_0^1 g(t, \omega) dt d\omega,$$

where

$$g(t, \omega) = -[(a \cdot x - t)_+ \cos(\|\omega\|_1 t + b(\omega)) + (-a \cdot x - t)_+ \cos(\|\omega\|_1 t - b(\omega))] \|\omega\|_1^2 |\mathcal{F}(f)(\omega)|.$$

Consider the probability measure on $\{-1, 1\} \times [0, 1] \times \mathbb{R}^d$ defined by

$$dP(z, t, \omega) = \frac{1}{v} |\cos(z\|\omega\|_1 t + b(\omega))| \|\omega\|_1^2 |\mathcal{F}(f)(\omega)| dt d\omega, \quad (3.20)$$

where

$$v = \int_{\mathbb{R}^d} \int_0^1 [|\cos(\|\omega\|_1 t + b(\omega))| + |\cos(\|\omega\|_1 t - b(\omega))|] \|\omega\|_1^2 |\mathcal{F}(f)(\omega)| dt d\omega \leq 2v_{f,2}.$$

Define a function $h(z, t, a)(x)$ that equals

$$(za \cdot x - t)_+ \eta(z, t, \omega),$$

where $\eta(z, t, \omega) = -\text{sgn} \cos(\|\omega\|_1 zt + b(\omega))$. Note that $h(z, t, a)(x)$ has the form $\pm(\pm a \cdot x - t)_+$.

Thus, we see that

$$f(x) - x \cdot \nabla f(0) - f(0) = v \int_{\{-1, 1\} \times [0, 1] \times \mathbb{R}^d} h(z, t, a)(x) dP(z, t, \omega). \quad (3.21)$$

The result follows from an application of Theorem 6. □

Proof of Theorem 8. For the result in Theorem 8, we will use exactly the same techniques.

The function $f(x) - x^T \nabla \nabla^T f(0)x/2 - x \cdot \nabla f(0) - f(0)$ can be written as the real part of

$$\int_{\mathbb{R}^d} (e^{i\omega \cdot x} + (\omega \cdot x)^2/2 - i\omega \cdot x - 1) \mathcal{F}(f)(\omega) d\omega. \quad (3.22)$$

As before, the integrand in (3.22) admits an integral representation given by

$$(i/2) \|\omega\|_1^3 \int_0^1 [(-a \cdot x - t)_+^2 e^{-i\|\omega\|_1 t} - (a \cdot x - t)_+^2 e^{i\|\omega\|_1 t}] dt,$$

which can be used to show that $f(x) - x^T \nabla \nabla^T f(0)x/2 - x \cdot \nabla f(0) - f(0)$ equals

$$\frac{v}{2} \int_{\{-1,1\} \times [0,1] \times \mathbb{R}^d} h(z, t, a)(x) dP(z, t, \omega), \quad (3.23)$$

where

$$h(z, t, a) = \text{sgn} \sin(z \|\omega\|_1 t + b(\omega)) (za \cdot x - t)_+^2$$

and

$$dP(z, t, \omega) = \frac{1}{v} |\sin(z \|\omega\|_1 t + b(\omega))| \|\omega\|_1^3 |\mathcal{F}(f)(\omega)| dt d\omega,$$

$$v = \int_{\mathbb{R}^d} \int_0^1 [|\sin(\|\omega\|_1 t + b(\omega))| + |\sin(\|\omega\|_1 t - b(\omega))|] \|\omega\|_1^3 |\mathcal{F}(f)(\omega)| dt d\omega \leq 2v_{f,3}.$$

The result follows from an application of Theorem 6. □

Remark 3. By slightly modifying the definition of h from the proofs of Theorem 7 and Theorem 8 (in particular, multiplying it by a sinusoidal function of ω and t), it suffices to sample instead from the density $dP(t, \omega) = \frac{\|\omega\|_1^s |\mathcal{F}(f)(\omega)|}{v_{f,s}} dt d\omega$ on $[0, 1] \times \mathbb{R}^d$.

Remark 4. For unit bounded x , the expression $e^{i\omega \cdot x} - i\omega \cdot x - 1$ is bounded in magnitude by $\|\omega\|_1^2$, so one only needs Fourier representation of $f(x) - x \cdot \nabla f(0) - f(0)$ when using the integrability with the $\|\omega\|_1^2$ factor. Similarly, $e^{i\omega \cdot x} + (\omega \cdot x)^2/2 - i\omega \cdot x - 1$ is bounded in magnitude by $\|\omega\|_1^3$, so one only needs Fourier representation of $f(x) - x^T \nabla \nabla^T f(0)x - x \cdot \nabla f(0) - 1$ when using the integrability with the $\|\omega\|_1^3$ factor.

Remark 5. Note that in Theorem 7 and Theorem 8, we work with integrals with respect to the absolutely continuous measure $d\mathcal{F}(f)(\omega)$. In general, a (complex) Fourier measure $d\mathcal{F}(f)(\omega)$ does not need to be absolutely continuous. For instance, it can be discrete on a lattice of values of ω , associated with a multivariate Fourier series representation for bounded domains x (and periodic extensions thereof). Indeed, for bounded domains, one might have access to both Fourier series and Fourier transforms of extensions of f to \mathbb{R}^d . The best extension is one that gives the smallest Fourier norm $\int_{\mathbb{R}^d} \|\omega\|_1^s |d\mathcal{F}(f)(\omega)|$. For further discussion along these lines, see [23].

Next, we investigate the optimality of the rates from Section 7.4.4.

3.2.2 Lower bounds

Let $\mathcal{H}_s = \{x \mapsto \eta(a \cdot x - t)_+^{s-1} : \|a\|_1 \leq 1, 0 \leq t \leq 1, \eta \in \{-1, +1\}\}$ and for $p \in [2, +\infty]$ let \mathcal{F}_p^s denote the closure of the convex hull of \mathcal{H}_s with respect to the $\|\cdot\|_p$ norm on $L^p(D, P)$ for p finite, where P is the uniform probability measure on D , and $\|\cdot\|_\infty$ (the supremum norm over D) for $p = +\infty$. We let \mathcal{C}_m^s denote the collection of all convex combinations of m terms from \mathcal{H}_s . By Theorem 7 and Theorem 8, after possibly subtracting a linear or quadratic term, $f/(2v_{f,2})$ and $f/v_{f,3}$ belongs to \mathcal{F}_p^2 and \mathcal{F}_p^3 , respectively. For $p \in [2, +\infty]$ and $\epsilon > 0$, we define the ϵ -covering number $N_p(\epsilon)$ by

$$\min\{n : \exists \mathcal{F} \subset \mathcal{F}_p^s, |\mathcal{F}| = n, \text{ s.t. } \inf_{f' \in \mathcal{F}} \sup_{f \in \mathcal{F}_p^s} \|f - f'\|_p \leq \epsilon\}.$$

and the ϵ -packing number $M_p(\epsilon)$ by

$$\max\{n : \exists \mathcal{F} \subset \mathcal{F}_p^s, |\mathcal{F}| = n, \text{ s.t. } \inf_{f, f' \in \mathcal{F}} \|f - f'\|_p > \epsilon\}.$$

Theorem 6 implies that $\inf_{f_m \in \mathcal{C}_m^s} \sup_{f \in \mathcal{F}_\infty^s} \|f - f_m\|_\infty$ achieves the bounds as stated therein.

Theorem 9. For $p \in [2, +\infty]$ and $s \in \{2, 3\}$,

$$\inf_{f_m \in \mathcal{C}_m^s} \sup_{f \in \mathcal{F}_p^s} \|f - f_m\|_p \geq (Amd^{2s+1} \log(md))^{-1/2-s/d},$$

for some universal positive constant A .

Ignoring the dependence on d and logarithmic factors in m , this result coupled with Theorem 6 implies that $\inf_{f_m \in \mathcal{C}_m^2} \sup_{f \in \mathcal{F}_p^2} \|f - f_m\|_p$ is between $m^{-1/2-2/d}$ and $m^{-1/2-1/d}$; for large d , the rates are essentially the same. Compare this with [55, Theorem 4] or [3, Theorem 3], where a lower bound of $c(\delta, d) m^{-1/2-1/d-\delta}$, $\delta > 0$ arbitrary, was obtained for approximants of the form (3.1) with Lipschitz ϕ , but with inner parameter vectors of unbounded ℓ^1 norm.

We only give the proof of Theorem 9 for $s = 2$, since the other case $s = 3$ is handled similarly. First, we provide a few ancillary results that will be used later on. The next result is contained in [65, Lemma 4.2] and is useful for giving a lower bound on $M_p(\epsilon)$.

Lemma 12. *Let H be a Hilbert space equipped with a norm $\|\cdot\|$ and containing a finite set \mathcal{H} with the following properties.*

- (i) $|\mathcal{H}| \geq 3$,
- (ii) $\sum_{h, h' \in \mathcal{H}, h \neq h'} |\langle h, h' \rangle| \leq \delta^2$
- (iii) $\delta^2 \leq \min_{h \in \mathcal{H}} \|h\|^2$

Then there exists a collection $\Omega \subset \{0, 1\}^{|\mathcal{H}|}$ with cardinality at least $2^{(1-H(1/4))|\mathcal{H}|-1}$, where $H(1/4)$ is the entropy of a Bernoulli random variable with success probability $1/4$, such that each pair of elements in the set $\mathcal{F} = \left\{ \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \omega_h h : (\omega_h : h \in \mathcal{H}) \in \Omega \right\}$ is separated by at least $\frac{1}{2} \sqrt{\frac{\min_{h \in \mathcal{H}} \|h\|^2 - \delta^2}{|\mathcal{H}|}}$ in $\|\cdot\|$.

Lemma 13. *If θ belongs to $[R]^d = \{1, 2, \dots, R\}^d$, $R \in \mathbb{Z}^+$, then the collection of functions*

$$\mathcal{H} = \{x \mapsto \sin(\pi \theta \cdot x) / (4\pi \|\theta\|_1^2) : \theta \in [R]^d\}$$

satisfies the assumption of Lemma 12 with $H = L^2(D, P)$, where P is the uniform probability measure on D . Moreover, $|\mathcal{H}| = R^d$, $\delta = 0$, $\min_{h \in \mathcal{H}} \|h\| = 1/(4\sqrt{2}\pi d^2 R^2)$, and $\mathcal{F} \subset \mathcal{F}_p^1$ for

all $p \in [2, +\infty]$. Consequently, if $\epsilon = 1/(8\sqrt{2}\pi d^2 R^{2+d/2})$, then

$$\begin{aligned} \log M_p(\epsilon) &\geq (\log 2)(1 - H(1/4)) \left(8\epsilon\sqrt{2}\pi d^2\right)^{-\frac{2d}{4+d}} - 1 \\ &\geq (c\epsilon d^2)^{-\frac{2d}{4+d}}, \end{aligned} \tag{3.24}$$

for some universal constant $c > 0$.

Proof. We first observe the identity

$$\begin{aligned} \sin(\pi\theta \cdot x)/(4\pi\|\theta\|_1^2) &= \theta \cdot x/(4\pi\|\theta\|_1^2) + \\ &\frac{\pi}{4} \int_0^1 [(-a \cdot x - t)_+ - (a \cdot x - t)_+] \sin(\pi\|\theta\|_1 t) dt, \end{aligned}$$

where $a = a(\theta) = \theta/\|\theta\|_1$. Note that above integral can also be written as an expectation of

$$-z \operatorname{sgn}(\sin(\pi\|\theta\|_1 t)) (za \cdot x - t)_+ \in \mathcal{H}_2$$

with respect to the density

$$p_\theta(z, t) = \frac{\pi}{4} |\sin(\pi\|\theta\|_1 t)|,$$

on $\{-1, 1\} \times [0, 1]$. The fact that p_θ integrates to one is a consequence of the identity

$$\int_0^1 |\sin(\pi\|\theta\|_1 t)| dt = 2/\pi.$$

Since $\int_D |\sin(\pi\theta \cdot x)|^2 dP(x) = 1/2$, each member of \mathcal{H} has norm equal to $1/(4\sqrt{2}\pi\|\theta\|_1^2)$ and each pair of elements is orthogonal so that $\delta = 0$. Integrations over D involving $\sin(\pi\theta \cdot x)$ are easiest to see using an instance of Euler's formula, viz., $\sin(\alpha \cdot x) = \frac{1}{2i} (\prod_{k=1}^d e^{i\alpha(k)x(k)} - \prod_{k=1}^d e^{-i\alpha(k)x(k)})$. \square

Proof of Theorem 9. Let $A > 0$ be arbitrary. Suppose contrary to the hypothesis,

$$\inf_{f_m \in \mathcal{C}_m^2} \sup_{f \in \mathcal{F}_p^2} \|f - f_m\|_p < (Amd^5 \log(md))^{-1/2-2/d} \\ \triangleq \epsilon_0/3.$$

Note that each element of \mathcal{C}_m^2 has the form $\sum_{k=1}^m \lambda_k h_k$, where $\sum_{k=1}^m \lambda_k = 1$ and $h_k \in \mathcal{H}_s$. Next, consider the subcollection $\tilde{\mathcal{C}}_m^2$ with elements of the form $\sum_{k=1}^m \tilde{\lambda}_k \tilde{h}_k$, where $\tilde{\lambda}_k$ belongs to an $\epsilon_0/3$ -net $\tilde{\mathcal{P}}$ of the $m-1$ dimensional probability simplex \mathcal{P}_m and \tilde{h}_k belongs to an $\epsilon_0/3$ -net $\tilde{\mathcal{H}}$ of \mathcal{H}_s . By a stars and bars argument, there are at most $|\tilde{\mathcal{P}}| \binom{m+|\mathcal{H}|}{m}^{m-1}$ such functions. Furthermore, since $\sup_{h \in \mathcal{H}_s} \|h\|_\infty \leq 1$, we have

$$\inf_{f_m \in \tilde{\mathcal{C}}_m^2} \sup_{f \in \mathcal{F}_p^2} \|f - f_m\|_2 \leq \inf_{f_m \in \mathcal{C}_m^2} \sup_{f \in \mathcal{F}_p^2} \|f - f_m\|_2 + \\ \inf_{\tilde{h} \in \tilde{\mathcal{H}}} \sup_{h \in \mathcal{H}_s} \|h - \tilde{h}\|_2 + \\ \inf_{\tilde{\lambda} \in \tilde{\mathcal{P}}} \sup_{\lambda \in \mathcal{P}_m} \|\lambda - \tilde{\lambda}\|_1 \\ < \epsilon_0/3 + \epsilon_0/3 + \epsilon_0/3 = \epsilon_0.$$

Since $|\tilde{\mathcal{H}}| \asymp \epsilon_0^{-d-1}$ and $|\tilde{\mathcal{P}}| \asymp \epsilon_0^{-m+1}$, it follows that

$$\log N_p(\epsilon_0) \leq \log |\tilde{\mathcal{C}}_m^2| \\ \leq c_0 \log \left[\epsilon_0^{-m-1} \binom{m + c_1 \epsilon_0^{-d-1} - 1}{m} \right] \\ \leq c_2 dm \log(1/\epsilon_0) \\ \leq c_3 dm \log(Adm), \tag{3.25}$$

for some positive universal constants $c_0 > 0$, $c_1 > 0$, $c_2 > 0$, and $c_3 > 0$.

On the other hand, using (3.24) from Lemma 13 coupled with the fact that $N_p(\epsilon_0) \geq$

$M_p(2\epsilon_0)$, we have

$$\begin{aligned}
\log N_p(\epsilon_0) &\geq \log M_p(2\epsilon_0) \\
&\geq (2c\epsilon_0 d^2)^{-\frac{2d}{4+d}} \\
&\geq c_4 Adm \log(dm),
\end{aligned} \tag{3.26}$$

for some universal constant $c_4 > 0$. Combining (3.25) and (3.26), we find that

$$c_4 Adm \log(dm) \leq c_3 dm \log(Adm).$$

If A is large enough (independent of m or d), we reach a contradiction. This proves the lower bound. \square

3.3 L^2 approximation with bounded ℓ^0 and ℓ^1 norm

In Section 7.4.4, we explored conditions for which good approximation in $L^\infty(D)$ could be achieved even with ℓ^1 controls on the inner parameter vectors. In this section, we show how similar statements can be made in $L^2(D)$, but with control on the ℓ^0 norm as well. Note that unlike Theorem 6, we see in Theorem 10 how the smoothness of the activation function directly affects the rate of approximation. The proof is obtained by applying the Jones-Barron probabilistic method in two stages (similar to two-stage cluster sampling), first on the outer layer coefficients, and then on the inner layer coefficients.

Theorem 10. *Suppose f admits an integral representation*

$$f(x) = v \int_{[0,1] \times \{a: \|a\|_1=1\}} \eta(t, a) (a \cdot x - t)_+^{s-1} dP(t, a),$$

for x in $D = [-1, 1]^d$ and $s \in \{2, 3\}$, where P is a probability measure on $[0, 1] \times \{a \in \mathbb{R}^d : \|a\|_1 = 1\}$ and $\eta(t, a)$ is either -1 or $+1$. There exists a linear combination of ridge functions of the form

$$f_{m,m_0}(x) = \frac{v}{m} \sum_{k=1}^m b_k (a_k \cdot x - t_k)_+^{s-1},$$

where $\|a_k\|_0 \leq m_0$, $\|a_k\|_1 = 1$, and $b_k \in \{-1, +1\}$ such that

$$\|f - f_{m,m_0}\|_2 \leq v \sqrt{\frac{1}{m} + \frac{1}{m_0^{s-1}}}.$$

Furthermore, the same rates for $s = 2$ or $s = 3$ are achieved for general f adjusted by a linear or quadratic term with $v = 2v_{f,2} < +\infty$ or $v = v_{f,3} < +\infty$, respectively.

Remark 6. In particular, taking $m_0 = \sqrt{m}$, it follows that there exists an m -term linear combination of squared ReLU ridge functions, with \sqrt{m} -sparse inner parameter vectors, that approximates f with $L^2(D)$ error at most $\sqrt{2}vm^{-1/2}$. In other words, the $L^2(D)$ approximation error is inversely proportional to the inner layer sparsity and it need only be sublinear in the outer layer sparsity.

Proof. Take a random sample $\underline{a} = \{(t_k, a_k)'\}_{1 \leq k \leq m}$ from P . Given \underline{a} , take a random sample $\tilde{\underline{a}} = \{\tilde{a}_{\ell,k}\}_{1 \leq \ell \leq m_0, 1 \leq k \leq m}$, where $\mathbb{P}[\tilde{a}_{\ell,k} = \text{sgn}(a_k(j))e_j] = |a_k(j)|$ for $j = 1, \dots, d$, $a_k = (a_k(1), \dots, a_k(d))'$, and e_j is the j -th standard basis vector for \mathbb{R}^d . Note that

$$\mathbb{E}_{\tilde{\underline{a}}|\underline{a}}[\tilde{a}_{\ell,k}] = a_k \quad (3.27)$$

and

$$\begin{aligned} \text{Var}_{\tilde{\underline{a}}|\underline{a}}[\tilde{a}_{\ell,k} \cdot x] &\leq \mathbb{E}_{\tilde{\underline{a}}|\underline{a}}[\tilde{a}_{\ell,k} \cdot x]^2 = \sum_{j=1}^d |a_k(j)| |x(j)|^2 \\ &\leq \|a_k\|_1 \|x\|_\infty^2 \leq 1. \end{aligned} \quad (3.28)$$

Define

$$\bar{f}_{m,m_0}(x) = \frac{v}{m} \sum_{k=1}^m \eta(t_k, a_k) \left(\frac{1}{m_0} \sum_{\ell=1}^{m_0} \tilde{a}_{\ell,k} \cdot x - t_k \right)_+^{s-1}. \quad (3.29)$$

By the bias-variance decomposition,

$$\mathbb{E}\|f - \bar{f}_{m,m_0}\|_2^2 = \mathbb{E}\|\bar{f}_{m,m_0} - \mathbb{E}\bar{f}_{m,m_0}\|_2^2 + \|f - \mathbb{E}\bar{f}_{m,m_0}\|_2^2.$$

Note that $\mathbb{E}\|\bar{f}_{m,m_0} - \mathbb{E}\bar{f}_{m,m_0}\|_2^2 \leq \frac{v^2}{m}$. Next, observe that

$$f(x) - \mathbb{E}\bar{f}_{m,m_0}(x) = \frac{v}{m} \sum_{k=1}^m \mathbb{E}_{\underline{a}} \left[\eta(t_k, a_k) \times \right. \\ \left. \mathbb{E}_{\tilde{a}|\underline{a}} \left((a_k \cdot x - t_k)_+^{s-1} - \left(\frac{1}{m_0} \sum_{\ell=1}^{m_0} \tilde{a}_{\ell,k} \cdot x - t_k \right)_+^{s-1} \right) \right],$$

which, by an application of the triangle inequality, implies that

$$|f(x) - \mathbb{E}\bar{f}_{m,m_0}(x)| \leq \frac{v}{m} \sum_{k=1}^m \mathbb{E}_{\underline{a}} \left| (a_k \cdot x - t_k)_+^{s-1} - \mathbb{E}_{\tilde{a}|\underline{a}} \left(\frac{1}{m_0} \sum_{\ell=1}^{m_0} \tilde{a}_{\ell,k} \cdot x - t_k \right)_+^{s-1} \right|.$$

Next, we use the following two properties of $(z)_+^{s-1}$: for all z and z' in \mathbb{R} ,

$$|(z)_+ - (z')_+| \leq |z - z'|, \quad (3.30)$$

$$|(z)_+^2 - (z')_+^2 - 2(z - z')(z')_+| \leq |z - z'|^2. \quad (3.31)$$

If $s = 2$, we have by (3.30), (3.27), and (3.28) that

$$\begin{aligned} & \mathbb{E}_{\underline{a}} \left| (a_k \cdot x - t_k)_+ - \mathbb{E}_{\tilde{a}|\underline{a}} \left(\frac{1}{m_0} \sum_{\ell=1}^{m_0} \tilde{a}_{\ell,k} \cdot x - t_k \right)_+ \right| \leq \\ & \mathbb{E}_{\underline{a}} \mathbb{E}_{\tilde{a}|\underline{a}} \left| a_k \cdot x - \frac{1}{m_0} \sum_{\ell=1}^{m_0} \tilde{a}_{\ell,k} \cdot x \right| \leq \\ & \mathbb{E}_{\underline{a}} \sqrt{\mathbb{E}_{\tilde{a}|\underline{a}} \left| a_k \cdot x - \frac{1}{m_0} \sum_{\ell=1}^{m_0} \tilde{a}_{\ell,k} \cdot x \right|^2} = \\ & \mathbb{E}_{\underline{a}} \sqrt{\frac{\text{Var}_{\tilde{a}|\underline{a}}[\tilde{a}_{\ell,k} \cdot x]}{m_0}} \leq \frac{1}{\sqrt{m_0}}. \end{aligned}$$

This shows that $\|f - \mathbb{E}\bar{f}_{m,m_0}\|_2^2 \leq \frac{v^2}{m_0}$. If $s = 3$, we have from (3.31), (3.27), and (3.28) that

$$\begin{aligned} & \mathbb{E}_{\underline{a}} \left| (a_k \cdot x - t_k)_+^2 - \mathbb{E}_{\underline{\tilde{a}}|\underline{a}} \left(\frac{1}{m_0} \sum_{\ell=1}^{m_0} \tilde{a}_{\ell,k} \cdot x - t_k \right)_+^2 \right| \leq \\ & \mathbb{E}_{\underline{a}} \mathbb{E}_{\underline{\tilde{a}}|\underline{a}} \left| a_k \cdot x - \frac{1}{m_0} \sum_{\ell=1}^{m_0} \tilde{a}_{\ell,k} \cdot x \right|^2 = \\ & \mathbb{E}_{\underline{a}} \left[\frac{\text{Var}_{\underline{\tilde{a}}|\underline{a}}[\tilde{a}_{\ell,k} \cdot x]}{m_0} \right] \leq \frac{1}{m_0}. \end{aligned}$$

This shows that $\|f - \mathbb{E}\bar{f}_{m,m_0}\|_2^2 \leq \frac{v^2}{m_0^2}$. Since these bounds hold on average, there exists a realization of (3.29) for which the bounds are also valid. Note that the vector $\frac{1}{m_0} \sum_{\ell=1}^{m_0} \tilde{a}_{\ell,k}$ has ℓ^0 norm at most m_0 and unit ℓ^1 norm.

The fact that the bounds also hold for f adjusted by a linear or quadratic term (under an assumption of finite $v_{f,2}$ or $v_{f,3}$) follows from (3.21) and (3.23). \square

Chapter 4

Minimax lower bounds for ridge combinations including neural nets

4.1 Introduction

As seen in Chapter 2 and Chapter 3, ridge combinations provide flexible classes for fitting functions of many variables. The ridge activation function may be a general Lipschitz function. When the ridge activation function is a sigmoid, these are single-hidden layer artificial neural nets. When the activation is a sine or cosine function, it is a sinusoidal model in a ridge combination form. We consider also a class of polynomial nets which are combinations of Hermite ridge functions. Ridge combinations are also the functions used in projection pursuit regression fitting. What distinguishes these models from other classical functional forms is the presence of parameters internal to the ridge functions which are free to be adjusted in the fit. In essence, it is a parameterized, infinite dictionary of functions from which we make linear combinations. This provides a flexibility of function modeling not present in the case of a fixed dictionary. Here we discuss results on risk properties of estimation of functions using these models and we develop new minimax lower bounds.

For a given activation function $\phi(z)$ on \mathbb{R} , consider the parameterized family \mathcal{F}_m of functions

$$f_m(x) = f_m(x, c_0, c_1, b) = \sum_{k=1}^m c_{1,k} \phi(\sum_{j=1}^d c_{0,j,k} x_j - b_k), \quad (4.1)$$

where $c_1 = (c_{1,1}, \dots, c_{1,m})'$ is the vector of outer layer parameters and $c_{0,k} = (c_{0,1,k}, \dots, c_{0,d,k})'$ are the vectors of inner parameters for the single hidden-layer of functions $\phi(c_{0,k} \cdot x - b_k)$ with horizontal shifts $b = (b_1, \dots, b_m)$, $k = 1, \dots, m$. For positive v_0 , let

$$\mathcal{D}_{v_0} = \mathcal{D}_{v_0, \phi} = \{\phi(\theta \cdot x - t), x \in B : \|\theta\|_1 \leq v_0, t \in \mathbb{R}\} \quad (4.2)$$

be the dictionary of all such inner layer ridge functions $\phi(\theta \cdot x - t)$ with parameter restricted to the ℓ_1 ball of size v_0 and variables x restricted to the cube $[-1, 1]^d$. The choice of the ℓ_1 norm on the inner parameters is natural as it corresponds to $\|\theta\|_B = \sup_{x \in B} |\theta \cdot x|$ for $B = [-1, 1]^d$.

Let $\mathcal{F}_{v_0, v_1} = \mathcal{F}_{v_0, v_1, \phi} = \ell_1(v_1, \mathcal{D}_{v_0})$ be the closure of the set of all linear combinations of functions in \mathcal{D}_{v_0} with ℓ_1 norm of outer coefficients not more than v_1 . These v_0 and v_1 control the freedom in the size of this function class. They can either be fixed for minimax evaluations, or adapted in the estimation (as reflected in some of the upper bounds on risk for penalized least square estimation). The functions of the form (4.1) are in $\ell_1(v_1, \mathcal{D})$ when $\|c_{0,k}\|_1 \leq v_0$ and $\|c_1\|_1 \leq v_1$. Indeed, let $\mathcal{F}_{m, v_0, v_1} = \ell_1(m, v_1, \mathcal{D}_{v_0})$ be the subset of such functions in $\ell_1(v_1, \mathcal{D}_{v_0})$ that use m terms.

Data are of the form $\{(X_i, Y_i)\}_{i=1}^n$, drawn independently from a joint distribution $P_{X,Y}$ with P_X on $[-1, 1]^d$. The target function is $f(x) = \mathbb{E}[Y|X = x]$, the mean of the conditional distribution $P_{Y|X=x}$, optimal in mean square for the prediction of future Y from corresponding input X . In some cases, assumptions are made on the error of the target function $\epsilon_i = Y_i - f(X_i)$ (i.e. bounded, Gaussian, or sub-Gaussian).

From the data, estimators $\hat{f}(x) = \hat{f}(x, \{(X_i, Y_i)\}_{i=1}^n)$ are formed and the loss at a target f is the $L_2(P_X)$ square error $\|f - \hat{f}\|^2$ and the risk is the expected squared error $\mathbb{E}\|f - \hat{f}\|^2$. For any class of functions \mathcal{F} on $[-1, 1]^d$, the minimax risk is

$$R_{n,d}(\mathcal{F}) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}\|f - \hat{f}\|^2, \quad (4.3)$$

where the infimum runs over all estimators \hat{f} of f based on the data $\{(X_i, Y_i)\}_{i=1}^n$.

It is known that for certain complexity penalized least squares estimators [22], [38], [33],

[37] the risk satisfies

$$\mathbb{E}\|f - \hat{f}\|^2 \leq \inf_{f_m \in \mathcal{F}_m} \{\|f - f_m\|^2 + \frac{cmd \log n}{n}\}, \quad (4.4)$$

where the constant c depends on parameters of the noise distribution and on properties of the activation function ϕ , which can be a step function or a fixed bounded Lipschitz function. The $d \log n$ in the second term is from the log-cardinality of customary d -dimensional covers of the dictionary. The right side is an index of resolvability expressing the tradeoff between approximation error $\|f - f_m\|^2$ and descriptive complexity $md \log n$ relative to sample size, in accordance with risk bounds for minimum description length criteria [66], [28], [29], [67]. When the target f is in \mathcal{F}_{v_1, v_0} , it is known as in [51], [23], [5] that $\|f - f_m\|^2 \leq v_1^2/m$ with slight improvements possible depending on the dimension $\|f - f_m\|^2 \leq v_1^2/m^{1/2+1/d}$ as in [55], [57], [32]. When f is not in \mathcal{F}_{v_0, v_1} , let f_{v_0, v_1} be its projection onto this convex set of functions. Then the additional error beyond $\|f - f_{v_0, v_1}\|^2$ is controlled by the bound [22]

$$\inf_m \left\{ \frac{v_1^2}{m} + \frac{c_1 m d \log n}{n} \right\} = 2v_1 \left(\frac{c_1 d \log n}{n} \right)^{1/2}. \quad (4.5)$$

Moreover, with \hat{f} restricted to \mathcal{F}_{v_0, v_1} , this bounds the mean squared error $\mathbb{E}\|\hat{f} - f_{v_0, v_1}\|^2$ from the projection. The same risk is available from ℓ_1 penalized least square estimation [33], [28], [29], [57] and from greedy implementations of complexity and ℓ_1 penalized estimation [33], [57]. The slight approximation improvements (albeit not known whether available by greedy algorithms) provide the risk bound [57]

$$R_{n,d}(\mathcal{F}_{v_0, v_1}) \leq c_2 \left(\frac{dv_0^2 v_1^2}{n} \right)^{1/2+1/(2(d+1))}, \quad (4.6)$$

for bounded Lipschitz activation functions ϕ , improving a similar result in [54], [32]. This fact can be shown through improved upper bounds on the metric entropy from [53].

A couple of lower bounds on the minimax risk in \mathcal{F}_{v_0, v_1} are known [32] and, improving on [32], the working paper [57] states the lower bound

$$R_{n,d}(\mathcal{F}_{v_0, v_1}) \geq c_3 v_1^{d/(d+2)} \left(\frac{1}{d^4 n} \right)^{1/2+1/(d+2)} \quad (4.7)$$

for an unconstrained v_0 .

Note that for large d , these exponents are near $1/2$. Indeed, if d is large compared to $\log n$, then the bounds in (4.6) and (4.7) are of the same order as with exponent $1/2$. It is desirable to have improved lower bounds which take the form d/n to a fractional power as long as d is of smaller order than n .

Good empirical performance of neural net (and neural net like) models has been reported as in [25] even when d is much larger than n , though theoretical understanding has been lacking. In Chapter 2, we obtained risk upper bounds of the form

$$R_{n,d}(\mathcal{F}_{v_0,v_1}) \leq c_4 \left(\frac{v_0^2 v_1^4 \log(d+1)}{n} \right)^\gamma, \quad (4.8)$$

for fixed positive γ , again for bounded Lipschitz ϕ . These allow d much larger than n , as long as $d = e^{o(n)}$. With greedy implementations of least squares over a discretization of the parameter with complexity or ℓ_1 penalty, such upper bounds are obtained in Chapter 2 with $\gamma = 1/3$ and $\gamma = 2/5$. At the expense of a slightly worse exponent on v_1 and an additional smoothness assumption on ϕ , the rate with $\gamma = 1/3$ or $\gamma = 2/5$ is also possible when the greedy algorithm selects candidate neurons from a continuum of choices.

It is desirable likewise to have lower bounds on the minimax risk for this setting that show that it depends primarily on $v_0^\alpha v_1^{2\alpha}/n$ to some power (within $\log d$ factors). It is the purpose of this chapter to obtain such lower bounds. Here with $\gamma = 1/2$. Thereby, this chapter on lower bounds is to provide a companion to (refinement of) Chapter 2 or [57]. Lower bounding minimax risk in non-parametric regression is primarily an information-theoretic problem. This was first observed by [68] and then [69], [70] who adapted Fano's inequality in this setting. Furthermore, [32] showed conditions such that the minimax risk ϵ_n^2 is characterized (to within a constant factor) by solving for the approximation error ϵ^2 that matches the metric entropy relative to the sample size $(\log N(\epsilon))/n$, where $N(\epsilon)$ is the size of the largest ϵ -packing set. Accordingly, the core of our analysis is providing packing sets for \mathcal{F}_{v_0,v_1} for specific choices of ϕ .

4.2 Results for sinusoidal nets

We now state our main result. In this section, it is for the sinusoidal activation function $\phi(z) = \sqrt{2}\sin(\pi z)$. We consider two regimes: when d is larger than v_0 and visa-versa. In each case, this entails putting a non-restrictive technical condition on either quantity. For d larger than v_0 , this condition is

$$\frac{d}{v_0} + 1 > (c_4 \frac{v_1^2 n}{v_0 \log(1+d/v_0)})^{1/v_0}, \quad (4.9)$$

and when v_0 is larger than d ,

$$\frac{v_0}{d} + 1 > (c_5 \frac{v_1^2 n}{d \log(1+v_0/d)})^{1/d}, \quad (4.10)$$

for some positive constants c_4, c_5 . Note that when d is large compared to $\log n$, condition (4.10) holds. Indeed, the left side is at least 2 and the right side is $e^{\frac{1}{d} \log(\frac{v_1^2 n}{d \log(1+v_0/d)})}$, which is near 1. Likewise, (4.9) holds when v_0 is large compared to $\log n$.

Theorem 11. *Consider the model $Y = f(X) + \varepsilon$ for $f \in \mathcal{F}_{v_0, v_1, \text{ sine}}$, where $\varepsilon \sim N(0, 1)$ and $X \sim \text{Uniform}[-1, 1]^d$. If d is large enough so that (4.9) is satisfied, then*

$$R_{n,d}(\mathcal{F}_{v_0, v_1, \text{ sine}}) \geq c_6 \left(\frac{v_0 v_1^2 \log(1+d/v_0)}{n} \right)^{1/2}, \quad (4.11)$$

for some universal constant $c_6 > 0$. Furthermore, if v_0 is large enough so that (4.10) is satisfied, then

$$R_{n,d}(\mathcal{F}_{v_0, v_1, \text{ sine}}) \geq c_7 \left(\frac{d v_1^2 \log(1+v_0/d)}{n} \right)^{1/2}. \quad (4.12)$$

for some universal constant $c_7 > 0$.

Before we prove Theorem 11, we first state a lemma which is contained in the proof of Theorem 1 (pp. 46-47) in [71].

Lemma 14. For integers M, L with $M \geq 10$ and $1 \leq L \leq M/10$, define the set

$$\mathcal{S} = \{\omega \in \{0, 1\}^M : \|\omega\|_1 = L\}.$$

There exists a subset $\mathcal{A} \subset \mathcal{S}$ with cardinality at least $\sqrt{\binom{M}{L}}$ such that the Hamming distance between any pairs of \mathcal{A} is at least $L/5$.

Note that the elements of the set \mathcal{A} in Lemma 14 can be interpreted as binary codes of length M , constant Hamming weight L , and minimum Hamming distance $L/5$. These are called constant weight codes and the cardinality of the largest such codebook, denoted by $A(M, L/5, L)$, is also given a combinatorial lower bound in [72]. The conclusion of Lemma 14 is $A(M, L/5, L) \geq \sqrt{\binom{M}{L}}$.

Proof of Theorem 11. For simplicity, we henceforth write \mathcal{F}_{v_0, v_1} instead of $\mathcal{F}_{v_0, v_1, \text{ sine}}$. Define the collection $\Lambda = \{\theta \in \mathbb{Z}^d : \|\theta\|_1 \leq v_0\}$. Without loss of generality, assume that v_0 is an integer so that $M := \#\Lambda \geq \binom{d+v_0}{d}$. Consider sinusoidal ridge functions $\sqrt{2} \sin(\pi \theta \cdot x)$ with θ in Λ . Note that these functions (for $\theta \neq 0$) are orthonormal with respect to the uniform probability measure P on $B = [-1, 1]^d$. This fact is easily established using an instance of Euler's formula $\sin(\pi \theta \cdot x) = \frac{1}{2i} (\prod_{k=1}^d e^{i\pi \theta_k x_k} - \prod_{k=1}^d e^{-i\pi \theta_k x_k})$.

For an enumeration $\theta_1, \dots, \theta_M$ of Λ , define a subclass of \mathcal{F}_{v_0, v_1} by

$$\mathcal{F}_0 = \{f_\omega = \frac{v_1}{L} \sum_{k=1}^M \omega_k \sqrt{2} \sin(\pi \theta_k \cdot x) : \omega \in \mathcal{A}\},$$

where \mathcal{A} is the set in Lemma 14. Any distinct pairs $f_\omega, f_{\omega'}$ in \mathcal{F}_0 have $L_2(P)$ squared distance at least $\|f_\omega - f_{\omega'}\|^2 \geq v_1^2 \|\omega - \omega'\|_2^2 / L^2 \geq v_1^2 / (5L)$. A separation of ϵ^2 determines $L = (v_1 / (\sqrt{5}\epsilon))^2$. Depending on the size of d relative to v_0 , there are two different behaviors of M . For $d > v_0$, we use $M \geq \binom{d+v_0}{v_0} \geq (1 + d/v_0)^{v_0}$ and for $d < v_0$, $M \geq \binom{d+v_0}{d} \geq (1 + v_0/d)^d$.

By Lemma 14, a lower bound on the cardinality of \mathcal{A} is $\sqrt{\binom{M}{L}}$ with logarithm lower bounded by $(L/2) \log(M/L)$. To obtain a cleaner form that highlights the dependence on L , we assume that $L \leq \sqrt{M}$, giving $\log(\#\mathcal{A}) \geq (L/4) \log M$. Since L is proportional to $(v_1/\epsilon)^2$, this condition puts a lower bound on ϵ of order $v_1 M^{-1/4}$. If $\epsilon > v_1 / (1 + d/v_0)^{v_0/4}$, it follows

that a lower bound on the logarithm of the packing number is of order $\log N_{d>v_0}(\epsilon) = v_0(v_1/\epsilon)^2 \log(1 + d/v_0)$. If $\epsilon > v_1/(1 + v_0/d)^{d/4}$, a lower bound on the logarithm of the packing number is of order $\log N_{v_0>d}(\epsilon) = d(v_1/\epsilon)^2 \log(1 + v_0/d)$. Thus we have found an ϵ -packing set of these cardinalities. As such, they are lower bounds on the metric entropy of \mathcal{F}_{v_0, v_1} .

Next we use the information-theoretic lower bound techniques in [32] or [73]. Let $p_\omega(x, y) = p(x)\psi(y - f_\omega(x))$, where p is the uniform density on $[-1, 1]^d$ and ψ is the $N(0, 1)$ density. Then

$$R_{n,d}(\mathcal{F}_{v_0, v_1}) \geq (\epsilon^2/4) \inf_{\hat{f}} \sup_{f \in \mathcal{F}_0} \mathbb{P}(\|f - \hat{f}\|^2 \geq \epsilon^2),$$

where the estimators \hat{f} are now restricted to \mathcal{F}_0 . The supremum is at least the uniformly weighted average over $f \in \mathcal{F}_0$. Thus a lower bound on the minimax risk is a constant times ϵ^2 provided the minimax probability is bounded away from zero, as it is for sufficient size packing sets. Indeed, by Fano's inequality as in [32], this minimax probability is at least

$$1 - \frac{\alpha \log(\#\mathcal{F}_0) + \log 2}{\log(\#\mathcal{F}_0)},$$

for α in $(0, 1)$, or by an inequality of Pinsker, as in Theorem 2.5 in [73], it is at least

$$\frac{\sqrt{\#\mathcal{F}_0}}{1 + \sqrt{\#\mathcal{F}_0}} (1 - 2\alpha - \sqrt{\frac{2\alpha}{\log(\#\mathcal{F}_0)}}),$$

for some α in $(0, 1/8)$. These inequalities hold provided we have the following

$$\frac{1}{\#\mathcal{F}_0} \sum_{\omega \in \mathcal{A}} D(p_\omega^n || q) \leq \alpha \log(\#\mathcal{F}_0),$$

bounding the mutual information between ω and the data $\{(X_i, Y_i)\}_{i=1}^n$, where q is any fixed joint density for $\{(X_i, Y_i)\}_{i=1}^n$. When suitable metric entropy upper bounds on the log-cardinality of covers $\mathcal{F}_{\omega' \in \mathcal{A}'} := \{f : \|f - f_{\omega'}\| < \epsilon'\}$ of \mathcal{F}_0 are available, one may use q as a uniform mixture of $p_{\omega'}^n$, for ω' in \mathcal{A}' as in [32], as long as ϵ and ϵ' are arranged to be of the same order. In the special case that \mathcal{F}_0 has small radius already of order ϵ , one has the simplicity of taking \mathcal{A}' to be the singleton set consisting of $\omega' = 0$. In the present

case, since each element in \mathcal{F}_0 has squared norm $v_1^2/L = 5\epsilon^2$ and pairs of elements in \mathcal{F}_0 have squared separation ϵ^2 , these function are near $f_0 \equiv 0$ and hence we choose $q = p_0^n$. A standard calculation yields

$$D(p_\omega^n || p_0^n) \leq \frac{n}{2} \|f_\omega\|^2 \leq \frac{nv_1^2}{2L} = (5/2)n\epsilon^2.$$

We choose ϵ_n such that this $(5/2)n\epsilon_n^2 \leq \alpha \log(\#\mathcal{F}_0)$. Thus, in accordance with [32], if $N_{d>v_0}(\epsilon_n)$ and $N_{v_0>d}(\epsilon_n)$ are available lower bounds on $\#\mathcal{F}_0$, to within a constant factor, a minimax lower bound ϵ_n^2 on the $L_2(P)$ squared error risk is determined by matching

$$\epsilon_n^2 = \frac{\log N_{d>v_0}(\epsilon_n)}{n},$$

and

$$\epsilon_n^2 = \frac{\log N_{v_0>d}(\epsilon_n)}{n}.$$

Solving in either case, we find that

$$\epsilon_n^2 = \left(\frac{v_0 v_1^2 \log(1+d/v_0)}{n} \right)^{1/2},$$

and

$$\epsilon_n^2 = \left(\frac{d v_1^2 \log(1+v_0/d)}{n} \right)^{1/2}.$$

These quantities are valid lower bounds on $R_{n,d}(\mathcal{F}_{v_0,v_1})$ to within constant factors, provided $N_{d>v_0}(\epsilon_n)$ and $N_{v_0>d}(\epsilon_n)$ are valid lower bounds on the ϵ_n -packing number of \mathcal{F}_{v_0,v_1} . Checking that $\epsilon_n > v_1/(1+d/v_0)^{v_0/2}$ and $\epsilon_n > v_1/(1+v_0/d)^{d/2}$ yields conditions (4.9) and (4.10), respectively. \square

Remark. Conditions (4.9) and (4.10) are needed to ensure that the lower bounds for the packing numbers take on the form $L \log M$ instead of $L \log(M/L)$. We accomplish this by imposing $L \leq \sqrt{M}$. Alternatively, any upper bound of the form M^ρ , $\rho \in (0, 1)$ will work with similar conclusion, adjusting lower bounds (4.11) and (4.12) by a factor of $\sqrt{1-\rho}$, with corresponding adjustment to the requirements on d/v_0 in (4.9) and v_0/d in (4.10).

4.3 Implications for neural nets

The variation of a function f with respect to a dictionary \mathcal{D} [3], also called the atomic norm of f with respect to \mathcal{D} , denoted $V_f(\mathcal{D})$, is defined as the infimum of all v such that f is in $\ell_1(v, \mathcal{D})$. Here the closure in the definition of $\ell_1(v, \mathcal{D})$ is taken in L_∞ .

Define $\phi(z) = \sqrt{2}\sin(\pi z)$. On the interval $[-v_0, v_0]$, it can be shown that $\phi(z)$ has variation $V_\phi = 2\sqrt{2}\pi v_0$ with respect to the dictionary of unit step activation functions $\pm\text{step}(z' - t')$, where $\text{step}(z) = \mathbb{I}\{z > 0\}$, or equivalently, variation $\sqrt{2}\pi v_0$ with respect to the dictionary of signum activation functions with shifts $\pm\text{sgn}(z' - t')$, where $\text{sgn}(z) = 2\text{step}(z) - 1$. This can be seen directly from the identity

$$\sin z = \frac{v}{2} \int_0^1 \cos(vt) [\text{sgn}(z/v - t) - \text{sgn}(-z/v - t)] dt,$$

for $|z| \leq v$. Evaluation of $\int_0^1 |\cos(vt)| dt$ gives the exact value of ϕ with respect to sgn as $\sqrt{2}\pi v_0$ for integer $v = v_0$. Accordingly, $\mathcal{F}_{v_0, v_1, \phi}$ is contained in $\mathcal{F}_{1, \sqrt{2}\pi v_0 v_1, \text{sgn}}$.

Likewise, for the clipped linear function $\text{clip}(z) = \text{sgn}(z) \min\{1, |z|\}$ a similar identity holds:

$$\begin{aligned} \sin z = z + \frac{v^2}{2} \int_0^1 \sin(vt) [\text{clip}(-2z/v - 2t - 1) - \\ \text{clip}(2z/v - 2t - 1)] dt, \end{aligned}$$

for $|z| \leq v$. The above form arises from integrating

$$\begin{aligned} \cos w = \cos v - \frac{v}{2} \int_0^1 \sin(vt) [\text{sgn}(-w/v - t) + \\ \text{sgn}(w/v - t)] dt, \end{aligned}$$

from $w = 0$ to $w = z$. And likewise, evaluation of $\int_0^1 |\sin(vt)| dt$ gives the exact variation of ϕ with respect to the dictionary of clip activation functions $\pm\text{clip}(z' - t')$ as $V_\phi = \sqrt{2}\pi(v_0^2 + 1)$ for integer $v = v_0$. Accordingly, $\mathcal{F}_{v_0, v_1, \phi}$ is contained in $\mathcal{F}_{2, \sqrt{2}\pi(v_0^2 + 1)v_1, \text{clip}}$ and hence we have the following corollary.

Corollary 1. *Using the same setup and conditions (4.9) and (4.10) as in Theorem 11, the minimax risk for the sigmoid classes $\mathcal{F}_{1,\sqrt{2}\pi v_0 v_1, \text{sgn}}$ and $\mathcal{F}_{2,\sqrt{2}\pi(v_0^2+1)v_1, \text{clip}}$ have the same lower bounds (4.11) and (4.12) as for $\mathcal{F}_{v_0, v_1, \text{sine}}$.*

4.4 Implications for polynomial nets

It is also possible to give minimax lower bounds for the function classes $\mathcal{F}_{v_0, v_1, \phi_\ell}$ with activation function ϕ_ℓ equal to the standardized Hermite polynomial $H_\ell/\sqrt{\ell!}$, where $H_\ell(z) = (-1)^\ell e^{\frac{z^2}{2}} \frac{d^\ell}{dz^\ell} e^{-\frac{z^2}{2}}$. As with Theorem 11, this requires a lower bound on d :

$$\frac{d}{v_0^2} > (c_8 \frac{v_1^2 n}{v_0^2 \log(d/v_0^2)})^{2/v_0^2}. \quad (4.13)$$

for some constant $c_8 > 0$. Moreover, we also need a growth condition on the order of the polynomial ℓ :

$$\ell > c_9 \log(\frac{v_1^2 n}{v_0^2 \log(d/v_0^2)}), \quad (4.14)$$

for some constant $c_9 > 0$. In light of (4.13), condition (4.14) is also satisfied if ℓ is at least a constant multiple of $v_0^2 \log(d/v_0^2)$.

Theorem 12. *Consider the model $Y = f(X) + \varepsilon$ for $f \in \mathcal{F}_{v_0, v_1, \phi_\ell}$, where $\varepsilon \sim N(0, 1)$ and $X \sim N(0, I_d)$. If d and ℓ are large enough so that conditions (4.13) and (4.14) are satisfied, respectively, then*

$$R_{n,d}(\mathcal{F}_{v_0, v_1, \phi_\ell}) \geq c_{10} (\frac{v_0^2 v_1^2 \log(d/v_0^2)}{n})^{1/2}, \quad (4.15)$$

for some universal constant $c_{10} > 0$.

Proof of Theorem 12. By Lemma 14, if $d \geq 10$ and $1 \leq d' \leq d/10$, there exists a subset \mathcal{C} of $\{0, 1\}^d$ with cardinality at least $M := \sqrt{\binom{d}{d'}}$ such that each element has Hamming weight d' and pairs of elements have minimum Hamming distance $d'/5$. Thus, if a and a' belong to this codebook, $|a \cdot a'| \leq (9/10)d'$. Choose $d' = v_0^2$ (assuming that v_0^2 is an integer less than d), and form the collection $\mathcal{B} = \{\theta = a/v_0 : a \in \mathcal{C}\}$. Note that each member of \mathcal{B} has unit ℓ_2 norm and ℓ_1 norm v_0 . Moreover, the Euclidean inner product between each pair has

magnitude bounded by $9/10$. Next, we use the fact that if $X \sim N(0, I_d)$ and θ, θ' have unit ℓ_2 norm, then $\mathbb{E}[\phi_\ell(\theta \cdot X)\phi_\ell(\theta' \cdot X)] = (\theta \cdot \theta')^\ell$. For an enumeration $\theta_1, \dots, \theta_M$ of \mathcal{B} , define a subclass of $\mathcal{F}_{v_0, v_1, H_\ell}$ by

$$\mathcal{F}_0 = \{f_\omega = \frac{v_1}{L} \sum_{k=1}^M \omega_k \phi_\ell(\theta_k \cdot x) : \omega \in \mathcal{A}\},$$

where \mathcal{A} is the set from Lemma 14. Moreover, since each θ_k has unit norm, $\|\omega - \omega'\|_1 \geq L/5$, and $\|\omega - \omega'\|_1^2 \leq 2L\|\omega - \omega'\|_1$,

$$\begin{aligned} \|f_\omega - f_{\omega'}\|^2 &= \frac{v_1^2}{L^2} [\|\omega - \omega'\|_1 + \\ &\quad \sum_{i \neq j} (\omega_i - \omega'_i)(\omega_j - \omega'_j)(\theta_i \cdot \theta_j)^\ell] \\ &\geq \frac{v_1^2}{L^2} [\|\omega - \omega'\|_1 - \|\omega - \omega'\|_1^2 (9/10)^\ell] \\ &\geq \frac{v_1^2}{L^2} \|\omega - \omega'\|_1 (1 - 2L(9/10)^\ell) \\ &\geq \frac{v_1^2}{L} (1 - 2L(9/10)^\ell) \\ &\geq \frac{v_1^2}{10L}, \end{aligned}$$

provided $\ell > \frac{\log(4L)}{\log(10/9)}$. A separation of ϵ^2 determines $L = (v_1/(\sqrt{10}\epsilon))^2$. If $L \leq \sqrt{M}$, or equivalently, $\epsilon \geq v_1 M^{-1/4}$, then $\log(\#\mathcal{F}_0)$ is at least a constant multiple of $\log N_{d>v_0}(\epsilon) = (v_0 v_1/\epsilon)^2 \log(d/v_0^2)$. As before in Theorem 11, a minimax lower bound ϵ_n^2 on the $L_2(P)$ squared error risk is determined by matching

$$\epsilon_n^2 = \frac{\log N_{d>v_0}(\epsilon_n)}{n},$$

which yields

$$\epsilon_n^2 = \left(\frac{v_0^2 v_1^2 \log(d/v_0^2)}{n} \right)^{1/2}.$$

If conditions (4.13) and (4.14) are satisfied, $N_{d>v_0}(\epsilon_n)$ is a valid lower bound on the ϵ_n -packing number of $\mathcal{F}_{v_0, v_1, \phi_\ell}$. □

4.5 Discussion

Our risk lower bound of the form $(\frac{v_0 v_1^2 \log(1+d/v_0)}{n})^{1/2}$ shows that in the very high-dimensional case, it is the $v_0 v_1^2/n$ to a half-power that controls the rate (to within a logarithmic factor). The v_0 and v_1 , as ℓ_1 norms of the inner and outer coefficient vectors, have the interpretations as the effective dimensions of these vectors. Indeed, a vector in \mathbb{R}^d with bounded coefficients that has v_0 non-negligible coordinates has ℓ_1 norm of this order. These rates confirm that it is a power of these effective dimensions over sample size n (instead of the full ambient dimension d) that controls the main behavior of the statistical risk. Our lower bounds on packing numbers complement the upper bound covering numbers in [36] and [57]. Our rates are akin to those obtained in [31] for high-dimensional linear regression. However, there is an important difference. The richness of \mathcal{F}_{v_0, v_1} is largely determined by the sizes of v_0 and v_1 and \mathcal{F}_{v_0, v_1} more flexibly represents a larger class of functions. It would be interesting to see if the gap between the powers $1/2$ and $1/3$ could be closed by improving either the lower bound in (4.11) or the upper bound in (4.8).

Chapter 5

Estimating the coefficients of a mixture of two linear regressions by expectation maximization

5.1 Introduction

The Expectation-Maximization (EM) algorithm is a widely used technique for parameter estimation. It is an iterative procedure that monotonically increases the likelihood. When the likelihood is not concave, it is well known that EM can converge to a non-global optimum. However, recent work has side-stepped the question of whether EM reaches the likelihood maximizer, instead by directly working out statistical guarantees on its loss. These explorations have identified regions of initialization for which the EM estimate approaches the true parameter in probability, assuming the model is well-specified.

This line of research was spurred by [1] which established general conditions for which a ball centered at the true parameter would be a basin of attraction for the population version of the EM operator. For a large enough sample size, the difference (in that ball) between the sample EM operator and the population EM operator can be bounded such that the EM estimate approaches the true parameter with high probability. That bound is the sum of two terms with distinct interpretations. There is an *algorithmic convergence*

term $\kappa^t \|\theta^0 - \theta^*\|$ for initializer θ^0 , truth θ^* , and some modulus of contraction $\kappa \in (0, 1)$; this comes from the analysis of the population EM operator. The second term captures *statistical convergence* and is proportional to the supremum norm of $M - M_n$, the difference between the population and sample EM operators, over the ball. This result is also shown for a “sample-splitting” version of EM, where the sample is partitioned into batches and each batch governs a single step of the algorithm.

That article also detailed three specific simple models in which their analysis is easily seen to apply: symmetric mixture of two spherical Gaussians, symmetric mixture of two linear models with Gaussian covariates and error, and linear regression with data missing completely at random.

The performance of EM for their first example, a symmetric mixture of two spherical Gaussians, has since received further attention. [6] showed that the intersection of a suitable half-space and ball about the origin is also a basin of attraction for the population EM in that model when the component means are separated well enough relative to the noise. Exact probabilistic bounds on the error of the EM estimate were also derived when the initializer is in the region. The authors also proposed a random initialization strategy that has a high probability of finding the basin of attraction when the component means are well-separated as $\sqrt{d \log d}$. Concurrently, [2] revealed that the entirety of \mathbb{R}^d (except the hyperplane perpendicular to θ^*) is a basin of attraction for the population EM operator (in addition to asymptotic consistency of the empirical iterates). Subsequently in [74], a more explicit expression for the contraction constant and its dependence on the initializer was obtained through an elegant argument.

The second example of [1], the symmetric mixture of two linear models with Gaussian covariates and error, can be seen as a generalization of the symmetric mixture. This model, also known as Hierarchical Mixture of Experts (HME) in the machine learning community [75], has drawn recent attention (e.g. [76], [77], [78], [79], [80]). The analysis of the two-mixture case was generalized to arbitrary multiple components in [79], but initialization is still required to be in a ball around each of the true coefficient vectors.

Our purpose here is to follow up the analysis of [1] by proving a larger basin of attraction for the mixture of two linear models and by establishing an exact probabilistic bound on the

error of the sample-splitting EM estimate when the initializer falls in the specified region. In related works, typically some variant of the mean value theorem is employed to establish contractivity and the rate of geometric decay is then determined by relying heavily on the fact that initializer belongs to a bounded set and is not too far from the truth (i.e. a ball centered at the truth). Our technique relies on Stein’s Lemma, which allows us to reduce the problem to the two-dimensional case and exploit certain monotonicity properties of the EM operator. Such methods allow one to be very careful and explicit in the analysis and more cleanly reveal the role of the initialization conditions. These results cannot be deduced from other works, even by sharpening their analysis. Our improvements are not solely in terms of constants – as long as the cosine angle between the initializer and the truth is sufficiently large, contractivity holds. In particular, the norm of the initializer can be arbitrarily large, provided the cosine angle condition is met.

In Section 5.2, we explain the model and derive a basin of attraction for the population version of the EM operators and also show that it is not contractive in certain regions of \mathbb{R}^d . Section 5.3 looks at the behavior of the sample-splitting EM operator in this region and proves statistical guarantees. Section 5.4 considers a more general model that doesn’t require symmetry. We point out that estimation for that model can be handled by an estimator related to the symmetric case’s EM; this estimator essentially inherits the statistical guarantees derived for EM in the symmetric case. Finally, the more technical proofs are in the supplementary material in Appendix 8.7.

5.2 The population EM operator

Let data $(X_i, Y_i)_{i=1}^n$ be *i.i.d.* with $X_i \sim N(0, I_d)$ and

$$Y_i = R_i \langle \theta^*, X_i \rangle + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$, $R_i \sim \text{Rademacher}$, and X_i, ε_i, R_i are independent of each other. In other words, each predictor variable is normal, and the response is centered at either the θ^* or $-\theta^*$ linear combination of the predictor. The two classes are equally probable, and the

label of each observation is unknown. We seek to estimate θ^* (or $-\theta^*$, which produces the same model distribution).

The likelihood function is multi-model, and direct maximization is intractable. The EM algorithm has been used to estimate the model coefficients [75], and simulation studies have shown that it has desirable empirical performance [81], [82], [83]. The EM operator for estimating θ^* (see [1, page 6] for a derivation) is

$$M_n(\theta) = \left(\frac{1}{n} \sum X_i X_i^T\right)^{-1} \left[\frac{1}{n} \sum (2\phi(Y_i \langle \theta, X_i \rangle / \sigma^2) - 1) X_i Y_i\right] \quad (5.1)$$

where $\phi(t) = \frac{1}{1+e^{-2t}}$ is a horizontally stretched logistic sigmoid. The population EM operator replaces sample averages with expectations, thus

$$M(\theta) = 2\mathbb{E}[\phi(Y \langle \theta, X \rangle / \sigma^2) XY]. \quad (5.2)$$

Conveniently, this estimation can be reduced to the $\sigma = 1$ case. If we divide each response datum by σ :

$$Y_i/\sigma = R_i \langle \theta^*/\sigma, X_i \rangle + \varepsilon_i/\sigma,$$

the unknown parameter to estimate becomes θ^*/σ , and the noise has variance 1. Inspection of (5.1) and (5.2) confirms that the EM operators for the new problem are equal to $1/\sigma$ times the EM operators for the original problem. For instance, denoting the population EM operator of the new problem by \widetilde{M} ,

$$\begin{aligned} \widetilde{M}(\theta/\sigma) &= 2\mathbb{E}[\phi((Y/\sigma) \langle \theta/\sigma, X \rangle) X (Y/\sigma)] \\ &= \frac{2}{\sigma} \mathbb{E}[\phi(Y \langle \theta, X \rangle / \sigma^2) XY] \\ &= \frac{1}{\sigma} M(\theta). \end{aligned}$$

The transformed problem's error is easily related to the original problem's error:

$$\begin{aligned}\|\widetilde{M}(\theta/\sigma) - \theta^*/\sigma\| &= \left\| \frac{1}{\sigma} M(\theta) - \theta^*/\sigma \right\| \\ &= \frac{1}{\sigma} \|M(\theta) - \theta^*\|\end{aligned}$$

Thus, in the general case, the estimation error is exactly σ times the estimation error of the normalized problem. We use this observation to simplify the proof of Lemma 15, while stating our results for general σ .

In [1], it was shown that if the EM algorithm is initialized in a ball around θ^* with radius proportional θ^* , the EM algorithm converges with high probability. The purpose of this chapter is to relax these conditions and show that if the cosine angle between θ^* and the initializer is not too small, the EM algorithm also converges. We also simplify the analysis, using only elementary facts about multivariate normal distributions. This improvement is manifested in the set containment

$$\{\theta : \|\theta - \theta^*\| \leq \sqrt{1 - \rho^2} \|\theta^*\|\} \subseteq \{\theta : \langle \theta, \theta^* \rangle \geq \rho \|\theta\| \|\theta^*\|\}, \quad \rho \in [0, 1],$$

since for all θ in the set on the left side,

$$\begin{aligned}\langle \theta, \theta^* \rangle &= \frac{1}{2} (\|\theta\|^2 + \|\theta^*\|^2 - \|\theta - \theta^*\|^2) \\ &\geq \frac{1}{2} (\|\theta\|^2 + \rho^2 \|\theta^*\|^2) \\ &\geq \rho \|\theta\| \|\theta^*\|.\end{aligned}$$

The authors of [1] required the initializer θ^0 to be at most $\|\theta^*\|/32$ away from θ^* , while our condition allows for the norm of θ^0 to be unbounded. We will also show how the analysis relates to the one-dimensional mixture of two Gaussians by exploiting the self-consistency property of its population EM operator.

Let θ_0 be the unit vector in the direction of θ and let θ_0^\perp be the unit vector that belongs to the hyperplane spanned by $\{\theta^*, \theta\}$ and orthogonal to θ (i.e. $\theta_0^\perp \in \text{span}\{\theta, \theta^*\}$ and $\langle \theta, \theta_0^\perp \rangle = 0$). Let $\theta^\perp = \|\theta\| \theta_0^\perp$. We will later show that $M(\theta)$ belongs to $\text{span}\{\theta, \theta^*\}$,

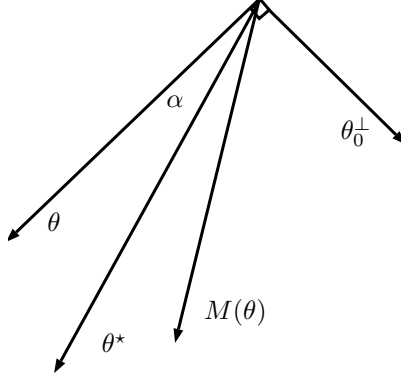


Figure 5.1: The population EM operator $M(\theta)$ lies in the space spanned by θ and θ^* . The unit vector θ_0^\perp lies in the space spanned by θ and θ^* and is perpendicular to θ . The vector θ forms an angle α with θ^* .

as in Fig. 5.1. Denote the angle between θ^* and θ_0 as α , with $\|\theta^*\| \cos \alpha = \langle \theta_0, \theta^* \rangle$ and $\rho = \cos \alpha$. As we will see from the following results, as long as $\cos \alpha$ is not too small, $M(\theta)$ is a contracting operation that is always closer to the truth θ^* than θ .

Lemma 15. *For any θ in \mathbb{R}^d with $\langle \theta, \theta^* \rangle > 0$,*

$$\|M(\theta) - \theta^*\| \leq \sqrt{\kappa} \sqrt{1 + 4 \left(\frac{|\langle \theta^\perp, \theta^* \rangle| + \sigma^2}{\langle \theta, \theta^* \rangle} \right)^2} \|\theta - \theta^*\|, \quad (5.3)$$

where

$$\kappa^2 = \max \left\{ 1 - \frac{|\langle \theta_0, \theta^* \rangle|^2}{\sigma^2 + \|\theta^*\|^2}, 1 - \frac{\langle \theta, \theta^* \rangle}{\sigma^2 + \langle \theta, \theta^* \rangle} \right\} \leq 1. \quad (5.4)$$

As we will see, this constant κ is closely related to the contraction constant γ of the operator $M(\theta)$.

If we write the signal to noise ratio as $\eta = \|\theta^*\|/\sigma$ and use the fact that $\|\theta^*\| \cos \alpha = \langle \theta_0, \theta^* \rangle$, the contractivity constant can be written as

$$\max \left\{ \left(1 - \frac{\eta^2 \cos^2 \alpha}{1 + \eta^2} \right)^{1/4}, \left(1 - \frac{\|\theta\| \eta \cos \alpha}{\sigma + \|\theta\| \eta \cos \alpha} \right)^{1/4} \right\} \sqrt{1 + 4 \left(\tan \alpha + \frac{\sigma}{\|\theta\|} \frac{1}{\eta \cos \alpha} \right)^2}. \quad (5.5)$$

Remark 7. *If $\|\theta\| \geq 10\sigma$, $\|\theta^*\| \geq 20\sigma$ and $\cos \alpha \geq 0.9$, the quantity (5.5) is bounded by a universal constant $\gamma < 1$, implying the population EM operator $\theta^{t+1} \leftarrow M(\theta^t)$ converges to the truth θ^* exponentially fast.*

The next theorem shows that the above conditions are essentially necessary in the sense that contractivity of M can fail for certain initializers that do not meet our cosine angle criterion. In contrast, it is known that the population EM operator for a symmetric mixture of two Gaussians is globally contractive [74], [2]. The disparity is likely due to the additional variability coming from the input design matrix X .

Theorem 13. *There are points θ satisfying $\langle \theta, \theta^\star \rangle > 0$ such that*

$$\|M(\theta) - \theta^\star\| > \|\theta - \theta^\star\|.$$

While this result does not generally imply that the empirical iterates $\theta^{t+1} \leftarrow M_n(\theta^t)$ will fail to converge to θ^\star for $\langle \theta^0, \theta^\star \rangle > 0$, it does suggest that difficulties may arise in this regime. Indeed, the discussion at the end of this chapter gives empirical evidence for this observation.

5.3 The sample EM operator

As in [1], we analyze a sample-splitting version of the EM algorithm, where for an allocation of n samples and T iterations, we divide the data into T subsets of size $\lfloor n/T \rfloor$. We then perform the updates $\theta^{t+1} \leftarrow M_{n/T}(\theta^t)$, using a new subset of samples to compute $M_{n/T}(\theta)$ at each iteration.

Theorem 14. *Let $\langle \theta^0, \theta^\star \rangle > \rho \|\theta^0\| \|\theta^\star\|$, $10\sigma \leq \|\theta^0\| \leq L\sigma$, and $\|\theta^\star\| \geq 20\sigma$ for $\rho \in (0.9, 1)$ and $L > \sqrt{1 + 3\|\theta^\star\|^2/\sigma^2}$. Suppose furthermore that $n \geq cd \log(1/\delta)$ for $\delta \in (0, 1)$ and some constant $c = c(\rho, \sigma, \|\theta^\star\|, L) \geq 1$. Then there exists $\gamma = \gamma(\rho, \sigma, \|\theta^\star\|) \in (0, 1)$ such that the sample-splitting empirical EM iterates $\{\theta^t\}_{t=1}^T$ based on n/T samples per step satisfy*

$$\|\theta^t - \theta^\star\| \leq \gamma^t \|\theta^0 - \theta^\star\| + \frac{C\sqrt{\|\theta^\star\|^2 + \sigma^2}}{1 - \gamma} \sqrt{\frac{dT \log(T/\delta)}{n}},$$

with probability at least $1 - \delta$.

We will prove this result at the end of the chapter. The main aspect of the analysis lies in showing that M_n satisfies an invariance property, i.e. $M_n(\mathcal{A}) \subseteq \mathcal{A}$, where \mathcal{A} is the basin

of attraction. The algorithmic error $\gamma^t \|\theta^0 - \theta^*\|$ follows from Lemma 15 and the stochastic error

$$\frac{C\sqrt{\|\theta^*\|^2 + \sigma^2}}{1-\gamma} \sqrt{\frac{dT \log(T/\delta)}{n}} \text{ from the proof of Corollary 4 in [1].}$$

Remark 8. *Theorem 14 requires the initializer to have a good inner product with θ^* . But how to initialize in practice? There is considerable literature showing the efficacy of initialization based on spectral [77], [78], [79] or Bayesian [82] methods.*

5.4 Without assuming symmetry

Without requiring symmetry, we can still derive statistical guarantees for a variant on the EM estimation procedure described above. In this section, we assume that data $(X_i, Y_i)_{i=1}^n$ is *i.i.d.* with $X_i \sim N(0, I_d)$ and

$$Y_i = \mathbb{1}\{R_i = 1\} \langle \theta_1^*, X_i \rangle + \mathbb{1}\{R_i = -1\} \langle \theta_2^*, X_i \rangle + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$, $R_i \sim \text{Rademacher}$, and X_i, ε_i, R_i are independent of each other.

This time each model distribution is specified (uniquely up to class labels) by two parameters: θ_1^* and θ_2^* . Our previous analysis was for the restriction of this model to the slice in which $\theta_2^* = -\theta_1^*$.

Our first step is to reformulate the model as a shifted version of the symmetric case:

$$Y_i = R_i \langle \theta^*, X_i \rangle + \langle s, X_i \rangle + \varepsilon_i,$$

where $\theta^* := (\theta_1^* - \theta_2^*)/2$ and the shift is $s := (\theta_1^* + \theta_2^*)/2$. The shift can be estimated by $\hat{s} = \frac{1}{n} \sum_{i=1}^n X_i Y_i$ (or alternative by $(\frac{1}{n} \sum_{i=1}^n X_i X_i^T)^{-1} \hat{s}$) which concentrates around s . We construct a shifted version of the response vector and define an estimate for it:

$$\tilde{Y}_i := Y_i - \langle s, X_i \rangle \quad \text{and} \quad Y_i^{(s)} := Y_i - \langle \hat{s}, X_i \rangle$$

We use the symmetric model version of the EM algorithm on the approximately symmetric data $(X_i, Y_i^{(s)})$ to define the estimator $\hat{\theta}$ for θ^* . The error incurred by the use of the

estimated \widehat{s} can be handled separately from the performance of EM on the truly symmetric (X_i, \widetilde{Y}_i) , via the triangle inequality:

$$\begin{aligned} \|M_n(\theta, \underline{X}, \underline{Y}^{(s)}) - \theta^*\| &\leq \|M_n(\theta, \underline{X}, \underline{Y}^{(s)}) - M_n(\theta, \underline{X}, \underline{\widetilde{Y}})\| \\ &\quad + \|M_n(\theta, \underline{X}, \underline{\widetilde{Y}}) - \theta^*\|. \end{aligned} \quad (5.6)$$

where each underlined letter represents the corresponding vector of n variables. Theorem 14 provides guarantees for good control on the second term of (5.6). The first term is small since the update procedure M_n is a smooth function of the data; it is of asymptotically smaller order than the second term. Finally, if desired, one can estimate the original parameters by $\theta_1^t := \theta^t + \widehat{s}$ and $\theta_2^t := \widehat{s} - \theta^t$. The proof for the asymmetric case is below.

Theorem 15. *Apply the sample-splitting version of EM described in Section 5.3 on the shifted data $\underline{\widetilde{Y}}$ defined above and assume that θ_0 satisfies the same initialization conditions with $\theta^* = (\theta_1^* - \theta_2^*)/2$. There exists constant $C > 0$ for which the EM iterates $\{\theta^t\}_{t=1}^T$ satisfy*

$$\begin{aligned} \|\theta^t - \theta^*\| &\leq \gamma^t \|\theta^0 - \theta^*\| + \frac{C\sqrt{\|\theta^*\|^2 + \|s\|^2 + \sigma^2}}{1 - \gamma} \sqrt{\frac{dT \log(T/\delta)}{n}} \\ &= \gamma^t \|\theta^0 - \theta^*\| + \frac{C\sqrt{(\|\theta_1^*\|^2 + \|\theta_2^*\|^2)/2 + \sigma^2}}{1 - \gamma} \sqrt{\frac{dT \log(T/\delta)}{n}}. \end{aligned}$$

with probability at least $1 - \delta$.

Remark 9. *Combine Lemma 16 and Theorem 15 to deduce the error rates on the original centers.*

$$\|\theta_i^t - \theta_i^*\| \leq \gamma^t \|\theta^0 - \theta^*\| + \frac{(C + D_1)\sqrt{(\|\theta_1^*\|^2 + \|\theta_2^*\|^2)/2 + \sigma^2}}{1 - \gamma} \sqrt{\frac{dT \log(T/\delta)}{n}},$$

for $i = 1, 2$ with probability at least $1 - \delta$.

5.5 Proofs of main theorems

Proof of Lemma 15. For simplification, we assume throughout this proof that $\sigma^2 = 1$. If $W = \langle \theta^*, X \rangle + \varepsilon$, a few applications of Stein's Lemma [84, Lemma 1] yields

$$\begin{aligned} M(\theta) &= \mathbb{E}[(2\phi(W\langle \theta, X \rangle) - 1)XW] \\ &= \theta^*(\mathbb{E}[2\phi(W\langle \theta, X \rangle) + 2W\langle \theta, X \rangle\phi'(W\langle \theta, X \rangle) - 1]) \\ &\quad + \theta\mathbb{E}[2W^2\phi'(W\langle \theta, X \rangle)]. \end{aligned}$$

In what follows, we let

$$A = \mathbb{E}[2\phi(W\langle \theta, X \rangle) + 2W\langle \theta, X \rangle\phi'(W\langle \theta, X \rangle) - 1]$$

and

$$B = 2W^2\phi'(W\langle \theta, X \rangle).$$

Thus, we see that $M(\theta) = \theta^*A + \theta B$ belongs to $\text{span}\{\theta, \theta^*\} = \{\lambda_1\theta + \lambda_2\theta^*, : \lambda_1, \lambda_2 \in \mathbb{R}\}$.

This is a crucial fact that will exploit multiple times.

Observe that for any a in $\text{span}\{\theta, \theta^*\}$,

$$a = \langle \theta_0, a \rangle \theta_0 + \langle \theta_0^\perp, a \rangle \theta_0^\perp,$$

and

$$\|a\|^2 = |\langle \theta_0, a \rangle|^2 + |\langle \theta_0^\perp, a \rangle|^2.$$

Specializing this to $a = M(\theta) - \theta^*$ yields

$$\|M(\theta) - \theta^*\|^2 = |\langle \theta_0, M(\theta) - \theta^* \rangle|^2 + |\langle \theta_0^\perp, M(\theta) - \theta^* \rangle|^2.$$

The strategy for establishing contractivity of $M(\theta)$ will be to show that the sum of $|\langle \theta_0, M(\theta) - \theta^* \rangle|^2$ and $|\langle \theta_0^\perp, M(\theta) - \theta^* \rangle|^2$ is less than $\gamma^2\|\theta - \theta^*\|^2$. This idea was used in [74] to obtain global contractivity of the population EM operator for the mixture of two

Gaussians problem.

To reduce this $(d+1)$ -dimensional problem (as seen from the joint distribution of (X, Y)) to a 2-dimensional problem, we note that

$$W\langle\theta, X\rangle \stackrel{\mathcal{D}}{=} \Lambda Z_1 Z_2 + \Gamma Z_2^2,$$

where $Z_1, Z_2 \stackrel{i.i.d.}{\sim} N(0, 1)$. The coefficients Γ and Λ are

$$\Gamma = \langle\theta, \theta^\star\rangle$$

and

$$\Lambda^2 = \|\theta\|^2(1 + \|\theta^\star\|^2) - \Gamma^2 = \|\theta\|^2(1 + |\langle\theta_0^\perp, \theta^\star\rangle|^2).$$

This is because we have

$$(W, \langle\theta, X\rangle) \stackrel{\mathcal{D}}{=} (\sqrt{1 + \|\theta^\star\|^2} Z_2, \frac{\Lambda}{\sqrt{1 + \|\theta^\star\|^2}} Z_1 + \frac{\Gamma}{\sqrt{1 + \|\theta^\star\|^2}} Z_2).$$

Note that $\Lambda Z_1 Z_2 + \Gamma Z_2^2 \stackrel{\mathcal{D}}{=} \Lambda Z_1 |Z_2| + \Gamma Z_2^2$ because they have the same moment generating function. Deduce that

$$W\langle\theta, X\rangle \stackrel{\mathcal{D}}{=} \Lambda Z_1 |Z_2| + \Gamma Z_2^2.$$

Lemma 23 implies that

$$(1 - \kappa)\langle\theta_0^\perp, \theta^\star\rangle \leq \langle\theta_0^\perp, M(\theta)\rangle \leq (1 + \sqrt{\kappa})\langle\theta_0^\perp, \theta^\star\rangle,$$

and consequently,

$$|\langle\theta_0^\perp, M(\theta) - \theta^\star\rangle| \leq \sqrt{\kappa}|\langle\theta_0^\perp, \theta - \theta^\star\rangle| \leq \sqrt{\kappa}\|\theta - \theta^\star\|. \quad (5.7)$$

Next, we note that $\Lambda^2 \rightarrow \Gamma$ as $\theta \rightarrow \theta^*$. In fact,

$$\begin{aligned} |\Lambda^2 - \Gamma| &= ||\theta||^2(1 + |\langle \theta_0^\perp, \theta^* \rangle|^2) - \langle \theta, \theta^* \rangle| \\ &\leq ||\theta||^2 |\langle \theta_0^\perp, \theta^* \rangle|^2 + |\langle \theta, \theta - \theta^* \rangle| \\ &\leq ||\theta||(|\langle \theta_0^\perp, \theta^* \rangle| + 1) \|\theta - \theta^*\|. \end{aligned}$$

Finally, define

$$h(\alpha, \beta) = \mathbb{E}[(2\phi(\alpha Z_2(Z_1 + \beta Z_2)) - 1)(Z_2(Z_1 + \beta Z_2))].$$

Note that by definition of h and Lemma 20, $h(\Lambda, \frac{\Gamma}{\Lambda}) = \frac{\langle \theta, M(\theta) \rangle}{\Lambda}$. In fact, h is the one-dimensional population EM operator for this model. By the self-consistency property of EM [85, page 79], $h(\beta, \beta) = \beta$. Translating this to our problem, we have that $h(\frac{\Gamma}{\Lambda}, \frac{\Gamma}{\Lambda}) = \frac{\Gamma}{\Lambda} = \frac{\langle \theta, \theta^* \rangle}{\Lambda}$. Since $h(\Lambda, \frac{\Gamma}{\Lambda}) - h(\frac{\Gamma}{\Lambda}, \frac{\Gamma}{\Lambda}) = \int_{\frac{\Gamma}{\Lambda}}^{\Lambda} \frac{\partial h}{\partial \alpha} h(\alpha, \frac{\Gamma}{\Lambda}) d\alpha$, we have from Lemma 24,

$$\begin{aligned} |\langle \theta_0, M(\theta) - \theta^* \rangle| &\leq \frac{\Lambda}{||\theta||} \left| \int_{\frac{\Gamma}{\Lambda}}^{\Lambda} \frac{\partial h}{\partial \alpha} h(\alpha, \frac{\Gamma}{\Lambda}) d\alpha \right| \\ &\leq \frac{2\Lambda}{||\theta||} \sqrt{\kappa} \left| \int_{\frac{\Gamma}{\Lambda}}^{\Lambda} \frac{d\alpha}{\alpha^2} \right| \\ &= 2\sqrt{\kappa} \frac{|\Lambda^2 - \Gamma|}{\Gamma ||\theta||} \\ &\leq 2\sqrt{\kappa} \left(\frac{|\langle \theta_0^\perp, \theta^* \rangle| + 1}{\langle \theta, \theta^* \rangle} \right) \|\theta - \theta^*\|. \end{aligned}$$

Combining this with inequality (5.7) yields (5.3). □

Remark 10. *The function h is related to the EM operator for the one-dimensional symmetric mixture of two Gaussians model*

$$Y = R\beta + \varepsilon,$$

$R \sim \text{Rademacher}(1/2)$ and $\varepsilon \sim N(0, 1)$. One can derive that (see [6, page 4]) the population EM operator is

$$T(\alpha, \beta) = \mathbb{E}[(2\phi(\alpha(Z_1 + \beta)) - 1)(Z_1 + \beta)].$$

Then $h(\alpha, \beta)$ is a “smoothed” version of $T(\alpha, \beta)$ as seen through the identity

$$h(\alpha, \beta) = \mathbb{E}[|Z_2|T(\alpha|Z_2, \beta|Z_2)].$$

In light of this relationship, it is perhaps not surprising that the EM operator for the mixture of linear regressions problem also enjoys a large basin of attraction.

Remark 11. Recently in [59], the authors analyzed gradient descent for a single-hidden layer convolutional neural network structure with no overlap and Gaussian input. In this setup, we observe i.i.d. data $(X_i, Y_i)_{i=1}^n$, where $Y_i = f(X_i, w) + \varepsilon_i$ and $X_i \sim N(0, I_d)$ and $\varepsilon_i \sim N(0, \sigma^2)$ are independent of each other. The neural network has the form $f(x, w) = \frac{1}{k} \sum_{j=1}^k \max\{0, \langle w_j, x \rangle\}$ and the only nonzero coordinates of w_j are in the j -th successive block of d/k coordinates and are equal to a fixed d/k dimensional filter vector w . One desires to minimize the risk $\ell(w) = \mathbb{E}(f(X, w) - f(X, w^*))^2$. Interestingly, the gradient of $\ell(w)$ belongs to the linear span of w and w^* , akin to our $M(\theta) \in \text{span}\{\theta, \theta^*\}$ (and also in the Gaussian mixture problem [6]). This property plays a critical role in the analysis.

One can use an alternative scheme to gradient descent using a simple method of moments estimator based on the identity $2\mathbb{E}[X \max\{0, \langle w, X \rangle\}] = w$. We observe that $\hat{w} = \frac{2}{n} \sum_{i=1}^n X_i Y_i$ is an unbiased estimator of $\frac{1}{k} \sum_{j=1}^k w_j^*$ (in fact, w^* need not be the same across successive blocks) and its mean square error is less than a multiple of $\frac{d}{n}(\|w^*\|^2 + \sigma^2) \log(1/\delta)$ with probability at least $1 - \delta$. Our problem, however, is not directly amenable to such a method.

Proof of Theorem 14. The conditions on ρ , $\|\theta\|$, and $\|\theta^*\|$ ensure that the factor on the right side of inequality (5.3) multiplying $\|\theta - \theta^*\|$ is less than 1.

Consider the set $\mathcal{A} = \{\theta : \langle \theta, \theta^* \rangle > \rho \|\theta\| \|\theta^*\|, 10\sigma \leq \|\theta\| \leq L\sigma\}$. We will show that the empirical EM updates stay in this set. That is, $M_n(\mathcal{A}) \subseteq \mathcal{A}$. This is based on Lemma 18 which shows that

$$M(\mathcal{A}) \subseteq \{\theta : \langle \theta, \theta^* \rangle > (1 + \Delta)\rho \|\theta\| \|\theta^*\|, \|\theta^*\|(1 - \kappa) \leq \|\theta\| \leq \sqrt{\sigma^2 + 3\|\theta^*\|^2}\}.$$

This statement is what allows us to say that $M_n(\mathcal{A}) \subseteq \mathcal{A}$; in particular when M_n is close

to M . To be precise, assume $\sup_{\theta \in \mathcal{A}} \|M_n(\theta) - M(\theta)\| < \epsilon$. That implies

$$\sup_{\theta \in \mathcal{A}} \left\| \frac{M_n(\theta)}{\|M_n(\theta)\|} - \frac{M(\theta)}{\|M(\theta)\|} \right\| \leq \sup_{\theta \in \mathcal{A}} \frac{2\|M_n(\theta) - M(\theta)\|}{\|M(\theta)\|} < \frac{2\epsilon}{(1-\kappa)\|\theta^*\|}.$$

For the last inequality, we used the fact that $\|M(\theta)\| \geq \|\theta^*\|(1-\kappa)$ for all θ in \mathcal{A} . It follows from Lemma 18 that

$$\begin{aligned} \sup_{\theta \in \mathcal{A}} \langle \theta^*, \frac{M_n(\theta)}{\|M_n(\theta)\|} \rangle &\geq \sup_{\theta \in \mathcal{A}} \langle \theta^*, \frac{M(\theta)}{\|M(\theta)\|} \rangle - \frac{2\epsilon}{(1-\kappa)} \\ &\geq \|\theta^*\|(1+\Delta)\rho - \frac{2\epsilon}{(1-\kappa)} \\ &\geq \|\theta^*\|\rho, \end{aligned}$$

provided $\epsilon < (\frac{1-\kappa}{2})\Delta\rho\|\theta^*\|$ and

$$\begin{aligned} \sup_{\theta \in \mathcal{A}} \|M_n(\theta)\| &\geq \sup_{\theta \in \mathcal{A}} \|M(\theta)\| - \epsilon \\ &\geq \|\theta^*\|(1-\kappa) - \epsilon \\ &\geq 20\sigma(1-\kappa) - \epsilon \\ &\geq 10\sigma, \end{aligned}$$

provided $\epsilon < 10\sigma(1-2\kappa)$. Also, note that

$$\begin{aligned} \sup_{\theta \in \mathcal{A}} \|M_n(\theta)\| &\leq \sup_{\theta \in \mathcal{A}} \|M(\theta)\| + \epsilon \\ &\leq \sqrt{\sigma^2 + 3\|\theta^*\|^2} + \epsilon \\ &\leq L\sigma, \end{aligned}$$

provided $\epsilon < L\sigma - \sqrt{\sigma^2 + 3\|\theta^*\|^2}$. For this to be true, we also require that L be large enough so that $L\sigma - \sqrt{\sigma^2 + 3\|\theta^*\|^2} > 0$.

For $\delta \in (0, 1)$, let $\epsilon_M(n, \delta)$ be the smallest number such that for any fixed θ in \mathcal{A} , we have

$$\|M_n(\theta) - M(\theta)\| \leq \epsilon_M(n, \delta),$$

with probability at least $1 - \delta$. Moreover, suppose n is large enough so that

$$\epsilon_M(n, \delta) \leq \min\{10\sigma(1 - 2\kappa), (\frac{1 - \kappa}{2})\Delta\rho\|\theta^\star\|, L\sigma - \sqrt{\sigma^2 + 3\|\theta^\star\|^2}\},$$

which guarantees that $M_n(\mathcal{A}) \subseteq \mathcal{A}$. For any iteration $t \in [T]$, we have

$$\|M_{n/T}(\theta^t) - M(\theta^t)\| \leq \epsilon_M(n/T, \delta/T),$$

with probability at least $1 - \delta/T$. Thus by a union bound and $M_n(\mathcal{A}) \subseteq \mathcal{A}$,

$$\max_{t \in [T]} \|M_{n/T}(\theta^t) - M(\theta^t)\| \leq \epsilon_M(n/T, \delta/T),$$

with probability at least $1 - \delta$.

Hence if θ^0 belongs to \mathcal{A} , then by Lemma 15,

$$\begin{aligned} \|\theta^{t+1} - \theta^\star\| &= \|M_{n/T}(\theta^t) - \theta^\star\| \\ &\leq \|M(\theta^t) - \theta^\star\| + \|M_{n/T}(\theta^t) - M(\theta^t)\| \\ &\leq \gamma\|\theta^t - \theta^\star\| + \max_{t \in [T]} \|M_{n/T}(\theta) - M(\theta)\| \\ &\leq \gamma\|\theta^t - \theta^\star\| + \epsilon_M(n/T, \delta/T). \end{aligned}$$

Solving this recursive inequality yields,

$$\begin{aligned} \|\theta^t - \theta^\star\| &\leq \gamma^t\|\theta^0 - \theta^\star\| + \epsilon_M(n/T, \delta/T) \sum_{i=0}^{t-1} \gamma^i \\ &\leq \gamma^t\|\theta^0 - \theta^\star\| + \frac{\epsilon_M(n/T, \delta/T)}{1 - \gamma}, \end{aligned}$$

with probability at least $1 - \delta$.

Finally, it was shown in [1] that

$$\epsilon_M(n/T, \delta/T) \leq C\sqrt{\|\theta^\star\|^2 + \sigma^2} \sqrt{\frac{dT \log(T/\delta)}{n}}$$

with probability at least $1 - \delta/T$. □

5.6 Extensions to other models

In this section, we discuss how the theory we developed can be used to study the following nonlinear mixture models:

$$Y_i = R_i(\langle \theta^*, X_i \rangle)_+ + \epsilon_i, \quad (5.8)$$

or

$$\tilde{Y}_i = R_i \mathbb{1} \{ \langle \theta^*, X_i \rangle > 0 \} + \epsilon_i, \quad \|\theta^*\| = 1. \quad (5.9)$$

The first model is a symmetric mixture of two ramp activation functions and the second model is a symmetric mixture of two unit step functions. It turns out that the empirical iterates

$$\theta^{t+1} = L_n(\theta^t),$$

where

$$L_n(\theta) = L_n(\theta, \underline{Y}, \underline{X}) = 2 \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^T \right)^{-1} \left[\frac{1}{n} \sum_{i=1}^n (2\phi(Y_i(\langle \theta, X_i \rangle)_+ / \sigma^2) - 1) X_i Y_i \right]$$

can be used for *either* Model (5.9) via $\hat{\theta} = \theta^t$ or Model (5.8) via $\hat{\theta} = \theta^t / \|\theta^t\|$, provided the norm of the initializer $\|\theta^0\|$ is sufficiently large. More precisely, with high probability,

$$\left\| \frac{L_n(\theta, \tilde{\underline{Y}}, \underline{X})}{\|L_n(\theta, \tilde{\underline{Y}}, \underline{X})\|} - \frac{\theta^*}{\|\theta^*\|} \right\| \leq \left\| \frac{L_n(s\theta, \underline{Y}, \underline{X})}{\|L_n(s\theta, \underline{Y}, \underline{X})\|} - \frac{\theta^*}{\|\theta^*\|} \right\| + O(1/s).$$

The analogous population operator is

$$L(\theta) = 4\mathbb{E} [\phi(Y(\langle \theta, X \rangle)_+ / \sigma^2) XY].$$

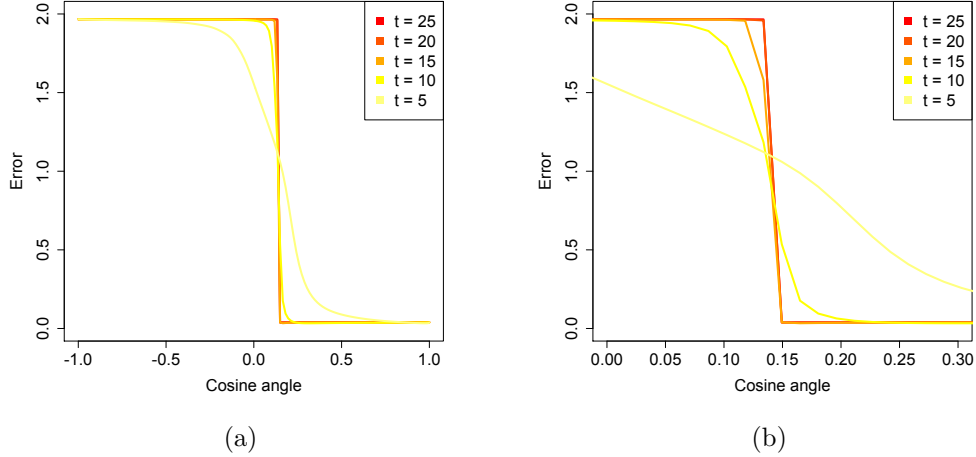
Note that these are *not* the EM operators for the respective problems; in fact, there is no unique solution to the "maximization" part of the algorithm. It can be shown that $\|L(\theta) - \theta^*\| = \|M(\theta) - \theta^*\|$ and hence the results from Lemma 15, Theorem 13, and Theorem 14 hold verbatim. What is important is that our basin of attraction is a cone, and thus as long as the cosine angle of the initializer θ^0 with θ^* is sufficiently large, irrespective of the size of θ^0 , we are guaranteed convergence to θ^* . Note that the previously established

basin of attraction equal to a ball around θ^* does not suffice for this purpose.

5.7 Discussion

In this chapter, we showed that the empirical EM iterates converge to true coefficients of a mixture of two linear regressions as long as the initializer lies within a cone (see the condition on Theorem 14: $\langle \theta^0, \theta^* \rangle > \rho \|\theta^0\| \|\theta^*\|$).

In Fig. 5.2a, we perform a simulation study of $\theta^{t+1} \leftarrow M_n(\theta^t)$ with $\sigma = 1$, $n = 1000$, $d = 2$, and $\theta^* = (1, 0)'$. All entries of the design matrix X and the noise ε are generated *i.i.d.* from a standard normal distribution. We consider the error $\|\theta^t - \theta^*\|$ plotted as a function of $\cos \alpha = \frac{\langle \theta^0, \theta^* \rangle}{\|\theta^0\| \|\theta^*\|}$ at iterations $t = 5, 10, 15, 20, 25$ (corresponding the shaded curves). For each t , we choose a unit vector θ^0 so that $\cos \alpha$ ranges between -1 and $+1$. In accordance with the theory we have developed, increasing the iteration size and cosine angle decreases the overall error. According to Theorem 13, the algorithm should suffer from small $\cos \alpha$. Indeed, we observe a sharp transition at $\cos \alpha \approx 0.15$. The algorithm converges to $(-1, 0)'$ for initializers with cosine angle smaller than this. The plot in Fig. 5.2b is a zoomed-in version of Fig. 5.2a near this transition point.



5.8 Additional proofs

Proof of Theorem 15. Follow the proof of Theorem 14 which gives convergence rates for the symmetric EM iterates. We have

$$\begin{aligned} \|\theta^{t+1} - \theta^*\| &\leq \gamma\|\theta^t - \theta^*\| + \max_{t \in [T]} \|M_{n/T}(\theta^t, \underline{X}, \tilde{Y}) - M(\theta^t)\| + \\ &\quad \max_{t \in [T]} \|M_{n/T}(\theta^t, \underline{X}, \tilde{Y}) - M_{n/T}(\theta^t, \underline{X}, Y^{(s)})\|. \end{aligned}$$

The second term was handled in the proof of Theorem 14. We only need to bound the third term. It suffices to show that

$$\|M_{n/T}(\theta^t, \underline{X}, \tilde{Y}) - M_{n/T}(\theta^t, \underline{X}, Y^{(s)})\| \leq \epsilon_S(n/T, \delta/T)$$

with probability at least $1 - \delta/T$. We need

$$\epsilon_S(n, \delta) \leq D_3 \sqrt{\frac{d}{n} (\|s\|^2 + \|\theta^*\|^2 + \sigma^2) \log \frac{1}{\delta}}$$

for some $D_3 > 0$. That is an easy consequence of Lemma 16 and Lemma 17. The rest of the proof follows exactly as that of Theorem 14. \square

Proof of Theorem 13. Note that in general, $M(\theta) = \theta^* A + \theta B$, where

$$A = \mathbb{E}[2\phi(W\langle\theta, X\rangle/\sigma^2) + 2(W\langle\theta, X\rangle/\sigma^2)\phi'(W\langle\theta, X\rangle/\sigma^2) - 1],$$

$$B = 2\mathbb{E}[(W^2/\sigma^2)\phi'(W\langle\theta, X\rangle/\sigma^2)].$$

Suppose $\langle\theta, \theta^*\rangle = 0$. This implies that $A = 0$. To see this, note that

$$\mathbb{E}\phi(W\langle\theta, X\rangle) = \mathbb{E}\phi(\Lambda Z_1 Z_2) = \phi(0) = 1/2,$$

and

$$\mathbb{E}[W\langle\theta, X\rangle\phi'(W\langle\theta, X\rangle)] = \mathbb{E}[\Lambda Z_1 Z_2 \phi'(\Lambda Z_1 Z_2)] = 0.$$

Next, observe that $B = 2(1 + \|\theta^\star\|^2/\sigma^2)\mathbb{E}[Z_2^2\phi'(Z_1Z_2\|\theta\|\sqrt{\sigma^2 + \|\theta^\star\|^2/\sigma^2})] \rightarrow 1 + \|\theta^\star\|^2/\sigma^2 > 1$ as $\theta \rightarrow 0$. By continuity, there exists $a > 0$ such that if $\|\theta\| = a$, then $B > 1$ and hence

$$\begin{aligned}\|M(\theta) - \theta^\star\|^2 &= \|\theta - \theta^\star\|^2 + (B^2 - 1)\|\theta\|^2 \\ &> \|\theta - \theta^\star\|^2.\end{aligned}$$

This shows that

$$\lim_{\langle \theta, \theta^\star \rangle \rightarrow 0, \|\theta\|=a} [\|M(\theta) - \theta^\star\|^2 - \|\theta - \theta^\star\|^2] > 0.$$

By continuity, it follows that there are choices of θ with $\langle \theta, \theta^\star \rangle > 0$ such that $\|M(\theta) - \theta^\star\|^2 > \|\theta - \theta^\star\|^2$. \square

Lemma 16. *There exists constant $D_1 > 0$, such that*

$$\mathbb{P} \left\{ \|\hat{s} - s\| \leq D_1 \sqrt{\frac{d}{n} ((\|\theta_1^\star\|^2 + \|\theta_2^\star\|^2)/2 + \sigma^2) \log(1/\delta)} \right\} \geq 1 - \delta$$

for all $\delta \in (0, 1)$.

Proof. Denote $\hat{\Sigma} = \frac{1}{n} X_i X_i^T$. Recall that $\hat{s} = \frac{1}{n} \sum X_i Y_i$. We have

$$\begin{aligned}\|\hat{s} - s\| &= \left\| \frac{1}{n} \sum_{i=1}^n X_i (\langle s, X_i \rangle + R_i \langle \theta^\star, X_i \rangle + \varepsilon_i) - s \right\| \\ &\leq \left\| \left(\hat{\Sigma} - I \right) s \right\| + \left\| \frac{1}{n} \sum_{i=1}^n R_i X_i X_i^T \theta^\star \right\| + \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\| \\ &\leq c \sqrt{\frac{d}{n} \log(1/\delta) (\|s\|^2 + \|\theta^\star\|^2 + \sigma^2)} \\ &= c \sqrt{\frac{d}{n} ((\|\theta_1^\star\|^2 + \|\theta_2^\star\|^2)/2 + \sigma^2) \log(1/\delta)},\end{aligned}$$

for some $c > 0$ with probability at least $1 - \delta$. \square

Lemma 17. *There exists constant $D_2 > 0$ for which*

$$\mathbb{P} \left\{ \|M_n(\theta, \underline{X}, \underline{Y}^{(s)}) - M_n(\theta, \underline{X}, \tilde{\underline{Y}})\| \leq D_2 \|\hat{s} - s\| \right\} \rightarrow 1$$

for all $\theta \in \mathbb{R}^d$.

Proof. Write

$$\begin{aligned} & M_n(\theta, \underline{X}, \underline{Y}^{(s)}) - M_n(\theta, \underline{X}, \underline{\tilde{Y}}) \\ &= \widehat{\Sigma}^{-1} \frac{2}{n} \sum_{i=1}^n \left[\phi \left(Y_i^{(s)} \langle \theta, X_i \rangle \right) X_i Y_i^{(s)} - \phi \left(\tilde{Y}_i \langle \theta, X_i \rangle \right) X_i \tilde{Y}_i \right] \\ & \quad + \widehat{\Sigma}^{-1} \frac{1}{n} \sum_{i=1}^n X_i \left(\tilde{Y}_i - Y_i^{(s)} \right). \end{aligned}$$

Use triangle inequality to deduce that

$$\begin{aligned} & \|M_n(\theta, \underline{X}, \underline{Y}^{(s)}) - M_n(\theta, \underline{X}, \underline{\tilde{Y}})\| \\ & \leq \left\| \widehat{\Sigma}^{-1} \frac{2}{n} \sum_{i=1}^n \phi \left(Y_i^{(s)} \langle \theta, X_i \rangle \right) X_i \langle \widehat{s} - s, X_i \rangle \right\| \\ & \quad + \left\| \widehat{\Sigma}^{-1} \frac{2}{n} \sum_{i=1}^n \left(\phi \left(Y_i^{(s)} \langle \theta, X_i \rangle \right) - \phi \left(\tilde{Y}_i \langle \theta, X_i \rangle \right) \right) X_i \tilde{Y}_i \right\| \\ & \quad + \left\| \widehat{\Sigma}^{-1} \frac{1}{n} \sum_{i=1}^n X_i X_i^T (\widehat{s} - s) \right\|. \end{aligned}$$

The first and the third term can be bounded by a constant multiple of $\|\widehat{s} - s\|$ with high probability. Simply notice that $\mathbb{P}\{\|\widehat{\Sigma}^{-1}\|_{op} > 2\} \rightarrow 0$ and $|\phi| \leq 1$. For the second term, use the mean-value theorem and the basic inequality $|u\phi'(u)| < e^{-|u|}$ for all $u \in \mathbb{R}$ to bound this term by

$$\|\widehat{\Sigma}^{-1}\|_{op} \frac{2}{n} \sum_{i=1}^n \left| \frac{\tilde{Y}_i}{Y_i^{(m)}} \right| \exp(-|Y_i^{(m)} \langle \theta, X_i \rangle|) \cdot \|X_i\| \|\widehat{s} - s\|$$

for some $Y_i^{(m)}$ that lies between $Y_i^{(s)}$ and \tilde{Y}_i . The above is bounded by a constant multiple of $\|\widehat{s} - s\|$ with high probability. \square

For the following lemmata, let

$$A = \mathbb{E}[2\phi(W\langle\theta, X\rangle/\sigma^2) + 2(W\langle\theta, X\rangle/\sigma^2)\phi'(W\langle\theta, X\rangle/\sigma^2) - 1],$$

$$B = 2\mathbb{E}[(W^2/\sigma^2)\phi'(W\langle\theta, X\rangle/\sigma^2)],$$

and

$$\kappa^2 = \frac{1}{\frac{\Gamma}{\Lambda} \min \left\{ \Lambda, \frac{\Gamma}{\Lambda} \right\} + 1} = \max \left\{ 1 - \frac{|\langle \theta_0, \theta^* \rangle|^2}{\sigma^2 + \|\theta^*\|^2}, 1 - \frac{\langle \theta, \theta^* \rangle}{\sigma^2 + \langle \theta, \theta^* \rangle} \right\}.$$

Lemma 18. *The cosine angle between θ^* and $M(\theta)$ is equal to*

$$\frac{\|\theta^*\|^2 A + \langle \theta, \theta^* \rangle B}{\sqrt{(\|\theta^*\|^2 A + \langle \theta, \theta^* \rangle B)^2 + B^2(\|\theta\|^2 \|\theta^*\|^2 - |\langle \theta, \theta^* \rangle|^2)}}.$$

If $\langle \theta, \theta^* \rangle \geq \rho \|\theta\| \|\theta^*\|$ and $3\sigma \leq \|\theta\| \leq L\sigma$, then there exists positive $\Delta = \Delta(\rho, \sigma, \|\theta^*\|, L)$ such that this cosine angle is at least $(1 + \Delta)\rho$. Moreover,

$$\|\theta^*\|^2 (1 - \kappa)^2 \leq \|M(\theta)\|^2 = \|\theta^*\|^2 A^2 + \|\theta\|^2 B^2 + 2\langle \theta, \theta^* \rangle AB \leq \sigma^2 + 3\|\theta^*\|^2,$$

and

$$\langle \theta^*, M(\theta) \rangle = \|\theta^*\|^2 A + \langle \theta, \theta^* \rangle B \geq \|\theta^*\|^2 (1 - \kappa).$$

Proof. We will prove the first statement. Let $\tau = \frac{\|\theta^*\|}{\|\theta\|} \frac{A}{B}$. Observe that

$$\begin{aligned} \frac{\|\theta^*\|^2 A + \langle \theta, \theta^* \rangle B}{\sqrt{(\|\theta^*\|^2 A + \langle \theta, \theta^* \rangle B)^2 + B^2(\|\theta\|^2 \|\theta^*\|^2 - |\langle \theta, \theta^* \rangle|^2)}} &= \frac{1}{\sqrt{1 + \frac{\|\theta\|^2 \|\theta^*\|^2 - |\langle \theta, \theta^* \rangle|^2}{(\|\theta^*\|^2 \frac{A}{B} + \langle \theta, \theta^* \rangle)^2}}} \\ &\geq \frac{1}{\sqrt{1 + \frac{1 - \rho^2}{(\tau + \rho)^2}}} \\ &= \frac{\rho}{\sqrt{1 - (1 - \rho^2) \frac{\tau(\tau + 2\rho)}{(\tau + \rho)^2}}} \\ &\geq \frac{\rho}{\sqrt{1 - (1 - \rho^2) \frac{\tau}{\tau + \rho}}} \\ &\geq \rho \left(1 + \frac{1}{2}(1 - \rho^2) \frac{\tau}{\tau + \rho}\right), \end{aligned}$$

where the last line follows from the inequality $1/\sqrt{1 - a} \geq 1 + a/2$ for all $a \in (0, 1)$.

Finally, note that from Lemma 23,

$$\frac{A}{B} \geq \frac{\sigma^2(1 - \kappa)}{2(\|\theta^*\|^2 + \sigma^2)\kappa^3}.$$

Thus, $\tau \geq \tau_0 := \frac{\sigma \|\theta^\star\| (1-\kappa)}{2L(\|\theta^\star\|^2 + \sigma^2) \kappa^3}$ and so we can set

$$\Delta = \frac{1}{2}(1 - \rho^2) \frac{\tau_0}{\tau_0 + \rho} > 0.$$

For the second claim, the identity

$$\|M(\theta)\|^2 = \|\theta^\star\|^2 A^2 + \|\theta\|^2 B^2 + 2\langle \theta, \theta^\star \rangle AB$$

is an immediate consequence of $M(\theta) = A\theta^\star + B\theta$. By Lemma 23, $A \geq 1 - \kappa$ and hence since $\langle \theta, \theta^\star \rangle \geq 0$, we have $\|M(\theta)\|^2 \geq \|\theta^\star\|^2 A^2 \geq \|\theta^\star\|^2 (1 - \kappa)^2$.

Next, we will show that $\|M(\theta)\|^2 \leq \sigma^2 + 3\|\theta^\star\|^2$. To see this, note that by Lemma 20 and Jensen's inequality,

$$\begin{aligned} \langle \theta, M(\theta) \rangle &= \mathbb{E}[(2\phi(W\langle \theta, X \rangle) - 1)W\langle \theta, X \rangle] \\ &\leq \mathbb{E}|W\langle \theta, X \rangle| \\ &\leq \sqrt{\mathbb{E}|W\langle \theta, X \rangle|^2} \\ &= \sqrt{\Lambda^2 + 3\Gamma^2} \\ &= \|\theta\| \sqrt{\sigma^2 + \|\theta^\star\|^2 + 2|\langle \theta_0, \theta^\star \rangle|^2}. \end{aligned}$$

Next, it can be shown that $A \leq \sqrt{2}$ and hence

$$\begin{aligned} \langle \theta_0^\perp, M(\theta) \rangle &= A\langle \theta_0^\perp, \theta^\star \rangle \\ &\leq \sqrt{2}\langle \theta_0^\perp, \theta^\star \rangle. \end{aligned}$$

Putting these two facts together, we have

$$\begin{aligned} \|M(\theta)\|^2 &= |\langle \theta_0^\perp, M(\theta) \rangle|^2 + |\langle \theta_0, M(\theta) \rangle|^2 \\ &\leq \sigma^2 + \|\theta^\star\|^2 + 2|\langle \theta_0^\perp, \theta^\star \rangle|^2 + 2|\langle \theta_0, \theta^\star \rangle|^2 \\ &= \sigma^2 + 3\|\theta^\star\|^2. \end{aligned}$$

The final statement

$$\langle \theta^*, M(\theta) \rangle = \|\theta^*\|^2 A + \langle \theta, \theta^* \rangle B \geq \|\theta^*\|^2 (1 - \kappa).$$

follows from similar arguments. □

Lemma 19. *If $\langle \theta, \theta^* \rangle \geq 0$ and $\sigma^2 = 1$, then*

$$\mathbb{E}[W \langle \theta, X \rangle \phi'(W \langle \theta, X \rangle)] \geq 0.$$

Proof. Note that the statement is true if

$$\mathbb{E}[(\alpha Z + \beta) \phi'(\alpha Z + \beta)] \geq 0,$$

where $Z \sim N(0, 1)$ and $\alpha \geq 0$ and $\beta \geq 0$. This fact is proved in Lemma 5 in [6] or Lemma 1 in [74]. □

Lemma 20. *Assume $\sigma^2 = 1$. Then*

$$\begin{aligned} \langle \theta, M(\theta) \rangle &= \mathbb{E}[(2\phi(W \langle \theta, X \rangle) - 1)W \langle \theta, X \rangle] \\ &= \mathbb{E}[(2\phi(\Lambda Z_1 Z_2 + \Gamma Z_2^2) - 1)(\Lambda Z_1 Z_2 + \Gamma Z_2^2)], \end{aligned}$$

and

$$\langle \theta_0^\perp, M(\theta) \rangle = \langle \theta_0^\perp, \theta^* \rangle \mathbb{E}[2\phi(W \langle \theta, X \rangle) + 2W \langle \theta, X \rangle \phi'(W \langle \theta, X \rangle) - 1].$$

Lemma 21. *The following inequalities hold for all $x \in \mathbb{R}$:*

$$|2\phi(x) + 2x\phi'(x) - 1| \leq 1 + \sqrt{2(1 - \phi(x))},$$

and

$$x^2 \phi'(x) \leq \sqrt{2(1 - \phi(x))}.$$

Lemma 22. *Let $\alpha, \beta > 0$ and $Z \sim N(0, 1)$. Then*

$$\mathbb{E}2(1 - \phi(\alpha(Z + \beta))) \leq \exp\{-\frac{\beta}{2} \min\{\alpha, \beta\}\}.$$

Moreover,

$$\mathbb{E}2(1 - \phi(\alpha Z_2(Z_1 + \beta Z_2))) \leq \frac{1}{\sqrt{\beta \min\{\alpha, \beta\} + 1}}.$$

Proof. The second conclusion follows immediately from the first since

$$\begin{aligned} \mathbb{E}2(1 - \phi(\alpha Z_2(Z_1 + \beta Z_2))) &= \mathbb{E}_{Z_2} \mathbb{E}_{Z_1} 2(1 - \phi(\alpha |Z_2| (Z_1 + \beta |Z_2|))) \\ &\leq \mathbb{E}_{Z_2} \exp\{-\frac{Z_2^2}{2} \beta \min\{\alpha, \beta\}\} \\ &= \frac{1}{\sqrt{\beta \min\{\alpha, \beta\} + 1}}. \end{aligned}$$

The last equality follows from the moment generating function of $\chi^2(1)$.

For the first conclusion, we first observe that the mapping $\alpha \mapsto \mathbb{E}\phi(\alpha(Z + \beta))$ is increasing (Lemma 5 in [6] or Lemma 1 in [74]). Next, note the inequality

$$2(1 - \phi(x)) \leq e^{-x},$$

which is equivalent to $(e^x - 1)^2 \geq 0$. If $\alpha \geq \beta$, then

$$\begin{aligned} \mathbb{E}2(1 - \phi(\alpha(Z + \beta))) &\leq \mathbb{E}2(1 - \phi(\beta(Z + \beta))) \\ &\leq \mathbb{E}e^{-(\beta(Z + \beta))} \\ &= e^{-\beta^2/2}. \end{aligned}$$

If $\alpha \leq \beta$, then

$$\begin{aligned} \mathbb{E}2(1 - \phi(\alpha(Z + \beta))) &\leq \mathbb{E}e^{-(\alpha(Z + \beta))} \\ &= e^{\alpha^2/2 - \alpha\beta} \\ &\leq e^{-\alpha\beta/2}. \end{aligned}$$

In each case, we used the moment generating function of a normal distribution to evaluate the expectations. \square

Lemma 23. *Assume $\sigma^2 = 1$. We have*

$$1 - \kappa \leq A \leq 1 + \sqrt{\kappa},$$

and

$$B \leq 2(1 + \|\theta^*\|^2)\kappa^3.$$

Proof. By Lemma 19 and Lemma 22,

$$\begin{aligned} A &= \mathbb{E}[2\phi(W\langle\theta, X\rangle) + 2W\langle\theta, X\rangle\phi'(W\langle\theta, X\rangle) - 1] \\ &\geq \mathbb{E}[2\phi(W\langle\theta, X\rangle) - 1] \\ &\geq 1 - \kappa. \end{aligned}$$

By Lemma 21, Jensen's inequality, and Lemma 22,

$$\begin{aligned} A &= \mathbb{E}[2\phi(W\langle\theta, X\rangle) + 2W\langle\theta, X\rangle\phi'(W\langle\theta, X\rangle) - 1] \\ &\leq \mathbb{E}[1 + \sqrt{2(1 - \phi(W\langle\theta, X\rangle))}] \\ &\leq 1 + \sqrt{\mathbb{E}2(1 - \phi(W\langle\theta, X\rangle))} \\ &\leq 1 + \sqrt{\kappa}. \end{aligned}$$

By Lemma 22,

$$\begin{aligned}
B &= 2\mathbb{E}[W^2\phi'(W\langle\theta, X\rangle)] \\
&\leq 2\mathbb{E}[2W^2(1 - \phi(W\langle\theta, X\rangle))] \\
&= 2(1 + \|\theta^\star\|^2)\mathbb{E}_{Z_2}Z_2^2\mathbb{E}_{Z_1}[2(1 - \phi(\Lambda Z_2(Z_1 + \frac{\Gamma}{\Lambda}Z_2)))] \\
&\leq 2(1 + \|\theta^\star\|^2)\mathbb{E}_{Z_2}[Z_2^2 \exp\{-\frac{Z_2^2}{2}\frac{\Gamma}{\Lambda} \min\left\{\frac{\Gamma}{\Lambda}, \Lambda\right\}\}] \\
&= 2(1 + \|\theta^\star\|^2) \left(\frac{1}{\frac{\Gamma}{\Lambda} \min\left\{\Lambda, \frac{\Gamma}{\Lambda}\right\} + 1} \right)^{3/2} \\
&= 2(1 + \|\theta^\star\|^2)\kappa^3.
\end{aligned}$$

□

Lemma 24. *Define*

$$h(\alpha, \beta) = \mathbb{E}[(2\phi(\alpha Z_2(Z_1 + \beta Z_2)) - 1)(Z_2(Z_1 + \beta Z_2))].$$

Let $\alpha, \beta > 0$. Then

$$\frac{\partial}{\partial \alpha} h(\alpha, \beta) \leq \frac{2}{\alpha^2} \left(\frac{1}{\beta \min\{\alpha, \beta\} + 1} \right)^{1/4}.$$

Proof. First, observe that

$$\frac{\partial}{\partial \alpha} h(\alpha, \beta) = \mathbb{E}[2\phi'(\alpha Z_2(Z_1 + \beta Z_2))(Z_2(Z_1 + \beta Z_2))^2].$$

By Lemma 21, Jensen's inequality, and Lemma 22,

$$\begin{aligned}
&\mathbb{E}[2\phi'(\alpha Z_2(Z_1 + \beta Z_2))(Z_2(Z_1 + \beta Z_2))^2] \\
&= \frac{1}{\alpha^2} \mathbb{E}[2\phi'(\alpha Z_2(Z_1 + \beta Z_2))(\alpha Z_2(Z_1 + \beta Z_2))^2] \\
&\leq \frac{2}{\alpha^2} \mathbb{E}\sqrt{2(1 - \phi(\alpha Z_2(Z_1 + \beta Z_2)))} \\
&\leq \frac{2}{\alpha^2} \sqrt{\mathbb{E}2(1 - \phi(\alpha Z_2(Z_1 + \beta Z_2)))} \\
&\leq \frac{2}{\alpha^2} \left(\frac{1}{\beta \min\{\alpha, \beta\} + 1} \right)^{1/4}.
\end{aligned}$$



Chapter 6

Recovering the endpoint of a density from noisy data with application to convex body estimation

6.1 Preliminaries

6.1.1 Introduction

The problem of estimating the support of a distribution, given i.i.d. samples, poses both statistical and computational questions. When the support of the distribution is known to be convex, geometric methods have been borrowed from stochastic and convex geometry with the use of random polytopes since the seminal works [86,87]. When the distribution of the samples is uniform on a convex body, estimation in a minimax setup has been tackled in [88] (see also the references therein). There, the natural estimator defined as the convex hull of the samples (which is referred to as *random polytope* in the stochastic geometry literature) is shown to attain the minimax rate of convergence on the class of convex bodies, under the Nikodym metric.

When the samples are still supported on a convex body but their distribution is no

longer uniform, [89] studies the performance of the random polytope as an estimator of the convex support under the Nikodym metric, whereas [90] focuses on the Hausdorff metric. In the latter, computational issues are addressed in higher dimensions. Namely, determining the list of vertices of the convex hull of n points in dimension $d \geq 2$ is very expensive, namely, exponentially in $d \log n$ (see [91]). In [90], a randomized algorithm produces an approximation of the random polytope that achieves a trade off between computational cost and statistical accuracy. The approximation is given in terms of a membership oracle, which is a very desirable feature for the computation/approximation of a convex body.

Both works [89, 90] assume that one has access to direct samples. What if these samples are contaminated, e.g., subject to measurement errors? In [92], a closely related problem is studied, where two independent contaminated samples are observed, and one wants to estimate the set where $f - g$ is positive, where f and g are the respective densities of the two samples. In that work, the contamination is modeled as an additive noise with known distribution, and some techniques borrowed from inverse problems are used. The main drawback is that the estimator is not tractable and it only gives a theoretical benchmark for minimax estimation.

Goldenshluger and Tsybakov [9] study the problem of estimating the endpoint of a univariate distribution, given samples contaminated with additive noise. Their analysis suggests that their estimator is optimal in a minimax sense and its computation is straightforward. The simplicity of their procedure is due to the dominating bias phenomenon. In our work, we use this phenomenon in order to extend their result, which then we lift to a higher dimensional setup: that of estimating the convex support of a uniform distribution, given samples that are contaminated with additive Gaussian noise. Our method relies on projecting the data points along a finite collection of unit vectors. Unlike in [92], we give an explicit form for our estimator. In addition, our estimator is tractable when the ambient dimension is not too large. If the dimension is too high, the number of steps required to compute a membership oracle for our estimator becomes exponentially large: namely, of order $(O(\ln n))^{(d-1)/2}$.

6.1.2 Notation

6.1.3 Notation

In this work, $d \geq 2$ is a fixed integer standing for the dimension of the ambient Euclidean space \mathbb{R}^d . The Euclidean ball with center $a \in \mathbb{R}^d$ and radius $r \geq 0$ is denoted by $B_d(a, r)$. The unit sphere in \mathbb{R}^d is denoted by \mathbb{S}^{d-1} and β_d stands for the volume of the unit Euclidean ball.

We refer to convex and compact sets with nonempty interior in \mathbb{R}^d as convex bodies. The collection of all convex bodies in \mathbb{R}^d is denoted by \mathcal{K}_d . Let $\sigma^2 > 0$ and $n \geq 1$. If X_1, \dots, X_n are i.i.d. random uniform points in a convex body G and ξ_1, \dots, ξ_n are i.i.d. d -dimensional centered Gaussian random vectors with covariance matrix $\sigma^2 I$, where I is the $d \times d$ identity matrix, independent of the X_i 's, we denote by \mathbb{P}_G the joint distribution of $X_1 + \varepsilon_1, \dots, X_n + \varepsilon_n$ and by \mathbb{E}_G the corresponding expectation operator (we omit the dependency on n and σ^2 for simplicity).

The support function of a convex set $G \subseteq \mathbb{R}^d$ is defined as $h_G(u) = \sup_{x \in G} \langle u, x \rangle$, $u \in \mathbb{R}^d$, where $\langle \cdot, \cdot \rangle$ is the canonical scalar product in \mathbb{R}^d .

The Hausdorff distance between two sets $A, B \subseteq \mathbb{R}^d$ is

$$d_H(A, B) = \inf\{\varepsilon > 0 : G_1 \subseteq G_2 + \varepsilon B_d(0, 1) \text{ and } G_2 \subseteq G_1 + \varepsilon B_d(0, 1)\}.$$

If A and B are convex bodies, then the Hausdorff distance between them can be written in terms of their support functions, namely,

$$d_H(A, B) = \sup_{u \in \mathbb{S}^{d-1}} |h_A(u) - h_B(u)|.$$

For f in $L^1(\mathbb{R}^d)$, let $\mathcal{F}[f](t) = \int_{\mathbb{R}^d} e^{i\langle t, x \rangle} f(x) dx$ denote the Fourier transform of f .

The total variation distance between two distributions P and Q having densities p and q with respect to a dominating measure μ is defined by $\text{TV}(P, Q) = \int |p - q| d\mu$.

The Lebesgue measure of a measurable, bounded set A in \mathbb{R}^d is denoted by $|A|$. For a vector $x = (x_1, x_2, \dots, x_d)'$, we define $\|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$ for $p \geq 1$ and $\|x\|_\infty =$

$\sup_{1 \leq i \leq d} |x_i|$. For a function, f defined on a set A , let $\|f\|_\infty = \sup_{x \in A} |f(x)|$. The Nikodym distance between two measurable, bounded sets A and B is defined by $\mathbf{d}_\Delta(A, B) = |A \Delta B|$.

We use standard big- O notations, e.g., for any positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for some absolute constant $C > 0$, $a_n = o(b_n)$ or $a_n \ll b_n$ if $\lim a_n/b_n = 0$. Finally, we write $a_n \asymp b_n$ when both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. Furthermore, the subscript in $a_n = O_r(b_n)$ means $a_n \leq C_r b_n$ for some constant C_r depending on the parameter r only. We write $a_n \propto b_n$ when $a_n = Cb_n$ for some absolute constant C . We let ϕ_σ denote the Gaussian density with mean zero and variance σ^2 , i.e., $\phi_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}$ for all $x \in \mathbb{R}$.

6.1.4 Model and outline

A popular class of problems in statistics literature are the so-called inverse or deconvolution problems. Here, the experimenter only has access to contaminated versions of the original variables: $Y = X + \epsilon$, where ϵ follows a known distribution. This problem is usually considered in density or regression contexts [93], [12], [94], but other functionals of the distribution have also been studied [95]. In our setting, a naïve estimator is to take the convex hull of Y_1, \dots, Y_n . However, there is a positive probability that at least one Y will land outside G and these outliers enlarge the boundary of the convex hull so that it overestimates G .

In what follows, we consider the problem of estimating a convex body from noisy observations. More formally, suppose we have access to independent observations

$$Y_i = X_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (6.1)$$

where X_1, \dots, X_n are i.i.d. uniform random points in an unknown convex body G and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. Gaussian random vectors with zero mean and covariance matrix $\sigma^2 I$, independent of X_1, \dots, X_n . In the sequel, we assume that σ^2 is a fixed and known positive number. The goal is to estimate G using Y_1, \dots, Y_n . This can be seen as an inverse problem: the object of interest is a special feature (here, the support) of a density that is observed up to a convolution with a Gaussian distribution. Our approach will not use the path of

inverse problems, but instead, will be essentially based on geometric arguments.

Given an estimator \hat{G}_n of G , we measure its error using the Hausdorff distance. Namely, it is defined as $\mathbb{E}_G [\mathbf{d}_H(\hat{G}_n, G)]$. Let $\mathcal{C} \subseteq \mathcal{K}_d$ be a subclass of convex bodies. The risk of an estimator \hat{G}_n on the class \mathcal{C} is $\sup_{G \in \mathcal{C}} \mathbb{E}_G [\mathbf{d}_H(\hat{G}_n, G)]$ and the minimax risk on \mathcal{C} is defined as

$$\mathcal{R}_n(\mathcal{C}) = \inf_{\hat{G}} \sup_{G \in \mathcal{C}} \mathbb{E}_G [\mathbf{d}_H(\hat{G}, G)],$$

where the infimum is taken over all estimators \hat{G} based on Y_1, \dots, Y_n . The minimax rate on the class \mathcal{C} is the speed at which $\mathcal{R}_n(\mathcal{C})$ goes to zero.

Because the Nikodym distance $\mathbf{d}_\Delta(G_1, G_2)$ is equal to the squared $L^2(\mathbb{R}^d)$ norm between $\mathbb{1}_{G_1}$ and $\mathbb{1}_{G_2}$, it is not surprising that techniques from deconvolution in density and function estimation can be applied. These are usually implemented as plug-in estimators [96], [97], where the density is first estimated using Fourier transforms to form a kernel density estimator and then the support estimator is obtained by thresholding. A pitfall of this method is that the bandwidth parameter must be selected and it is not always clear how to do this in practice. Furthermore, the Fourier transform of the noise distribution must never vanish and hence this excludes compactly supported noise. For example, borrowing ideas from [98], [99], and [92], if $G \subset [-\delta, \delta]^d$, one can consider an estimator \hat{G}_n defined by

$$\hat{G}_n = \arg \max_{G' \in \mathcal{F}_n} \left\{ \frac{1}{n} \sum_{i=1}^n \phi_{G'}(Y_i) - \frac{|G'|}{2} \right\},$$

where $\phi_{G'}$ is a function for which $\mathbb{E} \phi_{G'}(Y) = \mathbb{E} \mathcal{K}_\lambda * \mathbb{1}_{G'}(X) \rightarrow |G \cap G'|/|G|$ as the bandwidth λ of the kernel \mathcal{K}_λ goes to zero and \mathcal{F}_n is a suitably chosen covering of \mathcal{C}_d . One can show [14] that this estimator has an order $1/\sqrt{\ln n}$ convergence with respect to \mathbf{d}_Δ . In addition to being incomputable, a pitfall of this estimator is that the bandwidth parameter must be selected and it is not always clear how to do this in practice.

Our strategy for estimating G avoids standard methods from inverse problems that would require Fourier transforms and tuning parameters. To give intuition for our procedure, first observe that a convex set can be represented in terms of its support function

via

$$G = \{x \in \mathbb{R}^d : \langle u, x \rangle \leq h_G(u) \text{ for all } u \in \mathbb{S}^{d-1}\}.$$

If we can find a suitable way of estimating h_G , say by \hat{h}_n , then there is hope that an estimator of the form

$$\hat{G}_n = \{x \in \mathbb{R}^d : \langle u, x \rangle \leq \hat{h}_n(u) \text{ for all } u \in \mathbb{S}^{d-1}\}$$

will perform well. This is the core idea of our procedure: We project the data points Y_1, \dots, Y_n along unit vectors and for all such $u \in \mathbb{S}^{d-1}$, we estimate the endpoint of the distribution of $\langle u, X_1 \rangle$ given the one dimensional sample $\langle u, Y_1 \rangle, \dots, \langle u, Y_n \rangle$.

A first pass would be to estimate $h_G(u)$ by projecting the data onto a hyperplane $\langle Y, u \rangle$ and then taking the maximum over all observations $\max_{1 \leq i \leq n} \langle Y_i, u \rangle$. However, this estimator will on average overshoot h_G because of the influence of the noise in the variables. We will see that this problem can be overcome by subtracting a suitable, *explicitly defined* sequence b_n to form $\hat{h}_n(u) = \max_{1 \leq i \leq n} \langle Y_i, u \rangle - b_n$. Note that $\hat{h}_n(u)$ is neither subadditive nor positive homogeneous and thus it is not the support function of \hat{G}_n . We will show that \hat{G}_n is still a suitable estimator and that it converges to G at a rate of $\ln \ln n / \sqrt{\ln n}$ in Hausdorff distance. This logarithmic rate is considerably worse than in [90] and is consistent with the sort of slow rates encountered in Gaussian deconvolution problems (more generally known as the ill-posed regime).

Part of our analysis also involves the optimality of our proposed estimator. In other words, we provide a minimax lower bound for this estimation problem. Our strategy boils down to applying Le Cam's two point method and lower bounding the optimization problem

$$\sup_{G_1, G_2 \in \mathcal{C}_d} \{d_H(G_1, G_2) : \text{TV}(P_{G_1}^{\otimes n}, P_{G_2}^{\otimes n}) = O(1/n)\},$$

where $P_{G_k}^{\otimes n}$ denotes the joint distribution of Y_1, \dots, Y_n if X_1, \dots, X_n are sampled uniformly from G_k . We select two sets G_1 and G_2 with equal Lebesgue measure for which $|\mathcal{F}[\mathbb{1}_{G_1} - \mathbb{1}_{G_2}]|$ is small in some ball around the origin. In general, analysis of $|\mathcal{F}[\mathbb{1}_{G_1} - \mathbb{1}_{G_2}]|$ is extremely challenging, but we will choose the sets in such a way that this d -dimensional integral

evaluates to a product of one-dimensional integrals (which are more amenable). A similar construction was needed to obtain a lower bound for deconvolution in manifold estimation under Hausdorff loss in [11].

Section 6.3 is devoted to the study of the one dimensional case, where we extend the results proven in [9]. The one-dimensional case reduces to estimating the end-point of a univariate density. This problem has been extensively studied in the noiseless case [8, 100] and more recently as an inverse problem [9, 10]. In [9], it is assumed that the density of the (one-dimensional) X_i 's is *exactly* equal to a polynomial in a neighborhood of the endpoint of the support. We extend their results to the case when the distribution function is only bounded by two polynomials whose degrees may differ, in the vicinity of the endpoint.

In Section 6.4, we use these one dimensional results in order to define our estimator of the support G of the X_i 's and to bound its risk on a certain subclass of \mathcal{K}_d . We show that our estimator nearly attains the minimax rate on that class, up to logarithmic factors.

Finally, Section 6.5 is devoted to some proofs.

6.2 Estimation when $d = 1$

When $d = 1$, the convex sets take the form $G = [a, b]$ and we assume that $a + \delta \leq b$ for some fixed $\delta > 0$ (i.e., we assume that a and b are uniformly separated). We prove the following theorem:

Theorem 16. *Let $\mathcal{C} = \{G = [a, b] \subseteq [-1, 1] : a + \delta \leq b\}$. Then*

$$\inf_{\hat{G}} \sup_{G \in \mathcal{C}} d_H(G, \hat{G}) \asymp \frac{1}{\sqrt{n}},$$

for all estimators \hat{G} based on Y_1, \dots, Y_n .

Remark 12. *Note that when \hat{G} is based on the direct observations X_1, \dots, X_n ,*

$$\inf_{\hat{G}} \sup_{G \in \mathcal{C}} d_H(G, \hat{G}) \asymp \frac{1}{n}.$$

Proof. The sample mean and variance are unbiased estimators of their respective population

counterparts:

$$\mathbb{E}\bar{Y}_n = (a + b)/2, \quad \mathbb{E}S_n^2 = (b - a)^2/12 + \sigma^2,$$

where

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i, \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2.$$

This suggests the MME

$$\hat{a}_n = \bar{Y}_n - \sqrt{3(S_n^2 - \sigma^2)} \mathbb{1}\{S_n \geq \sigma\}, \quad \hat{b}_n = \bar{Y}_n + \sqrt{3(S_n^2 - \sigma^2)} \mathbb{1}\{S_n \geq \sigma\}$$

and the set estimator $\hat{G}_n = [\hat{a}_n, \hat{b}_n]$. Indeed, it is not hard to show that $\mathbb{E}|\hat{a}_n - a| = O(1/\sqrt{n})$ and $\mathbb{E}|\hat{b}_n - b| = O(1/\sqrt{n})$. In fact, one cannot estimate G better than this, in a minimax sense. For the lower bound, we use Le Cam's two point method. To this end, let $G_1 = [0, \delta + \gamma]$ and $G_2 = [0, \delta]$. Note that $d_H(G_1, G_2) = \gamma$ and furthermore if $\chi^2(P_{G_1} \parallel P_{G_2}) = O(\gamma^2)$, then choosing $\gamma \asymp \frac{1}{\sqrt{n}}$ finishes the proof since

$$\chi^2(P_{G_1}^{\otimes n} \parallel P_{G_2}^{\otimes n}) = (1 + \chi^2(P_{G_1} \parallel P_{G_2}))^n - 1.$$

and hence

$$\int \min\{dP_{G_1}^{\otimes n}, dP_{G_2}^{\otimes n}\} \geq \frac{1}{2} \exp\{-(1 + \chi^2(P_{G_1} \parallel P_{G_2}))^n + 1\} \geq c > 0,$$

for some universal positive constant c . We now show that $\chi^2(P_{G_1} \parallel P_{G_2}) = O(\gamma^2)$. Note that

$$\chi^2(P_{G_1} \parallel P_{G_2}) = \int \frac{(f_{G_1}(y) - f_{G_2}(y))^2}{f_{G_2}(y)} dy,$$

where

$$f_{G_1}(y) = \frac{1}{\delta + \gamma} \int_0^{\delta + \gamma} \phi_\sigma(y - x) dx, \quad f_{G_2}(y) = \frac{1}{\delta} \int_0^\delta \phi_\sigma(y - x) dx,$$

and ϕ_σ is the density of the normal errors ε . We can write

$$f_{G_1}(y) - f_{G_2}(y) = \frac{1}{\delta + \gamma} \int_\delta^{\delta + \gamma} \phi_\sigma(y - x) dx - \frac{\gamma}{\delta + \gamma} f_{G_2}(y). \quad (6.2)$$

Using the inequality $(a + a)^2 \leq 2a^2 + 2b^2$ and (6.2), we have that

$$\begin{aligned} \int \frac{(f_{G_1}(y) - f_{G_2}(y))^2}{f_{G_2}(y)} dy &\leq \frac{2\gamma^2}{(\delta + \gamma)^2} + \frac{2}{(\delta + \gamma)^2} \int \frac{(\int_{\delta}^{\delta+\gamma} \phi_{\sigma}(y-x) dx)^2}{f_{G_2}(y)} \\ &\leq \frac{2\gamma^2}{(\delta + \gamma)^2} (1 + \int \frac{(\sup_{\delta \leq x \leq \delta+\gamma} \phi_{\sigma}(y-x) dx)^2}{f_{G_2}(y)}). \end{aligned} \quad (6.3)$$

Finally, observe that $\sup_{\delta \leq x \leq \delta+\gamma} \phi_{\sigma}(y-x) dx \propto \exp\{-y^2/(2\sigma^2) + O(|y|)\}$ and $f_{G_2}(y) \propto \exp\{-y^2/(2\sigma^2) + O(|y|)\}$. Thus,

$$\frac{(\sup_{\delta \leq x \leq \delta+\gamma} \phi_{\sigma}(y-x) dx)^2}{f_{G_2}(y)} \propto \exp\{-y^2/(2\sigma^2) + O(|y|)\}$$

and hence the integral in (6.3) is bounded by a constant. \square

Thus, even with error-in variables, the rates are still parametric (c.f., order $1/n$ rates without measurement error).

6.3 Dominating bias in endpoint estimation

Let $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. centered Gaussian random variables. Then, the maximum $\max_{i=1, \dots, n} \varepsilon_i$ concentrates around $\sqrt{2\sigma^2 \ln n}$, where $\sigma^2 = \mathbb{E}[\varepsilon_1^2]$. Our first result shows the same remains true if one adds i.i.d. nonpositive random variables to the ε_i 's, as long as their cumulative distribution function increases polynomially near zero. As a byproduct, one can estimate the endpoint of a distribution with polynomial decay near its boundary by subtracting a deterministic bias from the maximum of the observations. In the sequel, denote by $b_n = \sqrt{2\sigma^2 \ln n}$.

Theorem 17. *Let X be a random variable with cumulative distribution function F and ε be a centered Gaussian random variable with variance $\sigma^2 > 0$, independent of X . Let $Y = X + \varepsilon$ and consider a sequence Y_1, Y_2, \dots of independent copies of Y and define $M_n = \max\{Y_1, \dots, Y_n\}$, for all $n \geq 1$. Assume that there exist real numbers $\theta_F \in \mathbb{R}$, $\alpha \geq \beta \geq 0$, $r > 0$ and $L > 0$ such that the following is true:*

$$L^{-1}t^{\alpha} \leq 1 - F(\theta_F - t) \leq Lt^{\beta}, \forall t \in [0, r].$$

Then, there exist $n_0 \geq 1$ and $c_0, c_1, c_2 > 0$ that depend on α, β, L and r only, such that for all $n \geq n_0$ and $t > 0$,

$$\mathbb{P} \left[|M_n - b_n - \theta_F| > \frac{t + c_0 \ln \ln n}{b_n} \right] \leq c_1 e^{-\frac{t}{2\sigma^2}} + e^{-c_2 n}.$$

The expressions of n_0 and of the constants c_1 and c_2 can be found in the proof of the theorem.

When α and β are equal and known, it is possible to account for the deterministic bias at a higher order and get a more accurate estimate of θ_F .

Theorem 18. *Let assumptions of Theorem 17 hold with $\alpha = \beta$. Set $\tilde{b}_n = \sqrt{2\sigma^2 \ln n} \left(1 - \frac{(\alpha + 1) \ln \ln n}{4 \ln n} \right)$. Then, there exist $n_0 \geq 1$ and $c_1, c_2 > 0$ that depend on α, L and r only, such that for all $n \geq n_0$ and $t > 0$,*

$$\mathbb{P} \left[|M_n - \tilde{b}_n - \theta_F| > \frac{t}{\tilde{b}_n} \right] \leq c_1 e^{-\frac{t}{2\sigma^2}} + e^{-c_2 n}.$$

In Theorem 17, θ_F is the endpoint of the distribution of the X_i 's. When θ_F is unknown, it can be estimated using $\hat{\theta}_n := M_n - b_n$ (or $\tilde{\theta}_n := M_n - \tilde{b}_n$ if $\alpha = \beta$ is known). Theorems 17 and 18 show that $\hat{\theta}_n$ and $\tilde{\theta}_n$ are consistent estimators of θ_F , but that they concentrate very slowly around θ_F , at a polylogarithmic rate. We actually show that this rate is optimal (up to a sublogarithmic factor in the case of $\hat{\theta}_n$) in a minimax sense.

For every collection of parameters $\alpha \geq \beta \geq 0, r > 0$ and $L > 0$, let $\mathcal{F}(\alpha, \beta, r, L)$ the class of all cumulative distribution functions F satisfying $L^{-1}t^\alpha \leq 1 - F(\theta_F - t) \leq Lt^\beta, \forall t \in [0, r]$.

The following result is a simple consequence of Theorem 17.

Corollary 2. *For all $\alpha \geq \beta \geq 0, r > 0$ and $L > 0$,*

$$\inf_{\hat{T}_n} \sup_{F \in \mathcal{F}(\alpha, \beta, r, L)} \mathbb{E} \left[|\hat{T}_n - \theta_F| \right] \lesssim \begin{cases} \frac{\ln \ln n}{\sqrt{\ln n}} & \text{if } \alpha > \beta, \\ \frac{1}{\sqrt{\ln n}} & \text{if } \alpha = \beta, \end{cases}$$

where the infimum is taken over all estimators \hat{T}_n . All the constants depend only on the parameters α, β, r, L and σ^2 .

Theorem 2 in [9] suggests that the upper bound in Corollary 2 is optimal, up to a sublogarithmic factor. However, note that Theorem 2 in [101] only deals with a modified version of the model and hence does not show a lower bound that matches their upper bound.

As a conclusion, these results suggest that in the presence of Gaussian errors, the endpoint θ_F of the distribution of the contaminated data can only be estimated at a polylogarithmic rate, in a minimax sense. In the next section, we prove a lower bound for the multivariate case, whose rate is polylogarithmic in the sample size.

6.4 Application to convex support estimation from noisy data

6.4.1 Definition of the estimator

In this section, we apply Theorem 17 to the problem of estimating a convex body from noisy observations of independent uniform random points. Let G be a convex body in \mathbb{R}^d and let X be uniformly distributed in G . Let ε be a d -dimensional centered Gaussian random variable with covariance matrix $\sigma^2 I$, where σ^2 is a known positive number and I is the $d \times d$ identity matrix. Let $Y = X + \varepsilon$ and assume that a sample Y_1, \dots, Y_n of n independent copies of Y is available to estimate G .

Our estimation scheme consists in reducing the d -dimensional estimation problem to a 1-dimensional one, based on the following observation. Let $u \in \mathcal{S}^{d-1}$. Then, $\langle u, Y \rangle = \langle u, X \rangle + \langle u, \varepsilon \rangle$ and:

- $\langle u, \varepsilon \rangle$ is a centered Gaussian random variable with variance σ^2 ,
- $h_G(u)$ is the endpoint of the distribution of $\langle u, X \rangle$.

In the sequel, we denote by F_u the cumulative distribution function of $\langle u, X \rangle$.

Consider the following assumption:

Assumption 3. $B(a, r) \subseteq G \subseteq B(0, R)$, for some $a \in \mathbb{R}^d$.

Then, we have the following lemma, which allows us to use the one dimensional results of the previous section.

Lemma 25. *Let G satisfy Assumption 3. Then, for all $u \in \mathcal{S}^{d-1}$, $\theta_{F_u} = h_G(u)$ and $F_u \in \mathcal{F}(d, 1, r, L)$, where $L = (2R)^{d-1} r^d \beta_d \max \left(1, \frac{d}{r^{d-1} \beta_{d-1}} \right)$.*

We are now in a position to define an estimator of G . For $u \in \mathbb{R}^d$, let $\hat{h}(u)$ be the estimator of $h_G(u)$ defined as $\hat{h}(u) = \max_{i=1, \dots, n} \langle u, Y_i \rangle - b_n$, where we recall that $b_n = \sqrt{2\sigma^2 \ln n}$.

Let M be a positive integer and U_1, \dots, U_M be independent uniform random vectors on the sphere \mathcal{S}^{d-1} and define

$$\hat{G}_M = \{x \in \mathbb{R}^d : \langle U_j, x \rangle \leq \hat{h}(U_j), \forall j = 1, \dots, M\}. \quad (6.4)$$

We also define a truncated version of \hat{G}_M . Let $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n Y_i$. Define

$$\tilde{G}_M = \begin{cases} \hat{G}_M \cap B(\hat{\mu}, \ln n) & \text{if } \hat{G}_M \neq \emptyset \\ \{\hat{\mu}\} & \text{otherwise.} \end{cases} \quad (6.5)$$

First, we give a deviation inequality for the estimator \hat{G}_M . Then, as a corollary, we prove that for some choice of M (independent of G), the truncated estimator \tilde{G}_M has risk of order $(\ln n)^{-1/2}$.

Theorem 19. *Let $n > 3$, $b_n = \sqrt{2\sigma^2 \ln n}$ and M be a positive integer with $(\ln M)/b_n \leq \min(r/(4\sigma^2), 1/2)$. Then, there exist positive constants c_0, c_1, c_2 and c_3 such that the following holds. For all convex bodies G that satisfy Assumption 3, for all positive x with $x \leq \frac{rb_n}{4\sigma^2} - \ln M$,*

$$d_H(\hat{G}_M, G) \leq c_0 \frac{x + \ln M}{b_n}$$

with probability at least $1 - c_1 e^{-x} - M e^{-c_2 n} - (6b_n)^d e^{-c_3 M (\ln M)^{d-1} b_n^{-(d-1)}}$.

This yields a uniform upper bound on the risk of \tilde{G}_M , which we derive for some special choice of M . Denote by $\mathcal{K}_{r,R}$ the collection of all convex bodies satisfying Assumption 3.

Corollary 3. *Let $A = 2d(d+1)8^{(d-1)/2}$ and $M = \lfloor Ab_n^{d-1}(\ln b_n)^{-(d-2)} \rfloor$. Then, the truncated estimator \tilde{G}_M satisfies*

$$\sup_{G \in \mathcal{K}_{r,R}} \mathbb{E}_G[d_H(\tilde{G}_M, G)] = O\left(\frac{\ln \ln n}{\sqrt{\ln n}}\right).$$

Remark 13. *Suppose that for all $x \in \partial G$, there exist $a, b \in \mathbb{R}^d$ such that $B(a, r) \subseteq G \subseteq B(b, R)$, $x \in B(a, r)$ and $x \in \partial B(b, R)$. In particular, this means that the complement of G has reach at least r , i.e., one can roll a Euclidean ball of radius r inside G along its boundary (see, e.g., [102, Definition 11]). In addition, G can roll freely inside a Euclidean ball of radius R , along its boundary. This ensures that for all $u \in \mathbb{S}^{d-1}$, the random variable $\langle u, X \rangle - h_G(u)$ satisfies the assumption of Theorem 18 with $\alpha = (d+1)/2$ and some $L > 0$ that depends on r and R only.*

Hence, we are in the case where $\alpha = \beta$ in Theorem 18, which shows that the rate of estimation of the support function of G at a single unit vector can be improved by a sublogarithmic factor. However, a close look at the proof of Theorem 19 suggests that a sublogarithmic factor is still unavoidable in our proof technique, because of the union bound on a covering of the unit sphere.

Remark 14. *Theorem 19 can be easily extended to cases where the X_i 's are not uniformly distributed on G . What matters to the proof is that uniformly over unit vectors u , the cumulative distribution function F_u of $\langle u, X \rangle - h_G(u)$ increases polynomially near 0. Examples of such distributions are given in [90].*

6.4.2 Lower bound for the minimax risk

Theorem 20. *For each τ in $(0, 1)$, there are choices of r and R and positive constants c and C depending only on $d > 1$, σ , τ , r , and R such that*

$$\inf_{\hat{G}_n} \sup_{G \in \mathcal{K}_{r,R}} \mathbb{P}_G[d_H(G, \hat{G}_n) > c(\ln n)^{-2/\tau}] \geq C,$$

and

$$\inf_{\hat{G}_n} \sup_{G \in \mathcal{K}_{r,R}} \mathbb{E}_G[d_H(G, \hat{G}_n)] \geq C(\ln n)^{-2/\tau},$$

where the infimum runs over all estimators \widehat{G}_n of G based on Y_1, \dots, Y_n .

Proof of Theorem 20. In the following, we assume that c and C are generic positive constants, depending only on d, σ, τ, r , and R .

Let $\delta > 0$ and m be a positive integer. Let ψ be chosen as in Lemma 34 and $\gamma_m = (4/3)\delta^{-1}\pi m$. Replacing ψ by $x \mapsto 2\delta\psi(x/(2\delta))$, we can assume that ψ is supported in the interval $[-\delta, \delta]$ and $\inf_{|x| \leq \delta(3/4)} \psi(x) > 0$. Note that this transformation does not affect the bound on its derivatives (6.42) and hence the decay of its Fourier transform.

Define $h_m(x) = \psi(x) \sin(\gamma_m x)$, $H_m(x_1, \dots, x_{d-1}) = \prod_{k=1}^{d-1} h_m(x_k)$, and for $L > 0$ and $\omega \in \{-1, +1\}$, let

$$b_\omega(x_1, \dots, x_{d-1}) = \sum_{k=1}^{d-1} g(x_k) + \omega(L/\gamma_m^2) H_m(x_1, \dots, x_{d-1}),$$

where g satisfies:

$$\max_{x \in [-\delta, \delta]} g''(x) < 0, \quad \text{and} \tag{6.6}$$

$$|\mathcal{F}[g](t)| \leq C e^{-c|t|^\tau}, \quad \text{for some positive constants } c \text{ and } C \tag{6.7}$$

For concreteness, one can take an appropriately scaled Cauchy density, $g(x) \propto \frac{1}{1+x^2/\delta_0^2}$, which is strictly concave in the region where $|x| < \delta_0/\sqrt{3}$ and satisfies (6.6) with $\delta_0 > \sqrt{3}\delta$ and (6.7) with $\tau = 1$.

By (6.6) and Lemma 35, we ensure that the Hessian of b_ω , i.e., $\nabla^2 b_\omega$, is negative-semidefinite and so that the sets

$$G_\omega = \{(x_1, \dots, x_d)' \in [-\delta, \delta]^d : 0 \leq x_d \leq b_\omega(x_1, \dots, x_{d-1})\}$$

are convex. Since the G_ω have nonempty interior and are bounded, there are choices of r and R such that $G_\omega \in \mathcal{K}_{r,R}$.

Note that h_m is an odd function about the origin. Thus $\int_{[-\delta, \delta]^{d-1}} H_m(x) dx = 0$ because we are integrating an odd function about the origin. Therefore, $|G_\omega| = (d-1) \int_{[-\delta, \delta]} g(x) dx$.

Also, note that

$$\begin{aligned}
d_{\Delta}(G_{+1}, G_{-1}) &= \int_{[-\delta, \delta]^{d-1}} |b_{+1}(x) - b_{-1}(x)| dx \\
&= \frac{2L}{\gamma_m^2} \int_{[-\delta, \delta]^{d-1}} |H_m(x)| dx \\
&= \frac{2L}{\gamma_m^2} \prod_{k=1}^{d-1} \int_{[-\delta, \delta]} |\sin(\gamma_m x_k) \psi(x_k)| dx_k.
\end{aligned}$$

The factor $\prod_{k=1}^{d-1} \int_{[-\delta, \delta]} |\sin(\gamma_m x_k) \psi(x_k)| dx_k$ in the above expression can be lower bounded by a constant, independent of m . In fact,

$$\begin{aligned}
\int_{[-\delta, \delta]} |\sin(\gamma_m x_k) \psi(x_k)| dx_k &\geq \int_{|x_k| \leq \delta(3/4)} |\sin(\gamma_m x_k) \psi(x_k)| dx_k \\
&\geq 3\delta/4 \inf_{|x| \leq \delta(3/4)} |\psi(x)| \int_{|x_k| \leq 1} |\sin(\pi m x_k)| dx_k \\
&= 3\delta/\pi \inf_{|x| \leq \delta(3/4)} |\psi(x)| \\
&> 0.
\end{aligned}$$

Here, we used the fact that

$$\begin{aligned}
\int_{[-1, 1]} |\sin(\pi m x)| dx &= 4m \int_{[0, 1/(2m)]} |\sin(\pi m x)| dx \\
&= (4/\pi) \int_{[0, \pi/2]} \sin(x) dx \\
&= 4/\pi,
\end{aligned}$$

for any non-zero integer m . Thus, there exists a constant $C_1 > 0$, independent of m , such that

$$d_{\Delta}(G_{+1}, G_{-1}) \geq \frac{C_1}{m^2}. \quad (6.8)$$

For $\omega = \pm 1$, define $f_\omega = \mathbb{1}_{G_\omega}/|G_\omega|$. Note that for all $y > 0$,

$$\begin{aligned}
\text{TV}(\mathbb{P}_{G_{+1}}, \mathbb{P}_{G_{-1}}) &= \int_{\mathbb{R}^d} |(f_{+1} - f_{-1}) * \phi_\sigma(x)| dx \\
&= \int_{\|x\| > y} |(f_{+1} - f_{-1}) * \phi_\sigma(x)| dx + \int_{\|x\| \leq y} |(f_{+1} - f_{-1}) * \phi_\sigma(x)| dx \\
&\leq 2 \int_{\|x\| > y} \sup_{z \in [-\delta, \delta]^d} \phi_\sigma(x - z) dx + \\
&\quad \sqrt{|B_d(0, y)|} \sqrt{\int_{\mathbb{R}^d} |\mathcal{F}[f_+ - f_-](t) \mathcal{F}[\phi_\sigma](t)|^2 dt} \\
&\leq C_2 e^{-c_2 y^2} + C_2 y^{d/2} \sqrt{\int_{\mathbb{R}^d} |\mathcal{F}[f_+ - f_-](t) \mathcal{F}[\phi_\sigma](t)|^2 dt},
\end{aligned}$$

for some positive constants c_2 and C_2 that depend only on δ , σ , and d . Set $y \propto \sqrt{\log \frac{1}{\int_{\mathbb{R}^d} |\mathcal{F}[f_+ - f_-](t) \mathcal{F}[\phi_\sigma](t)|^2 dt}}$ so that $\text{TV}(\mathbb{P}_{G_{+1}}, \mathbb{P}_{G_{-1}})$ can be bounded by a fixed power of $\int_{\mathbb{R}^d} |\mathcal{F}[f_+ - f_-](t) \mathcal{F}[\phi_\sigma](t)|^2 dt$.

Split $\int_{\mathbb{R}^d} |\mathcal{F}[f_{+1} - f_{-1}](t) \mathcal{F}[\phi_\sigma](t)|^2 dt$ into two integrals with domains of integration $\|t\|_\infty \leq am^\tau$ and $\|t\|_\infty > am^\tau$. Using the fact that $\mathcal{F}[\phi_\sigma](t) = \sigma^d e^{-\sigma^2 \|t\|_2^2/2}$, we have

$$\int_{\|t\|_\infty > am^\tau} |\mathcal{F}[f_{+1} - f_{-1}](t) \mathcal{F}[\phi_\sigma](t)|^2 dt \leq C_3 e^{-c_3 m^{2\tau}}.$$

By Lemma 32, we have

$$|\mathcal{F}[f_{+1} - f_{-1}](t)| \leq C e^{-cm^\tau},$$

whenever $\|t\|_\infty \leq am^\tau$. Thus

$$\begin{aligned}
&\int_{\|t\|_2 \leq am^\tau} |\mathcal{F}[f_{+1} - f_{-1}](t) \mathcal{F}[\phi_\sigma](t)|^2 dt \\
&\leq C e^{-cm^\tau} \int_{\mathbb{R}^d} |\mathcal{F}[\phi_\sigma](t)|^2 dt.
\end{aligned}$$

This shows that

$$\text{TV}(\mathbb{P}_{G_{+1}}, \mathbb{P}_{G_{-1}}) \leq C_4 e^{-c_4 m^\tau},$$

for some positive constants c_4 and C_4 that depend only on d , σ , τ , r , and R .

The lower bound is a simple two point statistical hypothesis test. By Lemma 31,

$$\inf_{\widehat{G}_n} \sup_{G \in \mathcal{K}_{r,R}} \mathbb{P}_G[C_5 \mathbf{d}_H(G, \widehat{G}_n) > c_5(\ln n)^{-2/\tau}] \geq$$

$$\inf_{\widehat{G}_n} \sup_{G \in \mathcal{K}_{r,R}} \mathbb{P}_G[\mathbf{d}_\Delta(G, \widehat{G}_n) > c_6(\ln n)^{-2/\tau}].$$

In summary, we have shown that $\mathbf{d}_\Delta(G_{+1}, G_{-1}) \geq \frac{C_1}{m^2}$ and $\text{TV}(\mathbb{P}_{G_{+1}}, \mathbb{P}_{G_{-1}}) \leq C_4 e^{-c_4 m^\tau}$, where the constants depend only on d, σ, τ, r , and R . Choosing $m \asymp (\ln n)^{1/\tau}$ and applying Theorem 2.2(i) in [73] finishes the proof of the lower bound on the minimax probability. To get the second conclusion of the theorem, apply Markov's inequality.

□

6.5 Proofs

6.5.1 Proof of Theorem 17

Denote by G the cumulative distribution function of $Y_1 - \theta_F$. We use the following lemma, which we prove in Section 6.5.5.

Lemma 26. *There exist two positive constants c and C that depend only on r, L and α , such that for all $x \geq \sigma^2/r$,*

$$\frac{c e^{-\frac{x^2}{2\sigma^2}}}{x^{\alpha+1}} \leq 1 - G(x) \leq \frac{C e^{-\frac{x^2}{2\sigma^2}}}{x^{\beta+1}}.$$

Let x be a positive number and n be a positive integer. Write that

$$\mathbb{P}[|M_n - \theta_F - b_n| > x] = 1 - G(b_n + x)^n + G(b_n - x)^n. \quad (6.9)$$

Let us first bound from below $G(b_n + x)^n$. Assume that n is sufficiently large so that

$b_n \geq r/\sigma^2$. By Lemma 26,

$$\begin{aligned} G(b_n + x) &\geq 1 - \frac{Ce^{-\frac{(b_n+x)^2}{2\sigma^2}}}{(b_n + x)^{\beta+1}} \geq 1 - \frac{Ce^{-\frac{(b_n+x)^2}{2\sigma^2}}}{b_n^{\beta+1}} \\ &= 1 - C \exp\left(-\frac{x^2}{2\sigma^2} - \frac{xb_n}{\sigma^2} - \frac{b_n^2}{2\sigma^2} - (\beta+1) \ln b_n\right) \end{aligned} \quad (6.10)$$

$$\begin{aligned} &\geq 1 - C \exp\left(-\frac{xb_n}{\sigma^2} - \frac{b_n^2}{2\sigma^2}\right) \\ &= 1 - \frac{C}{n} \exp\left(-\frac{xb_n}{\sigma^2}\right), \end{aligned} \quad (6.11)$$

as long as n is large enough so $\ln b_n \geq 0$.

Note that for all $u \in [0, 1/2]$, $1 - u \geq e^{-2(\ln 2)u} \geq 1 - 2(\ln 2)u$. Hence, if n is large enough, (6.11) implies

$$G(b_n + x)^n \geq 1 - 2(\ln 2)Ce^{-\frac{xb_n}{2\sigma^2}}. \quad (6.12)$$

Let us now bound from above $G(b_n - x)^n$. First, if $x \leq b_n - r/\sigma^2$, Lemma 26 yields

$$\begin{aligned} G(b_n - x) &\leq 1 - \frac{ce^{-\frac{(b_n-x)^2}{2\sigma^2}}}{(b_n - x)^{\alpha+1}} \leq 1 - \frac{c}{b_n^{\alpha+1}} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{xb_n}{\sigma^2} - \frac{b_n^2}{2\sigma^2}\right) \\ &\leq 1 - c \exp\left(\frac{b_n x}{2\sigma^2} - \frac{b_n^2}{2\sigma^2} - (\alpha+1) \ln b_n\right) \end{aligned} \quad (6.13)$$

$$= 1 - \frac{ce^{B_1}}{n} \exp\left(\frac{b_n x}{2\sigma^2} - \frac{\alpha+1}{2} \ln \ln n\right), \quad (6.14)$$

where $B_1 = (1/2)(\alpha+1) \ln(2\sigma^2)$. Together with the inequalities $1 - u \leq e^{-u} \leq 1/u, \forall u > 0$, (6.14) implies

$$G(b_n - x)^n \leq c^{-1}e^{-B_1}e^{-\frac{xb_n}{2\sigma^2} + \frac{\alpha+1}{2} \ln \ln n}. \quad (6.15)$$

Now, if $x > b_n - r/\sigma^2$, one can simply bound

$$\begin{aligned} G(b_n - x)^n &\leq G(r/\sigma^2)^n \\ &\leq e^{-c_2 n}, \end{aligned} \quad (6.16)$$

using Lemma 26, with $c_2 = -\ln\left(1 - \frac{c\sigma^{2\alpha+2}e^{-\frac{r^2}{2\sigma^6}}}{r^{\alpha+1}}\right)$. Finally, combining (6.15) and (6.16)

yields

$$G(b_n - x)^n \leq c^{-1} e^{-B_1} e^{-\frac{x b_n}{2\sigma^2} + \frac{\alpha+1}{2} \ln \ln n} + e^{-c_2 n}, \quad (6.17)$$

for all positive number x . Now, plugging (6.12) and (6.17) into (6.9) yields

$$\mathbb{P}[|M_n - \theta_F - b_n| > x] \leq c_1 e^{-\frac{x b_n}{2\sigma^2} + \frac{\alpha+1}{2} \ln \ln n} + e^{-c_2 n}, \quad (6.18)$$

where $c_1 = 2(\ln 2)C + c^{-1}e^{-B_1}$. Taking x of the form $\frac{t + c_0 \ln \ln n}{b_n}$ for $t \geq 0$ and $c_0 = (\alpha+1)\sigma^2$ yields Theorem 17. \square

6.5.2 Proof of Theorem 18

The proof of Theorem 18 follows the same lines as that of Theorem 17, where b_n is replaced with \tilde{b}_n . The main modification occurs in (6.10) and (6.13), where we note that $\ln n - B \leq \frac{\tilde{b}_n^2}{2\sigma^2} + (\alpha+1) \ln \tilde{b}_n \leq \ln n + B$, for some positive constant B . \square

6.5.3 Proof of Theorem 19

The proof relies on Lemma 7 in [103], which we state here in a simpler form.

Lemma 27. *Let $\delta \in (0, 1/2]$ and \mathcal{N} be a δ -net of \mathcal{S}^{d-1} . Let G be a convex body in \mathbb{R}^d and h_G its support function. Let $a \in \mathbb{R}^d$ and $0 < r \leq R$ such that $B(a, r) \subseteq G \subseteq B(a, R)$. Let $\hat{h} : \mathcal{S}^{d-1} \rightarrow \mathbb{R}$ and $\hat{G} = \{x \in \mathbb{R}^d : \langle u, x \rangle \leq \hat{h}(u), \forall u \in \mathcal{N}\}$. Let $\phi_\sigma = \max_{u \in \mathcal{N}} |\hat{h}(u) - h_G(u)|$. If $\phi_\sigma \leq r/2$, then $d_H(\hat{G}, G) \leq \frac{3\phi_\sigma R}{2r} + 4R\delta$.*

Let G satisfy Assumption 3. Combining Lemma 25 and Theorem 17, we have that for all $u \in \mathbb{S}^{d-1}$, and all $t \geq 0$,

$$\mathbb{P}_G \left[|\hat{h}(u) - h_G(u)| > t \right] \leq c_1 e^{-\frac{b_n t}{2\sigma^2}} + e^{-c_2 n}, \quad (6.19)$$

with c_1 and c_2 as in Theorem 17 with $\alpha = (d+1)/2$. Hence, by a union bound,

$$\mathbb{P}_G \left[\max_{j=1, \dots, M} |\hat{h}(U_j) - h_G(U_j)| > t \right] \leq c_1 M e^{-\frac{b_n t}{2\sigma^2}} + M e^{-c_2 n}. \quad (6.20)$$

Let $t < r/2$. Consider the event \mathcal{A} where U_1, \dots, U_M form a δ -net of \mathcal{S}^{d-1} , where $\delta \in (0, 1/2)$. By Lemma 27, if \mathcal{A} holds and if $|\hat{h}(U_j) - h_G(U_j)| \leq t$ for all $j = 1, \dots, M$, then $d_H(\hat{G}, G) \leq \frac{3tR}{r} + 4R\delta$. Hence, by (6.20) and Lemma 10 in [103],

$$\begin{aligned} \mathbb{P} \left[d_H(\hat{G}, G) > \frac{3tR}{r} + 4R\delta \right] \\ \leq c_1 M e^{-\frac{bn t}{2\sigma^2}} + M e^{-c_2 n} + 6^d \exp \left(-c_3 M \delta^{d-1} + d \ln \left(\frac{1}{\delta} \right) \right), \end{aligned} \quad (6.21)$$

where $c_3 = (2d8^{(d-1)/2})^{-1}$. By taking $\delta = (\ln M)/b_n$, this ends the proof of Theorem 19. \square

6.5.4 Proof of Corollary 3

In the sequel, let $a \in B_d(0, R)$ coming from Assumption 3. Note that since $\tilde{G}_M \subseteq B(\hat{\mu}, \ln n)$ and $G \subseteq B(0, R)$,

$$d_H(\tilde{G}_M, G) \leq |\hat{\mu}_n - a| + \ln n + R \leq |\hat{\mu}_n - \mu| + \ln n + 2R, \quad (6.22)$$

where μ is the centroid of G . Consider the events \mathcal{A} : " $\tilde{G}_M \neq \emptyset$ " and \mathcal{B} : " $|\hat{\mu} - \mu| \leq 5R$ ". Write

$$\mathbb{E}_G[d_H(\tilde{G}_M, G)] = E_1 + E_2 + E_3, \quad (6.23)$$

where $E_1 = \mathbb{E}_G[d_H(\tilde{G}_M, G)\mathbb{1}_{\mathcal{A} \cap \mathcal{B}}]$, $E_2 = \mathbb{E}_G[d_H(\tilde{G}_M, G)\mathbb{1}_{\mathcal{A}^c \cap \mathcal{B}}]$ and $E_3 = \mathbb{E}_G[d_H(\tilde{G}_M, G)\mathbb{1}_{\mathcal{B}^c}]$. In order to bound E_1 , let us state the following lemma, which is a simple application of Fubini's lemma.

Lemma 28. *Let Z be a nonnegative random variable and A a positive number. Then,*

$$\mathbb{E}[Z\mathbb{1}_{Z < A}] \leq \int_0^A \mathbb{P}[Z \geq t] dt.$$

This lemma yields, together with (6.22), with the same notation as in (6.21),

$$\begin{aligned}
E_1 &\leq \int_0^{\ln n + 7R} \mathbb{P}[\mathbf{d}_H(\tilde{G}_M, G) \geq t] \\
&\leq 4R\delta + \int_0^{\ln n + 7R - 4R\delta} \mathbb{P}[\mathbf{d}_H(\tilde{G}_M, G) \geq t + 4R\delta] \\
&= 4R\delta + \frac{3R}{r} \int_0^{r(\ln n)/(3R) + 7r/3 - 4r\delta/3} \mathbb{P}[\mathbf{d}_H(\tilde{G}_M, G) \geq \frac{3Rt}{r} + 4R\delta]. \tag{6.24}
\end{aligned}$$

Now, we split the last integral in (6.24) in two terms: First, the integral between 0 and $r/2$, where we can apply (6.21), and then between $r/2$ and $r(\ln n)/(3R) + 7r/3 - 4r\delta/3$, where we bound the probability term by the value it takes for $t = r/2$. This yields

$$E_1 \leq \frac{C_1 \ln \ln n}{\sqrt{\ln n}}, \tag{6.25}$$

for some positive constant C_1 that depends neither on n nor on G . For E_2 , note that if \mathcal{A} is not satisfied, then $\tilde{G}_M = \{\hat{\mu}\}$ and $\mathbf{d}_H(\tilde{G}_M, G) \leq |\hat{\mu} - \mu| + 2R$, which is bounded from above by $7R$ if \mathcal{B} is satisfied. Hence,

$$\begin{aligned}
E_2 &\leq 7R\mathbb{P}[\hat{G}_M = \emptyset] \\
&\leq 7R\mathbb{P}[a \notin \hat{G}_M] \\
&= 7R\mathbb{P}[\exists j = 1, \dots, M : \hat{h}(U_j) < \langle U_j, a \rangle] \\
&\leq 7RM\mathbb{P}[\hat{h}(U_1) < \langle U_1, a \rangle] \\
&\leq 7RM\mathbb{P}[\hat{h}(U_1) < h_G(U_1) - r/2] \\
&\leq 7RMc_1 e^{-\frac{b_n r/2}{2\sigma^2}} + e^{-c_2 n}
\end{aligned}$$

by (6.19). Hence,

$$E_2 \leq \frac{C_2 \ln \ln n}{\sqrt{\ln n}}, \tag{6.26}$$

where C_2 is a positive constant that depends neither on n nor on G . Now, using (6.22),

$$E_3 \leq \mathbb{E}_G [(|\hat{\mu} - \mu| + \ln n + 2R) \mathbf{1}_{|\hat{\mu} - \mu| > R}]. \tag{6.27}$$

To bound the latter expectation from above, we use the following lemma, which is also an direct application of Fubini's lemma.

Lemma 29. *Let Z be a nonnegative random variable and A a positive number. Then,*

$$\mathbb{E}[Z \mathbf{1}_{Z > A}] \leq A + \int_A^\infty \mathbb{P}[Z \geq t] dt.$$

Hence, (6.27) yields

$$E_3 \leq (\ln n + 3R) \mathbb{P}[|\hat{\mu} - \mu| > 5R] + \int_{5R}^\infty \mathbb{P}[|\hat{\mu} - \mu| \geq t] dt. \quad (6.28)$$

We now use the following lemma.

Lemma 30. *For all $t \geq 5R$,*

$$\mathbb{P}[|\hat{\mu} - \mu| > t] \leq 6^d e^{-9nt^2/200}.$$

Proof. Let \mathcal{N} be a $(1/2)$ -net of the unit sphere. Let $u \in \mathcal{S}^{d-1}$ such that $|\hat{\mu} - \mu| = \langle u, \hat{\mu} - \mu \rangle$. Let $u^* \in \mathcal{N}$ such that $|u^* - u| \leq 1/2$. Then, by Cauchy-Schartz inequality,

$$\begin{aligned} \langle u^*, \mu \rangle &\geq \langle u, \hat{\mu} - \mu \rangle - (1/2)|\hat{\mu} - \mu| \\ &= \frac{1}{2}|\hat{\mu} - \mu|. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{P}[|\hat{\mu} - \mu| > t] &\leq \mathbb{P}[\exists u^* \in \mathcal{N} : \langle u^*, \hat{\mu} - \mu \rangle \geq t/2] \\ &\leq 6^d \max_{u \in \mathcal{N}} \mathbb{P}[\langle u, \hat{\mu} - \mu \rangle \geq t/2] \\ &\leq 6^d \max_{u \in \mathcal{S}^{d-1}} \mathbb{P}[\langle u, \hat{\mu} - \mu \rangle \geq t/2]. \end{aligned} \quad (6.29)$$

Let $u \in \mathcal{S}^{d-1}$. Then, by Markov's inequality, and using the fact that $|X_1 - \mu| \leq 2R$ almost

surely, for all $\lambda > 0$,

$$\begin{aligned}\mathbb{P}[\langle u, \hat{\mu} - \mu \rangle \geq t/2] &\leq \mathbb{E} \left[e^{\frac{\lambda \langle u, Y_1 - \mu \rangle}{n}} \right]^n e^{-\lambda t/2} \\ &\leq \mathbb{E} \left[e^{\frac{\lambda \langle u, X_1 - \mu \rangle}{n}} \right]^n \mathbb{E} \left[e^{\frac{\lambda \langle u, \varepsilon_1 \rangle}{n}} \right]^n e^{-\lambda t/2} \\ &\leq e^{2R\lambda + \lambda^2 \sigma^2 / (2n)} e^{-\lambda t/2}.\end{aligned}$$

Choosing $\lambda = \frac{3nt}{10\sigma^2}$ and plugging in (6.29) yields the desired result. \square

Applying Lemma 30 to (6.28) entails

$$E_3 \leq \frac{C_3 \ln \ln n}{\sqrt{\ln n}}. \quad (6.30)$$

Applying (6.25), (6.26) and (6.30) to (6.23) ends the proof of the corollary. \square

6.5.5 Proofs of the lemmas and corollaries

Proof of Lemma 26: Without loss of generality, let us assume that $\theta_F = 0$. For all $x \in \mathbb{R}$,

$$1 - G(x) = \int_{-\infty}^0 (1 - F(t)) \frac{e^{-\frac{(x-t)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dt. \quad (6.31)$$

Let us split the latter integral into two parts: Denote by I_1 the integral between $-\infty$ and $-r$ and by I_2 the integral between $-r$ and 0, so $1 - G(x) = I_1 + I_2$.

Assume that $x \geq \sigma^2/r$. First, using the assumption about F , one has:

$$\begin{aligned}I_1 &= \int_0^r (1 - F(-t)) \frac{e^{-\frac{(x+t)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dt \\ &\leq \frac{L}{\sqrt{2\pi\sigma^2}} \int_0^r t^\alpha e^{\frac{(x-t)^2}{2\sigma^2}} dt \\ &= \frac{L e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \int_0^r t^\alpha e^{\frac{-xt}{\sigma^2}} e^{\frac{-t^2}{2\sigma^2}} dt \\ &\leq \frac{L \sigma^{2\alpha+2} e^{-\frac{x^2}{2\sigma^2}}}{x^{\alpha+1} \sqrt{2\pi\sigma^2}} \int_0^{rx/\sigma^2} t^\alpha e^{-t} dt \\ &\leq \frac{L \Gamma(\alpha+1) \sigma^{2\alpha+2} e^{-\frac{x^2}{2\sigma^2}}}{x^{\alpha+1} \sqrt{2\pi\sigma^2}},\end{aligned}$$

where Γ is Euler's gamma function. Hence,

$$I_1 \leq \frac{C' e^{-\frac{x^2}{2\sigma^2}}}{x^{\alpha+1}}, \quad (6.32)$$

where C' is the positive constant given by

$$C' = \frac{L\Gamma(\alpha+1)\sigma^{2\alpha+2}}{\sqrt{2\pi\sigma^2}}.$$

On the other hand,

$$\begin{aligned} I_1 &= \int_0^r (1 - F(-t)) \frac{e^{-\frac{(x+t)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dt \\ &\geq \frac{L^{-1}}{\sqrt{2\pi\sigma^2}} \int_0^r t^\alpha e^{\frac{(x-t)^2}{2\sigma^2}} dt \\ &= \frac{L^{-1} e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \int_0^r t^\alpha e^{\frac{-xt}{\sigma^2}} e^{\frac{-t^2}{2\sigma^2}} dt \\ &\geq \frac{L^{-1} \sigma^{2\alpha+2} e^{-\frac{r^2}{2\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}}{x^{\alpha+1} \sqrt{2\pi\sigma^2}} \int_0^{rx/\sigma^2} t^\alpha e^{-t} dt \\ &\geq \frac{L^{-1} e^{-\frac{r^2}{2\sigma^2}} \sigma^{2\alpha+2} e^{-\frac{x^2}{2\sigma^2}}}{x^{\alpha+1} \sqrt{2\pi\sigma^2}} \int_0^1 t^\alpha e^{-t} dt, \end{aligned}$$

since $rx/\sigma^2 \geq 1$. Hence,

$$I_1 \geq \frac{ce^{-\frac{x^2}{2\sigma^2}}}{x^{\alpha+1}}, \quad (6.33)$$

where c is the positive constant given by

$$c = \frac{L^{-1} e^{-\frac{r^2}{2\sigma^2}} \sigma^{2\alpha+2}}{\sqrt{2\pi\sigma^2}} \int_0^1 t^\alpha e^{-t} dt.$$

Now, let us bound the nonnegative integral I_2 from above.

$$\begin{aligned}
I_2 &= \int_r^\infty (1 - F(-t)) \frac{e^{-\frac{(x+t)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dt \\
&\leq \int_r^\infty \frac{e^{-\frac{(x+t)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dt \\
&= \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \int_r^\infty e^{-\frac{xt}{\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} dt \\
&\leq e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{xr}{\sigma^2}} \int_r^\infty \frac{e^{-\frac{t^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dt \\
&= \frac{1}{2} e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{xr}{\sigma^2}}.
\end{aligned}$$

Since for all $t \geq 0$, $e^{-t} t^{\alpha+1} \leq \left(\frac{\alpha+1}{e}\right)^{\alpha+1}$,

$$I_2 \leq \frac{C'' e^{-\frac{x^2}{2\sigma^2}}}{x^{\alpha+1}}, \quad (6.34)$$

with C'' being the positive constant

$$C'' = \frac{\sigma^{2\alpha+2}}{2r^{\alpha+1}} \left(\frac{\alpha+1}{e}\right)^{\alpha+1}.$$

Hence, (6.32), (6.33) and (6.34) yield

$$\frac{c e^{-\frac{x^2}{2\sigma^2}}}{x^{\alpha+1}} \leq 1 - G(x) \leq (C' + C'') \frac{e^{-\frac{x^2}{2\sigma^2}}}{x^{\alpha+1}}, \quad (6.35)$$

for all $x \geq \sigma^2/r$. This proves Lemma 26. \square

Proof of Lemma 25: Let $u \in \mathcal{S}^{d-1}$. For $t \geq 0$, denote by $C_G(u, t) = \{x \in G : \langle u, x \rangle \geq h_G(u) - t\}$. Then, for all $t \geq 0$, $1 - F_u(t) = \frac{|C_G(u, t)|}{|G|}$. Let $x^* \in G$ such that $\langle u, x^* \rangle = h_G(u)$: G has a supporting hyperplane passing through x^* that is orthogonal to u .

By Assumption 3, there is a ball $B = B(a, r)$ included in G . Consider the section B_u of B passing through a , orthogonal to u : $B_u = B \cap (a_u^\perp)$. Denote by **cone** the smallest cone with apex x^* that contains B_u . Then, for all $t \in [0, r]$, $|C_G(u, t)| \geq |C_{\text{cone}}(u, t)| = \left(\frac{r}{t}\right)^{d-1} \frac{\beta_{d-1} t^d}{d}$,

where $\ell = \langle u, x^* - a \rangle$. Since $G \subseteq B(0, R)$ by Assumption 3, $\ell \leq 2R$ and since $B(a, r) \subseteq G$, $|G| \geq r^d \beta_d$, which altogether proves the lower bound of Lemma 25. For the upper bound, note that Assumption 3 implies that G can be included in a hypercube with edge length $2R$ that has one of its $(d - 1)$ -dimensional faces that contains x^* and is orthogonal to u . Hence, $|C_G(u, t)| \leq 2Rt$, for all $t \in [0, 2R]$. This proves the upper bound of Lemma 25.

Lemma 31. *If G and G' are convex sets satisfying Assumption 3, then there exists a constant C that depends only on d and R such that*

$$d_\Delta(G, G') \leq C d_H(G, G').$$

Proof. See Lemma 2 in [104]. □

Lemma 32. *There exists constants $a > 0$, $c > 0$ and $C > 0$, depending only on d , τ , r , and R , such that if $\|t\|_\infty \leq am^\tau$, then*

$$|\mathcal{F}[\mathbb{1}_{G_{+1}} - \mathbb{1}_{G_{-1}}](t)| \leq Ce^{-cm^\tau}.$$

Proof. The ideas we use here are inspired by the proof of Theorem 8 in [11]. Let $t = (t_1, \dots, t_d)'$ belong to the product set

$$[-\gamma_m/2, \gamma_m/2]^{d-1} \times [-am^\tau, am^\tau].$$

Note that

$$\begin{aligned}
& \mathcal{F}[\mathbb{1}_{G_{+1}} - \mathbb{1}_{G_{-1}}](t) \\
&= \int_{[-\delta, \delta]^{d-1}} e^{i(t_1 x_1 + \dots + t_{d-1} x_{d-1})} \frac{e^{ib_{+1}(x_1, \dots, x_{d-1})t_d} - e^{ib_{-1}(x_1, \dots, x_{d-1})t_d}}{it_d} dx \\
&= 2 \int_{[-\delta, \delta]^{d-1}} e^{i(t_1 x_1 + \dots + t_{d-1} x_{d-1})} e^{it_d \sum_{k=1}^{d-1} g(x_k)} \frac{\sin((Lt_d/\gamma_m^2)H(x))}{t_d} dx \\
&= 2 \sum_{j=0}^{\infty} \frac{(Lt_d/\gamma_m^2)^{2j+1} (-1)^j}{t_d (2j+1)!} \prod_{k=1}^{d-1} \int_{\mathbb{R}} e^{it_k x_k} e^{it_d g(x_k)} h^{2j+1}(x_k) dx_k \\
&= 2 \sum_{j=0}^{\infty} \frac{(Lt_d/\gamma_m^2)^{2j+1} (-1)^j}{t_d (2j+1)!} \prod_{k=1}^{d-1} (\mathcal{F}[\sin^{2j+1}(\gamma_m x_k) e^{it_d g(x_k)} \psi^{2j+1}(x_k)])(t_k). \tag{6.36}
\end{aligned}$$

Next, write

$$\begin{aligned}
\sin^{2j+1}(\gamma_m x_k) &= \left(\frac{e^{ix_k \gamma_m} - e^{-ix_k \gamma_m}}{2i} \right)^{2j+1} \\
&= \left(\frac{1}{2i} \right)^{2j+1} \sum_{s=0}^{2j+1} \binom{2j+1}{s} (-1)^s e^{-ix_k w_s},
\end{aligned}$$

where $w_s = \gamma_m(2s - 2j - 1)$.

Using this expression and linearity of the Fourier transform, we can write

$$\begin{aligned}
& (\mathcal{F}[\sin^{2j+1}(\gamma_m x_k) e^{it_d g(x_k)} \psi^{2j+1}(x_k)])(t_k) \\
&= \left(\frac{1}{2i} \right)^{2j+1} \sum_{s=0}^{2j+1} \binom{2j+1}{s} (-1)^s (\mathcal{F}[e^{it_d g(x_k) - ix_k w_s} \psi^{2j+1}(x_k)])(t_k) \\
&= \left(\frac{1}{2i} \right)^{2j+1} \sum_{s=0}^{2j+1} \binom{2j+1}{s} (-1)^s (\mathcal{F}[e^{it_d g(x_k)} \psi^{2j+1}(x_k)])(t_k - w_s),
\end{aligned}$$

and hence by the triangle inequality,

$$\begin{aligned}
& |(\mathcal{F}[\sin^{2j+1}(\gamma_m x_k) e^{it_d g(x_k)} \psi^{2j+1}(x_k)])(t_k)| \\
&\leq \left(\frac{1}{2} \right)^{2j+1} \sum_{s=0}^{2j+1} \binom{2j+1}{s} |\mathcal{F}[e^{it_d g(x_k)} \psi^{2j+1}(x_k)](t_k - w_s)|. \tag{6.37}
\end{aligned}$$

The function $x \mapsto e^{it_d g(x)}$ can be expanded as

$$\sum_{\ell=0}^{\infty} \frac{(it_d g(x))^\ell}{\ell!},$$

and hence

$$|\mathcal{F}[e^{it_d g(x_k)} \psi^{2j+1}(x_k)](t_k - w_s)| \leq \sum_{\ell=0}^{\infty} \frac{|t_d|^\ell}{\ell!} |\mathcal{F}[g^\ell(x_k) \psi^{2j+1}(x_k)](t_k - w_s)|. \quad (6.38)$$

By (6.7), g is chosen so that its Fourier transform has the same decay as the Fourier transform of ψ . We deduce from Lemma 33 that there exists constants $c > 0$ and $B > 0$, independent of j and ℓ , such that

$$|\mathcal{F}[g^\ell(x_k) \psi^{2j+1}(x_k)](t_k - w_s)| \leq B^{\ell+2j+1} e^{-c|t_k - w_s|^\tau}.$$

Applying this inequality to each term in the sum in (6.38) and summing over ℓ , we find that

$$|\mathcal{F}[e^{it_d g(x_k)} \psi^{2j+1}(x_k)](t_k - w_s)| \leq B^{2j+1} e^{B|t_d| - c|t_k - w_s|^\tau}.$$

Since we restricted the t_k ($k = 1, \dots, d-1$) to be in the interval $[-\gamma_m/2, \gamma_m/2]$, it follows that $|t_k - w_s| \geq \gamma_m/2$. Hence if $\|t\|_\infty \leq am^\tau$, then

$$|\mathcal{F}[e^{*it_d g(x_k)} \psi^{2j+1}(x_k)](t_k - w_s)| \leq B^{2j+1} e^{Bam^\tau - c\gamma_m^\tau/2}.$$

Set $a = c\gamma_m^\tau/(4Bm^\tau)$, which is independent of m . Thus there exists a positive constant c_1 such that

$$|\mathcal{F}[e^{it_d g(x_k)} \psi^{2j+1}(x_k)](t_k - w_s)| \leq B^{2j+1} e^{-c_1 m^\tau}. \quad (6.39)$$

Finally, we apply the inequality (6.39) to each term in the sum in (6.37) and use the identity $(\frac{1}{2})^{2j+1} \sum_{s=0}^{2j+1} \binom{2j+1}{s} = 1$ which yields

$$|(\mathcal{F}[\sin^{2j+1}(\gamma_m x_k) e^{it_d g(x_k)} \psi^{2j+1}(x_k)])(t_k)| \leq B^{2j+1} e^{-c_1 m^\tau}. \quad (6.40)$$

Returning to (6.36), we can use (6.40) to arrive at the bound

$$|\mathcal{F}[\mathbb{1}_{G_{+1}} - \mathbb{1}_{G_{-1}}](t)| \leq 2e^{-c_1(d-1)m^\tau} \sum_{j=0}^{\infty} \frac{(L|t_d|B^{d-1}/\gamma_m^2)^{2j+1}}{|t_d|(2j+1)!}.$$

Note that $\sum_{j=0}^{\infty} \frac{(L|t_d|B^{d-1}/\gamma_m^2)^{2j+1}}{|t_d|(2j+1)!}$ is further bounded by

$$LB^{d-1}(1/\gamma_m^2) \sinh(L|t_d|B^{d-1}/\gamma_m^2)$$

since

$$\begin{aligned} \sum_{j=0}^{\infty} \frac{(L|t_d|B^{d-1}/\gamma_m^2)^{2j+1}}{|t_d|(2j+1)!} &= LB^{d-1}(1/\gamma_m^2) \sum_{j=0}^{\infty} \frac{(L|t_d|B^{d-1}/\gamma_m^2)^{2j}}{(2j+1)!} \\ &\leq LB^{d-1}(1/\gamma_m^2) \sum_{j=0}^{\infty} \frac{(L|t_d|B^{d-1}/\gamma_m^2)^{2j}}{(2j)!} \\ &= LB^{d-1}(1/\gamma_m^2) \sinh(L|t_d|B^{d-1}/\gamma_m^2). \end{aligned}$$

The last term is bounded by a constant since $|t_d| \leq am^\tau = O(\gamma_m^2)$. □

Lemma 33. *Let $\{\psi_j\}$ be a sequence of real-valued functions on \mathbb{R} . Suppose there exists positive constants $C > 0$ and $c > 0$ such that*

$$|\mathcal{F}[\psi_j](t)| \leq Ce^{-c|t|^\tau},$$

for all $t \in \mathbb{R}$ and $j \geq 1$, where $\tau \in (0, 1]$. Then for each $k \geq 1$ and all $t \in \mathbb{R}$,

$$|\mathcal{F}[\prod_{1 \leq j \leq k} \psi_j](t)| \leq C^k B^{k-1} e^{-c|t|^\tau/2}, \tag{6.41}$$

where $B = \int_{\mathbb{R}} e^{-c|s|^\tau/2} ds$.

Proof. We will proof the claim using induction. To this end, suppose (6.41) holds. Then, using the fact that the Fourier transform of a product is the convolution of the individual

Fourier transforms, we have

$$\begin{aligned}
|\mathcal{F}[\prod_{1 \leq j \leq k+1} \psi_j](t)| &= |\mathcal{F}[\prod_{1 \leq j \leq k} \psi_j] * \mathcal{F}[\psi_{k+1}](t)| \\
&= \left| \int_{\mathbb{R}} \mathcal{F}[\prod_{1 \leq j \leq k} \psi_j](s) \mathcal{F}[\psi_{k+1}](t-s) ds \right| \\
&\leq \int_{\mathbb{R}} |\mathcal{F}[\prod_{1 \leq j \leq k} \psi_j](s) \mathcal{F}[\psi_{k+1}](t-s)| ds \\
&\leq C^{k+1} B^{k-1} \int_{\mathbb{R}} e^{-c|s|^\tau/2 - c|t-s|^\tau} ds.
\end{aligned}$$

Next, note that the mapping $x \mapsto |x|^\tau$ is Hölder continuous in the sense that

$$||x|^\tau - |y|^\tau| \leq |x - y|^\tau,$$

for all x, y in \mathbb{R} . Using this, we have that

$$\int_{\mathbb{R}} e^{-c|s|^\tau/2 - c|t-s|^\tau} ds \leq e^{-c|t|^\tau/2} \int_{\mathbb{R}} e^{-c|s|^\tau/2} ds = B e^{-c|t|^\tau/2}.$$

Thus we have shown that

$$|\mathcal{F}[\prod_{1 \leq j \leq k+1} \psi_j](t)| \leq C^k B^{k-1} e^{-c|t|^\tau/2}.$$

□

Lemma 34. *Let $a_1 \geq a_2 \geq \dots$ be a positive sequence with $\sum_{j=1}^{\infty} a_j = 1$. There exists a non-negative function ψ defined on \mathbb{R} that is symmetric (i.e., $\psi(-x) = \psi(x)$), infinitely many times differentiable, integrates to one (i.e., $\int_{\mathbb{R}} \psi = 1$), support equal to $(-1/2, 1/2)$, and such that*

$$\sup_{x \in [-1/2, 1/2]} \left| \frac{d^k \psi}{dx^k}(x) \right| \leq \frac{2^k}{a_1 \dots a_k}, \quad k = 1, 2, \dots \quad (6.42)$$

In particular, for $\tau \in (0, 1)$ and $a_j = \frac{1}{a j^{1/\tau}}$, where $a = \sum_{j=1}^{\infty} \frac{1}{j^{1/\tau}}$, the function ψ satisfies

$$|\mathcal{F}[\psi](t)| \leq \exp \left\{ -\frac{1}{e\tau} \left(\frac{|t|}{2a} \right)^\tau \right\}, \quad \forall t \in \mathbb{R}.$$

Furthermore, $\|\psi\|_{\infty} \leq 1$, $\|\psi'\|_{\infty} \leq 2/(1-\tau)$, and $\|\psi'\|_{\infty} \leq 8/(1-\tau)^2$.

Proof. The existence of ψ can be found in Theorem 1.3.5 of [105]. For the second conclusion, note that the identity

$$(-it)^k \mathcal{F}[\psi](t) = \int_{-1/2}^{1/2} e^{itx} \frac{d^k \psi}{dx^k}(x) dx, \quad k = 1, 2, \dots$$

holds. Using this and the upper bound for $\frac{d^k \psi}{dx^k}$, we see that

$$|t|^k |\mathcal{F}[\psi](t)| \leq (2a)^k (k!)^{1/\tau}.$$

Next, use the fact that $k! \leq e^{k \ln k}$ to upper bound $(2a)^k (k!)^{1/\tau}$ by $\exp\{k \ln(2a) + (1/\tau)k \ln k\}$.

We have thus shown that

$$|\mathcal{F}[\psi](t)| \leq \exp\{k \ln(2a) + (1/\tau)k \ln k\} / |t|^k,$$

for $t \neq 0$ and $k = 1, 2, \dots$. Choose $k = \frac{1}{e} \left(\frac{|t|}{2a} \right)^\tau$ so that

$$|\mathcal{F}[\psi](t)| \leq \exp \left\{ -\frac{1}{e^\tau} \left(\frac{|t|}{2a} \right)^\tau \right\}.$$

The estimates on the L_∞ norms of ψ , ψ' , and ψ'' follow from the fact that $a \leq 1/(1-\tau)$. \square

Lemma 35. *If $\max_{x \in [-\delta, \delta]} g''(x) < 0$, there exists $L > 0$, depending only on τ and γ_m , such that the sets G_ω are convex.*

Proof. As discussed in the proof of Theorem 20, the sets G_ω are convex if the Hessian of b_ω is negative-semidefinite. This is equivalent to showing that the largest eigenvalue of $\nabla^2 b_\omega$

is nonpositive. We can bound the maximum eigenvalue of $\nabla^2 b_\omega$ via

$$\begin{aligned}
\lambda_{\max} &= \max_{\|u\|_2=1} u' \nabla^2 b_\omega u \\
&= \max_{\|u\|_2=1} \left[\sum_k g''(x_k) u_k^2 + \sum_{i,j} \omega(L/\gamma_m^2) \frac{\partial^2 H_m}{\partial x_i \partial x_j}(x_1, \dots, x_{d-1}) u_i u_j \right] \\
&\leq \max_{x \in [-\delta, \delta]} g''(x) + (L/\gamma_m^2) \max\{\|h_m\|_\infty^{d-3} \|h'_m\|_\infty^2, \|h_m\|_\infty^{d-2} \|h''_m\|_\infty\} \\
&\leq \max_{x \in [-\delta, \delta]} g''(x) + (L/\gamma_m^2) \max\{\|h'_m\|_\infty^2, \|h''_m\|_\infty\}
\end{aligned}$$

Now, from Lemma 34 we have the estimates $\|\psi\|_\infty \leq 1$, $\|\psi'\|_\infty \leq 2/(1-\tau)$, and $\|\psi''\|_\infty \leq 8/(1-\tau)^2$. Thus,

$$\begin{aligned}
|h'_m(x)| &= |\psi'(x) \sin(\gamma_m x) - \gamma_m \psi(x) \cos(\gamma_m x)| \\
&\leq 2/(1-\tau) + \gamma_m,
\end{aligned}$$

and

$$\begin{aligned}
|h''_m(x)| &= |\psi''(x) \cos(\gamma_m x) - 2\gamma_m \psi'(x) \sin(\gamma_m x) - \gamma_m^2 \psi(x) \cos(\gamma_m x)| \\
&\leq 8/(1-\tau)^2 + 4\gamma_m/(1-\tau) + \gamma_m^2.
\end{aligned}$$

It thus follows that

$$\max\{\|h'_m\|_\infty^2, \|h''_m\|_\infty\} \leq 8/(1-\tau)^2 + 4\gamma_m/(1-\tau) + \gamma_m^2.$$

Next, choose L , depending only on τ and γ_m , such that

$$(L/\gamma_m^2)[8/(1-\tau)^2 + 4\gamma_m/(1-\tau) + \gamma_m^2] \leq -(1/2) \max_{x \in [-\delta, \delta]} g''(x).$$

This means that $\lambda_{\max} \leq (1/2) \max_{x \in [-\delta, \delta]} g''(x) < 0$.

□

Chapter 7

Estimating the number of connected components in a graph via subgraph sampling

7.1 Introduction

Counting the number of features in a graph – ranging from basic local structures like motifs or graphlets (e.g., edges, triangles, wedges, stars, cycles, cliques) to more global features like the number of connected components – is an important task in network analysis. For example, the global clustering coefficient of a graph (i.e. the fraction of closed triangles) is a measure of the tendency for nodes to cluster together and a key quantity used to study cohesion in various networks [106]. To learn these graph properties, applied researchers typically collect data from a random sample of nodes to construct a representation of the true network. We refer to these problems collectively as *statistical inference on sampled networks*, where the goal is to infer properties of the parent network (population) from a subsampled version. Below we mention a few examples that arise in various fields of study.

- Sociology: Social networks of the Hadza hunter-gatherers of Tanzania were studied in [107] by surveying 205 individuals in 17 Hadza camps (from a population of 517). Another study [108] of farmers in Ghana used network data from a survey of 180

households in three villages from a population of 550 households.

- Economics and business: Low sampling ratios have been used in applied economics (such as 30% in [109]), particularly for large scale studies [110, 111]. A good overview of various experiments in applied economics and their corresponding sampling ratios can be found in [112, Appendix F, p. 11]. Word of mouth marketing in consumer referral networks was studied in [113] using 158 respondents from a potential subject pool of 238.
- Genomics: The authors of [114] use protein-protein interaction data and demonstrate that it is possible to arrive at a reliable statistical estimate for the number of interactions (edges) from a sample containing approximately 1500 vertices.
- World Wide Web and Internet: Informed random IP address probing was used in [115] in an attempt to obtain a router-level map of the Internet.

As mentioned earlier, a primary concern of these studies is how well the data represent the true network and how to reconstruct the relevant properties of the parent graphs from samples. These issues and how they are addressed broadly arise from two perspectives:

- The full network is unknown due to the lack of data, which could arise from the underlying experimental design and data collection procedure, e.g., historical or observational data. In this case, one needs to construct statistical estimators (i.e., functions of the sampled graph) to conduct sound inference. These estimators must be designed to account for the fact that the sampled network is only a partial observation of the true network, and thus subject to certain inherent biases and variability.
- The full network is either too large to scan or too expensive to store. In this case, approximation algorithms can overcome such computational or storage issues that would otherwise be unwieldy. For example, for massive social networks, it is generally impossible to enumerate the whole population. Rather than reading the entire graph, query-based algorithms randomly (or deterministically) sample parts of the graph or adaptively explore the graph through a random walk [116]. Some popular instances

of traversal based procedures are snowball sampling [117] and respondent-driven sampling [118]. Indeed, sampling (based on edge and degree queries) is a commonly used primitive to speed up computation, which leads to various sublinear-time algorithms for testing or estimating graph properties such as the average degree [119], triangle and more general subgraph counts [120, 121], expansion properties [122]; we refer the readers to the monograph [123].

Learning properties of graphs from samples has been an important problem in statistical network analysis since the early work of Goodman [15] and Frank [16]. Estimation of various properties such as graph totals [124] and connectivity [16, 125] has been studied in a variety of sample models. However, most of the analysis has been confined to obtaining unbiased estimators for certain classes of graphs and little is known about their optimality. The purpose of this chapter is to initiate a systematic study of statistical inference on sampled networks, with the goal of determining their statistical limits in terms of minimax risks and sample complexity, achieved by computationally efficient procedures.

As a first step towards this end, in this chapter we focus on a representative problem introduced in [16], namely, estimating the number of connected components in a graph from a partial sample of the population network. We study this problem for two reasons. First, it encapsulates many challenging aspects of statistical inference on sampled graphs, and we believe the mathematical framework and machinery developed in this chapter will prove useful for estimating other graph properties as well. Second, the number of connected components is a useful graph property that quantifies the connectivity of a network. In addition, it finds use in data-analytic applications related to determining the number of classes in a population [15]. Another example is the recent work [126], which studies the estimation of the number of documented deaths in the Syrian Civil War from a subgraph induced by a set of vertices obtained from an adaptive sampling process (similar to subgraph sampling). There, the goal is to estimate the number of unique individuals in a population, which roughly corresponds to the number of connected components in a network of duplicate records connected by shared attributes.

Next we discuss the sampling model, which determines how reflective the data is of the population graph and therefore the quality of the estimation procedure. There are many ways to sample from a graph (see [127, 128] for a list of techniques and [129–131] for comprehensive reviews). For simplicity, this chapter focuses on the simplest sampling model, namely, *subgraph sampling*, where we randomly sample a subset of the vertices and observe their induced subgraph; in other words, only the edges between the sampled vertices are revealed. For results on the related neighborhood sampling model we refer to the companion Chapter 8 or [18]. One of the earliest works that adopts the subgraph sampling model is by Frank [16], which is the basis for the theory developed in this chapter. Drawing from previous work on estimating population total using vertex sampling [124], Frank obtained unbiased estimators of the number of connected components and performance guarantees (variance calculations) for graphs whose connected components are either all trees or all cliques. Extensions to more general graphs are briefly discussed, although no unbiased estimators are proposed. This generality is desirable since it is more realistic to assume that the objects in each class (component) are in between being weakly and strongly connected to each other, corresponding to having the level of connectivity between a tree and clique. While the results of Frank are interesting, questions of their generality and optimality remain open and we therefore address these matters in the sequel. Specifically, the main goals of this chapter are as follows:

- Characterize the sample complexity, i.e., the minimal sample size to achieve a given accuracy, as a function of graph parameters.
- Devise computationally efficient estimators that provably achieve the optimal sample complexity bound.

Of particular interest is the *sublinear regime*, where only a vanishing fraction of the vertices are sampled. In this case, it is impossible to reconstruct the entire graph, but it might still be possible to accurately estimate the desired graph property.

The problem of estimating the number of connected components in a large graph has also been studied in the computer science literature, where the goal is to design randomized algorithms with sublinear (in the size of the graph) time complexity. The celebrated work

[132] proposed a randomized algorithm to estimate the number of connected components in a general graph (motivated by computing the weight of the minimum spanning tree) within an additive error of ϵN for graphs with N vertices and average degree d_{avg} , with runtime $O(\frac{d_{\text{avg}}}{\epsilon^2} \log \frac{d_{\text{avg}}}{\epsilon})$. Their method relies on data obtained from a random sample of vertices and then performing a breadth first search on each vertex which ends according to a random stopping criterion. The algorithm requires knowledge of the average degree d_{avg} and must therefore be known or estimated a priori. The runtime was further improved to $O(\epsilon^{-2} \log \frac{1}{\epsilon})$ by modifying the stopping criterion [133]. In these algorithms, the breadth first search may visit many of the edges and explore a larger fraction of the graph at each round. From an applied perspective, such traversal based procedures can be impractical or impossible to implement in many statistical applications due to limitations inherent in the experimental design and it is more realistic to treat the network data as a random sample from a parent graph.

Finally, let us compare, conceptually, the framework in the present chapter with the work on *model-based* network analysis, where networks are modeled as random graphs drawn from specific generative models, such as the stochastic block model [134], graphons [135], or exponential random graph models [136] (cf. the recent survey [129]), and performance analysis of statistical procedures for parameter estimation or clustering are carried out for these models. In contrast, in network sampling we adopt a *design-based* framework [131], where the graph is assumed to be deterministic and the randomness comes from the sampling process.

7.1.1 Organization

The chapter is organized as follows. In Section 8.1.1, we formally define the estimation problem, the subgraph sampling model, and describe what classes of graphs we will be focusing on. To motivate our attention on specific classes of graphs (chordal graphs with maximum degree constraints), we show that in the absence of such structural assumptions, sublinear sample complexity is impossible in the sense that at least a constant fraction of the vertices need to be sampled. Section 7.3 introduces the definition of chordal graphs and states our main results in terms of the minimax risk and sample complexity. In Section 7.4,

after introducing the relevant combinatorial properties of chordal graphs, we define the estimator of the number of connect components and provide its statistical guarantees. We also propose a heuristic for constructing an estimator on non-chordal graphs. In Section 8.3, we develop a general strategy for proving minimax lower bound for estimating graph properties and particularize it to obtain matching lower bounds for the estimator constructed in Section 7.4. Finally, in Section 8.5, we perform a numerical study of the proposed estimators on simulated data for various graphs. Some of the technical proofs are deferred till Appendix 7.7.

7.1.2 Notations

We use standard big- O notations, e.g., for any positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for some absolute constant $C > 0$, $a_n = o(b_n)$ or $a_n \ll b_n$ or if $\lim a_n/b_n = 0$. Furthermore, the subscript in $a_n = O_r(b_n)$ means $a_n \leq C_r b_n$ for some constant C_r depending on the parameter r only. For positive integer k , let $[k] = \{1, \dots, k\}$. Let $\text{Bern}(p)$ denote the Bernoulli distribution with mean p and $\text{Bin}(N, p)$ the binomial distribution with N trials and success probability p .

Next we introduce some graph-theoretic notations that will be used throughout the chapter. Let $G = (V, E)$ be a simple undirected graph. Let $\mathbf{e} = \mathbf{e}(G) = |E(G)|$ denote the number of edges, $\mathbf{v} = \mathbf{v}(G) = |V(G)|$ denote the number of vertices, and $\mathbf{cc} = \mathbf{cc}(G)$ be the number of connected components in G . The neighborhood of a vertex u is denoted by $N_G(u) = \{v \in V(G) : \{u, v\} \in E(G)\}$.

Two graphs G and G' are isomorphic, denoted by $G \simeq G'$, if there exists a bijection between the vertex sets of G and G' that preserves adjacency, i.e., if there exists a bijective function $g : V(G) \rightarrow V(G')$ such that $\{g(u), g(v)\} \in E(G')$ if and only if $\{u, v\} \in E(G)$. The disjoint union of two graphs G and G' , denoted $G + G'$, is the graph whose vertex (resp. edge) set is the disjoint union of the vertex (resp. edge) sets of G and of G' . For brevity, we denote by kG to the disjoint union of k copies of G .

We use the notation K_n , P_n , and C_n to denote the complete graph, path graph, and cycle graph on n vertices, respectively. Let $K_{n,n'}$ denote the complete bipartite graph with nn' edges and $n + n'$ vertices. Let S_n denote the star graph $K_{1,n}$ on $n + 1$ vertices.

We need two types of subgraph counts: Denote by $s(H, G)$ (resp. $n(H, G)$) the number of vertex (resp. edge) induced subgraphs of G that are isomorphic to H .¹ For example, $s(\text{---}\circ\text{---}\circ, \text{---}\circ\text{---}\circ) = 2$ and $n(\text{---}\circ\text{---}\circ, \text{---}\circ\text{---}\circ) = 8$. Let $\omega(G)$ denote the clique number, i.e., the size of the largest clique in G .

7.2 Model

7.2.1 Subgraph sampling model

To fix notations, let $G = (V, E)$ be a simple, undirected graph on N vertices. In the subgraph sampling model, we sample a set of vertices denoted by $S \subset V$, and observe their induced subgraph, denoted by $G[S] = (S, E[S])$, where the edge set is defined as $E[S] = \{\{i, j\} \in S^2 : \{i, j\} \in E\}$. See Fig. 7.1 for an illustration. To simplify notations, we abbreviate the sampled graph $G[S]$ as \tilde{G} .

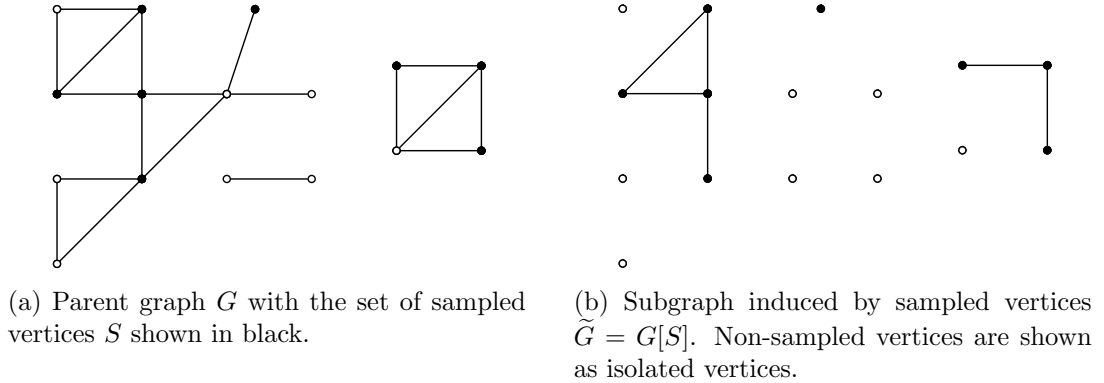


Figure 7.1: Subgraph sampling.

According to how the set S of sampled vertices is generated, there are two variations of the subgraph sampling model [16]:

- *Uniform sampling*: Exactly n vertices are chosen uniformly at random without replacement from the vertex set V . In this case, the probability of observing a subgraph

1. The subgraph counts are directly related to the graph homomorphism numbers [137, Sec 5.2]. Denote by $\text{inj}(H, G)$ the number of injective homomorphisms from H to G and $\text{ind}(H, G)$ the number of injective homomorphisms that also preserve non-adjacency. Then $\text{ind}(H, G) = s(H, G)\text{aut}(H)$ and $\text{inj}(H, G) = n(H, G)\text{aut}(H)$, where $\text{aut}(H)$ denotes the number of automorphisms (i.e. isomorphisms to itself) for H .

isomorphic² to H with $v(H) = n$ is equal to

$$\mathbb{P}[\tilde{G} \simeq H] = \frac{s(H, G)}{\binom{N}{n}}. \quad (7.1)$$

- *Bernoulli sampling*: Each vertex is sampled independently with probability p , where p is called the *sampling ratio*. Thus, the sample size $|S|$ is distributed as $\text{Bin}(N, p)$, and the probability of observing a subgraph isomorphic to H is equal to

$$\mathbb{P}[\tilde{G} \simeq H] = s(H, G)p^{v(H)}(1-p)^{v(G)-v(H)}. \quad (7.2)$$

The relation between these two models is analogous to that between sampling without replacements and sampling with replacements. In the sublinear sampling regime where $n \ll N$, they are nearly equivalent. For technical simplicity, we focus on the Bernoulli sampling model and we refer to $n \triangleq pN$ as the *effective sample size*. Extensions to the uniform sampling model will be discussed in Section 7.4.4.

A number of previous work on subgraph sampling is closely related with the theory of graph limits [138], which is motivated by the so-called property testing problems in graphs [123]. According to [138, Definition 2.11], a graph parameter f is “testable” if for any $\epsilon > 0$, there exists a sample size n such that for any graph G with at least n vertices, there is an estimator $\hat{f} = \hat{f}(\tilde{G})$ such that $\mathbb{P}[|f(G) - \hat{f}| > \epsilon] < \epsilon$. In other words, testable properties can be estimated with sample complexity that is *independent* of the size of the graph. Examples of testable properties include the edge density $e(G)/\binom{v(G)}{2}$ and the density of maximum cuts $\frac{\text{MaxCut}(G)}{v(G)^2}$, where $\text{MaxCut}(G)$ is the size of the maximum edge cut-set in G [139]; however, the number of connected components $\text{cc}(G)$ or its normalized version $\frac{\text{cc}(G)}{v(G)}$ are not testable.³ Instead, our focus is to understand the dependency of sample complexity of estimating $\text{cc}(G)$ on the graph size N as well as other graph parameters. It turns out for

2. Note that it is sufficient to describe the sampled graph up to isomorphism since the property cc we want to estimate is invariant under graph isomorphisms.

3. To see this, recall from [138, Theorem 6.1(b)] an equivalent characterization of f being testable is that for any $\epsilon > 0$, there exists a sample size n such that for any graph G with at least n vertices, $|f(G) - \mathbb{E}f(\tilde{G})| < \epsilon$. This is violated for star graphs $G = S_N$ as $N \rightarrow \infty$

certain classes of graphs, the sample complexity grows *sublinearly* in N , which is the most interesting regime.

7.2.2 Classes of graphs

Before introducing the classes of graphs we consider in this chapter, we note that, unless further structures are assumed about the parent graph, estimating many graph properties, including the number of connected components, has very high sample complexity that scales linearly with the size of the graph. Indeed, there are two main obstacles in estimating the number of connected components in graphs, namely, *high-degree vertices* and *long induced cycles*. If either is allowed to be present, we will show that even if we sample a constant fraction of the vertices, any estimator of $\text{cc}(G)$ has a worst-case additive error that is almost linear in the network size N . Specifically,

- For any sampling ratio p bounded away from 1, as long as the maximum degree is allowed to scale as $\Omega(N)$, even if we restrict the parent graph to be acyclic, the worst-case estimation error for any estimator is $\Omega(N)$.
- For any sampling ratio p bounded away from $1/2$, as long as the length of the induced cycles is allowed to be $\Omega(\log N)$, even if we restrict the parent graph to have maximum degree 2, the worst-case estimation error for any estimator is $\Omega(\frac{N}{\log N})$.

The precise statements follow from the minimax lower bounds in Theorem 34 and Theorem 32. Below we provide an intuitive explanation for each scenario.

For the first claim involving large degree, consider a pair of acyclic graphs G and G' , where G is the star graph on N vertices and G' consisting of N isolated vertices. Note that as long as the center vertex in G is not sampled, the sampling distributions of G and G' are identical. This implies that the total variation between the sampled graph under G and G' is at most p . Since the numbers of connected components in G and G' differ by $N - 1$, this leads to a minimax lower bound for the estimation error of $\Omega(N)$ whenever p is bounded away from one.

The effect of long induced cycles is subtler. The key observation is that a cycle and a path (or a cycle versus two cycles) locally look exactly the same. Indeed, let G (resp. G')

consists of $N/(2r)$ disjoint copies of the smaller graph H (resp. H'), where H is a cycle of length $2r$ and H' consists of two disjoint cycles of length r (see Fig. 7.2). Both G and G' have maximum degree 2 and contain induced cycles of length at most $2r$. The local structure of G and G' is the same (e.g., each connected subgraph with at most $r - 1$ vertices appears exactly N times in each graph) and the sampled versions of H and H' are identically distributed provided at most $r - 1$ vertices are sampled. Thus, we must sample at least r vertices (which occurs with probability at most $e^{-r(1-2p)^2}$) for the distributions to be different. By a union bound, it can be shown that the total variation between the sampled graphs \tilde{G} and \tilde{G}' is $O((N/r)e^{-r(1-2p)^2})$. Thus, whenever the sampling ratio p is bounded away from $1/2$, choosing $r = \Theta(\log N)$ leads to a near-linear lower bound $\Omega(\frac{N}{\log N})$.

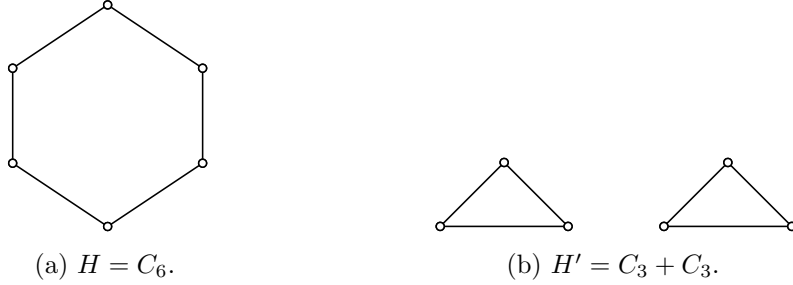


Figure 7.2: Examples of G (resp. G') consisting multiple copies of H (resp. H') with $r = 3$. Both graphs have 6 vertices and 6 edges.

The difficulties caused by high-degree vertices and long induced cycles motivate us to consider classes of graphs defined by two key parameters, namely, the maximum degree d and the length of the longest induced cycles c . The case of $c = 2$ corresponds to forests (acyclic graphs), which have been considered by Frank [16]. The case of $c = 3$ corresponds to *chordal graphs*, i.e., graphs without induced cycle of length four or above, which is the focus of this chapter. It is well-known that various computation tasks that are intractable in the worst case, such as maximal clique and graph coloring, are easy for chordal graphs; it turns out that the chordality structure also aids in both the design and the analysis of computationally efficient estimators which provably attain the optimal sample complexity.

7.3 Main results

This section summarizes our main results in terms of the minimax risk of estimating the number of connected components over various class of graphs. As mentioned before, for ease of exposition, we focus on the Bernoulli sampling model, where each vertex is sampled independently with probability p . Similar conclusions can be obtained for the uniform sampling model upon identifying $p = n/N$, as given in Section 7.4.4.

When p grows from 0 to 1, an increasing fraction of the graph is observed and intuitively the estimation problem becomes easier. Indeed, all forthcoming minimax rates are inversely proportional to powers of p . Of particular interest is whether accurate estimation in the sublinear sampling regime, i.e., $p = o(1)$. The forthcoming theory will give explicit conditions on p for this to hold true.

As mentioned in the previous section, the main class of graphs we study is the so-called *chordal graphs* (see Fig. 7.3 for an example):

Definition 1. A graph G is *chordal* if it does not contain induced cycles of length four or above, i.e., $s(C_k, G) = 0$ for $k \geq 4$.

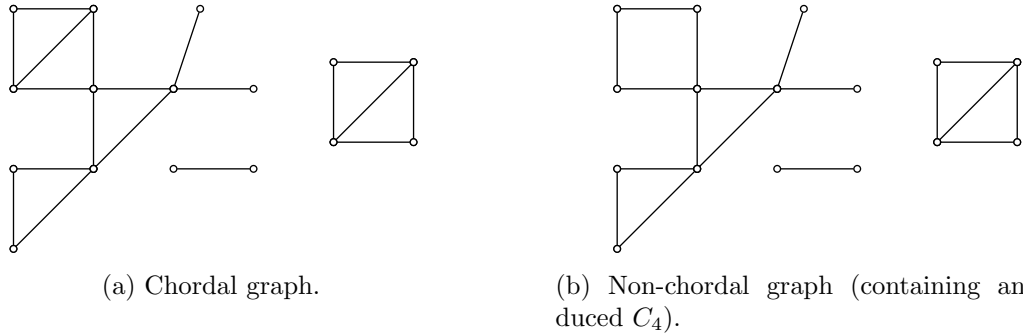


Figure 7.3: Examples of chordal and non-chordal graphs both with three connected components.

We emphasize that chordal graphs are allowed to have arbitrarily long cycles but no induced cycles longer than three. The class of chordal graphs encompasses *forests* and *disjoint union of cliques* as special cases, the two models that were studied in Frank's original paper [16]. In addition to constructing estimators that adapt to larger collections of graphs (for which forests and unions of cliques are special cases), we also provide theoretical

analysis and optimality guarantees – elements that were not considered in past work.

Next, we characterize the rate of the minimax mean-squared error for estimating the number of connected components in a chordal graph, which turns out to depend on the number of vertices, the maximum degree, and the clique number. The upper and lower bounds differ by at most a multiplicative factor depending only on the clique number. To simplify the notation, henceforth we denote $q = 1 - p$.

Theorem 21 (Chordal graphs). *Let $\mathcal{G}(N, d, \omega)$ denote the collection of all chordal graphs on N vertices with maximum degree and clique number at most d and $\omega \geq 2$, respectively. Then*

$$\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{G}(N, d, \omega)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 = \Theta_\omega \left(\left(\frac{N}{p^\omega} \vee \frac{Nd}{p^{\omega-1}} \right) \wedge N^2 \right),$$

where the lower bound holds provided that $p \leq p_0$ for some constant $p_0 < \frac{1}{2}$ that only depends on ω .

Furthermore, if $p \geq 1/2$, then for any ω ,

$$\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{G}(N, d, \omega)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 \leq Nq(d+1). \quad (7.3)$$

Specializing Theorem 21 to $\omega = 2$ yields the minimax rates for estimating the number of trees in forests for small sampling ratio p . The next theorem shows that the result holds verbatim even if p is arbitrarily close to 1, and, consequently, shows minimax rate-optimality of the bound in (7.3).

Theorem 22 (Forests). *Let $\mathcal{F}(N, d) \triangleq \mathcal{G}(N, d, 2)$ denote the collection of all forests on N vertices with maximum degree at most d . Then for all $0 \leq p \leq 1$ and $1 \leq d \leq N$,*

$$\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{F}(N, d)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 \asymp \left(\frac{Nq}{p^2} \vee \frac{Nqd}{p} \right) \wedge N^2. \quad (7.4)$$

The upper bounds in the previous results are achieved by unbiased estimators. As (7.3) shows, they work well even when the clique number ω grow with N , provided we sample more than half of the vertices; however, if the sample ratio p is below $\frac{1}{2}$, especially in the sublinear regime of $p = o(1)$ that we are interested in, the variance is exponentially large. To deal

with large d and ω , we must give up unbiasedness to achieve a good bias-variance tradeoff. Such biased estimators, obtained using the smoothing technique introduced in [140], lead to better performance as quantified in the following theorem. The proofs of these bounds are given in Theorem 27 and Theorem 29.

Theorem 23 (Chordal graphs). *Let $\mathcal{G}(N, d)$ denote the collection of all chordal graphs on N vertices with maximum degree at most d . Then, for any $p < 1/2$,*

$$\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{G}(N, d)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 \lesssim N^2 (N/d^2)^{-\frac{p}{2-3p}}.$$

Finally, for the special case of graphs consisting of disjoint union of cliques, as the following theorem shows, there are enough structures so that we no longer need to impose any condition on the maximal degree. Similar to Theorem 23, the achievable scheme is a biased estimator, significantly improving the unbiased estimator in [15, 16] which has exponentially large variance.

Theorem 24 (Cliques). *Let $\mathcal{C}(N)$ denote the collection of all graphs on N vertices consisting of disjoint unions of cliques. Then, for any $p < 1/2$,*

$$\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{C}(N)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 \leq N^2 (N/4)^{-\frac{p}{2-3p}}.$$

Alternatively, the above results can be summarized in terms of the *sample complexity*, i.e., the minimum sample size that allows an estimator $\widehat{\text{cc}}(G)$ within an additive error of ϵN with probability, say, at least 0.99, uniformly for all graphs in a given class. Here the sample size is understood as the average number of sampled vertices $n = pN$. We have the following characterization:

Table 7.1: Sample complexity for various classes of graphs

Graph	Sample complexity n
Chordal	$\Theta_{\omega} \left(\max \left\{ N^{\frac{\omega-2}{\omega-1}} d^{\frac{1}{\omega-1}} \epsilon^{-\frac{2}{\omega-1}}, N^{\frac{\omega-1}{\omega}} \epsilon^{-\frac{2}{\omega}} \right\} \right)$
Forest	$\Theta \left(\max \left\{ \frac{d}{\epsilon^2}, \frac{\sqrt{N}}{\epsilon} \right\} \right)$
Cliques	$\Theta \left(\frac{N}{\log N} \log \frac{1}{\epsilon} \right), \quad \epsilon \geq N^{-1/2+\Omega(1)} *$

* The lower bound part of this statement follows from [141, Section 3], which shows the optimality of Theorem 24.

A consequence of Theorem 22 is that if the effective sample size n scales as $O(\max(\sqrt{N}, d))$, for the class of forests $\mathcal{F}(N, d)$ the worse-case estimation error for any estimator is $\Omega(N)$, which is within a constant factor to the trivial error bound when no samples are available. Conversely, if $n \gg \max(\sqrt{N}, d)$, which is sublinear in N as long as the maximal degree satisfies $d = o(N)$, then it is possible to achieve a non-trivial estimation error of $o(N)$. More generally for chordal graphs, Theorem 21 implies that if $n = O(\max(N^{\frac{\omega-1}{\omega}}, d^{\frac{1}{\omega-1}} N^{\frac{\omega-2}{\omega-1}}))$, the worse-case estimation error in $\mathcal{G}(N, d, \omega)$ for any estimator is at least $\Omega_{\omega}(N)$,

7.4 Algorithms and performance guarantees

In this section we propose estimators which provably achieve the upper bounds presented in Section 7.3 for the Bernoulli sampling model. In Section 7.4.1, we highlight some combinatorial properties and characterizations of chordal graphs that underpin both the construction and the analysis of the estimators in Section 7.4.2. The special case of disjoint unions of cliques is treated in Section 7.4.3, where the estimator of Frank [16] is recovered and further improved. Analogous results for the uniform sampling model are given in Section 7.4.4. Fi-

nally, in Section 7.4.5, we discuss a heuristic to generalize the methodology to non-chordal graphs.

7.4.1 Combinatorial properties of chordal graphs

In this subsection we discuss the relevant combinatorial properties of chordal graphs which aid in the design and analysis of our estimators. We start by introducing a notion of vertex elimination ordering.

Definition 2. A *perfect elimination ordering (PEO)* of a graph G on N vertices is a vertex labelling $\{v_1, v_2, \dots, v_N\}$ such that, for each j , $N_G(v_j) \cap \{v_1, \dots, v_{j-1}\}$ is a clique.

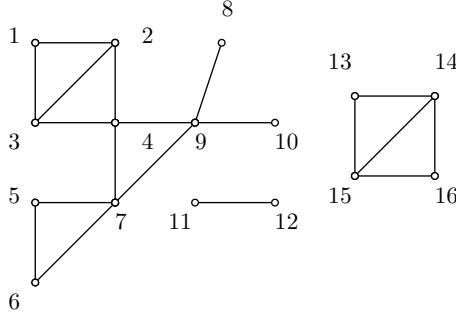


Figure 7.4: A chordal graph G with PEO labelled. In this example, $\text{cc}(G) = 3 = 16 - 19 + 6 = s(K_1, G) - s(K_2, G) + s(K_3, G)$.

In other words, if one eliminates the vertices sequentially according to a PEO starting from the last vertex, at each step, the neighborhood of the vertex to be eliminated forms a clique; see Fig. 7.4 for an example. A classical result of Dirac asserts that the existence of a PEO is in fact the defining property of chordal graphs (cf. e.g., [142, Theorem 5.3.17]).

Theorem 25. A graph is chordal if and only if it admits a PEO.

In general a PEO of a chordal graph is not unique; however, it turns out that the size of each neighborhood in the vertex elimination process is unique up to permutation, a fact that we will exploit later on. The next theorem makes this claim precise.

Lemma 36. Let $\{v_1, \dots, v_N\}$ and $\{v'_1, \dots, v'_N\}$ be two PEOs of a chordal graph G . Let c_j and c'_j denote the cardinalities of $N_G(v_j) \cap \{v_1, \dots, v_{j-1}\}$ and $N_G(v'_j) \cap \{v'_1, \dots, v'_{j-1}\}$,

respectively. Then there is a bijection between the set of numbers $\{c_j : j \in [N]\}$ and $\{c'_j : j \in [N]\}$.

Proof. By [142, Theorem 5.3.26], the chromatic polynomial of G is

$$\chi(G; x) = (x - c_1) \cdots (x - c_N) = (x - c'_1) \cdots (x - c'_N).$$

The conclusion follows from the uniqueness of the chromatic polynomial (and its roots). \square

Recall that $s(K_i, G)$ denotes the number of cliques of size i in G . For any chordal graph G , it turns out that the number of components can be expressed as an alternating sum of clique counts (cf. e.g., [142, Exercise 5.3.22, p. 231]); see Fig. 7.4 for an example. Instead of the topological proof involving properties of the clique simplex of chordal graphs [143, 144], in the next lemma we provide a combinatorial proof together with a sandwich bound. The main purpose of this exposition is to explain how to enumerate cliques in chordal graphs using vertex elimination, which plays a key role in analyzing the statistical estimator developed in the next subsection.

Lemma 37. *For any chordal graph G ,*

$$\text{cc}(G) = \sum_{i \geq 1} (-1)^{i+1} s(K_i, G). \quad (7.5)$$

Furthermore, for any $r \geq 1$,

$$\sum_{i=1}^{2r} (-1)^{i+1} s(K_i, G) \leq \text{cc}(G) \leq \sum_{i=1}^{2r-1} (-1)^{i+1} s(K_i, G). \quad (7.6)$$

Proof. Since G is chordal, by Theorem 25, it has a PEO $\{v_1, \dots, v_N\}$. Define

$$C_j \triangleq N_G(v_j) \cap \{v_1, \dots, v_{j-1}\}, \quad c_j \triangleq |C_j|. \quad (7.7)$$

Since the neighbors of v_j among v_1, \dots, v_{j-1} form a clique, we obtain $\binom{c_j}{i-1}$ new cliques of

size i when we adjoin the vertex v_j to the subgraph induced by v_1, \dots, v_{j-1} . Thus,

$$s(K_i, G) = \sum_{j=1}^N \binom{c_j}{i-1}. \quad (7.8)$$

Moreover, note that

$$cc(G) = \sum_{j=1}^N \mathbb{1}\{c_j = 0\}.$$

Hence, it follows that

$$\begin{aligned} \sum_{i=1}^{2r-1} (-1)^{i+1} s(K_i, G) &= \sum_{i=1}^{2r-1} (-1)^{i+1} \sum_{j=1}^N \binom{c_j}{i-1} = \sum_{j=1}^N \sum_{i=1}^{2r-1} (-1)^{i+1} \binom{c_j}{i-1} \\ &= \sum_{j=1}^N \sum_{i=0}^{2(r-1)} (-1)^i \binom{c_j}{i} = \sum_{j=1}^N \left(\binom{c_j-1}{2(r-1)} \mathbb{1}\{c_j \neq 0\} + \mathbb{1}\{c_j = 0\} \right) \\ &\geq \sum_{j=1}^N \mathbb{1}\{c_j = 0\} = cc(G), \end{aligned}$$

and

$$\begin{aligned} \sum_{i=1}^{2r} (-1)^{i+1} s(K_i, G) &= \sum_{i=1}^{2r} (-1)^{i+1} \sum_{j=1}^N \binom{c_j}{i-1} = \sum_{j=1}^N \sum_{i=1}^{2r} (-1)^{i+1} \binom{c_j}{i-1} \\ &= \sum_{j=1}^N \sum_{i=0}^{2r-1} (-1)^i \binom{c_j}{i} = \sum_{j=1}^N \left(-\binom{c_j-1}{2r-1} \mathbb{1}\{c_j \neq 0\} + \mathbb{1}\{c_j = 0\} \right) \\ &\leq \sum_{j=1}^N \mathbb{1}\{c_j = 0\} = cc(G). \quad \square \end{aligned}$$

7.4.2 Estimators for chordal graphs

Bounded clique number: unbiased estimators

In this subsection, we consider unbiased estimation of the number of connected components in chordal graphs. As we will see, unbiased estimators turn out to be minimax rate-optimal for chordal graphs with bounded clique size. The subgraph count identity (7.5) suggests

the following unbiased estimator

$$\hat{\mathbf{c}}\mathbf{c} = - \sum_{i \geq 1} \left(-\frac{1}{p}\right)^i \mathbf{s}(K_i, \tilde{G}). \quad (7.9)$$

Indeed, since the probability of observing any given clique of size i is p^i , (7.9) is clearly unbiased in the same spirit of the Horvitz-Thompson estimator [145]. In the case where the parent graph G is a forest, (7.9) reduces to the estimator $\hat{\mathbf{c}}\mathbf{c} = \mathbf{v}(\tilde{G})/p - \mathbf{e}(\tilde{G})/p^2$, as proposed by Frank [16].

A few comments about the estimator (7.9) are in order. First, it is completely adaptive to the parameters ω , d and N , since the sum in (7.9) terminates at the clique number of the subsampled graph. Second, it can be evaluated in time that is linear in $\mathbf{v}(\tilde{G}) + \mathbf{e}(\tilde{G})$. Indeed, the next lemma gives a simple formula for computing (7.9) using the PEO. Since a PEO of a chordal graph G can be found in $O(\mathbf{v}(G) + \mathbf{e}(G))$ time [146] and any induced subgraph of a chordal graph remains chordal, the estimator (7.9) can be evaluated in linear time. Recall that $q = 1 - p$.

Lemma 38. *Let $\{\tilde{v}_1, \dots, \tilde{v}_m\}$, $m = |S|$, be a PEO of \tilde{G} . Then*

$$\hat{\mathbf{c}}\mathbf{c} = \frac{1}{p} \sum_{j=1}^m \left(-\frac{q}{p}\right)^{\tilde{c}_j}, \quad (7.10)$$

where $\tilde{c}_j \triangleq |N_{\tilde{G}}(\tilde{v}_j) \cap \{\tilde{v}_1, \dots, \tilde{v}_{j-1}\}|$ can be calculated from \tilde{G} in linear time.

Proof. Because the subsampled graph \tilde{G} is also chordal, by (7.8), we have $\mathbf{s}(K_i, \tilde{G}) = \sum_{j=1}^m \binom{\tilde{c}_j}{i-1}$. Thus, (7.9) can also be written as

$$\begin{aligned} \hat{\mathbf{c}}\mathbf{c} &= - \sum_{i=1}^m \left(-\frac{1}{p}\right)^i \mathbf{s}(K_i, \tilde{G}) = - \sum_{i=1}^m \left(-\frac{1}{p}\right)^i \sum_{j=1}^m \binom{\tilde{c}_j}{i-1} \\ &= - \sum_{j=1}^m \sum_{i=1}^m \left(-\frac{1}{p}\right)^i \binom{\tilde{c}_j}{i-1} = \frac{1}{p} \sum_{j=1}^m \sum_{i=0}^{m-1} \left(-\frac{1}{p}\right)^i \binom{\tilde{c}_j}{i} \\ &= \frac{1}{p} \sum_{j=1}^m \left(-\frac{q}{p}\right)^{\tilde{c}_j}. \end{aligned} \quad \square$$

Using elementary enumerative combinatorics, in particular, the vertex elimination struc-

ture of chordal graphs, the next theorem provides a performance guarantee for the estimator (7.9) in terms of a variance bound and a high-probability bound, which, in particular, settles the upper bound of the minimax mean squared error in Theorem 21 and Theorem 22.

Theorem 26. *Let G be a chordal graph on N vertices with maximum degree and clique number at most d and $\omega \geq 2$, respectively. Suppose \tilde{G} is generated by the $\text{Bern}(p)$ sampling model. Then $\hat{\text{cc}}$ defined in (7.9) is an unbiased estimator of $\text{cc}(G)$. Furthermore,*

$$\text{Var}[\hat{\text{cc}}] \leq N \left(\frac{q}{p} + d \right) \left(\left(\frac{q}{p} \right)^{\omega-1} \vee \frac{q}{p} \right) \leq \frac{N}{p^\omega} + \frac{Nd}{p^{\omega-1}}, \quad (7.11)$$

and for all $t \geq 0$,

$$\mathbb{P}[|\hat{\text{cc}} - \text{cc}(G)| \geq t] \leq 2 \exp \left\{ -\frac{8p^\omega t^2}{25(d\omega + 1)(N + t/3)} \right\}. \quad (7.12)$$

To prove Theorem 26 we start by presenting a useful lemma. Note that Lemma 38 states that $\hat{\text{cc}}$ is a linear combination of $(-q/p)\tilde{\text{c}}_j$; here c_j is computed using a PEO of the sampled graph, which itself is random. The next result allows us rewrite the same estimator as a linear combination of $(-q/p)\hat{\text{c}}_j$, where $\hat{\text{c}}_j$ depends on the PEO of the parent graph (which is deterministic). Note that this is only used in the course of analysis since the population level PEO is not observed. This representation is extremely useful in analyzing the performance of $\hat{\text{cc}}$ and its biased variant in Section 7.4.2. More generally, we prove the following result.

Lemma 39. *Let $\{v_1, \dots, v_N\}$ be a PEO of G and let $\{\tilde{v}_1, \dots, \tilde{v}_m\}$, $m = |S|$, be a PEO of \tilde{G} . Furthermore, let $\hat{\text{c}}_j = |N_{\tilde{G}}(v_j) \cap \{v_1, \dots, v_{j-1}\}|$ and $\tilde{\text{c}}_j = |N_{\tilde{G}}(\tilde{v}_j) \cap \{\tilde{v}_1, \dots, \tilde{v}_{j-1}\}|$. Let $\hat{\text{g}} = \hat{\text{g}}(\tilde{G})$ be a linear estimator of the form*

$$\hat{\text{g}} = \sum_{j=1}^m g(\tilde{\text{c}}_j). \quad (7.13)$$

Then

$$\hat{\text{g}} = \sum_{j=1}^N b_j g(\hat{\text{c}}_j),$$

where $b_j \triangleq \mathbb{1}\{v_j \in S\}$.

Proof. Note that $\{v_1, \dots, v_N\}$ is also a PEO⁴ of \tilde{G} and hence by Lemma 36, there is a bijection between $\{\tilde{c}_j : j \in [m]\}$ and $\{\hat{c}_j : j \in [N]\}$. Therefore

$$\hat{g} = \sum_{j=1}^m g(\tilde{c}_j) = \sum_{j=1}^N b_j g(\hat{c}_j). \quad \square$$

We also need a couple of ancillary results whose proofs are given in Appendix 7.7:

Lemma 40 (Orthogonality). *Let⁵*

$$f(k) = \left(-\frac{q}{p}\right)^k, \quad k \geq 0. \quad (7.14)$$

Let $\{b_v : v \in V\}$ be independent Bern(p) random variables. For any $S \subset V$, define $N_S = \sum_{v \in S} b_v$. Then

$$\mathbb{E}[f(N_S)f(N_T)] = \mathbb{1}\{S = T\}(q/p)^{|S|}.$$

In particular, $\mathbb{E}[f(N_S)] = 0$ for any $S \neq \emptyset$.

Lemma 41. *Let $\{v_1, \dots, v_N\}$ be a PEO of a chordal graph G on N vertices with maximum degree and clique number at most d and ω , respectively. Let $C_j \triangleq N_G(v_j) \cap \{v_1, \dots, v_{j-1}\}$. Then⁶*

$$|\{(i, j) : i \neq j, C_j = C_i \neq \emptyset\}| \leq N(d-1). \quad (7.15)$$

Furthermore, let

$$A_j = \{v_j\} \cup C_j. \quad (7.16)$$

Then for each $j \in [N]$,

$$|\{i \in [N] : i \neq j, A_i \cap A_j \neq \emptyset\}| \leq d\omega. \quad (7.17)$$

4. When we say a PEO $\{v_1, \dots, v_N\}$ of G is also a PEO of $\tilde{G} = G[S]$, it is understood in the following sense: for any $v_j \in S$, $N_{\tilde{G}}(v_j) \cap \{v_i \in S : i < j\}$ is a clique in $G[S]$.

5. In fact, the function $f(N_S) = (-\frac{q}{p})^{N_S}$ is the (unnormalized) orthogonal basis for the binomial measure that is used in the analysis of Boolean functions [147, Definition 8.40].

6. The bound in (7.15) is almost optimal, since the left-hand side is equal to $N(d-2)$ when G consists of $N/(d+1)$ copies of stars S_d .

To prove a high-probability bound for the proposed estimator we also need a concentration inequality for sum of dependent random variables due to Janson [148]. The following result can be distilled from [148, Theorem 2.3]. The two-sided version of the concentration inequality therein also holds; see the paragraph before [148, Equation (2.3)].

Lemma 42. *Let $X = \sum_{j \in [N]} Y_j$, where $|Y_j - \mathbb{E}[Y_j]| \leq b$ almost surely. Let $S = \sum_{j \in [N]} \text{Var}[Y_j]$. Let $\Gamma = ([N], E(\Gamma))$ be a dependency graph for $\{Y_j\}_{j \in [N]}$ in the sense that if $A \subset [N]$, and $i \in [N] \setminus A$ does not belong to the neighborhood of any vertex in A , then Y_i is independent of $\{Y_j\}_{j \in A}$. Furthermore, suppose Γ has maximum degree d_{\max} . Then, for all $t \geq 0$,*

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq 2 \exp \left\{ -\frac{8t^2}{25(d_{\max} + 1)(S + bt/3)} \right\}.$$

Proof of Theorem 26. For a chordal graph G on N vertices, let $\{v_1, \dots, v_N\}$ be a PEO of G . Recall from (7.7) that C_j denote the set of neighbors of v_j among v_1, \dots, v_{j-1} and c_j denotes its cardinality. That is,

$$c_j = |N_G(v_j) \cap \{v_1, \dots, v_{j-1}\}| = \sum_{k=1}^{j-1} \mathbb{1}\{v_k \sim v_j\}.$$

As in Lemma 39, let \widehat{c}_j denote the sample version, i.e.,

$$\widehat{c}_j \triangleq |N_{\widetilde{G}}(v_j) \cap \{v_1, \dots, v_{j-1}\}| = b_j \sum_{k=1}^{j-1} b_k \mathbb{1}\{v_k \sim v_j\},$$

where $b_k \triangleq \mathbb{1}\{v_k \in S\} \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p)$. By Lemma 38 and Lemma 39, \widehat{c} can be written as

$$\widehat{c} = \frac{1}{p} \sum_{j=1}^m f(\widetilde{c}_j) = \frac{1}{p} \sum_{j=1}^N b_j f(\widehat{c}_j), \quad (7.18)$$

where the function f is defined in (7.14).

To show the variance bound (7.11), we note that

$$\text{Var}[\widehat{c}] = \frac{1}{p^2} \sum_{j=1}^N \text{Var}[b_j f(\widehat{c}_j)] + \frac{1}{p^2} \sum_{j \neq i}^N \text{Cov}[b_j f(\widehat{c}_j), b_i f(\widehat{c}_i)]. \quad (7.19)$$

Note that $\widehat{c}_j \mid \{b_j = 1\} \sim \text{Bin}(c_j, p)$. Using Lemma 40, it is straightforward to verify that

$$\text{Var}[b_j f(\widehat{c}_j)] = \begin{cases} p \left(\frac{q}{p}\right)^{c_j} & \text{if } c_j > 0 \\ pq & \text{if } c_j = 0 \end{cases}. \quad (7.20)$$

Since $c_j \leq \omega - 1$, it follows that

$$\text{Var}[b_j f(\widehat{c}_j)] \leq p \left[\left(\frac{q}{p}\right)^{\omega-1} \vee \frac{q}{p} \right]. \quad (7.21)$$

The covariance terms are less obvious to bound; but thanks to the orthogonality property in Lemma 40, many of them are zero or negative. Let $N_C \triangleq \sum b_j \mathbb{1}\{v_j \in C\}$. For any j , since $v_j \notin C_j$ by definition, applying Lemma 40 yields

$$\mathbb{E}[b_j f(\widehat{c}_j)] = p \mathbb{E}[f(N_{C_j})] = p \mathbb{1}\{C_j = \emptyset\}. \quad (7.22)$$

Without loss of generality, assume $j < i$. By the definition of C_j , we have $v_i \notin C_j$. Next, we consider two cases separately:

Case I: $v_j \notin C_i$. If either C_j or C_i is nonempty, Lemma 40 yields

$$\text{Cov}[b_j f(\widehat{c}_j), b_i f(\widehat{c}_i)] \stackrel{(7.22)}{=} \mathbb{E}[b_i b_j f(\widehat{c}_j) f(\widehat{c}_i)] = p^2 \mathbb{E}[f(N_{C_j}) f(N_{C_i})] = p^2 \mathbb{1}\{C_j = C_i\} \left(\frac{q}{p}\right)^{c_j}.$$

If $C_j = C_i = \emptyset$, then $\text{Cov}[b_j f(\widehat{c}_j), b_i f(\widehat{c}_i)] = \text{Cov}[b_j, b_i] = 0$.

Case II: $v_j \in C_i$. Then $\mathbb{E}[b_i f(\widehat{c}_i)] = 0$ by (7.22). Using Lemma 40 again, we have

$$\begin{aligned} \text{Cov}[b_j f(\widehat{c}_j), b_i f(\widehat{c}_i)] &= p \mathbb{E} \left[b_j \left(-\frac{q}{p}\right)^{b_j} \right] \mathbb{E}[f(N_{C_j}) f(N_{C_i \setminus \{v_j\}})] \\ &= -pq \mathbb{E}[f(N_{C_j}) f(N_{C_i \setminus \{v_j\}})] \\ &= -pq \mathbb{1}\{C_j = C_i \setminus \{v_j\}\} \left(\frac{q}{p}\right)^{c_j}. \end{aligned}$$

To summarize, we have shown that

$$\text{Cov}[b_j f(\widehat{c}_j), b_i f(\widehat{c}_i)] = \begin{cases} p^2 \left(\frac{q}{p}\right)^{c_j} & \text{if } C_j = C_i \neq \emptyset \\ -pq \left(\frac{q}{p}\right)^{c_j} & \text{if } C_j = C_i \setminus \{v_j\} \text{ and } v_j \in C_i \\ 0 & \text{otherwise} \end{cases}$$

Thus,

$$\sum_{j \neq i} \text{Cov}[b_j f(\widehat{c}_j), b_i f(\widehat{c}_i)] \leq \sum_{j \neq i: C_j = C_i \neq \emptyset} p^2 \left(\frac{q}{p}\right)^{c_j} \stackrel{(7.15)}{\leq} N(d-1)p^2 \left[\left(\frac{q}{p}\right)^{\omega-1} \vee \frac{q}{p} \right]. \quad (7.23)$$

Finally, combining (7.19), (7.21) and (7.23) yields the desired (7.11).

The high-probability bound (7.12) for $\widehat{c}c$ follows from the concentration inequality in Lemma 42. To apply this result, note that $\widehat{c}c$ is a sum of dependent random variables

$$\widehat{c}c = \sum_{j \in [N]} Y_j, \quad (7.24)$$

where $Y_j = \frac{1}{p} b_j f(\widehat{c}_j)$ satisfies $\mathbb{E}[Y_j] = 0$ for $c_j > 0$ and $|Y_j| \leq b \triangleq \left(\frac{1}{p}\right)^\omega$ almost surely. Also, $S \triangleq \sum_{j \in [N]} \text{Var}[Y_j] \leq N\left(\frac{1}{p}\right)^\omega$ by (7.20). To control the dependency between $\{Y_j\}_{j \in [N]}$, note that $\widehat{c}_j = b_j \sum_{k: v_k \in C_j} b_k$. Thus Y_j only depends on $\{b_k : k \in A_j\}$, where $A_j = \{v_j\} \cup C_j$. Define a dependency graph Γ , where $V(\Gamma) = [N]$ and

$$E(\Gamma) = \{\{i, j\} : i \neq j, A_i \cap A_j \neq \emptyset\}.$$

Then Γ has maximum degree bounded by $d\omega$, by Lemma 41. □

Unbounded clique number: smoothed estimators

Up to this point, we have only considered unbiased estimators of the number of connected components. If the sample ratio p is at least $\frac{1}{2}$, Theorem 21 implies its variance is

$$\text{Var}[\widehat{c}c] \leq N(d+1),$$

regardless of the clique number ω of the parent graph. However, if the clique number ω grows with N , for small sampling ratio p the coefficients of the unbiased estimator (7.9) are as large as $\frac{1}{p^\omega}$ which results in exponentially large variance. Therefore, in order to deal with graphs with large cliques, we must give up unbiasedness to achieve better bias-variance tradeoff. Using a technique known as *smoothing* introduced in [140], next we modify the unbiased estimator to achieve a good bias-variance tradeoff.

To this end, consider a discrete random variable $L \in \mathbb{N}$ independent of everything else. Define the following estimator by discarding those terms in (7.10) for which \tilde{c}_j exceeds L , and then averaging over the distribution of L . In other words, let

$$\hat{c}_L \triangleq \mathbb{E}_L \left[\frac{1}{p} \sum_{j=1}^m \left(-\frac{q}{p} \right)^{\tilde{c}_j} \mathbb{1}_{\{\tilde{c}_j \leq L\}} \right] = \frac{1}{p} \sum_{j=1}^m \left(-\frac{q}{p} \right)^{\tilde{c}_j} \mathbb{P}[L \geq \tilde{c}_j]. \quad (7.25)$$

Effectively, smoothing acts as soft truncation by introducing a tail probability that modulates the exponential growth of the original coefficients. The variance can then be bounded by the maximum magnitude of the coefficients in (7.25). Like (7.9), (7.25) can be computed in linear time.

The next theorem bounds the mean-square error of \hat{c}_L , which implies the minimax upper bound previously announced in Theorem 23.

Theorem 27. *Let $L \sim \text{Poisson}(\lambda)$ with $\lambda = \frac{p}{2-3p} \log \left(\frac{Np}{1+d\omega} \right)$. If the maximum degree and clique number of G is at most d and ω , respectively, then when $p < 1/2$,*

$$\mathbb{E}_G |\hat{c}_L - \text{cc}(G)|^2 \leq 2N^2 \left(\frac{Np}{1+d\omega} \right)^{-\frac{p}{2-3p}}.$$

Proof. Let $\{v_1, \dots, v_N\}$ be a PEO of the parent graph G and let $\{\tilde{v}_1, \dots, \tilde{v}_m\}$, $m = |S|$, be a PEO of \tilde{G} and $\tilde{c}_j = |N_{\tilde{G}}(\tilde{v}_j) \cap \{\tilde{v}_1, \dots, \tilde{v}_{j-1}\}|$. Let $\hat{c}_j = |N_{\tilde{G}}(v_j) \cap \{v_1, \dots, v_{j-1}\}|$ and $c_j = |N_G(v_j) \cap \{v_1, \dots, v_{j-1}\}|$. By Lemma 39, we can rewrite \hat{c}_L as

$$\hat{c}_L = \frac{1}{p} \sum_{j \geq 1} b_j \left(-\frac{q}{p} \right)^{\hat{c}_j} \mathbb{P}[L \geq \hat{c}_j],$$

where $\hat{c}_j \sim \text{Bin}(c_j, p)$ conditioned on $\{b_j = 1\}$.

We compute the bias and variance of $\widehat{\mathbf{c}}_L$ and then optimize over λ . First,

$$\begin{aligned}
\mathbb{E}[\mathbf{cc}(G) - \widehat{\mathbf{c}}_L] &= \frac{1}{p} \sum_{j=1}^N \mathbb{E}[b_j \left(-\frac{q}{p}\right)^{\widehat{\mathbf{c}}_j} \mathbb{P}[L < \widehat{\mathbf{c}}_j]] = \sum_{j=1}^N \sum_{i=0}^{\mathbf{c}_j} \binom{\mathbf{c}_j}{i} p^i q^{\mathbf{c}_j-i} \left(-\frac{q}{p}\right)^i \mathbb{P}[L < i] \\
&= \sum_{j=1}^N q^{\mathbf{c}_j} \sum_{i=0}^{\mathbf{c}_j} \binom{\mathbf{c}_j}{i} (-1)^i \mathbb{P}[L < i] = \sum_{j=1}^N q^{\mathbf{c}_j} \sum_{i=0}^{\mathbf{c}_j} \binom{\mathbf{c}_j}{i} (-1)^i \sum_{\ell=0}^{i-1} \mathbb{P}[L = \ell] \\
&= \sum_{j=1}^N q^{\mathbf{c}_j} \sum_{\ell=0}^{\mathbf{c}_j-1} \mathbb{P}[L = \ell] \sum_{i=\ell+1}^{\mathbf{c}_j} \binom{\mathbf{c}_j}{i} (-1)^i \\
&\stackrel{(a)}{=} \sum_{j=1}^N q^{\mathbf{c}_j} \mathbb{E}_L \left[\binom{\mathbf{c}_j-1}{L} (-1)^{L+1} \right] \\
&\stackrel{(b)}{=} -e^{-\lambda} \sum_{j=1}^N q^{\mathbf{c}_j} L_{\mathbf{c}_j-1}(\lambda),
\end{aligned}$$

where (a) follows from the fact that $\sum_{i=\ell+1}^k \binom{k}{i} (-1)^i = \binom{k-1}{\ell} (-1)^{\ell+1}$, and (b) follows from

$$\mathbb{E}_L \left[\binom{k-1}{L} (-1)^{L+1} \right] = e^{-\lambda} L_{k-1}(\lambda), \quad (7.26)$$

where L_m is the Laguerre polynomial of degree m , which satisfies $|L_m(x)| \leq e^{x/2}$ for all $m \geq 0$ and $x \geq 0$ [149]. Thus

$$|\mathbb{E}[\widehat{\mathbf{c}}_L - \widehat{\mathbf{c}}]| \leq N e^{-\lambda/2}. \quad (7.27)$$

To bound the variance, write $\widehat{\mathbf{c}}_L = \frac{1}{p} \sum_{j=1}^N W_j$, where $W_j = b_j \left(-\frac{q}{p}\right)^{\widehat{\mathbf{c}}_j} \mathbb{P}[L \geq \widehat{\mathbf{c}}_j]$. Thus

$$\text{Var}[\widehat{\mathbf{c}}_L] = \frac{1}{p^2} \sum_{j \in [N]} \text{Var}[W_j] + \frac{1}{p^2} \sum_{i \neq j} \text{Cov}[W_i, W_j] \quad (7.28)$$

Note that W_j is a function of $\{b_\ell : v_\ell \in A_j, \ell \in [N]\}$, where A_j is defined in (7.16). Using Lemma 41, we have

$$|\{(i, j) \in [N]^2 : i \neq j, A_i \cap A_j \neq \emptyset\}| \leq N d \omega. \quad (7.29)$$

Thus the number of cross terms in (7.28) is at most $Nd\omega$ thanks to (7.29). Thus,

$$\text{Var}[\widehat{\text{cc}}_L] \leq \frac{N(1+d\omega)}{p^2} \max_{1 \leq j \leq N} \text{Var}[W_j]. \quad (7.30)$$

Finally, note that if $p < 1/2$, then

$$\text{Var}[W_j] \leq p \left(\sup_{k \geq 0} \left\{ \left(\frac{q}{p} \right)^k \mathbb{P}[L \geq k] \right\} \right)^2 \leq p \left(\mathbb{E}_L \left[\left(\frac{q}{p} \right)^L \right] \right)^2 = p \exp \left\{ 2\lambda \left(\frac{q}{p} - 1 \right) \right\}. \quad (7.31)$$

Combining (7.27), (7.30), and (7.31), we have

$$\mathbb{E}_G |\widehat{\text{cc}}_L - \text{cc}(G)|^2 \leq N^2 e^{-\lambda} + \frac{N(1+d\omega)}{p} \exp \left\{ 2\lambda \left(\frac{q}{p} - 1 \right) \right\}.$$

The choice of λ yields the desired bound. \square

7.4.3 Unions of cliques

If the parent graph G consists of disjoint union of cliques, so does the sampled graph \tilde{G} . Counting cliques in each connected components, we can rewrite the estimator (7.9) as

$$\widehat{\text{cc}} = \sum_{r \geq 1} \left(1 - \left(-\frac{q}{p} \right)^r \right) \tilde{\text{cc}}_r = \text{cc}(\tilde{G}) - \sum_{r \geq 1} \left(-\frac{q}{p} \right)^r \tilde{\text{cc}}_r, \quad (7.32)$$

where $\tilde{\text{cc}}_r$ is the number of components in the sampled graph \tilde{G} that have r vertices. This coincides with the unbiased estimator proposed by Frank [16] for cliques, which is, in turn, based on the estimator of Goodman [15]. The following theorem provides an upper bound on its variance, recovering the previous result in [16, Corollary 11]:

Theorem 28. *Let G be a disjoint union of cliques with clique number at most ω . Then $\widehat{\text{cc}}$ is an unbiased estimator of $\text{cc}(G)$ and*

$$\mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 = \text{Var}[\widehat{\text{cc}}] = \sum_{r=1}^N \left(\frac{q}{p} \right)^r \text{cc}_r \leq N \left(\left(\frac{q}{p} \right)^\omega \wedge \frac{q}{p} \right),$$

where cc_r is the number of connected components in G of size r .

Proof. The estimator (7.9) can also be written as $\widehat{\text{cc}} = \sum_{k=1}^{\text{cc}(G)} [1 - (-\frac{q}{p})^{\widetilde{N}_k}]$, where \widetilde{N}_k is the number of sampled vertices from the k^{th} component. Then $\widetilde{N}_k \stackrel{\text{ind.}}{\sim} \text{Bin}(N_k, p)$. Thus,

$$\text{Var}[\widehat{\text{cc}}] = \sum_{k=1}^{\text{cc}(G)} \left(\frac{q}{p}\right)^{N_k} = \sum_{r=1}^N \left(\frac{q}{p}\right)^r \text{cc}_r.$$

The upper bound follows from the fact that $\text{cc}_r = 0$ for all $r > \omega$ and $\sum_{r=1}^N \text{cc}_r = \text{cc}(G) \leq N$. \square

Theorem 28 implies that as long as we sample at least half of the vertices, i.e., $p \geq \frac{1}{2}$, for any G consisting of disjoint cliques, the unbiased estimator (7.32) satisfies

$$\mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 \leq N,$$

regardless of the clique size. However, if $p < 1/2$, the variance can be exponentially large in N . Next, we use the smoothing technique again to obtain a biased estimator with near-optimal performance. To this end, consider a discrete random variable $L \in \mathbb{N}$ and define the following estimator by truncating (7.32) at the random location L and average over its distribution:

$$\widetilde{\text{cc}}_L \triangleq \text{cc}(\widetilde{G}) - \mathbb{E}_L \left[\sum_{r=1}^L \left(-\frac{q}{p}\right)^r \widetilde{\text{cc}}_r \right] = \text{cc}(\widetilde{G}) - \sum_{r \geq 1} \left(-\frac{q}{p}\right)^r \mathbb{P}[L \geq r] \widetilde{\text{cc}}_r. \quad (7.33)$$

The following result, proved in Appendix 7.7, bounds the mean squared error of $\widetilde{\text{cc}}_L$ and, consequently, bounds the minimax risk in Theorem 24. It turns out that the smoothed estimator (7.33) with appropriately chosen parameters is nearly optimal. In fact, Theorem 29 gives an upper bound on the sampling complexity (see Table 7.1), which, in view of [141, Theorem 4], is seen to be optimal.

Theorem 29. *Let G be a disjoint union of cliques. Let $L \sim \text{Pois}(\lambda)$ with $\lambda = \frac{p}{2-3p} \log(N/4)$. If $p < 1/2$, then*

$$\mathbb{E}_G |\widetilde{\text{cc}}_L - \text{cc}(G)|^2 \leq N^2 (N/4)^{-\frac{p}{2-3p}}.$$

Remark 15. *Alternatively, we could specialize the estimator $\widehat{\text{cc}}_L$ in (7.25) that is designed*

for general chordal graphs to the case when G is a disjoint union of cliques; however, the analysis is less clean and the results are slightly weaker than Theorem 29.

7.4.4 Extensions to uniform sampling model

As we mentioned earlier, the uniform sampling model where n vertices are selected uniformly at random from G is similar to Bernoulli sampling with $p = n/N$. For this model, the unbiased estimator analogous to (7.9) is

$$\widehat{\mathbf{c}}\mathbf{c}_U = \sum_{i \geq 1} \frac{(-1)^{i+1}}{p_i} \mathbf{s}(K_i, \tilde{G}), \quad (7.34)$$

where $p_i \triangleq \frac{\binom{N-i}{n-i}}{\binom{N}{n}}$. Next we show that this unbiased estimator enjoys the same variance bound in Theorem 26 up to constant factors that only depend on ω . The proof of this result is given in Appendix 7.7.

Theorem 30. *Let \tilde{G} be generated from the uniform sampling model with $n = pN$. Then*

$$\text{Var}[\widehat{\mathbf{c}}\mathbf{c}_U] = O_\omega \left(\frac{N}{p^\omega} + \frac{Nd}{p^{\omega-1}} \right).$$

7.4.5 Non-chordal graphs

A general graph can always be made chordal by adding edges. Such an operation is called a *chordal completion* or *triangulation* of a graph, henceforth denoted by TRI. There are many ways to triangulate a graph and this is typically done with the goal of minimizing some objective function (e.g., number of edges or the clique number). Without loss of generality, triangulations do not affect the number of connected components, since the operation can be applied to each component.

In view of the various estimators and their performance guarantees developed so far for chordal graphs, a natural question to ask is how one might generalize those to non-chordal graphs. One heuristic is to first triangulate the subsampled graph and then apply the estimator such as (7.10) and (7.25) that are designed for chordal graphs. Suppose a trian-

gulation operation commutes with subgraph sampling in distribution,⁷ then the modified estimator would inherit all the performance guarantees proved for chordal graphs; unfortunately, this does not hold in general. Thus, so far our theory does not readily extend to non-chordal graphs. Nevertheless, the empirical performance of this heuristic estimator is competitive with $\widehat{\mathbf{C}}$ in both performance (see Fig. 7.13) and computational efficiency. Indeed, there are polynomial time algorithms that add at most $8k^2$ edges if at least k edges must be added to make the graph chordal [151]. In view of the theoretical guarantees in Theorem 26, it is better to be conservative with adding edges so as the maximal degree d and the clique number ω are kept small.

It should be noted that blindly applying estimators designed for chordal graphs to the subsampled non-chordal graph without triangulation leads to nonsensical estimates. Thus, preprocessing the graph appears to be necessary for producing good results. We will leave the task of rigorously establishing these heuristics for future work.

7.5 Lower bounds

7.5.1 General strategy

Next we give a general lower bound for estimating additive graph properties (e.g. the number of connected components, subgraph counts) under the Bernoulli sampling model. The proof uses the method of two fuzzy hypotheses [152, Theorem 2.15], which, in the context of estimating graph properties, entails constructing a pair of random graphs whose properties have different average values, and the distributions of their subsampled versions are close in total variation, which is ensured by matching lower-order subgraph counts or sampling certain configurations on their vertices. The utility of this result is to use a pair of smaller graphs (which can be found in an ad hoc manner) to construct a bigger pair of graphs on N vertices and produce a lower bound that scales with N .

Theorem 31. *Let f be a graph parameter that is invariant under isomorphisms and addi-*

7. By “commute in distribution” we mean the random graphs $\text{TRI}(\widetilde{G})$ and $\widehat{\text{TRI}}(G)$ have the same distribution. That is, the triangulated sampled graph is statistically identical to a sampled graph from a triangulation of the parent graph.

tive under disjoint union, i.e., $f(G + H) = f(G) + f(H)$ [137, p. 41]. Let \mathcal{G} be a class of graphs with at most N vertices. Let m and $M = N/m$ be integers. Let H and H' be two graphs with m vertices. Assume that any disjoint union of the form $G_1 + \dots + G_M$ is in \mathcal{G} where G_i is either H or H' . Suppose $M \geq 300$ and $\text{TV}(P, P') \leq 1/300$, where P (resp. P') denote the distribution of the isomorphism class of the sampled graph \tilde{H} (resp. \tilde{H}'). Let \tilde{G} denote the sampled version of G under the Bernoulli sampling model with probability p . Then

$$\inf_{\hat{f}} \sup_{G \in \mathcal{G}} \mathbb{P} \left[|\hat{f}(\tilde{G}) - f(G)| \geq \Delta \right] \geq 0.01. \quad (7.35)$$

where

$$\Delta = \frac{|f(H) - f(H')|}{8} \left(\sqrt{\frac{N}{m \text{TV}(P, P')}} \wedge \frac{N}{m} \right).$$

Proof. Fix $\alpha \in (0, 1)$. Let $M = N/m$ and $G = G_1 + G_2 + \dots + G_M$, where $G_i \simeq H$ or H' with probability α and $1 - \alpha$, respectively. Let \mathbb{P}_α denote the law of G and \mathbb{E}_α the corresponding expectation. Assume without loss of generality that $f(H) > f(H')$. Note that $\mathbb{E}_\alpha f(G) = M[\alpha f(H) + (1 - \alpha)f(H')]$.

Let \tilde{G}_i be the sample version of G_i . Then $\tilde{G} = \tilde{G}_1 + \dots + \tilde{G}_M$. For each subgraph h , by (7.2), we have

$$\mathbb{P} \left[\tilde{G}_i \simeq h \mid G_i \simeq H \right] = \mathfrak{s}(h, H) p^{\mathfrak{v}(h)} (1 - p)^{m - \mathfrak{v}(h)},$$

and

$$\mathbb{P} \left[\tilde{G}_i \simeq h \mid G_i \simeq H' \right] = \mathfrak{s}(h, H') p^{\mathfrak{v}(h)} (1 - p)^{m - \mathfrak{v}(h)}.$$

Let $P \triangleq P_{\tilde{H}} = \mathcal{L}(\tilde{G}_i \mid G_i \simeq H)$ and $P' \triangleq P_{\tilde{H}'} = \mathcal{L}(\tilde{G}_i \mid G_i \simeq H')$. Then the law of each \tilde{G}_i is simply a mixture $P_\alpha \triangleq \mathcal{L}(\tilde{G}_i) = \alpha P + (1 - \alpha)P'$. Furthermore, $(\tilde{G}_1, \tilde{G}_2, \dots, \tilde{G}_M)' \sim P_\alpha^{\otimes M}$.

To lower bound the minimax risk of estimating the functional $f(G)$, we apply the method of two fuzzy hypotheses [152, Theorem 2.15(i)]. To this end, consider a pair of priors, that is, the distribution of G with $\alpha = \alpha_0 = 1/2$ and $\alpha_1 = 1/2 + \delta$, respectively, where $\delta \in [0, 1/2]$ is to be determined. To ensure that the values of $f(G)$ are separated under the two priors, note that $f(G) \stackrel{\text{D}}{=} (f(H) - f(H')) \text{Bin}(M, \alpha) + f(H')M$. Define $L = f(H)(1/2 + \delta/4)M +$

$f(H')(1/2 - \delta/4)M$ and

$$\Delta \triangleq \frac{1}{4}(\mathbb{E}_{\alpha_1} f(G) - \mathbb{E}_{\alpha_0} f(G)) = \frac{M\delta}{4}(f(H) - f(H')).$$

By Hoeffding's inequality, for any $\delta \geq 0$,

$$\mathbb{P}_{\alpha_0}[f(G) \leq L] = \mathbb{P}[\text{Bin}(M, \alpha_0) \leq M\alpha_0 + M\delta/4] \geq 1 - e^{-\delta^2 M/8} \triangleq 1 - \beta_0.$$

and

$$\mathbb{P}_{\alpha_1}[f(G) \geq L + 2\Delta] = \mathbb{P}[\text{Bin}(M, \alpha_1) \geq M\alpha_1 - M\delta/4] \geq 1 - e^{-\delta^2 M/8} \triangleq 1 - \beta_1.$$

Invoking [152, Theorem 2.15(i)], we have

$$\inf_{\hat{f}} \sup_{G \in \mathcal{G}} \mathbb{P} \left[|\hat{f}(\tilde{G}) - f(G)| \geq \Delta \right] \geq \frac{1 - \text{TV}(P_{\alpha_0}^{\otimes M}, P_{\alpha_1}^{\otimes M}) - \beta_0 - \beta_1}{2}. \quad (7.36)$$

The total variation term can be bounded as follows:

$$\begin{aligned} \text{TV}(P_{\alpha_0}^{\otimes M}, P_{\alpha_1}^{\otimes M}) &\stackrel{(a)}{\leq} 1 - \frac{1}{2} \exp\{-\chi^2(P_{\alpha_0}^{\otimes M} \| P_{\alpha_1}^{\otimes M})\} \\ &= 1 - \frac{1}{2} \exp\{-(1 + \chi^2(P_{\alpha_0} \| P_{\alpha_1}))^M + 1\} \\ &\stackrel{(b)}{\leq} 1 - \frac{1}{2} \exp\{-(1 + 4\delta^2 \text{TV}(P, P'))^M + 1\}, \end{aligned}$$

where (a) follows from the inequality between the total variation and the χ^2 -divergence $\chi^2(P \| Q) \triangleq \int (\frac{dP}{dQ} - 1)^2 dQ$ [152, Eqn. (2.25)]; (b) follows from

$$\begin{aligned} \chi^2(P_{\alpha_0} \| P_{\alpha_1}) &= \chi^2 \left(\frac{P + P'}{2} + \delta(P - P') \left\| \frac{P + P'}{2} \right\| \right) \\ &= \delta^2 \int \frac{(P - P')^2}{\frac{P + P'}{2}} \leq 4\delta^2 \text{TV}(P, P'). \end{aligned}$$

Choosing $\delta = \frac{1}{2} \wedge \sqrt{\frac{1}{4M\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}')}} and in view of the assumptions that $M \geq 300$ and$

$\text{TV}(P, P') \leq 1/300$, the right-hand size of (7.36) is at least

$$\frac{1}{4} \exp\{-(1 + 4\delta^2 \text{TV}(P, P'))^M + 1\} - e^{-\delta^2 M/8} \geq 0.01,$$

which proves (7.35). \square

7.5.2 Bounding total variations between sampled graphs

The application of Theorem 31 relies on the construction of a pair of small graphs H and H' whose sampled versions are close in total variation. To this end, we offer two schemes to bound $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'})$ from above.

Matching subgraphs

Since $\text{cc}(G)$ is invariant with respect to isomorphisms, it suffices to describe the sampled graph \tilde{G} up to isomorphisms. It is well-known that a graph G can be determined up to isomorphisms by its homomorphism numbers that count the number of ways to embed a smaller graph in G . Among various versions of graph homomorphism numbers (cf. [137, Sec 5.2]) the one that is most relevant to the present chapter is $\mathbf{s}(H, G)$, which, as defined in Section 7.1.2, is the number of *vertex-induced* subgraphs of G that are isomorphic to H . Specifically, the relevance of induced subgraph counts to the subgraph sampling model is two-fold:

- The list of vertex-induced subgraph counts $\{\mathbf{s}(H, G) : v(H) \leq N\}$ determines G up to isomorphism and hence constitutes a sufficient statistic for \tilde{G} . In fact, it is further sufficient to summarize \tilde{G} into the list of numbers⁸

$$\{\mathbf{s}(H, \tilde{G}) : v(H) \leq N, H \text{ is connected}\},$$

since the counts of disconnected subgraphs is a fixed polynomial of connected subgraph counts. This is a well-known result in the graph reconstruction theory [153–155]. For

8. This statistic cannot be further reduced because it is known that the connected subgraphs counts do not fulfill any predetermined relations in the sense that the closure of the range of their normalized version (subgraph densities) has nonempty interior [153].

example, for any graph G , we have

$$s(\circ \circ, G) = \binom{s(\circ, G)}{2} - s(\circ - \circ, G)$$

and

$$\begin{aligned} s(\circ - \circ, G) &= \binom{s(\circ - \circ, G)}{2} - s(\circ \circ, G) - 3s(\circ \triangle, G) - s(\circ - \circ - \circ, G) \\ &\quad - 2s(\circ \square, G) - s(\circ \triangle - \circ, G) - 2s(\circ \nabla, G) - 3s(\circ \boxtimes, G), \end{aligned}$$

which can be obtained by counting pairs of vertices or edges in two different ways, respectively. See [156, Section 2] for more examples.

- Under the Bernoulli sampling model, the probabilistic law of the isomorphism class of the sampled graph is a polynomial in the sampling ratio p , with coefficients given by the induced subgraph counts. Indeed, recall from (7.2) that $\mathbb{P}[\tilde{G} \simeq H] = s(H, G)p^{v(H)}(1-p)^{v(G)-v(H)}$. Therefore two graphs with matching subgraph counts for all (connected) graphs of n vertices are statistically indistinguishable unless more than n vertices are sampled.

We begin with a refinement of the classical result that says disconnected subgraph counts are fixed polynomials of connected subgraph counts. Below we provide a more quantitative version by showing that only those connected subgraphs which contain no more vertices than the disconnected subgraph involved. The proof of this next result is given in Appendix 7.7.

Lemma 43. *Let H be a disconnected graph of v vertices. Then for any G , $s(H, G)$ can be expressed as a polynomial, independent of G , in $\{s(g, G) : g \text{ is connected and } v(g) \leq v\}$.*

Corollary 4. *Suppose H and H' are two graphs in which $s(h, H) = s(h, H')$ for all connected h with $v(h) \leq v$. Then $s(h, H) = s(h, H')$ for all h with $v(h) \leq v$.*

Lemma 44. *Let H and H' be two graphs on m vertices. If*

$$s(h, H) = s(h, H') \tag{7.37}$$

for all connected graphs h with at most k vertices with $k \in [m]$, then

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq \mathbb{P}[\text{Bin}(m, p) \geq k+1] \leq \binom{m}{k+1} p^{k+1}. \quad (7.38)$$

Furthermore, if $p \leq (k+1)/m$, then

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq \exp \left\{ -\frac{2(k+1-pm)^2}{m} \right\}. \quad (7.39)$$

Proof. By Corollary 4, we have

$$\mathbf{s}(h, H) = \mathbf{s}(h, H'), \quad (7.40)$$

for all h (not necessarily connected) with $\mathbf{v}(h) \leq k$. Note that conditioned on ℓ vertices are sampled, \tilde{H} is uniformly distributed over the collection of all induced subgraphs of H with ℓ vertices. Thus

$$\mathbb{P}[\tilde{H} \simeq h \mid \mathbf{v}(\tilde{H}) = \ell] = \frac{\mathbf{s}(h, H)}{\binom{m}{\ell}}.$$

In view of (7.40), we conclude that the isomorphism class of \tilde{H} and \tilde{H}' have the same distribution provided that no more than k vertices are sampled. Hence the first inequality in (7.38) follows, while the last inequality therein follows from the union bound $\mathbb{P}[\text{Bin}(m, p) \geq \ell] \leq \binom{m}{\ell} p^\ell$. The bound (7.39) follows directly from Hoeffding's inequality on the binomial tail probability in (7.38). \square

In Fig. 7.5, we give an example of two graphs H and H' on 8 vertices that have matching counts of connected subgraphs with at most 4 vertices. Thus, by Lemma 44, they also have matching counts of *all* subgraphs with at most 4 vertices, and if $p \leq 5/8$, then $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq e^{-\frac{25}{4}(1-\frac{8p}{5})^2}$.

Labeling-based coupling

It is well-known that for any probability distributions P and P' , the total variation is given by $\text{TV}(P, P') = \inf \mathbb{P}[X \neq X']$, where the infimum is over all couplings, i.e., joint distributions of X and X' that are marginally distributed as P and P' respectively. There is a natural coupling between the sampled graphs \tilde{H} and \tilde{H}' when we define the parent graph

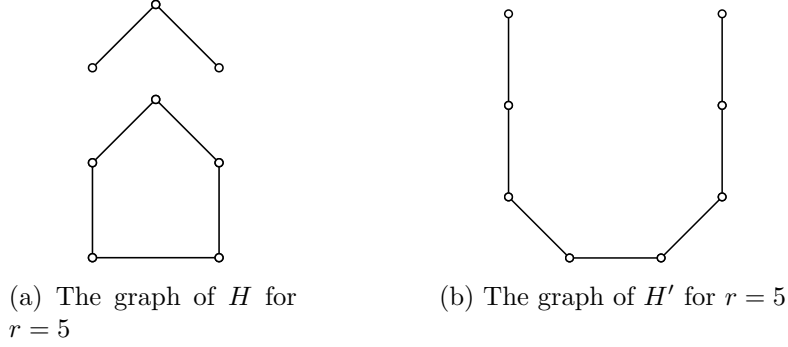


Figure 7.5: Each connected subgraph with $k \leq 4$ vertices appears exactly $9 - k$ times in each graph.

H and H' on the same set of labelled vertices. In some of the applications of Theorem 31, the constructions of H and H' are such that if certain configurations of the vertices are included or excluded in the sample, the resulting graphs are isomorphic. This property allows us to bound the total variation between the sampled graphs as follows.

Lemma 45. *Let H and H' be graphs defined on the same set of vertices V . Let U be a subset of V and suppose that for any $u \in U$, we have $H[V \setminus \{u\}] \simeq H'[V \setminus \{u\}]$. Then, the total variation $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'})$ can be bounded by the probability that every vertex in U is sampled, viz.,*

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq 1 - \mathbb{P}[\tilde{H} \simeq \tilde{H}'] \leq p^{|U|}.$$

If, in addition, $H[U] \simeq H'[U]$, then the total variation $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'})$ can be bounded by the probability that every vertex in U is sampled and at least one vertex in $V \setminus U$ is sampled, viz.,

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq p^{|U|}(1 - (1 - p)^{|V| - |U|}).$$

In Fig. 7.6, we give an example of two graphs H and H' satisfying the assumption of Lemma 45. In this example, $|U| = 2$, and $|V| = 8$. Note that if any of the vertices in U are removed along with all their incident edges, then the resulting graphs are isomorphic. Also, since $H[U] \simeq H'[U]$, Lemma 45 implies that $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq p^2(1 - (1 - p)^6)$.

In the remainder of the section, we apply Theorem 31, Lemma 44, and Lemma 45 to derive lower bounds on the minimax risk for graphs that contain cycles and general chordal graphs, respectively. The main task is to handcraft a pair of graphs H and H' that either

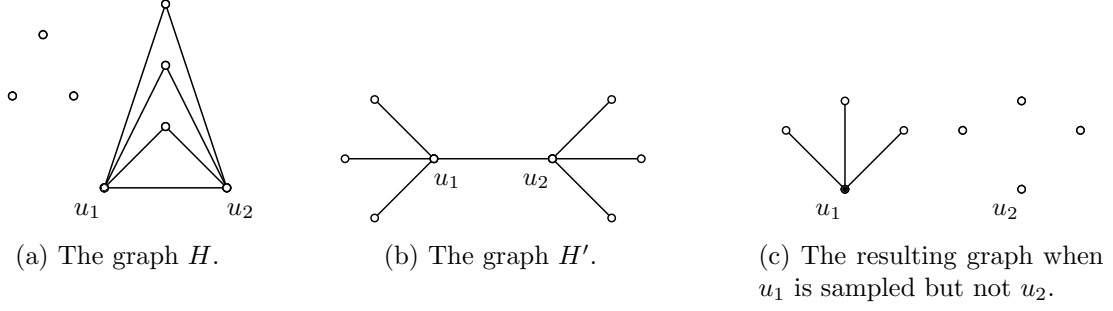


Figure 7.6: Example where $U = \{u_1, u_2\}$ is an edge. If any of these vertices are not sampled and all incident edges are removed, the resulting graphs are isomorphic.

have matching counts of small subgraphs *or* for which certain configurations of their vertices induce subgraphs that are isomorphic.

7.5.3 Lower bound for graphs with long induced cycles

Theorem 32. *Let $\mathcal{G}(N, r)$ denote the collection of all graphs on N vertices with longest induced cycle at most r , $r \geq 4$. Suppose $p < 1/2$ and $r \geq \frac{6}{(1-2p)^2}$. Then*

$$\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{G}(N, r)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 \gtrsim N e^{r(1-2p)^2} \wedge \frac{N^2}{r^2}.$$

In particular, if $p < 1/2$ and $r = \Theta(\log N)$, then

$$\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{G}(N, r)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)| \gtrsim \frac{N}{\log N}.$$

Proof. We will prove the lower bound via Theorem 31 with $m = 2(r-1)$. Let $H = C_r + P_{r-2}$ and $H' = P_{2(r-1)}$. Note that $s(P_i, H) = s(P_i, H') = 2r - 1 - i$ for $i = 1, 2, \dots, r-1$. For an illustration of the construction when $r = 3$, see Fig. 7.5. Since paths of length at most $r-1$ are the only connected subgraphs of H and H' with at most $r-1$ vertices, Corollary 4 implies that H and H' have matching subgraph counts up to order $r-1$.

In the notation of Theorem 31, $k = r-1$, $m = 2(r-1)$, and $|\text{cc}(H) - \text{cc}(H')| = 1$. By Theorem 31,

$$\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{G}(N, r)} \mathbb{P}[|\widehat{\text{cc}} - \text{cc}(G)| \geq \Delta] \geq 0.10,$$

where

$$\Delta \asymp |\text{cc}(H) - \text{cc}(H')| \left(\sqrt{\frac{N}{m\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'})}} \wedge \frac{N}{m} \right) = \left(\sqrt{\frac{N}{m\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'})}} \wedge \frac{N}{m} \right).$$

Furthermore, by (7.39), the total variation between the sampled graphs \tilde{H} and \tilde{H}' satisfies

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq e^{-\frac{r^2}{r-1}(1-2p+\frac{2p}{r})^2} \leq e^{-r(1-2p)^2} < 1/300,$$

provided $p < 1/2$ and $r \geq \frac{6}{(1-2p)^2}$. The desired lower bound on the squared error follows from Markov's inequality. \square

7.5.4 Lower bound for chordal graphs

Theorem 33 (Chordal graphs). *Let $\mathcal{G}(N, d, \omega)$ denote the collection of all chordal graphs on N vertices with maximum degree and clique number at most d and $\omega \geq 2$, respectively. Assume that $p < \frac{1}{2^{\omega-100}}$. Then*

$$\inf_{\hat{\text{cc}}} \sup_{G \in \mathcal{G}(N, d, \omega)} \mathbb{E}_G |\hat{\text{cc}} - \text{cc}(G)|^2 = \Theta_{\omega} \left(\left(\frac{N}{p^{\omega}} \vee \frac{Nd}{p^{\omega-1}} \right) \wedge N^2 \right).$$

Proof. There are two different constructions we give, according to whether $d \geq 2^{\omega}$ or $d < 2^{\omega}$.

Case I: $d \geq 2^{\omega}$. For every $\omega \geq 2$ and $m \in \mathbb{N}$, we construct a pair of graphs H and H' , such that

$$v(H) = v(H') = \omega - 1 + m2^{\omega-2} \tag{7.41}$$

$$d_{\max}(H) = d_{\max}(H') = m2^{\omega-3} + \omega - 2, \quad \omega \geq 3 \tag{7.42}$$

$$d_{\max}(H) = 0, \quad d_{\max}(H') = m, \quad \omega = 2 \tag{7.43}$$

$$\text{cc}(H) = m + 1, \quad \text{cc}(H') = 1 \tag{7.44}$$

$$|\text{s}(K_{\omega}, H) - \text{s}(K_{\omega}, H')| = m \tag{7.45}$$

Fix a set of $\omega - 1$ vertices U that forms a clique. We first construct H . For every subset $S \subset U$ such that $|S|$ is even, let V_S be a set of m distinct vertices such that the neighborhood

of every $v \in V_S$ is given by $\partial v = S$. Let the vertex set $V(H)$ be the union of U and all V_S such that $|S|$ is even. In particular, because of the presence of $S = \emptyset$, H always has exactly m isolated vertices (unless $\omega = 2$, in which case H consists of $m + 1$ isolated vertices). Repeat the same construction for H' with $|S|$ being odd. Then both H and H' are chordal and have the same number of vertices as in (7.41), since

$$v(H) = \omega - 1 + m \sum_{0 \leq i \leq \omega-1, i \text{ even}} \binom{\omega-1}{i} = v(H') = \omega - 1 + m \sum_{0 \leq i \leq \omega-1, i \text{ odd}} \binom{\omega-1}{i}$$

which follows from the binomial summation formula. Similarly, (7.42)–(7.45) can be readily verified.

We also have that

$$\begin{aligned} s(K_i, H) &= \binom{\omega-1}{i} + m \sum_{0 \leq j \leq \omega-1, j \text{ even}} \binom{\omega-1}{j} \binom{j}{i-1} = \\ s(K_i, H') &= \binom{\omega-1}{i} + m \sum_{0 \leq j \leq \omega-1, j \text{ odd}} \binom{\omega-1}{j} \binom{j}{i-1} = \\ &\quad \binom{\omega-1}{i} + m \binom{\omega-1}{i-1} 2^{\omega-1-i}, \end{aligned}$$

for $i = 1, 2, \dots, \omega - 1$. This follows from the fact that

$$\sum_{0 \leq j \leq \omega-1} (-1)^j \binom{\omega-1}{j} \binom{j}{i-1} = 0,$$

and

$$\sum_{0 \leq j \leq \omega-1} \binom{\omega-1}{j} \binom{j}{i-1} = \binom{\omega-1}{i-1} 2^{\omega-i}.$$

To compute the total variation distance between the sampled graphs, we first assume that H and H' are defined on the same set of labelled vertices V . The key observation is the following: by construction, $H[U] \simeq H'[U]$ (since U induces a clique) and, furthermore, failing to sample any vertex in U results in an isomorphic graph, i.e., $H[V \setminus \{u\}] \simeq H'[V \setminus \{u\}]$ for any $u \in U$. Indeed, the structure of the induced subgraph $H[V \setminus \{u\}]$ can be described as follows. First, let U form a clique. Next, for every nonempty subset $S \subset U \setminus \{u\}$, attach a set of m distinct vertices (denoted by V_S) so that the neighborhood of every $v \in V_S$ is given

by $\partial v = S$. Finally, add $m + 1$ isolated vertices. See Fig. 7.6 ($\omega = 3$) and Fig. 7.7 ($\omega = 4$) for illustrations of this property and the iterative nature of this construction, in the sense that the construction of H (resp. H') for $\omega = k + 1$ can be obtained from the construction of H (resp. H') for $\omega = k$ by adding another vertex u to U such that $\partial u = U$ and then adjoining m distinct vertices to every even (resp. odd) cardinality set $S \subset U$ containing u .

Thus by Lemma 45,

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq p^{|U|} \left(1 - (1 - p)^{|V| - |U|}\right) = p^{\omega-1} (1 - (1 - p)^{m2^{\omega-2}}).$$

According to (7.42), we choose $m = \lfloor (d - \omega + 2)2^{-\omega+3} \rfloor \geq d2^{-\omega+2}$ if $\omega \geq 3$ and $m = d$ if $\omega = 2$. Then we have,

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = p^{\omega-1} (1 - (1 - p)^d) \leq p^{\omega-1} (pd \wedge 1).$$

The condition on p ensures that $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq p < 1/300$. In view of Theorem 31 and (7.44), we have

$$\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{G}(N, d, \omega)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 = \Theta_\omega \left(\left(\frac{N}{p^\omega} \vee \frac{Nd}{p^{\omega-1}} \right) \wedge N^2 \right),$$

provided $d \geq 2^\omega$.

Case II: $d \leq 2^\omega$. In this case, the previous construction is no longer feasible and we must construct another pair of graphs that have a smaller maximum degree. To this end, we consider graphs H and H' consisting of disjoint cliques of size at most $\omega \geq 2$, such that

$$\begin{aligned} v(H) &= v(H') = \omega 2^{\omega-2} \\ d_{\max}(H) &= d_{\max}(H') = \omega - 1 \\ |\text{cc}(H) - \text{cc}(H')| &= 1. \end{aligned} \tag{7.46}$$

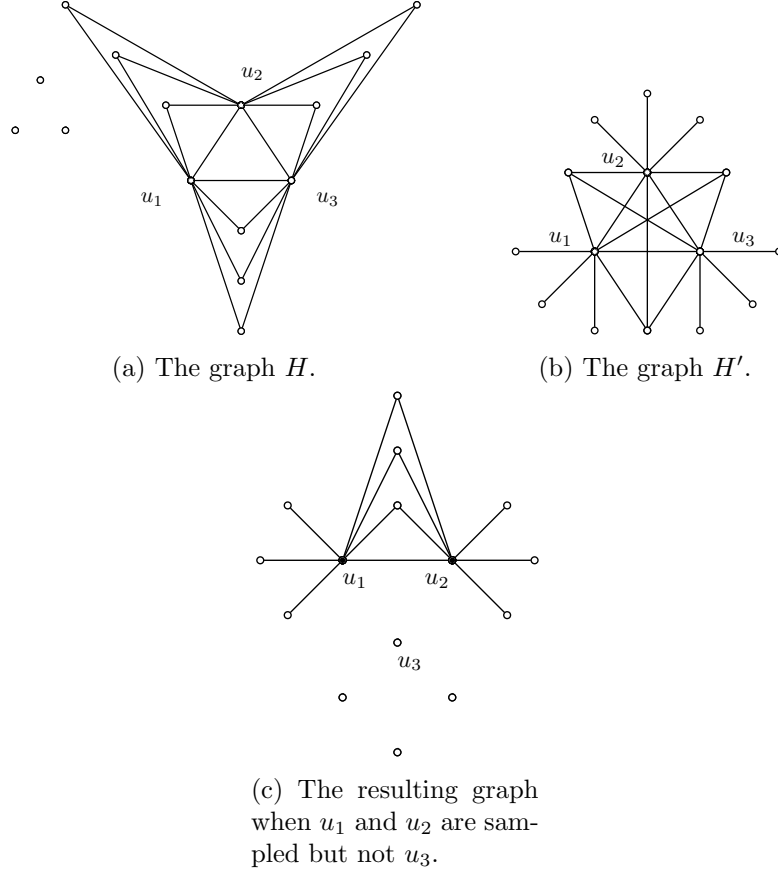


Figure 7.7: Example for $\omega = 4$ and $m = 3$, where $U = \{u_1, u_2, u_3\}$ form a triangle. If any one or two (as shown in the figure) of these vertices are not sampled and all incident edges are removed, the graphs are isomorphic.

If ω is odd, we set

$$\begin{aligned}
 H &= \binom{\omega}{\omega} K_{\omega} + \binom{\omega}{\omega-2} K_{\omega-2} + \cdots + \binom{\omega}{3} K_3 + \binom{\omega}{1} K_1 \\
 H' &= \binom{\omega}{\omega-1} K_{\omega-1} + \binom{\omega}{\omega-3} K_{\omega-3} + \cdots + \binom{\omega}{4} K_4 + \binom{\omega}{2} K_2.
 \end{aligned} \tag{7.47}$$

If ω is even, we set

$$\begin{aligned}
 H &= \binom{\omega}{\omega} K_{\omega} + \binom{\omega}{\omega-2} K_{\omega-2} + \cdots + \binom{\omega}{4} K_4 + \binom{\omega}{2} K_2 \\
 H' &= \binom{\omega}{\omega-1} K_{\omega-1} + \binom{\omega}{\omega-3} K_{\omega-3} + \cdots + \binom{\omega}{3} K_3 + \binom{\omega}{1} K_1.
 \end{aligned} \tag{7.48}$$

See Fig. 7.8 and Fig. 7.9 for examples of this construction.



Figure 7.8: Illustration for the construction in (7.47) for $\omega = 3$. Each graph contains a matching number of cliques of size up to 2.

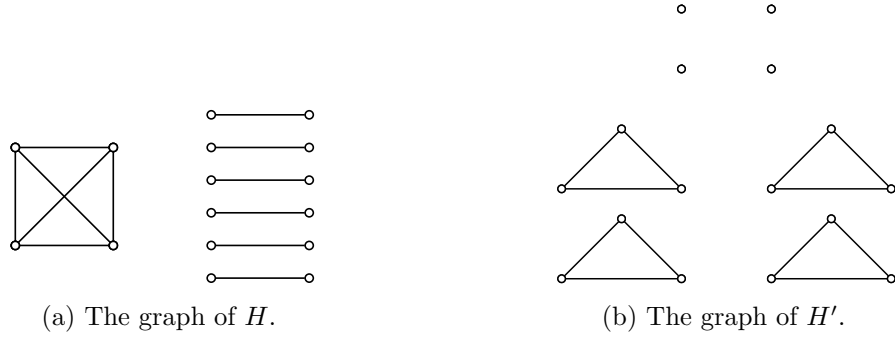


Figure 7.9: Illustration for the construction in (7.48) for $\omega = 4$. Each graph contains a matching number of cliques of size up to 3.

Next we verify that H and H' have matching subgraph counts. For $i = 1, 2, \dots, \omega - 1$,

$$s(K_i, H) - s(K_i, H') = \sum_{k=i}^{\omega} (-1)^k \binom{\omega}{k} \binom{k}{i} = 0,$$

and

$$s(K_i, H) = s(K_i, H') = \frac{1}{2} \sum_{k=i}^{\omega} \binom{\omega}{k} \binom{k}{i} = 2^{\omega-1-i} \binom{\omega}{i}.$$

Hence H and H' contain matching number of cliques up to size $\omega - 1$. Note that the only connected induced subgraphs of H and H' with at most $\omega - 1$ vertices are cliques. Consequently, by (7.38), $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq \binom{\omega-2}{\omega} p^\omega$ and together with Theorem 31 and (7.46), we have

$$\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{G}(N, d, \omega)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 \geq \Omega_\omega \left(\frac{N}{p^\omega} \wedge N^2 \right) = \Theta_\omega \left(\left(\frac{N}{p^\omega} \vee \frac{Nd}{p^{\omega-1}} \right) \wedge N^2 \right),$$

where the last inequality follows from the current assumption that $d \leq 2^\omega$. The condition on p ensures that $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq p 2^{\omega-2} < 1/300$. \square

7.5.5 Lower bounds for forests

Particularizing Theorem 33 to $\omega = 2$, we obtain a lower bound which shows that the estimator for forests $\widehat{\text{cc}} = \text{v}(\tilde{G})/p - \text{e}(\tilde{G})/p^2$ proposed by Frank [16] is minimax rate-optimal. As opposed to the general construction in Theorem 33, Fig. 7.10 illustrates a simple construction of H and H' for forests. However, we still require that p is less than some absolute constant. Through another argument, we show that this constant can be arbitrarily close to one.

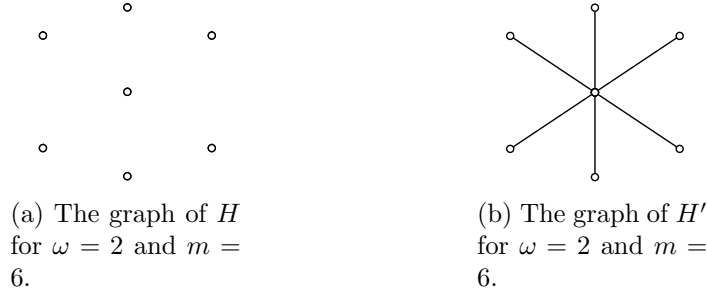


Figure 7.10: The two graphs are isomorphic if the center vertex is not sampled and all incident edges are removed. Thus, $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = p(1 - q^6)$.

Theorem 34 (Forests). *Let $\mathcal{F}(N, d) = \mathcal{G}(N, d, 2)$ denote the collection of all forests on N vertices with maximum degree at most d . Then for all $0 < p < 1$,*

$$\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{F}(N, d)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)|^2 \gtrsim \left(\frac{Nq}{p^2} \vee \frac{Nqd}{p} \right) \wedge N^2.$$

In particular, if $d = \Theta(N)$ and $\omega \geq 2$, then

$$\inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{G}(N, d, \omega)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)| \geq \inf_{\widehat{\text{cc}}} \sup_{G \in \mathcal{F}(N, d)} \mathbb{E}_G |\widehat{\text{cc}} - \text{cc}(G)| \gtrsim N.$$

Proof. The strategy is to choose a one-parameter family of forests \mathcal{F}_0 and reduce the problem to estimating the total number of trials in a binomial experiment with a given success probability. To this end, define $M = N/(d + 1)$ and let

$$\mathcal{F}_0 = \{(N - m(d + 1))S_0 + mS_d : m \in \{0, 1, \dots, M\}\}.$$

Let $G \in \mathcal{F}_0$. Because we do not observe the labels $\{b_v : v \in V(G)\}$, the distribution of

\tilde{G} can be described by the vector (T_0, T_1, \dots, T_d) , where T_j is the observed number of S_j . Since $T_0 = N - \sum_{j \geq 1} (j+1)T_j$, it follows that (T_1, \dots, T_d) is sufficient for \tilde{G} . Next, we will show that $T = T_1 + \dots + T_d \sim \text{Bin}(m, p')$, where $p' \triangleq p(1 - q^d)$ is sufficient for \tilde{G} . To this end, note that conditioned on $T = n$, the probability mass function of (T_1, \dots, T_d) at (n_1, \dots, n_d) is equal to

$$\begin{aligned} \frac{\mathbb{P}[T_1 = n_1, \dots, T_d = n_d, T = n]}{\mathbb{P}[T = n]} &= \frac{\binom{m}{n} \binom{n}{n_1, \dots, n_d} p_1^{n_1} \dots p_d^{n_d} (1 - p')^{m-n}}{\binom{m}{n} (p')^n (1 - p')^{m-n}} \\ &= \binom{n}{n_1, \dots, n_d} (p_1/p')^{n_1} \dots (p_d/p')^{n_d}, \end{aligned}$$

where $p_j \triangleq \binom{d}{j} p^j q^{d-j}$. Thus, $(T_1, \dots, T_d) \mid T = n \sim \text{Multinomial}(n, p_1/p', \dots, p_d/p')$, whose distribution is independent of m . Thus, since $\text{cc}(G) = N - md$, we have that

$$\begin{aligned} \inf_{\hat{\text{cc}}} \sup_{G \in \mathcal{F}(N, d)} \mathbb{E}_G |\hat{\text{cc}} - \text{cc}(G)|^2 &\geq \inf_{\hat{\text{cc}}} \sup_{G \in \mathcal{F}_0} \mathbb{E}_G |\hat{\text{cc}} - \text{cc}(G)|^2 \\ &= d^2 \inf_{\hat{m}(T)} \sup_{m \in \{0, 1, \dots, M\}} \mathbb{E}_{T \sim \text{Bin}(m, p')} |\hat{m}(T) - m|^2 \\ &\gtrsim \left(\frac{Nq}{p^2} \vee \frac{Nqd}{p} \right) \wedge N^2, \end{aligned}$$

which follows applying Lemma 46 below with $\alpha = p'$ and $M = N/(d+1)$ and the fact that $p' = p(1 - q^d) \leq p \wedge (p^2 d)$. \square

The proof of Lemma 46 is given in Appendix 7.7.

Lemma 46 (Binomial experiment). *Let $X \sim \text{Bin}(m, \alpha)$. For all $0 \leq \alpha \leq 1$ and $M \in \mathbb{N}$ known a priori,*

$$\inf_{\hat{m}} \sup_{m \in \{0, 1, \dots, M\}} \mathbb{E} |\hat{m}(X) - m|^2 \asymp \frac{(1 - \alpha)M}{\alpha} \wedge M^2.$$

7.6 Numerical experiments

In this section, we study the empirical performance of the estimators proposed in Section 8.5 using synthetic data from various random graphs. The error bars in the following plots show the variability of the relative error $\frac{|\hat{\text{cc}} - \text{cc}(G)|}{\text{cc}(G)}$ over 20 independent experiments of subgraph

sampling on a fixed parent graph G . The solid black horizontal line shows the sample average and the whiskers show the mean \pm the standard deviation.

Chordal graphs Both Fig. 7.11 and Fig. 7.12 focus on chordal graphs, where the parent graph is first generated from a random graph ensemble then triangulated by calculating a fill-in of edges to make it chordal (using a maximum cardinality search algorithm from [157]). In Fig. 7.11a, the parent graph G is a triangulated Erdős-Rényi graph $\mathcal{G}(N, \delta)$, with $N = 2000$ and $\delta = 0.0005$ which is below the connectivity threshold $\delta = \frac{\log N}{N}$ [158]. In Fig. 7.11b, we generate G with $N = 20000$ vertices by taking the disjoint union of 200 independent copies of $\mathcal{G}(100, 0.2)$ and then apply triangulation. In accordance with Theorem 26, the better performance in Fig. 7.11b is due to moderately sized d and ω , and large $\text{cc}(G)$.

In Fig. 7.12 we perform a simulation study of the smoothed estimator $\hat{\text{cc}}_L$ from Theorem 27. The parent graph is equal to a triangulated realization of $\mathcal{G}(1000, 0.0015)$ with $d = 88$, $\omega = 15$, and $\text{cc}(G) = 325$. The plots in Fig. 7.12b show that the sampling variability is significantly reduced for the smoothed estimator, particularly for small values of p (to show detail, the vertical axes are plotted on different scales). This behavior is in accordance with the upper bounds furnished in Theorem 26 and Theorem 27. Large values of ω inflate the variance of $\hat{\text{cc}}$ considerably by an exponential factor of $1/p^\omega$, whereas the effect of large ω on the variance of $\hat{\text{cc}}_L$ is polynomial, viz., $\omega^{\frac{p}{2-3p}}$. We chose the smoothing parameter λ to be $p \log N$, but other values that improve the performance can be chosen through cross-validation on various known graphs.

The non-monotone behavior of the relative error in Fig. 7.12a can be explained by the tradeoff between increasing p (which improves the accuracy) and increasing probability of observing a clique (which increases the variability, particularly in this case of large ω). Such behavior is apparent for moderate values of p (e.g., $p < 0.25$), but less so as p increases to 1 since the mean squared error tends to zero as more of the parent graph is observed. The plots also suggest that the marginal benefit (i.e., the marginal decrease in relative error) from increasing p diminishes for moderate values of p . Future research would address the selection of p , if such control was available to the experimenter.

Non-chordal graphs Finally, in Fig. 7.13 we experiment with sampling non-chordal graphs. As proposed in Section 7.4.5, one heuristic is to modify the original estimator by first triangulating the subsampled graph \tilde{G} to $\text{TRI}(\tilde{G})$ and then applying the estimator $\hat{c}\hat{c}$ in (7.10). The plots in Fig. 7.13 show that this strategy works well; in fact the performance is competitive with the same estimator in Fig. 7.11, where the parent graph is first triangulated and then subsampled.

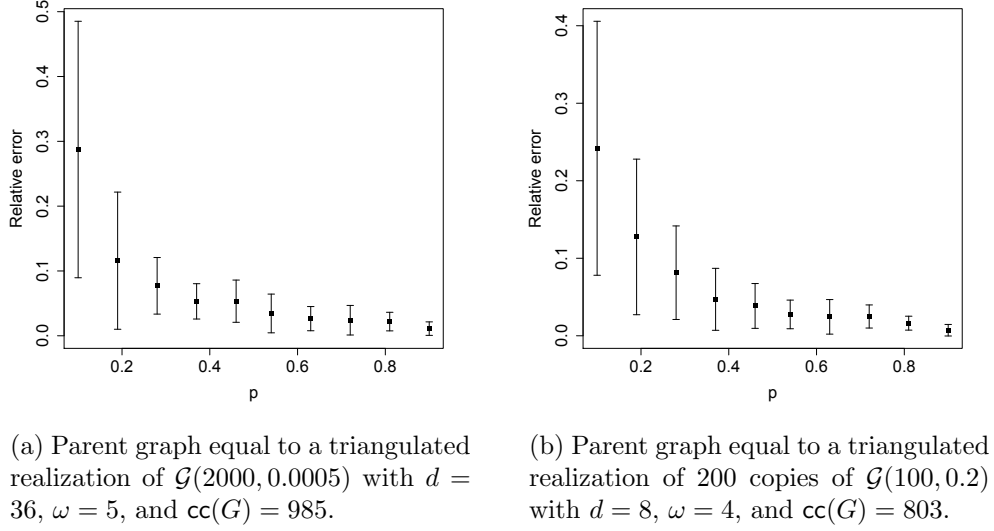


Figure 7.11: The relative error of $\hat{c}\hat{c}$ with moderate values of d and ω .

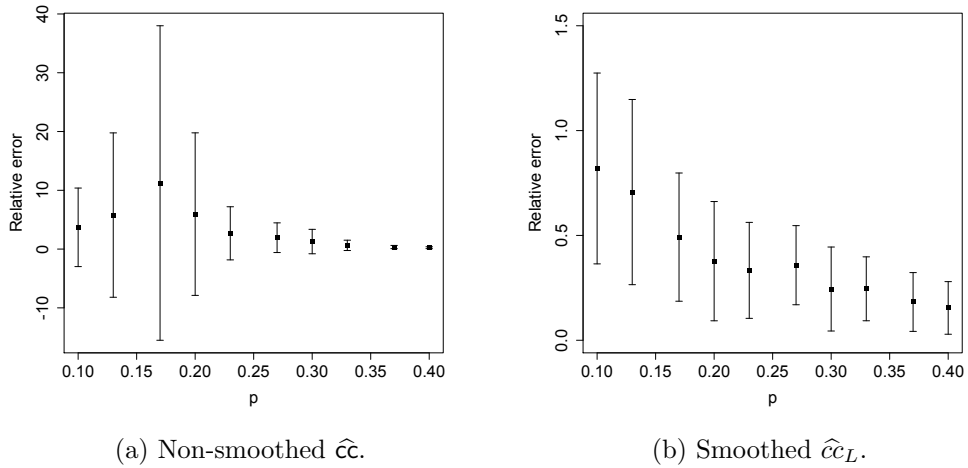
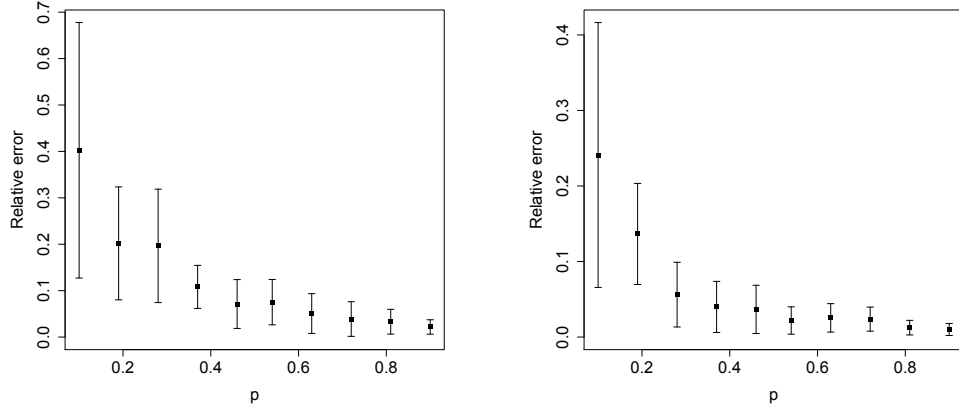


Figure 7.12: A comparison of the relative error of the unbiased estimator $\hat{c}\hat{c}$ in (7.10) and its smoothed version $\hat{c}\hat{c}_L$ in (7.25). The parent graph is a triangulated realization of $\mathcal{G}(1000, 0.0015)$ with $d = 88$, $\omega = 15$, and $\text{cc}(G) = 325$.



(a) Parent graph equal to a realization of $\mathcal{G}(2000, 0.0005)$ with $d = 8$, $\omega = 3$, and $\text{cc}(G) = 756$.

(b) Parent graph equal to a realization of 200 copies of $\mathcal{G}(100, 0.2)$ with $d = 7$, $\omega = 4$, and $\text{cc}(G) = 532$.

Figure 7.13: The estimator $\hat{\text{cc}}(\text{TRI}(\tilde{G}))$ applied to non-chordal graphs.

7.7 Additional proofs

In this appendix, we give the proofs of Lemma 40, Lemma 41, Theorem 29, Theorem 30, Lemma 43, and Lemma 46.

Proof of Lemma 40. Note that $N_S + N_T = N_{S \setminus T} + N_{T \setminus S} + 2N_{S \cap T}$, where $N_{S \setminus T}$, $N_{T \setminus S}$, and $N_{S \cap T}$ are independent binomially distributed random variables. By independence, we have

$$\begin{aligned} \mathbb{E}[f(N_S)f(N_T)] &= \mathbb{E} \left[\left(-\frac{q}{p} \right)^{N_S + N_T} \right] = \mathbb{E} \left[\left(-\frac{q}{p} \right)^{N_{S \setminus T} + N_{T \setminus S} + 2N_{S \cap T}} \right] \\ &= \mathbb{E} \left[\left(-\frac{q}{p} \right)^{N_{S \setminus T}} \right] \mathbb{E} \left[\left(-\frac{q}{p} \right)^{N_{T \setminus S}} \right] \mathbb{E} \left[\left(-\frac{q}{p} \right)^{2N_{S \cap T}} \right]. \end{aligned}$$

Finally, note that if $S \neq T$, then at least one of $\mathbb{E}[(-\frac{q}{p})^{N_{S \setminus T}}]$ or $\mathbb{E}[(-\frac{q}{p})^{N_{T \setminus S}}]$ is zero. If $S = T$, we have

$$\mathbb{E}[f(N_S)^2] = \mathbb{E} \left[\left(-\frac{q}{p} \right)^{2N_S} \right] = \left(\frac{q}{p} \right)^{|S|}.$$

□

Proof of Lemma 41. Let $c_j = |C_j|$. To prove (7.15), we will show that for any fixed j ,

$$|\{i \in [N] : i \neq j, C_j = C_i \neq \emptyset\}| \leq d - c_j \leq d - 1.$$

By definition of the PEO, $|N_G(v)| \geq c_j$ for all $v \in C_j$. For any $i \in [N]$ such that $C_j = C_i \neq \emptyset$, $v_i \in N_G(v)$ for all $v \in C_j$. Also, the fact that $C_j = C_i \neq \emptyset$ makes it impossible for $v_i \in C_j$. This shows that

$$c_j + |\{i \in [N] : i \neq j, C_j = C_i \neq \emptyset\}| \leq |N_G(v)| \leq d,$$

and hence the desired (7.15).

Next, we show (7.17). Let $a_j = |A_j|$. We will prove that for any fixed j ,

$$|\{i \in [N] : i \neq j, A_i \cap A_j \neq \emptyset\}| \leq da_j - (a_j - 1)^2. \quad (7.49)$$

This fact immediately implies (7.29) by noting that $a_j \leq \omega$. To this end, note that

$$\begin{aligned} |\{i \in [N] : i \neq j, A_i \cap A_j \neq \emptyset\}| &= |\{i \in [N] : i \neq j, v_i \notin A_j, A_i \cap A_j \neq \emptyset\}| + \\ &\quad |\{i \in [N] : i \neq j, v_i \in A_j\}|, \end{aligned}$$

where the second term is obviously at most $a_j - 1$. Next we prove that the first term is at most $(d + 1 - a_j)a_j$, which, in view of $(d + 1 - a_j)a_j + (a_j - 1) = da_j - (a_j - 1)^2$, implies the desired (7.49). Suppose, for the sake of contradiction, that

$$|\{i \in [N] : i \neq j, v_i \notin A_j, A_i \cap A_j \neq \emptyset\}| \geq (d + 1 - a_j)a_j + 1$$

Then at least $(d + 1 - a_j)a_j + 1$ of the A_i have nonempty intersection with A_j , meaning that at least $(d + 1 - a_j)a_j + 1$ vertices outside A_j are incident to vertices in A_j . By the pigeonhole principle, there is at least one vertex $u \in A_j$ which is incident to $d + 2 - a_j$ of those vertices outside A_j . Moreover, the vertices in A_j form a clique of size a_j in G by definition of the PEO. This implies that $|N_G(u)| \geq (a_j - 1) + (d - a_j + 2) = d + 1$, contradicting the maximum

degree assumption and completing the proof. \square

Proof of Theorem 29. The bias of this estimator is seen to be

$$\mathbb{E}[\text{cc}(G) - \tilde{\text{cc}}_L] = \sum_{k=1}^{\text{cc}(G)} \mathbb{E} \left[\mathbb{P}[L < \tilde{N}_k] \left(-\frac{q}{p} \right)^{\tilde{N}_k} \right].$$

Note that

$$\begin{aligned} \mathbb{E} \left[\mathbb{P}[L < \tilde{N}_k] \left(-\frac{q}{p} \right)^{\tilde{N}_k} \right] &= \sum_{r=1}^N \mathbb{P}[L < r] \left(-\frac{q}{p} \right)^r \mathbb{P}[\tilde{N}_k = r] \\ &= \sum_{i=0}^{N-1} \mathbb{P}[L = i] \sum_{r=i+1}^N \left(-\frac{q}{p} \right)^r \mathbb{P}[\tilde{N}_k = r]. \end{aligned}$$

Since $\tilde{N}_k \sim \text{Bin}(N_k, p)$, it follows that

$$\sum_{r=i+1}^N \left(-\frac{q}{p} \right)^r \mathbb{P}[\tilde{N}_k = r] = q^{N_k} \sum_{r=i+1}^N \binom{N_k}{r} (-1)^r = q^{N_k} (-1)^{i+1} \binom{N_k-1}{i}.$$

Putting these facts together, we have

$$\mathbb{E}[\text{cc}(G) - \tilde{\text{cc}}_L] = - \sum_{k=1}^{\text{cc}(G)} q^{N_k} P_{N_k-1}(\lambda) = \sum_{k=1}^{\text{cc}(G)} q^{N_k} \mathbb{E}_L \left[\binom{N_k-1}{L} (-1)^{L+1} \right],$$

Analogous to (7.26), we have $\left| \mathbb{E}_L \left[\binom{N_k-1}{L} (-1)^{L+1} \right] \right| \leq e^{-\lambda/2}$, and hence by the Cauchy-Schwarz inequality,

$$|\mathbb{E}[\text{cc}(G) - \tilde{\text{cc}}_L]| \leq e^{-\lambda/2} \sqrt{N \sum_{k=1}^{\text{cc}(G)} q^{N_k}}. \quad (7.50)$$

For the variance of $\tilde{\text{cc}}_L$, note that $\tilde{\text{cc}}_L = \sum_{k=1}^{\text{cc}(G)} W_k$, where $W_k \triangleq 1 - \mathbb{P}[L \geq \tilde{N}_k] \left(-\frac{q}{p} \right)^{\tilde{N}_k}$. The W_k are independent random variables and hence

$$\text{Var}[\tilde{\text{cc}}_L] = \sum_{k=1}^{\text{cc}(G)} \text{Var}[W_k] \leq \sum_{k=1}^{\text{cc}(G)} \mathbb{E}W_k^2.$$

Also,

$$W_k^2 \leq \max_{1 \leq r \leq N} \left\{ 1 - \mathbb{P}[L \geq r] \left(-\frac{q}{p} \right)^r \right\}^2 \mathbb{1}\{\tilde{N}_k \geq 1\}.$$

This means that

$$\text{Var}[\tilde{\text{cc}}_L] \leq \max_{1 \leq r \leq N} \left\{ 1 - \mathbb{P}[L \geq r] \left(-\frac{q}{p} \right)^r \right\}^{2 \text{cc}(G)} \sum_{k=1}^N (1 - q^{N_k}).$$

Since $p < 1/2$, we have

$$\begin{aligned} \mathbb{P}[L \geq r] \left(\frac{q}{p} \right)^r &= \sum_{i=r}^{\infty} \mathbb{P}[L = i] \left(\frac{q}{p} \right)^r \leq \sum_{i=r}^{\infty} \mathbb{P}[L = i] \left(\frac{q}{p} \right)^i \\ &\leq \sum_{i=0}^{\infty} \mathbb{P}[L = i] \left(\frac{q}{p} \right)^i = \mathbb{E}_L \left(\frac{q}{p} \right)^L = e^{\lambda(\frac{q}{p}-1)}. \end{aligned}$$

Thus, it follows that

$$\text{Var}[\tilde{\text{cc}}_L] \leq 4e^{2\lambda(\frac{q}{p}-1)} \sum_{k=1}^{\text{cc}(G)} (1 - q^{N_k}). \quad (7.51)$$

Combining (7.50) and (7.51) yields

$$\mathbb{E}|\tilde{\text{cc}}_L - \text{cc}(G)|^2 \leq 4e^{2\lambda(\frac{q}{p}-1)} \sum_{k=1}^{\text{cc}(G)} (1 - q^{N_k}) + Ne^{-\lambda} \sum_{k=1}^{\text{cc}(G)} q^{N_k} \leq \text{cc}(G) \max \left\{ 4e^{2\lambda(\frac{q}{p}-1)}, Ne^{-\lambda} \right\}.$$

Choosing $\lambda = \frac{p}{2-3p} \log(N/4)$ leads to $4e^{2\lambda(\frac{q}{p}-1)} = Ne^{-\lambda}$ and completes the proof. \square

Proof of Theorem 30. Using $(a_1 + \dots + a_k)^2 \leq k(a_1^2 + \dots + a_k^2)$, we have

$$\text{Var}[\widehat{\text{cc}}_U] \leq \omega \cdot \sum_{i=1}^{\omega} \frac{\text{Var}[\text{s}(K_i, \tilde{G})]}{p_i^2}. \quad (7.52)$$

Next, each variance term can be bounded as follows. Let $b_v = \mathbb{1}\{v \in S\} \sim \text{Bern}(p)$. Note that

$$\begin{aligned} \text{Var}[\text{s}(K_i, \tilde{G})] &= \text{Var} \left[\sum_{T: G[T] \simeq K_i} \prod_{v \in T} b_v \right] \\ &= \sum_{T: G[T] \simeq K_i} \text{Var} \left[\prod_{v \in T} b_v \right] + \sum_{k=0}^{i-1} \sum_{\substack{T \neq T': |T \cap T'| = k, \\ G[T] \simeq K_i, G[T'] \simeq K_i}} \text{Cov} \left[\prod_{v \in T} b_v, \prod_{v' \in T'} b_{v'} \right] \\ &= \text{s}(K_i, G)p_{i,i} + 2 \sum_{k=0}^{i-1} \text{n}(T_{i,k}, G)p_{i,k}, \end{aligned} \quad (7.53)$$

where

$$p_{i,k} \triangleq p_{2i-k} - p_i^2 = \frac{\binom{N-2i+k}{n-2i+k}}{\binom{N}{n}} - \left(\frac{\binom{N-i}{n-i}}{\binom{N}{n}} \right)^2, \quad 0 \leq k \leq i \leq n,$$

$T_{i,k}$ denotes two K_i 's sharing k vertices, and we recall that $n(H, G)$ notes the number of embeddings of (edge-induced subgraphs isomorphic to) H in G . It is readily seen that $\frac{p_{i,k}}{p_i^2} \leq \frac{i!}{p^k}$ since

$$\frac{p_{i,k}}{p_i^2} \leq \frac{p_{2i-k}}{p_i^2} = \frac{\binom{N-2i+k}{n-2i+k}}{\binom{N-i}{n-i}} \frac{\binom{N}{n}}{\binom{N-i}{n-i}} = \frac{\prod_{j=i+1}^{2i-k} \frac{n-j+1}{N-j+1}}{\prod_{j=1}^i \frac{n-j+1}{N-j+1}} \leq \frac{\prod_{j=i+1}^{2i-k} \frac{n}{N}}{\prod_{j=1}^i \frac{n}{N}} = \frac{i!}{p^k},$$

where we used $p = n/N$ and the inequalities $\frac{n}{N} \leq \frac{n-j+1}{N-j+1} \leq \frac{n}{N}$ for $1 \leq j \leq (1 + \frac{1}{N})n$.

Furthermore, from the same steps, for $k = 0$ we have

$$\frac{p_{2i}}{p_i^2} = \prod_{j=1}^i \frac{\frac{n-j+1-i}{N-j+1-i}}{\frac{n-j+1}{N-j+1}} \leq 1,$$

or equivalently, $p_{i,0} \leq 0$, which also follows from negative association.

Substituting $p_{i,0} \leq 0$ and $\frac{p_{i,k}}{p_i^2} \leq \frac{i!}{p^k}$ into (7.53) yields

$$\begin{aligned} \frac{1}{p_i^2} \text{Var}[s(K_i, \tilde{G})] &= \frac{s(K_i, G)p_{i,i}}{p_i^2} + 2 \sum_{k=0}^{i-1} n(T_{i,k}, G) \frac{p_{i,k}}{p_i^2} \\ &\leq \frac{s(K_i, G)p_{i,i}}{p_i^2} + 2 \sum_{k=1}^{i-1} n(T_{i,k}, G) \frac{p_{i,k}}{p_i^2} \\ &\leq i! \left(\frac{s(K_i, G)}{p^i} + 2 \sum_{k=1}^{i-1} \frac{n(T_{i,k}, G)}{p^k} \right). \end{aligned} \quad (7.54)$$

To finish the proof, we establish two combinatorial facts:

$$s(K_i, G) = O_\omega(N), \quad i = 1, 2, \dots, \omega \quad (7.55)$$

$$n(T_{i,k}, G) = O_\omega(Nd), \quad k = 1, 2, \dots, i-1 \quad (7.56)$$

Here (7.55) follows from the fact that for any chordal graph G with clique number bounded by ω , the number of cliques of any size is at most $O_\omega(|V(G)|) = O_\omega(N)$. This can be seen from the PEO representation in (7.8) since $c_j \leq \omega - 1$. To show (7.56), note that to

enumerate $T_{i,k}$, we can first enumerate cliques of size i , then for each clique, choose $i - k$ other vertices in the neighborhood of k vertices of the clique. Note that for each $v \in V(G)$, the neighborhood of v is also a chordal graph of at most d vertices and clique number at most ω . Therefore, by (7.55), the number of K_{i-k} 's in the neighborhood of any given vertex is at most $O_\omega(d)$.

Finally, applying (7.55)–(7.56) to each term in (7.54), we have

$$\frac{1}{p_i^2} \text{Var}[\mathbf{s}(K_i, \tilde{G})] = O_\omega \left(\frac{N}{p^i} + \sum_{k=1}^{i-1} \frac{Nd}{p^k} \right) = O_\omega \left(\frac{N}{p^i} + \frac{Nd}{p^{i-1}} \right),$$

which, in view of (7.52), yields the desired result. \square

Proof of Lemma 43. We use Kocay's Vertex Theorem [155] which says that if \mathcal{H} is a collection of graphs, then

$$\prod_{h \in \mathcal{H}} \mathbf{s}(h, G) = \sum_g a_g \mathbf{s}(g, G),$$

where the sum runs over all graphs g such that $v(g) \leq \sum_{h \in \mathcal{H}} v(h)$ and a_g is the number of decompositions of $V(g)$ into $\cup_{h \in \mathcal{H}} V(h)$ such that $g[V(h)] \simeq h$.

In particular, if \mathcal{H} consists of the connected components of H , then the only disconnected g with $v(g) = v$ satisfying the above decomposition property is $g \simeq H$. Hence

$$\mathbf{s}(H, G) = \frac{1}{a_H} \left[\prod_{h \in \mathcal{H}} \mathbf{s}(h, G) - \sum_g a_g \mathbf{s}(g, G) \right],$$

where the sum runs over all g that are either connected and $v(g) \leq v$ or disconnected and $v(g) \leq v - 1$. This shows that $\mathbf{s}(H, G)$ can be expressed as a polynomial, independent of G , in $\mathbf{s}(g, G)$ where either g is connected and $v(g) \leq v$ or g is disconnected and $v(g) \leq v - 1$.

The proof proceeds by induction on v . The base case of $v = 1$ is clearly true. Suppose that for any disconnected graph h with at most v vertices, $\mathbf{s}(h, G)$ can be expressed as a polynomial, independent of G , in $\mathbf{s}(g, G)$ where g is connected and $v(g) \leq v$. By the first part of the proof, if H is a disconnected graph with $v + 1$ vertices, then $\mathbf{s}(H, G)$ can be expressed as a polynomial, independent of G , in $\mathbf{s}(h, G)$ where either h is connected and $v(h) \leq v + 1$ or h is disconnected and $v(h) \leq v$. By $S(v)$, each $\mathbf{s}(h, G)$ with h disconnected

and $v(h) \leq v$ can be expressed as a polynomial, independent of G , in $\mathfrak{s}(g, G)$ where g is connected and $v(g) \leq v$. Thus, we can express $\mathfrak{s}(H, G)$ as a polynomial, independent of G , in terms of $\mathfrak{s}(g, G)$ where g is connected and $v(g) \leq v + 1$. \square

Proof of Lemma 46. The upper bound follows from choosing $\hat{m} = X/\alpha$ when $\alpha > (1 - \alpha)/M$ and $\hat{m} = (M + 1)/2$ when $\alpha \leq (1 - \alpha)/M$.

For the lower bound, let $\gamma > 0$. Consider the two hypothesis $H_1 : m_1 = M$ and $H_2 : m_2 = M - \sqrt{\frac{\gamma M}{\alpha}} \wedge M$. By Le Cam's two point method [152, Theorem 2.2(i)],

$$\begin{aligned} \inf_{\hat{m}} \sup_{m \in \{0, 1, \dots, M\}} \mathbb{E} |\hat{m}(X) - m|^2 &\geq \frac{1}{2} |m_1 - m_2|^2 [1 - \text{TV}(\text{Bin}(m_1, \alpha), \text{Bin}(m_2, \alpha))] \\ &\geq \frac{\frac{\gamma M}{\alpha} \wedge M^2}{2} [1 - d_{\text{H}}(\text{Bin}(m_1, \alpha), \text{Bin}(m_2, \alpha))], \end{aligned}$$

where we used the inequality between total variation and the Hellinger distance $\text{TV} \leq d_{\text{H}}$ [152, Lemma 2.3]. Finally, choosing $\gamma = (1 - \alpha)/16$ and using the bound in [159, Lemma 21] on Hellinger distance between two binomials, we obtain $d_{\text{H}}(\text{Bin}(m_1, \alpha), \text{Bin}(m_2, \alpha)) \leq 1/2$ as desired. \square

Chapter 8

Counting motifs with graph sampling

8.1 Introduction

As we saw in Chapter 7, counting the number of features in a graph is an important statistical and computational problem. These features are typically basic local structures like motifs [160] or graphlets [161] (e.g., patterns of small subgraphs). Seeking to capture the interactions and relationships between groups and individuals, applied researchers typically construct a network from data that has been collected from a random sample of nodes. This scenario is sometimes due to resource constraints (e.g., massive social network, surveying a hidden population) or an inability to gain access the full population (e.g., historical data, corrupted data). Most of the problems encountered in practice are motivated by the need to infer global properties of the parent network (population) from the sampled version. For specific motivations and applications of statistical inference on sampled graphs, we refer the reader to the [127–129] for comprehensive reviews as well as applications in computer networks and social networks.

From a computational and statistical perspective, it is desirable to design sublinear time (in the size of the graph) algorithms which typically involves random sampling as a primitive to reduce both time and sample complexities. Various sublinear-time algorithms based on

edge and degree queries have been proposed to estimate graph properties such as the average degree [119, 162], triangles [120], stars [121], and more general subgraph counts [163]. In all of these works, however, some form of adaptive queries, e.g. breadth or depth first search, is performed, which can be impractical or unrealistic in the context of certain applications such as social network analysis [107] or econometrics [112], where a sampled graph is obtained and statistical analysis is to be conducted on the basis of this dataset alone. In this work, we focus on data arising from specific sampling models, in particular, subgraph sampling and neighborhood sampling [137], two of the most popular and commonly used sampling models in part due to their simplicity and ease of implementation. In subgraph sampling, we sample each vertex independently with equal probability and observe the subgraph induced by these sampled vertices. In neighborhood sampling, we additionally observe the edges between the sampled vertices and their neighbors. Despite their ubiquity, theoretical understanding of these sampling models in the context of statistical inference and estimation has been lacking.

In this chapter, we study the problem of estimating the counts of various classes of motifs, such as edges, triangles, cliques, and wedges, from a statistical perspective. Network motifs are important local properties of a graph. Detecting and counting motifs have diverse applications in a suite of scientific applications including gene regulation networks [160], protein-protein interaction networks [164], and social networks [165]. Throughout this chapter, motifs will be viewed as *induced subgraphs* of the parent graph. For a subgraph H , the number of copies of H contained in G as induced subgraphs is denoted by $s(H, G)$. Many useful graph statistics can often be expressed in terms of induced subgraph counts, e.g., the global clustering coefficients, which is the density of induced open triangles. It is worth pointing out that in some literature motifs are also understood as (not necessarily induced) subgraphs [160]. In fact, it is well-known that the number of a given subgraph can be expressed as a linear combination of induced subgraph counts. For instance, if we denote the number of copies of H contained in G as subgraphs by $n(H, G)$, then for wedges,

we have $n(\text{open triangle}, G) = s(\text{open triangle}, G) + 3s(\text{closed triangle}, G)$.¹ For this reason, we focus on counting motifs as induced subgraphs. Furthermore, while we make no assumption about the connectivity of the parent graph, we focus on motifs being *connected* subgraphs which is the most relevant case for applications. It is a classical result that subgraph count of disconnected subgraphs can be expressed as a fixed polynomials in terms of the connected ones; cf. [137, 154]. Additionally, motifs in directed graphs have also been considered [160]; in this chapter we focus on undirected simple graphs.

The purpose of this chapter is to develop a statistical theory for estimating motif counts in sampled graph. We will be concerned with both methodologies as well as their statistical optimality, with focus on large graphs and the sublinear sample regime, where only a vanishing fraction of vertices are sampled. In particular, a few questions we want to address quantitatively are as follows:

- How does the sample complexity depend on the motif itself? For example, is estimating the count of open triangles as easy as estimating the closed triangles? How does the sample complexity of counting 4-cycles compare with that of counting 4-cliques?
- How much of the graph must be observed to ensure accurate estimation? For example, severe under-coverage issues have been observed in the study of protein-protein interaction networks [166].
- How much more informative is neighborhood sampling than subgraph sampling from the perspective of reducing the sample complexity?
- To what extent does additional structures of the parent graph, e.g., tree or planarity, impact the sample complexity?

Finally, let us also mention that motif counts e.g., triangles [167], wheels [168], and cycles [169] have been used as useful test statistics for generative network models such as

1. More generally, we have (cf. [137, Eq. (5.15)]):

$$n(H, G) = \sum_{H'} n(H, H') s(H', G), \quad (8.1)$$

where the summation ranges over all simple graphs H' (up to isomorphisms) obtained from H by adding edges.

the stochastic block models. Furthermore, edges counts of similarity and dependency graphs have been used in the context of testing and estimating change-point detection [170, 171]. In this chapter we do not assume any generative network model, and the randomness of the problem comes solely from the sampling mechanism.

8.1.1 Sampling model

In this subsection, we formally describe the two graph sampling models we will study in the remainder of the chapter.

Subgraph sampling. Fix a simple graph $G = (V, E)$ on $v(G)$ vertices. For $S \subset V$, we denote by $G[S]$ the vertex induced subgraph. If S represents a collection of vertices that are randomly sampled according to a sampling mechanism, we denote $G[S]$ by \tilde{G} . The first and simplest sampling model we consider is the subgraph sampling model, where each vertex is sampled with equal probability. In particular, we sample each vertex independently with probability p , where p is called the *sampling ratio* and can be thought of as the fraction of the graph that is observed. Thus, the sample size $|S|$ is distributed as $\text{Bin}(v(G), p)$, and the probability of observing a subgraph isomorphic to H is equal to

$$\mathbb{P}[\tilde{G} \simeq H] = s(H, G) p^{v(H)} (1 - p)^{v(G) - v(H)}. \quad (8.2)$$

There is also a variant of this model where exactly $n = pv(G)$ vertices are chosen uniformly at random without replacement from the vertex set V . In the sublinear sampling regime where $n \ll v(G)$, they are nearly equivalent.

Neighborhood sampling. In this model, in addition to observing $G[S]$, we also observe the labelled neighbors of all vertices in S , denoted by $G\{S\}$. That is, $G\{S\}$ is equal to $\tilde{G} = (V, \tilde{E})$, where $\tilde{E} = \cup_{v \in S} \cup_{u \in N_G(v)} \{u, v\}$ together with the colors $b_v \in \{0, 1\}$ for each $v \in V(\tilde{G})$, indicating which vertices were sampled. We refer to such bicolored graphs as *neighborhood subgraphs*, which is a union of stars with the root vertex of each star colored. This model is also known in the literature as ego-centric [131] or star sampling [130, 172].

In other words, we randomly sample the rows of the adjacency matrix of G independently with probability p and then observe the rows together with the row indices. The graph then consists of unions of star graphs (not necessarily disjoint) together with colors indicating the root of the stars. Neighborhood sampling operates like subgraph sampling but neighborhood information is acquired for each sampled vertex. Hence neighborhood sampling is more informative in the sense that, upon sampling the same set of vertices, considerably more edges are observed. For an illustration and comparison of both subgraph and neighborhood sampling, see Fig. 8.1. Thus it is reasonable to expect (and indeed we will prove in the sequel) that for the same statistical task, neighborhood sampling typically has significantly lower sample complexity than the subgraph sampling scheme. Note that in many cases, neighborhood sampling is more realistic than subgraph sampling (e.g., social network crawling), where sampling a vertex means that its immediate connections (e.g., friends list) are obtained for free.

A more general version of the neighborhood sampling model is described by Lovász in [137, Section 1.7], where each sample consists of a radius- r (labeled) neighborhood rooted at a randomly chosen vertex. Since from a union of marked stars one can disassemble each star individually, our model is equivalent to this one with $r = 1$.

It turns out that the knowledge of the colors provides crucial information about the sampled graph and affects the quality of estimation (see Appendix 8.9). In practice, the model with labels is more realistic since the experimenter would know which nodes were sampled. We henceforth assume that all sampled graphs obtained from neighborhood sampling are bicolored, with black and white vertices corresponding to sampled and non-sampled vertices, respectively. For a neighborhood subgraph h , let $V_b(h)$ (resp. $v_b(h)$) denote the collection (resp. number) of black vertices. Suppose H is a bicolored subgraph of G . Let $N(H, G)$ be the number of ways that H can appear (isomorphic as a vertex-colored graph) in G from neighborhood sampling with $v_b(H)$ vertices. Thus,

$$\mathbb{P}[\tilde{G} \cong H] = N(H, G) p^{v_b(H)} q^{v(G) - v_b(H)}.$$

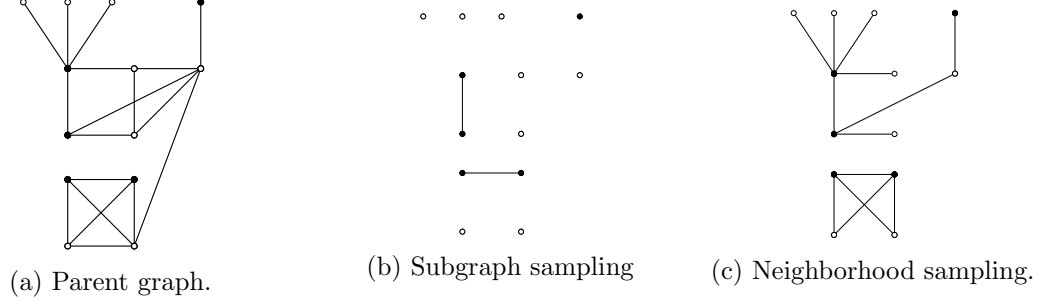


Figure 8.1: A comparison of subgraph and neighborhood sampling: Five vertices are sampled in the parent graph, and the observed graph is shown in Fig. 8.1b and Fig. 8.1c for the subgraph and neighborhood sampling, respectively.

8.1.2 Main results

Let h denote a motif, which is a connected graph on k vertices. As mentioned earlier, we do not assume any generative model or additional structures on the parent graph G , except that the maximal degree is at most d ; this parameter, however, need not be bounded, and one of the goals is to understand how the sample complexity depends on d . The goal is to estimate the motif count $\mathbf{s}(h, G)$ based on the sampled graph \tilde{G} obtained from either subgraph or neighborhood sampling.

Methodologically speaking, Horvitz-Thompson (HT) estimator [145] is perhaps the most natural idea to apply here. The HT estimator is an unbiased estimator of the population total by weighting the empirical count of a given item by the inverse of the probability of observing said item. To be precise, consider estimate the edge count in a graph with m edges and maximal degree d , the sampling ratio required by the HT estimator to achieve a relative error of ϵ scales as $\Theta(\max\{\frac{1}{\sqrt{m\epsilon}}, \frac{d}{m\epsilon^2}\})$, which turns out to be minimax optimal. For ϵ being a small constant, this yields a sublinear sample complexity when m is large and $m \gg d$.

For neighborhood sampling, which is more informative than subgraph sampling since more edges are observed, we show that the optimal sampling ratio can be improved to $\Theta(\min\{\frac{1}{\sqrt{m\epsilon}}, \frac{d}{m\epsilon^2}\})$, which, perhaps surprisingly, is not always achieved by the HT estimator. The main reason for its suboptimality in the high degree regime is the correlation between observed edges. To reduce correlation, we propose a family of linear estimators encompassing and outperforming the Horvitz-Thompson estimator. The key idea is to use

the color information indicating which vertices are sampled. For example, in a neighborhood sampled graph it is possible to observe two types of edges: $\bullet-\circ$ and $\bullet-\bullet$. The estimator takes a linear combination of the count of these two types of edges with a negative weight on the latter, which, as counterintuitive as it sounds, significantly reduces the variance and achieves the optimal sample complexity.

For general motifs h on k vertices, for subgraph sampling, it turns out the simple HT scheme for estimating $s = \mathbf{s}(h, G)$ achieves a multiplicative error of ϵ with the optimal sampling fraction

$$\Theta_k \left(\max \left\{ \frac{1}{(s\epsilon^2)^{\frac{1}{k}}}, \frac{d^{k-1}}{s\epsilon^2} \right\} \right),$$

which only depends on the size of the motif but *not* its actual topology. For neighborhood sampling, the situation is more complicated and the picture is less complete. For general h , we propose a family of estimators that achieves the sample ratio:

$$\Theta_k \left(\min \left\{ \left(\frac{d}{s\epsilon^2} \right)^{\frac{1}{k-1}}, \sqrt{\frac{d^{k-2}}{s\epsilon^2}} \right\} \right)$$

which again only depends on the size of h . We conjecture that this is optimal for neighborhood sampling and we indeed prove this for (a) all motifs up to 4 vertices; (b) cliques of all sizes.

Let us conclude this part by providing some intuition on proving the impossibility results. The main apparatus is matching subgraph counts: If two graphs have matching subgraph counts for all induced (resp. neighborhood) subgraphs up to size k , then the total variation of the sampled versions obtained from subgraph (resp. neighborhood) sampling are at $O(p^k)$. At a high level, this idea is akin to the method of moment matching, which have been widely used to prove statistical lower bound for functional estimation [173–176]; in comparison, in the graph-theoretic context, moments correspond to graph homomorphism numbers which are indexed by subgraphs instead of integers [177]. To give a concrete example, consider the triangle motif and take

$$H = \triangle \quad H' = \square \quad (8.3)$$

which have matching subgraph counts up to size two (equal number of vertices and edges)

but distinct number of triangles. Then with subgraph sampling, the sampled graph satisfies $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = O(p^3)$. For neighborhood sampling, we can take

$$H = \triangle - \circ - \circ \quad H' = \square - \circ \quad (8.4)$$

which have *matching degree sequences* $(3, 2, 2, 2, 1)$ but distinct number of triangles. In general, these pairs of graphs can be either shown to exist by the strong independence of graph homomorphism numbers for connected subgraphs [153] or explicitly constructed by a linear algebra argument [178]; however, for neighborhood sampling it is significantly more involved as we need to relate the neighborhood subgraph counts to the injective graph homomorphism numbers. Based on these small pairs of graphs, the lower bound in general is constructed by using either H or H' as its connected components.

8.1.3 Notations

We use standard big- O notations, e.g., for any positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for some absolute constant $C > 0$, $a_n = o(b_n)$ or $a_n \ll b_n$ or if $\lim a_n/b_n = 0$. Furthermore, the subscript in $a_n = O_r(b_n)$ indicates that $a_n \leq C_r b_n$ for some constant C_r depending on r only. For nonnegative integer k , let $[k] = \{1, \dots, k\}$.

Next we establish some graph-theoretic notations that will be used throughout the chapter. Let $G = (V, E)$ be a simple, undirected graph. Let $\mathbf{e} = \mathbf{e}(G) = |E(G)|$ denote the number of edges, $\mathbf{v} = \mathbf{v}(G) = |V(G)|$ denote the number of vertices, and $\mathbf{cc} = \mathbf{cc}(G)$ be the number of connected components in G . The open neighborhood of a vertex u is denoted by $N_G(u) = \{v \in V(G) : \{u, v\} \in E(G)\}$. The closed neighborhood is defined by $N_G[u] = \{u\} \cup N_G(u)$. Two vertices u and v are said to be adjacent to each other, denoted by $u \sim v$, if $\{u, v\} \in E(G)$.

Two graphs G and G' are isomorphic, denoted by $G \simeq G'$, if there exists a bijection between the vertex sets of G and G' that preserves adjacency, i.e., if there exists a bijective function $g : V(G) \rightarrow V(G')$ such that $\{g(u), g(v)\} \in E(G')$ whenever $\{u, v\} \in E(G)$. If G and G' are vertex-colored graphs with colorings c and c' (i.e., a function that assigns a color to each vertex), then G and G' are isomorphic as vertex-colored graphs, denoted

tion 8.2.1) and neighborhood (Section 8.2.2) sampling. Section 8.3 discusses converse results and states counterpart minimax lower bounds for subgraph (Section 8.3.2) and neighborhood (Section 8.3.3) sampling. We further restrict the class of graphs to be acyclic or planar in Section 8.4 and explore whether such additional structure can be exploited to improve the quality of estimation. In Section 8.5, we perform a numerical study of the proposed estimators for counting edges, triangles, and wedges on both simulated and real-world data. Finally, in Appendix 8.7, we prove some of the auxiliary lemmas and theorems that were stated in the main body of the chapter.

8.2 Methodologies and performance guarantees

8.2.1 Subgraph sampling

The motivation for our estimation scheme is based on the observation that any motif count $s(h, G)$ can be written as a sum of indicator functions as in (8.5). Note that for a fixed subset of vertices $T \subset V(G)$, the probability it induces a subgraph in the sampled graph \tilde{G} that is isomorphic to h is

$$\mathbb{P}[\tilde{G}[T] \simeq h] = p^{v(h)} \mathbb{1}\{G[T] \simeq h\}.$$

In view of (8.5), this suggests the following unbiased estimator of $s(h, G)$:

$$\hat{s}_h \triangleq s(h, \tilde{G})/p^{v(h)}. \tag{8.7}$$

We refer to this estimator as the Horvitz-Thompson (HT) estimator [145] since it also uses inverse probability weighting to achieve unbiasedness. The next theorem gives an upper bound on the mean-squared error for this simple scheme, which, somewhat surprisingly, turns out to be minimax optimal within a constant factor as long as the motif h is *connected*.

Theorem 35 (Subgraph sampling). *Let h be an arbitrary connected graph with k vertices. Let G be a graph with maximum degree at most d . Consider the subgraph sampling model*

with sampling ratio p . Then the estimator (8.7) satisfies

$$\mathbb{E}_G |\widehat{s}_h - s(h, G)|^2 \leq s(h, G) \cdot k 2^k \left(\frac{1}{p^k} \vee \frac{d^{k-1}}{p} \right).$$

Furthermore,

$$\inf_{\widetilde{s}} \sup_{\substack{G: d(G) \leq d \\ s(h, G) \leq s}} \mathbb{E}_G |\widetilde{s} - s(h, G)|^2 = \Theta_k \left(\left(\frac{s}{p^k} \vee \frac{s d^{k-1}}{p} \right) \wedge s^2 \right).$$

The above result establishes the optimality of the HT estimator for classes of graphs with degree constraints. Since the lower bound construction actually uses instances of graphs containing many cycles, it is a priori unclear whether additional assumptions such as tree structures can help. Indeed, for the related problem of estimating the number of connected components with subgraph sampling, it has been shown that for parent graphs that are forests the sample complexity is strictly smaller [17]. Nevertheless, for counting motifs such as edges or wedges, in Theorem 41 and Theorem 43 we show that the HT estimator (8.7) cannot be improved up to constant factors even if the parent graph is known to be a forest.

The proof of the lower bound of Theorem 35 is given in Section 8.3.2. Below we prove the upper bound of the variance:

Proof. Since \widehat{s} is unbiased, it remains to bound its variance. Let $b_v \triangleq \mathbb{1}\{v \in S\}$, which are iid as $\text{Bern}(p)$. For any $T \subset V(G)$, let $b_T \triangleq \prod_{v \in T} b_v$. Then

$$\widehat{s} = p^{-k} \sum_{T \subset V(G)} b_T \mathbb{1}\{G[T] \simeq h\}.$$

Hence

$$\begin{aligned}
\text{Var}[\widehat{\mathbf{s}}] &= p^{-2k} \sum_{T \cap T' \neq \emptyset} \text{Cov}(b_T, b_{T'}) \mathbb{1} \{G[T] \simeq h, G[T'] \simeq h\} \\
&\leq p^{-2k} \sum_{T \cap T' \neq \emptyset} \mathbb{E}[b_{T \cup T'}] \mathbb{1} \{G[T] \simeq h, G[T'] \simeq h\} \\
&= \sum_{t=1}^k p^{-t} \sum_{|T \cap T'|=t} \mathbb{1} \{G[T] \simeq h, G[T'] \simeq h\} \\
&\leq \sum_{t=1}^k p^{-t} \mathbf{s}(h, G) \binom{k}{t} d^{k-t} \leq \mathbf{s}(h, G) (2d)^k \cdot k \max\{(pd)^{-k}, (pd)^{-1}\},
\end{aligned}$$

where the penultimate step follows from the fact that the maximum degree of G is d and, crucially, h is connected. \square

8.2.2 Neighborhood sampling

Our methodology is again motivated by (8.5) which represents neighborhood subgraph counts as a sum of indicators. In contrast to subgraph sampling, a motif can be observed in the sampled graph by sampling only some, but not all, of its vertices. For example, we only need to sample one vertex of an edge, or two vertices of a triangle to observe the full motif. More generally, for a subset T vertices in G , we can determine whether $H \simeq G[T]$ or not if at least $\mathbf{v}(H) - 1$ vertices from T are sampled. This reduces the variance but introduces more correlation at the same time.

Throughout this subsection, the neighborhood sampled graph is again denoted by $\tilde{G} = G\{S\}$, and $b_v = \mathbb{1} \{v \in S\}$ indicates whether a given vertex v is sampled.

Edges

We begin by discussing the Horvitz-Thompson type of estimator and why it falls short for the neighborhood sampling model. Analogously to the estimator (8.7) designed for subgraph sampling, for neighborhood sampling, we can take the observed number of edges and re-weight it according to the probability of observing an edge. Note that with neighborhood sampling, a given edge is observed if and only if at least one of the end points is sampled.

Thus, the corresponding Horvitz-Thompson type edge estimator is

$$\hat{e}_{\text{HT}} = \frac{e(\tilde{G})}{p^2 + 2pq}, \quad (8.8)$$

which is again an unbiased estimator for $e(G)$. To bound the variance, put $\tau = p^2 + 2pq \in [p, 2p]$ and write

$$e(\tilde{G}) = \sum_{A \in E(G)} r_A.$$

where $A = \{u, v\}$ and $r_A \triangleq \mathbb{1}\{b_u = 1 \text{ or } b_v = 1\} \sim \text{Bern}(\tau)$. For another edge $A' = \{v, w\}$ intersecting A , we have $\text{Cov}[r_A, r_{A'}] = \mathbb{P}[b_v = 1 \text{ or } b_u = b_w = 1] \leq 3p$, by the union bound. Thus the number of non-zero covariance terms is determined by $n(\text{⌔}, G)$, the number of ⌔ contained in G as subgraphs, and we have

$$\text{Var}[e(\tilde{G})] \leq e(G)\tau + 2n(\text{⌔}, G)(3p) \leq 2e(G)p(1 + 3d), \quad (8.9)$$

where we used the fact that $n(\text{⌔}, G) \leq e(G)d$. Therefore, the variance of the Horvitz-Thompson estimator satisfies

$$\text{Var}[\hat{e}_{\text{HT}}] \lesssim \frac{e(G)d}{p}. \quad (8.10)$$

However, as we show next, this estimator is suboptimal when $p > \frac{1}{d}$, or equivalently, when the maximum degree exceeds $\frac{1}{p}$. In fact, the bound (8.10) itself is tight which can be seen by considering a star graph G with d leaves, and the suboptimality of the HT estimator is largely due to the *heavy correlation* between the observed edges. For example, for the star graph, the correlation is introduced through the root vertex, since with probability p we observe a full star, and with probability q a star with $\text{Bin}(d, p)$ number of black leaves. Thus, the key observation is to incorporate the colors of the vertices to reduce (or eliminate) correlation.

Next, we describe a class of estimators, encompassing and improving the Horvitz-Thompson estimator. Consider

$$\hat{e} = \sum_{A \in E(\tilde{G})} \mathcal{K}_A, \quad (8.11)$$

where \mathcal{K}_A has the form

$$\mathcal{K}_A = b_u(1 - b_v)f(d_u) + b_v(1 - b_u)f(d_v) + b_ub_vg(d_u, d_v); \quad (8.12)$$

here $A = \{u, v\}$ and f and g are functions of the degree of sampled vertices. For the neighborhood sampling model, this estimator is well-defined since the degree of any sampled vertex is observed without error. It is easy to see that





$$\mathbb{E}[\hat{\mathbf{e}}] = \sum_{\{u,v\} \in E(G)} [pq(f(d_u) + f(d_v)) + p^2g(d_u, d_v)]. \quad (8.13)$$

For simplicity, next we choose f and g to be constant; in other words, we do not use the degree information of the sampled vertices. This strategy works as long as the maximal degree d of the parent graph is known. To illustrate the main idea, we postpone the discussion on adapting to the unknown d to Section 8.2.2. With $f \equiv \alpha$ and $g \equiv \beta$, the estimator (8.11) reduces to

$$\hat{\mathbf{e}} = \alpha \mathbf{N}(\bullet \circ, \tilde{G}) + \beta \mathbf{N}(\bullet \bullet, \tilde{G}), \quad (8.14)$$

which is a linear combination of the counts of the two types of observed edges. In contrast to the HT estimator (8.8) which treats the two types of edges equally, the optimal choice will weigh them differently. Furthermore, somewhat counter-intuitively, the weights can be negative, which serves to reduce the correlation.

Table 8.1: Probability mass function of $\mathcal{K}_A\mathcal{K}_{A'}$ for two distinct intersecting edges (excluding zero values).

Graph				
Probability	pq^2	$2p^2q$	p^2q	p^3
Value	α^2	$\alpha\beta$	α^2	β^2

In view of (8.13), one way of making $\hat{\mathbf{e}}$ unbiased is to set

$$pq(f(d_u) + f(d_v)) + p^2g(d_u, d_v) = 2pq\alpha + p^2\beta = 1. \quad (8.15)$$

Since the unbiased estimator is not unique, we set out to find the one with the minimum variance. Similar to (8.9), we have

$$\text{Var}[\widehat{\mathbf{e}}] = \mathbf{e}(G)\text{Var}[\mathcal{K}_A] + 2\mathbf{n}(\curvearrowright, G)\text{Cov}[\mathcal{K}_A, \mathcal{K}_{A'}] \leq \mathbf{e}(G)(\text{Var}[\mathcal{K}_A] + 2d\text{Cov}[\mathcal{K}_A, \mathcal{K}_{A'}]), \quad (8.16)$$

where $A = \{u, v\}$ and $A' = \{v, w\}$ are distinct intersecting edges. Using Table 8.1, we find

$$\begin{aligned} \text{Var}[\mathcal{K}_A] &= 2pq\alpha^2 + p^2\beta^2 - 1 \\ \text{Cov}[\mathcal{K}_A, \mathcal{K}_{A'}] &= \alpha^2(pq^2 + p^2q) + p^3\beta^2 + 2p^2q\alpha\beta - 1. \end{aligned}$$

In fact, when the unbiased condition (8.15) is met, the covariance simplifies to $\text{Cov}[\mathcal{K}_A, \mathcal{K}_{A'}] = \frac{q}{p}(1 - p\alpha)^2 \geq 0$. Finally, optimizing the RHS of (8.16) over α, β subject to the constraint (8.15), we arrive the following performance guarantee for $\widehat{\mathbf{e}}$:

Theorem 36. *Set*

$$\alpha = \frac{1 + dp}{p(2 + (d - 1)p)} \quad \beta = \frac{1 - d(1 - 2p)}{p(2 + (d - 1)p)}. \quad (8.17)$$

Then

$$\mathbb{E}[(\widehat{\mathbf{e}} - \mathbf{e}(G))^2] = \text{Var}[\widehat{\mathbf{e}}] \leq \frac{\mathbf{e}(G)(d + 1)q^2}{p(2 + (d - 1)p)} \lesssim \mathbf{e}(G) \left(\frac{1}{p^2} \wedge \frac{d}{p} \right). \quad (8.18)$$

Furthermore, if p is bounded from one, then

$$\inf_{\widehat{\mathbf{e}}} \sup_{\substack{G: d(G) \leq d \\ \mathbf{e}(h, G) \leq m}} \mathbb{E}_G |\widehat{\mathbf{e}} - \mathbf{e}(G)|^2 = \Theta \left(\left(\frac{md}{p} \wedge \frac{m}{p^2} \right) \wedge m^2 \right)$$

The optimal weights in (8.17) appear somewhat mysterious. In fact, the following more transparent choice also achieves the optimal risk within constant factors:

- $p \leq 1/d$: we can set either $\alpha = \beta = \frac{1}{p^2 + 2pq}$ or $\alpha = \frac{1}{2pq}$ and $\beta = 0$, that is, we can use either the full HT estimator (8.8), or the HT estimator restricted to only edges of type $\bullet \text{---} \circ$, which is the more probable one.
- $p > 1/d$: we choose $\alpha = \frac{1}{p}$ and $\beta = \frac{1 - 2q}{p^2}$. This is the unique weights that simulta-

neously kill all covariance terms and, at the same time, achieve zero bias. Note that although zero covariance is always possible, it is at a price of setting $\beta \approx -\frac{1}{p^2}$, which inflates the variance too much when p is small and hence suboptimal when $p \ll \frac{1}{d}$.

It is a priori unclear whether additional structures such as tree or planarity helps for estimating motif counts with neighborhood sampling. Nevertheless, for counting edges, in Theorem 42 we show that the Horvitz-Thompson estimator (8.7) can only be marginally improved, in the sense that the lower bound continues to hold up to a sub-polynomial factor $p^{o(1)}$ where $o(1)$ is uniformly vanishing as $p \rightarrow 0$. Similarly, for planar graphs, Theorem 47 shows a similar statement.

Cliques and general motifs

For ease of exposition, we start by developing the methodology for estimating cliques counts. Both the procedure and the performance guarantee readily extend to general motifs.

We now generalize the techniques for counting edges to estimate the number of cliques of size ω in a given graph. Note that there are two types of colored cliques one observe: (a) K_ω° : all but one vertex are sampled; (b) K_ω^\bullet : all vertices are sampled, with the first one being more probable when the sampling ratio is small. In the case of triangles, we have $K_3^\circ = \triangleleft$ and $K_3^\bullet = \blacktriangleleft$. Analogous to the estimator (8.14), we take a linear combination of these two types of clique counts as the linear estimator:

$$\hat{s} = \alpha N(K_\omega^\circ, \tilde{G}) + \beta N(K_\omega^\bullet, \tilde{G}). \quad (8.19)$$

Similar to the design principles for counting edges, in the low sampling ratio regime $p < \frac{1}{d}$, we implement the Horvitz-Thompson estimator, so that the coefficients scale like $p^{-\omega}$; in the high sampling ratio regime $p > \frac{1}{d}$, we choose a *negative* β , which scale as $-p^{-2\omega}$, to reduce the correlation between various observed cliques. However, unlike the case of counting edges, we cannot perfectly eliminate all covariance terms but will be able to remove the leading one.

The following result, which includes Theorem 36 as a special case ($\omega = 2$), gives the performance guarantee of the estimator (8.19) and establishes its optimality in the worst

case:

Theorem 37 (Cliques). *Set*

$$\begin{cases} \alpha = \frac{1}{p^{\omega-1}}, & \beta = \frac{1-\omega q}{p^\omega} & \text{if } p > 1/d \\ \alpha = \frac{1}{\omega p^{\omega-1}}, & \beta = \frac{1}{p^\omega} & \text{if } p \leq 1/d. \end{cases} \quad (8.20)$$

Then

$$\mathbb{E}_G |\widehat{\mathbf{s}} - \mathbf{s}(K_\omega, G)|^2 = \text{Var}_G[\widehat{\mathbf{s}}] \leq \mathbf{s}(K_\omega, G) \cdot \omega^3 2^{\omega+1} \left(\frac{\mathbf{d}(G)}{p^{\omega-1}} \wedge \frac{\mathbf{d}(G)^{\omega-2}}{p^2} \right).$$

Furthermore,

$$\inf_{\widehat{\mathbf{s}}} \sup_{\substack{G: \mathbf{d}(G) \leq d \\ \mathbf{s}(K_\omega, G) \leq s}} \mathbb{E}_G |\widehat{\mathbf{s}} - \mathbf{s}(K_\omega, G)|^2 = \Theta_\omega \left(\frac{sd}{p^{\omega-1}} \wedge \frac{sd^{\omega-2}}{p^2} \wedge s^2 \right)$$

Proof. Let $b_v \triangleq \mathbb{1}\{v \in S\} \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p)$. For any $T \subset V(G)$, let $b_T \triangleq \prod_{v \in T} b_v$. Write

$$\begin{aligned} \widehat{\mathbf{s}} &= \sum_{T \subset V(G)} \alpha \mathbb{1}\{\widetilde{G}\{T\} \simeq K_\omega^\circ\} + \beta \mathbb{1}\{\widetilde{G}\{T\} \simeq K_\omega^\bullet\} \\ &= \sum_{T \subset V(G)} f(T) \mathbb{1}\{G[T] \simeq K_\omega\}, \end{aligned} \quad (8.21)$$

where

$$f(T) \triangleq \alpha \sum_{v \in T} b_{T \setminus \{v\}} (1 - b_v) + \beta b_T.$$

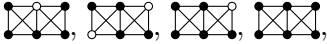
Similar to (8.15), enforcing unbiasedness, we have the constraint $\mathbb{E}[f(T)] = 1$, i.e.,

$$\omega p^{\omega-1} q \alpha + p^\omega \beta = 1 \quad (8.22)$$

Furthermore, whenever $|T \cap T'| = t \in [\omega]$, we have

$$\mathbb{E}[f(T)f(T')] = \alpha^2 (tqp^{2\omega-t-1} + (\omega-t)^2 q^2 p^{2\omega-t-2}) + 2(\omega-t)qp^{2\omega-t-1}\alpha\beta + \beta^2 p^{2\omega-t} \quad (8.23)$$

$$\begin{aligned} &= p^{-t} \left[\alpha^2 tqp^{2\omega-1} + (\alpha(\omega-t)qp^{\omega-1} + \beta p^\omega)^2 \right] \\ &\stackrel{(8.22)}{=} p^{-t} \left[\alpha^2 tqp^{2\omega-1} + (1 - tq\alpha p^{\omega-1})^2 \right] \end{aligned} \quad (8.24)$$

This follows from evaluating the probability of observing a pair of intersecting cliques with two, one, or zero unsampled vertices. For example, the four summands in (8.23), in the case of $\omega = 4$ and $t = 2$, correspond to , respectively.

Let $c_t \triangleq \text{Cov}[f(T), f(T')] = p^{-t} \left[\alpha^2 tqp^{2\omega-1} + (\alpha(\omega-t)qp^{\omega-1} + \beta p^\omega)^2 \right] - 1$ for $|T \cap T'| = t$. Denote by $T_{\omega,t}$ the subgraph correspond to two intersecting ω -cliques sharing t vertices. Then

$$\begin{aligned} \text{Var}[\widehat{S}] &= \sum_{T \cap T' \neq \emptyset} \text{Cov}(f(T), f(T')) \mathbb{1} \{G[T] \simeq K_\omega, G[T'] \simeq K_\omega\} \\ &= \sum_{t=1}^{\omega} c_t \sum_{|T \cap T'|=t} \mathbf{n}(T_{\omega,t}, G) \leq \mathbf{s}(K_\omega, G) d^\omega \sum_{t=1}^{\omega} c_t \binom{\omega}{t} d^{-t}. \end{aligned} \quad (8.25)$$

Next consider two cases separately.

Case I: $p \leq \frac{1}{d}$. In this case we choose $\alpha = \frac{1}{\omega p^{\omega-1}}$ and $\beta = \frac{1}{p^\omega}$. Then $c_t = p^{-t} \left(\frac{tpq}{\omega^2} + (1 - \frac{tq}{\omega})^2 \right) \leq 2p^{-t}$. Furthermore, for the special case of $t = \omega$, we have $c_\omega \leq p^{-(\omega-1)}$. Thus,

$$\text{Var}[\widehat{S}] \leq \mathbf{s}(K_\omega, G) \left(d^\omega \sum_{t=1}^{\omega} \binom{\omega}{t} (pd)^{-t} + p^{-(\omega-1)} \right) \leq \mathbf{s}(K_\omega, G) \omega 2^{\omega+1} dp^{-(\omega-1)}. \quad (8.26)$$

Case II: $p \leq \frac{1}{d}$. In this high-degree regime, the pairs of cliques sharing one vertex ($t = 1$) dominates (i.e., open triangle for counting edge and bowties for counting triangles). Thus our strategy is to choose the coefficients to eliminate the these covariance terms. In fact,

(8.24) for $t = 1$ simplifies wonderfully to

$$c_1 = \frac{q}{p}(1 - \alpha p^{\omega-1})^2.$$

Thus we choose $\alpha = \frac{1}{p^{\omega-1}}$ and $\beta = \frac{1-\omega q}{p^\omega}$. Hence $c_t \leq 2\omega^2 p^{-t}$ for all $t \geq 2$, and

$$\text{Var}[\widehat{\mathbf{s}}] \leq \mathbf{s}(K_\omega, G) d^\omega 2\omega^2 \sum_{t=2}^{\omega} \binom{\omega}{t} (pd)^{-t} \leq \mathbf{s}(K_\omega, G) 2^{\omega+1} \omega^3 d^{\omega-2} p^{-2}. \quad (8.27)$$

Combining (8.26) and (8.27) completes the proof. \square

To extend to general motif h on k vertices, note that in the neighborhood sampled graph, again it is possible to observe fully sampled or partially sampled (with one unsampled vertices) motifs. Consider the following estimator analogous to (8.19):

$$\widehat{\mathbf{s}}_h = \alpha \mathbf{N}(h^\circ, \widetilde{G}) + \beta \mathbf{N}(h^\bullet, \widetilde{G}), \quad (8.28)$$

where $\mathbf{N}(h^\circ, \widetilde{G})$ is the count of h with all vertices sampled and $\mathbf{N}(h^\bullet, \widetilde{G})$ is the total count of h with exactly one unsampled vertices. For instance, if $h = \text{triangle}$, then $\mathbf{N}(h^\bullet, \widetilde{G}) = \mathbf{N}(\text{triangle}, G)$ and $\mathbf{N}(h^\circ, \widetilde{G}) = \mathbf{N}(\text{triangle}, G) + \mathbf{N}(\text{triangle}, G)$. This example shows that in general, for motifs with less symmetry, there exist multiple partially sampled motifs and in principle they can be weighted differently. However, in (8.28) we elect to treat them equally, which turns out to be optimal for a wide class of motifs. Let us point out that if the parent graph has more structures, e.g., forest, then distinguishing different partially sampled motifs can lead to strict improvement; see Theorem 44.

The estimator (8.28) turns out to satisfy the same bound as in the clique case. To see this, note that in (8.25), the covariance terms are given in (8.24) which do not depend on the actual motif h . Furthermore, the sum of the indicators satisfies the same bound in terms of maximal degree provided that h is connected. Using the same optimized coefficients as in (8.20), the guarantee in Theorem 37 holds verbatim:

$$\mathbb{E}_G |\widehat{\mathbf{s}}_h - \mathbf{s}(h, G)|^2 = \text{Var}_G[\widehat{\mathbf{s}}_h] \leq \mathbf{s}(h, G) \cdot k^3 2^{k+1} \left(\frac{d(G)}{p^{k-1}} \wedge \frac{d(G)^{k-2}}{p^2} \right). \quad (8.29)$$

We conjecture that, similar to Theorem 35, this rate is optimal as long as the motif h is connected. So far we are able to prove this for cliques of all sizes (Theorem 40) and motifs on at most 4 vertices (Appendix 8.10).

Adaptation to the maximum degree

In practice, the bound on the maximum degree d is likely unknown to the observer and obtaining a consistent estimate might be difficult if the high-degree vertices are rare. For example, in a star, most of the vertices have degree one except for the root. Even if a consistent estimate is obtained, it is unclear how to avoid it correlating with the data used to form \hat{e} . Because \hat{e} has the form of a sum, such correlations increase the number of cross terms in its variance decomposition.

To overcome these difficulties, we weight each observed edge according to the size of the neighborhood of its incident vertices. Once a vertex is sampled, its degree is exactly determined and thus incorporating this information does not introduce any additional randomness. This observations leads to the following adaptive estimator which achieves a risk that is similar to the optimal risk in Theorem 36:

Theorem 38. *Let \hat{e} be given in (8.11) with $f(x) = \frac{px+q}{p(px+2q)}$ and $g(x, y) = \frac{1-pq(f(x)+f(y))}{p^2}$. Then for any graph G on N vertices and maximum degree bounded by d , \hat{e} is an unbiased estimator of $e(G)$ and*

$$\text{Var}[\hat{e}] \lesssim \frac{Nd}{p^2} \wedge \frac{e(G)d}{p}.$$

Remark 16. *The variance bound from Theorem 38 is weaker than Theorem 36 in the $p > 1/d$ regime – $\frac{Nd}{p^2}$ versus $\frac{e(G)}{p^2}$. They have the largest disparity when G consists of $N/(d+1)$ copies of the star graph S_{d+1} , in which case $e(G) = Nd/(d+1)$. This is due to the fact that with high probability $1-p$, all sampled vertices from S_{d+1} have degree one. Ideally, we would like to know the degree of the root of the star; however this is impossible unless the root is sampled. Nonetheless, we can still find a good estimate. More generally, in addition to using the degree d_u from a sampled vertex u , we may modify the estimator to incorporate degree information from a non-sampled vertex via an unbiased estimate, i.e.,*

$\widehat{d}_u = \frac{|N_{\widetilde{G}}(u)|}{p} = \sum_{v \in N_G(u)} \frac{b_v}{p}$. For example, we can redefine \mathcal{K}_A from (8.11) as

$$\mathcal{K}_A = b_u(1 - b_v)f(d_u \vee \widehat{d}_v) + b_v(1 - b_u)f(d_u \vee \widehat{d}_v) + b_u b_v g(d_u, d_v).$$

8.3 Lower bounds

Throughout this section we assume that the sampling ratio p is bounded away from one.

8.3.1 Auxiliary results

We start with a result which is the general strategy of proving all lower bounds in this chapter. A variant of this result was proved in [17] for the Bernoulli sampling model, however, an examination of the proof reveals that the conclusions also hold for neighborhood sampling. In the context of estimating motif counts, the essential ingredients involve constructing a pair of random graphs whose motif counts have different average values, and the distributions of their sampled versions are close in total variation, which is ensured by matching lower-order subgraphs counts in terms of \mathbf{s} for subgraph sampling or \mathbf{N} for neighborhood sampling. The utility of this result is to use a pair of smaller graphs (which can be found in an ad hoc manner) to construct a bigger pair of graphs and produce a lower bound that scales with an arbitrary positive integer s .

Theorem 39 (Theorem 11 in [17]). *Let f be a graph parameter that is invariant under isomorphisms and additive under disjoint union, i.e., $f(G + H) = f(G) + f(H)$. Fix a subgraph h . Let d, s, m and $M = s/m$ be integers. Let H and H' be two graphs such that $\mathbf{s}(h, H) \vee \mathbf{s}(h, H') \leq m$ and $\mathbf{d}(H) \vee \mathbf{d}(H') \leq d$. Suppose $M \geq 300$ and $\text{TV}(P, P') \leq 1/300$, where P (resp. P') denote the distribution of the isomorphism class of the (subgraph or neighborhood) sampled graph \widetilde{H} (resp. \widetilde{H}'). Let \widetilde{G} denote the sampled version of G under the Bernoulli or neighborhood sampling models with probability p . Then*

$$\inf_{\widehat{f}} \sup_{G: \substack{\mathbf{d}(G) \leq d \\ \mathbf{s}(h, G) \leq s}} \mathbb{P}_G \left[|\widehat{f}(\widetilde{G}) - f(G)| \geq \Delta \right] \geq 0.01. \quad (8.30)$$

where

$$\Delta = \frac{|f(H) - f(H')|}{8} \left(\sqrt{\frac{s}{m\text{TV}(P, P')}} \wedge \frac{s}{m} \right).$$

Next we recall the well-known fact [154, 155] that disconnected subgraphs counts are determined by (fixed polynomials of) connected subgraph counts. The following version is from [17, Corollary 1 and Lemma 9]:

Lemma 47. *Let H and H' be two graphs with m vertices and $v \leq m$. Suppose $\mathbf{s}(h, H) = \mathbf{s}(h, H')$ for all connected h with $\mathbf{v}(h) \leq v$. Then $\mathbf{s}(h, H) = \mathbf{s}(h, H')$ for all h with $\mathbf{v}(h) \leq v$ and, furthermore,*

$$\text{TV}(P, P') \leq \binom{m}{v+1} p^{v+1},$$

where P (resp. P') denote the distribution of the isomorphism class of the subgraph sampled graph \tilde{H} (resp. \tilde{H}') with sampling ratio p .

The following version is for neighborhood sampling, which will be used in the proof of Theorem 40. We need to develop an analogous result that expresses disconnected neighborhood subgraph counts as polynomials of the connected cones. This is done in Lemma 52 in Appendix 8.7.

Lemma 48. *Let H and H' be two graphs with m vertices and $v \leq m$. Suppose $\mathbf{N}(h, H) = \mathbf{N}(h, H')$ for all connected, bicolored h with $\mathbf{v}_b(h) \leq v$. Then*

$$\mathbf{N}(h, H) = \mathbf{N}(h, H') \tag{8.31}$$

for all h with $\mathbf{v}_b(h) \leq v$ and, furthermore,

$$\text{TV}(P, P') \leq \binom{m}{v+1} p^{v+1}, \tag{8.32}$$

where P (resp. P') denote the distribution of the isomorphism class of the sampled graph \tilde{H} (resp. \tilde{H}') generated from neighborhood sampling with sampling ratio p .

Proof. The first conclusion (8.31) follows from Lemma 52. For the second conclusion (8.32), we note that conditioned on ℓ vertices are sampled, \tilde{H} is uniformly distributed over the

collection of all bicolored neighborhood subgraphs h with $\mathbf{v}_b(h) = \ell$. Thus,

$$\mathbb{P} \left[\tilde{H} \cong h \mid \mathbf{v}_b(h) = \ell \right] = \frac{\mathbf{N}(h, H)}{\binom{m}{\ell}}.$$

By (8.31), we conclude that the isomorphism class of \tilde{H} and \tilde{H}' have the same distribution provided that no more than v vertices are sampled. Thus, $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq \mathbb{P}[\text{Bin}(m, p) \leq v + 1]$, and consequently, $\mathbb{P}[\text{Bin}(m, p) \leq v + 1] \leq \binom{m}{v+1} p^{v+1}$ follows from a union bound. \square

Lemma 49. *For any connected graph h with k vertices, there exists a pair of (in fact, connected) graphs H and H' , such that $\mathbf{s}(h, H) \neq \mathbf{s}(h, H')$ and $\mathbf{s}(g, H) = \mathbf{s}(g, H')$ for all connected g with $\mathbf{v}(g) \leq k - 1$.*

Proof. The existence of such a pair H and H' follows from the strong independence² of connected subgraph counts [153, Theorem 1]. For example, for $h = \triangle_{\circ}$, we can take the ad hoc construction in (8.3), which have equal number of vertices and edges but distinct number of triangles. Alternatively, next we provide an explicit construction using a linear algebra argument which is similar to that of [153, Theorem 3] and [178, Section 2]. Let $\{h_1, \dots, h_m\}$ denote all distinct (up to isomorphism) induced *connected* subgraph of h , ordered in increasing number of edges (arbitrarily among graphs with the same number of edges) so that h_1 is an isolated vertex and $h_m = h$. Then the matrix $B = (b_{ij})$ with $b_{ij} = \mathbf{s}(h_i, h_j)$ is upper triangular with strictly positive diagonals. Thus B is invertible and the entries of B^{-1} are rational. Let $x = B^{-1}e_m$, where $e_m = (0, \dots, 0, 1)$. Then $x_m = 1$ since $b_{mm} = \mathbf{s}(h, h) = 1$. Let $w = \alpha x \in \mathbb{Z}^m$, where $\alpha \in \mathbb{N}$ is the lowest common denominator of the entries of x . Now define H and H' as the disjoint union with weights given by the vector w :

$$H = \sum_{i=1}^m \max\{w_i, 0\} h_i, \quad H' = \sum_{i=1}^m \max\{-w_i, 0\} h_i. \quad (8.33)$$

By design, any connected induced subgraph of H and H' with at most $k - 1$ vertices belongs to $\{h_1, \dots, h_{m-1}\}$. For any $1 \leq i \leq m - 1$, since h_i is connected, we have $\mathbf{s}(h_i, H) -$

². This means that the closure of the range of their normalized version (subgraph densities) has nonempty interior.

$\mathbf{s}(h_i, H') = \sum_{j=1}^m w_j \mathbf{s}(h_i, h_j) = 0$, and $\mathbf{s}(h, H') = 0$ and $\mathbf{s}(h, H) = \alpha \geq 1$. For example, for $h = \text{triangle}$, such a solution is given by:

$$H = \text{triangle} + 6 \times \text{edge} \quad H' = 4 \times \text{triangle} + 4 \times \text{edge}$$

which have matching subgraphs of order three (vertices, edges and triangles). For a construction for general cliques, see [17, Eq. (47)]. \square

Next we present graph-theoretic results that are needed for proving lower bound under the neighborhood sampling model. First, we relate the neighborhood subgraph counts \mathbf{N} to the usual subgraph \mathbf{n} . Since \mathbf{N} is essentially subgraph counts with prescribed degree for the sampled vertices (cf. [137, p. 62]), this can be done by inclusion-exclusion principle similar to (8.1) that expresses the induced subgraph counts \mathbf{s} in terms of the subgraph counts \mathbf{n} ; however, the key difference here is that the size of the subgraphs that appear in the linear combination is not bounded a priori. For example,

$$\begin{aligned} \mathbf{N}(\text{degree-2 vertex}, G) &= \text{number of degree-2 vertices in } G \\ &= \sum_{k \geq 2} (-1)^{k-2} \binom{k}{2} \mathbf{n}(S_{k+1}, G), \end{aligned}$$

where S_{k+1} is the star graph with k leaves. The following lemma is a general statement:

Lemma 50. *Let h be a bicolored connected neighborhood graph and h_0 denote the uncolored version. Then for any G ,*

$$\mathbf{N}(h, G) = \sum_g c(g, h) \mathbf{n}(g, G) \tag{8.34}$$

where the sum is over all (uncolored) g obtained from h by either adding edges incident to the black vertices in h or adding vertices connected to black vertices in h . In particular, the coefficients $c(g, h)$ do not depend on G .

Proof. The proof is by the inclusion-exclusion principle and essentially similar to the argument in Section 5.2, in particular, the proof of Proposition 5.6(b) in [137].

Recall the definition of the subgraph count $\mathbf{n}(H, G)$ in (8.6) in terms of counting distinct subsets. It will be convenient to work with the labeled version counting graph homo-

morphisms. The following definitions are largely from [137, Chapter 5]. We say ψ is an injective homomorphism from H to G , if $\psi : V(H) \rightarrow V(G)$ is injective, and $(u, v) \in E(H)$ if $(\psi(u), \psi(v)) \in E(G)$. Denote by $\text{inj}(H, G)$ the number of injective homomorphisms from H to G . Then $\text{inj}(H, G) = \mathbf{n}(H, G)\text{aut}(H)$, where $\text{aut}(H)$ denotes the number of automorphisms (i.e. isomorphisms to itself) for H . Furthermore, for neighborhood subgraph h , $\text{aut}(h)$ denotes the number of automorphisms for h that also preserve the colors. For example, $\text{aut}(\bullet \square \bullet) = 2$ and $\text{aut}(\circ \square \circ) = 4$. Throughout the proof, ψ always denotes an injection.

We use the following version of the inclusion-exclusion principle [137, Appendix A.1]. Let S be a ground set and let $\{A_i : i \in S\}$ be a collection of sets. For each $I \subset S$, define $A_I \triangleq \cap_{i \in I} A_i$ and $B_I \triangleq A_I \setminus \cup_{i \notin I} A_i$; in words, B_I denotes those elements that belong to exactly those A_i for $i \in I$ and none other. Then we have

$$|A_I| = \sum_{J \subset I} |B_J| \quad (8.35)$$

$$|B_I| = \sum_{J \subset I} (-1)^{|J|-|I|} |A_J|. \quad (8.36)$$

Fix G . Let \mathcal{G} denote the collection of (uncolored) subgraphs that are “extensions” of h , obtained from h by either adding edges between the black vertices in h or adding vertices attached to black vertices in h . For example, for $h = \bullet \circ$, we have $\mathcal{G} = \{\bullet \circ, \bullet \circ \circ, \bullet \circ \circ \circ, \bullet \circ \circ \circ \circ, \dots\}$ is the collection of all stars. Let the g^* be the maximal subgraph of G that is in \mathcal{G} ; in other words, $\mathbf{n}(g, G) = 0$, for any other $g \in \mathcal{G}$ containing g^* as a subgraph.

Now we define the ground set to be the edge set of g^* . Let h_0 be the uncolored version of h , then $E(h_0) \subset E(g^*)$. For every $I \subset E(g^*)$, define $A_I \triangleq \{\psi : V(g^*) \rightarrow V(g) : (\psi(u), \psi(v)) \in E(G) \text{ if } (u, v) \in I\}$ and $B_I \triangleq \{\psi : V(g^*) \rightarrow V(g) : (\psi(u), \psi(v)) \in E(G) \text{ if and only if } (u, v) \in I\}$. The key observation is that $|B_{E(h_0)}| = \text{aut}(h)\mathbf{N}(h, G)$, and $|A_{E(g)}| = \text{inj}(g, G) = \text{aut}(g)\mathbf{n}(h, G)$. Applying the inclusion-exclusion principle (8.36) yields

$$\text{aut}(h)\mathbf{N}(h, G) = \sum_{g: g \supset h_0} (-1)^{|E(g)|-|E(h_0)|} \text{inj}(g, G).$$

proving the desired (8.34). \square

The next result is the counterpart of Lemma 49, which shows the existence of a pair of graphs with matching lower order neighborhood subgraph counts but contain distinct number of copies of a certain motif; however, unlike Lemma 49, so far we can only deal with the clique motifs. For example for $\omega = 3$, we can use the ad hoc construction in (8.3); both graphs have the same degree sequence but distinct number of triangles. For $\omega = 4$, we can choose

$$\begin{aligned} H &= \text{triangle} + 3 \times \text{square} + 12 \times \text{triangle with one isolated vertex} + 12 \times \text{path of length 2} \\ H' &= 6 \times \text{triangle} + 12 \times \text{path of length 3} + 4 \times \text{triangle with one isolated vertex} + 4 \times \text{path of length 2} \end{aligned} \quad (8.37)$$

It is straightforward (although extremely tedious!) to verify that $\mathbf{N}(h, H) = \mathbf{N}(h, H')$ for all neighborhood subgraphs h with at most 2 black vertices. The general result is as follows:

Lemma 51. *There exists two graphs H and H' such that $\mathbf{s}(K_\omega, H) - \mathbf{s}(K_\omega, H') \geq 1$ and $\mathbf{N}(h, H) = \mathbf{N}(h, H')$ for all neighborhood subgraphs h such that $\mathbf{v}_b(h) \leq \omega - 2$.*

Proof. First we show that there exist a pair of graphs H and H' such that $\mathbf{n}(g, H) = \mathbf{n}(g, H')$ for all connected graphs g with at most ω vertices except for the clique K_ω , and $\mathbf{n}(g, H) = \mathbf{n}(g, H') = 0$ for all connected graphs g with more than ω vertices. Analogous to the proof of Lemma 49, this either follows from the strong independence of injective graph homomorphism numbers [153], or from the following linear algebra argument. Let $\{h_1, \dots, h_m\}$ denote all distinct (up to isomorphism) *connected* graphs of at most ω vertices. Order the graphs in increasing number of edges (arbitrarily among graphs with the same number of edges) so that h_1 is an isolated vertex and $h_m = K_\omega$. Then the matrix $B = (b_{ij})$ with $b_{ij} = \mathbf{n}(h_i, h_j)$ is upper triangular with strictly positive diagonals. Then H and H' can be constructed from the vector $x = B^{-1}e_m$ similar to (8.33); see (8.37) for a concrete example for K_4 . By design, each connected component of H and H' has at most ω vertices, we have $\mathbf{n}(g, H) = \mathbf{n}(g, H') = 0$ for all connected g with $\mathbf{v}(g) > \omega$.

Next we show that the neighborhood subgraph counts are matched up to order $\omega - 2$. For each neighborhood subgraph h with $\mathbf{v}_b(h) \leq \omega - 2$, by Lemma 50, we have $\mathbf{N}(h, H) = \sum_{g \in \mathcal{G}} c(g, h) \mathbf{n}(g, H)$, where the coefficients $c(g, h)$ are independent of H , and \mathcal{G} contains all subgraphs obtained from h by adding edges incident to black vertices in h or attaching

vertices to black vertices in h . The crucial observation is two-fold: (a) since $v_b(h) \leq \omega - 2$, there exists at least a pair of white vertices in h , which are not connected. Since no edges are added between white vertices, the collection \mathcal{G} excludes the full clique K_ω ; (b) for each $g \in \mathcal{G}$, if g contains more than ω vertices, then $n(g, H) = n(g, H) = 0$; if g contains at most ω vertices (and not K_ω by the previous point), then $n(g, H) = n(g, H)$ by design. Therefore we conclude that $N(h, H) = N(h, H')$ for all neighborhood subgraphs h with $v_b(h) \leq \omega - 2$. \square

8.3.2 Subgraph sampling

Next we prove the lower bound part of Theorem 35:

Proof. Throughout the proof, we assume that both d and s are at least some sufficiently large constant that only depends on $k = v(h)$ and we use c, c', c_0, c_1, \dots to denote constants that possibly depend on k only. We consider two cases separately.

Case I: $p \leq 1/d$. Let H and H' be the pair of graphs from Lemma 49, such that $s(h, H) - s(h, H') \geq 1$ and $s(g, H) = s(g, H')$ for all induced subgraphs g with $v(g) \leq k - 1$. Therefore, by Lemma 48, we have $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = O_k(p^{k-1})$. Let $r = s(h, H)$ which is a constant only depending on k . Applying Theorem 39 with $M = \lfloor s/r \rfloor$ yields the lower bound

$$\inf_{\tilde{s}} \sup_{\substack{G: d(G) \leq d \\ s(h, G) \leq s}} \mathbb{E}_G |\tilde{s} - s(h, G)|^2 = \Omega_k \left(\frac{s}{p^k} \wedge s^2 \right). \quad (8.38)$$

Case II: $p > 1/d$. To apply Lemma 55, we construct a pair of graphs H and H' with maximum degree d such that $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq 1/2$, $s(h, H') = 0$ and $c_1 \ell^{k-1}/p \leq s(h, H) \leq c_2 \ell^{k-1}/p$. Choosing $\ell = c_3((sp)^{\frac{1}{k-1}} \wedge d)$ for some small constant c_3 and applying Theorem 39, we obtain

$$\inf_{\tilde{s}} \sup_{\substack{G: d(G) \leq d \\ s(h, G) \leq s}} \mathbb{E}_G |\tilde{s}(H) - s(h, G)|^2 = \Omega \left(\frac{\ell^{k-1}s}{p} \right) = \Theta_k \left(\frac{sd^{k-1}}{p} \wedge s^2 \right). \quad (8.39)$$

Combining (8.38) and (8.39) completes the proof of the lower bound of Theorem 35.

It remains to construct H and H' . The idea of the construction is to expand each vertex in h into an independent set, which was used in the proof of [153, Lemma 5]. Here, we also need to consider the possibility of expanding into a clique. Next consider two cases:

Suppose h satisfies the “distinct neighborhood” property, that is, for each $v \in V(h)$, $N_h(v)$ is a distinct subset of $V(h)$. Such h includes cliques, paths, cycles, etc. Pick an arbitrary vertex $u \in V(h)$. Let $\{S_v : v \in V(h)\}$ be a collection of disjoint subsets, so that $|S_u| = \lceil c/p \rceil$ and $|S_v| = \lceil cd \rceil$, where c is a constant that only depends on $v(h) = k$ such that $ck \leq 1$. Define a graph H with vertex set $\cup_{v \in V(h)} S_v$ by connecting each pair of $a \in S_u$ and $b \in S_v$ whenever $(u, v) \in E(h)$. In other words, H is obtained by blowing up each vertex in h into an independent set and each edge into a complete bipartite graph. Repeating the same construction with h replaced by $h - u$ yields H' , in which case S_u consists of isolated vertices. By construction, the maximum degree of both graph satisfies is at most d . Note that $H - S_u = H' - S_u$. Thus the sampled graph of H and H' have the same law provided that none of the vertices in S_u is sampled. Applying Lemma 55, we conclude that $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq (1 - p)^{c/p} \leq c'$ for all $p \leq 1/2$, where c' is a constant depending only on k .

Furthermore,

$$\begin{aligned} \mathfrak{s}(h, H') &= \sum_{T \cap S_u = \emptyset} \mathbb{1}\{H'[T] \simeq h\} + \sum_{T \cap S_u \neq \emptyset} \mathbb{1}\{H'[T] \simeq h\} \\ &\stackrel{(a)}{=} \sum_{T \cap S_u = \emptyset} \mathbb{1}\{H'[T] \simeq h\} = \sum_{T \cap S_u = \emptyset} \mathbb{1}\{H[T] \simeq h\}, \end{aligned}$$

where (a) follows from the fact that $H'[T]$ contains isolated vertices whenever $T \cap S_u \neq \emptyset$ while h is connected by assumption. Note that since $|T| = k$, if $T \cap S_u = \emptyset$, then there exists $t, t' \in T$ such that t, t' belong to the same independent set S_v for some v . By construction, t and t' have the same neighborhood, contradicting $H[T] \simeq h$. Thus, we conclude that $\mathfrak{s}(h, H') = 0$. For H , we have

$$\mathfrak{s}(h, H) = \sum_{T \cap S_u \neq \emptyset} \mathbb{1}\{H[T] \simeq h\} \geq |S_u| \prod_{v \neq u} |S_v| \geq c^k \ell^{k-1}/p,$$

and, similarly, $\mathfrak{s}(h, H) \leq |S_u| (\sum_{v \neq u} |S_v|)^{k-1} \leq (2ck)^k \ell^{k-1}/p$.

Next suppose that h does have distinct neighborhoods, thus there exist $\{u_1, \dots, u_\ell\} \subset V(h)$ with $\ell \geq 2$ such that the neighborhood $N_{u_i}(h)$ are identical, denoted by T . Let $g \triangleq h[T] \vee \ell K_1$ is an induced (by $T \cup \{u_1, \dots, u_\ell\}$) subgraph of h . We define H with vertex set $\cup_{v \in v(h)} S_v$ by the same procedure as above, except now all vertices are expanded into a clique, with $|S_{u_1}| = \lceil c/p \rceil$ and $|S_v| = \lceil cd \rceil$ for $v \neq u$. Finally, as before, we connect each pair of $a \in S_u$ and $b \in S_v$ whenever $(u, v) \in E(h)$. Define H' by repeating the same construction with h replaced by $h - u_1$. Analogous to the above we have $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq c$ and it remains to show that $\mathfrak{s}(h, H') = 0$. Indeed, for any set T of k vertices that does not include any vertex from S_{u_1} , since S_{u_i} forms a clique and u_1, \dots, u_ℓ form an independent set in h , the number of induced g in $H'[T]$ is strictly less than that in h . Thus, there exists no $T \subset \cup_{v \neq u_1} S_v$ such that $H'[T]$ is isomorphic to h , and hence $\mathfrak{s}(h, H') = 0$. Entirely analogously, we have $\mathfrak{s}(h, H) = \Theta_k(\ell^{k-1}/p)$. \square

8.3.3 Neighborhood sampling

To illustrate the main idea, we only prove the lower bound cliques. The proof for other motifs (of size up to four) is similar but involves several ad hoc constructions; see Appendix 8.10.

Theorem 40 (Cliques). *For neighborhood sampling with sampling ratio p ,*

$$\inf_{\hat{\mathfrak{s}}} \sup_{\substack{G: \mathfrak{d}(G) \leq d \\ \mathfrak{s}(h, G) \leq s}} \mathbb{E}_G |\hat{\mathfrak{s}} - \mathfrak{s}(K_\omega, G)|^2 = \Theta_\omega \left(\left(\frac{sd}{p^{\omega-1}} \wedge \frac{sd^{\omega-2}}{p^2} \right) \wedge s^2 \right)$$

Proof. For the lower bound, consider two cases. For simplicity, denote the minimax risk on the left-hand side by R .

Case I: $p > 1/d$. Applying Lemma 55 with G being the complete $(\omega - 2)$ -partite graph of $(\omega - 2)\ell$ vertices, $H_1 = K_{1/p, 1/p}$, and $H_2 = (2/p)K_1$, we obtain two graphs H and H' with $\mathfrak{s}(K_\omega, H) \asymp \frac{\ell^{\omega-2}}{p^2}$ and $\mathfrak{s}(K_\omega, H') = 0$, and $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq c < 1$ for all $p \leq 1/2$. By Theorem 39 with $M = s/(\ell^{\omega-2}/p^2)$, we obtain the lower bound $R \gtrsim \frac{s\ell^{\omega-2}}{p^2}$. Let $\ell = cd$ if $\frac{d^{\omega-2}}{p^2} \leq s$ and $\ell = c(p^2s)^{\frac{1}{\omega-2}}$ if $\frac{d^{\omega-2}}{p^2} > s$, for some small constant c . In either case, we find that $\mathfrak{s}(K_\omega, H) \leq s$, $\mathfrak{s}(K_\omega, H') \leq s$, and $R \asymp \frac{s\ell^{\omega-2}}{p^2} \asymp \frac{sd^{\omega-2}}{p^2} \wedge s^2$.

Case II: $p \leq 1/d$. We use a different construction. Let $\ell = c(d \wedge s^{1/\omega})$ for some small constant c . Let H and H' be the two graphs from Lemma 51 such that $\mathfrak{s}(K_\omega, H) - \mathfrak{s}(K_\omega, H') \geq 1$ and $\mathbf{N}(h, H) = \mathbf{N}(h, H')$ for all neighborhood subgraphs h with $\mathbf{v}_b(h) \leq \omega - 2$. By Lemma 48, we have $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = O_\omega(p^{\omega-1}) < 1$. Next we amplify the gap $|\mathfrak{s}(K_\omega, H) - \mathfrak{s}(K_\omega, H')| = \Omega(\ell^\omega)$ by expanding each vertex into an independent set, similar in what is done in the proof of Theorem 35. For each vertex in H , we associate ℓ distinct isolated vertices, and connect each pair of vertices by an edge if and only if they were connected in H . This defines a new graph F with $\ell \mathbf{v}(H)$ vertices and similarly we construct F' from H' . In this way, the subgraph counts of F and F' also match up to order $\omega - 2$, and, in view of Lemma 48, $\text{TV}(P_{\tilde{F}}, P_{\tilde{F}'}) = O_\omega((\ell p)^{\omega-1})$. Furthermore, the number of cliques satisfies $\mathfrak{s}(K_\omega, F) = \mathfrak{s}(K_\omega, H)\ell^\omega$ and $\mathfrak{s}(K_\omega, F') = \mathfrak{s}(K_\omega, H')\ell^\omega$. Thus, $\mathfrak{s}(K_\omega, F) \asymp \mathfrak{s}(K_\omega, F') \asymp |\mathfrak{s}(K_\omega, H) - \mathfrak{s}(K_\omega, H')| = \ell^\omega$. Applying Theorem 39 with $M = s/\ell^\omega$ yields $R \gtrsim (\ell^\omega (\sqrt{\frac{s/\ell^\omega}{(p\ell)^{\omega-1}}} \wedge \frac{s}{\ell^\omega}))^2 \asymp \frac{s\ell}{p^{\omega-1}} \wedge s^2 \asymp \frac{sd}{p^{\omega-1}} \wedge \frac{s^{1+1/\omega}}{p^{\omega-1}} \wedge s^2 \asymp \frac{sd}{p^{\omega-1}} \wedge s^2$, where the last step follows from the assumption that $p \leq 1/d$. \square

8.4 Graphs with additional structures

In this section, we explore how estimation of motif counts can be improved by prior knowledge of the parent graph structure. In particular, for counting edges, we show that even if the parent graph is known to be a forest a priori, for neighborhood sampling, the bound in Theorem 36 remains optimal up to a subpolynomial factor in p . Similarly, for subgraph sampling, we cannot improve the rate in Theorem 35. We also discuss some results for planar graphs. In what follows, we let \mathcal{F} and \mathcal{P} denote the collection of all forests and planar graphs, respectively.

The next results shows that for estimating edge counts, even if it is known a priori that the parent graph is a forest, the risk in Theorem 35 and Theorem 36 cannot be improved in terms of the exponents on p . The proofs of all the following results are given in Appendix 8.8.

Theorem 41. *For subgraph sampling with sampling ratio p ,*

$$\inf_{\widehat{\mathbf{e}}} \sup_{G \in \mathcal{F}: \substack{\mathbf{d}(G) \leq d \\ \mathbf{e}(G) \leq m}} \mathbb{E}_G |\widehat{\mathbf{e}} - \mathbf{e}(G)|^2 \asymp \left(\frac{m}{p^2} \vee \frac{md}{p} \right) \wedge m^2. \quad (8.40)$$

Theorem 42. *For neighborhood sampling with sampling ratio p ,*

$$\inf_{\widehat{\mathbf{e}}} \sup_{G \in \mathcal{F}: \substack{\mathbf{d}(G) \leq d \\ \mathbf{e}(G) \leq m}} \mathbb{E}_G |\widehat{\mathbf{e}} - \mathbf{e}(G)|^2 = \Omega \left(\frac{m}{p^{2+o(1)}} \wedge \frac{md}{p^{1+o(1)}} \wedge \frac{m^2}{p^{o(1)}} \right),$$

where $o(1) = 1/\sqrt{\log \frac{1}{p}}$ is with respect to $p \rightarrow 0$ and uniform in all other parameters.

For estimating the wedge count under subgraph sampling, the following result shows that the risk in Theorem 35 cannot be improved even if we know the parent graph is a forest.

Theorem 43. *For subgraph sampling with sampling ratio p ,*

$$\inf_{\widehat{\mathbf{w}}} \sup_{G \in \mathcal{F}: \substack{\mathbf{d}(G) \leq d \\ \mathbf{w}(G) \leq w}} \mathbb{E}_G |\widehat{\mathbf{w}} - \mathbf{w}(G)|^2 \asymp \left(\frac{w}{p^3} \vee \frac{wd^2}{p} \right) \wedge w^2. \quad (8.41)$$

On the other hand, for neighborhood sampling, the tree structure can be exploited to improve the rate. Analogous to (8.14), we consider an estimator of the form

$$\widehat{\mathbf{w}} = \lambda \mathbf{N}(\bullet \frown \bullet, \widetilde{G}) + \alpha \mathbf{N}(\bullet \frown_{\circ} \bullet, \widetilde{G}) + \beta \mathbf{N}(\bullet \frown \bullet, \widetilde{G}), \quad (8.42)$$

If we weight $\bullet \frown \bullet$ and $\bullet \frown_{\circ} \bullet$ equally, i.e., $\alpha = \lambda$, this estimator reduces to (8.28) and hence inherits the same performance guarantee in (8.29), which by Theorem 49, is optimal. However, as will be seen in Theorem 45, there is added flexibility by this three-parameter family of estimators that produces improved bounds when the parent graphs satisfies certain additional structure. It should also be mentioned that the alternative choices $\lambda = \frac{5-8p}{p^2(4p-3)}$, $\alpha = \frac{1}{p^2}$, and $\beta = \frac{3p-2}{p^3(4p-3)}$ yield the same performance bound as in (8.29).

For this next result, we show that we can improve the performance of the wedge estimator (8.42) if the parent graph is a forest by choosing alternate values of the parameters: $\alpha = \frac{1}{2pq}$,

$\lambda = \frac{1}{p^2}$, and $\beta = 0$. These choices eliminate the largest term in the variance of (8.42), which is proportional to $\mathfrak{n}(S_4, G)(4\alpha^2 p^3 q^2 + 4\lambda\beta p^4 q + \beta^2 p^5 + p^4 q \lambda^2 - 1)$. We immediately get the following variance bound:

$$\text{Var}[\widehat{\mathbf{w}}] \lesssim \frac{\mathbf{w}(G)}{p^2} \vee \frac{\mathbf{w}(G)d}{p}. \quad (8.43)$$

Note also that $\mathfrak{s}(P_3, G) = \sum_u \binom{d_u}{2}$ whenever G is a forest. Hence another estimator we can use is $\sum_u \frac{b_u}{p} \binom{d_u}{2}$ which has variance of order $\frac{\mathbf{w}(G)d^2}{p}$. Putting this all together, we obtain the following result.

Theorem 44. *For neighborhood sampling with sampling ratio p ,*

$$\inf_{\widehat{\mathbf{w}}} \sup_{G \in \mathcal{F}: \substack{\mathbf{d}(G) \leq d \\ \mathbf{w}(G) \leq w}} \mathbb{E}_G |\widehat{\mathbf{w}} - \mathbf{w}(G)|^2 \lesssim \left(\frac{w}{p^2} \vee \frac{wd}{p} \right) \wedge \left(\frac{wd^2}{p} \right) \wedge w^2.$$

The next theorem shows that the minimax bound from Theorem 44 is optimal.

Theorem 45. *For neighborhood sampling with sampling ratio p and $w \geq d$,*

$$\inf_{\widehat{\mathbf{w}}} \sup_{G \in \mathcal{F}: \substack{\mathbf{d}(G) \leq d \\ \mathbf{w}(G) \leq w}} \mathbb{E}_G |\widehat{\mathbf{w}} - \mathbf{w}(G)|^2 = \Omega \left(\left(\frac{w}{p^2} \vee \frac{wd}{p} \right) \wedge \left(\frac{wd^2}{p} \right) \wedge w^2 \right).$$

In the context of estimating triangles, the next set of results show that planarity improves the rates of estimation for both sampling models. Despite the smaller risk however, for subgraph sampling, the optimal estimator is still the Horvitz-Thompson type.

Theorem 46. *For subgraph sampling with sampling ratio p ,*

$$\inf_{\widehat{\mathbf{t}}} \sup_{G \in \mathcal{P}: \substack{\mathbf{d}(G) \leq d \\ \mathbf{t}(G) \leq t}} \mathbb{E}_G |\widehat{\mathbf{t}} - \mathbf{t}(G)|^2 \asymp \left(\frac{t}{p^3} \vee \frac{td}{p^2} \right) \wedge t^2.$$

Theorem 47. *For neighborhood sampling with sampling ratio p ,*

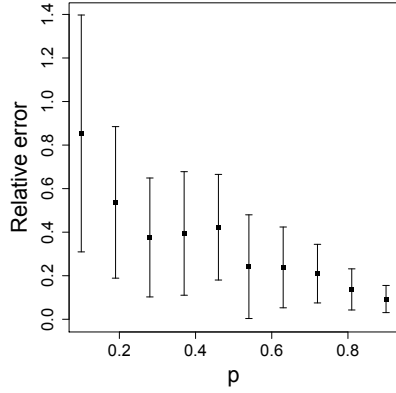
$$\left(\left(\frac{t}{p^{7/3}} \wedge \frac{td}{p^2} \right) \vee \frac{td}{p} \right) \wedge t^2 \lesssim \inf_{\widehat{\mathbf{t}}} \sup_{G \in \mathcal{P}: \substack{\mathbf{d}(G) \leq d \\ \mathbf{t}(G) \leq t}} \mathbb{E}_G |\widehat{\mathbf{t}} - \mathbf{t}(G)|^2 \lesssim \left(\left(\frac{t}{p^3} \wedge \frac{td}{p^2} \right) \vee \frac{td}{p} \right) \wedge t^2.$$

8.5 Numerical experiments

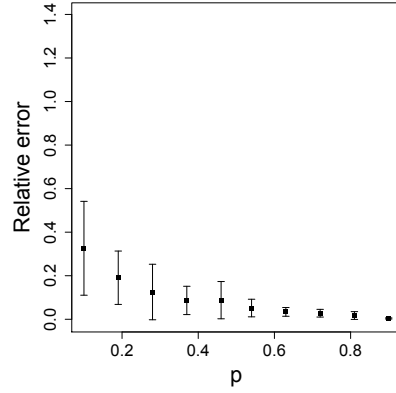
We perform our experiments on both synthetic and real-world data. For the synthetic data, we take as our parent graph G a realization of an Erdős-Rényi graph $\mathcal{G}(N, \delta)$ for various choices of parameters. For the real-world experiment, we study the social networks of survey participants using a Facebook app [179]. This dataset contains 10 ego-networks (the closed neighborhood of a focal vertex (“ego”) and any edges between vertices in its neighborhood) of various sizes, although we only use three of them as our parent graphs G . The error bars in the following figures show the variability of the relative error of edges, triangles, and wedges over 10 independent experiments of subgraph and neighborhood sampling on a fixed parent graph G . The solid black horizontal line shows the sample average and the whiskers show the mean \pm the standard deviation.

Specifically, for subgraph sampling, we always use the HT estimator (8.7). For neighborhood sampling, for counting triangles or wedges, we use the estimator (8.28) with choice of parameters given in Theorem 37 and for counting edges we use the adaptive estimator in Theorem 38. The relative error for estimating the number of edges, triangles, and wedges are given in Fig. 8.2– Fig. 8.4, respectively.

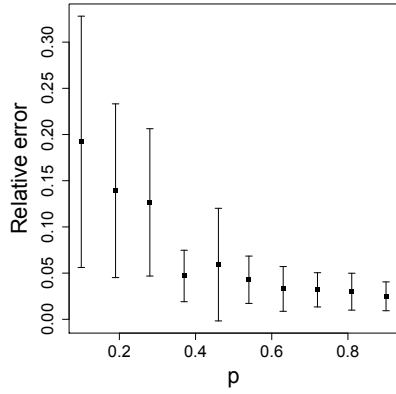
As predicted by the variance bounds, the estimators based on neighborhood sampling perform better than subgraph sampling. Furthermore, there is markedly less variability across the 10 independent experiments in neighborhood sampling. In all plots, however, this variability decreases as p grows. Furthermore, in accordance with our theory, counting bigger motifs (involving more vertices) is subject to more variability, which is evidenced in the plots for triangles and wedges by the wider spread in the whiskers.



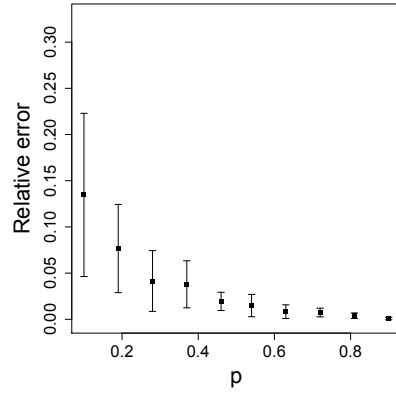
(a) Facebook network (subgraph sampling).



(b) Facebook network (neighborhood sampling).

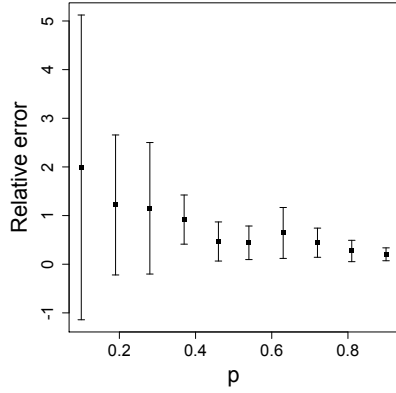


(c) Erdős-Rényi graph (subgraph sampling).

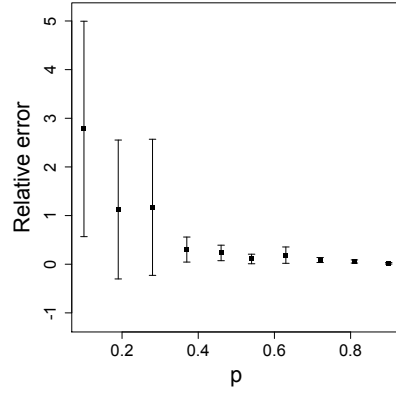


(d) Erdős-Rényi graph (neighborhood sampling).

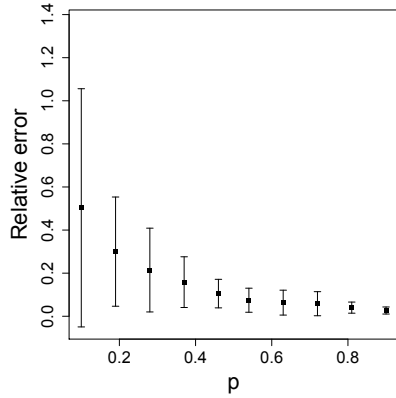
Figure 8.2: Relative error of estimating the edge count. In Fig. 8.2a and Fig. 8.2b, the parent graph G is the Facebook network with $d = 77$, $v(G) = 333$, $e(G) = 2519$. In Fig. 8.2c and Fig. 8.2d, G is a realization of the Erdős-Rényi graph $\mathcal{G}(1000, 0.05)$ with $d = 12$, and $e(G) = 2536$.



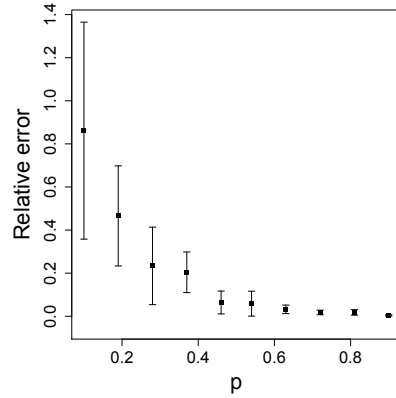
(a) Facebook network (subgraph sampling).



(b) Facebook network (neighborhood sampling).

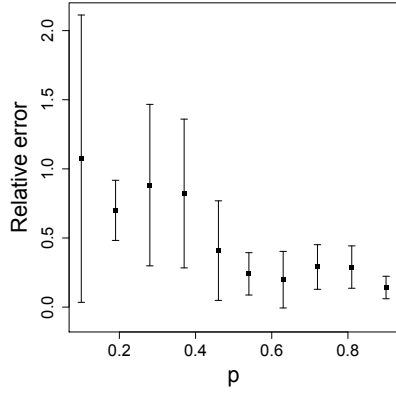


(c) Erdős-Rényi graph (subgraph sampling).

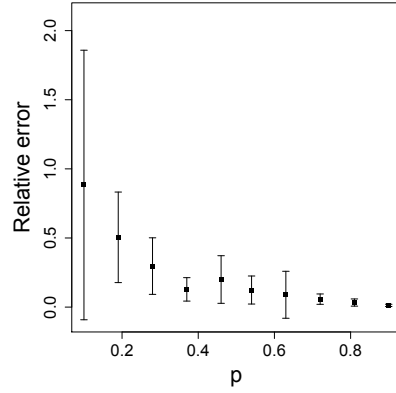


(d) Erdős-Rényi graph (neighborhood sampling).

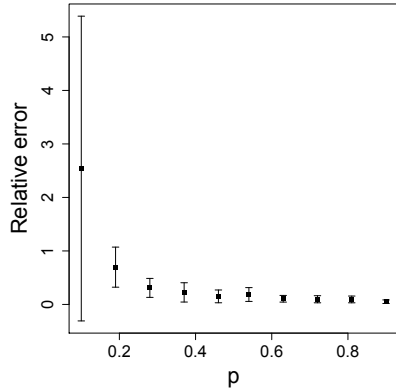
Figure 8.3: Relative error of counting triangles. In Fig. 8.3a and Fig. 8.3b, the parent graph is the Facebook network with $d = 77$, $v(G) = 168$, $t(G) = 7945$. In Fig. 8.3c and Fig. 8.3d, the parent graph is a realization of the Erdős-Rényi graph $\mathcal{G}(1000, 0.02)$ with $d = 35$, and $t(G) = 1319$.



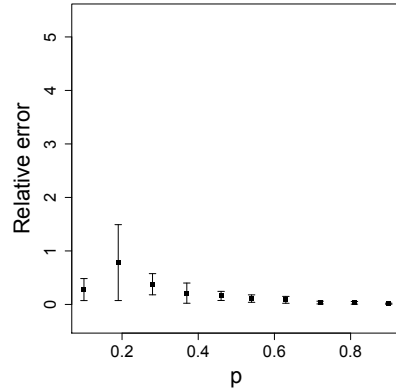
(a) Facebook network (subgraph sampling).



(b) Facebook network (neighborhood sampling).



(c) Erdős-Rényi graph (subgraph sampling).



(d) Erdős-Rényi graph (neighborhood sampling).

Figure 8.4: Relative error of counting wedges. In Fig. 8.4a and Fig. 8.4b, the parent graph is a Facebook network with $d = 29$, $v(G) = 61$, $w(G) = 1039$. In Fig. 8.4c and Fig. 8.4d, the parent graph is a realization of the Erdős-Rényi graph $\mathcal{G}(1000, 0.001)$ with $d = 7$, and $w(G) = 514$.

8.6 Discussion

We conclude the chapter by mentioning a number of interesting questions that remain open:

- As mentioned in the introduction, a more general (and powerful) version of the neighborhood sampling model is to observe a labeled radius- r ball rooted at a randomly chosen vertex [137]. The current chapter focuses on the case of $r = 1$. For $r = 2$, we note for example that a triangle could be observed simply by sampling only one of its

vertices, i.e., \triangle_{\bullet} . Thus, a Horvitz-Thompson type of estimator is $\frac{1}{3p} \mathbf{N}(\triangle_{\bullet}, \tilde{G})$ and the variance scales as $1/p$. When p is small, this outperforms the neighborhood sampling counterpart ($r = 1$) in Theorem 37, where the variance scales as $1/p^2$. Understanding the statistical limits of r -hop neighborhood sampling is an interesting and challenging research direction. In particular, the lower bound will potentially involve more complicated graph statistics as opposed neighborhood subgraph counts.

- In this chapter we have focused on counting motifs as induced subgraphs. As shown in (8.1), subgraph counts can be expressed linear combinations of induced subgraph counts. However, this does not necessarily mean their sample complexity are the same. Although we do not have a systematic understanding so far, here is a concrete example that demonstrates this: consider estimating the number of (not necessarily) 4-cycles with neighborhood sampling. Note that to observe a C_4 one only need to sample the two diagonal vertices. Thus, a simple unbiased estimator is $\frac{1}{2p^2} \mathbf{n}(\square_{\bullet\bullet}, \tilde{G})$, whose variance scales as $O(1/p^2)$ and is much smaller than the best error rate for estimating induced C_4 's which scales as $1/p^3$, as given by Theorem 40. The explanation for this phenomenon is that although we have the deterministic relationship $\mathbf{n}(\square_{\bullet\bullet}, G) = \mathbf{s}(\square_{\bullet\bullet}, G) + \mathbf{s}(\square_{\bullet\bullet}, G) + \mathbf{s}(\square_{\bullet\bullet}, G)$ and each of the three subgraph counts can be estimated at the rate of p^{-3} , the statistical errors cancel each other and result in a faster rate.

8.7 Auxiliary lemmas

Lemma 52 (Kocay's Edge Theorem for Colored Graphs). *Let h be a bicolored disconnected graph. Then $\mathbf{N}(h, G)$ can be expressed as a polynomial, independent of G , in $\mathbf{N}(g, G)$, where g is bicolored, connected, and $\mathbf{v}_b(g) \leq \mathbf{v}_b(h)$. Moreover, if $\prod_{g \in \mathcal{G}} \mathbf{N}(g, G)$ is a term in the polynomial, then $\sum_{g \in \mathcal{G}} \mathbf{v}_b(g) \leq \mathbf{v}_b(h)$ and the corresponding coefficient is bounded by $3^{[\mathbf{v}_b(h)]^2}$. The number of terms in the polynomial representation is bounded by the number of $\mathbf{v}_b(h)$ -tuples $(g_1, \dots, g_{\mathbf{v}_b(h)})$ of all bicolored neighborhood subgraphs such that $\sum_{i=1}^{\mathbf{v}_b(h)} \mathbf{v}_b(g_i) \leq \mathbf{v}_b(h)$ and $\mathbf{N}(g_i, h) \neq 0$.*

Proof. For a disconnected graph g' , note that g' can be decomposed into two graphs g'_1 and

g'_2 , where g'_1 is connected and $\mathbf{v}_b(g'_2) \leq \mathbf{v}_b(g') - 1$. Then,

$$\mathbf{N}(g'_1, G)\mathbf{N}(g'_2, G) = \sum_g a_g \mathbf{N}(g, G), \quad (8.44)$$

where the sum runs over all graphs g with $\mathbf{v}_b(g) \leq \mathbf{v}_b(g'_1) + \mathbf{v}_b(g'_2) = \mathbf{v}_b(g')$ and a_g is the number of decompositions of $V(g)$ into $V(g'_1) \cup V(g'_2)$ and $V_b(g)$ into $V_b(g'_1) \cup V_b(g'_2)$ (not necessarily disjoint) such that $g\{V_b(g'_1)\} \cong g'_1$ and $g\{V_b(g'_2)\} \cong g'_2$.

The only disconnected graph satisfying the above decomposition property for $\mathbf{v}_b(g) = \mathbf{v}_b(g')$ is $g \cong g'$, and hence

$$\mathbf{N}(g', G) = \frac{1}{a_{g'}} \left[\mathbf{N}(g'_1, G)\mathbf{N}(g'_2, G) - \sum_g a_g \mathbf{N}(g, G) \right], \quad (8.45)$$

where $\mathbf{v}_b(g'_2) \leq \mathbf{v}_b(g') - 1$ and the sum ranges over all g that are either connected and $\mathbf{v}_b(g) \leq \mathbf{v}_b(g')$ or disconnected and $\mathbf{v}_b(g) \leq \mathbf{v}_b(g') - 1$. Furthermore, each a_g can be bounded by the number of ways of decomposing a set of size $\mathbf{v}_b(g')$ into two sets (with possible overlap), or $3^{\mathbf{v}_b(g')}$.

We will now prove the following claim using induction. Let h be a bicolored disconnected graph. For each $k < \mathbf{v}_b(h)$,

$$\mathbf{N}(h, G) = \sum_{\mathcal{G}} c_{\mathcal{G}} \prod_{g \in \mathcal{G}} \mathbf{N}(g, G), \quad (8.46)$$

where \mathcal{G} contains at least one disconnected g' for which $\mathbf{v}_b(g') \leq \mathbf{v}_b(h) - k$, $\sum_{g \in \mathcal{G}} \mathbf{v}_b(g) \leq \mathbf{v}_b(h)$, $|c_{\mathcal{G}}| \leq 3^{k\mathbf{v}_b(h)}$, and the number of terms is bounded by the number of k -tuples (g_1, \dots, g_k) of all bicolored neighborhood graphs such that $\sum_{i=1}^k \mathbf{v}_b(g_i) \leq \mathbf{v}_b(h)$ and $\mathbf{N}(g_i, h) \neq 0$.

The base case $k = 1$ is established by decomposing h into two graphs h_1 and h_2 with h_1 connected and $\mathbf{v}_b(h_2) \leq \mathbf{v}_b(h) - 1$ and applying (8.45) with $g' \cong h$, $g'_1 \cong h_1$, and $g'_2 \cong h_2$.

Next, suppose (8.46) holds. Then applying (8.45) to each disconnected g' , we have

$$\begin{aligned}
\mathbf{N}(h, G) &= \sum_{\mathcal{G}} c_{\mathcal{G}} \mathbf{N}(g', G) \prod_g \mathbf{N}(g, G) \\
&= \sum_{\mathcal{G}} \frac{c_{\mathcal{G}}}{c_{g'}} [\mathbf{N}(g'_1, G) \mathbf{N}(g'_2, G) - \sum_{h'} c_{h'} \mathbf{N}(h', G)] \prod_g \mathbf{N}(g, G) \\
&= \sum_{\mathcal{G}} \frac{c_{\mathcal{G}}}{c_{g'}} \mathbf{N}(g'_1, G) \mathbf{N}(g'_2, G) \prod_g \mathbf{N}(g, G) - \sum_{\mathcal{G}} \sum_{h'} \frac{c_{\mathcal{G}} c_{h'}}{c_{g'}} \mathbf{N}(h', G) \prod_g \mathbf{N}(g, G). \quad (8.47)
\end{aligned}$$

Note that $\mathbf{v}_b(g'_2) \leq \mathbf{v}_b(g') - 1 \leq \mathbf{v}_b(h) - (k+1)$ and if h' is disconnected, then $\mathbf{v}_b(h') \leq \mathbf{v}_b(g') - 1 \leq \mathbf{v}_b(h) - (k+1)$. Finally, we observe that (8.47) has the form

$$\sum_{\tilde{\mathcal{G}}} c_{\tilde{\mathcal{G}}} \prod_g \mathbf{N}(g, G), \quad (8.48)$$

where $\tilde{\mathcal{G}}$ contains at least one disconnected g' for which $\mathbf{v}_b(g') \leq \mathbf{v}_b(h) - (k+1)$, $\mathbf{v}_b(g') \leq \mathbf{v}_b(h) - (k+1)$, $\sum_{g \in \tilde{\mathcal{G}}} \mathbf{v}_b(g) \leq \mathbf{v}_b(h)$, and $|c_{\tilde{\mathcal{G}}}| \leq \left| \frac{c_{\mathcal{G}}}{c_{g'}} \right| \vee \left| \frac{c_{\mathcal{G}} c_{h'}}{c_{g'}} \right| \leq 3^{(k+1)\mathbf{v}_b(h)}$. The number of terms is bounded by the number of $(k+1)$ -tuples (g_1, \dots, g_{k+1}) of all bicolored neighborhood graphs such that $\sum_{i=1}^{k+1} \mathbf{v}_b(g_i) \leq \mathbf{v}_b(h)$ and $\mathbf{N}(g_i, h) \neq 0$. Repeat this until $k = \mathbf{v}_b(h)$ and so that the right hand side of (8.46) contains no disconnected g in its terms. \square

Lemma 53. *Let H and H' be two graphs on M vertices. Suppose there exists a constant $B > 0$ and positive integer k such that for each connected subgraph h ,*

$$|\mathbf{N}(h, H) - \mathbf{N}(h, H')| \leq B M^{\mathbf{v}_b(h)-k}.$$

Then for each subgraph h ,

$$|\mathbf{N}(h, H) - \mathbf{N}(h, H')| \leq B Q_h \mathbf{v}_b(h) 3^{[\mathbf{v}_b(h)]^2} M^{\mathbf{v}_b(h)-k},$$

where Q_h is the number of $\mathbf{v}_b(h)$ -tuples $(g_1, \dots, g_{\mathbf{v}_b(h)})$ of all bicolored neighborhood graphs such that $\sum_{i=1}^{\mathbf{v}_b(h)} \mathbf{v}_b(g_i) \leq \mathbf{v}_b(h)$ and $\mathbf{N}(g_i, H) \neq 0$ or $\mathbf{N}(g_i, H') \neq 0$.

Proof. Let h be a disconnected subgraph. By Lemma 52, $\mathbf{N}(h, H) = \sum_{\mathcal{G}_h} c_{\mathcal{G}_h} \prod_{g \in \mathcal{G}_h} \mathbf{N}(g, H)$, where $\sum_{g \in \mathcal{G}_h} \mathbf{v}_b(g) \leq \mathbf{v}_b(h)$, g is connected, $|c_{\mathcal{G}_h}| \leq 3^{[\mathbf{v}_b(h)]^2}$, and the number of terms is

bounded by Q_h . Thus, using the fact that if x_1, \dots, x_n and y_1, \dots, y_n are positive real numbers, $|x_1 \cdots x_n - y_1 \cdots y_n| \leq \sum_{i=1}^n |x_i - y_i| x_1 \cdots x_{i-1} y_{i+1} \cdots y_n$, we have

$$\begin{aligned} |\mathbf{N}(h, H) - \mathbf{N}(h, H')| &\leq \sum_{\mathcal{G}_h} |c_{\mathcal{G}_h}| \left| \prod_{g \in \mathcal{G}_h} \mathbf{N}(g, H) - \prod_{g \in \mathcal{G}_h} \mathbf{N}(g, H') \right| \\ &\leq \sum_{\mathcal{G}_h} |c_{\mathcal{G}_h}| \sum_i |\mathbf{N}(g_i, H) - \mathbf{N}(g_i, H')| \prod_{j \leq i-1} \mathbf{N}(g_j, H) \prod_{j \geq i+1} \mathbf{N}(g_j, H'), \end{aligned}$$

where $\{g_i\}$ is an ordering of $\{g\}_{g \in \mathcal{G}_h}$. Next, we use the fact that $\max\{\mathbf{N}(g, H), \mathbf{N}(g, H')\} \leq \binom{M}{\mathbf{v}_b(g)} \leq M^{\mathbf{v}_b(g)}$ to bound

$$\sum_i |\mathbf{N}(g_i, H) - \mathbf{N}(g_i, H')| \prod_{j \leq i-1} \mathbf{N}(g_j, H) \prod_{j \geq i+1} \mathbf{N}(g_j, H') \leq B |\mathcal{G}_h| M^{\sum_{g \in \mathcal{G}_h} \mathbf{v}_b(g) - k}$$

Since $|\mathcal{G}_h| \leq \sum_{g \in \mathcal{G}_h} \mathbf{v}_b(g) \leq \mathbf{v}_b(h)$, the above is further bounded by $B \mathbf{v}_b(h) M^{\mathbf{v}_b(h) - k}$. Thus,

$$\begin{aligned} |\mathbf{N}(h, H) - \mathbf{N}(h, H')| &\leq B \mathbf{v}_b(h) M^{\mathbf{v}_b(h) - k} \sum_{\mathcal{G}_h} |a_{\mathcal{G}_h}| \\ &\leq B Q_h \mathbf{v}_b(h) 3^{[\mathbf{v}_b(h)]^2} M^{\mathbf{v}_b(h) - k}. \end{aligned}$$

□

Next we present two results on the total variation that will be used in the regime of $p > \frac{1}{d}$. The main idea is the following: if a subset T of vertices are not sampled, for subgraph sampling, in the observed graph we delete all edges incident to T , i.e., the edge set of $G\{T\}$, and for neighborhood sampling, we delete all edges within T , that is, the edge set of $G[T]$. Therefore, for two parent graphs, if missing T leads to isomorphic graphs, then by a natural coupling, the total variation between the sampled graphs is at most the probability that T is not completely absent in the sample.

Lemma 54. *Let $G_\theta = K_{A, \Delta - \theta} + K_{B, \Delta + \theta}$ for integer θ between zero and Δ . Consider the neighborhood sampling model with sampling ratio p . Suppose $|\theta - \theta'| \asymp \sqrt{\frac{\Delta}{p}}$ and both A and B are at most $1/p$. For neighborhood sampling with sampling ratio p , there exists $0 < c < 1$*

such that

$$\text{TV}(P_{\tilde{G}_\theta}, P_{\tilde{G}_{\theta'}}) \leq c.$$

Proof. Note that G_θ is the union of two complete bipartite graphs. Suppose that none of the $A + B$ “left” side vertices are sampled. Then G_θ can be described by $K_{A,X} + K_{B,Y} + (2\Delta - (X + Y))K_1$, where $(X, Y) \sim \text{Bin}(\Delta - \theta, p) \otimes \text{Bin}(\Delta + \theta, p)$. Thus, if $(X', Y') \sim \text{Bin}(\Delta - \theta', p) \otimes \text{Bin}(\Delta + \theta', p)$, then

$$\text{TV}(P_{\tilde{G}_\theta}, P_{\tilde{G}_{\theta'}}) \leq 1 - q^{A+B} + q^{A+B} \text{TV}(P_{(X,Y)}, P_{(X',Y')}).$$

Furthermore, observe that

$$\text{TV}(P_{(X,Y)}, P_{(X',Y')}) \leq \text{TV}(P_X, P_{X'}) + \text{TV}(P_Y, P_{Y'}),$$

where

$$\text{TV}(P_X, P_{X'}) = \text{TV}(\text{Bin}(\Delta - \theta, p), \text{Bin}(\Delta - \theta', p)),$$

$$\text{TV}(P_Y, P_{Y'}) = \text{TV}(\text{Bin}(\Delta + \theta, p), \text{Bin}(\Delta + \theta', p)).$$

This shows that if $|\theta - \theta'| \asymp \sqrt{\frac{\Delta}{p}}$ and both A and B are $O(\frac{1}{p})$, then $\text{TV}(P_{\tilde{G}_\theta}, P_{\tilde{G}_{\theta'}})$ is less than a constant less than one. \square

Lemma 55. *Let G , H_1 , and H_2 be an arbitrary graphs and let $H = G \vee H_1$ for and $H' = G \vee H_2$. If $v = v(H_1) = v(H_2) \leq 1/p$, then for neighborhood sampling with sampling ratio p ,*

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq 1 - q^v \leq 1 - q^{1/p}, \quad (8.49)$$

More generally, for $H = (V, E)$ and $H' = (V, E')$ defined on the same set V of vertices, if $T \subset V$ is such that $(V \setminus T, E \setminus E(H[T]))$ and $(V \setminus T, E' \setminus E(H'[T]))$ are isomorphic, then (8.49) holds with $v = |T|$.

Proof. Suppose that none of the v vertices in H_1 or H_2 are sampled. Then H_1 and H_2 are

isomorphic to each other. Thus,

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq \mathbb{P}[\text{at least one vertex in } H_1 \text{ or } H_2 \text{ is sampled}] = 1 - q^v.$$

The second claim follows from the same argument. \square

The following lemma, which was used in the proof of Theorems 42 and 44, relies on a number-theoretic fact:

Lemma 56. *There exist two sequences of integers $(\alpha_1, \dots, \alpha_{k+1})$ and $(\beta_1, \dots, \beta_{k+1})$ such that*

$$\sum_{x \in [k+1]} x^i \alpha_x = 0 \quad i = 0, 2, 3, \dots, k,$$

$$\sum_{x=1}^{k+1} x^i \beta_x = 0 \quad i = 0, 1, 3, \dots, k,$$

and

$$\sum_{x \in [k+1]} x \alpha_x = \text{lcm}(1, \dots, k+1),$$

$$\sum_{x \in [k+1]} x^2 \beta_x = \text{lcm}^2(1, \dots, k+1),$$

where lcm stands for the least common multiple. Moreover, there exists universal constants A and B such that

$$\sum_{x \in [k+1]} |\alpha_x| \leq A^k, \quad \sum_{x \in [k+1]} |\beta_x| \leq B^k. \quad (8.50)$$

Proof. We first introduce the quantity

$$\gamma_i = \sum_{x=1}^k \frac{(-1)^{x+1}}{x^i} \binom{k}{x}.$$

The key observation is that $\sum_{x=0}^{k+1} (-1)^x \binom{k+1}{x} D(x) = 0$ for all polynomials D with degree less than or equal to k . Hence we can set

$$\alpha_x = \left(\gamma_1 - \frac{1}{x} \right) (-1)^x \binom{k+1}{x} \text{lcm}(1, \dots, k+1)$$

and

$$\beta_x = \left(\gamma_1^2 - \gamma_2 - \frac{\gamma_1}{x} + \frac{1}{x^2} \right) (-1)^x \binom{k+1}{x} \text{lcm}^2(1, \dots, k+1),$$

where $x = 1, 2, \dots, k+1$. A well-known number theoretic fact is that the least common multiple of the k integers is in fact significantly smaller than their product. In fact, we have the estimates [180], [181]

$$2^{k-1} \leq \text{lcm}(1, \dots, k) \leq 3^k, \text{ for all } k \geq 1,$$

which shows (8.50). □

Lemma 57. *For the two graphs H and H' from Theorem 42 constructed with $(\alpha_1, \dots, \alpha_{k+1})$ from Lemma 56, we have for neighborhood sampling with sampling ratio p ,*

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = O(pA^k + (p\ell A^k)^k),$$

provided $p\ell A^k < 1$.

Proof. There are four types of connected subgraphs of H and H' : edge with one black vertex, edge with two black vertices, S_u , $u > 1$ with white center, S_u , $u > 1$ with black center. If g is an edge with one black vertex $\mathbf{N}(g, H) = 2\ell\alpha + \ell \sum_{x=1}^{k+1} xw_x$ and $\mathbf{N}(g, H') = 2\ell\alpha' + \ell \sum_{x=1}^{k+1} xw'_x$. If g is an edge with two black vertices $\mathbf{N}(g, H) = \ell\alpha$ and $\mathbf{N}(g, H') = \ell\alpha'$. If $g \cong S_u$ with white center, then $\mathbf{N}(g, H) = \sum_{x=1}^{k+1} w_x \binom{\ell x}{\mathbf{v}_b(g)}$ and $\mathbf{N}(g, H') = \sum_{x=1}^{k+1} w'_x \binom{\ell x}{\mathbf{v}_b(g)}$ and furthermore,

$$\begin{aligned} |\mathbf{N}(g, H) - \mathbf{N}(g, H')| &= \left| \sum_{x=1}^{k+1} w_x \binom{\ell x}{\mathbf{v}_b(g)} - \sum_{x=1}^{k+1} w'_x \binom{\ell x}{\mathbf{v}_b(g)} \right| \\ &= \frac{\ell}{\mathbf{v}_b(g)} \left| \sum_{x=1}^{k+1} xw_x - \sum_{x=1}^{k+1} xw'_x \right|. \end{aligned}$$

If $g \cong S_u$ with black center, then

$$\begin{aligned}\mathbf{N}(g, H) &= \sum_{x=1}^{k+1} w_x \binom{\ell x}{\mathbf{v}_b(g) - 1} \mathbb{1}\{\ell x = u\} \\ \mathbf{N}(g, H') &= \sum_{x=1}^{k+1} w'_x \binom{\ell x}{\mathbf{v}_b(g) - 1} \mathbb{1}\{\ell x = u\}\end{aligned}$$

We find that $|\mathbf{N}(g, H) - \mathbf{N}(g, H')| \leq 2a^k(\ell(k+1))^{\mathbf{v}_b(g)-1}$ and $|\mathbf{N}(g, H)| \leq 2a^k(\ell(k+1))^{\mathbf{v}_b(g)}$.

Let $v = \mathbf{v}(H) = \mathbf{v}(H') \leq (\ell(k+1) + 1)a^k$. Then

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq \frac{1}{2} \sum_{h: \mathbf{v}_b(h) \leq k} |\mathbf{N}(h, H) - \mathbf{N}(h, H')| p^{\mathbf{v}_b(h)} q^{v - \mathbf{v}_b(h)} + \mathbb{P}[\text{Bin}(v, p) \geq k+1],$$

where the sum runs over all bicolored graphs with at most k black vertices. By Lemma 53, for each subgraph h ,

$$|\mathbf{N}(h, H) - \mathbf{N}(h, H')| \leq \mathbf{v}_b(h) 3^{[\mathbf{v}_b(h)]^2} (2\mathbf{v}_b(h)a^k(k+3))^{\mathbf{v}_b(h)} (\ell(k+1))^{\mathbf{v}_b(h)-1},$$

where we used the bound $Q_h \leq [\mathbf{v}_b(h)(k+3)]^{\mathbf{v}_b(h)}$. Hence,

$$\begin{aligned}\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) &\leq \frac{1}{2} \sum_{h: 1 \leq \mathbf{v}_b(h) \leq k} |\mathbf{N}(h, H) - \mathbf{N}(h, H')| p^{\mathbf{v}_b(h)} q^{v - \mathbf{v}_b(h)} + \mathbb{P}[\text{Bin}(v, p) \geq k+1] \\ &\leq \frac{1}{2} \sum_{h: 1 \leq \mathbf{v}_b(h) \leq k} \mathbf{v}_b(h) 3^{[\mathbf{v}_b(h)]^2} (2\mathbf{v}_b(h)a^k(k+3))^{\mathbf{v}_b(h)} (\ell(k+1))^{\mathbf{v}_b(h)-1} p^{\mathbf{v}_b(h)} q^{v - \mathbf{v}_b(h)} + \\ &\quad \mathbb{P}[\text{Bin}((\ell(k+1) + 1)a^k, p) \geq k+1] \\ &\leq (pA^k) \sum_{v=0}^k (p\ell A^k)^v + \sum_{v=k+1}^{\infty} (p\ell A^k)^v \\ &= O(pA^k + (p\ell A^k)^{k+1}),\end{aligned}$$

for some constant $A > 0$ and provided $p\ell A^k < 1$. □

Lemma 58. *For the two graphs H and H' from Theorem 45 constructed with $(\beta_1, \dots, \beta_{k+1})$*

from Lemma 56, we have for neighborhood sampling with sampling ratio p ,

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = O(pA^k + (p\ell A^k)^2 + (p\ell A^k)^k),$$

provided $p\ell A^k < 1$.

Proof. There are two types of connected subgraphs of H and H' : S_u , $u > 1$ with white center and S_u , $u > 1$ with black center. If $g \cong S_u$ with white center, then $\mathbf{N}(g, H) = \sum_{x=1}^{k+1} w_x \binom{\ell x}{\mathbf{v}_b(g)}$ and $\mathbf{N}(g, H') = \sum_{x=1}^{k+1} w'_x \binom{\ell x}{\mathbf{v}_b(g)}$ and furthermore, since $\sum_{x=1}^{k+1} x^i w_x = \sum_{x=1}^{k+1} x^i w'_x$ for $i = 0, 1, 3, \dots, \mathbf{v}_b(g)$,

$$\begin{aligned} |\mathbf{N}(g, H) - \mathbf{N}(g, H')| &= \left| \sum_{x=1}^{k+1} w_x \binom{\ell x}{\mathbf{v}_b(g)} - \sum_{x=1}^{k+1} w'_x \binom{\ell x}{\mathbf{v}_b(g)} \right| \\ &= \frac{\ell^2}{\mathbf{v}_b(g)(\mathbf{v}_b(g) - 1)} \left| \sum_{x=1}^{k+1} x^2 w_x - \sum_{x=1}^{k+1} x^2 w'_x \right|. \end{aligned}$$

If $g \cong S_u$ with black center, then

$$\begin{aligned} \mathbf{N}(g, H) &= \sum_{x=1}^{k+1} w_x \binom{\ell x}{\mathbf{v}_b(g) - 1} \mathbb{1}\{\ell x = u\} \\ \mathbf{N}(g, H') &= \sum_{x=1}^{k+1} w'_x \binom{\ell x}{\mathbf{v}_b(g) - 1} \mathbb{1}\{\ell x = u\} \end{aligned}$$

We find that $|\mathbf{N}(g, H) - \mathbf{N}(g, H')| \leq 2a^k(\ell(k+1))^{\mathbf{v}_b(g)-1}$ and $|\mathbf{N}(g, H)| \leq a^k(\ell(k+1))^{\mathbf{v}_b(g)}$.

Let $v = \mathbf{v}(H) = \mathbf{v}(H') \leq (\ell(k+1) + 1)a^k$. Then

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq \frac{1}{2} \sum_{h: \mathbf{v}_b(h) \leq k} |\mathbf{N}(h, H) - \mathbf{N}(h, H')| p^{\mathbf{v}_b(h)} q^{v - \mathbf{v}_b(h)} + \mathbb{P}[\text{Bin}(v, p) \geq k+1],$$

where the sum runs over all bicolored graphs with at most k black vertices. By Lemma 53, for each subgraph h with $\mathbf{v}_b(h) \neq 2$,

$$|\mathbf{N}(h, H) - \mathbf{N}(h, H')| \leq \mathbf{v}_b(h) 3^{[\mathbf{v}_b(h)]^2} (2\mathbf{v}_b(h) a^k (k+3))^{\mathbf{v}_b(h)} (\ell(k+1))^{\mathbf{v}_b(h)-1},$$

where we used the bound $Q_h \leq [\mathbf{v}_b(h)(k+3)]^{\mathbf{v}_b(h)}$. Hence,

$$\begin{aligned}
\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) &\leq \frac{1}{2} \sum_{h: 1 \leq \mathbf{v}_b(h) \leq k} |\mathbf{N}(h, H) - \mathbf{N}(h, H')| p^{\mathbf{v}_b(h)} q^{v - \mathbf{v}_b(h)} + \mathbb{P}[\text{Bin}(v, p) \geq k+1] \\
&\leq \frac{1}{2} \sum_{h: \mathbf{v}_b(h) \neq 2, \mathbf{v}_b(h) \leq k} \mathbf{v}_b(h) 3^{[\mathbf{v}_b(h)]^2} (2\mathbf{v}_b(h) a^k (k+3))^{\mathbf{v}_b(h)} (\ell(k+1))^{\mathbf{v}_b(h)-1} p^{\mathbf{v}_b(h)} q^{v - \mathbf{v}_b(h)} + \\
&\quad a^k \ell^2 p^2 + \mathbb{P}[\text{Bin}((\ell(k+1)+1)a^k, p) \geq k+1] \\
&\leq (pA^k) \sum_{v=0}^k (p\ell A^k)^v + (p\ell A^k)^2 + \sum_{v=k+1}^{\infty} (p\ell A^k)^v \\
&= O(pA^k + (p\ell A^k)^2 + (p\ell A^k)^{k+1}),
\end{aligned}$$

for some constant $A > 0$ and provided $p\ell A^k < 1$. \square

Lemma 59. *There exists two planar graphs H and H' on order ℓ vertices with matching degree sequences and maximum degree equal to $\ell+1$ such that for neighborhood sampling with sampling ratio p , $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = O(p^2 + p^3 \ell^3)$ and $|\mathbf{w}(H) - \mathbf{w}(H')| = 3|\mathbf{t}(H) - \mathbf{t}(H')| \asymp \ell$ provided $p = O(1/\ell)$. Furthermore, there exists two planar graphs H and H' on order ℓ vertices such that for neighborhood sampling with sampling ratio p , $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = O(p)$ and $|\mathbf{t}(H) - \mathbf{t}(H')| \asymp \ell$.*

Proof. The proof follows from an examination of the two graphs below. Note that $\mathbf{N}(h, H) = \mathbf{N}(h, H')$ for all connected h with $\mathbf{v}_b(h) = 1$ and since $|\mathbf{N}(h, H) - \mathbf{N}(h, H')| = O(1)$ for all connected h with $\mathbf{v}_b(h) = 2$, it follows from Lemma 53 with $k = 2$ that $|\mathbf{N}(h, H) - \mathbf{N}(h, H')| = O(1)$ for all h with $\mathbf{v}_b(h) = 2$. Thus,

$$\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = \sum_h |\mathbf{N}(h, H) - \mathbf{N}(h, H')| p^{\mathbf{v}_b(h)} q^{v - \mathbf{v}_b(h)} = O(p^2 + \sum_{k=3}^{\infty} \ell^k p^k) = O(p^2 + p^3 \ell^3),$$

provided $p = O(1/\ell)$. The identity $|\mathbf{w}(H) - \mathbf{w}(H')| = 3|\mathbf{t}(H) - \mathbf{t}(H')| = \ell - 2$ follows from the fact that H and H' have matching degree sequences (corresponding to matching subgraphs from neighborhood sampling with one vertex).

For the second statement, consider two planar graphs H and H' on $\ell+2$ vertices, where H consists of ℓ triangles sharing a common edge, and H' consists of ℓ wedges sharing a pair

Table 8.2: The graph H with $\ell = 5$ and $d(H) = \ell + 1 = 6$

Copies	Components
1	
2	
2	
$\frac{\ell+1}{3}$	
$2(\ell + 1)$	

Table 8.3: The graph H' with $\ell = 5$ and $d(H') = \ell + 1 = 6$

Copies	Components
1	
2	
1	
$\ell + 1$	
$2(\ell + 1)$	

of non-adjacent vertices; see Fig. ?? for an illustration for $\ell = 5$.

Table 8.4: The graph H with $\ell = 5$

Copies	Graph
1	

Table 8.5: The graph H' with $\ell = 5$

Copies	Graph
1	

Note that if neither of the two highest-degree vertices in each graph (degree $\ell + 1$ in H and degree ℓ in H') are sampled and all incident edges removed, the two graphs are isomorphic. This shows that $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) \leq 1 - q^2 = O(p)$. Also, note that $\mathfrak{t}(H) = \ell$ and $\mathfrak{t}(H') = 0$. \square

8.8 Additional proofs

Proof of Theorems 41, 43, and 46. The upper bounds are achieved by Horvitz-Thompson estimation as in Theorem 35. However, for Theorem 43, we are able to achieve a smaller variance because $n(\text{triangle}, G)$ is of order td for planar G instead of td^2 and hence $\text{Var}[\widehat{t}_{\text{HT}}] \lesssim \frac{n(\text{triangle}, G)}{p^3} + \frac{n(\text{triangle}, G)}{p^2} + \frac{n(\text{triangle}, G)}{p} \lesssim \frac{t}{p^3} + \frac{td}{p^2} + \frac{td}{p} \asymp \frac{t}{p^3} \vee \frac{td}{p^2}$. For the lower bound, the proof follows the same lines as Section 8.3.2 in that we use two different constructions depending on whether $p \leq 1/d$ or $p > 1/d$.

For edges, let $H = S_\ell$ and $H' = (\ell + 1)S_1$ with $\ell = c(d \wedge m)$ for some small constant $c > 0$. Then $\text{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq p(1 - q^\ell) \leq p \wedge (\ell p^2)$.

For wedges, when $p \leq 1/d$, let $H = P_4 + K_1$ and $H' = P_3 + P_2$. Then $\text{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq O(p^3)$. When $p > 1/d$, let $H = S_\ell$ and $H' = (\ell + 1)K_1$. Then $\text{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq p$. Finally set $\ell = c(d \wedge w)$ for some universal constant $c > 0$.

Finally, for triangles, let H be the graph which consists of ℓ triangles that share the same edge plus ℓ isolated vertices. Let H' be the graph which consists of two S_ℓ star graphs with an edge between their roots. Choose $\ell = c(d \wedge t)$ for some small universal constant $c > 0$. Then $\text{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq p^2(1 - q^\ell) \leq p^2 \wedge (p^3 \ell)$. \square

Proof of Theorem 42. Let (w_1, \dots, w_{k+1}) and (w'_1, \dots, w'_{k+1}) be two sequences of integers defined by $w_x = \max\{\alpha_x, 0\}$ and $w'_x = \max\{-\alpha_x, 0\}$, where $(\alpha_1, \dots, \alpha_{k+1})$ is as in Lemma 56. Consider the disjoint union of stars

$$H \simeq \sum_{x=1}^{k+1} w_x S_{\ell_x} + \ell \alpha S_1 \quad \text{and} \quad H' \simeq \sum_{x=1}^{k+1} w'_x S_{\ell_x} + \ell \alpha' S_1,$$

for integer $\ell > 1$.

Note, for example, that $\mathbf{e}(H) = \ell(\sum_{x=1}^{k+1} x w_x + \alpha)$ and $\mathbf{v}(H) = \mathbf{e}(H) + \sum_{x=1}^{k+1} w_x + \ell \alpha =$

$\sum_{x=1}^{k+1}(\ell x + 1)w_x + 2\alpha\ell$. Thus, $\mathbf{e}(H) \vee \mathbf{e}(H') \leq \ell a^k$ for some universal $a > 0$. Note that

$$\mathbf{e}(H) - \mathbf{e}(H') = \ell(\alpha - \alpha') = \frac{\ell}{2} \left(\sum_{x=1}^{k+1} xw_x - \sum_{x=1}^{k+1} xw'_x \right) \geq \frac{\ell}{2},$$

and by Lemma 57 there exists universal $A > 0$ such that

$$\mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = \frac{1}{2} \sum_h |\mathbf{N}(h, H) - \mathbf{N}(h, H')| p^{\mathbf{v}_b(h)} q^{v - \mathbf{v}_b(h)} = O(pA^k + (p\ell A^k)^k),$$

provided $p\ell A^k < 1$.

By Theorem 39, we have

$$\inf_{\hat{\mathbf{e}}} \sup_{G \in \mathcal{F}: \mathbf{d}(G) \leq d, \mathbf{e}(G) \leq m} \mathbb{P} [|\hat{\mathbf{e}} - \mathbf{e}(G)| \geq \Delta_\ell] \geq c.$$

where

$$\begin{aligned} \Delta_\ell &\gtrsim |\mathbf{e}(H) - \mathbf{e}(H')| \left(\sqrt{\frac{m}{\mathbf{e}(H) \vee \mathbf{e}(H') \mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'})}} \wedge \frac{m}{\mathbf{e}(H) \vee \mathbf{e}(H')} \right) \\ &\gtrsim \sqrt{\frac{m\ell}{pc^k + (p\ell c^k)^k}} \wedge \frac{m}{c^k}, \end{aligned}$$

for some universal constants $c > 0$ provided $p\ell c^k < 1$. Next, choose

$$\ell = \begin{cases} \left(\frac{1}{pc^k} \right)^{1-1/k} \wedge \frac{m}{a^k} & \text{if } p > \left(\frac{1}{dc^k} \right)^{k/(k-1)} \\ d \wedge \frac{m}{a^k} & \text{if } p \leq \left(\frac{1}{d^k} \right)^{k/(k-1)} \end{cases}. \quad (8.51)$$

Taking $k = \sqrt{\log \frac{1}{p}}$ yields the desired lower bound. \square

Proof of Theorem 45. Let (w_1, \dots, w_{k+1}) and (w'_1, \dots, w'_{k+1}) be two sequences of integers defined by $w_x = \max\{\beta_x, 0\}$ and $w'_x = \max\{-\beta_x, 0\}$, where $(\beta_1, \dots, \beta_{k+1})$ is as in Lemma 56.

Let

$$H \simeq \sum_{x=1}^{k+1} w_x S_{\ell x} \quad \text{and} \quad H' \simeq \sum_{x=1}^{k+1} w'_x S_{\ell x},$$

for integer $\ell > 1$. Note, for example, that $\mathbf{e}(H) = \ell \sum_{x=1}^{k+1} xw_x$, $\mathbf{v}(H) = \sum_{x=1}^{k+1} (\ell x + 1)w_x$,

and $w(H) = \sum_{x=1}^{k+1} \binom{\ell x}{2} w_x$. This means that $w(H) \vee w(H') \leq \ell^2 a^{2k}$ for some universal $a > 0$.

Note that

$$w(H) - w(H') = \frac{\ell^2}{2} \left(\sum_{x=1}^{k+1} x^2 w_x - \sum_{x=1}^{k+1} x^2 w'_x \right) \geq \frac{\ell^2}{2}.$$

By Lemma 58, we have that $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = O(pA^k + (p\ell A^k)^2 + (p\ell A^k)^k)$ for some universal $A > 0$. By Theorem 39, we have

$$\inf_{\widehat{w}} \sup_{G \in \mathcal{F}: \substack{d(G) \leq d \\ w(G) \leq w}} \mathbb{P} [|\widehat{w} - w(G)| \geq \Delta_\ell] \geq c.$$

where

$$\begin{aligned} \Delta_\ell &\gtrsim |w(H) - w(H')| \left(\sqrt{\frac{w}{w(H) \vee w(H') \text{TV}(P_{\tilde{H}}, P_{\tilde{H}'})}} \wedge \frac{w}{w(H) \vee w(H')} \right) \\ &\gtrsim \sqrt{\frac{w\ell^2}{pc^k + (p\ell c^k)^2 + (p\ell c^k)^k}} \wedge \frac{w}{c^k}, \end{aligned}$$

for some universal constant $c > 0$. Next, choose $k = 2$ and $\ell = c(d \wedge w^{1/2})$ when $p \leq 1/d$ for some universal constant $c > 0$. For $p > 1/d$ and $w \geq d$, we use Lemma 54 with $A = B = 1$ and $\Delta = cd$. Then $w(H) \asymp w(H') \asymp d^2$ and $|w(H) - w(H')| \asymp d\sqrt{\frac{d}{p}}$, and $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) < c < 1$. By Theorem 39, we have $\inf_{\widehat{w}} \sup_{G \in \mathcal{F}: \substack{d(G) \leq d \\ w(G) \leq w}} \mathbb{E}_G |\widehat{w} - w(G)|^2 \gtrsim \frac{wd}{p}$. \square

Proof of Theorem 47. Let R denote the minimax risk. The bound $R \lesssim \frac{td}{p^2}$ follows immediately from Theorem 40 with $\omega = 3$. For the other regimes, we modify the estimator (8.21) from Theorem 37. To accomplish this, observe that $\mathfrak{n}(\text{triangle}, G)$ is of order td for planar G , since the number of triangles that share a common vertex is at most d . Choosing $\alpha = \frac{1}{2qp^2}$ so that, in the notation of the proof of Theorem 37, $c_1 - 1 \asymp \frac{1}{p}$ and $c_2 - 1 = p^{-2} \left[2\alpha^2 qp^5 + (1 - 2q\alpha p^2)^2 \right] \asymp \frac{1}{p}$, we have $\text{Var}[\widehat{s}] \lesssim \frac{t}{p^3} \vee \frac{td}{p}$. This yields the bound $R \lesssim \frac{t}{p^3} \vee \frac{td}{p}$. Thus, $R \lesssim \left(\frac{t}{p^3} \vee \frac{td}{p} \right) \wedge \frac{td}{p^2} = \left(\frac{t}{p^3} \wedge \frac{td}{p^2} \right) \vee \frac{td}{p}$. For the lower bound, consider two cases:

Case I: $p \leq 1/d$. By Lemma 59, there exists two planar graphs H and H' on order ℓ vertices such that $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = O(p^2 + p^3 \ell^3)$ and $\mathfrak{t}(H) \asymp \mathfrak{t}(H') \asymp |\mathfrak{t}(H) - \mathfrak{t}(H')| \asymp \ell$

provided $p = O(1/\ell)$. We choose $\ell = p^{-1/3} \wedge t$ if $p > 1/d^3$. Otherwise, if $p \leq 1/d^3$, we choose $\ell = d \wedge t$. By Theorem 39, this produces a lower bound of $R \gtrsim \left(\frac{t}{p^{7/3}} \wedge \frac{td}{p^2}\right) \wedge t^2$.

Case II: $p > 1/d$. We use the second statement of Lemma 59 which guarantees the existence of two planar graphs H and H' on order ℓ vertices such that $\text{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = O(p)$ and $\mathfrak{t}(H) \asymp |\mathfrak{t}(H) - \mathfrak{t}(H')| \asymp \ell$. Choosing $\ell = d \wedge t$ yields the lower bound $R \gtrsim \frac{td}{p} \wedge t^2$. \square

Proof of Theorem 38. To make $\hat{\mathbf{e}}$ unbiased, in view of (8.13), we set

$$1 = \mathbb{E}[\mathcal{K}_A] = pq(f(d_u) + f(d_v)) + p^2 g(d_u, d_v).$$

This determines

$$g(d_u, d_v) = \frac{1 - pq(f(d_u) + f(d_v))}{p^2}.$$

An easy calculation shows that

$$\text{Var}[\mathcal{K}_A] = \frac{(1 - pq(f(d_u) + f(d_v)))^2}{p^2} + pq(f^2(d_u) + f^2(d_v)) - 1$$

and if $A = \{u, w\}$ and $A' = \{w, v\}$ in G , then

$$\text{Cov}[\mathcal{K}_A, \mathcal{K}_{A'}] = \frac{q}{p}(1 - pf(d_u))(1 - pf(d_v)).$$

Otherwise, $\text{Cov}[\mathcal{K}_A, \mathcal{K}_{A'}] = 0$ if A and A' do not intersect. Thus,

$$\begin{aligned} \text{Var}[\hat{\mathbf{e}}] &= \frac{q}{p} \sum_{u \neq v} d_{uv} (1 - pf(d_u))(1 - pf(d_v)) \\ &\quad + \sum_{\{u, v\} \in E(G)} \left[\frac{(1 - pq(f(d_u) + f(d_v)))^2}{p^2} + pq(f^2(d_u) + f^2(d_v)) - 1 \right], \end{aligned} \quad (8.52)$$

where d_{uv} denotes the cardinality of $N_G(u) \cap N_G(v)$. To gain a better idea for how to choose f , we first suppose that $f \equiv \alpha$. Thus, (8.52) reduces to the mean square error of (8.11) or

$$\text{Var}[\hat{\mathbf{e}}] = \frac{2q}{p} \mathfrak{n}(P_3, G)(1 - p\alpha)^2 + \mathfrak{e}(G) \frac{q}{p^2} (1 + p(1 - 2\alpha((p-2)p\alpha + 2)))$$

Next, let us minimize the above expression over all α . Doing so with

$$\alpha' = \left(\frac{1}{p}\right) \frac{pe(G) + pn(P_3, G)}{pn(P_3, G) + e(G)(2-p)} + \left(\frac{1}{2p}\right) \frac{2qe(G)}{pn(P_3, G) + e(G)(2-p)}.$$

yields

$$\text{Var}[\widehat{e}] = \frac{q^2}{p} \frac{e(G)(e(G) + n(P_3, G))}{(2-p)e(G) + pn(P_3, G)}. \quad (8.53)$$

Note that α' is a convex combination of $\frac{1}{p}$ and $\frac{1}{2p}$. These are the values that yield the risk bound for the non-adaptive estimator (8.11) in Theorem 36, viz.,

$$\alpha = \left(\frac{1}{p}\right) \mathbb{1}\left\{d > \frac{1}{p}\right\} + \left(\frac{1}{2p}\right) \mathbb{1}\left\{d \leq \frac{1}{p}\right\}.$$

Of course, this choice of α' is not feasible since it depends on the unknown quantities $e(G)$ and $n(P_3, G)$. However, noting that $e(G) = \sum_u d_u/2$ and $n(P_3, G) = \sum_u \binom{d_u}{2}$ inspires us to define

$$\begin{aligned} f(d_u) &= \left(\frac{1}{p}\right) \frac{p\binom{d_u}{2} + p\binom{d_u}{2}}{p\binom{d_u}{2} + \binom{d_u}{2}(2-p)} + \left(\frac{1}{2p}\right) \frac{2q\binom{d_u}{2}}{p\binom{d_u}{2} + \binom{d_u}{2}(2-p)} \\ &= \left(\frac{1}{2p}\right) \frac{2pd_u}{p(d_u - 1) + (2-p)} + \left(\frac{1}{2p}\right) \frac{2q}{p(d_u - 1) + (2-p)} \\ &= \frac{pd_u + q}{p(pd_u + 2q)}. \end{aligned}$$

With this choice of f , we will verify that the variance and covariance terms in (8.52) also yield the rate (8.18). Note that

$$\begin{aligned} \frac{q}{p} \sum_{u \neq v} d_{uv} \left[\frac{q}{pd_u + 2q} \right] \left[\frac{q}{pd_v + 2q} \right] &\leq \frac{q}{p} \sum_{u \neq v} \frac{d_{uv}}{pd_u + 2q} \leq \frac{dq}{p} \sum_u \frac{d_u}{pd_u + 2q} \\ &\leq \frac{Ndq}{p} \frac{e(G)}{pe(G) + qN} \leq \frac{Nd}{p^2} \wedge \frac{e(G)d}{p}, \end{aligned}$$

where the second last inequality follows from the concavity of $x \mapsto \frac{x}{px+2q}$ for $x \geq 0$. The variance term has the bound

$$\sum_{\{u,v\} \in E(G)} \left[\frac{(1 - pq(f(d_u) + f(d_v)))^2}{p^2} + pq(f^2(d_u) + f^2(d_v)) - 1 \right] \lesssim e(G) \left(\left(\frac{1}{p^2} \wedge d^2 \right) \vee \frac{1}{p} \right),$$

which follows from $\frac{(1-pq(f(d_u)+f(d_v)))^2}{p^2} \lesssim \frac{1}{p^2} \wedge (d_u^2 + d_v^2)$ and $pq(f^2(d_u) + f^2(d_v)) \lesssim \frac{1}{p}$. \square

8.9 Neighborhood sampling without colors

In this appendix we demonstrate the usefulness of the color information (namely, which vertices are sampled) in neighborhood sampling by showing that without observing the colors, the performance guarantees in Theorem 36 are no longer unattainable in certain regimes.

Theorem 48. *Let \mathcal{F} denote the collection of all forests. Consider the neighborhood sampling model without observing the colors $\{b_v : v \in V\}$. Then*

$$\inf_{\hat{\mathbf{e}}} \sup_{G \in \mathcal{F}: \mathbf{d}(G) \leq d, \mathbf{e}(G) \leq m} \mathbb{E}_G |\hat{\mathbf{e}} - \mathbf{e}(G)|^2 \gtrsim mp(d \wedge m). \quad (8.54)$$

Proof. Let $M = m/k$, where $k = d \wedge m$ and set $\mathcal{F}_0 = \{G_{\underline{\theta}} : G_{\underline{\theta}} = S_{\theta_1} + \dots + S_{\theta_M}, \underline{\theta} = (\theta_1, \dots, \theta_M) \in [k]^M\}$. Note that for each $\underline{\theta} \in [k]^M$, $\mathbf{e}(G_{\underline{\theta}}) = \|\underline{\theta}\|_1$. Thus, if $\underline{X} = (X_1, \dots, X_M)$, where $\{X_i\}$ are independent and $X_i \sim p\delta_{\theta_i} + q\text{Bin}(\theta_i, p)$ for $i \in [M]$, then

$$\inf_{\hat{\mathbf{e}}} \sup_{G \in \mathcal{F}: \mathbf{d}(G) \leq d, \mathbf{e}(G) \leq m} \mathbb{E}_G |\hat{\mathbf{e}} - \mathbf{e}(G)|^2 \geq \inf_g \sup_{\underline{\theta} \in [d]^M} \mathbb{E}_{\underline{\theta}} \|\underline{\theta}\|_1 - g(\underline{X})|^2.$$

By the minimax theorem,

$$\begin{aligned} \inf_g \sup_{\underline{\theta} \in [k]^M} \mathbb{E}_{\underline{\theta}} \|\underline{\theta}\|_1 - g(\underline{X})|^2 &= \sup_{\underline{\theta} \in \pi} \inf_g \mathbb{E}_{\underline{\theta}} \|\underline{\theta}\|_1 - g(\underline{X})|^2 = \sup_{\underline{\theta} \in \pi} \mathbb{E}_{\underline{X}} \mathbb{E}_{\underline{\theta}|\underline{X}} \|\underline{\theta}\|_1 - \mathbb{E}_{\underline{\theta}|\underline{X}} \|\underline{\theta}\|_1|^2 \\ &\geq \sup_{\underline{\theta} \in \pi^{\otimes M}} \mathbb{E}_{\underline{X}} \mathbb{E}_{\underline{\theta}|\underline{X}} \|\underline{\theta}\|_1 - \mathbb{E}_{\underline{\theta}|\underline{X}} \|\underline{\theta}\|_1|^2 = M \sup_{\theta \in \pi} \mathbb{E}_X \mathbb{E}_{\theta|X} |\theta - \mathbb{E}_{\theta|X} \theta|^2 \\ &= M \inf_g \sup_{\theta \in [d \wedge m]} \mathbb{E}_{\theta} |\theta - g(X)|^2 \asymp m \left(pk \vee \left(\frac{1}{p} \wedge k \right) \right) \\ &\gtrsim mp(d \wedge m), \end{aligned}$$

where $X \sim \delta_{\theta} + q\text{Bin}(\theta, p)$ and the second to last line follows from Lemma 60 below. \square

Remark 17. *Note that when $p > (1/d)^{1/3}$ and $m \geq d$, the minimax lower bound (8.54) is*

strictly greater than the minimax risk in Theorem 42, thus confirming the intuition that the knowledge of which vertices are sampled provide useful information. On the other hand, the Horvitz-Thompson estimator (8.8) can be implemented without the color information and achieve the error bound $O(\frac{md}{p})$ in (8.10). Comparing with Theorem 35, we conclude that neighborhood sampling is at least as informative as subgraph sampling, even if the colors are not observed. This is intuitive because neighborhood sampling reveals more edges from the parent graph.

Lemma 60. *Given $\theta \in [k]$, let X be distributed according to $p\delta_\theta + q\text{Bin}(\theta, p)$. Assume that $p \leq 1/2$. Then*

$$\inf_g \sup_{\theta \in [k]} \mathbb{E}_\theta[|\theta - g(X)|^2] \asymp pk^2 \vee \left(\frac{k}{p} \wedge k^2\right). \quad (8.55)$$

Moreover, the minimax rate is achieved by the estimator $\hat{g}(X) = k \wedge \frac{X}{p}$.

Proof. Denote the minimax risk by R . Let id denote the identity map. Given any estimator g , without loss of generality, we assume $g : \{0, \dots, k\} \rightarrow [0, k]$. Since $\mathbb{E}_\theta[(\theta - g(X))^2] = p(\theta - g(\theta))^2 + q\mathbb{E}_{X \sim \text{Bin}(\theta, p)}[(\theta - g(X))^2]$, we have

$$\sup_{\theta \in [k]} \mathbb{E}_\theta[|\theta - g(X)|^2] \geq p\|\text{id} - g\|_\infty^2. \quad (8.56)$$

Also, $(\theta - g(X))^2 \geq -(X - g(X))^2 + (\theta - X)^2/2$, and hence

$$\mathbb{E}_{X \sim \text{Bin}(\theta, p)}[(\theta - g(X))^2] \geq -\|\text{id} - g\|_\infty^2 + \frac{1}{2}(q^2\theta^2 + pq\theta).$$

Therefore

$$\sup_{\theta \in [k]} \mathbb{E}_\theta[|\theta - g(X)|^2] \geq -q\|\text{id} - g\|_\infty^2 + \frac{q}{2}(q^2k^2 + pqk). \quad (8.57)$$

Combining (8.56) and (8.57), we get

$$R \geq \frac{pq}{2}(q^2k^2 + pqk) \asymp pk^2. \quad (8.58)$$

Next by the minimax theorem,

$$R = \sup_{\pi} \inf_g \mathbb{E}_{\pi} [|\theta - g(X)|^2] = \sup_{\pi} \inf_g \left(\underbrace{p \mathbb{E}_{\theta \in \pi} [|\theta - g(\theta)|^2]}_{\in [0, k^2]} + q \mathbb{E}_{\theta \in \pi, X \sim \text{Bin}(\theta, p)} [|\theta - g(X)|^2] \right).$$

We also know that

$$\sup_{\pi} \inf_g \mathbb{E}_{\theta \in \pi, X \sim \text{Bin}(\theta, p)} [|\theta - g(X)|^2] = \inf_g \sup_{\theta \in [k]} \mathbb{E}_{\theta \in \pi, X \sim \text{Bin}(\theta, p)} [|\theta - g(X)|^2] \asymp \frac{k}{p} \wedge k^2.$$

Therefore we have

$$\frac{k}{p} \wedge k^2 \lesssim R \lesssim pk^2 + \frac{k}{p} \wedge k^2.$$

Combining with (8.58) yields the characterization (8.55). \square

8.10 Lower bounds for other motifs

Theorem 49 (Wedges). *For neighborhood sampling with sampling ratio p ,*

$$\inf_{\widehat{\mathbf{w}}} \sup_{G: \substack{\mathbf{d}(G) \leq d \\ \mathbf{w}(G) \leq w}} \mathbb{E}_G |\widehat{\mathbf{w}} - \mathbf{w}(G)|^2 \asymp \frac{wd}{p^2} \wedge w^2.$$

Proof. For the lower bound, consider two cases:

Case I: $p \leq 1/d$. Let $h = P_5$ and $h' = K_3 + K_2$. For each node in the original graph, we associate ℓ distinct isolated vertices and connect each pair of vertices by an edge if and only if they were connected in the original graph. Call these expanded graphs H and H' . Note that H and H' have matching degree sequences $(2, 2, 2, 1, 1)$ and hence $\text{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) = O(\ell^2 p^2)$. Furthermore, $\mathbf{s}(P_3, H) \asymp \mathbf{s}(P_3, H') \asymp |\mathbf{s}(P_3, H) - \mathbf{s}(P_3, H')| \asymp \ell^3$. If $\ell = c(d \wedge w^{1/3})$, then by Theorem 39 with $M = w/\ell^3$, $\inf_{\widehat{\mathbf{w}}} \sup_{G \in \mathcal{G}(w, d)} \mathbb{E}_G |\widehat{\mathbf{w}} - \mathbf{w}(G)|^2 \gtrsim \frac{w\ell}{p^2} \wedge w^2 \asymp \frac{wd}{p^2} \wedge w^2$.

Case II: $p > 1/d$. We use Lemma 55 with $G = K_{\ell}$, $H_1 = K_{1/p} + K_{1/p}$, and $H_2 = K_{2/p}$. This gives us two graphs H and H' with $\mathbf{s}(P_3, H) = |\mathbf{s}(P_3, H) - \mathbf{s}(P_3, H')| \asymp \ell/p^2$. By Theorem 39 with $M = w/(\ell/p^2)$, $\inf_{\widehat{\mathbf{w}}} \sup_{G \in \mathcal{G}(w, d)} \mathbb{E}_G |\widehat{\mathbf{w}} - \mathbf{w}(G)|^2 \gtrsim \frac{w\ell}{p^2} \wedge w^2$. Let $\ell = cd$

if $\frac{d}{p^2} \leq w$ and $\ell = cp^2w$ if $\frac{d}{p^2} > w$, for some small constant c . In either case, we find that $w(H) \leq w$, $w(H') \leq w$, and $\inf_{\widehat{w}} \sup_{G \in \mathcal{G}(w,d)} \mathbb{E}_G |\widehat{w} - w(G)|^2 \asymp \frac{wd}{p^2} \wedge w^2$. \square

Lower bound for motifs of size four It remains to show that holds the result in Theorem 40 that holds for K_4 , namely,

$$\inf_{\widehat{s}} \sup_{\substack{G: d(G) \leq d \\ s(h,G) \leq s}} \mathbb{E}_G |\widehat{s} - s(K_\omega, G)|^2 = \Theta \left(\frac{sd}{p^3} \wedge \frac{sd^2}{p^2} \wedge s^2 \right) \quad (8.59)$$

continues to hold for $h = \text{---}\circ\text{---}\circ\text{---}\circ, \text{---}\circ\text{---}\circ\text{---}\circ, \text{---}\circ\text{---}\circ\text{---}\circ$ and $\text{---}\circ\text{---}\circ\text{---}\circ$. For the case of $p < 1/d$, the construction for K_4 in (8.37) works simultaneously for all motifs, because each motif is contained in one of H and H' and not the other. Next we consider the case of $p > 1/d$. The construction is ad hoc and similar to those in Theorem 35 and Theorem 37.

- For $h = \text{---}\circ\text{---}\circ\text{---}\circ$, we use the clique construction: label the root as v_1 and the leaves as v_2, v_3, v_4 . Define the graph H as follows: Expand v_1 into a clique S_1 of size ℓ , and for $i = 2, 3, 4$, expand each v_i into a clique S_i of size $1/p$. Connect each pair of vertices $u_i \in S_i$ and $u_j \in S_j$ for $i \neq j$ if and only if v_i and v_j are connected in the motif h . This defines a graph H on $\ell + 3/p$ vertices. Repeat the same construction with h replaced by $\text{---}\circ\text{---}\circ\text{---}\circ$, where the degree-one vertex is v_1 . Note that if we remove the edges between the set of vertices $T \triangleq S_2 \cup S_3 \cup S_4$, for H and H' the resulting graph is isomorphic. Thus by Lemma 55, we have $\text{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq 1 - (1-p)^{3/p} \leq 0.9$ if $p \leq 1/2$. Furthermore, note that $s(\text{---}\circ\text{---}\circ\text{---}\circ, H') = 0$ and $s(\text{---}\circ\text{---}\circ\text{---}\circ, H) = \ell/p^3$. Finally, taking $\ell = c(d \wedge \frac{s}{\ell/p^3})$ for some small constant c and invoking Theorem 39, we obtain the desired lower bound $\frac{sd}{p^3} \wedge s^2$ in (8.59).
- For $h = \text{---}\circ\text{---}\circ\text{---}\circ$, use the same construction as above with H and H' swapped.
- For $h = \text{---}\circ\text{---}\circ\text{---}\circ$, we repeat the clique construction of H with v_1 being any of the degree-three vertices in h , and of H' with $h' = \text{---}\circ\text{---}\circ\text{---}\circ$; in other words, we simply have $H' = K_{\ell+3/p}$.
- For $h = \text{---}\circ\text{---}\circ\text{---}\circ$, we repeat the clique construction of H with v_1 being any vertex in h , and of H' with $h' = \text{---}\circ\text{---}\circ\text{---}\circ$, with v_1 being the degree-two vertices.

Bibliography

- [1] S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.*, 45(1):77–120, 2017.
- [2] J. Xu, D. J. Hsu, and A. Maleki. Global analysis of expectation maximization for mixtures of two Gaussians. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2676–2684. Curran Associates, Inc., 2016.
- [3] A. R. Barron. Neural net approximation. *Yale Workshop on Adaptive and Learning Systems, Yale University Press*, 1992.
- [4] J. E. Yukich, M. B. Stinchcombe, and H. White. Sup-norm approximation bounds for networks through probabilistic methods. *IEEE Trans. Inform. Theory*, 41(4):1021–1027, 1995.
- [5] L. Breiman. Hinging hyperplanes for regression, classification, and function approximation. *IEEE Trans. Inform. Theory*, 39(3):999–1013, 1993.
- [6] **Klusowski, Jason M.** and W. D. Brinda. Statistical guarantees for estimating the centers of a two-component Gaussian mixture by EM. *arXiv preprint arXiv:1608.02280*, 2016.
- [7] **Klusowski, Jason M.**, W. D. Brinda, and D. Yang. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *arXiv preprint arXiv:1704.08231*, 2017.

- [8] P. Hall. On estimating the endpoint of a distribution. *Ann. Statist.*, 10(2):556–568, 1982.
- [9] A. Goldenshluger and A. Tsybakov. Estimating the endpoint of a distribution in the presence of additive observation errors. *Statist. Probab. Lett.*, 68(1):39–49, 2004.
- [10] P. Hall and L. Simar. Estimating a changepoint, boundary, or frontier in the presence of observation error. *J. Amer. Statist. Assoc.*, 97(458):523–534, 2002.
- [11] C. R. Genovese, M. Perone-Pacifco, I. Verdinelli, and L. Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.*, 40(2):941–963, 2012.
- [12] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272, 1991.
- [13] V.-E. Brunel, **Klusowski, Jason M.**, and D. Yang. Estimation of convex supports from noisy measurements. <https://sites.google.com/a/yale.edu/jason-klusowski/research>, 2017.
- [14] V.-E. Brunel, **Klusowski, Jason M.**, and D. Yang. Estimation of a convex density support using contaminated data. <https://sites.google.com/a/yale.edu/jason-klusowski/research>, 2017.
- [15] L. A. Goodman. On the estimation of the number of classes in a population. *Ann. Math. Statistics*, 20:572–579, 1949.
- [16] O. Frank. Estimation of the number of connected components in a graph by using a sampled subgraph. *Scand. J. Statist.*, 5(4):177–188, 1978.
- [17] **Klusowski, Jason M.** and Y. Wu. Estimating the number of connected components in a graph via subgraph sampling. *arXiv preprint arXiv:1801.04339*, 2018.
- [18] **Klusowski, Jason M.** and Y. Wu. Counting motifs with graph sampling. *arXiv preprint arXiv:1802.07773*, 2018.

- [19] A. J. Cookson and M. Niessner. Why Don't We Agree? Evidence from a Social Network of Investors. *Available at <http://dx.doi.org/10.2139/ssrn.2754086>*, 2016.
- [20] W. D. Brinda and **Klusowski, Jason M.** Finite-sample risk bounds for maximum likelihood estimation with arbitrary penalties. To appear in *IEEE Transactions on Information Theory*, 2018.
- [21] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *J. Amer. Statist. Assoc.*, 76(376):817–823, 1981.
- [22] A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.
- [23] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945, 1993.
- [24] W. S. Lee, P. L. Bartlett, and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inform. Theory*, 42(6, part 2):2118–2132, 1996.
- [25] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [26] M. Anthony and P. L. Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press, Cambridge, 1999.
- [27] **Klusowski, Jason M.** and A. R. Barron. Approximation by combinations of relu and squared relu ridge functions with ℓ^1 and ℓ^0 controls. Revise and resubmit to *IEEE Transactions on Information Theory*, *arXiv preprint arXiv:1607.07819*, 2018.
- [28] A. R. Barron, C. Huang, J. Li, and X. Luo. The MDL principle, penalized likelihoods, and statistical risk. *Workshop on Information Theory Methods in Science and Engineering, Tampere, Finland*, 2008.

- [29] A. R. Barron, C. Huang, J. Li, and X. Luo. The MDL principle, penalized likelihoods, and statistical risk. *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*, Tampere University Press, Tampere, Finland. Editor Ioan Tabus, 2008.
- [30] G. L. Cheang. Neural network approximation and estimation of functions. *Yale University, Department of Statistics PhD Thesis*, 1998.
- [31] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011.
- [32] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.
- [33] C. Huang, G. L. Cheang, and A. R. Barron. Risk of penalized least squares, greedy selection and ℓ_1 penalization for flexible function libraries. *Yale University, Department of Statistics technical report*, 2008.
- [34] C. Huang. Risk of penalized least squares, greedy selection and ℓ_1 -penalization for flexible function libraries. *Yale University, Department of Statistics PhD thesis*, 2008.
- [35] **Klusowski, Jason M.** and A. R. Barron. Minimax lower bounds for ridge combinations including neural nets. *Proceedings IEEE International Symposium on Information Theory*, Aachen, Germany, pages 1377–1380, June, 2017.
- [36] P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Trans. Inform. Theory*, 44(2):525–536, 1998.
- [37] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.
- [38] A. R. Barron, A. Cohen, W. Dahmen, and R. A. DeVore. Approximation and learning by greedy algorithms. *Ann. Statist.*, 36(1):64–94, 2008.

- [39] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.
- [40] M. Anthony and P. L. Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press, Cambridge, 1999.
- [41] T. Zhang. Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.*, 37(5A):2109–2144, 2009.
- [42] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712, 2000.
- [43] J. Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*, 2017.
- [44] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A tensor approach to learning mixed membership community models. *J. Mach. Learn. Res.*, 15:2239–2312, 2014.
- [45] M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.
- [46] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.
- [47] S. Ioannidis and A. Montanari. Learning combinations of sigmoids through gradient estimation. *arXiv preprint arXiv:1708.06678*, 2017.
- [48] A. Andoni, R. Panigrahy, G. Valiant, and L. Zhang. Learning polynomials with neural networks. In *International Conference on Machine Learning*, pages 1908–1916, 2014.
- [49] Y. Zhang, J. D. Lee, and M. I. Jordan. ℓ_1 -regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning*, pages 993–1001, 2016.
- [50] Y. Zhang, J. D. Lee, M. J. Wainwright, and M. I. Jordan. Learning halfspaces and neural networks with random initialization. *arXiv preprint arXiv:1511.07948*, 2015.

- [51] L. K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist.*, 20(1):608–613, 1992.
- [52] A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 148–155, 1998.
- [53] S. Mendelson. On the size of convex hulls of small sets. *J. Mach. Learn. Res.*, 2(1):1–18, 2002.
- [54] X. Chen and H. White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Trans. Inform. Theory*, 45(2):682–691, 1999.
- [55] Y. Makovoz. Random approximants and neural networks. *J. Approx. Theory*, 85(1):98–109, 1996.
- [56] J. Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.
- [57] **Klusowski, Jason M.** and A. R. Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv: 1607.01434*, 2018.
- [58] G. H. L. Cheang and A. R. Barron. A better approximation for balls. *J. Approx. Theory*, 104(2):183–203, 2000.
- [59] A. Brutzkus and A. Globerson. Globally optimal gradient descent for a ConvNet with Gaussian inputs. *arXiv Preprint*, February, 2017.
- [60] V. N. Vapnik and A. J. Červonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, 16:264–279, 1971.
- [61] D. Haussler. Sphere packing numbers for subsets of the Boolean n -cube with bounded Vapnik-Chervonenkis dimension. *J. Combin. Theory Ser. A*, 69(2):217–232, 1995.

- [62] Y. Makovoz. Uniform approximation by neural networks. *J. Approx. Theory*, 95(2):215–228, 1998.
- [63] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.
- [64] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.
- [65] V. Kůrková and M. Sanguinetti. Estimates of covering numbers of convex sets with slowly decaying orthogonal subsets. *Discrete Appl. Math.*, 155(15):1930–1942, 2007.
- [66] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054, 1991.
- [67] A. R. Barron and S. Chatterjee. Information theoretic validity of penalized likelihood. *Proceedings IEEE International Symposium on Information Theory, Honolulu, HI*, pages 3027–3031, June 2014.
- [68] I. A. Ibragimov and R. Z. Hasminskii. Nonparametric regression estimation. *Dokl. Akad. Nauk SSSR*, 252(4):780–784, 1980.
- [69] L. Birgé. Approximation dans les espaces métriques et théorie de l’estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237, 1983.
- [70] L. Birgé. On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields*, 71(2):271–291, 1986.
- [71] F. Gao, C.-K. Ing, and Y. Yang. Metric entropy and sparse linear approximation of ℓ_q -hulls for $0 < q \leq 1$. *J. Approx. Theory*, 166:42–55, 2013.
- [72] R. L. Graham and N. J. A. Sloane. Lower bounds for constant weight codes. *IEEE Trans. Inform. Theory*, 26(1):37–43, 1980.

- [73] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.
- [74] C. Daskalakis, C. Tzamos, and M. Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. *arXiv Preprint*, September, 2016.
- [75] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214, 1994.
- [76] Y. Chen, X. Yi, and C. Caramanis. A convex formulation for mixed regression with two components: minimax optimal rates. *In COLT*, pages 560–604, 2014.
- [77] X. Yi, C. Caramanis, and S. Sanghavi. Alternating minimization for mixed linear regression. *In ICML*, pages 613–621, 2014.
- [78] A. T. Chaganty and P. Liang. Spectral experts for estimating mixtures of linear regressions. *In ICML*, (2013):1040–1048, 2013.
- [79] K. Zhong, P. Jain, and I. S. Dhillon. Mixed linear regression with multiple components. *In Advances in Neural Information Processing Systems*, pages 2190–2198, 2016.
- [80] H. Sedghi, M. Janzamin, and A. Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. *arXiv Preprint*, 2014.
- [81] R. D. De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, Nov 1989.
- [82] K. Viele and B. Tong. Modeling with mixtures of linear regressions. *Stat. Comput.*, 12(4):315–330, 2002.
- [83] S. Faria and G. Soromenho. Fitting mixtures of linear regressions. *J. Stat. Comput. Simul.*, 80(1-2):201–225, 2010.
- [84] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.

- [85] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.
- [86] A. Rényi and R. Sulanke. Über die konvexe Hülle von n zufällig gewählten Punkten. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 2:75–84 (1963), 1963.
- [87] A. Rényi and R. Sulanke. Über die konvexe Hülle von n zufällig gewählten Punkten. II. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 3:138–147 (1964), 1964.
- [88] A. P. Korostelëv and A. B. Tsybakov. *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993.
- [89] V.-E. Brunel. Adaptive estimation of convex and polytopal density support. *Probab. Theory Related Fields*, 164(1-2):1–16, 2016.
- [90] V.-E. Brunel. On uniform performances of random polytopes and their functionals in convex bodies. *Preprint*, 2015.
- [91] B. Chazelle. An optimal convex hull algorithm in any fixed dimension. *Discrete & Computational Geometry*, 10(4):377–409, 1993.
- [92] S. Loustau and C. Marteau. Minimax fast rates for discriminant analysis with errors in variables. *Bernoulli*, 21(1):176–208, 2015.
- [93] R. J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83(404):1184–1186, 1988.
- [94] J. Fan and Y. K. Truong. Nonparametric regression with errors in variables. *Ann. Statist.*, 21(4):1900–1925, 1993.
- [95] P. Hall and S. N. Lahiri. Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Statist.*, 36(5):2110–2134, 2008.
- [96] A. Meister. Estimating the support of multivariate densities under measurement error. *J. Multivariate Anal.*, 97(8):1702–1717, 2006.
- [97] A. Meister. Deconvolving compactly supported densities. *Math. Methods Statist.*, 16(1):63–76, 2007.

- [98] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.
- [99] S. Loustau and C. Marteau. Noisy classification with boundary assumptions. *Preprint*, 2013.
- [100] S. Csörgő and D. M. Mason. Simple estimators of the endpoint of a distribution. In *Extreme value theory (Oberwolfach, 1987)*, volume 51 of *Lecture Notes in Statist.*, pages 132–147. Springer, New York, 1989.
- [101] A. Goldenshluger and A. Zeevi. Recovering convex boundaries from blurred and noisy observations. *Ann. Statist.*, 34(3):1375–1394, 2006.
- [102] C. Thäle. 50 years sets with positive reach-a survey-. *Surveys in Mathematics & its Applications*, 3, 2008.
- [103] V.-E. Brunel. Concentration of the empirical level sets of tukey’s halfspace depth. *Preprint*, 2016.
- [104] V.-E. Brunel. A universal deviation inequality for random polytopes. *Preprint*, 2013.
- [105] L. Hörmander. *The analysis of linear partial differential operators. I.* Classics in Mathematics. Springer-Verlag, Berlin, 2003. Distribution theory and Fourier analysis, Reprint of the second (1990) edition [Springer, Berlin; MR1065993 (91m:35001a)].
- [106] R. D. Luce and A. D. Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.
- [107] C. L. Apicella, F. W. Marlowe, J. H. Fowler, and N. A. Christakis. Social networks and cooperation in hunter-gatherers. *Nature*, 481(7382):497–501, 01 2012.
- [108] T. G. Conley and C. R. Udry. Learning about a new technology: Pineapple in ghana. *American Economic Review*, 100(1):35–69, March 2010.
- [109] M. Fafchamps and S. Lund. Risk-sharing networks in rural philippines. *Journal of development Economics*, 71(2):261–287, 2003.

- [110] O. Bandiera and I. Rasul. Social networks and technology adoption in northern mozambique. *The Economic Journal*, 116(514):869–902, 2006.
- [111] B. Feigenberg, E. M. Field, and R. Pande. Building social capital through microfinance. Technical report, National Bureau of Economic Research, 2010.
- [112] A. Chandrasekhar and R. Lewis. Econometrics of sampled networks. 2011.
- [113] P. H. Reingen and J. B. Kernan. Analysis of referral networks in marketing: Methods and illustration. *Journal of Marketing Research*, pages 370–378, 1986.
- [114] M. P. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008.
- [115] R. Govindan and H. Tangmunarunkit. Heuristics for internet map discovery. In *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 3, pages 1371–1380. IEEE, 2000.
- [116] A. Ben-Hamou, R. I. Oliveira, and Y. Peres. Estimating graph parameters via random walks with restarts. *arXiv preprint arXiv:1709.00869*, 2017.
- [117] L. A. Goodman. Snowball sampling. *Ann. Math. Statist.*, 32:148–170, 1961.
- [118] M. J. Salganik and D. D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology*, 34(1):193–240, 2004.
- [119] O. Goldreich and D. Ron. Approximating average parameters of graphs. *Random Structures Algorithms*, 32(4):473–493, 2008.
- [120] T. Eden, A. Levi, D. Ron, and C. Seshadhri. Approximately counting triangles in sublinear time. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015*, pages 614–633. IEEE Computer Soc., Los Alamitos, CA, 2015.

- [121] M. Aliakbarpour, A. S. Biswas, T. Gouleakis, J. Peebles, R. Rubinfeld, and A. Yodpinyanee. Sublinear-time algorithms for counting star subgraphs via edge sampling. *Algorithmica*, pages 1–30.
- [122] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 68–75. Springer, 2011.
- [123] O. Goldreich. *Introduction to Property Testing*. Cambridge University, 2017.
- [124] O. Frank. Estimation of graph totals. *Scand. J. Statist.*, 4(2):81–89, 1977.
- [125] M. Capobianco. Estimating the connectivity of a graph. *Graph Theory and Applications*, pages 65–74, 1972.
- [126] B. Chen, A. Shrivastava, and R. C. Steorts. Unique entity estimation with application to the Syrian conflict. *arXiv preprint arXiv:1710.02690*, 2017.
- [127] G. Cormode and N. Duffield. Sampling for big data: a tutorial. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1975–1975. ACM, 2014.
- [128] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 631–636. ACM, 2006.
- [129] E. D. Kolaczyk. *Topics at the Frontier of Statistics and Network Analysis: (Re)Visiting the Foundations*. SemStat Elements. Cambridge University Press, 2017.
- [130] E. D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer Science & Business Media, 2009.
- [131] M. S. Handcock and K. J. Gile. Modeling social networks from sampled data. *Ann. Appl. Stat.*, 4(1):5–25, 2010.
- [132] B. Chazelle, R. Rubinfeld, and L. Trevisan. Approximating the minimum spanning tree weight in sublinear time. *SIAM J. Comput.*, 34(6):1370–1379, 2005.

- [133] P. Berenbrink, B. Krayenhoff, and F. Mallmann-Trenn. Estimating the number of connected components in sublinear time. *Inform. Process. Lett.*, 114(11):639–642, 2014.
- [134] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.
- [135] C. Gao, Y. Lu, and H. H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.
- [136] P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.
- [137] L. Lovász. *Large Networks and Graph Limits*, volume 60. American Mathematical Society, 2012.
- [138] C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztegombi. Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing. *Adv. Math.*, 219(6):1801–1851, 2008.
- [139] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998.
- [140] A. Orlitsky, A. T. Suresh, and Y. Wu. Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA*, 113(47):13283–13288, 2016.
- [141] Y. Wu and P. Yang. Sample complexity of the distinct element problem. *arxiv preprint arxiv:1612.03375*, Apr 2016.
- [142] D. B. West. *Introduction to graph theory*. Prentice Hall, Inc., Upper Saddle River, NJ, 1996.
- [143] E. W. McMahon, B. A. Shimkus, and J. A. Wolfson. Chordal graphs and the characteristic polynomial. *Discrete Math.*, 262(1-3):211–219, 2003.

- [144] K. Dohmen. Lower bounds for the probability of a union via chordal graphs. *Electronic Communications in Probability*, 18, 2013.
- [145] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.
- [146] D. J. Rose, R. E. Tarjan, and G. S. Lueker. Algorithmic aspects of vertex elimination on graphs. *SIAM J. Comput.*, 5(2):266–283, 1976.
- [147] R. O’Donnell. *Analysis of Boolean functions*. Cambridge University Press, New York, 2014.
- [148] S. Janson. Large deviations for sums of partly dependent random variables. *Random Structures Algorithms*, 24(3):234–248, 2004.
- [149] M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications, Inc., New York, 1992. Reprint of the 1972 edition.
- [150] Y. Zhang, E. D. Kolaczyk, and B. D. Spencer. Estimating network degree distributions under sampling: an inverse problem, with applications to monitoring social media networks. *Ann. Appl. Stat.*, 9(1):166–199, 2015.
- [151] A. Natanzon, R. Shamir, and R. Sharan. A polynomial approximation algorithm for the minimum fill-in problem. *SIAM J. Comput.*, 30(4):1067–1079, 2000.
- [152] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.
- [153] P. Erdős, L. Lovász, and J. Spencer. Strong independence of graphcopy functions. In *Graph theory and related topics (Proc. Conf., Univ. Waterloo, Waterloo, Ont., 1977)*, pages 165–172. Academic Press, New York-London, 1979.
- [154] H. Whitney. The coloring of graphs. *Ann. of Math. (2)*, 33(4):688–718, 1932.

- [155] W. L. Kocay. Some new methods in reconstruction theory. In *Combinatorial mathematics, IX (Brisbane, 1981)*, volume 952 of *Lecture Notes in Math.*, pages 89–114. Springer, Berlin-New York, 1982.
- [156] B. D. McKay and S. a. P. Radziszowski. Subgraph counting identities and Ramsey numbers. *J. Combin. Theory Ser. B*, 69(2):193–209, 1997.
- [157] R. E. Tarjan and M. Yannakakis. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM J. Comput.*, 13:566–579, 1984.
- [158] P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.
- [159] Y. Polyanskiy, A. T. Suresh, and Y. Wu. Sample complexity of population recovery. In *Proceedings of Conference on Learning Theory (COLT)*, Amsterdam, Netherland, Jul 2017. arXiv:1702.05574.
- [160] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.
- [161] N. Pržulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.
- [162] U. Feige. On sums of independent random variables with unbounded variance and estimating the average degree in a graph. *SIAM J. Comput.*, 35(4):964–984, 2006.
- [163] M. Gonen, D. Ron, and Y. Shavitt. Counting stars and other small subgraphs in sublinear-time. *SIAM J. Discrete Math.*, 25(3):1365–1411, 2011.
- [164] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, et al. A comprehensive analysis of protein–protein interactions in *saccharomyces cerevisiae*. *Nature*, 403(6770):623, 2000.
- [165] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

- [166] J.-D. J. Han, D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, 23(7):839, 2005.
- [167] C. Gao and J. Lafferty. Testing network structure using relations between small subgraph probabilities. *arXiv preprint arXiv:1704.06742*, 2017.
- [168] P. J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *Ann. Statist.*, 39(5):2280–2301, 2011.
- [169] E. Mossel, J. Neeman, and A. Sly. Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields*, 162(3-4):431–461, 2015.
- [170] H. Chen, N. Zhang, et al. Graph-based change-point detection. *The Annals of Statistics*, 43(1):139–176, 2015.
- [171] L. Chu and H. Chen. Asymptotic distribution-free change-point detection for modern data. *arXiv preprint arXiv:1707.00167*, 2017.
- [172] M. Capobianco. Estimating the connectivity of a graph. In Y. Alavi, D. R. Lick, and A. T. White, editors, *Graph Theory and Applications*, pages 65–74, Berlin, Heidelberg, 1972. Springer Berlin Heidelberg.
- [173] O. Lepski, A. Nemirovski, and V. Spokoiny. On estimation of the L_r norm of a regression function. 113(2):221–253, 1999.
- [174] T. Cai and M. G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. 39(2):1012–1041, 2011.
- [175] Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.
- [176] J. Jiao, K. Venkat, Y. Han, and T. Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

- [177] L. Lovász and B. Szegedy. The graph theoretic moment problem. *arXiv preprint arXiv:1010.5159*, 2010.
- [178] N. Biggs. On cluster expansions in graph theory and physics. *Quart. J. Math. Oxford Ser. (2)*, 29(114):159–173, 1978.
- [179] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. <http://snap.stanford.edu/data/egonets-Facebook.html>, June 2014.
- [180] M. Nair. On Chebyshev-type inequalities for primes. *Amer. Math. Monthly*, 89(2):126–129, 1982.
- [181] D. Hanson. On the product of the primes. *Canad. Math. Bull.*, 15:33–37, 1972.