Combinatorial Methods in Statistics

by

Paxton Mark Turner

B.S., Louisiana State University (2015)

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Department of Mathematics
May 7, 2021

Certified by .....
Philippe Rigollet
Professor of Mathematics
Thesis Supervisor

Accepted by .....
Jonathan Kelner
Graduate Chair, Applied Mathematics

Combinatorial Methods in Statistics

by

Paxton Mark Turner

Submitted to the Department of Mathematics
on May 7, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis explores combinatorial methods in random vector balancing, nonparametric estimation, and network inference. First, motivated by problems from controlled experiments, we study random vector balancing from the perspective of discrepancy theory, a classical topic in combinatorics, and give sharp statistical results along with improved algorithmic guarantees. Next, we focus on the problem of density estimation and investigate the fundamental statistical limits of coresets, a popular framework for obtaining algorithmic speedups by replacing a large dataset with a representative subset. In the following chapter, motivated by the problem of fast evaluation of kernel density estimators, we demonstrate how a multivariate interpolation scheme from finite-element theory based on the combinatorial-geometric properties of a certain mesh can be used to significantly improve the storage and query time of a nonparametric estimator while also preserving its accuracy. Our final chapter focuses on pedigree reconstruction, a combinatorial inference task of recovering the latent network of familial relationships of a population from its extant genetic data.

Thesis Supervisor: Philippe Rigollet
Title: Professor of Mathematics

Acknowledgments

I am very grateful to have been advised by Philippe Rigollet. Thanks to his guidance I learned a lot of interesting mathematics and statistics, and I also learned how to tell a good story about a project and to view it in the context of broader themes. Philippe helped me develop a vision for my research and taught me to care deeply about the meaning and substance of a result. His energy and enthusiasm are contagious and made for a dynamic atmosphere in our meetings and his group, and this helped me grow as a researcher. Thank you, Philippe.

I am grateful to my committee members Michel Goemans and Ankur Moitra for their thoughtful feedback. I also thank Michel for his dedicated efforts and constant friendly spirit as department head. I thank Ankur for his inspiring course on machine learning and for a helpful conversation.

I am thankful to Elchanan Mossel for his guidance and group meetings early in my Ph.D. His interests influenced my own and showed me a world of interesting problems.

I am very grateful to my fantastic collaborators Youn Kim, Jingbo Liu, Raghu Meka, Elchanan, Govind Ramnaryan, and Philippe. I also thank Ben Demeo, Cole Franks, Fred Koehler, Ashwin Narayan, and Tselil Schramm for many interesting research conversations. I thoroughly enjoyed exploring together and learning from you. I am thankful to the Rigollet group for being a constant source of inspiration, especially our *csoport* on discrepancy theory with Kwangjun Ahn, Sinho Chewi, Patrik Gerber, Chen Lu, and Philippe. I'm also grateful to Jan-Christian Hütter, Cheng Mao, and Jonathan Niles–Weed for their examples and guidance.

I thank Karene Chu for being so great and friendly to work with on the MITx Fundamentals of Statistics course over the years and for the many enjoyable meetings during summer 2018 with Karene, Youn, Eren Kizildag, Regina's pizza, and statistics.

I am grateful to all of the friends I met during graduate school, including Nilin Abrahamsen, Kwangjun Ahn, Morris Ang, Pranam Chatterjee, Sinho Chewi, Julien Clancy, Guangyan Du, Robin Elliott, Eren, Christian Gaetz, Ted Grunberg, Yiwen Huang, Pro Jiradilok, Youn, Fred, Brandon Levy, Ashwin, Shourav Pednekar, Govind, Mingxing Liu, Vishal Patil, Xiaoyu Peng, Ahaan Rungta, Andy Senger, Xingyi Shi, Dom Skinner, Felipe Suarez, Peihao Sun, Ethan Sussman, Julie Takagi, Sarah Tammen, Jonathan Tidor, Yoichiro Tsurimaki, Yadu Vijay, and Hong Yue. You made me feel welcome here, and I cherish our memories together.

I am especially grateful for the time I could spend in graduate school with my friend Matt Brennan, who passed away unexpectedly at the beginning of this year. He was an amazingly kind person and inspirational researcher and is sorely missed.

I am thankful for my longtime friends Kate, Chris, Gargi, Xin, Indy, Hanif, Avery, and Howard, and I look forward to more good times to come.

I am grateful to my family, especially my mom, Jackie; my dad, Mark; and my brother, Dex for their constant love and support throughout my life and for always encouraging my interests, as well as my grandparents Deanie and the late Jack, and Beverly and Win. I am grateful to my girlfriend Boya for her love and always being there for me.

Contents

Chapter 1

Introduction

Modern machine learning problems present the dual challenges of achieving statistical accuracy and computational efficiency. There is an inherent tension between these two objectives. From a statistical point of view, oftentimes more data is better: good estimators become more accurate with growing sample size. On the other hand, from a computational perspective, larger datasets are more costly to process and thus estimators with attractive statistical properties may be impractical to evaluate. The purpose of this thesis is to demonstrate how *combinatorial methods* can be used to leverage the structure of data and produce computationally tractable estimators with fast rates of convergence.

Combinatorial principles have a rich history in statistics and continue to be central to the field. The birth of the central limit theorem in the classical work of de Moivre–Laplace relies on understanding the asymptotics of binomial coefficients. The fundamental Vapnik–Chervonenkis combinatorics establishes uniform central limit theorems and hence also rates of estimation for many learning problems. Over the last twenty years, statistical problems involving discrete structures such as inference problems in networks have received a lot of attention, spurred by computational, mathematical, biological, and technological applications, among others. In other situations, combinatorial structure in an estimator itself leads to attractive statistical and computational properties. For example, one may choose estimators based on solutions to combinatorial optimization problems.

This thesis investigates the role of combinatorial structure in statistical problems arising in random vector balancing, nonparametric estimation, and network inference.

In Chapter 2, motivated by a connection with controlled experiments, we explore the statistical and computational aspects of random vector balancing. Vector balancing is a combinatorial optimization problem central to discrepancy theory that consists of dividing a set of vectors into two groups that have approximately the same sum. We provide a sharp analysis of the discrepancy of Gaussian vectors as well as an efficient algorithm achieving the best-known guarantees in an interesting regime. This chapter is based on a joint work [Turner et al., 2020] with Raghu Meka and Philippe Rigollet.

In Chapter 3, we develop a statistical perspective on coresets. Coresets provide a useful framework for summarizing data by extracting a small representative subset.

We focus on the problem of density estimation and study the fundamental limitations of coreset-based estimators, providing a sharp minimax lower bound. We also introduce new coreset kernel density estimators based on Carathéodory's theorem that are near-minimax optimal and improve on prior methods. This chapter is based on a joint work [Turner et al., 2021b] with Jingbo Liu and Philippe Rigollet.

In Chapter 4 we demonstrate how interpolation can be used to improve on the computational aspects of nonparametric estimators. We construct a piecewise multivariate interpolation of a density estimator on a combinatorially structured mesh arising in finite element methods. The resulting interpolant has sublinear space and near-constant query time while preserving statistical accuracy. This chapter is based on a joint work [Turner et al., 2021a] with Jingbo Liu and Philippe Rigollet.

Our final chapter focuses on pedigree reconstruction, a combinatorial inference task of recovering the latent network of familial relationships of a population. We introduce a natural generative model for the pedigree structure and provide an efficient algorithm that approximately recovers the genealogical network from extant genetic data. This chapter is based on a joint work [Kim et al., 2020] with Younhun Kim, Elchanan Mossel, and Govind Ramnarayan.

The remainder of this introduction describes these works in more detail.

1.1 Balancing Gaussian vectors

Motivation

Random vector balancing has a natural connection to experimental design. In controlled experiments, the subjects are assigned to two groups, a treatment and control. Next, an intervention—for example administering a drug—is applied to the treatment group, and inference is conducted to assess efficacy. Pure randomization of the assignment to treatment and controls is often cited as the 'gold standard' because it results in similar covariate structure between the two groups, and indeed this helps to remove confounding variables for the purpose of inference. However, the uniformly random assignment does not result in *optimal* balance between the covariates of each group, and in many cases it is desirable to aim for a higher degree of balance, as we demonstrate in a simple example.

Suppose that the experimenter divides the subjects into a treatment group + and control group − and at the conclusion of the experiment observes an additive linear response

$$Y_i = \alpha^+ \mathbb{1}(\sigma_i = 1) + \alpha^- \mathbb{1}(\sigma_i = -1) + \beta X_i + \varepsilon_i, \quad 1 \le i \le n \tag{1.1}$$

where $\alpha^+, \alpha^-, \beta$ are unknowns; $X_1, \dots, X_n \overset{iid}{\sim} N(0,1)$ are the covariates of the n subjects; $\sigma_i \in \{-1, 1\}$ is the assignment; and $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$ is the noise which we assume is independent of $X_1, \dots, X_n$.

We allow σ to be random and to depend on $X_1, \dots, X_n$, and we enforce that $\mathbb{P}[\sigma_i = 1 | X_1, \dots, X_n] = \frac{1}{2}$. From the perspective of balancing, the natural treatment effect estimator to consider is the Horwitz–Thompson estimator $\hat{\tau}$ [Horvitz and Thompson,

1952, Harshaw et al., 2019], which takes the following form under this assumption:

$$\hat{\tau} = \frac{2}{n} \sum_{i=1}^{n} \sigma_i Y_i.$$

By our setup, $\hat{\tau}$ is unbiased, so the rate of estimation of the treatment effect is controlled by the variance of $\hat{\tau}$. If the treatment and control group are approximately the same size, it holds that

$$\mathbb{E}[(\alpha^+ - \alpha^- - \hat{\tau})^2] = \mathsf{var}\left(\hat{\tau}\right) = \frac{4\beta^2}{n^2} \mathsf{var}\left(\sum_{i=1}^{n} \sigma_i X_i\right) + \frac{4\sigma^2}{n} + o\left(\frac{1}{n}\right). \tag{1.2}$$

The expression above suggests that the treatment and controls should be assigned in order to achieve the best possible aggregate covariate balance between the two groups. What results is a well-studied combinatorial optimization problem known as the *discrepancy* of the one-dimensional covariates $X_1, \ldots, X_n$:

$$\mathcal{D}(X_1, \ldots, X_n) = \min_{\sigma \in \{\pm 1\}^n} \left| \sum_{i=1}^{n} \sigma_i X_i \right| \tag{1.3}$$

This derivation raises two fundamental questions:

1. How large is the discrepancy?

2. Can we find $\sigma \in \{-1, 1\}^n$ with small objective value in polynomial time?

Before revisiting these questions, we first give a very brief overview of discrepancy theory.

Discrepancy theory

Discrepancy theory is a rich area originating in combinatorics that centers on the combinatorial optimization problem

$$\mathcal{D}(X_1, \ldots, X_n) = \min_{\sigma \in \{\pm 1\}^n} \left| \sum_{i=1}^{n} \sigma_i X_i \right|_\infty = \min_{\sigma \in \{\pm 1\}^n} \left| \mathbf{X}\sigma \right|_\infty, \tag{1.4}$$

where $\mathbf{X}$ is the matrix with columns $X_1, \ldots, X_n \in \mathbb{R}^d$. Observe that (1.3) is a special case of this problem when $d = 1$. Discrepancy first arose in the context of hypergraph coloring problems. In this setting, the matrix $\mathbf{X}$ is deterministic, has $\{0, 1\}$-valued entries, and is interpreted as the adjacency matrix of a set system as follows. The columns of $\mathbf{X}$ represent a universe of elements $\{1, \ldots, n\}$, and the rows represent sets $S_1, \ldots, S_d \subset \{1, \ldots, n\}$ where $\mathbf{X}_{ij} = 1$ if and only if $j \in S_i$. Hence, discrepancy is the problem of coloring the elements $\{1, \ldots, n\}$ with two colors, say red and blue, so that every set in our set system has roughly as many red points as blue points.

It is instructive to evaluate the performance of the random coloring $\sigma \sim \mathsf{Unif}\{-1, 1\}^n$ on a set system given by $X \in \{0, 1\}^{n \times n}$. By Hoeffding's inequality and the union

bound,

$$\left|\sum_{i=1}^{n} \sigma_i X_i\right|_\infty \lesssim \sqrt{n \log n}.$$

Is it possible to improve upon this naive assignment? A famous result of Spencer [1985] shows that, remarkably, it is possible to beat the union bound with a clever choice of coloring based on a delicate nonconstructive application of the probabilistic method and the pigeonhole principle. If $d = n$, Spencer's main result 'Six standard deviations suffice' states that

$$\mathcal{D}(X_1, \ldots, X_n) \leq 6\sqrt{n}.$$

This inequality is optimal up to constant factor, and the more general bound

$$\mathcal{D}(X_1, \ldots, X_n) = O\left(\sqrt{n \log \frac{2d}{n}}\right)$$

holds over all $d \times n$ matrices with $d \geq n$ and entries in $[-1, 1]$.

The last decade has witnessed a flurry of algorithmic progress in discrepancy theory, starting with Bansal [2010] who discovered the first algorithm achieving Spencer's bound, and since then several other algorithms [Lovett and Meka, 2012, Harvey et al., 2014, Rothvoss, 2017, Levy et al., 2017, Eldan and Singh, 2018] have been shown to attain $O(\sqrt{n})$ discrepancy. Still many questions remain wide open such as the Beck–Fiala conjecture, which states that

$$\mathcal{D}(X_1, \ldots, X_n) \lesssim \sqrt{t}, \quad \text{for all } t\text{-sparse } X_1, \ldots, X_n \in \{0, 1\}^d,$$

and the tantalizingly stronger Komlós conjecture, which states that

$$\mathcal{D}(X_1, \ldots, X_n) \lesssim 1, \quad \text{if } \max_{1 \leq i \leq n} |X_i|_2 \leq 1.$$

Our problem and contributions

Spencer's result suggests that in the problem of experimental design the optimal allocation improves over pure randomization. While Spencer's focus was on a worst-case setting, from the statistical point of view, it is natural for us to interpret the covariates $X_1, \ldots, X_n$ as i.i.d. copies of a random vector $X \in \mathbb{R}^d$, with $X \sim N(0, I_d)$ as a canonical example. Our results focus on the regime $d \ll n$ and demonstrate a striking gap between the performance of the optimal assignment and pure randomization.

Our first result gives a sharp characterization of the discrepancy of Gaussian random vectors in a regime that was not previously explored.

Theorem 1.1 (Informal)**.** *Let* $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0, I_d)$*. Then as* $n \to \infty$

$$\mathcal{D}(X_1, \ldots, X_n) \sim \sqrt{\frac{\pi n}{2}}\, 2^{-n/d}$$

with high probability, for $\omega(1) = d = o(n)$.

The cases $d = 1$, $d = O(1)$, and $d = \Omega(n)$ were analyzed in Karmarkar et al. [1986], Costello [2009], and Chandrasekaran and Vempala [2014], respectively, and combining these results with ours provides a complete understanding of the discrepancy of Gaussian random vectors. Under the slightly stronger assumption $d = O(n/\log n)$, Theorem 1.1 also applies to more general distributions with i.i.d. coordinates that are absolutely continuous with respect to the Lebesgue measure and satisfy some additional regularity properties (see Chapter 2 for more details).

The proof of Theorem 1.1 is based on the second moment method, a nonconstructive technique from probabilistic combinatorics for deriving existence results [Alon and Spencer, 2008]. The second moment method states that for a nonnegative random variable S, we have

$$\mathbb{P}[S > 0] \geq \frac{\mathbb{E}[S]^2}{\mathbb{E}[S^2]}. \tag{1.5}$$

As described in detail in Chapter 2, our strategy is to let S count the number of signings with discrepancy at most $\gamma 2^{-n/d}\sqrt{\pi n/2}$ and show that the right-hand-side of (1.5) tends to 1 asymptotically for $\gamma > 1$. This requires a careful analysis which we carry out using Laplace's method. The lower bound is simpler and is proved through the first moment method, which is a straightforward application of Markov's inequality (see Proposition 2.2).

The nonconstructive nature of Theorem 1.1 raises the question of the performance of polynomial-time algorithms in our average-case setting. The next result provides a partial answer to this question and establishes the first efficient algorithm that achieves quasi-polynomially small discrepancy (*i.e.*, discrepancy decaying faster than $n^{-\Omega(1)}$) in dimensions two and higher.

Theorem 1.2. *Let* $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(0, I_d)$ *where* $d = O(\sqrt{\frac{\log n}{\log\log n}})$. *Then there exists an absolute constant* $c > 0$ *and a randomized polynomial-time algorithm that with high probability outputs a signing* $\sigma \in \{-1, +1\}^n$ *such that*

$$\left\| \sum_{i=1}^{n} \sigma_i X_i \right\|_\infty \leq \exp\left(-\frac{c \log^2 n}{d}\right).$$

The univariate case was shown by Karmarkar and Karp [1982]. As with our first result, Theorem 1.2 also applies to a more general family of distributions on $\mathbb{R}^d$ with i.i.d. entries that are absolutely continuous with respect to the Lebesgue measure.

We give a very simplified description of our algorithm, which we call generalized Karmarkar–Karp (GKK), applied to i.i.d. uniformly distributed points in $[-1, 1]^d$. As in Karmarkar and Karp [1982], our algorithm is a *differencing method*, which means that throughout the algorithm, we maintain a set of vectors S, and our basic operations consist of removing two vectors, say x and y, from S and then adding the difference to S : $S \leftarrow S \cup \{x - y\} \backslash \{x, y\}$. We perform a sequence of these differencing operations in a judicious way until there is a single vector v remaining in S. Note that at any given time, the elements of S correspond to (disjoint) partial signed sums

of the original vectors $X_1, \dots, X_n$. Hence, the final vector $v \in S$ is indeed a signed sum of the original vectors.

Now we describe our particular differencing method. First we divide the cube into small boxes of side length $n^{-\frac{1}{4d}}$. In each box, we repeatedly sample two points at random and take their difference. The outcomes of these differences now lie in $[-n^{-\frac{1}{4d}}, n^{-\frac{1}{4d}}]$. We can again repeat this procedure of subdividing by $n^{-\frac{1}{4d}}$ and random differencing to obtain outcomes in an even smaller box $[-n^{-\frac{2}{4d}}, n^{-\frac{2}{4d}}]$. Note that after each phase of subdividing and random differencing, the number of remaining points is reduced by 50%. Hence after $O(\log n)$ phases of subdividing and differencing, only one point remains, and it lies in the cube $[-n^{-\frac{\log n}{4d}}, n^{-\frac{\log n}{4d}}]$, which matches the guarantee of Theorem 1.2.

This heuristic analysis gives the correct bound, but it ignores that there may be an odd number of points left over in a box after random differencing. Our algorithm handles these leftover points by using a delicate clean-up step that brings these leftover points into a smaller range using a careful choice of differencing operations. There are also other technical difficulties that arise in controlling the distribution of the outcomes of random differencing in each phase. The full details of our algorithm and its analysis can be found in Chapter 2.

In comparison with Theorem 1.1, which proves that the discrepancy is exponentially small, and Theorem 1.2, which attains quasi-polynomially small discrepancy algorithmically, the uniformly random signing performs drastically worse and has discrepancy $\Theta(\sqrt{n \log d})$. These results along with the discussion at the beginning of this section suggest that in the setting of controlled experiments with a large sample size and a small number of covariates that are strongly predictive, it can be highly advantageous to select a design that optimizes balance. Our findings also highlight a theme in this thesis of using combinatorial methods to beat randomization, a notion we revisit in the next section.

Further directions

There is a stark contrast between the bounds of Theorems 1.1 and 1.2 that suggests the presence of a statistical-to-computational gap as has been observed in sparse PCA and many other planted problems [Berthet and Rigollet, 2013a, Brennan et al., 2018, Kunisky et al., 2019, Brennan and Bresler, 2020]. In the univariate case, it is a longstanding open question as to whether this gap can be closed, and there is evidence from statistical physics [Boettcher and Mertens, 2008], worst-case reductions [Hoberg et al., 2017], and most recently the overlap gap property [Gamarnik and Kızıldağ, 2021] suggesting a negative answer.

Next, the result of Theorem 1.2 only allows for very mildly growing dimension, and our algorithm breaks down when $d \gg \log n$. While the optimization problem has value $\sqrt{n}2^{-n/d}$, for $\log n \ll d \ll n$ it is not even clear how to design algorithms attaining $o(\sqrt{d})$ discrepancy, which appears to be the natural barrier for standard algorithms based on partial coloring. Can $o(1)$ discrepancy be attained efficiently when $n = \mathsf{poly}(d)$? This remains an interesting and challenging open question.

Finally, our algorithm GKK from Theorem 1.2 gives rise to an experimental design

that can be constructed efficiently. By our development earlier in this section, we expect that the GKK design leads to a treatment effect estimator $\hat{\tau}$ with very small variance in the simple model of (1.1). We also conjecture that the GKK design performs well on models with more complicated correlation structures and unobserved covariates as is typical in problems in causal inference. It would be interesting to analyze the performance of GKK in such settings and compare it on real and synthetic data to other methods in the literature such as the matching-based design of Greevy et al. [2004], the rerandomization based-design of Li et al. [2018], as well as the discrepancy-based designs of Bertsimas et al. [2015], Krieger et al. [2019], and Harshaw et al. [2019].

1.2 Coreset density estimation

Background

A coreset is a small weighted subset of the data that captures most of its information. Algorithmic tasks on large datasets raise fundamental computational challenges, so replacing the original observations with a coreset can improve on the time and space complexity while maintaining a high quality of approximation. This is a well-studied paradigm that has been applied to many problems including Bayesian posterior estimation [Campbell and Broderick, 2019], clustering [Feldman et al., 2013], herding [Harvey and Samadi, 2014], logistic regression [Huggins et al., 2016, Munteanu et al., 2018], stochastic gradient descent [Mirzasoleiman et al., 2020], graph sparsification [Spielman and Srivastava, 2011], neural networks [Dubey et al., 2018], matrix-column subset selection [Yang et al., 2017], and kernel density estimation [Phillips, 2013, Phillips and Tai, 2018b, Karnin and Liberty, 2019].[1]

The term 'coreset' was first coined in Bādoiu et al. [2002], though the conception of the idea goes back to Vapnik and Chervonenkis [1971]. In their seminal work, Vapnik and Chervonenkis consider the notion of a range space $(X, \mathcal{R})$ where X is a dataset and $\mathcal{R}$ is a collection of finite subsets of X. They introduce the notion of an ε-sample as a subset $\mathcal{C} \subset X$ that satisfies

$$\sup_{R \in \mathcal{R}} \left| \frac{|R|}{|X|} - \frac{|\mathcal{C} \cap R|}{|\mathcal{C}|} \right| \leq \varepsilon, \tag{1.6}$$

One of their main results shows that a random sample of size $\tilde{\Omega}(\frac{D}{\varepsilon^2})$ from X yields an ε-sample with high probability, where D is the VC-dimension of $(X, \mathcal{R})$. Here we interpret $\mathcal{C}$ as a coreset for the task of counting the relative number of points in a given subset $R \in \mathcal{R}$.

The main application of interest to us is kernel density estimation. The kernel

[1]The definition of coreset is not consistent in the literature, but the one we give here captures many important applications. Also several works use coresets implicitly or refer to them by different names, for example, as sketches, sparsifiers, or ε-samples. Here we distinguish coresets from dimensionality-reduction techniques that embed all of the dataset in a lower dimensional space (which are sometimes also referred to as sketches).

density estimator (KDE) with bandwidth h on the dataset $X_1, \dots, X_n$ is defined to be

$$\hat{f}(y) = \frac{1}{n} \sum_{i=1}^{n} k_h(X_i - y), \tag{1.7}$$

where $k_h(x) = h^{-d} k(\frac{x}{h})$ and $k : \mathbb{R}^d \to \mathbb{R}$. The goal of coreset kernel density estimation is to approximate $\hat{f}$ by the coreset KDE

$$\hat{f}_{\mathcal{C}}(y) = \sum_{X_j \in \mathcal{C}} \lambda_j k_h(X_j - y). \tag{1.8}$$

Motivated by problems in computational geometry [Joshi et al., 2011, Phillips, 2013], clustering [Bachem et al., 2018, Feldman et al., 2013] and outlier detection [Schubert et al., 2014], among others, the existing literature mostly considers the case of deterministic observations and bandwidth $h = \Theta(1)$. In this setting, the primary methods for constructing a coreset are based on random sampling, the Frank–Wolfe algorithm, and discrepancy theory.

Random sampling If the kernel k satisfies certain structural properties, *i.e.*, is Lipschitz, positive definite, or satisfies a certain VC-dimension constraint, then a coreset sampled uniformly at random from the data of size $\tilde{O}(\frac{1}{\varepsilon^2})$ satisfies $\|\hat{f} - \hat{f}_{\mathcal{C}}\|_\infty \leq \varepsilon$. KDE coresets based on importance sampling appear to have not been studied, and as we show, such coresets are provably far from optimal in certain cases. On the other hand, importance sampling based on sensitivity scores and the closely related leverage scores has shown great success in many other problems [Spielman and Srivastava, 2011, Feldman et al., 2013, Cohen et al., 2017, see e.g.].

Frank–Wolfe The Frank–Wolfe algorithm is a greedy iterative method related to gradient descent that is useful for producing sparse approximations to the sample mean in high-dimensional spaces. For kernel density estimation, if we assume that $(x, y) \mapsto k(x - y)$ is positive-definite, there exists a reproducing kernel Hilbert space $\mathcal{H}$, and the KDE $\hat{f}$ is simply the sample average of the feature vectors $\{\phi_i\}_{i=1}^n := \{k_h(X_i - \cdot)\}_{i=1}^n$. The Frank–Wolfe algorithm is defined as follows. Let x_0 denote an arbitrary ϕ_i. Then for $1 \leq t \leq |\mathcal{C}|$, define recursively

$$\begin{aligned} i_t &= \arg\min_{1 \leq i \leq n} \langle x_t - \hat{f}, \phi_i - \hat{f} \rangle_{\mathcal{H}} \\ x_{t+1} &= \frac{1}{t} \phi_{i_t} + \frac{t-1}{t} x_t, \end{aligned}$$

and set $x_{\mathcal{C}} = \hat{f}_{\mathcal{C}}$. It is known that $\|\hat{f} - \hat{f}_{\mathcal{C}}\|_{\mathcal{H}} = O(|\mathcal{C}|^{-1/2})$, and by Cauchy–Schwarz this implies the same bound in $\|\cdot\|_\infty$ if $h = \Omega(1)$.

Discrepancy In a very general sense, low discrepancy implies the existence of small coresets. Define the discrepancy of a class of functions $\mathscr{F}$ to be

$$\mathcal{D}(\mathscr{F}) = \inf_{\sigma \in \{-1,1\}^n} \sup_{f \in \mathscr{F}} \left| \sum_{i=1}^{n} \sigma_i f(X_i) \right|.$$

Note that if the class $\mathscr{F} = \{f_1, \dots, f_d\}$ is finite, then $\mathcal{D}(\mathscr{F})$ is the same as the discrepancy of the matrix $(f_i(X_j))_{1 \le i \le d, 1 \le j \le n}$, as defined in Section 1.1. A classical fact [see Matoušek, 1999, Lemma 1.6] states that if $\mathscr{F}$ contains the function $\mathbb{1}(x) \equiv 1$ for all x and $\mathcal{D}(\mathscr{F}) \le \varepsilon n$, then there exists a coreset of size $\frac{n}{2}$ such that

$$\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{(n/2)} \sum_{X_i \in \mathcal{C}} f(X_i) \right| \le \varepsilon. \tag{1.9}$$

The idea behind (1.9) is roughly that $\{X_i : \sigma_i = 1\}$ and $\{X_i : \sigma_i = -1\}$ serve as good ε-approximations. This procedure can be iterated in a straightforward manner to yield much smaller coresets [for a simple proof see e.g. Karnin and Liberty, 2019, Fact 5]: if for all $S \subset \{X_1, \dots, X_n\}$

$$\inf_{\sigma \in \{-1,1\}^n} \sup_{f \in \mathscr{F}} \left| \sum_{i \in S} \sigma_i f(X_i) \right| \le D,$$

then there exists a coreset $\mathcal{C}$ of size $O(\frac{D}{\varepsilon})$ such that

$$\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{|\mathcal{C}|} \sum_{X_i \in \mathcal{C}} f(X_i) \right| \le \varepsilon. \tag{1.10}$$

The previous display may be regarded as a weighted version of the original notion of ε-approximation considered by Vapnik and Chervonenkis [1971] as in (1.6).

This connection between coresets and discrepancy theory was first applied to the kernel density estimation problem by Phillips [2013]. Here the class is specified by the translates of the kernel: $\mathscr{F} = \{k_h(X_i - \cdot)\}_{i=1}^n$, and hence the resulting error metric is the sup-norm. The state-of-the-art [Phillips and Tai, 2018b] shows that if $h = \Theta(1)$ and k is a PSD kernel that is Lipschitz and decays sufficiently fast, then there exists a coreset $\mathcal{C}$ of size $\tilde{O}(\frac{\sqrt{d}}{\varepsilon})$ such that $\|\hat{f} - \hat{f}_{\mathcal{C}}\|_\infty \le \varepsilon$.

This result is nearly optimal in worst-case over the dataset X, and it improves upon the guarantees of the Frank–Wolfe algorithm and random sampling, which have size $\Omega(\frac{1}{\varepsilon^2})$. However, as we show, by imposing certain smoothness constraints on the dataset, namely that the data is generated from a Hölder smooth probability density function, it is possible to significantly improve on the guarantees of Phillips and Tai [2018b] in a classical statistical setting as we describe next.

Density estimation and our problem

Density estimation is the problem of reconstructing an unknown probability density function from data. Classical nonparametric statistics considers the rates of estimation of densities from certain smoothness classes. Here we consider a Hölder class $\mathcal{P}_{\mathcal{H}}(\beta)$ that contains all probability density functions that are supported on $[0,1]^d$ and have bounded partial derivatives of order β (see Section 3.1.2 for a formal definition). Let $f \in \mathcal{P}_{\mathcal{H}}(\beta)$, and assume that the observations $X_1, \dots, X_n$ are drawn i.i.d. from the probability distribution $\mathbb{P}_f$ associated to f. Let $\mathbb{E}_f$ denote the expectation

over the data with respect to $\mathbb{P}_f$. Then the minimax rate in L_2, which is natural from the statistical point of view, has the form

$$\inf_{\hat{f}} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta)} \mathbb{E}_f \|f - \hat{f}\|_2 = O_{\beta,d}\left(n^{-\frac{\beta}{2\beta+d}}\right). \tag{1.11}$$

For appropriate choice of kernel, the KDE $\hat{f}$ with bandwidth $h \asymp n^{-1/(2\beta+d)}$ attains the minimax rate over the class $\mathcal{P}_{\mathcal{H}}(\beta)$. In particular, there is simple Fourier-theoretic condition characterizing a large class of such kernels. Let $\mathcal{F}[k]$ denote the Fourier transform of k. If for some $\delta > 0$ it holds that

$$\mathcal{F}[k](\omega) \equiv 1, \quad \forall \omega \in [-\delta, \delta]^d, \tag{1.12}$$

then the corresponding KDE with bandwidth $h \asymp n^{-1/(2\beta+d)}$ attains the minimax rate [see Tsybakov, 2009, Chapter 1]. A simple example of such a kernel is constructed as follows. Let $\psi : \mathbb{R} \to [0, 1]$ denote a cutoff function that has the following properties: $\psi \in \mathcal{C}^\infty$, $\psi\big|_{[-1,1]} \equiv 1$, and ψ is supported on $[-2, 2]$. Then the kernel

$$k_s(x) = \prod_{i=1}^{d} \mathcal{F}[\psi](x_i) \tag{1.13}$$

satisfies the condition (1.12), and thus the corresponding KDE attains the minimax rate. Also note that the kernel k is smooth, which will be useful for us later in constructing KDE coresets.

Our main goal is to extend this understanding of minimax rates for density estimation to coreset density estimators of the form (1.8). Specifically, we are interested in finding the smallest possible coresets yielding minimax optimal estimators.

Our methods and contributions

Our estimation strategy is to first compute a minimax optimal standard KDE and then approximate it using a coreset KDE. To motivate our techniques, let us first consider a naive approach based on the Frank–Wolfe algorithm.

The standard KDE (1.7) can be thought of as the average of the infinite-dimensional vectors $k_h(X_1 - \cdot), \ldots, k_h(X_n - \cdot)$. Our goal is to obtain a sparse approximation for the KDE, and one approach we discussed for doing so is to apply Frank–Wolfe in the reproducing kernel Hilbert space corresponding to k. Let $m = |\mathcal{C}|$. Factoring in the bandwidth $h \asymp n^{-1/(2\beta+d)}$ of the standard KDE and the guarantees for Frank–Wolfe previously discussed in Section 1.2, we obtain a coreset $\hat{f}_{\mathcal{C}}^{\mathsf{FW}}$ that satisfies

$$\|\hat{f} - \hat{f}_{\mathcal{C}}^{\mathsf{FW}}\|_\infty \leq h^{-d} m^{-\frac{1}{2}} \lesssim n^{\frac{d}{2\beta+d}} m^{-\frac{1}{2}}.$$

However, setting the right-hand-side to be the minimax rate $n^{-\beta/(2\beta+d)}$ requires $m \gtrsim n$, and there is essentially no compression.

Although Frank–Wolfe fails to yield sublinear coresets for density estimation, it

gives a hint towards the correct approach. While Frank–Wolfe yields sparse approximation of points in convex bodies in infinite-dimensional space, our method is to use Carathéodory's theorem, a result that yields *exact* sparse representation of points in convex bodies in *finite*-dimensional space. Indeed, approximate Carathéodory is one of the main applications of the Frank–Wolfe algorithm, but looking in the opposite direction nicely motivates our approach.

Carathéodory's theorem states that every point in the convex hull of a set $S \subset \mathbb{R}^D$ can be expressed as a convex combination of $D+1$ points of S, and the proof yields an algorithm that runs in $O(nD^3 + n^2)$ arithmetic operations [Carathéodory, 1907, Hiriart-Urruty and Lemaréchal, 2004]. Therefore, if we can embed the functions $k_h(X_1 - \cdot), \ldots, k_h(X_n - \cdot)$ in $\mathbb{R}^D$, then Carathéodory immediately yields a coreset of cardinality $D+1$. For simplicity, let us see how to apply this strategy in the univariate case $X_1, \ldots, X_n \in [0, 2\pi]$ and for a smooth periodic kernel k. Then k has a Fourier expansion

$$k(y) = \sum_{\omega \in \mathbb{Z}} \mathcal{F}[k](\omega) e^{iy\omega},$$

and by the smoothness of k, basic Fourier analysis implies that $\mathcal{F}[k](\omega) = O(|\omega|^{-\gamma})$ for every $\gamma > 0$ [Katznelson, 2004]. Recall that in the univariate case, $k_h(y) = h^{-1}k(\frac{y}{h})$. Therefore,

$$\left| k_h(y) - \sum_{|\omega| \leq T} \mathcal{F}[k_h](\omega) e^{iy\omega} \right| = \left| k_h(y) - \sum_{|\omega| \leq Th} \mathcal{F}[k](\omega) e^{iy\omega} \right| \leq \varepsilon$$

if $T \geq h^{-1}\varepsilon^{-\frac{1}{\gamma}}$. In our setting, we have $h \asymp n^{-1/(2\beta+1)}$ and $\varepsilon = n^{-\beta/(2\beta+1)}$, which implies that for any fixed $\delta > 0$, we can take $\gamma > 0$ sufficiently large so that the Fourier expansion truncated at $T = O(n^{\frac{1}{2\beta+1}+\delta})$ gives a good approximation to $k_h(y)$. Applying this reasoning to the translates $k_h(X_1 - \cdot), \ldots, k_h(X_n - \cdot)$ gives the desired embedding into $\mathbb{R}^D$ with $D = O(n^{\frac{1}{2\beta+1}+\delta})$, where we map $k_h(X_i - \cdot)$ to the D-dimensional vector of its Fourier coefficients up to $T = O(n^{\frac{1}{2\beta+1}+\delta})$. Now applying Carathéodory's theorem in $\mathbb{R}^D$ yields a coreset $\mathcal{C}$ of size $O(n^{\frac{1}{2\beta+1}+\delta})$ and weights $\{\lambda_j\}$ such that

$$\left| \hat{f} - \hat{f}_{\mathcal{C}} \right|_\infty = \left| \frac{1}{n} \sum_{i=1}^n k_h(X_i - \cdot) - \sum_{X_j \in \mathcal{C}} \lambda_j k_h(X_j - \cdot) \right|_\infty \leq n^{-\frac{\beta}{2\beta+1}}.$$

Therefore if $\hat{f}$ is a also good approximation to f, we have derived the rate of the Carathéodory coreset KDE. Conditions for $\hat{f}$ to be a minimax optimal estimator are well-known from classical nonparametric statistics [Tsybakov, 2009], and as discussed earlier, k_s gives an example of a kernel satisfying these conditions that is also smooth enough so that Carathéodory can be applied effectively. It is possible to adapt the above heuristic argument to work in the nonperiodic and high-dimensional case, which yields our first main result.

Theorem 1.3. *Let $\delta > 0$, and let k_s denote the kernel from* (1.13)*. Then there*

exists a coreset $\mathcal{C}$ *of cardinality* $O_{\beta,d,\delta}(n^{\frac{d}{2\beta+d}+\delta})$ *with nonnegative weights* $\{\lambda_i\}$ *where* $\sum \lambda_i = 1$ *and*

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta)} \mathbb{E}_f \|\hat{f}_{\mathcal{C}} - f\|_2 \leq O_{\beta,d}(n^{-\frac{\beta}{2\beta+d}}).$$

Moreover, the coreset $\mathcal{C}$ *can be constructed in time* $O(n^{1+\frac{3d}{2\beta+d}} + n^2)$.

Our techniques imply that similar results hold for any smooth kernel whose corresponding KDE attains the minimax rate. Also for a coreset of size m, we show that our construction attains the rate $m^{-\frac{\beta}{d}+\delta}$ for all fixed $\delta > 0$. See Corollary 3.1 in Chapter 3 for a more general statement.

We also prove a minimax lower bound on a more general class of estimators that demonstrates Theorem 1.3 is nearly optimal with respect to the coreset size. A *decorated coreset* $(\mathcal{C}, \sigma)$ consists of a coreset along with a bit string of length B. A decorated coreset-based estimator consists of all estimators that are measurable with respect to a decorated-coreset; informally, such estimators are only allowed to use information contained in the decorated coreset. We show that coreset KDEs with bandwidth $h = \frac{1}{\mathsf{poly}(n)}$ and weights $|\lambda_j| = \mathsf{poly}(n)$ are decorated coreset-based estimators with $B = \mathsf{poly}(n)$ bits. Our minimax lower bound shows that for $B = \mathsf{poly}(n)$, decorated coreset-based estimators require $|\mathcal{C}| = \Omega(n^{d/(2\beta+d)})$ to attain the rate of estimation $n^{-\beta/(2\beta+d)}$ over the Hölder class $\mathcal{P}_{\mathcal{H}}(\beta)$, and this shows that the Carathéodory coreset is nearly optimal. More generally, we prove a minimax lower bound $\Omega((m \log n)^{-\beta/d})$ on the rate of estimation of decorated coreset-based estimators with coreset cardinality m and bit complexity $B = \mathsf{poly}(n)$, and this is also nearly matched by the Carathéodory construction on m points. Refer to Theorems 3.1 and 3.4 in Chapter 3 for more details.

Another remark is that the Carathéodory coreset uses nonuniform weights, in contrast to many of the previously studied coreset KDE methods such as those based on importance sampling, Frank–Wolfe, and discrepancy theory. For the univariate Gaussian kernel, we prove strong lower bounds demonstrating the power of nonuniform weights. It is known that the Gaussian KDE achieves the minimax rate over the univariate Lipschitz densities $\mathcal{P}_{\mathcal{H}}(1)$. For any choice of bandwidth, we show that the Gaussian coreset KDE with uniform weights requires $\tilde{\Omega}(n^{\frac{2}{3}})$ coreset points to attain the minimax rate $n^{-\frac{1}{3}}$, in contrast to the Gaussian Carathéodory KDE, which requires only $n^{\frac{1}{3}+\delta}$ coreset points. The same lower bounds hold for any smooth nonnegative univariate kernel. More generally, we expect that coreset KDEs with uniform weights require $\Omega(n^{\frac{\beta+d}{2\beta+d}})$ coreset points to attain the minimax rate $n^{-\beta/(2\beta+d)}$ over $\mathcal{P}_{\mathcal{H}}(\beta)$. Additionally, we show that for a large class of kernels, the discrepancy-based approach of Phillips and Tai [2018b] attains the rate $n^{-\beta/(2\beta+d)}$ over $\mathcal{P}_{\mathcal{H}}(\beta)$ using $O(n^{\frac{\beta+d}{2\beta+d}})$ coreset points, matching our conjectured lower bound.

Further directions

Faster Carathéodory As stated in Theorem 1.3, for all parameters our algorithm requires at least n^2 time to construct the coreset, and when the dimension is large

relative to the smoothness, the time-complexity can be as large as n^4. The main computational bottleneck is the standard implementation of Carathéodory's theorem, which for n vectors in dimension D runs in time $O(nD^3 + n^2)$. Recent work of Maalouf et al. [2019] on fast least mean squares solvers gives a new algorithm for Carathéodory's theorem that when applied to n vectors in dimension D runs in time $O(nD+D^4 \log n)$, which reduces to time $O(nD)$ in low dimensions. It is an interesting and practical question as to what is the optimal run-time for Carathéodory's theorem.

Coresets for many tasks In this work, we studied coresets for the density estimation problem, and it is an interesting direction to study coresets that perform well on many different problems. Let us consider a class $\mathscr{F}$ of test functions that we interpret as a collection of tasks to be performed on a dataset. It is natural to regard a coreset as a measure $\mathbb{P}_{\mathcal{C}}$ that is sparsely supported on the dataset and approximates the empirical measure $\mathbb{P}_n$. The integral probability metric

$$d(\mathbb{P}_n, \mathbb{P}_{\mathcal{C}}) = \sup_{f \in \mathscr{F}} \left| \int f \, \mathrm{d}\mathbb{P}_n - \int f \, \mathrm{d}\mathbb{P}_{\mathcal{C}} \right| = \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \sum_{X_j \in \mathcal{C}} \lambda_j f(X_j) \right| \tag{1.14}$$

quantifies the performance of the coreset on the set of tasks $\mathscr{F}$. Can we characterize this performance in terms of the complexity of the class $\mathscr{F}$? Recent work of Karnin and Liberty [2019] studies this question from the point of view of discrepancy theory using the halving approach in Section 1.2. Discrepancy-based techniques produce uniformly weighted coresets, and we would like to find a complexity measure that provides additional power by accounting for the flexibility of the weights in our formulation.

1.3 Interpolation of density estimators

As described in Section 1.2, the kernel density estimator is statistically optimal in many situations, yet from a computational point of view, naive evaluation of the KDE suffers linear evaluation time. This tension inspired numerous approaches for fast evaluation of KDEs including methods based on the fast Gauss transform [Greengard and Rokhlin, 1987, Greengard and Strain, 1991], locality sensitive hashing [Charikar and Siminelakis, 2017, Backurs et al., 2019], binning [Scott and Sheather, 1985], and interpolation [Jones, 1989, Kogure, 1998]. The rich study of the KDE motivates a more general question: if we know that a density estimator is statistically accurate, can we improve on its computational properties?

In Chapter 4, we provide an affirmative answer by using piecewise multivariate interpolation to efficiently convert a minimax estimator $\hat{f}$ of an unknown Hölder density f of smoothness β to a new nearly-minimax estimator $\hat{g}$ that has near-constant query time and storage $\tilde{O}_{\beta,d}(n^{d/(2\beta+d)})$.

Interestingly, the storage derived here roughly matches the number of points in the minimax optimal Carathéodory coreset KDE (see Theorem 1.3). An intuitive explanation for this is that $n^{d/(2\beta+d)}$ represents the number of degrees of freedom in the Hölder class; concretely, any $f \in \mathcal{P}_{\mathcal{H}}(\beta)$ can be approximated to minimax-

accuracy via a Fourier expansion with $O(n^{d/(2\beta+d)})$ terms. Moreover, our minimax lower bounds for decorated coreset-based estimators described in the last section imply that $\tilde{O}(n^{d/(2\beta+d)})$ is the optimal storage that can be achieved if $\hat{g}$ is required to be near-minimax optimal (see Theorem 3.4).

The near-constant query time is a consequence of our interpolation scheme, which uses piecewise multivariate polynomials. We divide the domain, which is assumed to be $[0,1]^d$, into $n^{d/(2\beta+d)}$ boxes of side length $n^{-1/(2\beta+d)}$. In each box, $\hat{g}$ is specified by a degree β polynomial interpolant of $\hat{f}$ over a carefully chosen set of points known as the *principal lattice*. Thus to evaluate $\hat{g}$ at a query y, it suffices to identify the box containing y—this is a simple search requiring $O(\log n)$ time–and to evaluate the degree β polynomial corresponding to that box.

The minimax accuracy is a consequence of two important facts: (i) Hölder smooth densities of order β are well-approximated by piecewise polynomials of degree β with pieces given by the $n^{d/(2\beta+d)}$ boxes described above, and (ii) our interpolation scheme is stable enough to recover the 'true' polynomial in each box from the noisy queries $\hat{f}$ of f on the principal lattice.

Interpolation on the principal lattice

This section provides an exposition of some of the main ideas in Chung and Yao [1977] [Nicolaides, 1972, see also]. The principal lattice is defined as follows. Let $\Delta_d \subset [0,1]^d$ denote the convex hull of $v_0 = 0$ and $v_i = e_i$ for $1 \le i \le d$, where e_i denotes the i-th standard basis vector in $\mathbb{R}^d$. Let $\ell \ge 1$ denote an integer. The *principal lattice* $\mathcal{P}(\ell) = \{U_1, \ldots, U_M\}$ of order ℓ consists of all convex combinations $\sum_{j=0}^{d} \lambda_j v_j$ where $\lambda_j \in \frac{1}{\ell}\mathbb{Z}$ for all $0 \le j \le d$. We also define $\mathcal{P}(0) = 0 \in \mathbb{R}^d$. By a simple balls-and-bins counting argument, it holds that the cardinality of $\mathcal{P}(\ell)$ is $M = \binom{\ell+d}{\ell}$.

The principal lattice $\mathcal{P}(\ell)$ has the remarkable property that for any values specified on its elements $U_1, \ldots, U_M$, there exists a unique polynomial of degree ℓ interpolating those values. Hence we say that $\mathcal{P}(\ell)$ admits unique interpolants. This property is crucial to our approach described above because it allows us to uniquely and approximately recover the degree ℓ Taylor expansion of the unknown density f in each box.

The aforementioned unique interpolation property of the principal lattice is quite special and does not hold for all point configurations. By simple linear-algebraic considerations, it is not hard to see that the cardinality of such a set must be $M = \binom{\ell+d}{d}$. A theorem of Chung and Yao [1977] gives an elegant combinatorial-geometric condition known as GC characterizing point configurations that admit unique interpolants: a set of points $\mathcal{P}$ satisfies the GC condition if every $u \in \mathcal{P}$ has an associated set of ℓ affine $(d-1)$-dimensional hyperplanes whose union contains every point in $\mathcal{P}$ except for u. Under the condition GC, it is straightforward to build a Lagrangian interpolant. To see this, an equivalent interpretation of GC is that for every $u \in \mathcal{P}$ there exist ℓ linear functions $h_1^u, \ldots, h_\ell^u$ such that

$$\prod_{i=1}^{\ell} h_i^u(u') = \mathbb{1}(u = u') \quad \forall u' \in \mathcal{P}.$$

Therefore,

$$p_g(y) = \sum_{u \in \mathcal{P}} g(u) \prod_{i=1}^{\ell} h_i^u(y)$$

is the unique polynomial of degree ℓ such that $p_g(u) = g(u)$ for all $u \in \mathcal{P}$.

Let us apply this framework to the principal lattice. Let $\lambda_j(x)$ denote the j-th barycentric coordinate of $x \in \mathbb{R}^d$ with respect to $v_0, \ldots, v_d$. Concretely,

$$\begin{aligned} \lambda_j(x) &= x_j \quad \forall 1 \le j \le d \\ \lambda_0(x) &= 1 - x \cdot \mathbb{1} \end{aligned}$$

We see that $\mathcal{P}(\ell)$ satifies GC: the ℓ associated hyperplanes to $U \in \mathcal{P}(\ell)$ are given by

$$\left\{ x : \lambda_j(x) = \frac{s}{\ell} \right\}$$

where j satisfies $\lambda_j(U) > 0$ and $s = 0, 1, \ldots, \ell\lambda_j(U) - 1$. To see this, observe that if $U' \in \mathcal{P}(\ell)$ has barycentric coordinates majorizing those of U (*i.e.*, $\lambda_j(U') \ge \lambda_j(U)$ for all $0 \le j \le d$), then in fact $U' = U$ — otherwise $\sum_j \lambda_j > 1$, which is a contradiction. Therefore,

$$p_g(y) = \sum_{i=1}^{M} g(U_i) \cdot \prod_{\substack{j=0 \\ \lambda_j(U_i) > 0}}^{d} \prod_{s=0}^{\ell\lambda_j(U_i)-1} \frac{\lambda_j(x) - \frac{s}{\ell}}{\lambda_j(U_i) - \frac{s}{\ell}}$$

is the unique polynomial interpolant of degree ℓ such that $p_g(U_i) = g(U_i)$ for $U_i \in \mathcal{P}(\ell)$.

Our method and contributions

We now formally describe our method for converting the given estimator $\hat{f}$ of the unknown density $f \in \mathcal{P}_{\mathcal{H}}(\beta)$ to a more computationally tractable form. Let $\ell = \lfloor \beta \rfloor$ denote the greatest integer *strictly* less than β.

Construction of $\hat{g}$

1. Partition: Divide $[0,1]^d$ into h^{-d} boxes $\{B_k\}$ of side length $h = n^{-1/(2\beta+d)}$
2. Mesh: Let $\mathcal{P}^k(\ell) \subset B_k$ denote a shifted copy of the rescaled principal lattice $h \cdot \mathcal{P}(\ell)$
3. Interpolate: For each box, construct the unique degree ℓ polynomial interpolant p_k such that $p_k(U) = \hat{f}(u)$ for all $u \in \mathcal{P}^k(\ell)$

Return: $\hat{g} : [0,1]^d \to \mathbb{R}$ defined by

$$\hat{g}(y) = \sum_k p_k(y) \mathbb{1}(y \in B_k).$$

Our main result is the following theorem.

Theorem 1.4. *Let $f \in \mathcal{P}_{\mathcal{H}}(\beta)$, and let $\hat{f}$ denote an estimator that is pointwise minimax optimal:*

$$\sup_{y \in [0,1]^d} \mathbb{P}_f \left[\left| \hat{f}(y) - f(y) \right| > t \right] \leq 2 \exp\left(-\frac{t^2}{\varepsilon^2} \right),$$

where $\varepsilon = O_{\beta,d}(n^{-\beta/(2\beta+d)})$. Let Q denote the amount of time it takes to query $\hat{f}$. Then $\hat{g}$ has the following properties:

- *Sublinear space: $\tilde{O}_{\beta,d}(n^{\frac{d}{2\beta+d}})$*
- *Near-constant query time: $\tilde{O}_{\beta,d}(1)$*
- *Near-minimax error: $\|f - \hat{g}\|_\infty = \tilde{O}_{\beta,d}(n^{-\frac{\beta}{2\beta+d}})$*

The assumption on $\hat{f}$ is satisfied for many density estimators, including large families of kernel density estimators, local polynomial estimators, and projection estimators [Tsybakov, 2009]. Though our work here targets the minimax rate while using minimal space, more generally, by tuning the parameter h in our construction above we can guarantee $\|f - \hat{g}\|_\infty \leq \delta$ with $\tilde{O}_{\beta,d}(\delta^{-d/\beta})$ space and near-constant query time. Also, in this work we focus on density estimation, but our methods are readily applicable to other nonparametric settings such as regression.

Further directions

The implicit constants in Theorem 1.4 scale roughly as $\binom{\beta+d}{\beta}$ where β is the smoothness of the unknown density and d is the dimension. When $\beta \asymp d$, this binomial coefficient scales exponentially, so it would be useful to improve these constants. However, it is important to keep in mind that typically in nonparametric statistics $\beta = O(1)$, and hence we must have $d = O(\log n)$ for even consistent estimation to be possible.

Another interesting problem is to find an adaptive method achieving similar guarantees to Theorem 1.4 when the smoothness parameter β is not known in advance. It is possible to use existing methods to estimate the smoothness from the data, but it is unclear if this information can be extracted from a black-box estimator satisfying an accuracy assumption such as in Theorem 1.4.

Finally, the estimator from Theorem 1.4 is a multivariate piecewise polynomial and is thus discontinuous. It would be interesting to see if a smoothed version of our estimator can be constructed using splines.

1.4 Pedigree reconstruction

In our final chapter, we investigate a combinatorial inference problem arising from genomics. A *pedigree* is a graph that describes the genealogy of a population. The nodes of this graph represent individuals, and the edges indicate parent-child relationships. A fundamental problem in bioinformatics is to reconstruct the pedigree on

prior generations given the genetic information of previously sequenced individuals. In real-world applications, companies such as MyHeritage, 23andMe, and Ancestry.com provide similar services using large databases. There is also a host of algorithms from the computational biology literature demonstrating promising empirical and theoretical performance [Thompson, 2000, Steel and Hein, 2006, Thatte and Steel, 2008, He et al., 2014, Huisman, 2017, Wang, 2019]. In this work, we study an idealized model that generates random pedigrees with a large number of generations and develop an efficient recursive algorithm for reconstructing the unknown pedigree from the observed gene sequences of extant individuals.

Our model

We consider a simple model of how a child inherits genetic information from its parents. Each parent possesses a gene sequence, which we assume to be a string of characters of length B from an alphabet. For each entry or *block* of the child's gene sequence, we flip a fair coin. If heads, the block is filled with the mother's corresponding symbol, and otherwise the child inherits its father's symbol in that block. In our model, we assume that the pedigree is graded by generations so that only individuals from the same generation form couples. Thus, once the symbols of the *founders*, the highest level nodes in the pedigree, are specified, we can continue repeating the above procedure in a Markovian way to generate an inheritance process on the entire pedigree. We observe the gene sequences of the *extant*, the lowest nodes in the pedigree, and our problem is to reconstruct the latent pedigree from these observations.

Unfortunately, pedigree reconstruction in this model is ill-posed due to unidentifiability (see Section 5.2 for examples). However, examples of unidentifiable pedigrees shed light on the phenomena that make inference challenging. The difficulty in pedigree reconstruction stems from *inbreeding* which essentially amounts to cycles in the pedigree. For example, there may be pairs of extant nodes with multiple lowest common ancestors or couples that have a common ancestor. On the other end of the spectrum of difficulty, one can consider pedigrees with no cycles. In this situation, a simple approach is to estimate the pairwise distances among all of the extant nodes by counting the number of common symbols shared by two individuals. If the gene sequences are long enough, then by Hoeffding–Chernoff bounds this procedure reconstructs all of the pairwise distance correctly, which in turn suffices to correctly reconstruct the pedigree [Steel and Hein, 2006].

In this work, we formulate a natural generative model for the pedigree structure with a mild degree of inbreeding that makes the inference task challenging yet tractable for approximate recovery. The generative model is specified by a branching factor α that represents the average number of children per couple, an integer T that denotes the number of generations, and an integer N_T that denotes the size of the founding population. First, the individuals in the founding population randomly pair up into couples, and each couple generates $\mathsf{Poisson}(\alpha)$ children independently that compose generation $T-1$. Iterating this procedure of random pairing and generating $\mathsf{Poisson}(\alpha)$ children per couple for T generations yields the pedigree structure. The

inheritance procedure operates as described above, where we initialize the founders to have no common symbols. Equivalently, this can be viewed as sampling the gene sequences of the founders i.i.d. and uniformly from a very large or infinite alphabet.

Next we give some intuition about the effect of the parameters in our model on the degree of inbreeding and the difficulty of the reconstruction problem. The block length B can be viewed as the sample complexity of our model by considering the i-th entry of each extant gene sequence of the pedigree as a single draw from the inheritance distribution resulting from the network structure. Thus the larger that B is, the more information we have in the sample regarding the pedigree. As the branching factor α increases, the inference task also intuitively becomes easier because more information is transmitted to the extant. As the size N_T of the founding population grows large relative to α, the amount of inbreeding reduces—for example, siblings are less likely to form a couple because the average number of children is small relative to the size of the generation. On the other hand, as the number of generations T grows, cycles become more likely and this increases the degree of inbreeding. These phenomena are further illustrated by our main result.

Our methods and contributions

In this work, we develop an efficient algorithm for recovering the pedigree that recursively reconstructs it generation by generation, starting with the parents of the extant nodes. At a high level, the algorithm operates in the following manner. Once generation t is reconstructed, we reconstruct the next level $t+1$ by first recovering the gene sequences at level t. Next, we determine which nodes in level t are siblings with one another by comparing common blocks among their recovered strings. Finally, collections of nodes determined to be mutual siblings in generation t are then iteratively assigned parents from generation $t+1$. The idea for this recursive scheme is simple, but several complications arise in its implementation and analysis due to the presence of inbreeding (see Chapter 5). Our main result is the following.

Theorem 1.5 (Informal)**.** *Assume that the branching factor α is a sufficiently large absolute constant, that the number of generations is $T = O(\frac{\log N_T}{\log \alpha})$, and that the block length is $B = \Omega(\log N_T)$. There exists an algorithm that runs in time $\mathsf{poly}(N_T, B)$ and recovers 90% of the true pedigree in every generation, with high probability as $N_T \to \infty$.*

By the bound imposed on T, with high probability, the pedigree can be shown to have size on the order $N_T^{1+\delta}$, for a small constant δ that is independent of the branching factor α. In other words, regardless of α, the deepest pedigrees that can be handled by Theorem 1.5 all have the same number of vertices. On the other hand, as α grows, we show that our algorithm recovers a growing fraction of the pedigree that can be made arbitrarily close to 1 for α large enough. Finally, our algorithm has sample complexity $B = \mathsf{poly}(T)$, which is one of its main advantages. In comparison, the naive approach discussed for reconstruction of tree pedigrees based on estimating pairwise differences requires $B = \exp(\Omega(T))$ to recover a large fraction of the truth.

Further directions

Relaxing the assumptions in our model to reflect the properties of more realistic pedigrees arising in real-world applications is an important direction for future work. For example, our approach here requires the branching factor α to be a very large constant, but it may be possible to achieve approximate recovery for all $\alpha > 2$. Notably, our model here does not allow for mutations, a basic phenomenon in biological inheritance, and it would be interesting to investigate the impact of mutations on the inference problem. As mentioned above, our prior on the gene sequences of the founders is essentially i.i.d. over a large enough alphabet so that *all* of the generated symbols are distinct. It is an intriguing question as to whether similar guarantees can be attained with a binary or ternary alphabet. Another compelling problem is to achieve strong recovery guarantees for pedigrees with a higher degree of inbreeding than the ones we encounter here.

Chapter 2

Balancing Gaussian vectors

2.1 Introduction

Randomized controlled experiments are often dubbed the “gold standard" for estimating treatment effects because of their ability to create a treatment and a control group that have the same features on average. Indeed, pure randomization, i.e., assigning each observation uniformly at random between the treatment and control group, leads to two groups with approximately the same size, the same average age, the same average height, etc. Unfortunately, because of random fluctuations, this approach may not lead to the best balance between the attributes of the control group and those of the treatment group. Yet, near perfect balance is highly desirable since it often leads to a more accurate estimator of the treatment effect. This quest for balance was initiated at the dawn of controlled experiments. Indeed, W.S. Gosset, a.k.a Student (of t-test fame) already questioned the use of pure randomization when it leads to unbalanced covariates [Student, 1938], and R.A. Fisher proposed randomized block designs as a better solution in certain cases [Fisher, 1935]. One traditional approach to overcome this limitation is to simply *rerandomize* the allocation until the generated assignment is deemed balanced enough [Morgan and Rubin, 2012, Li et al., 2018]. Rerandomization is effectively a primitive form of optimization that consists in keeping the best of several random solutions. However, it was not until recently that covariate balancing was recognized for the combinatorial optimization problem that it really is. With this motivation, Bertsimas et al. [2015], Kallus [2018] proposed algorithms based on mixed integer programming that, while flexible, did not come with theoretical guarantees. More recently, Harshaw et al. [2019] used new algorithms from Bansal et al. [2018] with theoretical guarantees to generate experimental designs with a tunable degree of randomization versus covariate balance and characterized the resulting trade-off between model robustness and efficiency for a specific treatment effect estimator computed on data collected in such experiments.

In this work, we investigate both the theoretical and algorithmic aspects associated to this question by framing it in the broader scope of *vector balancing*. In particular, this question bears strong theoretical footing in discrepancy theory.[1]

[1]The recent work Harshaw et al. [2019] takes a similar point of view, though here our purpose is

Let $X_1, \ldots, X_n \in \mathbb{R}^m$ denote a collection of vectors and let $\mathbf{X}$ denote the $m \times n$ matrix whose column i is X_i. The *discrepancy* $\mathcal{D}(X_1, \ldots, X_n)$ of this collection is defined as follows.[2]

$$\mathcal{D}_n := \mathcal{D}(X_1, \ldots, X_n) = \min_{\sigma \in \{\pm 1\}^n} \left| \sum_{i=1}^n \sigma_i X_i \right|_\infty = \min_{\sigma \in \{\pm 1\}^n} |\mathbf{X}\sigma|_\infty \tag{2.1}$$

Discrepancy theory is a rich and well-studied area with applications to combinatorics, optimization, geometry, and statistics, among many others [see the comprehensive texts Matoušek, 1999, Chazelle, 2000]. A fundamental result in the area due to Spencer [1985] states that if $\max_i |X_i|_\infty \leq 1$ and $m = n$, then $\mathcal{D}_n \leq 6\sqrt{n}$. Spencer's proof is nonconstructive and relies on a technique known as *partial coloring*. In the last decade, starting with the breakthrough work of Bansal [2010], several algorithmic versions of the partial coloring method have been introduced to efficiently find a signing σ that approximately attains the minimum in (2.1). These include approaches based on random walks [Bansal, 2010, Lovett and Meka, 2012], random projections [Rothvoss, 2017], and multiplicative weights [Levy et al., 2017]. In the regime where $m \geq n$, these algorithms can be used to compute a signing (or allocation) $\sigma \in \{-1, 1\}^n$ with objective value $O(\sqrt{n \log(2m/n)})$. Moreover, this guarantee is tight in the sense that examples are known with discrepancy matching this bound.

The aforementioned results make minimal structural assumptions on the vectors $X_1, \ldots, X_n$ and treat the input as worst-case. However, in the context of controlled experiments, it is natural to assume that $X_1, \ldots, X_n$ are, in fact, independent copies of a random vector $X \in \mathbb{R}^m$. While more general results are possible, the reader should keep in mind the canonical example where $X \sim \mathcal{N}_m(0, I_m)$ is a standard Gaussian vector, and in particular where the entries of X are of order 1. We dub the study of $\mathcal{D}_n$ in this context *average-case discrepancy*.

It was first shown in Karmarkar et al. [1986] via a nonconstructive application of the second moment method that when $m = 1$, the average-case discrepancy is $\mathcal{D}_n = \Theta(\sqrt{n}\, 2^{-n})$ with high probability, assuming that the underlying distribution has a sufficiently regular density. This result was extended to specific multidimensional regimes. First, Costello [2009] showed that $\mathcal{D}_n = \Theta(\sqrt{n}\, 2^{-n/m})$ in the constant dimension regime $m = O(1)$. The optimal discrepancy is also known in the super-linear regime $m \geq 2n$ where it was shown that $\mathcal{D}_n = O(\sqrt{n \log(2m/n)})$.[3] In particular, there is a striking gap between this benchmark and the discrepancy $|\mathbf{X}\sigma^{\mathsf{rdm}}|_\infty = \Theta(\sqrt{n \log m})$ achieved by a random signing σ^{rdm}, especially in the sublinear regime. Motivated by applications to controlled experiments, Krieger et al. [2019] studied the average-case discrepancy problem with the aim to improve on this gap. The authors devised a simple and efficient greedy scheme that, in the univariate

to focus purely on optimal covariate balance.

[2]In the interest of clarity, we free ourselves from important considerations in the practical design of controlled experiments such as having two groups of exactly the same size.

[3]The upper bound established in Chandrasekaran and Vempala [2014] presents additional polylogarithmic terms that are negligible for most of the range $m \geq 2n$. This is also the regime considered by Harshaw et al. [2019].

case, outputs an allocation σ^{gree} satisfying $|\mathbf{X}\sigma^{\mathsf{gree}}| = O(n^{-2})$. In addition, Krieger et al. [2019] argue that $|\mathbf{X}\sigma^{\mathsf{gree}}| = O(n^{-2/m})$ for any *constant* dimension m.

This state of the art leaves three important questions open:

1. Can a sub-polynomial discrepancy be achieved in polynomial time even in dimension 1?

2. What is the optimal discrepancy in the intermediate regime where $\omega(1) = m = o(n)$?

3. Do there exist efficient allocations that perform better than the random allocation in super-constant dimension?

The answer to the first question is well known. Indeed, the best known algorithm for number partitioning is due to Karmarkar and Karp [1982] and yields $\sigma \in \{-1, 1\}^n$ such that $|\mathbf{X}\sigma|_\infty = e^{-\Omega(\log^2 n)}$ with high probability [see also Boettcher and Mertens, 2008]. While this result provides a super-polynomial improvement over algorithms built for the worst case, a significant gap remains between the information-theoretic bounds and the algorithmic ones despite extensive work on the subject [Boettcher and Mertens, 2008, Borgs et al., 2001, Hoberg et al., 2017]. This suggests the possibility of a statistical-to-computational gap similar to those that have been observed starting with sparse PCA [Berthet and Rigollet, 2013a,b] and more recently in other planted problems [Brennan et al., 2018, Bandeira et al., 2018]. Moreover, while the greedy algorithm of Krieger et al. [2019] is loosely based on ideas from this algorithm, no multivariate extension of this algorithm is known even for the case $m = 2$. Note that in the super-linear regime $m \geq 2n$, the work of Chandrasekaran and Vempala [2014] also proposes a polynomial-time algorithm based on Lovett and Meka [2012] showing an absence of substantial statistical-to-computational gaps.

In this paper, we provide answers to the remaining two questions raised above. First, we show that the discrepancy of standard Gaussian vectors is $\Theta(\sqrt{n}\, 2^{-n/m})$ with high probability for the remaining regime $\omega(1) = m = o(n)$. Moreover, we complement this existential result by giving the first randomized polynomial-time algorithm that achieves discrepancy $e^{-\Omega(\log^2(n)/m)}$ when $2 \leq m = O(\sqrt{\log n})$. Note that while this remains an intrinsically low-dimensional result, it covers already super-constant dimension. This first algorithmic result paves the way for potential algorithmic advances in a wider range of high-dimensional problems. In particular, our existential result sets an information-theoretic benchmark against which future algorithmic results can be compared as well as a baseline to establish potential statistical-to-computational gaps in high dimensions. These improved discrepancy bounds also have direct applications to randomized control trials. For example, in the case of an additive linear response with all covariates observed, the discrepancy attained by the allocation controls the fluctuations of the difference-in-means treatment effect estimator [Krieger et al., 2019].

Another point of view on balancing covariates in randomized trails is that of pairwise matching. In this setup, the experimenter first divides the sample into two equal-sized groups and then pairs up individuals who have similar covariates.

For the unidimensional case, Greevy et al. [2004] proposed a scheme that consists of performing a minimum cost matching that leads to a bounded discrepancy. This result may be extended to yield a discrepancy of order $n^{1-1/m}$ in dimension m using results on random combinatorial optimization Steele [1992]. Unlike matching algorithms, bipartite matching algorithms can be implemented in near-linear time using modern tools from computational optimal transport [Cuturi and Peyré, 2018, Altschuler et al., 2017, 2019]. We leave it as an interesting open question to study allocation schemes based on random bipartite matching problems for which sharp results have recently been discovered [Ledoux and Zhu, 2019].

2.2 Main results

In this section, we give an overview of our main results. Detailed computations and proofs are postponed to subsequent sections.

2.2.1 Existential result

Our first main result shows that when $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(0, I_m)$ and $\omega(1) = m = o(n)$, then the discrepancy is asymptotically $\sqrt{\frac{\pi n}{2}}\, 2^{-n/m}$ with high probability. As in the one-dimensional case [Karmarkar et al., 1986], this result highlights that drastic cancellations are possible, with high probability, when the number of vectors grows asymptotically faster than the dimension.

Theorem 2.1. *Fix an absolute constant $\gamma > 1$ and suppose that $\omega(1) = m = o(n)$. Let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(0, I_m)$ be independent standard Gaussian random vectors. Then*

$$\lim_{n\to\infty} \mathbb{P}\Big[\mathcal{D}(X_1, \dots, X_n) \leq \gamma \sqrt{\frac{\pi n}{2}} 2^{-n/m}\Big] = 1\,. \tag{2.2}$$

If $\gamma' < 1$, then

$$\lim_{n\to\infty} \mathbb{P}\left[\mathcal{D}(X_1, \dots, X_n) \geq \gamma' \sqrt{\frac{\pi n}{2}} 2^{-n/m}\right] = 1. \tag{2.3}$$

The work of Costello [2009] handles the case $m = O(1)$, and shows that the limiting probability in (2.2) is exactly $1 - \exp(-2\gamma^m)$. We also note that the series of papers by Borgs et al. [2001, 2008a,b] provides an even more complete description of the unidimensional case.

Our results are not limited specifically to Gaussian distributions. A mild extension of our techniques allows us to derive a similar result for a more general family of distributions, assuming that $m = O(n/\log n)$.

Remark 2.1. *Let $C > 0$ denote a sufficiently small absolute constant, and suppose that $m \leq Cn/\log n$. Let $\mathbf{X}$ denote an $m \times n$ random matrix whose entries are i.i.d random variables having a common density $f : \mathbb{R} \to \mathbb{R}$ such that*

$$\int f(x)^2 \mathrm{d}x < \infty, \qquad \int x^4 f(x) \mathrm{d}x < \infty, \quad \textit{and} \quad f(x) = f(-x), \forall\, x \in \mathbb{R}.$$

Then there exist absolute positive constants $c \le c'$ such that

$$\lim_{n\to\infty} \mathbb{P}\left[c\sqrt{n}2^{-n/m} \le \mathcal{D}(X_1,\dots,X_n) \le c'\sqrt{n}2^{-n/m}\right] = 1.$$

We omit the proof of the above remark and focus on the Gaussian case for simplicity and because for Gaussian vectors, our analysis covers the whole range $m = o(n)$.

The proof of the upper bound in Theorem 2.1 is a nonconstructive application of the second moment method, in a similar spirit to the analysis of Karmarkar et al. [1986] on the one-dimensional case as well as Achlioptas–Moore's analysis of the threshold for random k-SAT [Achioptas and Moore, 2002]. Recall that the second moment method states that for a nonnegative random variable S, we have

$$\mathbb{P}[S > 0] \ge \frac{\mathbb{E}[S]^2}{\mathbb{E}[S^2]}. \tag{2.4}$$

As described in more detail in Section 2.3, our strategy is to let S count the number of signings with discrepancy at most $\gamma 2^{-n/m}\sqrt{\pi n/2}$ and show that the right-hand-side of (2.4) tends to 1 asymptotically. We also note that the lower bound in Theorem 2.1 is a straightforward consequence of the Markov inequality (first moment method) applied to S (see Proposition 2.2).

In addition to our result for $m = o(n)$, using similar techniques we also provide a precise characterization of Gaussian discrepancy in the linear regime $m \le \delta n$, where δ is a sufficiently small absolute constant. In Appendix 2.5.2, we show that the discrepancy is $\Theta(\sqrt{n}2^{-1/\delta})$ with probability at least 99%, asymptotically as $n \to \infty$. This provides further evidence of a conjecture of Aubin et al. [2019] that the discrepancy when $m = \delta n$ is asymptotically $c(\delta)\sqrt{n}$ with high probability for an explicit function $c(\delta)$.[4] In particular, our result combined with those of Chandrasekaran and Vempala [2014] confirms that the discrepancy is $\Theta(c(\delta)\sqrt{n})$ with asymptotic probability at least 99% when $m = \delta n$ for all $\delta > 0$.

Complementary to our work, we discuss recent existential results on average-case discrepancy in the discrete case when $X_1,\dots,X_n$ are i.i.d vectors in $\{0,1\}^m$. Extending prior work of Ezra and Lovett [2016], Franks and Saks [2018] and Hoberg and Rothvoss [2018] use a nonconstructive Fourier-analytic argument to show, for two different models of random sparse binary vectors, that the discrepancy is $O(1)$ if $n = \tilde{\Omega}(m^3)$ [Franks and Saks, 2018] and $n = \tilde{\Omega}(m^2)$ [Hoberg and Rothvoss, 2018]. In addition, for the continuous case, Franks and Saks [2018] show that the discrepancy of random unit vectors is $O(\exp(-\sqrt{n/m^3}))$. Potukuchi [2018] uses the second moment method to show the discrepancy is $O(1)$ if $n = \Omega(m \log m)$ in the specific case where the entries of X_1 are uniform on $\{0,1\}$. In other recent work, Bansal and Meka [2019] establish an average-case version of the Beck–Fiala conjecture, giving an algorithmic proof that the discrepancy of uniformly random t-sparse binary vectors is at most $O(\sqrt{t})$ for the entire range of parameters m, n if $t = \Omega(\log\log m)$. It is an open question as to whether there exists a polynomial-time algorithm achieving $O(1)$

[4] See Appendix 2.5.2 for a more precise description of their results.

discrepancy for random $\{-1,+1\}$ vectors or sparse $\{0,1\}$ vectors with $n = \mathsf{poly}(m)$ [Hoberg and Rothvoss, 2018, Franks and Saks, 2018].

2.2.2 Algorithmic result

Our second main result is algorithmic and applies to a large family of continuous distributions. We construct a randomized polynomial-time algorithm called Generalized Karmarkar–Karp (**GKK**) that achieves discrepancy $\exp(-\Omega(\log^2(n)/m))$ with high probability, assuming $m = O(\sqrt{\log n})$. This establishes the first such efficient algorithm achieving quasi-polynomially-small discrepancy for this regime. Our algorithm and analysis extend those of Karmarkar and Karp [1982] in the one-dimensional case to higher dimensions.[5]

Theorem 2.2. *Let $\mathbf{X}$ denote a random $m \times n$ matrix with iid entries having a common density $\rho : [-\Delta, \Delta] \to \mathbb{R}$ which is L-Lipschitz and bounded above by some constant $D > 0$. Suppose that*

$$m \leq C\sqrt{\frac{\log n}{\max(1, \log \Delta)}},$$

for some sufficiently small absolute constant $C = C(D, L) > 0$. Then the algorithm **GKK** *outputs, in polynomial time, a signing $\sigma \in \{-1,+1\}^n$ such that*

$$|\mathbf{X}\sigma|_\infty \leq \exp\left(-\frac{c\log^2 n}{m}\right),$$

with probability at least $1 - \exp(-cn^{1/4})$ for some absolute constant $c > 0$.

This result easily extends to distributions with unbounded support. For example, if $\mathbf{X}$ has i.i.d standard Gaussian entries, then setting $\Delta = O(\sqrt{\log n})$ and conditioning on the (high probability) event $\{|\mathbf{X}_{ij}| \leq \Delta \ \forall\, i, j\}$, we can apply Theorem 2.2 to show that **GKK** yields discrepancy $\exp(-c\log^2(n)/m)$ for the Gaussian matrix $\mathbf{X}$.

It is an open question as to whether or not the guarantee of Theorem 2.2 can be improved to achieve sub-quasi-polynomial discrepancy efficiently, even in dimension one. Note that for $m = 1$, Hoberg et al. [2017] provide evidence of hardness of a $O(2^{\sqrt{n}})$-approximation to the optimal discrepancy in worst case via a reduction from the Minkowski problem and the shortest vector problem. We leave the following question.

Question 2.1. *Suppose that $m = n^\gamma$ for some $\gamma \in (0, 1)$. Let $\mathbf{X}$ denote a random $m \times n$ matrix with independent standard Gaussian entries. What is the smallest possible value of $|\mathbf{X}\sigma|_\infty$ that can be achieved algorithmically in polynomial time?*

In particular, it is an open problem as to whether the partial coloring method can be used to guarantee subconstant discrepancy for standard Gaussians when $m = n^\gamma$.

[5] Karmarkar and Karp [1982] give two algorithms for number partitioning. The first one is a simple greedy heuristic, but its analysis was only performed for the uniform distribution over a decade later by Yakir [1996]. Our algorithm presented here generalizes the second one which was rigorously analyzed in the original paper of Karmarkar and Karp [1982].

We suspect that the answer is negative. It seems that even attaining discrepancy $o(\sqrt{m})$ serves as a natural bottleneck for such an approach.

2.3 Gaussian discrepancy in sub-linear dimension

The main goal of this section is to prove the following proposition. Throughout, we adopt the shorthand notation $u_n \lesssim_n v_n$ for $u_n \le v_n(1+o(1))$ and $u_n \simeq_n v_n$ for $u_n = v_n(1+o(1))$.

Proposition 2.1. *Fix $\gamma > 1$, $\omega(1) = m = o(n)$, and let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(0, I_m)$ be independent standard Gaussian random vectors. Then*

$$\lim_{n\to\infty} \mathbb{P}\Big[\mathcal{D}(X_1,\dots,X_n) \le \gamma\sqrt{\frac{\pi n}{2}}2^{-n/m}\Big] = 1\,.$$

We first outline our proof strategy based on the second moment method. Set $\varepsilon = \varepsilon(n) = \gamma 2^{-n/m}\sqrt{\pi n/2}$ and define S, the number of low discrepancy solutions, to be

$$S = \sum_{\sigma\in\{\pm1\}^n} \mathbb{1}\Big(\Big|\sum_{i=1}^n \sigma_i X_i\Big|_\infty \le \varepsilon\Big). \tag{2.5}$$

Our goal is to show that $\mathbb{E}[S^2]/\mathbb{E}[S]^2 = 1 + o(1)$. By the second moment method (2.4), this implies the desired result.

The next lemma gives a useful form for the first and second moments of S and follows from a straightforward calculation. Its proof is postponed to Appendix 2.5.1.

Lemma 2.1. *The random variable S defined as in* (2.5) *has its first two moments given by*

$$\mathbb{E}[S] = 2^n\mathbb{P}\Big(|Z| \le \frac{\varepsilon}{\sqrt{n}}\Big)^m \tag{2.6}$$

where $Z \sim \mathcal{N}(0,1)$, and

$$\mathbb{E}[S^2] = 2^n\sum_{k=0}^n \binom{n}{k}\mathbb{P}_{\rho_k}\left(\left|\sqrt{n}X\right| \le \varepsilon\,,\,\left|\sqrt{n}Y\right| \le \varepsilon\right)^m\,. \tag{2.7}$$

Here $\rho_k = 1 - 2k/n$ and $\mathbb{P}_{\rho_k}$ denotes the joint distribution of (X,Y) with $X, Y \sim \mathcal{N}(0,1)$ having correlation ρ_k.

Given this representation, we proceed in two steps to prove an upper bound on the second moment $\mathbb{E}[S^2]$:

(i) We first apply a truncation argument to show that the contribution from the $k \le n/4$ and $k \ge 3n/4$ terms in the summand of (2.7) is negligible. See Lemma 2.5 and its proof in Appendix 2.5.1 for details.

(ii) Then we show that the dominant contribution in the summation (2.7) is asymptotically bounded by $\mathbb{E}[S]^2$ and comes from an interval of length $\Theta(\sqrt{n})$ around

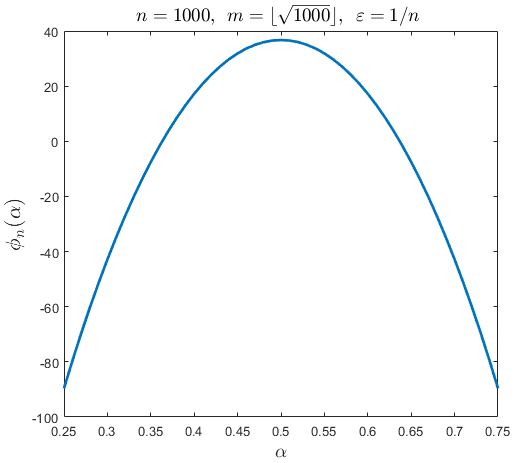


Figure 2-1: *$\alpha \mapsto \phi_n(\alpha)$ for $n = 1000, m = \lfloor\sqrt{1000}\rfloor$, and $\varepsilon = 1/n$.*

$k \simeq n/2$. This part is somewhat delicate and we apply the *Laplace method* to obtain sharp bounds.

By step (i), it suffices to control the leading term

$$L := 2^n \sum_{k=n/4}^{3n/4} \binom{n}{k} \mathbb{P}_{\rho_k}\left(\left|\sqrt{n}X\right| \le \varepsilon\,,\ \left|\sqrt{n}Y\right| \le \varepsilon\right)^m. \tag{2.8}$$

To that end, approximate the above binomial coefficient using Lemma C.2 in Berthet et al. [2018]: For any $l \in (0, 1/2]$, $\alpha \in (l, 1-l)$ such that $n\alpha$ is an integer, it holds

$$\exp\Big(-\frac{1}{12l^2n}\Big) \le \sqrt{2\pi n\alpha(1-\alpha)}\exp(-nh(\alpha))\binom{n}{\alpha n} \le \exp\Big(\frac{1}{12n}\Big),$$

where $h(\alpha) = -\alpha\log\alpha - (1-\alpha)\log(1-\alpha)$ denotes the binary entropy with $h(0) = h(1) = 0$. Therefore, it holds that

$$L \lesssim_n \frac{2^n}{\sqrt{2\pi n}} \sum_{k=n/4}^{3n/4} \exp(\phi_n(\alpha_k)) \tag{2.9}$$

where $\alpha_k = k/n$ and

$$\phi_n(\alpha) = nh(\alpha) + m\log(\mathbb{P}_{1-2\alpha}\left[\left|\sqrt{n}X\right| \le \varepsilon\,,\ \left|\sqrt{n}Y\right| \le \varepsilon\right]) - \frac{1}{2}\log\alpha(1-\alpha). \tag{2.10}$$

Moreover, as justified in Lemma 2.6 (see Appendix 2.5.1), for n sufficiently large, $\phi_n(\alpha)$ is a strictly concave function on $[0.25, 0.75]$ with a unique maximum at $\alpha = 0.5$. See Figure 2-1 for the graph of $\phi_n(\alpha)$ for a specific setting of the parameters. Thus

we can make the Riemann sum approximation

$$L \lesssim_n \frac{2^n}{\sqrt{2\pi n}} \sum_{k=n/4}^{3n/4} \exp(\phi_n(\alpha_k)) \lesssim_n \frac{\sqrt{n}2^n}{\sqrt{2\pi}} \int_{1/4}^{3/4} \exp(\phi_n(\alpha)) d\alpha. \tag{2.11}$$

Our goal now is to employ the Laplace method [see, e.g., Murray, 1984], a well-known technique from asymptotic analysis, to compute explicitly the asymptotic growth of the right-hand-side above. It consists in performing a second-order Taylor expansion of ϕ_n in order to reduce the problem to the computation of a Gaussian integral.

Lemma 2.2. *Suppose that $m = o(n)$ and set $\varepsilon = \gamma 2^{-n/m}\sqrt{n\pi/2}$. Recall the definition of S from (2.5). Then*

$$L \lesssim_n \mathbb{E}[S]^2. \tag{2.12}$$

Proof. We apply the Laplace method to the integral in (2.11). Let $\eta \in (0,1)$ be arbitrary, and define $g_n(\alpha) = \phi_n(\alpha)/n$. Since $h''(\alpha)$ is continuous, Lemma 2.6 implies that there exists $\delta = \delta(\eta)$ and $N = N(\eta)$ such that

$$\frac{1}{n}\,|\phi_n''(\alpha) - \phi_n''(1/2)| \le \eta\,, \quad \forall \alpha \in (1/2-\delta, 1/2+\delta),\, n \ge N. \tag{2.13}$$

The above inequality follows by writing $g_n''(\alpha) = h''(\alpha) + r_n(\alpha)$, where $r_n(\alpha)$ is a remainder term that goes to 0 uniformly in $\alpha \in (0.25, 0.75)$ as $n \to \infty$, using Lemma 2.6. Using that the remainder term is small and $h''(\alpha)$ is continuous at $\alpha = 1/2$, we arrive at (2.13).

By (2.13) and Taylor's theorem,

$$\phi_n(\alpha) - \phi_n(1/2) \le \frac{1}{2}(\phi_n''(1/2) + \eta n)(\alpha - 1/2)^2\,, \quad \forall \alpha \in (1/2-\delta, 1/2+\delta),\, n \ge N. \tag{2.14}$$

Moreover,

$$\phi_n''(1/2) + \eta n < 0 \tag{2.15}$$

for n sufficiently large because $\eta \in (0,1)$ and $\phi_n''(1/2) \simeq_n -4n$ by Lemma 2.6. Therefore, since ϕ_n is increasing on $(0.25, 0.75)$ for n sufficiently large,

$$\begin{aligned} \frac{\sqrt{n}}{\exp(\phi_n(1/2))} \int_{1/4}^{1/2-\delta} \exp(\phi_n(\alpha))\, d\alpha &\lesssim_n 10\sqrt{n}\exp(\phi_n(1/2-\delta) - \phi_n(1/2)) \\ &\lesssim_n 10\sqrt{n}\exp\left(\frac{1}{2}(\phi_n''(1/2)+\eta n)\delta^2\right) = o(1), \end{aligned} \tag{2.16}$$

where we applied (2.14) and (2.15). By symmetry of $\phi_n(\alpha)$ about $\alpha = 1/2$, the

integral as in (2.16) from $1/2+\delta$ to $3/4$ is negligible. Moreover, by (2.14),

$$\int_{1/2-\delta}^{1/2+\delta} \exp(\phi_n(\alpha))\,d\alpha \lesssim_n \int_{1/2-\delta}^{1/2+\delta} \exp\left(\phi_n(1/2) + \frac{1}{2}(\phi_n''(1/2)+\eta n)(\alpha-1/2)^2\right) d\alpha \tag{2.17}$$
$$\lesssim_n \exp(\phi_n(1/2))\sqrt{\frac{2\pi}{|\phi_n''(1/2)+\eta n|}} = 2^n f_n^m \sqrt{\frac{2\pi}{n(1-\eta/4)}},$$

where

$$f_n = \mathbb{P}_0(\left|\sqrt{n}X\right| \le \varepsilon, \left|\sqrt{n}Y\right| \le \varepsilon).$$

Since $\eta \in (0,1)$ was arbitrary, we conclude by (2.6), (2.9), (2.11), (2.16), (2.16), and the definition of f_n that

$$L \lesssim_n \frac{2^n}{\sqrt{2\pi n}} \cdot n \cdot \int_{1/4}^{3/4} \exp(\phi_n(\alpha))\,d\alpha \lesssim_n 2^{2n} f_n^m = \mathbb{E}[S]^2.$$

□

Proof of Proposition 2.1. We see that $\mathbb{E}[S^2]/\mathbb{E}[S]^2 \lesssim_n 1$ as $n \to \infty$ applying Lemma 2.1, Lemma 2.5, (2.8), (2.9), and Lemma 2.2. Proposition 2.1 follows by the second moment method. □

We complement Proposition 2.1 with a near-matching lower bound.

Proposition 2.2. *Let $\omega(1) = m = o(n)$, fix $\gamma < 1$, and let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(0, I_m)$ be independent standard Gaussian random vectors. Then*

$$\lim_{n\to\infty} \mathbb{P}\left[\mathcal{D}(X_1,\dots,X_n) \le \gamma\sqrt{\frac{\pi n}{2}}2^{-n/m}\right] = 0.$$

Proof. Recall the definition of S as in (2.5), which counts the number of signings with discrepancy $\varepsilon = \gamma 2^{-n/m}\sqrt{\pi n/2}$. By the Markov inequality, (2.20), and (2.6),

$$\mathbb{P}\left[S > 1\right] \le \mathbb{E}[S] = 2^n \mathbb{P}\left[|Z| < \gamma\sqrt{\frac{\pi n}{2}}2^{-n/m}\right]^m \lesssim_n \gamma^m \to 0$$

because $\omega(1) = m = o(n)$ and $\gamma < 1$. This completes the proof. □

Our first main result, Theorem 2.1, is a direct consequence of Propositions 2.1 and 2.2.

2.4 Algorithmic discrepancy minimization in low dimension

Now we describe our approach for proving Theorem 2.2. In this section we introduce the generalized Karmarkar–Karp algorithm **GKK**. Recall that the goal is to find

algorithmically $\sigma \in \{\pm 1\}^n$ such that $|\mathbf{X}\sigma|_\infty$ is small. As in Karmarkar and Karp [1982], our algorithm is a *differencing method*, which means that throughout the algorithm, we maintain a set of vectors S, and our basic operations consist of removing two vectors, say x and y, from S and then adding the difference to S : $S \leftarrow S \cup \{x-y\}\backslash\{x,y\}$. We perform a sequence of these differencing operations in a judicious way until there is a single vector v remaining in S. Note that at any given time, the elements of S correspond to (disjoint) partial signed sums of the original vectors $X_1, \ldots, X_n$. Hence, the final vector $v \in S$ is indeed a signed sum of the original vectors. It is possible to keep track of the final signing by tracking the differences, though we do not do so explicitly.

Next, we informally describe the **GKK** differencing method in detail. For simplicity, we assume that $\Delta = 1$ in this description. The algorithm **GKK** is a recursive procedure that consists of $\Theta(\log n)$ phases. For the first phase of the recursion, given a collection of n vectors lying in $[-1,1]^m$, we partition this cube into sub-cubes of side length $\alpha = n^{-\Omega(1/m)}$. The idea is that with sub-cubes of this size, we are likely to have multiple points in each sub-cube, and these points would be very close to each other. We then randomly difference the vectors in each sub-cube until there is at most one point left in each sub-cube. Next, we enter a *clean-up* step to deal with the leftover vectors. First we combine the leftover vectors (at most one per each sub-cube) via a standard differencing algorithm that we call **REDUCE** into a single 'bad' vector $v^{(0)}$ and let $G' \subset [-\alpha,\alpha]^m$ denote the vectors formed from random differencing. Next we make the entries of the bad vector small by adding signed combinations of a few vectors from G'. Namely, we draw at random points from G' and greedily difference them against $v^{(0)}$ until the resulting vector is sufficiently small in the Euclidean norm. Specially, our update procedure for this clean-up step is

$$\begin{aligned} v^{(k)} &= v^{(k-1)} + a^* \mathbf{u}_k \\ a^* &= \operatorname*{argmin}_{a \in \{\pm 1\}} \left| v^{(k-1)} + a\mathbf{u}_k \right|_2 . \end{aligned} \tag{2.18}$$

where $\mathbf{u}_k$ is drawn at random from the remaining vectors is G'.

Once we have $v^{(k)} \in [-O_m(\alpha), O_m(\alpha)]^m$, we stop drawing random vectors from G', and this ends the first phase of recursion. The remaining vectors form the input to the second phase, which applies the same procedure as above on the smaller cube $[-\alpha,\alpha]^m$. Moreover, subsequent phases follow the same pattern: **partition**, **difference**, and **clean-up**. After each phase, the input cube shrinks by a factor of $n^{-\Omega(1/m)}$. Hence, after a logarithmic number of phases, the remaining vectors lie in a cube of side length $n^{-\Omega(n/m)} = e^{-\Omega(\log^2 n/m)}$. We then apply **REDUCE** to combine the remaining vectors into a single vector with discrepancy as in Theorem 2.2.

We remark that our algorithm also features a resampling step that happens immediately after partitioning. In each phase, this resampling procedure labels points as 'good' or 'bad' so that the good points are independent and have independent coordinates that have a nice distribution. This same resampling trick was also used in Karmarkar and Karp [1982] and is essential for (most of) the remaining random vectors at the end of each phase to have a nice distribution facilitating a recursive

analysis. Moreover, the **partition** and **difference** steps of our algorithm are also similar to those used in Karmarkar and Karp [1982] for the one-dimensional case.

In summary, the algorithm **GKK** consists of several phases of a subroutine **PRDC**, which stands for partition, resample, difference, clean-up, that we now define explicitly. In the first part of the clean-up phase, we remark that the aforementioned algorithm **REDUCE** is applied. However, we defer the explicit description of this algorithm, which uses standard techniques, to Appendix 2.5.3, instead stating its key property of use.

Lemma 2.3. *Given $X_1, \ldots, X_N \in \mathbb{R}^m$, the algorithm* **REDUCE** *is polynomial-time and outputs $\sigma \in \{\pm 1\}^N$ such that*

$$\left|\sum_{i=1}^{N} \sigma_i X_i\right|_\infty \leq \max_{S \subset [N]:|S|=m} \sum_{j \in S} |X_j|_\infty . \tag{2.19}$$

In the explicit description of **PRDC** below, $\gamma > 0$ denotes a fixed absolute constant to be set later (see Appendix 2.5.5).

PRDC:

Input: A number $\alpha_t > 0$. A set of vectors $S_t \subset [-\alpha_t, \alpha_t]^m$. A single vector $v_t \subset \gamma m [-\alpha_t, \alpha_t]^m$. A pdf $g_t : [-\alpha_t, \alpha_t]^m \to \mathbb{R}$. Define $N_t = 2^m \lceil |S_t|^{1/(4m)} \rceil^m$.

1. **Partition:** Define $\alpha_{t+1} = \alpha_t / \lceil |S_t|^{1/(4m)} \rceil$. Divide the cube $[-\alpha_t, \alpha_t]^m$ into N_t disjoint sub-cubes $C_1, \ldots, C_{N_t}$ that are of the form $\alpha_{t+1} z + [0, \alpha_{t+1}]^m$ for some integer vector $z \in \mathbb{Z}^m$.

2. **Resample:** Independently for every vector x in S_t, if $x \in C_j$, then label x as 'good' with probability $(\min_{y \in C_j} g_t(y))/g_t(x)$. Otherwise, label x to be 'bad.' Let G_t denote the set of good points and B_t denote the set of bad points.

3. **Difference:** For every sub-cube C_j, pick uniformly at random two points in $G_t \cap C_j$, include their difference in G'_t, and remove them from G_t. Continue this until $G_t \cap C_j$ has at most 1 good point for every j. Let B'_t be the union of B_t, v_t, and the leftover good points.

4. **Clean-up:**

   (a) Apply **REDUCE** to the vectors in B'_t to obtain σ. Define $v_t^{(0)} = \sum_{b_i \in B'_t} \sigma_i b_i$.

   (b) For $k = 0, 1, 2, \ldots$

   If $\left|v_t^{(k)}\right|_2 \geq \gamma m \alpha_{t+1}$: remove uniformly at random a point $x \in G'_t$. Define $v_t^{(k+1)} = v_t^{(k)} + a^* x$ where $a^* = \operatorname{argmin}_{a \in \{\pm 1\}} |v_t^{(k)} + ax|_2$. Define $G'_t \leftarrow G'_t \backslash \{x\}$.

   Else: $v_{t+1} := v_t^{(k)}$. BREAK

Output: $S_{t+1} := G'_t$, v_{t+1}, $\alpha_{t+1} := \alpha_t / \lceil |S_t|^{1/(4m)} \rceil$

Now we explicitly describe our main algorithm **GKK** in terms of the subroutine **PRDC**. Recall that ρ is the density corresponding to a particular entry of $\mathbf{X}$. First we need the following definition.

Definition 2.1 (Triangular distribution)**.** *A random vector* $\mathbf{y} \in \mathbb{R}^m$ *follows a* triangular distribution *on the cube* $[-R, R]^m$ *if the distribution of* $\mathbf{y}$ *is given by* $\mathbf{u} - \mathbf{v}$*, where* $\mathbf{u}$ *and* $\mathbf{v}$ *are independent and uniformly distributed on* $[0, R]^m$*. Notationally, we write* $\mathbf{y} \sim \mathrm{Tri}[-R, R]^m$*.*

GKK:

Input: An $m \times n$ matrix $\mathbf{X}$. A probability density function $\rho : [-\Delta, \Delta] \to \mathbb{R}$. Let $T = \lceil C^* \log n \rceil$ where $C^* := (2 \log(10/3))^{-1}$.

1. Set $S_1 = \mathrm{col}(\mathbf{X}), \alpha_1 = \Delta$, $v_1 = \mathbf{0}$, and $g_1 = \rho^{\otimes m}$.
2. For $t = 1, 2, \ldots, T$:
   (a) Run **PRDC** on the input data S_t, v_t, α_t, g_t to output S_{t+1}, v_{t+1}, and α_{t+1}.
   (b) Set $g_{t+1}(x) = \frac{1}{\alpha_{t+1}} f(x/\alpha_{t+1})$, where $f(x)$ is the triangular density on $[-1, 1]^m$.
3. Apply **REDUCE** to the vectors in $S_T \cup \{v_T\}$ to obtain σ. Let $v = \sum_{s_i \in S_T \cup \{v_T\}} \sigma_i s_i$.

Output: $|v|_\infty$

We remark that the first three steps of **PRDC** are similar to those in the corresponding subroutine in Karmarkar and Karp [1982] for the one-dimensional case. The clean-up step and its analysis on the other hand are quite different. In particular, we use **REDUCE** to combine the 'bad' vectors left over from resampling into a single bad vector $v^{(0)}$. This subroutine is quite similar to the algorithm used by Beck–Fiala to show that t-sparse vectors have discrepancy at most $2t - 1$ [Beck and Fiala, 1981]. In contrast, Karmarkar and Karp [1982] use a greedy iterative algorithm for dealing with bad points in dimension 1, but it is not clear how to generalize their algorithm to also work in higher dimensions. In the next part of the clean-up step, we must bring the bad vector $v^{(0)}$ into a smaller range. Karmarkar and Karp [1982] do this by randomly sampling points from G' and greedily differencing them against $v^{(0)}$ until the resulting number is small. Here we use the same approach, but since we are working in higher dimensions, we measure the resulting vector in the Euclidean norm. In this part of the clean-up step, the key difference between our work and Karmarkar and Karp [1982] lies in our analysis, which includes elements of the analysis of stochastic gradient descent, as well as martingale concentration and the Khintchine inequality (see Appendix 2.5.5).

We also comment on the reason for the bound $m = O(\sqrt{\log n})$ in Theorem 2.2. First observe that by our choice of $\alpha = n^{-\Omega(1/m)}$ for the side-lengths of the sub-cubes at the first phase, it is necessary that $m = O(\log n)$; otherwise the sub-cubes are not smaller than the original cube. The reason we require the stronger condition $m = O(\sqrt{\log n})$ is so that not too many points are labeled 'bad' in the resampling

step of our algorithm. We direct the reader to Appendix 2.5.4 for the analysis and further discussion.

2.4.1 Analysis of GKK

The proof of Theorem 2.2 follows from a sequence of inductive assumptions. Recall that S_t denotes the points input to the t^{th} phase of **PRDC**, excluding the single 'bad' vector $v_t \in \gamma m[-\alpha_t, \alpha_t]^m$, where γ is a fixed absolute constant to be determined. Recall that $C^* = (2\log(10/3))^{-1}$, as set in the definition of **GKK**, and that $\Delta > 0$ is the side length of the cube containing the initial set of vectors S_1.

Proposition 2.3. *Let $X_1, \dots, X_n$ be iid random vectors, each having a joint density $g : [-\Delta, \Delta]^m \to \mathbb{R}$. Consider the output S_t, v_t, α_t that results after the $(t-1)$-th phase of* **PRDC** *in step 2 of* **GKK**. *Then conditioned on $|S_j| = n_j$ for $1 \le j \le t$, we have*

- *the n_t points in S_t are iid and follow a triangular distribution on $[-\alpha_t, \alpha_t]^m$, and*
- *the random vector v_t is independent of the vectors in S_t.*

Proposition 2.3 ensures that the distribution of the output of each phase of recursion is preserved, allowing us to apply induction. At the heart of this result is the following marginal calculation which implies that the good points have a uniform distribution on their respective sub-cubes. Conditioning on $X_1 \in C_1$, if L denotes the label of X_1 as 'good' or 'bad', then (X_1, L) has a mixed joint density $p(x, \ell)$ where $x \in C_1$ and $\ell \in \{\text{'good'}, \text{'bad'}\}$, which by Bayes' rule satisfies

$$p(x|L = \text{'good'}) = \frac{p(x, \text{'good'})}{\mathbb{P}[L = \text{'good'}]} = \frac{g(x) \cdot \frac{\min_{y \in C_1} g(y)}{g(x)}}{\int_{C_1} p(y, \text{'good'}) \mathrm{d}y} = \frac{1}{\mathrm{Vol}(C_1)},$$

for all $x \in C_1$.

The proofs of Propositions 2.4 and 2.5 below are postponed to Appendices 2.5.4 and 2.5.5, respectively. The former relies on showing that a large fraction of the points input to the t^{th} phase are labeled 'good' in the **resample** step, and the latter requires us to show that few of the random differences created in step 3 of **PRDC** are lost in the **clean-up** step.

Proposition 2.4. *Suppose that $1 \le t \le C^* \log n$ and $m \le C\sqrt{(\log n)/\max(1, \log \Delta)}$, where C is a sufficiently small absolute constant. Then for some fixed θ, conditioned on the events $|S_j| \ge \theta^{j-1} n$ for all $1 \le j \le t$, it holds that the set G_t' of random differences created in step 2 of the t^{th} phase of* **PRDC** *satisfies $|G_t'| \ge \beta |S_t|$ for some fixed β with probability at least $1 - \exp(-c_1\sqrt{n})$, where $c_1 > 0$ is an absolute constant. In particular, we may set $\theta = 0.3$ and $\beta = 0.4$.*

Proposition 2.5. *Suppose that $1 \le t \le C^* \log n$ and $m \le C\sqrt{\log n}$, where C is a sufficiently small absolute constant. Then conditioned on the events $|G_t'| \ge \beta |S_t|$ and $|S_j| \ge \theta^{j-1} n$ for $1 \le j \le t$, it holds that the set S_{t+1} (the input to the $(t+1)$-th*

iteration of **PRDC***) satisfies* $|S_{t+1}| \geq \theta |S_t|$ *with probability at least* $1 - \exp(-c_2 n^{1/4})$*, where* $c_2 > 0$ *is an absolute constant. In particular, we may choose* $\beta = 0.4$ *and* $\theta = 0.3$*.*

The proof of Theorem 2.2 follows easily from the previous two propositions and is found in Appendix 2.5.6.

2.5 Appendix

2.5.1 Proofs from Section 2.3

First, we calculate the first and second moments of S as defined in (2.5).

Proof of Lemma 2.1. Let $X_i^{(j)}$ denote the jth element of the vector X_i. Since these elements are independent, we get

$$\mathbb{E}[S] = \sum_{\sigma \in \{\pm 1\}^n} \prod_{j=1}^{m} \mathbb{P}\Big(\Big| \sum_{i=1}^{n} \sigma_i X_i^{(j)} \Big| \leq \varepsilon \Big) = 2^n \mathbb{P}\Big(|Z| \leq \frac{\varepsilon}{\sqrt{n}} \Big)^m$$

where $Z \sim \mathcal{N}(0,1)$. This completes the proof of (2.6).

To prove (2.7), let $d(\tau, \sigma)$ denotes the Hamming distance between σ and τ. Observe that if τ and σ satisfy $d(\tau, \sigma) = k$, then $X := \frac{1}{\sqrt{n}} \sum_{i=1}^n \sigma_i X_i^{(j)}$ and $Y := \frac{1}{\sqrt{n}} \sum_{i=1}^n \tau_i X_i^{(j)}$ are ρ_k-correlated standard Gaussians random variables. Thus

$$\begin{aligned}
\mathbb{E}[S^2] &= \sum_{\sigma, \tau \in \{\pm 1\}^n} \mathbb{P}\Big(\Big| \sum_{i=1}^{n} \sigma_i X_i \Big|_\infty \leq \varepsilon \,,\ \Big| \sum_{i=1}^{n} \tau_i X_i \Big|_\infty \leq \varepsilon \Big) \\
&= \sum_{\sigma} \sum_{k=0}^{n} \sum_{\tau \colon d(\tau,\sigma) = k} \mathbb{P}_{\rho_k}\Big(\Big| \sqrt{n} X \Big| \leq \varepsilon \,,\ \Big| \sqrt{n} Y \Big| \leq \varepsilon \Big)^m \\
&= 2^n \sum_{k=0}^{n} \binom{n}{k} \mathbb{P}_{\rho_k}\left(\Big| \sqrt{n} X \Big| \leq \varepsilon \,,\ \Big| \sqrt{n} Y \Big| \leq \varepsilon \right)^m ,
\end{aligned}$$

which proves the lemma. □

The following small-ball probability estimates are required for the proof of the truncation argument, Lemma 2.5.

Lemma 2.4. *Let* Z *denote a standard Gaussian random variable, and let* X, Y *denote* ρ*-correlated standard Gaussian random variables with* $\rho \in (-0.5, 0.5)$*. Then for* $0 < z < 1$*, we have for some absolute constant* $c > 0$ *that*

$$-cz^3 \leq \mathbb{P}\Big[|Z| \leq z \Big] - \sqrt{\frac{2}{\pi}} z \leq 0, \tag{2.20}$$

and for all $z \in (0,\infty)$, we have

$$\mathbb{P}_\rho\Big[|X| \le z, |Y| \le z\Big] \le \frac{2}{\pi\sqrt{1-\rho^2}} z^2. \tag{2.21}$$

Proof. Observe that $z \mapsto \mathbb{P}[|Z| \le z]$ is a concave function for $z \ge 0$. Hence, it lies below the tangent line to this curve at $z = 0$, which is precisely the function $z \mapsto \sqrt{2/\pi} z$. This proves the right-hand-side of (2.20). To prove the left-hand-side, we apply Taylor expansion and observe that for $|z| \le 1$, it holds that

$$\mathbb{P}[|Z| \le z] = \sqrt{\frac{2}{\pi}} z - \frac{1}{6}\sqrt{\frac{2}{\pi}} z^3 \pm O(z^5) \ge \sqrt{\frac{2}{\pi}} z - cz^3$$

for some absolute constant $c > 0$. To prove (2.21), note that the joint density $\psi_\rho(x,y)$ of a pair of standard normal ρ-correlated Gaussians satisfies

$$\psi_\rho(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big(-\frac{x^2 - 2\rho xy + y^2}{2-2\rho^2}\Big) \le \frac{1}{2\pi\sqrt{1-\rho^2}}.$$

The upper bound follows by positive-semidefiniteness of the covariance matrix. Hence, integrating over the rectangle $|x| \le z, |y| \le z$ and applying the above upper bound yields the desired result.

□

Lemma 2.5. *Suppose that $\omega(1) = m = o(n)$ and let $\varepsilon = \varepsilon(n) = \gamma 2^{-n/m}\sqrt{\pi n/2}$ for some $\gamma > 1$. Then*

$$2^n \sum_{k=0}^{n/4} \binom{n}{k} \mathbb{P}_{\rho_k}\Big(\Big|\sqrt{n}X\Big| \le \varepsilon\,,\ \Big|\sqrt{n}Y\Big| \le \varepsilon\Big)^m = o(\mathbb{E}[S]^2). \tag{2.22}$$

$$2^n \sum_{k=3n/4}^{n} \binom{n}{k} \mathbb{P}_{\rho_k}\Big(\Big|\sqrt{n}X\Big| \le \varepsilon\,,\ \Big|\sqrt{n}Y\Big| \le \varepsilon\Big)^m = o(\mathbb{E}[S]^2). \tag{2.23}$$

Proof. Note that (2.23) follows from (2.22) by symmetry, so it suffices to prove (2.22). We may write $m = n/g_n$ for some sequence g_n such that $\omega(1) = g_n = o(n)$. For notational convenience, define

$$f_n(\rho) = \mathbb{P}_\rho(\Big|\sqrt{n}X\Big| \le \varepsilon, \Big|\sqrt{n}Y\Big| \le \varepsilon).$$

By Lemma 2.1, we have

$$\begin{aligned}\frac{2^n \sum_{k=0}^{n/4} \binom{n}{k} \mathbb{P}_{\rho_k}\left(|\sqrt{n}X| \le \varepsilon\,,\ |\sqrt{n}Y| \le \varepsilon\right)^m}{\mathbb{E}[S]^2} \\ = \underbrace{\sum_{k=0}^{n/(g_n)^2} \frac{\binom{n}{k}}{2^n}\left(\frac{f_n(\rho_k)}{f_n(0)}\right)^m}_{=:A} + \underbrace{\sum_{k=n/(g_n)^2}^{n/4} \frac{\binom{n}{k}}{2^n}\left(\frac{f_n(\rho_k)}{f_n(0)}\right)^m}_{=:B}.\end{aligned} \tag{2.24}$$

For ε as above and $Z \sim N(0,1)$, we have by applying (2.20) that

$$2^n \mathbb{P}(|Z| < \varepsilon/\sqrt{n})^m \geq 2^n \left(\sqrt{\frac{2}{\pi n}}\varepsilon\right)^m (1 - c\varepsilon^2/n)^m \gtrsim_n \left(\frac{\gamma+1}{2}\right)^m, \tag{2.25}$$

where c is an absolute constant. To obtain the right-hand-side, note that $\varepsilon/\sqrt{n} \xrightarrow{n\to\infty} 0$ since $m = o(n)$. Thus, for n sufficiently large it holds that

$$1 - c\varepsilon^2/n \geq \frac{1}{2}\left(1 + \frac{1}{\gamma}\right),$$

which yields the right-hand-side of (2.25). Now using the crude bound $f_n(\rho_k) \leq \mathbb{P}(|\sqrt{n}Z| \leq \varepsilon)$, (2.25), the fact that $f_n(0) = \mathbb{P}(|\sqrt{n}Z| \leq \varepsilon)^2$, and the inequality

$$\sum_{k=0}^{j} \binom{n}{k} \leq \left(\frac{ne}{j}\right)^j,$$

we have

$$\begin{aligned} A &= \sum_{k=0}^{n/(g_n)^2} \frac{\binom{n}{k}}{2^n} \left(\frac{f_n(\rho_k)}{f_n(0)}\right)^m \\ &\lesssim_n \left(\frac{\gamma+1}{2}\right)^{-m} (e\, g_n^2)^{n/g_n^2} \\ &= \exp\left(-\frac{n \log \frac{1}{2}(1+\gamma)}{g_n} + \frac{n}{g_n^2} + \frac{2n \log g_n}{g_n^2}\right) = o(1) \end{aligned} \tag{2.26}$$

because $(1/2)(1+\gamma) > 1$, $g_n \to \infty$, and $n/g_n \to \infty$ as $n \to \infty$.

By (2.20) and (2.21) (noting again that $f_n(0) = \mathbb{P}(|\sqrt{n}Z| \leq \varepsilon)^2$), we have

$$B = \sum_{k=n/(g_n)^2}^{n/4} \frac{\binom{n}{k}}{2^n} \left(\frac{f_n(\rho_k)}{f_n(0)}\right)^m \lesssim_n (c')^m \sum_{k=n/(g_n)^2}^{n/4} \frac{\binom{n}{k}}{2^n} \left(\frac{n^2}{k(n-k)}\right)^{m/2} \tag{2.27}$$

where c' is an absolute constant. By the Hoeffding bound, letting c'' denote another absolute constant, we have

$$(2.27) \lesssim_n (c'')^m g_n^m e^{-n/8} = \exp\left(\frac{n \log c''}{g_n} + \frac{n \log g_n}{g_n} - \frac{n}{8}\right) = o(1)$$

since $g_n \to \infty$. Since $A, B = o(1)$, we conclude by (2.24) that (2.22) holds, as desired. □

Lemma 2.6. *Suppose that $m = o(n)$ and set $\varepsilon = \gamma 2^{-n/m}\sqrt{n\pi/2}$. Then the function $\alpha \mapsto \phi_n(\alpha)$ defined in (2.10) is asymptotically strictly concave on $(0.25, 0.75)$. More precisely,*

$$\lim_{n\to\infty} \frac{1}{n}\frac{\partial^2}{\partial\alpha^2}\phi_n(\alpha) = -\frac{1}{\alpha(1-\alpha)} < -4\,, \qquad \forall\, \alpha \in (0.25, 0.75)\,,$$

and the convergence is uniform over $\alpha \in (0.25, 0.75)$. Moreover, for n large enough, $\phi_n(\alpha)$ has a unique maximum over $(0.25, 0.75)$ located at $\alpha = 0.5$.

Proof. Because $|\partial_\alpha^2 \log \alpha(1-\alpha)| = O(1)$ for $\alpha \in (0.25, 0.75)$, $m = o(n)$, and

$$h''(\alpha) = -\frac{1}{\alpha(1-\alpha)},$$

to verify the strict concavity of $\phi_n(\alpha)$, it suffices to show that

$$\left|\frac{\partial^2}{\partial\alpha^2} \log \mathbb{P}_{1-2\alpha}\left[\left|\sqrt{n}X\right| \le \varepsilon\,,\ \left|\sqrt{n}Y\right| \le \varepsilon\right]\right| = O(1), \quad \alpha \in (0.25, 0.75). \tag{2.28}$$

For notational convenience, we write $f_n(\rho) = \mathbb{P}_\rho\left(|\sqrt{n}X| \le \varepsilon\,,\ |\sqrt{n}Y| \le \varepsilon\right)$. We study the logarithmic second derivative

$$J_n(\rho) := \frac{f_n''(\rho)}{f_n(\rho)} - \Big(\frac{f_n'(\rho)}{f_n(\rho)}\Big)^2 \tag{2.29}$$

by controlling each term individually.

First, recall that for any $\rho \in (-1, 1)$, the distribution $\mathbb{P}_\rho$ admits a density with respect to the Lebesgue measure over $\mathbb{R}^2$ given by

$$\psi_\rho(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big(-\frac{x^2 - 2\rho xy + y^2}{2 - 2\rho^2}\Big).$$

It holds that

$$f_n'(\rho) = \iint_{[-\frac{\varepsilon}{\sqrt{n}}, \frac{\varepsilon}{\sqrt{n}}]^2} \partial_\rho \psi_\rho(x, y)\mathrm{d}x\mathrm{d}y.$$

Thus since $\varepsilon = o(\sqrt{n})$ we get,

$$\lim_{n\to\infty} \frac{f_n'(\rho)}{f_n(\rho)} = \lim_{n\to\infty} \frac{\frac{\varepsilon^2}{n}\iint_{[-\frac{\varepsilon}{\sqrt{n}}, \frac{\varepsilon}{\sqrt{n}}]^2} \partial_\rho \psi_\rho(x, y)\mathrm{d}x\mathrm{d}y}{\frac{\varepsilon^2}{n}\iint_{[-\frac{\varepsilon}{\sqrt{n}}, \frac{\varepsilon}{\sqrt{n}}]^2} \psi_\rho(x, y)\mathrm{d}x\mathrm{d}y} = \frac{\partial_\rho \psi_\rho(0,0)}{\psi_\rho(0,0)} = \partial_\rho \log(\psi_\rho)(0,0). \tag{2.30}$$

Similarly,

$$\lim_{n\to\infty} \frac{f_n''(\rho)}{f_n(\rho)} = \frac{\partial_\rho^2 \psi_\rho(0,0)}{\psi_\rho(0,0)} = \partial_\rho^2 \log(\psi_\rho)(0,0) + \Big(\partial_\rho \log(\psi_\rho)(0,0)\Big)^2. \tag{2.31}$$

Together with (2.29) and (2.30), the above display yields

$$\lim_{n\to\infty} J_n(\rho) = \frac{1+\rho^2}{(1-\rho^2)^2} = O(1),$$

if $\rho \in (-0.5, 0.5)$. Moreover, the convergence in (2.30) and(2.31) is uniform over $\rho \in (-0.5, 0.5)$. This is because the functions $\psi_\rho, \partial_\rho\psi_\rho$, and $\partial_\rho^2\psi_\rho$ are all C-Lipschitz on $\mathbb{R}^2$ for some absolute constant $C > 0$, provided that we restrict $\rho \in (-0.5, 0.5)$. Next, changing variables via $\rho = 1 - 2\alpha$, this verifies (2.28). Thus we have shown that $\phi_n(\alpha)$ is strictly concave on $(0.25, 0.75)$ for n sufficiently large, completing the first part of the proof.

The strict concavity verifies that $\phi_n(\alpha)$ has a unique maximum on $(0.25, 0.75)$. We show that it occurs at $\alpha = 0.5$. It is easy to check that both $h(\alpha)$ and $\alpha \mapsto \log \frac{1}{\sqrt{\alpha(1-\alpha)}}$ have a critical point at $\alpha = 1/2$. So, applying the change of variables $\rho = 1 - 2\alpha$, we just need to verify that $f_n'(0) = 0$. Let $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ denote the density of a standard Gaussian and set $\ell = \varepsilon/\sqrt{n}$. Straightforward calculus shows that

$$\left.\frac{\partial}{\partial\rho}\right|_{\rho=0} \psi_\rho(x, y) = xy\phi(x)\phi(y).$$

Therefore,

$$\left.\frac{\partial}{\partial\rho}\right|_{\rho=0} f_n(\rho) = \left(\int_{-\ell}^{\ell} x\phi(x)\right)^2 = 0.$$

This proves the second part of the lemma, so we're done. □

2.5.2 Gaussian discrepancy in small linear dimension

The goal of this appendix is to prove the result below, which combined with Theorem 2.1 and Theorem 2 of Chandrasekaran and Vempala [2014] provides a precise characterization of asymptotic Gaussian discrepancy.

Theorem 2.3. *Let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(0, I_m)$ be independent standard Gaussian random vectors. Let $\gamma > 1$ denote an arbitrary absolute constant. Then there exists $\Delta = \Delta(\gamma)$ such that for $m \le \Delta n$,*

$$\liminf_{n\to\infty} \mathbb{P}\left[\mathcal{D}(X_1, \dots, X_n) \le \gamma\sqrt{\frac{\pi n}{2}} 2^{-n/m}\right] \ge 0.99\,. \tag{2.32}$$

In particular, combining Theorem 2.3 with Theorem 2 of Chandrasekaran and Vempala [2014], we can now estimate the discrepancy up to constant factor, with probability asymptotically larger than 99%, in the entire linear regime $m = \delta n$ where $\delta > 0$. Note that our guarantee on the probability here is weaker than that of the high-probability upper bound from Theorem 2.1. The constant 0.99 can be boosted to be arbitrarily close to 1 by choosing smaller Δ, though our techniques do not allow us to set the right-hand-side to be 1 for any fixed $\Delta > 0$.

The closely related work of Aubin et al. [2019] also considered Gaussian discrepancy in the linear regime $m = \delta n$ for fixed $\delta > 0$. Subject to a certain numerical hypothesis, the authors showed that

$$\liminf_{n\to\infty} \mathbb{P}\left[\mathcal{D}(X_1, \dots, X_n) \le c(\delta)\sqrt{n}\right] > 0, \tag{2.33}$$

where $c(\delta)$, as a function of δ, is the inverse of the function $x \mapsto \log(1/2)/\mathbb{P}[|Z| \le x]$ and $Z \sim N(0,1)$. Their proof is an application of the second moment method, similar to ours. They also showed the following high-probability lower bound using the first moment method:

$$\lim_{n\to\infty} \mathbb{P}\left[\mathcal{D}(X_1, \dots, X_n) \ge (c(\delta) - \varepsilon)\sqrt{n}\right] = 1, \tag{2.34}$$

where $\varepsilon > 0$ is an arbitrary absolute constant. Aubin et al. [2019] conjectures, with strong evidence using heuristics from statistical mechanics, that the event in (2.33) holds with probability tending to 1. We remark that as $\delta \to 0$, we have $c(\delta) = \Theta(2^{-1/\delta}) = \Theta(2^{-n/m})$. Theorem 2.3 shows that with a constant factor's worth of 'extra room' in the discrepancy threshold, the asymptotic probability in (2.33) can be boosted to be arbitrarily close to 1.

On the algorithmic side, using a mild extension of the techniques of Chandrasekaran and Vempala [2014], in dimension $m = \delta n$ with $\delta \in (0,1)$, one can show an algorithmic bound of $O(\sqrt{\delta n})$ on the discrepancy, and this is the best known result for this regime. Hence, Theorem 2.3 suggests the possibility of a statistical-to-computational gap in the small linear regime $m = \delta n$ for $\delta \in (0,1)$. Note that for $\delta > 1$, the results of Chandrasekaran and Vempala [2014] confirm an absence of statistical-to-computational gaps in the discrepancy.

The proof of Theorem 2.3 follows closely the steps from Section 2.3 with some modifications. We begin with a truncation argument as in Lemma 2.5.

Lemma 2.7. *Let $\gamma > 1$ denote an arbitrary absolute constant. Then there exists $\Delta = \Delta(\gamma)$ such that if $m = \delta n$ for $\delta \le \Delta$ and $\varepsilon = \varepsilon(n) = \gamma 2^{-1/\delta}\sqrt{\pi n/2}$, then*

$$2^n \sum_{k=0}^{n/4} \binom{n}{k} \mathbb{P}_{\rho_k}\left(\left|\sqrt{n}X\right| \le \varepsilon\,,\ \left|\sqrt{n}Y\right| \le \varepsilon\right)^m = o(\mathbb{E}[S]^2). \tag{2.35}$$

$$2^n \sum_{k=3n/4}^{n} \binom{n}{k} \mathbb{P}_{\rho_k}\left(\left|\sqrt{n}X\right| \le \varepsilon\,,\ \left|\sqrt{n}Y\right| \le \varepsilon\right)^m = o(\mathbb{E}[S]^2). \tag{2.36}$$

Proof. The proof follows closely that of Lemma 2.5, setting $g_n \equiv 1/\delta$. We set

$$f_\delta(\rho) = \mathbb{P}_\rho(\left|\sqrt{n}X\right| \le \varepsilon, \left|\sqrt{n}Y\right| \le \varepsilon) = \mathbb{P}_\rho(|X| \le \gamma 2^{-1/\delta}\sqrt{\pi/2},\ |Y| \le \gamma 2^{-1/\delta}\sqrt{\pi/2}).$$

Note that the function f_δ is independent of n by our choice of ε. As in (2.24) from

Lemma 2.5, we let

$$A = \sum_{k=0}^{\delta^2 n} \frac{\binom{n}{k}}{2^n} \left(\frac{f_\delta(\rho_k)}{f_\delta(0)} \right)^m, \quad B = \sum_{k=\delta^2 n}^{n/4} \frac{\binom{n}{k}}{2^n} \left(\frac{f_\delta(\rho_k)}{f_\delta(0)} \right)^m.$$

Note that for δ sufficiently small (depending on γ), it holds that $\varepsilon/\sqrt{n} \leq 1$. Therefore, similar to (2.25), we can apply the lower bound from Lemma 2.4 to conclude that

$$2^n \mathbb{P}[|Z| < \varepsilon/\sqrt{n}]^m \geq 2^n \left(\sqrt{\frac{2}{\pi n}} \varepsilon \right)^m (1 - c\varepsilon^2/n)^m \geq \left(\frac{\gamma+1}{2} \right)^m, \tag{2.37}$$

Hence, as in (2.26) we have

$$A \lesssim_n \left(\frac{\gamma+1}{2} \right)^{-m} (e\delta^{-2})^{\delta^2 n} = \exp\left(-\delta n \log\left(\frac{1}{2}(1+\gamma) \right) + \delta^2 n + 2\delta^2 n \log(1/\delta) \right). \tag{2.38}$$

Hence, if $\delta \leq \Delta(\gamma)$ for $\Delta(\gamma)$ sufficiently small, then we have that $A = o(1)$.

Similar to (2.27), we have by applying (2.20) and (2.21) that

$$B \lesssim_n (c'(\gamma))^m \sum_{k=\delta^2 n}^{n/4} \frac{\binom{n}{k}}{2^n} \left(\frac{n^2}{k(n-k)} \right)^{m/2}. \tag{2.39}$$

By the Hoeffding bound (letting $c''(\gamma)$ denote another constant depending on γ), we have

$$(2.39) \lesssim_n (c''(\gamma))^m \delta^{-m} e^{-n/8} = \exp\left(\delta n \log(c''(\gamma)) + \delta n \log(1/\delta) - n/8 \right) = o(1), \tag{2.40}$$

provided that $\delta \leq \Delta(\gamma)$ for $\Delta(\gamma)$ sufficiently small. Since $A = o(1)$ as well for this setting of parameters, the lemma follows. □

Our next lemma is a version of Lemma 2.6 corresponding to the linear regime. We use the log-concavity of the function ϕ_n when we apply the Laplace method to the second moment, as in the sub-linear regime.

Lemma 2.8. *Let $\eta > 0$ and $\gamma > 1$ be arbitrary constants, and let $\Delta = \Delta(\gamma, \eta)$ denote a sufficiently small absolute constant. Suppose that $m = \delta n$ for $\delta \leq \Delta$, and set $\varepsilon = \gamma 2^{-1/\delta} \sqrt{n\pi/2}$. Then the function $\alpha \mapsto \phi_n(\alpha)$ defined in (2.10) is strictly concave on $(0.25, 0.75)$. More precisely,*

$$\frac{1}{n} \frac{\partial^2}{\partial \alpha^2} \phi_n(\alpha) \leq -\frac{1}{\alpha(1-\alpha)} + \eta < -4 + \eta, \qquad \forall\, \alpha \in (0.25, 0.75). \tag{2.41}$$

Moreover, $\phi_n(\alpha)$ has a unique maximum over $(0.25, 0.75)$ located at $\alpha = 0.5$.

Proof. Recall that

$$f_\delta(\rho) = \mathbb{P}_\rho(|X| \le \gamma 2^{-1/\delta}\sqrt{\pi/2},\ |Y| \le \gamma 2^{-1/\delta}\sqrt{\pi/2}).$$

As in the proof of Lemma 2.6, it suffices to study the logarithmic second derivative with respect to ρ

$$J_\delta(\rho) := \frac{f_\delta''(\rho)}{f_\delta(\rho)} - \Big(\frac{f_\delta'(\rho)}{f_\delta(\rho)}\Big)^2 \tag{2.42}$$

and show that $|J_\delta(\rho)| = O(1)$ for $\rho \in (-0.5, 0.5)$. Recall that ψ_ρ denotes the density associated to $\mathbb{P}_\rho$.

Since $\varepsilon/\sqrt{n} \to 0$ as $\delta \to 0$, we have, similar to (2.30), that

$$\lim_{\delta\to 0} \frac{f_\delta'(\rho)}{f_\delta(\rho)} = \lim_{\delta\to 0} \frac{\frac{\varepsilon^2}{n}\iint_{[-\frac{\varepsilon}{\sqrt{n}},\frac{\varepsilon}{\sqrt{n}}]^2} \partial_\rho\psi_\rho(x,y)\mathrm{d}x\mathrm{d}y}{\frac{\varepsilon^2}{n}\iint_{[-\frac{\varepsilon}{\sqrt{n}},\frac{\varepsilon}{\sqrt{n}}]^2} \psi_\rho(x,y)\mathrm{d}x\mathrm{d}y} = \frac{\partial_\rho\psi_\rho(0,0)}{\psi_\rho(0,0)} = \partial_\rho \log(\psi_\rho)(0,0). \tag{2.43}$$

And similar to (2.31), we have

$$\lim_{\delta\to 0} \frac{f_\delta''(\rho)}{f_\delta(\rho)} = \frac{\partial_\rho^2\psi_\rho(0,0)}{\psi_\rho(0,0)} = \partial_\rho^2 \log(\psi_\rho)(0,0) + \Big(\partial_\rho \log(\psi_\rho)(0,0)\Big)^2. \tag{2.44}$$

It follows that

$$\lim_{\delta\to 0} J_\delta(\rho) = \frac{1+\rho^2}{(1-\rho^2)^2} = O(1)$$

for $\rho \in (-0.5, 0.5)$. Moreover, similar to the proof of Lemma 2.6, the convergence in (2.43) and (2.44) is uniform in δ by the Lipschitzness of $\psi_\rho, \partial_\rho\psi_\rho$, and $\partial_\rho^2\psi_\rho$ over the interval $\rho \in (-0.5, 0.5)$. Therefore, if we take δ sufficiently small with respect to γ, η, then (2.41) holds.

Note that independent of ε, we have that $\rho = 0$ is a critical point of ϕ_n, as shown at the end of the proof of Lemma 2.6. Applying this and making the change of variables $\rho = 1 - 2\alpha$ verifies the last statement of Lemma 2.8. □

Proof of Theorem 2.3. Recall from the definition in (2.8) that

$$L := 2^n \sum_{k=n/4}^{3n/4} \binom{n}{k} \mathbb{P}_{\rho_k}\left(\left|\sqrt{n}X\right| \le \varepsilon\,,\ \left|\sqrt{n}Y\right| \le \varepsilon\right)^m.$$

Applying Stirling's formula and a Riemann sum approximation as in (2.9) and (2.11), respectively, we have that

$$L \lesssim_n 2^n\sqrt{\frac{n}{2\pi}}\int_{1/4}^{3/4} \exp(\phi_n(\alpha))d\alpha. \tag{2.45}$$

Since $\phi_n(\alpha)/n$ is independent of n, we can apply the Laplace method directly [see

Murray, 1984] along with Lemma 2.8 to see that

$$\int_{1/4}^{3/4} \exp(\phi_n(\alpha))d\alpha \lesssim_n \sqrt{\frac{2\pi}{|\phi_n''(1/2)|}} \exp(\phi_n(1/2)) \leq \sqrt{\frac{2\pi}{n(4-\eta)}} 2^{n+1} f_\delta(0)^m. \quad (2.46)$$

assuming $\delta \leq \Delta$ for $\Delta(\gamma, \eta)$ sufficiently small.

Therefore, by Lemma 2.5, (2.45), (2.46), Lemma 2.1, the definition of f_δ, and assuming that $\delta \leq \Delta$ for $\Delta(\gamma, \eta)$ sufficiently small, we have

$$\mathbb{E}[S^2] \lesssim_n L \lesssim_n \sqrt{\frac{4}{4-\eta}} (2^n \mathbb{P}[|\sqrt{n}Z| \leq \varepsilon]^m)^2 = \sqrt{\frac{4}{4-\eta}} \mathbb{E}[S]^2.$$

Setting $\eta = 10^{-5}$, we have by the second moment method (2.4) that

$$\mathbb{P}[S > 0] \geq \frac{\mathbb{E}[S]^2}{\mathbb{E}[S^2]} \gtrsim_n \sqrt{1 - \eta/4} \geq 0.99,$$

completing the proof of Theorem 2.3. □

2.5.3 The REDUCE algorithm

In this appendix we define the **REDUCE** algorithm, a simple procedure for combining a set of points into a single point whose ℓ_∞-norm is not too large. This algorithm **REDUCE** is described explicitly below, and its main property of use is described in Lemma 2.3, whose proof is given below. The analysis of this algorithm uses feasibility as in the classical proof of the Beck-Fiala theorem [Alon and Spencer, 2008].

REDUCE:

Input: $m \times N$ matrix $\mathbf{X}$ with columns $X_1, \ldots, X_N$.

If $N < m$:

Choose $s \in \{\pm 1\}^N$ arbitrarily.

Else:

1. Let $s^{(0)} = \mathbf{0} \in \mathbb{R}^N$, and let $T_0 = \emptyset$.

2. For $k = 0, 1, 2, \ldots$

   If $|T_k| < N - m$

   (a) Find (e.g., using Gaussian elimination) a vector $v \neq \mathbf{0} \in \mathbb{R}^N$ such that $\mathbf{X}v = \mathbf{0}$ and $v_j = 0$ for all $j \in T_k$.

   (b) Define $s^{(k+1)} = s^{(k)} + \lambda v$, where $\lambda > 0$ is the smallest real number such that $|s_j^{(k)} + \lambda v_j| = 1$ for some $j \notin T_k$.

   (c) Define $T_{k+1} = \{j : \left|s^{(k+1)}\right| = 1\}$.

   Else: $s := s^{(k)}$. BREAK

Output: $\sigma := \mathrm{sgn}(s)$

Proof of Lemma 2.3. We suppose that $N > m$, otherwise, an arbitrary choice of signing gives the desired upper bound. Suppose that we are in the k-th iteration of Step 2 of **REDUCE**. If $|T_k| < N - m$, then there are at most $m + |T_k| < N$ linear constraints on the vector $v \in \mathbb{R}^N$ in step 2(a). So by dimension-counting, there exists a nonempty subspace of feasible v. Next if $s^{(k)} \in [-1, 1]^m$, then λ from step 2(b) exists and furthermore $s^{(k+1)} \in [-1, 1]^m$ by the choice of j in step 2(b). Also, we have that $T_k \subset T_{k+1}$; if $|(s^{(k)})_j| = 1$, then the j-th coordinate remains unchanged for future iterations of step 2. Finally, $|T_k|$ increases at least by 1 in each iteration, so the loop in step 2 is guaranteed to terminate after at most $N - m$ iterations.

It remains to verify that σ satisfies the upper bound from Lemma 2.3. Observe that $s \in [-1, 1]^m$, $T := |\{j : |s_j| = 1\}| \geq N - m$, and

$$\sum_{i=1}^{N} s_i X_i = \mathbf{0}.$$

Therefore,

$$\begin{aligned}\left|\sum_{i=1}^{N} \sigma_i X_i\right|_\infty &\leq \left|\sum_{i=1}^{N} s_i X_i\right|_\infty + \left|\sum_{i \notin T} (\mathrm{sgn}(s_i) - s_i) X_i\right|_\infty \\ &\leq \max_{S \subset [N]: |S| = m} \sum_{i \in S} |X_i|_\infty .\end{aligned}$$

□

2.5.4 Proof of Proposition 2.4

We need to show that at each application of resampling in **GKK**, a small number of points are labeled 'bad'. As discussed in the introduction, the restriction on the dimension $m = O(\sqrt{\log n})$ is needed in our analysis to show that the probability of a point being labeled 'bad' is small.

We briefly describe the intuition for this condition by considering the first phase of the algorithm **GKK**. Suppose, for example, that $X_1, \ldots, X_n$ are independent triangularly distributed vectors on $[-1, 1]^m$. In step 1 of **PRDC**, the cube $[-1, 1]^m$ is partitioned into sub-cubes of side length $\alpha' = n^{-\Omega(1/m)}$. Next, we enter the resampling step. We show below that the probability of a point being labeled 'bad' is at most $O(2^m m\alpha') = O(2^m m n^{-\Omega(1/m)})$. Roughly speaking, the reason for this is that there are $2^m(\alpha')^{-m}$ sub-cubes, and the probability of a point in a particular sub-cube being labeled 'bad' is controlled by the product of three terms: 1) the ℓ_1-Lipschitz constant of the density of X_1, which is 1, 2) the ℓ_1-diameter of the sub-cube, which is $m\alpha'$, and 3) the volume of the sub-cube, which is $(\alpha')^m$. Hence, the probability of a point being labeled 'bad' is a small constant, assuming that $m = O(\sqrt{\log n})$.

The next two lemmas present the above argument in full detail.

Lemma 2.9. *Let $\rho : [-\Delta, \Delta] \to \mathbb{R}$ denote a pdf that is L-Lipschitz and bounded above by some constant $D > 0$. Let $g = \rho^{\otimes m} : [-\Delta, \Delta]^m \to \mathbb{R}$ denote the density of the distribution of m independent random variables, each individually distributed according to ρ. Then g is L'-Lipschitz in the ℓ_1 norm:*

$$\forall\, x, y \in [-\Delta, \Delta]^m, \quad |g(x) - g(y)| \leq L' \left|x - y\right|_1,$$

where

$$L' = LD^{m-1}.$$

Proof. Define $x^1 = x$, and for $2 \leq k \leq m$, define

$$x^k = x^{k-1} + \mathbf{e}_k(y_k - x_k),$$

where $\mathbf{e}_k$ denotes the k-th elementary basis vector. Then we have

$$\begin{aligned} |g(x) - g(y)| &\leq \sum_{k=1}^{m} \left|g(x^k) - g(x^{k-1})\right| \left(\prod_{i<k} g(y_i)\right) \left(\prod_{i>k} g(x_i)\right) \\ &\leq \sum_{k=1}^{m} LD^{m-1}|x_k - y_k| \\ &= LD^{m-1} \left|x - y\right|_1. \end{aligned}$$

□

Lemma 2.10. *Let $S = X_1, \ldots, X_s \in [-\Delta, \Delta]^m$ denote a sample of iid random vectors, each having a joint density $g = \rho^{\otimes m}$, where ρ is L-Lipschitz and bounded above by $D > 0$. Let B denote the bad points created in step 2 of* **PRDC** *run on the input $S, v = 0, \alpha = \Delta$, and g. If $m \leq C\sqrt{\log(s)/\max(1, \log \Delta)}$ for a sufficiently small constant $C = C(D, L) > 0$, then*

$$\mathbb{P}[|B| > 0.1s] \leq \exp(-c_1 s),$$

where c_1 is an absolute constant.

Proof. Let $\alpha' = \Delta/\lceil s^{1/(4m)} \rceil$. Let $C_1, \ldots, C_N$ denote the sub-cubes of side length α' formed by partitioning (step 1 of **PRDC**), recalling that $N = (2\Delta)^m(\alpha')^{-m}$. Since $X_1, \ldots, X_s$ are independent, we first study the probability that X_1 is bad and then

apply a Hoeffding bound.

$$\begin{aligned}\mathbb{P}[X_1 \text{ is bad}] &= \sum_{j=1}^{N} \int_{C_j} \left(1 - \frac{\min_{y \in C_j} g(y)}{g(x)}\right) g(x)\, dx \\ &= \sum_{j=1}^{N} \int_{C_j} \left(g(x) - \min_{y \in C_j} g(y)\right) dx \\ &\leq \sum_{j=1}^{N} \mathrm{Vol}(C_j) L D^{m-1} \operatorname{diam}_{\ell_1}(C_j) \\ &= (2\Delta)^m L D^{m-1} m \alpha',\end{aligned}$$

where we measure the diameter in the ℓ_1 norm and applied Lemma 2.9. Since

$$m \leq C\sqrt{\log(s)/\max(1, \log \Delta)},$$

we have

$$p := (2\Delta)^m L D^{m-1} m \alpha' \leq (2\Delta)^m D^{m-1} m \Delta s^{-1/(4m)} \leq 0.05$$

for $C = C(D, L) > 0$ sufficiently small. Since the X_i's are independent, by Hoeffding's inequality,

$$\mathbb{P}[\,|B| \geq 0.1s] \leq \mathbb{P}[\,|B| - ps \geq 0.05s] \leq \exp\left(-\frac{2(0.05)^2 s^2}{s}\right),$$

which completes the proof. □

Proof of Proposition 2.4. The proof is by induction on t. We first handle the base case $t = 1$. By assumption the matrix $\mathbf{X}$ has independent entries, each having a pdf which is L-Lipschitz and bounded above by D. By Lemma 2.10, with probability at least $1 - \exp(-c_1 n)$, there are at most $0.1n$ points labeled 'bad'. Since $m \leq C\sqrt{\log(n)/\max(1, \log \Delta)}$, for C sufficiently small, there are at most $N_1 \leq (2\Delta)^m \alpha_2^{-m} \leq n^{0.6}$ sub-cubes created by partitioning (step 1 of **PRDC**). Thus, at most that many good points are leftover after random differencing in step 3 of **PRDC**. We conclude that with probability at least $1 - \exp(-c_1 n)$, there are at least

$$\frac{n - 0.01n - n^{0.6}}{2} \geq 0.4n \tag{2.47}$$

points in G_1', the set of random differences.

Now we show the inductive step. Let $\mathcal{E}$ denote the event $|S_j| = n_j$ where $n_j \geq (0.3)^{j-1} n$ for all $1 \leq j \leq t$. It suffices to show that

$$\mathbb{P}\left[\,|G_{t+1}'| \leq 0.4 n_t \,\middle|\, \mathcal{E}\right] \leq \exp\left(-c_1\sqrt{n}\right). \tag{2.48}$$

By Proposition 2.3 in Appendix 2.5.7, conditionally on $\mathcal{E}$, the distribution of the

points in $S_t = \mathbf{y}_1, \ldots, \mathbf{y}_{n_t}$ are iid and follow a triangular distribution on $[-\alpha_t, \alpha_t]^m$. Hence, we have by Lemma 2.9 that the density of $\alpha_t^{-1}\mathbf{y}_1, \ldots, \alpha_t^{-1}\mathbf{y}_{n_t}$ is 1-Lipschitz with respect to ℓ_1 and is bounded above by $D = 1$. Note that, by an application of the chain rule, the probability $\alpha_t^{-1}\mathbf{y}_j$ is labeled 'good' using the triangular density on $[-1, 1]^m$ for g in step 2 of **PRDC** is the same as the probability that $\mathbf{y}_j$ is labeled 'good' using the triangular density on $[-\alpha_t, \alpha_t]^m$ for g in step 2 of **PRDC**.

Since $t \le \lceil C^* \log n \rceil$ and $n_j \ge (0.3)^{j-1}n$ for $1 \le j \le t$, we have that $n_t \ge \sqrt{n}$. In particular, for $C > 0$ sufficiently small, $s = \sqrt{n}$ satisfies the required lower bound of Lemma 2.10. Therefore,

$$\mathbb{P}\left[|B_{t+1}| \ge 0.1 n_t \,\middle|\, \mathcal{E} \right] \le \exp(-c_1 n_t) \le \exp(-c_1\sqrt{n}).$$

For C sufficiently small and $m \le C\sqrt{\log n}$, there are at most $N_t \le 2^m n_t^{1/4} \le n_t^{0.6}$ subcubes formed in step 1 of **PRDC**. Hence, at most $n_t^{0.6}$ good points are leftover after the random differencing step of **PRDC**. Halving the number of remaining points as in (2.47) of the base case, we conclude that (2.48) holds with the desired probability in phase t. □

2.5.5 Proof of Proposition 2.5

The goal of this subsection is to prove Proposition 2.5. The next technical lemma implies that a negligible fraction of points are lost in step 4(b), the clean-up step of **PRDC**.

Lemma 2.11. *Let $\alpha = \lceil s^{1/(4m)} \rceil^{-1}$, and let $\mathcal{U} = \mathbf{u}_1, \ldots, \mathbf{u}_s \overset{iid}{\sim} \mathrm{Tri}[-\alpha, \alpha]^m$ denote a sample from a triangular distribution. Let $v^{(0)} \in \mathbb{R}^m$ denote a random vector independent of $\mathcal{U}$ satisfying $\left|v^{(0)}\right|_2 \le Rm^{3/2}$ for some absolute constant $R > 0$. For $k = 1, 2, \ldots,$ define a sequence of random vectors*

$$v^{(k)} = v^{(k-1)} + a^* \mathbf{u}_k$$

where

$$a^* = \operatorname*{argmin}_{a \in \{\pm 1\}} \left| v^{(k-1)} + a\mathbf{u}_k \right|_2.$$

Let c^* denote the absolute constant from Claim 2.1. Suppose that $R' \ge 2/c^*$ and

$$K \ge \frac{8R^2 m^2 \sqrt{s}}{R' c^*}.$$

Then with probability at least

$$1 - \exp\left(-\frac{(c^*)^2 K}{8m} \right)$$

there exists $k \le K$ such that

$$|v^{(k)}|_2 \le R' m \alpha.$$

Proof. By the definition of $v^{(k)}$, we have that

$$0 \leq \left|v^{(K+1)}\right|_2^2 = \left|v^{(0)}\right|_2^2 + \sum_{k=0}^{K} \left(-2\left|\langle v^{(k)}, \mathbf{u}_{k+1}\rangle\right| + |\mathbf{u}_{k+1}|_2^2\right).$$

Consider the event $\mathcal{E}$ that for all $1 \leq k \leq K$, we have $\left|v^{(k)}\right|_2 \geq R'm\alpha$. Let $\nu^{(k)} = v^{(k)}/\left|v^{(k)}\right|_2$. Observe that $|\mathbf{u}_k|_2^2 \leq \alpha^2 m$. Applying this and rearranging the inequality above, we have that the event $\mathcal{E}$ implies

$$\sum_{k=0}^{K} \left|\langle \nu^{(k)}, \mathbf{u}_{k+1}\rangle\right| \leq \frac{R^2m^3 + \alpha^2 mK}{2R'm\alpha}. \tag{2.49}$$

For $0 \leq j \leq K$, define a sequence of random variables

$$M_j := \sum_{k=0}^{j} \left(\left|\langle \nu^{(k)}, \mathbf{u}_{k+1}\rangle\right| - c^*\alpha\right).$$

For convenience, we also define $M_{-1} \equiv 0$. Note that M_j is measurable with respect to the sigma-field Ω_j generated by the random variables $v^{(0)}, v^{(1)}, \ldots, v^{(j+1)}$. Therefore, $\Omega_{-1} \subset \Omega_0 \subset \ldots$ defines a filtration for the sequence of random variables $\{M_j\}_{j\geq -1}$.

Claim 2.1. *There exists an absolute constant $c^* > 0$ such that $\{M_j\}_{j\geq -1}$ is a submartingale with respect to the filtration $\{\Omega_j\}_{j\geq -1}$.*

Proof. Since $v^{(0)}$ is independent of $\mathcal{U}$ and $\mathcal{U}$ is an independent sample, it follows that $\mathbf{u}_{k+1}$ is independent of $\nu^{(k)}$. Observe that the coordinates of $\mathbf{u}_{k+1}$ are subGaussian. By the Khintchine inequality for the ℓ_1 norm [see Exercises 2.6.5 and 2.6.6 of Vershynin, 2018], we have

$$\mathbb{E}\left[\left|\langle \nu^{(k)}, \mathbf{u}_{k+1}\rangle\right| \;\middle|\; v^{(k)}\right] = \mathbb{E}\left[\left|\langle \nu^{(k)}, \mathbf{u}_{k+1}\rangle\right| \;\middle|\; \nu^{(k)}\right] \geq \alpha c^* \left|\nu^{(k)}\right|_2 = \alpha c^* > 0$$

for an absolute constant $c^* > 0$. □

Let $c^* > 0$ denote the absolute constant from Claim 2.1, and set $R' \geq 2/c^*$. Next, note the equivalence between the following inequalities:

$$\begin{aligned} c^*\alpha K &\geq \frac{c^*\alpha K}{2} + \frac{R^2m^3 + \alpha^2 mK}{2R'm\alpha} \Leftrightarrow \\ K &\geq \frac{R^2m^2}{R'(c^* - 1/R')}\alpha^{-2}, \end{aligned} \tag{2.50}$$

assuming that $c^* - 1/R' > 0$. Setting $R' \geq 2/c^*$, it follows that if

$$K \geq \frac{8R^2m^2\sqrt{s}}{R'c^*},$$

then (2.50) holds. Next, note by Cauchy-Schwarz that the submartingale M_j has increments bounded by $\alpha\sqrt{m}$. Since (2.50) holds, we may apply the Hoeffding–Azuma inequality to conclude that for such choice of K and R' that

$$\mathbb{P}[\mathcal{E}] \le \mathbb{P}\left[M_K \le \frac{R^2 + \alpha^2 m^2 K}{2R'm\alpha} - c^*\alpha K\right] \le \mathbb{P}\left[M_K \le -\frac{c^*\alpha K}{2}\right] \le \exp\left(-\frac{(c^*)^2 K}{8m}\right),$$

as desired.

□

Proof of Proposition 2.5. Let $t \ge 1$ denote the current phase. Let $\mathcal{E}$ denote the event that $|S_j| = n_j$ for all $1 \le j \le t$ and $|G'_t| = g'_t$ where $n_j \ge (0.3)^{j-1}n$ for all $1 \le j \le t$ and $g'_t \ge (0.4)n_t$. By Proposition 2.3 and Lemma 2.16 in Appendix 2.5.7, conditionally on $\mathcal{E}$, the points $\mathbf{z}_1, \ldots, \mathbf{z}_{g'_t} \in G'_t$ are distributed as $\mathrm{Tri}[-\alpha_{t+1}, \alpha_{t+1}]^m$, and the leftover vector $v_t^{(0)}$ obtained in step 4(a) of **PRDC** is independent of this sample. Moreover, by Lemma 2.3 and the fact that $|v_t|_\infty \le |v_t|_2 \le \gamma m \alpha_t$, it follows that

$$\left|v_t^{(0)}\right|_\infty \le (\gamma + 1)m\alpha_t.$$

Hence, the Cauchy–Schwarz inequality yields that

$$\left|v_t^{(0)}\right|_2 \le (\gamma + 1)m^{3/2}\alpha_t.$$

Next, apply Lemma 2.11 with $\mathcal{U} = \frac{1}{\alpha_t}\mathbf{z}_1, \ldots, \frac{1}{\alpha_t}\mathbf{z}_{g'_t}$, $v^{(0)} = \frac{1}{\alpha_t}v_t^{(0)}$, $R = \gamma + 1$, $R' = \gamma$, and $K = (g'_t)^{3/4}$ where $\gamma \ge 2/c^*$. Recall that by assumption $g'_t \ge (0.4)n_t \ge (0.4)(0.3)^{t-1}n$. Since $t \le \lceil C^* \log n \rceil$, we have that $g'_t \ge \sqrt{n}$. So for C sufficiently small in the bound $m \le C\sqrt{\log n}$, we have that the lower bound

$$K = (g'_t)^{3/4} \ge \frac{8(\gamma+1)^2 m^2 \sqrt{g'_t}}{\gamma c^*}$$

holds, and so indeed Lemma 2.11 applies. Therefore, conditioned on $\mathcal{E}$, with probability at least

$$1 - \exp\left(-\frac{(c^*)^2 (g'_t)^{3/4}}{8m}\right) \ge 1 - \exp\left(-(c^*)^2 n^{1/4}\right)$$

there exists $k \le K = (g'_t)^{3/4}$ with

$$\left|v_t^{(k)}\right|_2 \le \gamma m \alpha_{t+1}.$$

By the lower bounds $n \ge e^{(1/C)m^2}$ and $g'_t \ge \sqrt{n}$, for C sufficiently small, it follows that $(g'_t)^{3/4} \le (0.01)g'_t$. Hence, conditioned on $\mathcal{E}$, with probability at least $1 - \exp\left(-(c^*)^2 n^{1/4}\right)$ we have $|S_{t+1}| \ge g'_t - (g'_t)^{3/4} \ge (0.3)n_t$, as desired. □

2.5.6 Proof of Theorem 2.2

Our main theorem is a direct consequence of Propositions 2.4 and 2.5.

Proof of Theorem 2.2. Recall that $T = \lceil C^* \log n \rceil$ where $C^* = (2\log(10/3))^{-1}$, and set $\theta = 0.3$. By the union bound over the T phases of **PRDC** in **GKK**, induction, and Propositions 2.4 and 2.5, we have that $|S_t| \geq \theta^{t-1} n$ for all $1 \leq t \leq T$ with probability at least $1 - \exp(-c_3 n^{1/4})$, for some absolute constant $c_3 > 0$. Since $\alpha_{t+1} = \alpha_t / \lceil |S_t|^{1/(4m)} \rceil$, this implies by induction that

$$\alpha_T \leq \max(1, \Delta) \theta^{-T^2/(4m)} n^{-T/(4m)} \leq \max(1, \Delta) \exp\left(-\frac{C^* \log^2 n}{8m} \right)$$

with probability at least $1 - \exp(-c_3 n^{1/4})$.

Moreover, by the stopping criterion from step 4(b) of **PRDC**, $|v_T|_\infty \leq |v_T|_2 \leq \gamma m \alpha_T$. Applying **REDUCE** to $S_T \cup \{v_T\}$, we see by Lemma 2.3 that the output $|v|_\infty$ of **GKK** satisfies

$$|v|_\infty \leq \max(1, \Delta)(\gamma m + m - 1) \exp\left(-\frac{C^* \log^2 n}{8m} \right) \leq \exp\left(-\frac{c \log^2 n}{m} \right)$$

for an absolute constant $c > 0$. Note that the right-hand-side follows if we take $C > 0$ sufficiently small in the bound $m \leq C\sqrt{\log(n)/\max(1, \log \Delta)}$. □

2.5.7 Distributional properties

Our analysis of **GKK** relies heavily on the fact that the operations in the algorithm preserve important features of the original distribution such as independence. Though not carefully proven in Karmarkar and Karp [1982], these features are crucial to our analysis, so we provide explicit justification of these properties below for completeness.

First we introduce some notation. Given $\alpha > 0$, a fixed collection of vectors $\mathbf{z}_1, \ldots, \mathbf{z}_s \subset [-\alpha, \alpha]^m$, and a density $g : [-\alpha, \alpha]^m$, divide the cube $[-\alpha, \alpha]^m$ into $N := 2^m (\lceil s^{1/(4m)} \rceil)^m$ sub-cubes $C_1, \ldots, C_N$ of side length $\alpha / \lceil s^{1/(4m)} \rceil$ as in step 1 of **PRDC**. Label the points $\mathbf{z}_1, \ldots, \mathbf{z}_s$ as in step (2) of **PRDC** using the density g. Define a random collection of ordered pairs $\mathcal{T}_{s,\alpha,g} \subset ([N] \times \{0, 1\})^s$ so that for $1 \leq i \leq s$,

$$(\mathcal{T}_{s,\alpha,g})_i = (j, 1)$$

if and only if $\mathbf{z}_i \in C_j$ and if $\mathbf{z}_i$ is labeled 'good', and

$$(\mathcal{T}_{s,\alpha,g})_i = (j, 0)$$

if and only if $\mathbf{z}_i \in C_j$ and $\mathbf{z}_i$ is labeled as 'bad'.

Usually s, α and g are clear from context, in which case we write $\mathcal{T}$ for $\mathcal{T}_{s,\alpha,g}$. Observe that $\mathcal{T}$ keeps track of which sub-cube v_i lands in and also whether it was labeled good or bad. We refer to $\mathcal{T}$ as the *configuration vector* corresponding to the input of **PRDC**.

We proceed by proving some preliminary lemmas, the first of which states roughly that given random vectors $\mathbf{z}_1, \ldots, \mathbf{z}_s$ with a nice conditional distribution, the good points in each sub-cube C_j have a uniform distribution.

Lemma 2.12. *Suppose that conditioned on an event $\mathcal{F}$,*

- *the random vectors $S = \mathbf{z}_1, \ldots, \mathbf{z}_s \in \mathbb{R}^m$ are iid, and each vector has the conditional joint density $g : [-\Delta, \Delta]^m \to \mathbb{R}$.*
- *$S \cup \{v\}$ is a collection of independent random vectors.*

Run the first two steps of **PRDC** *with input $S = \mathbf{z}_1, \ldots, \mathbf{z}_s$,$v$, $\alpha = \Delta$, and density g. Let G denote the good points, and let B denote the bad points. Then conditioned on $\mathcal{T}_{s,\Delta,g}$ and $\mathcal{F}$,*

- *the random vectors in $B \cup G$ are mutually independent.*
- *For $1 \le j \le N$, a given good point in C_j has a uniform distribution on C_j.*

Proof. The first statement follows because (1) $G \cup B = \mathbf{z}_1, \ldots, \mathbf{z}_s$ is an independent sample, conditioned on $\mathcal{F}$, and (2) the ordered pair $(\mathcal{T}_{s,\Delta,g})_i$ is generated independently for each $i \in [s]$. Thus it suffices to show, by symmetry and passing to conditional densities, that

$$g(z|\mathbf{z}_1 \in C_j, \mathbf{z}_1 \text{ good}) = \frac{1}{\mathrm{Vol}(C_j)}$$

for all $z \in C_j$. By Bayes' rule,

$$\begin{aligned} g(z|\mathbf{z}_1 \in C_j, \mathbf{z}_1 \text{ good}) &= \frac{\mathbb{P}[\mathbf{z}_1 \text{ good}|\mathbf{z}_1 = z, \mathbf{z}_1 \in C_j, \mathcal{F}]\, g(z|\mathbf{z}_1 \in C_j)}{\mathbb{P}[\mathbf{z}_1 \text{ good}|\mathbf{z}_1 \in C_j, \mathcal{F}]} \\ &= \left(\frac{\min_{x \in C_j} g(x)}{g(z)} \cdot \frac{g(z)}{\mathbb{P}[\mathbf{z}_1 \in C_j \,|\, \mathcal{F}]}\right) \Bigg/ \left(\frac{\mathrm{Vol}(C_j) \min_{x \in C_j} g(x)}{\mathbb{P}[\mathbf{z}_1 \in C_j | \mathcal{F}]}\right) \\ &= \frac{1}{\mathrm{Vol}(C_j)}, \end{aligned}$$

where the last line follows because

$$\mathbb{P}[\mathbf{z}_1 \text{ good}, \mathbf{z}_1 \in C_j | \mathcal{F}] = \int_{C_j} \mathbb{P}[\mathbf{z}_1 \text{ good}|\mathbf{z}_1 = z, \mathcal{F}] g(z)\, dz = \mathrm{Vol}(C_j) \min_{x \in C_j} g(x).$$

□

Lemma 2.13. *Consider the set-up of Lemma 2.12, and let $\alpha' = \alpha/\lceil s^{1/(4m)} \rceil$. Let G' denote the set of random differences constructed after step 3. of* **PRDC** *applied to S, v, $\alpha = \Delta$, and g. Then conditioned on the events $\mathcal{F}$ and $\mathcal{T} = \mathbf{T}$, the points in G' are iid and have a triangular distribution on $[-\alpha', \alpha']^m$.*

Proof. Observe that $\mathbf{T}$ determines the number of points in G'. The points in G' are independent by Lemma 2.12 and the fact that the points in G are randomly differenced in step 3. of **PRDC**. Since C_j is a translation of the sub-cube $[-\alpha', \alpha']^m$, the difference of two independent, uniformly sampled points from C_j have a triangular distribution on $[-\alpha', \alpha']^m$. □

Lemma 2.14. *Consider the set-up of Lemma 2.13, and let $\ell \in \mathbb{Z}_{\geq 0}$. Let the random variable $\mathcal{L}$ denote the number of points removed from G' in step 4(b) of* **PRDC** *applied to S, v, $\alpha = \Delta$, and g. Let S' and v' denote the vectors output by* **PRDC**. *Let $g' = |G'|$. Then conditioned on the events $\mathcal{F}$, $\mathcal{T} = \mathbf{T}$, and $\mathcal{L} = \ell$,*

- *The $g' - \ell$ points in S' are iid and follow a triangular distribution on $[-\alpha', \alpha']^m$.*
- *The random vector v' is independent of the vectors in S'.*

Proof. Recall that $|G'| = g'$ is determined by $\mathbf{T}$. Label the points in G' independently at random to be $G' = \mathbf{y}_1, \ldots, \mathbf{y}_{g'}$. The points in G' are independent and triangularly distributed on $[-\alpha', \alpha']^m$ by Lemma 2.13, conditionally on $\mathcal{F}$ and $\mathcal{T} = \mathbf{T}$. Recall the single vector v that was input initially to **PRDC**. In step 4(a), this is combined with vectors in B' to construct a single vector $v^{(0)}$. By Lemma 2.12, we have that $v^{(0)}$ is independent of G', conditionally on $\mathcal{T} = \mathbf{T}$ and $\mathcal{F}$.

Now in step 4(b) of **PRDC**, let us remove points from G' in the order $\mathbf{y}_{g'}, \mathbf{y}_{g'-1}, \ldots, \mathbf{y}_{g'-\ell+1}$. By the stopping criterion for step 4(b), we have

$$\{\mathcal{L} = \ell\} = \left\{ \left|v^{(k)}\right|_2 > \gamma m \alpha' \ \forall\, 1 \leq k \leq \ell - 1,\ \left|v^{(\ell)}\right|_2 < \gamma m \alpha' \right\}.$$

Since $v^{(k)} = v^{(k-1)} \pm \mathbf{y}_{g'-k+1}$ for $1 \leq k \leq \ell$, the random vector $v^{(k)}$ is independent of $\mathbf{y}_1, \ldots, \mathbf{y}_{g'-\ell}$. Therefore, the sample $S' = \mathbf{y}_1, \ldots, \mathbf{y}_{g'-\ell}$ is independent of the event $\mathcal{L} = \ell$. Hence, further conditioning on $\mathcal{L} = \ell$ does not affect the distribution of S', as desired. □

Summarizing the content of Lemmas 2.12, 2.13, and 2.14, we have the following proposition.

Proposition 2.6. *Suppose that conditioned on an event $\mathcal{F}$,*

- *the random vectors $S = \mathbf{z}_1, \ldots, \mathbf{z}_s \in \mathbb{R}^m$ are iid, and each vector has the conditional joint density $g : [-\Delta, \Delta]^m \to \mathbb{R}$.*
- *$S \cup \{v\}$ is a collection of independent random vectors.*

Let S', v' denote the vectors output by **PRDC** *applied to S, v, $\alpha = \Delta$, and g. Let $s' \in \mathbb{Z}_{\geq 0}$ and $\alpha' = \alpha / \lceil s^{1/(4m)} \rceil$. Then conditioned on $\mathcal{F}, \mathcal{T} = \mathbf{T}$, and $|S'| = s'$,*

- *the s' points in S' are iid and follow a triangular distribution on $[-\alpha', \alpha']^m$.*
- *The random vector v' is independent of the vectors in S'.*

Observe that Proposition 2.6 and induction imply the next lemma, which guarantees that we have a nice distribution after every phase of **PRDC**, conditionally on the data $\mathcal{T}^{(j)}$ at each step.

Lemma 2.15. *Let $X_1, \ldots, X_n$ be iid random vectors, each having a joint density $g : [-\Delta, \Delta]^m \to \mathbb{R}$, conditioned on some event $\mathcal{F}$. Consider the output S_t, v_t, α_t that results after the $(t-1)$-th phase of* **PRDC** *in step 2 of* **GKK**. *For $1 \leq j \leq t-1$, let*

$\mathcal{T}^{(j)}$ denote the configuration vector resulting from step 2 of the j-th phase of **PRDC**. *Then conditioned on $\mathcal{T}^{(j)} = \mathbf{T}^{(j)}$ for $1 \le j \le t-1$ and $|S_j| = n_j$ for $1 \le j \le t$, we have*

- *the n_t points in S_t are iid and follow a triangular distribution on $[-\alpha_t, \alpha_t]^m$.*
- *The random vector v_t is independent of the vectors in S_t.*

Next, marginalizing over all possible configuration vectors yields Proposition 2.3.

Proof of Proposition 2.3. We induct on the phase t. Consider the base case $t = 2$. Let $\mathbf{z}_1, \dots, \mathbf{z}_{n_2}$ denote the vectors in S_2, and let I_i denote a measurable subset of $[-\alpha_2, \alpha_2]^m$ for $1 \le i \le n_2$. Recall that $\mathbf{T}^{(1)}$ determines the number of differences in G_1', and $|S_2|$ determines the amount of points lost in step 4(b) of **PRDC**. Then we have, marginalizing over all possible choices of $\mathbf{T}^{(1)}$ compatible with $|S_2| = n_2$,

$$\begin{aligned}&\mathbb{P}\left[\mathbf{z}_i \in I_i \,\forall\, 1 \le i \le n_2 \,\middle|\, |S_2| = n_2\right] \\ &\quad = \sum_{\mathbf{T}^{(1)}} \mathbb{P}\left[\mathbf{z}_i \in I_i \,\forall\, 1 \le i \le n_2 \,\middle|\, \mathcal{T}^{(1)} = \mathbf{T}^{(1)},\ |S_2| = n_2\right] \mathbb{P}\left[\mathcal{T}^{(1)} = \mathbf{T}^{(1)} \,\middle|\, |S_2| = n_2\right]\end{aligned}$$

By Lemma 2.15,

$$\mathbb{P}\left[\mathbf{z}_i \in I_i \,\forall\, 1 \le i \le n_2 \,\middle|\, \mathcal{T}^{(1)} = \mathbf{T}^{(1)},\ |S_2| = n_2\right] = \mathbb{P}\left[\mathbf{u}_i \in I_i \,\forall\, 1 \le i \le n_2\right]$$

where $\mathbf{u}_1, \dots, \mathbf{u}_{n_2} \overset{iid}{\sim} \mathrm{Tri}[-\alpha_2, \alpha_2]^m$. Hence,

$$\mathbb{P}\left[\mathbf{z}_i \in I_i \,\forall\, 1 \le i \le n_2 \,\middle|\, |S_2| = n_2\right] = \mathbb{P}[\mathbf{u}_i \in I_i \,\forall\, 1 \le i \le n_2],$$

which confirms the first bullet point of Proposition 2.3 for the base case $t = 2$. Following a similar marginalization procedure, this also implies by Lemma 2.15 that v_2, the single vector output by **PRDC**, is independent of S_2 conditionally on $|S_2|$.

Now we handle the inductive step. Let $S_t = \mathbf{y}_1, \dots, \mathbf{y}_{n_t}$ and v_t denote the vectors output by the $(t-1)^{\text{th}}$ phase of **PRDC**. Suppose that conditionally on $\mathcal{F} := \{|S_2| = n_2, \dots, |S_t| = n_t\}$ that S_t is an iid sample of triangularly distributed vectors on $[-\alpha_t, \alpha_t]^m$, and v_t is independent of S_t. By Proposition 2.6, conditionally on $\mathcal{F}$, $|S_{t+1}| = n_{t+1}$, and the configuration vector $\mathcal{T}^{(t)} = \mathbf{T}^{(t)}$, the sample S_{t+1} is an iid collection of triangularly distributed vectors on $[-\alpha_{t+1}, \alpha_{t+1}]^m$. Hence, conditioning on $\mathcal{F} \cup \{|S_{t+1}| = n_{t+1}\}$ and applying the same marginalization over the configuration vector $\mathbf{T}^{(t)}$ as in the base case yields the first bullet point of Proposition 2.3 for the inductive step. The second bullet point follows similarly. □

The next lemma is used in Appendix 2.5.5. We omit its proof because it is similar to that of Proposition 2.3.

Lemma 2.16. *Let $X_1, \ldots, X_n$ be iid random vectors, each having a joint density $g : [-\Delta, \Delta]^m \to \mathbb{R}$. Apply* **GKK** *to the matrix* $\mathbf{X}$ *with columns $X_1, \ldots, X_n$, and consider the good points G'_t created from random differencing in step 3 of the t^{th} phase of* **PRDC**. *Also consider the random vector $v_t^{(0)}$ formed in step 4(a) of* **PRDC**. *Then conditioned on $|S_j| = n_j$ for $1 \leq j \leq t$ and $|G'_t| = g'_t$,*

- *the random vectors in G'_t form an independent sample of size g'_t from* $\text{Tri}[-\alpha_{t+1}, \alpha_{t+1}]^m$.
- *The random vector $v_t^{(0)}$ is independent of the vectors in G'_t.*

Chapter 3

Coreset density estimation

3.1 Introduction

The ever-growing size of datasets that are routinely collected has led practitioners across many fields to contemplate effective data summarization techniques that aim at reducing the size of the data while preserving the information that it contains. While there are many ways to achieve this goal, including standard data compression algorithms, they often prevent direct manipulation of data for learning purposes. *Coresets* have emerged as a flexible and efficient set of techniques that permit direct data manipulation. Coresets are well-studied in machine learning [Har-Peled and Kushal, 2007, Feldman et al., 2013, Bachem et al., 2017, 2018, Karnin and Liberty, 2019], statistics [Feldman et al., 2011, Zheng and Phillips, 2017, Munteanu et al., 2018, Huggins et al., 2016, Phillips and Tai, 2018a,b], and computational geometry [Agarwal et al., 2005, Clarkson, 2010, Frahling and Sohler, 2005, Gärtner and Jaggi, 2009, Claici et al., 2020].

Given a dataset $\mathcal{D} = \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ and task (density estimation, logistic regression, etc.) a coreset $\mathcal{C}$ is given by $\mathcal{C} = \{X_i : i \in S\}$ for some subset S of $\{1, \ldots, n\}$ of size $|S| \ll n$. A good coreset should suffice to perform the task at hand with the same accuracy as with the whole dataset $\mathcal{D}$.

In this work we study the canonical task of density estimation. Given i.i.d random variables $X_1, \ldots, X_n \sim \mathbb{P}_f$ that admit a common density f with respect to the Lebesgue measure over $\mathbb{R}^d$, the goal of density estimation is to estimate f. It is well known that the minimax rate of estimation over the L-Hölder smooth densities $\mathcal{P}_{\mathcal{H}}(\beta, L)$ of order β is given by

$$\inf_{\hat{f}} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f \|\hat{f} - f\|_2 = \Theta_{\beta,d,L}(n^{-\frac{\beta}{2\beta+d}}), \tag{3.1}$$

where the infimum is taken over all estimators based on the dataset $\mathcal{D}$. Moreover the minimax rate above is achieved by a kernel density estimator

$$\hat{f}_n(x) := \frac{1}{nh^d} \sum_{j=1}^{n} k\left(\frac{X_i - x}{h}\right) \tag{3.2}$$

for suitable choices of kernel $k : \mathbb{R}^d \to \mathbb{R}$ and bandwidth $h > 0$ [see e.g. Tsybakov, 2009, Theorem 1.2].

The main goal of this paper is to extend this understanding of rates for density estimation to estimators based on coresets. Specifically we would like to characterize the statistical performance of coresets in terms of their cardinality. To do so, we investigate two families of estimators built on coresets: one that is quite flexible and allows arbitrary estimators to be used on the coreset and another that is more structured and driven by practical considerations; it consists of weighted kernel density estimators built on coresets.

3.1.1 Two statistical frameworks for coreset density estimation

We formally define a coreset as follows. Throughout this work $m = o(n)$ denotes the cardinality of the coreset. Given $x \in \mathbb{R}^{d \times n}$, let $S = S(y|x)$ denote a conditional probability measure on the set $\binom{[n]}{m}$ of subsets of $[n] = \{1, 2, \ldots, n\}$ of cardinality m. In information theoretic language, S is a channel from $\mathbb{R}^{d \times n}$ to subsets of cardinality m. We refer to the channel S as a *coreset scheme* because it designates a data-driven method of choosing a subset of data points. In what follows, we abuse notation and let $S = S(x)$ denote an instantiation of a sample from the measure $S(y|x)$ for $x \in \mathbb{R}^{d \times n}$. A *coreset* X_S is then defined to be the projection of the dataset $X = (X_1, \ldots, X_n)$ onto the subset indicated by $S(X)$: $X_S := \{X_i\}_{i \in S(X)}$.

The first family of estimators that we investigate is quite general and allows the statistician to select a coreset and then employ an estimator that only manipulates data points in the coreset to estimate an unknown density. To study coresets, it is convenient to make the dependence of estimators on observations more explicit than in the traditional literature. More specifically, a density estimator $\hat{f}$ based on n observations $X_1, \ldots, X_n \in \mathbb{R}^d$ is a function $\hat{f} : \mathbb{R}^{d \times n} \to L^2(\mathbb{R}^d)$ denoted by $\hat{f}[X_1, \ldots, X_n](\cdot)$. Similarly, a *coreset-based estimator* $\hat{f}_S$ is constructed from a coreset scheme S of size m and an estimator (measurable function) $\hat{f} : \mathbb{R}^{d \times m} \to L^2(\mathbb{R}^d)$ on m observations. We enforce the additional restriction on $\hat{f}$ that for all $y_1, \ldots, y_m \in \mathbb{R}^d$ and for all bijections $\pi : [m] \to [m]$, it holds that $\hat{f}[y_1, \ldots, y_m](\cdot) = \hat{f}[y_{\pi(1)}, \ldots, y_{\pi(m)}](\cdot)$. Given S and $\hat{f}$ as above, we define the *coreset-based estimator* $\hat{f}_S : \mathbb{R}^{d \times n} \to L^2(\mathbb{R}^d)$ to be the function $\hat{f}_S[X](\cdot) := \hat{f}[X_S](\cdot) : \mathbb{R}^d \to \mathbb{R}$. We evaluate the performance of coreset-based estimators in Section 3.2 by characterizing their rate of estimation over Hölder classes.[1]

The symmetry restriction on $\hat{f}$ prevents the user from exploiting information about the ordering of data points to their advantage: the only information that can be used by the estimator $\hat{f}$ is contained in the unordered collection of distinct vectors given by the coreset X_S.

[1] Our notion of coreset-based estimators bares conceptual similarity to various notions of *compression schemes* as studied in the literature, e.g. Littlestone and Warmuth [1986], Ashtiani et al. [2020], Hanneke et al. [2019].

As evident from the results in Section 3.2, the information-theoretically optimal coreset estimator does not resemble coreset estimators employed in practice. To remedy this limitation, we also study *weighted coreset kernel density estimators* (KDEs) in Section 3.3. Here the statistician selects a kernel k, bandwidth parameter h, and a coreset X_S of cardinality m as defined above and then employs the estimator

$$\hat{f}_S(y) = \sum_{j \in S} \lambda_j h^{-d} k\left(\frac{X_j - y}{h}\right),$$

where the weights $\{\lambda_j\}_{j \in S}$ are nonnegative, sum to one and are allowed to depend on the full dataset.

In the case of uniform weights where $\lambda_j = \frac{1}{m}$ for all $j \in S$, coreset KDEs are well-studied [see e.g. Bach et al., 2012, Harvey and Samadi, 2014, Phillips and Tai, 2018a,b, Karnin and Liberty, 2019]. Interestingly, our results show that allowing flexibility in the weights gives a definitive advantage for the task of density estimation. By Theorems 3.2 and 3.5, the uniformly weighted coreset KDEs require a much larger coreset than that of weighted coreset KDEs to attain the minimax rate of estimation over univariate Lipschitz densities.

3.1.2 Setup and Notation

We reserve the notation $\|\cdot\|_2$ for the L^2 norm and $|\cdot|_p$ for the ℓ^p-norm. The constants $c, c_{\beta,d}, c_L$, etc. vary from line to line and the subscripts indicate parameter dependences.

Fix an integer $d \geq 1$. For any multi-index $s = (s_1, \ldots, s_d) \in \mathbb{Z}^d_{\geq 0}$ and $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, define $s! = s_1! \cdots s_d!$, $x^s = x_1^{s_1} \cdots x_d^{s_d}$ and let D^s denote the differential operator defined by

$$D^s = \frac{\partial^{|s|_1}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}.$$

We reserve the notation $|s|$ for the coordinate-wise application of $|\cdot|$ to the multi-index s.

Fix a positive real number β, and let $\lfloor \beta \rfloor$ denote the maximal integer *strictly* less than β. Given $L > 0$ we let $\mathcal{H}(\beta, L)$ denote the space of Hölder functions $f : \mathbb{R}^d \to \mathbb{R}$ that are supported on the cube $[-1/2, 1/2]^d$, are $\lfloor \beta \rfloor$ times differentiable, and satisfy

$$|D^s f(x) - D^s f(y)| \leq L\, |x - y|_2^{\beta - \lfloor \beta \rfloor},$$

for all $x, y \in \mathbb{R}^d$ and for all multi-indices s such that $|s|_1 = \lfloor \beta \rfloor$.

Let $\mathcal{P}_{\mathcal{H}}(\beta, L)$ denote the set of probability density functions contained in $\mathcal{H}(\beta, L)$. For $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$, let $\mathbb{P}_f$ (resp. $\mathbb{E}_f$) denote the probability distribution (resp. expectation) associated to f.

For $d \geq 1$ and $\gamma \in \mathbb{Z}_{\geq 0}$, we also define the Sobolev functions $\mathcal{S}(\gamma, L')$ that consist of all $f : \mathbb{R}^d \to \mathbb{R}$ that are γ times differentiable and satisfy

$$\|D^\alpha f\|_2 \leq L'$$

for all multi-indices α such that $|\alpha|_1 = \gamma$.

Given $f \in L^2$, we define the Fourier transform $\mathcal{F}[f] : \mathbb{R}^d \to \mathbb{R}$ by

$$\mathcal{F}[f](\omega) = \int_{\mathbb{R}^d} f(x) e^{-i\langle x,\omega\rangle} \mathrm{d}x.$$

3.2 Coreset-based estimators

In this section we study the performance of coreset-based estimators. Recall that coreset-based estimators are estimators that only depend on the data points in the coreset.

Define the *minimax risk for coreset-based estimators* $\psi_{n,m}(\beta, L)$ over $\mathcal{P}_{\mathcal{H}}(\beta, L)$ to be

$$\psi_{n,m}(\beta, L) = \inf_{\hat{f}, |S|=m} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f \, \|\hat{f}_S - f\|_2, \tag{3.3}$$

where the infimum above is over all choices of coreset scheme S of cardinality m and all estimators $\hat{f} : \mathbb{R}^{d\times m} \to L^2(\mathbb{R}^d)$.

Our main result on coreset-based estimators characterizes their minimax risk.

Theorem 3.1. *Fix $\beta, L > 0$ and an integer $d \geq 1$. Assume that $m = o(n)$. Then the minimax risk of coreset-based estimators satisfies*

$$\inf_{\hat{f}, |S|=m} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f \, \|\hat{f}_S - f\|_2 = \Theta_{\beta,d,L}(n^{-\frac{\beta}{2\beta+d}} + (m \log n)^{-\frac{\beta}{d}}).$$

The above theorem readily yields a characterization of the minimal size $m^*(\beta, d)$ that a coreset can have while still enjoying the minimax optimal rate $n^{-\frac{\beta}{2\beta+d}}$ from (3.1). More specifically, let $m^* = m^*(n)$ be such that

(i) if $m(n)$ is a sequence such that $m = o(m^*)$, then $\liminf_{n\to\infty} n^{\frac{\beta}{2\beta+d}} \psi_{n,m}(\beta, L) = \infty$, and

(ii) if $m = \Omega(m^*)$ then $\limsup_{n\to\infty} \psi_{n,m}(\beta, L) n^{\frac{\beta}{2\beta+d}} \leq C_{\beta,d,L}$ for some constant $C_{\beta,d,L} > 0$.

Then it follows readily from from Theorem 3.1 that $m^* = \Theta_{\beta,d,L}(n^{\frac{d}{2\beta+d}} / \log n)$.

Theorem 3.1 illustrates two different curses of dimensionality: the first stems from the original estimation problem, and the second stems from the compression problem. As $d \to \infty$, it holds that $m^* \sim n/\log n$, and in this regime there is essentially no compression, as the implicit constant in Theorem 3.1 grows rapidly with d.[2]

Our proof of the lower bound in Theorem 3.1 first uses a standard reduction from estimation to a multiple hypothesis testing problem over a finite function class. While Fano's inequality is the workhorse of our second step, note that the lower bound must hold only for coreset-based estimators and not *any* estimator as in standard minimax

[2]In fact, even for the classical estimation problem (3.1), this constant scales as d^d [see McDonald, 2017, Theorem 3].

lower bounds. This additional difficulty is overcome by a careful handling of the information structure generated by coreset scheme channels rather than using off-the-shelf results for minimax lower bounds. The full details of the lower bound are in the Appendix.

The estimator achieving the rate in Theorem 3.1 relies on an encoding procedure. It is constructed by building a dictionary between the subsets in $\binom{[n]}{m}$ and an ε-net on the space of Hölder functions. The key idea is that, for $1 \ll m \leq n/2$, the amount of subsets $\binom{n}{m}$ grows rapidly with m, so for m large enough, there is enough information to encode a nearby-neighbor in $L^2(\mathbb{R}^d)$ to the kernel density estimator on the entire dataset.

3.2.1 Proof of the upper bound in Theorem 3.1

Fix $\varepsilon = c^*(m \log n)^{-\frac{\beta}{d}}$ for c^* to be determined and let $\mathcal{N}_\varepsilon$ denote an ε-net of $\mathcal{P}_{\mathcal{H}}(\beta, L)$ with respect to the $L^2([-\frac{1}{2}, \frac{1}{2}]^d)$ norm. It follows from the classical Kolmogorov-Tikhomorov bound [see, e.g., Theorem XIV of Kolmogorov and Tikhomirov, 1993] that there exists a constant $C_{\mathsf{KT}}(\beta, d, L) > 0$ such that we can choose $\mathcal{N}_\varepsilon$ with $\log |\mathcal{N}_\varepsilon| \leq C_{\mathsf{KT}}(\beta, d, L)\, \varepsilon^{-d/\beta}$. In particular, there exists $\mathsf{f} \in \mathcal{N}_\varepsilon$ such that $\|\hat{f}_n - \mathsf{f}\|_2 \leq \varepsilon$ where $\hat{f}_n$ is the minimax optimal kernel density estimator defined in (3.2).

We now develop our encoding procedure for f. To that end, fix an integer $K \geq m$ such that $\binom{K}{m} \geq |\mathcal{N}_\varepsilon|$ and let $\phi : \binom{[K]}{m} \to \mathcal{N}_\varepsilon$ be any surjective map. Our procedure only looks at the first coordinates of the sample $X = \{X_1, \ldots, X_n\}$. Denote these coordinates by $x = \{x_1, \ldots, x_n\}$ and note that these n numbers are almost surely distinct. Let A denote a parameter to be determined, and define the intervals

$$B_{ik} = [(i-1)K^{-1}A + (k-1)A,\ (i-1)K^{-1}A + (k-1)A + K^{-1}A].$$

For $i = 1, \ldots, K$, define

$$B_i = \bigcup_{k=1}^{1/A} B_{ik}.$$

The next lemma, whose proof is in the Appendix, ensures that with high probability every bin B_i contains the first coordinate x_i of at least one data point.

Lemma 3.1. *Let $K^{-1} = c(\log n)/n$ for $c > 0$ a sufficiently large absolute constant, and let $A = A_{\beta,L,K}$ denote a sufficiently small constant. Then for all $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$ and $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}_f$, the event that for every $j = 1, \ldots, K$ there exists some x_i in bin B_j holds with probability at least $1 - O(n^{-2})$.*

In the high-probability event $\mathcal{E}$ that every bin B_i contains the first coordinate of some data point, choose a unique representative $x_j^\circ \in x$ such that $x_j^\circ \in B_j$ and pick any $T_{\mathsf{f}} \in \phi^{-1}(\mathsf{f})$. Then define $S = \{i : x_i = x_j^\circ, j \in T_{\mathsf{f}}\}$. If there exists a bin with no observation, then let X_S consist of two data points lying in the same bin and $m - 2$ random data points. Then set $\hat{f}_S \equiv 0$.

Note that $\hat{f}_S$ is indeed a coreset-based estimator. The function $\hat{f}$ such that $\hat{f}_S = \hat{f}[X_S]$ looks at the m data points in the coreset, and if their first coordinates lie in

distinct bins, then X_S is decoded as above to output the corresponding element f of the net $\mathcal{N}_\varepsilon$. Otherwise, $\hat{f} \equiv 0$.

Next, it suffices to show the upper bound of Theorem 3.1 in the case when $m \leq cn^{d/(2\beta+d)}$ for c a sufficiently small absolute constant. For $c^* = c^*_{\beta,d,L}$ sufficiently large, by Stirling's formula and our choice of K it holds that

$$\log \binom{K}{m} \geq C_{\mathsf{KT}}(\beta, d, L) \left(\frac{1}{\varepsilon}\right)^{\frac{d}{\beta}} \geq \log |\mathcal{N}_\varepsilon|.$$

Hence, the surjection ϕ and our encoding estimator $\hat{f}_S$ are well-defined.

Next we have

$$\mathbb{E}_f \|\hat{f}_S - f\|_2 = \mathbb{E}_f\Big[\|\mathsf{f} - f\|_2 \mathbb{1}_{\mathcal{E}}\Big] + \mathbb{E}_f\Big[\|0 - f\|_2 \mathbb{1}_{\mathcal{E}^c}\Big].$$

We control the first term as follows using (3.1) and the fact that $\|\mathsf{f} - \hat{f}_n\|_2 \leq \varepsilon$ on $\mathcal{E}$:

$$\begin{aligned}\mathbb{E}_f\Big[\|\mathsf{f} - f\|_2 \mathbb{1}_{\mathcal{E}}\Big] &\leq \mathbb{E}_f \|\hat{f}_n - f\|_2 + \mathbb{E}_f \|\mathsf{f} - \hat{f}_n\|_2 \\ &\leq c_{\beta,d,L} \left(n^{\frac{-\beta}{2\beta+d}} + (m \log n)^{-\frac{\beta}{d}}\right).\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}_f\Big[\|0 - f\|_2 \mathbb{1}_{\mathcal{E}^c}\Big] &\leq \left(\mathbb{E}_f \|f\|_2^2 \, \mathbb{P}(\mathcal{E}^c)\right)^{1/2} \\ &\leq c_{\beta,d,L} \, n^{-1}.\end{aligned}$$

Put together, the previous three displays yield the upper bound of Theorem 3.1.

3.3 Coreset kernel density estimators

In this section, we consider the family of weighted kernel density estimators built on coresets and study its rate of estimation over the Hölder densities. In this framework, the statistician first computes a minimax estimator $\hat{f}$ using the entire dataset and then approximates $\hat{f}$ with a weighted kernel density estimator over the coreset. Here we allow the weights to be a measurable function of the entire dataset rather than just the coreset.

As is typical in density estimation, we consider kernels $k : \mathbb{R}^d \to \mathbb{R}$ of the form $k(x) = \prod_{i=1}^d \kappa(x_i)$ where κ is an even function and $\int \kappa(x)\,\mathrm{d}x = 1$. Given bandwidth parameter h, we define $k_h(x) = h^{-d}\, k(\frac{x}{h})$.

3.3.1 Carathéodory coreset method

Given a KDE with uniform weights and bandwidth h defined by

$$\hat{f}(y) = \frac{1}{n} \sum_{j=1}^n k_h(X_j - y),$$

on a sample $X_1, \ldots, X_n$, we define a coreset KDE $\hat{g}_S$ as follows in terms of a cutoff frequency $T > 0$. Define $A = \{\omega \in \frac{\pi}{2}\mathbb{Z}^d : |\omega|_\infty \leq T\}$. Consider the complex vectors $(e^{i\langle X_j, \omega\rangle})_{\omega \in A}$. By Carathéodory's theorem [Carathéodory, 1907], there exists a subset $S \subset [n]$ of cardinality at most $2(1 + \frac{4T}{\pi})^d + 1$ and nonnegative weights $\{\lambda_j\}_{j \in S}$ with $\sum_{j \in S} \lambda_j = 1$ such that

$$\frac{1}{n} \sum_{j=1}^{n} (e^{i\langle X_j, \omega\rangle})_{\omega \in A} = \sum_{j \in S} \lambda_j (e^{i\langle X_j, \omega\rangle})_{\omega \in A}. \tag{3.4}$$

Then $\hat{g}_S(y)$ is defined to be

$$\hat{g}_S(y) = \sum_{j \in S} \lambda_j k_h(X_j - y).$$

Algorithmic considerations

For a convex polyhedron P with vertices $v_1, \ldots, v_n \in \mathbb{R}^D$, the proof of Carathéodory's theorem is constructive and yields a polynomial-time algorithm in n and D to find a convex combination of $D+1$ vertices that represents a given point in P [Carathéodory, 1907] [see also Hiriart-Urruty and Lemaréchal, 2004, Theorem 1.3.6]. For completeness, we describe below this algorithm applied to our problem. Note that, more generally, for a large class of convex bodies, Carathéodory's theorem may be implemented efficiently using standard tools from convex optimization [Grötschel et al., 2012, Chapter 6].

Set $D = 2|A| \leq 2(1 + \frac{4T}{\pi})^d$. For $j = 1, \ldots, n$, let

$$v_j = (\mathsf{Re}\, e^{i\langle X_j, \omega\rangle}, \mathsf{Im}\, e^{i\langle X_j, \omega\rangle})_{\omega \in A} \in \mathbb{R}^D.$$

Let M denote the matrix with columns $(v_1, 1)^T, \ldots, (v_n, 1)^T \in \mathbb{R}^{D+1}$, and let $\Delta_{n-1} \subset \mathbb{R}^n$ denote the standard simplex. Assume without loss of generality that $n \geq D + 2$. Next,

1. Find a nonzero vector $w \in \ker(M)$
2. Find $\alpha > 0$ so that $\lambda_1 := \frac{1}{n}\mathbb{1} + \alpha w$ lies on the boundary of Δ_{n-1}

Observe that $M\lambda_1 = (\frac{1}{n}\sum v_i, 1)^T$, and since $\lambda_1 \in \partial\Delta_{n-1}$ the average is now represented using a convex combination of at most $n-1$ of the vertices $v_1, \ldots, v_n$. As long as at least $D + 2$ vertices remain, we can continue reducing the number of vertices used to represent $\frac{1}{n}\sum v_j$ by applying steps 1 and 2. Thus after at most $n - D - 1$ iterations, we obtain a $(D+1)$-sparse vector $\lambda \in \Delta_{n-1}$ that satisfies $\sum \lambda_j v_j = \frac{1}{n}\sum v_i$, as desired.

3.3.2 Results on Carathéodory coresets

Proposition 3.1 is key to our results and specifies conditions on the kernel guaranteeing that the Carathéodory method yields an accurate estimator.

Proposition 3.1. *Let $k(x) = \prod_{i=1}^{d} \kappa(x_i)$ denote a kernel with $\kappa \in \mathcal{S}(\gamma, L')$ such that $|\kappa(x)| \leq c_{\beta,d} |x|^{-\nu}$ for some $\nu \geq \beta + d$ and such that the KDE*

$$\hat{f}(y) = \frac{1}{n} \sum_{i=1}^{n} k_h(X_i - y)$$

with bandwidth $h = n^{-\frac{1}{2\beta+d}}$ satisfies

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}\|f - \hat{f}\|_2 \leq c_{\beta,d,L}\, n^{-\frac{\beta}{2\beta+d}}. \tag{3.5}$$

Then the Carathéodory coreset estimator $\hat{g}_S$ constructed from $\hat{f}$ with $T = c_{d,\gamma,L'}\, n^{\frac{d/2+\beta+\gamma}{\gamma(2\beta+d)}}$ satisfies

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}\|\hat{g}_S - f\|_2 \leq c_{\beta,d,L}\, n^{-\frac{\beta}{2\beta+d}}.$$

There exists a kernel $k_s \in \mathcal{C}^\infty$ that satisfies the conditions above for all β and γ. We sketch the details here and postpone the full argument to the Proof of Theorem 3.2 in the Appendix. Let $\psi : \mathbb{R} \to [0,1]$ denote a cutoff function that has the following properties: $\psi \in \mathcal{C}^\infty$, $\psi\big|_{[-1,1]} \equiv 1$, and ψ is supported on $[-2,2]$. Define $\kappa_s(x) = \mathcal{F}[\psi](x)$, and let $k_s(x) = \prod_{i=1}^{d} \kappa_s(x_i)$ denote the resulting kernel. Observe that for all $\beta > 0$, the kernel k_s satisfies

$$\text{ess sup}_{\omega \neq 0} \frac{|1 - \mathcal{F}[k_s](\omega)|}{|\omega|^{\alpha}} \leq 1, \quad \forall \alpha \preceq \beta.$$

Using standard results from Tsybakov [2009], this implies that the resulting KDE $\hat{f}_s$ satisfies (3.5). Since $\psi = \mathcal{F}^{-1}[k_s] \in \mathcal{C}^\infty$, the Riemann–Lebesgue lemma guarantees that $|\kappa_s(x)| \leq c_{\beta,d} |x|^{-\nu}$ is satisfied for $\nu = \lceil \beta + d \rceil$. Since ψ is compactly supported, an application of Parseval's identity yields $\kappa_s \in \mathcal{S}(\gamma, c_\gamma)$. Applying Proposition 3.1 to k_s, we conclude that for the task of density estimation, weighted KDEs built on coresets are nearly as powerful as the coreset-based estimators studied in Section 3.2.

Theorem 3.2. *Let $\varepsilon > 0$. The Carathéodory coreset estimator $\hat{g}_S(y)$ built using the kernel k_s and setting $T = c_{d,\beta,\varepsilon}\, n^{\frac{\varepsilon}{d} + \frac{1}{2\beta+d}}$ satisfies*

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f\|\hat{g}_S - f\|_2 \leq c_{\beta,d,L}\, n^{-\frac{\beta}{2\beta+d}}.$$

The corresponding coreset has cardinality

$$m = c_{d,\beta,\varepsilon} n^{\frac{d}{2\beta+d}+\varepsilon}.$$

Theorem 3.2 shows that the Carathéodory coreset estimator achieves the minimax rate of estimation with near-optimal coreset size. In fact, a small modification yields a near-optimal rate of convergence for any coreset size as in Theorem 3.1.

Corollary 3.1. *Let $\varepsilon > 0$ and $m \leq c_{\beta,d,\varepsilon}\, n^{\frac{d}{2\beta+d}+\varepsilon}$. The Carathéodory coreset estimator $\hat{g}_S(y)$ built using the kernel k_s, setting $h = m^{-\frac{1}{d}+\frac{\varepsilon}{\beta}}$ and $T = c_d\, m^{1/d}$, satisfies*

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}\|\hat{g}_S - f\|_2 \leq c_{\beta,d,\varepsilon,L} \left(m^{-\frac{\beta}{d}+\varepsilon} + n^{-\frac{\beta}{2\beta+d}+\varepsilon} \right),$$

and the corresponding coreset has cardinality m.

Next we apply Proposition 3.1 to the popular Gaussian kernel $\phi(x) = (2\pi)^{-d/2} \exp(-\frac{1}{2}\,|x|_2^2)$. This kernel has rapid decay in the real domain and Fourier space, and is thus amenable to our techniques. Moreover, ϕ is a kernel of order $\ell = 1$, [Tsybakov, 2009, Definition 1.3 and Theorem 1.2] and so the standard KDE $\hat{f}_\phi$ on the full dataset attains the minimax rate of estimation $c_{d,L} n^{1/(2+d)}$ over the Lipschitz densities $\mathcal{P}_{\mathcal{H}}(1, L)$.

Theorem 3.3. *Let $\varepsilon > 0$. The Carathéodory coreset estimator $\hat{g}_\phi(y)$ built using the kernel ϕ and setting $T = c_{d,\varepsilon}\, n^{\frac{1}{2+d}+\frac{\varepsilon}{d}}$ satisfies*

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(1,L)} \mathbb{E}\|\hat{g}_\phi - f\|_2 \leq c_{d,L}\, n^{-\frac{1}{2+d}}.$$

The corresponding coreset has cardinality

$$m = c_{d,\varepsilon} n^{\frac{d}{2+d}+\varepsilon}.$$

In addition, we have a nearly matching lower bound to Theorem 3.2 for coreset KDEs. In fact, our lower bound applies to a generalization of coreset KDEs where the vector of weights $\{\lambda_j\}_{j \in S}$ is not constrained to be in the simplex but can range within a hypercube of width that may grow polynomially with n.

Theorem 3.4. *Let $A, B \geq 1$. Let k denote a kernel with $\|k\|_2 \leq n$. Let $\hat{g}_S$ denote a weighted coreset KDE with bandwidth $h \geq n^{-A}$ built from k with weights $\{\lambda_j\}_{j \in S}$ satisfying $\max_{j \in S} |\lambda_j| \leq n^B$. Then*

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f\|\hat{g}_S - f\|_2 \geq c_{\beta,d,L} \left[(A+B)^{-\frac{\beta}{d}} (m \log n)^{-\frac{\beta}{d}} + n^{-\frac{\beta}{2\beta+d}} \right].$$

This result is essentially a consequence of the lower bound in Theorem 3.1 because, in an appropriate sense, coreset KDEs with bounded weights are well-approximated by coreset-based estimators. Hence, in the case of bounded weights, allowing these weights to be measurable functions of the entire dataset rather than just the coreset, as would be required in Section 3.2, does not make a significant difference for the purpose of estimation. The full details of Theorem 3.4 are postponed to the Appendix.

3.3.3 Proof sketch of Proposition 3.1

Here we sketch the proof of Proposition 3.1, our main tool in constructing effective coreset KDEs. Full details of the argument may be found in the Appendix.

Let $k(x) = \prod_{i=1}^d \kappa(x_i)$ denote a kernel, and suppose that $\hat{f}(y) = \frac{1}{n}\sum_{i=1}^n k_h(X_i - y)$ is a good estimator for an unknown density f in that

$$\|f - \hat{f}\|_2 \leq \varepsilon := c_{\beta,d}\, n^{-\frac{\beta}{2\beta+d}}$$

on setting $h = n^{-1/(2\beta+d)}$. Our goal is to find a subset $S \subset [n]$ and weights $\{\lambda_j\}_{j\in S}$ such that

$$\frac{1}{n}\sum_{i=1}^n k_h(X_i - y) \approx \sum_{j\in S} \lambda_j k_h(X_j - y).$$

Suppose for simplicity that κ is compactly supported on $[-1/2, 1/2]$. By hypothesis and Parseval's theorem $\kappa \in \mathcal{S}(\gamma, L')$, and we can further show that $k \in \mathcal{S}(\gamma, c_{d,L'})$ and $k_h \in \mathcal{S}(\gamma, c_{d,L'}h^{-d/2-\gamma})$. Let $\bar{\mathcal{F}}[f] = 4^{-2d}\mathcal{F}[f]$ denote the rescaled Fourier transform. Using the Fourier expansion on the interval $[-2,2]^d$ and fast Fourier decay of k_h, we have

$$\|k_h(x) - \sum_{|\omega|_\infty < T} \bar{\mathcal{F}}[k_h](\omega) e^{i\langle x,\omega\rangle}\|_2 \leq \varepsilon \tag{3.6}$$

when $T = (\frac{c_{d,\gamma,L'}h^{-\frac{d}{2}-\gamma}}{\varepsilon})^{1/\gamma} = c_{d,\gamma,L'}\, n^{\frac{d/2+\beta+\gamma}{\gamma(2\beta+d)}}$. Observe that this matches the setting of T in Proposition 3.1.

The approximation (3.6) implies that for $X_i \in [-1/2, 1/2]^d$,

$$\hat{f}(y) \approx \sum_{|\omega|_\infty < T} \bar{\mathcal{F}}[k_h](\omega) \left(\frac{1}{n}\sum_{i=1}^n e^{i\langle X_i,\omega\rangle}\right) e^{-i\langle y,\omega\rangle}.$$

Using the Carathéodory coreset and weights $\{\lambda_j\}_{j\in S}$ constructed in Section 3.3.1, it follows that

$$\sum_{|\omega|_\infty < T} \bar{\mathcal{F}}[k_h](\omega) \left(\frac{1}{n}\sum_{i=1}^n e^{i\langle X_i,\omega\rangle}\right) e^{-i\langle y,\omega\rangle} = \sum_{|\omega|_\infty < T} \bar{\mathcal{F}}[k_h](\omega) \left(\sum_{i=1}^n \lambda_j e^{i\langle X_i,\omega\rangle}\right) e^{-i\langle y,\omega\rangle}.$$

Applying (3.6) again, we see that the right-hand-side is approximately equal to $\hat{g}_S(y)$, the estimator produced in Section (3.3.1). By the triangle inequality, we conclude that $\|\hat{g}_S(y) - f\|_2 \leq c_{\beta,d}\,\varepsilon$, as desired.

3.4 Lower bounds for coreset KDEs with uniform weights

In this section we study the performance of univariate uniformly weighted coreset KDEs

$$\hat{f}_S^{\mathsf{unif}}(y) = \frac{1}{m}\sum_{i\in S} k_h(X_i - y),$$

where X_S is the coreset and $|S| = m$. The next results demonstrate that for a large class of kernels, there is significant gap between the rate of estimation achieved

by $\hat{f}_S^{\mathrm{unif}}(y)$ and that of coreset KDEs with general weights. First we focus on the particular case of estimating the class $\mathcal{P}_{\mathcal{H}}(1, L)$ of univariate Lipschitz densities. For this class, the minimax rate of estimation (over all estimators) is $n^{-1/3}$, and this can be achieved by a weighted coreset KDE of cardinality $c_\varepsilon n^{1/3+\varepsilon}$ by Theorem 3.2, for all $\varepsilon > 0$.

Theorem 3.5. *Let k denote a nonnegative kernel satisfying*

$$k(t) = O(|t|^{-(k+1)}), \quad \textit{and} \quad \mathcal{F}[k](\omega) = O(|\omega|^{-\ell})$$

for some $\ell > 0$, $k > 1$. Suppose that $0 < \alpha < 1/3$. If

$$m \leq \frac{n^{\frac{2}{3}-2\left(\alpha(1-\frac{2}{\ell})+\frac{2}{3\ell}\right)}}{\log n},$$

then

$$\inf_{h,S:|S|\leq m} \sup_{f\in\mathcal{P}_{\mathcal{H}}(1,L)} \mathbb{E}\|\hat{f}_S^{\mathsf{unif}} - f\|_2 = \Omega_k\Big(\frac{n^{-\frac{1}{3}+\alpha}}{\log n}\Big). \tag{3.7}$$

The infimum above is over all possible choices of bandwidth h and all coreset schemes S of cardinality at most m.

By this result, if k has lighter than quadratic tails and fast Fourier decay, the error in (3.7) is a polynomial factor larger than the minimax rate $n^{-1/3}$ when $m \ll n^{2/3}$. Hence, our result covers a wide variety of kernels typically used for density estimation and shows that the uniformly weighted coreset KDE performs much worse than the encoding estimator or the Carathéodory method. In addition, for very smooth univariate kernels with rapid decay, we have the following lower bound that applies for all $\beta > 0$.

Theorem 3.6. *Fix $\beta > 0$ and a nonnegative kernel k on $\mathbb{R}$ satisfying the following fast decay and smoothness conditions:*

$$\lim_{s\to+\infty} \frac{1}{s} \log \frac{1}{\int_{|t|>s} k(t)dt} > 0, \tag{3.8}$$

$$\lim_{\omega\to\infty} \frac{1}{|\omega|} \log \frac{1}{|\mathcal{F}[k](\omega)|} > 0, \tag{3.9}$$

where we recall that $\mathcal{F}[k]$ denotes the Fourier transform. Let $\hat{f}_S^{\mathsf{unif}}$ be the uniformly weighted coreset KDE. Then there exists $L_\beta > 0$ such that for $L \geq L_\beta$ and any m and $h > 0$, we have

$$\inf_{h,S:|S|\leq m} \sup_{f\in\mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}\|\hat{f}_S^{\mathsf{unif}} - f\|_2 = \Omega_{\beta,k}\left(\frac{m^{-\frac{\beta}{1+\beta}}}{\log^{\beta+\frac{1}{2}} m}\right).$$

Therefore attaining the minimax rate with $\hat{f}_S^{\mathsf{unif}}$ requires $m \geq n^{\frac{\beta+1}{2\beta+1}}$ for such kernels. Next, note that the Gaussian kernel satisfies the hypotheses of Theorem

3.5 and 3.6. As we show in Theorem 3.7, results of Phillips and Tai [2018b] imply that our lower bounds are tight up to logarithmic factors: there exists a uniformly weighted Gaussian coreset KDE of size $m = \tilde{O}(n^{2/3})$ that attains the minimax rate $n^{-1/3}$ for estimating univariate Lipschitz densities ($\beta = 1$). In general, we expect a lower bound $m = \Omega(n^{\frac{\beta+d}{2\beta+d}})$ to hold for uniformly weighted coreset KDEs attaining the minimax rate. The proofs of Theorems 3.5 and 3.6 can be found in the Appendix.

3.5 Comparison to other methods

Three methods for constructing coreset kernel density estimators that have previously been explored include random sampling [Joshi et al., 2011, Lopez-Paz et al., 2015], the Frank–Wolfe algorithm [Bach et al., 2012, Harvey and Samadi, 2014, Phillips and Tai, 2018a], and discrepancy-based approaches [Phillips and Tai, 2018b, Karnin and Liberty, 2019]. These procedures all result in a uniformly weighted coreset KDE. To compare these results with ours on the problem of density estimation, for each method under consideration we raise the question: How large does m, the size of the coreset, need to be to guarantee that

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta, L)} \mathbb{E}_f \|\hat{g}_S - f\|_2 = O_{\beta,d,L}\left(n^{-\frac{\beta}{2\beta+d}}\right)? \tag{3.10}$$

Here $\hat{g}_S$ is the resulting coreset KDE and the right-hand-side is the minimax rate over all estimators on the full dataset $X_1, \ldots, X_n$.

Uniform random sampling of a subset of cardinality m yields an i.i.d dataset, so the rate obtained is at least $m^{-\beta/(2\beta+d)}$. Hence, we must take $m = \Omega(n)$ to achieve the minimax rate.

The Frank–Wolfe algorithm is a greedy method that iteratively constructs a sparse approximation to a given element in a convex set [Frank et al., 1956, Bubeck, 2015]. Thus Frank–Wolfe may be applied directly in the RKHS corresponding to a positive-semidefinite kernel as shown in Phillips and Tai [2018b] to approximate the KDE on the full dataset. However, due to the shrinking bandwidth in our problem, this approach also requires $m = \Omega(n)$ to guarantee the bound in (3.10). Another strategy is to approximately solve the linear equation (3.4) using the Frank–Wolfe algorithm. Unfortunately, a direct implementation again uses $m = \Omega(n)$ data points.

A more effective strategy utilizes discrepancy theory [Phillips, 2013, Phillips and Tai, 2018b, Karnin and Liberty, 2019] [see Matoušek, 1999, Chazelle, 2000, for a comprehensive exposition of discrepancy theory]. By the well-known halving algorithm [see e.g. Chazelle and Matoušek, 1996, Phillips and Tai, 2018b] if for all $N \leq n$, the *kernel discrepancy*

$$\mathsf{disc}_k = \sup_{x_1,\ldots,x_N} \min_{\substack{\sigma \in \{-1,+1\}^n \\ \mathbb{1}^T\sigma = 0}} \|\sum_{i=1}^{N} \sigma_i k(x_i - y)\|_\infty$$

is at most D, then there exists a coreset X_S of size $\tilde{O}_D(\varepsilon^{-1})$ such that

$$\|\frac{1}{n}\sum_{i=1}^{n} k(X_i - y) - \frac{1}{m}\sum_{j\in S} k(X_i - y)\|_\infty \leq \varepsilon. \tag{3.11}$$

The idea of the halving algorithm is to maintain a set of datapoints $\mathcal{C}_\ell$ at each iteration and then set $\mathcal{C}_{\ell+1}$ to be the set of vectors that receive sign $+1$ upon minimizing $\|\sum_{x\in\mathcal{C}_\ell} \sigma_x k(x-y)\|_\infty$. Starting with the original dataset and repeating this procedure $O(\log \frac{n}{m})$ times yields the desired coreset X_S satisfying (3.11).

Phillips and Tai [2018b, Theorem 4] use a state-of-the-art algorithm from Bansal et al. [2018] called the *Gram–Schmidt walk* to give strong bounds on the kernel discrepancy of bounded and Lipschitz kernels $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that are positive definite and decay rapidly away from the diagonal. With a careful handling of the Lipschitz constant and error in their argument when the bandwidth is set to be $h = n^{-1/(2\beta+d)}$, their techniques yield the following result applied to the kernel k_s. For completeness we give details of the argument in the Appendix.

Theorem 3.7. *Let k_s denote the kernel from Section 3.3.2. The algorithm of Phillips and Tai [2018b] yields in polynomial time a subset S with $|S| = m = \tilde{O}(n^{\frac{\beta+d}{2\beta+d}})$ such that the uniformly weighted coreset KDE $\hat{g}_S$ satisfies*

$$\sup_{f\in\mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}\|f - \hat{g}_S\|_2 \leq c_{\beta,d,L}\, n^{-\frac{\beta}{2\beta+d}}.$$

This result also applies to more general kernels, for example, the Gaussian kernel when $\beta = 1$. We suspect that this is the best result achievable by discrepancy-based methods. In particular for nonnegative univariate kernels with fast decay in the real and Fourier domains, such as the Gaussian kernel, Theorem 3.5 implies that this rate is optimal for estimating Lipschitz densities with uniformly weighted coreset KDEs.

In contrast, the Carathéodory coreset KDE as in Theorem 3.2 only needs cardinality $m = O_\varepsilon(n^{\frac{d}{2\beta+d}+\varepsilon})$ to be a minimax estimator. By Theorem 3.4, this result is nearly optimal for coreset KDEs with bounded kernels and weights. And as with the other three methods described, our construction is computationally efficient. Hence allowing more general weights results in more powerful coreset KDEs for the problem of density estimation.

3.6 Appendix

3.6.1 Proofs from Section 3.2

Proof of Lemma 3.1

Note that $f_1(x_1) \in \mathcal{P}_{\mathcal{H}}(\beta, L)$ as a univariate density because $f(x) \in \mathcal{P}_{\mathcal{H}}(\beta, L)$. Hence, f_1 satisfies

$$|f_1(x) - f_1(y)| \leq L|x - y|^\alpha$$

for some absolute constants $L > 0$ and $\alpha \in (0,1)$. If $B_{ik} = B_{jk} + s$ for $s \le A$, then

$$|\mathbb{P}(B_{ik}) - \mathbb{P}(B_{jk})| \le \int_{B_{ik}} |f(x_1) - f(x_1 + s)| \, \mathrm{d}x_1 \le LK^{-1}A^{1+\alpha}. \tag{3.12}$$

Thus for all i, j,

$$|\mathbb{P}(B_i) - \mathbb{P}(B_j)| \le \sum_{k=1}^{1/A} |\mathbb{P}(B_{ik}) - \mathbb{P}(B_{jk})| \le LK^{-1}A^{\alpha}. \tag{3.13}$$

It follows that for all $i = 1, \ldots, K$,

$$\lim_{A \to 0} \mathbb{P}(B_i) = K^{-1}. \tag{3.14}$$

Let $\mathcal{E}$ denote the event that every bin B_i contains at least one observation x_k. By the union bound,

$$\mathbb{P}(\mathcal{E}^c) \le \sum_{j=1} \mathbb{P}(X_{11} \notin B_j)^n \le K \max_j (1 - \mathbb{P}(B_j))^n.$$

By (3.14), choosing A small enough ensures that $\mathbb{P}[B_j] \ge (1/2)K^{-1}$ for all j. In fact, by (3.12) one may take $A = (\frac{1}{2K^{-2}L})^{1/\alpha}$. Hence, setting $K^{-1} = c(\log n)/n$ for c sufficiently large, we have

$$\mathbb{P}(\mathcal{E}^c) = O(n^{-2}).$$

Proof of the lower bound in Theorem 3.1

In this section, $X = X_1, \ldots, X_n \in \mathbb{R}^d$ denotes the sample. It is convenient to consider a more general family of *decorated coreset-based estimators*. A *decorated coreset* consists of a coreset X_S along with a data-dependent binary string σ of length R. A decorated coreset-based estimator is then given by $\hat{f}[X_S, \sigma]$, where $\hat{f} : \mathbb{R}^{d \times m} \times \{0,1\}^R \to L^2([-1/2, 1/2]^d)$ is a measurable function. As with coreset-based estimators, we require that $\hat{f}[x_1, \ldots, x_m, \sigma]$ is invariant under permutation of the vectors $x_1, \ldots, x_m \in \mathbb{R}^d$. We slightly abuse notation and refer to the channel $S : X \to Y_S = (X_S, \sigma)$ as a decorated coreset scheme and $\hat{f}_S$ as the decorated coreset-based estimator.

The next proposition implies the lower bound in Theorem 3.1 on setting $R = 0$, in which case a decorated coreset-based estimator is just a coreset-based estimator. This more general framework allows us to prove Theorem 3.4. on lower bounds for weighted coreset KDEs.

Proposition 3.2. *Let $\hat{f}_S$ denote a decorated coreset-based estimator with decorated coreset scheme S such that $\sigma \in \{0,1\}^R$. Then*

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta, L)} \mathbb{E}_f \|\hat{f}_S - f\|_2 \ge c_{\beta,d,L} \left((m \log n + R)^{-\frac{\beta}{d}} + n^{-\frac{\beta}{2\beta + d}} \right).$$

Choice of function class

Fix $h \in (0,1)$ such that $1/h^d$ is integral to be chosen later. Let $z_1, \dots, z_{1/h^d}$ label the points in $\{\frac{1}{2}h \cdot \mathbb{1}_d + h\mathbb{Z}^d\} \cap [-1/2, 1/2]^d$, where $\mathbb{1}_d$ denotes the all-ones vector of $\mathbb{R}^d$. We consider a class of functions of the form $f_\omega(x) = 1 + \sum_{j=1}^{1/h^d} \omega_j g_j(x)$ indexed by $\omega \in \{0,1\}^{1/h^d}$. Here, $g_j(x)$ is defined to be

$$g_j(x) = h^\beta \phi\left(\frac{x - z_j}{h}\right)$$

where $\phi : \mathbb{R}^d \to \mathbb{R}$ is L-Hölder smooth of order β, has $\|\phi\|_\infty = 1$, and has $\int \phi(x)\,\mathrm{d}x = 0$.

Informally, f_ω puts a bump on the uniform distribution with amplitude h^β over z_j if and only if $\omega_i = 1$. Using a standard argument [Tsybakov, 2009, Chapter 2] we can construct a packing $\mathcal{V}$ of $\{0,1\}^{1/h^d}$ which results $\mathcal{G} = \{f_\omega : \omega \in \mathcal{V}\}$ of the function class $\{f_\omega : \omega \in \{0,1\}^{1/h^d}\}$ such that

(i) $\|f - g\|_2 \geq c_{\beta,d,L}\, h^\beta$ for all $f, g \in \mathcal{G}, f \neq g$ and,

(ii) $\mathcal{G}$ is large in the sense that $M := |\mathcal{G}| \geq 2^{c_{\beta,d,L}/h^d}$.

Minimax lower bound

Using standard reductions from estimation to testing, we obtain that

$$\begin{aligned}\inf_{\substack{\hat{f}, |S|=m, \\ \sigma \in \{0,1\}^R}} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f \|\hat{f}_S - f\|_2 &\geq \inf_{\substack{\hat{f}, |S|=m, \\ \sigma \in \{0,1\}^R}} \max_{f \in \mathcal{G}} \mathbb{E}_f \|\hat{f}_S - f\|_2 \\ &\geq c_{\beta,d,L}\, h^\beta \cdot \inf_{\psi_S} \frac{1}{M} \sum_{\omega \in \mathcal{V}} \mathbb{P}_{f_\omega}[\psi_S(X) \neq \omega]. \qquad (3.15)\end{aligned}$$

where the infimum in the last line is over all tests $\psi_S : \mathbb{R}^{d \times n} \to [M]$ of the form $\psi_S(X) = \psi(Y_S)$ for a decorated coreset scheme S and a measurable function $\psi : \mathbb{R}^{d \times m} \times \{0,1\}^R \to [M]$.

Let V denote a random variable that is distributed uniformly over $\mathcal{V}$ and observe that

$$\frac{1}{M} \sum_{\omega \in \mathcal{V}} \mathbb{P}_{f_\omega}[\psi_S(X) \neq \omega] = \mathbb{P}[\psi_S(X) \neq V]$$

where $\mathbb{P}$ denotes the joint distribution of (X, V) characterized by the conditional distribution $X|V = \omega$ which is assumed to have density f_ω for all $\omega \in \mathcal{V}$.

Next, by Fano's inequality [Cover and Thomas, 2006, Theorem 2.10.1] and the chain rule, we have

$$\mathbb{P}[\psi_S(X) \neq V] \geq 1 - \frac{I(V; \psi_S(X)) + 1}{\log M}, \qquad (3.16)$$

where $I(V; \psi_S(X))$ denotes the mutual information between V and $\psi_S(X)$ and we used the fact that the entropy of V is $\log M$. Therefore, it remains to control

$I(V;\psi_S(X))$. To that end, note that it follows from the data processing inequality that

$$I(V;\psi_S(X)) \le I(V;(X_S,\sigma)) = I(V;Y_S) = \mathsf{KL}(P_{V,Y_S}\|P_V \otimes P_{Y_S})\,,$$

where P_{V,Y_S}, P_V and P_{Y_S} denote the distributions of (V,Y_S), V and Y_S respectively and observe that P_{Y_S} is the mixture distribution given by $P_{Y_S}(A,t) = M^{-1}\sum_{\omega\in\mathcal{V}} P_{f_\omega}(X_S \in A, \sigma = t)$ for $A \subset \mathbb{R}^{d\times m}$ and $t \in \{0,1\}^R$. Denote by f_{ω,Y_S} the mixed density of $P_{f_\omega}(X_S \in \cdot, \sigma = \cdot)$, where the continuous component is with respect to the Lebesgue measure on $[-1/2,1/2]^{d\times m}$. Denote by $\bar{f}_{Y_S}$ the mixed density of the uniform mixture of these:

$$\bar{f}_{Y_S} := \frac{1}{M}\sum_{\omega\in\mathcal{V}} f_{\omega,Y_S}\,.$$

By a standard information-theoretic inequality, for all measures $\mathbb{Q}$ it holds that

$$\mathsf{KL}(P_{V,Y_S}\|P_V\otimes P_{Y_S}) = \frac{1}{M}\sum_{\omega}\mathsf{KL}(P_{Y_S|\omega}\|\,P_{Y_S}) \le \frac{1}{M}\sum_{\omega}\mathsf{KL}(P_{Y_S|\omega}\|\,\mathbb{Q}). \tag{3.17}$$

In fact, we have equality precisely when $\mathbb{Q} = P_{Y_S}$, and (3.17) follows immediately from the nonnegativity of the KL-divergence. Setting $\mathbb{Q} = \mathsf{Unif}[-\frac{1}{2},\frac{1}{2}]^d \otimes \mathsf{Unif}\{0,1\}^R$, for all ω we have

$$\begin{aligned}\mathsf{KL}(P_{Y_S|\omega},\mathbb{Q}) &= \sum_{t\in\{0,1\}^R}\int_{[-\frac{1}{2},\frac{1}{2}]^d} f_{\omega,Y_S}(x,t)\log\frac{f_{\omega,Y_S}(x,t)}{2^{-R}}\,\mathrm{d}x\\ &\le \sum_{t\in\{0,1\}^R}\int_{[-\frac{1}{2},\frac{1}{2}]^d} f_{\omega,Y_S}(x,t)\log f_{\omega,Y_S}(x,t)\,\mathrm{d}x + R. \end{aligned}\tag{3.18}$$

Our next goal is to bound the first term on the right-hand-side above.

Lemma 3.2. *For any $\omega \in \mathcal{V}$, we have*

$$\sum_{t\in\{0,1\}^R}\int_{[-\frac{1}{2},\frac{1}{2}]^d} f_{\omega,Y_S}(x,t)\log f_{\omega,Y_S}(x,t)\,\mathrm{d}x \le 3m\log n.$$

Proof. Let $\mathbb{P}_{X_S}$ denote the distribution of the (undecorated) coreset X_S, and note that the density of this distribution is given by $f_{\omega,X_S}(x) := \sum_{t\in\{0,1\}^R} f_{\omega,Y_S}(x,t)$. Then because the logarithm is increasing,

$$\begin{aligned}\sum_{t\in\{0,1\}^R}\int_{[-\frac{1}{2},\frac{1}{2}]^d} f_{\omega,Y_S}(x,t)\log f_{\omega,Y_S}(x,t)\,\mathrm{d}x &\le \sum_{t\in\{0,1\}^R}\int_{[-\frac{1}{2},\frac{1}{2}]^d} f_{\omega,Y_S}(x,t)\log f_{\omega,X_S}(x)\,\mathrm{d}x\\ &= \int_{[-\frac{1}{2},\frac{1}{2}]^d} f_{\omega,X_S}(x)\log f_{\omega,X_S}(x)\,\mathrm{d}x.\end{aligned}$$

By the union bound,

$$\mathbb{P}_{X_S}(\cdot) \leq \sum_{s \in \binom{[n]}{m}} \mathbb{P}_{X_s}(\cdot) = \binom{n}{m} \mathbb{P}_{X_{[m]}}(\cdot)\,.$$

It follows readily that $f_{\omega,X_S}(\cdot) \leq \binom{n}{m} f_{\omega,X_{[m]}}(\cdot)$. Next, let $Z \in [-1/2, 1/2]^{d \times m}$ be a random variable with density f_{ω,X_S} and note that

$$\int f_{\omega,X_S} \log f_{\omega,X_S} = \mathbb{E} \log f_{\omega,X_S}(Z) \leq \log \binom{n}{m} + \mathbb{E} \log f_{\omega,X_{[m]}}(Z) \leq m \log\big(\frac{en}{m}\big) + m \log 2\,,$$

where in the last inequality, we use the fact that $f_{\omega,X_{[m]}} = f_\omega^m \leq 2^m$. The lemma follows. □

Since $\log M \geq c_{\beta,d,L} h^{-d}$, it follows from (3.16)–(3.18) and Lemma 3.2 that

$$\mathbb{P}[\psi_S(X) \neq V] \geq 1 - \frac{3m \log n + R + 1}{\log M} \geq 0.5$$

on setting $h = c_{\beta,d,L}(m \log n + R)^{-1/d}$. Plugging this value back into (3.15) yields

$$\inf_{\hat{f}, |S|=m} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f \, \|\hat{f}_S - f\|_2 \geq c_{\beta,d,L}(m \log n + R)^{-\beta/d}\,.$$

Moreover, it follows from standard minimax theory [see e.g. Tsybakov, 2009, Chapter 2] that

$$\inf_{\hat{f}, |S|=m} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f \, \|\hat{f}_S - f\|_2 \geq c_{\beta,d,L} n^{-\frac{\beta}{2\beta+d}}\,.$$

Combined together, the above two displays give the lower bound of Proposition 3.2.

3.6.2 Proofs from Section 3.3

Proof of Proposition 3.1

Let $\varphi : \mathbb{R}^d \to [0,1]$ denote a cutoff function that has the following properties: $\varphi \in \mathcal{C}^\infty$, $\varphi\big|_{[-1,1]^d} \equiv 1$, and φ is compactly supported on $[-2,2]^d$.

Lemma 3.3. *Let $\tilde{k}_h(x) = k_h(x)\varphi(x)$ where $|\kappa(x)| \leq c_{\beta,d}\,|x|^{-\nu}$. Then*

$$\|\tilde{k}_h - k_h\|_2 \leq c_{\beta,d}\, h^{-d+\nu}.$$

Proof.

$$\begin{aligned}
\|\tilde{k}_h - k_h\|_2 &= \|(1-\varphi)k_h\|_2 \\
&\le \|(1-\mathbb{1}_{[-1,1]^d})k_h\|_2 \\
&= h^{-d/2}\|(1-\mathbb{1}_{[-\frac{1}{h},\frac{1}{h}]^d})k\|_2 \\
&\le dh^{-d/2}\|\mathbb{1}_{|x_1|\ge\frac{1}{h}}\, k\|_2 \\
&\le c_{\beta,d}\, h^{-d/2}\sqrt{\int_{|x_1|\ge\frac{1}{h}} \kappa^2(x_1)\,\mathrm{d}x_1} \\
&\le c_{\beta,d}\, h^{-d+\nu}.
\end{aligned}$$

□

The triangle inequality and the previous lemma yield the next result.

Lemma 3.4. *Let k denote a kernel such that $|\kappa(x)| \le c_{\beta,d}\,|x|_2^{-\nu}$. Recall the definition of $\tilde{k}_h$ from Lemma 3.3. Let $\{X_j : j \in S\} \subset \mathbb{R}^d$ denote an arbitrary set of points (not necessarily from a sample), and let*

$$\hat{g}_S(y) = \sum_{j\in S} \lambda_j k_h(X_j - y)$$

denote a weighted KDE on the points labeled by S where $\lambda_j \ge 0$ and $\mathbb{1}^T\lambda = 1$. Let

$$\tilde{g_S}(y) = \sum_{j\in S} \lambda_j \tilde{k}_h(X_j - y).$$

Then

$$\|\hat{g}_S - \tilde{g_S}\|_2 \le c_{\beta,d} h^{-\nu+d}.$$

Next we show that $\tilde{k}_h$ is well approximated by its Fourier expansion on $[-2,2]^d$. Since $\tilde{k}_h$ is a smooth periodic function on $[-2,2]^d$, it is expressed in L^2 as a Fourier series on $\frac{\pi}{2}\mathbb{Z}^d$. Thus we bound the tail of this expansion. In what follows, $\alpha \in \mathbb{Z}^d_{\ge 0}$ is a multi-index and

$$\bar{\mathcal{F}}[f](\omega) = \frac{1}{4^{2d}}\int f(x)e^{-i\langle x,\omega\rangle}\,\mathrm{d}x$$

denotes the (rescaled) Fourier transform on $[-2,2]^d$, where $\omega \in \frac{\pi}{2}\mathbb{Z}^d$.

Lemma 3.5. *Suppose that $\kappa \in \mathcal{S}(\beta, L')$. Let $A = \{\omega \in \frac{\pi}{2}\mathbb{Z}^d : |\omega|_1 \le T\}$, and define*

$$\tilde{k}_h^T(y) = \sum_{\omega\in A} \bar{\mathcal{F}}[\tilde{k}_h](\omega)e^{i\langle y,\omega\rangle}.$$

Then

$$\|(\tilde{k}_h - \tilde{k}_h^T)\mathbb{1}_{[-2,2]^d}\|_2 \le c_{\gamma,d,L'}\, T^{-\gamma}h^{-d/2-\gamma}$$

Proof. Observe that for $\omega \notin A$, it holds that

$$\sum_{|\alpha|_1=\gamma} \frac{\gamma!}{\alpha!}\,|\omega|^\alpha = (|\omega_1| + \cdots + |\omega_d|)^\gamma \ge T^\gamma.$$

Therefore,

$$\begin{aligned}\|\bar{\mathcal{F}}[\tilde{k}_h](\omega)\mathbb{1}_{\omega\notin A}\|_{\ell_2} &\le T^{-\gamma}\|\sum_{|\alpha|_1=\gamma}\frac{\gamma!}{\alpha!}|\omega|^\alpha\,\bar{\mathcal{F}}[\tilde{k}_h](\omega)\mathbb{1}_{\omega\notin A}\|_{\ell_2}\\ &\le T^{-\gamma}\sum_{|\alpha|_1=\gamma}\frac{\gamma!}{\alpha!}\|\omega^\alpha\bar{\mathcal{F}}[\tilde{k}_h](\omega)\|_{\ell_2}\\ &= c_d\,T^{-\gamma}\sum_{|\alpha|_1=\gamma}\frac{\gamma!}{\alpha!}\|\frac{\partial^\alpha}{\partial x^\alpha}\tilde{k}_h(x)\|_2, \end{aligned} \tag{3.19}$$

where in the last line we used Parseval's identity. For any multi-index α with $|\alpha|_1=\gamma$,

$$\begin{aligned}\|\frac{\partial^\alpha}{\partial x^\alpha}\tilde{k}_h(x)\|_2 &= \|\sum_{\eta\preceq\alpha}\frac{\partial^\eta}{\partial x^\eta}k_h(x)\,\frac{\partial^{\alpha-\eta}}{\partial x^{\alpha-\eta}}\varphi(x)\|_2\\ &\le h^{-\frac{d}{2}-\gamma}\sum_{\eta\preceq\alpha}c_{d,\gamma}\,\|\frac{\partial^\eta}{\partial x^\eta}k(x)\|_2, \end{aligned} \tag{3.20}$$

where we used that the derivatives of φ are bounded. Next by Parseval's identity,

$$\|\frac{\partial^\eta}{\partial x^\eta}k(x)\|_2^2 = c_d\prod_{i=1}^d\|\omega_i^{\eta_i}\mathcal{F}[\kappa](\omega_i)\|_2^2. \tag{3.21}$$

For $0\le a\le\gamma$, we have

$$\int|\omega^a\mathcal{F}[\kappa](\omega)|^2\,\mathrm{d}\omega \le 2\|\kappa\|_1^2+\int_{|\omega|\ge1}|\omega^\gamma\mathcal{F}[\kappa](\omega)|^2\,\mathrm{d}\omega \le 2\|\kappa\|_1^2+L'. \tag{3.22}$$

By (3.19)–(3.22),

$$\|\bar{\mathcal{F}}[\tilde{k}_h](\omega)\mathbb{1}_{\omega\notin A}\|_{\ell_2}\le c_{d,\gamma,L'}\,T^{-\gamma}h^{-\frac{d}{2}-\gamma},$$

as desired.

□

Applying the previous lemma and linearity of the Fourier transform, we have the next corollary that gives an expansion for a general KDE on the smaller domain $[-\frac{1}{2},\frac{1}{2}]^d$.

Corollary 3.2. *Let $\tilde{g_S}$ denote the weighted KDE built from $\tilde{k}_h$ from Lemma 3.4 where $\{X_j : j\in S\}\subset[-\frac{1}{2},\frac{1}{2}]^d$ is an arbitrary set of points (not necessarily from a sample) and moreover $\kappa\in\mathcal{S}(\beta,L')$. Let $A=\{\omega\in\frac{\pi}{2}\mathbb{Z}^d:|\omega|_1\le T\}$, and define*

$$\tilde{g_S}^T(y)=\sum_{\omega\in A}\bar{\mathcal{F}}[\tilde{g_S}](\omega)e^{i\langle y,\omega\rangle}.$$

Then

$$\|(\tilde{g_S}-\tilde{g_S}^T)\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]^d}\|_2\le c_{d,\gamma,L'}\,T^{-\gamma}h^{-d/2-\gamma}L.$$

Now we have all the ingredients needed to prove Proposition 3.1.

Proof of Proposition 3.1 . Let

$$\tilde{f}(y) = \frac{1}{n}\sum_{j=1}^{n} \tilde{k}_h(X_j - y),$$

and

$$\tilde{g}_S(y) = \sum_{j\in S} \lambda_j \tilde{k}_h(X_j - y)$$

where the coreset $\{X_j : j \in S\}$ is constructed by Carathéodory's theorem as in Section 3.3.1 and $\tilde{k}_h$ is defined as in Lemma 3.3. Also consider their Fourier expansions $\tilde{f}^T$ and $\tilde{g}_S^T$ as defined in Corollary 3.2. Observe that, by construction of the Carathéodory coreset,

$$\tilde{f}^T(y) = \tilde{g}_S^T(y) \quad \forall y \in [-\frac{1}{2}, \frac{1}{2}]^d.$$

In what follows, $\|\cdot\|_2$ is computed on $[-\frac{1}{2}, \frac{1}{2}]^d$. By the triangle inequality,

$$\begin{aligned}
\|\hat{g}_S - \hat{f}\|_2 &\le \|\hat{g}_S - \tilde{g}_S\|_2 + \|\tilde{g}_S - \tilde{g}_S^T\|_2 + \|\tilde{g}_S^T - \tilde{f}^T\| \\
&\quad + \|\tilde{f}^T - \tilde{f}\|_2 + \|\tilde{f} - \hat{f}\|_2 \\
&\le c_{\beta,d}\, h^{-d+\nu} + c_{d,\gamma,L'}\, T^{-\gamma} h^{-d/2-\gamma} + 0 \\
&\quad + c_{d,\gamma,L'}\, T^{-\gamma} h^{-d/2-\gamma} + c_{\beta,d}\, h^{-d+\nu}
\end{aligned} \tag{3.23}$$

On the right-hand-side of the first line, the first and last terms are bounded via Lemma 3.4. The second and fourth terms are bounded via Lemma 3.5, and the third term is 0 by Carathéodory. By our choice of T and the decay properties of k, we have

$$\|\hat{g}_S - \hat{f}\|_2 \le c_{\beta,d,L}\, h^{\beta} \le c_{\beta,d,L}\, n^{-\beta/(2\beta+d)}.$$

The conclusion follows by the hypothesis on k, the previous display, and the triangle inequality. □

Proof of Theorem 3.2

Our goal is to apply Proposition 3.1 to k_s. First we show that the standard KDE built from k_s attains the minimax rate on $\mathcal{P}_{\mathcal{H}}(\beta, L)$. The Fourier condition

$$\text{ess sup}_{\omega\neq 0} \frac{|1 - \mathcal{F}[k_s](\omega)|}{|\omega|^{\alpha}} \le 1, \quad \forall \alpha \preceq \beta,$$

implies that k_s is a kernel of order β [Tsybakov, 2009, Definition 1.3]. Since $\mathcal{F}[k_s](0) = 1 = \int k_s(x)\,\mathrm{d}x$, it remains to show that the 'moments' of order at most β of k_s vanish. In fact all of the moments vanish. We have, expanding the exponential and using the

multinomial formula,

$$\begin{aligned}\psi(\omega) &= \mathcal{F}^{-1}[k_s](\omega) \\ &= \int k_s(x) e^{i\langle x,\omega\rangle} \mathrm{d}x \\ &= \sum_{t=0}^{\infty} \int k_s(x) \frac{(i\langle x,\omega\rangle)^t}{t!} \mathrm{d}x \\ &= \sum_{t=0}^{\infty} \sum_{|\alpha|_1=t} \frac{i^t}{\alpha!} w^\alpha \left\{ \int k_s(x) x^\alpha \mathrm{d}x \right\}.\end{aligned}$$

Since $\psi(\omega) \equiv 1$ in a neighborhood near the origin, it follows that all of the terms $\int k_s(x) x^\alpha \mathrm{d}x = 0$. Thus k_s is a kernel of order β for all $\beta \in \mathbb{Z}_{\geq 0}$, and the standard KDE on all of the dataset with bandwidth $h = n^{-1/(2\beta+d)}$ attains the rate of estimation $n^{-\beta/(2\beta+d)}$ over $\mathcal{P}_{\mathcal{H}}(\beta, L)$ [see e.g. Tsybakov, 2009, Theorem 1.2].

Next, $|\kappa_s(x)| \leq c_{\beta,d} |x|^\nu$ for $\nu = \lceil \beta + d \rceil$. This is because

$$x^\nu \kappa_s(x) = x^\nu \mathcal{F}[\psi](x) = \mathcal{F}\left[\frac{\mathrm{d}^\nu}{\mathrm{d}x^\nu} \psi \right](x) \leq \| \frac{\mathrm{d}^\nu}{\mathrm{d}x^\nu} \psi \|_1 \leq c_{\beta,d}.$$

Moreover for all $\gamma \in \mathbb{Z}_{>0}$, $\kappa_s \in \mathcal{S}(\gamma, c_\gamma)$. By Parseval's identity,

$$\| \frac{\mathrm{d}^\gamma}{\mathrm{d}x^\gamma} \kappa_s \|_2 = \frac{1}{\sqrt{2\pi}} \| \mathcal{F}[\frac{\mathrm{d}^\gamma}{\mathrm{d}x^\gamma} \kappa_s] \|_2 = \frac{1}{\sqrt{2\pi}} \| \omega^\gamma \psi(\omega) \|_2 \leq c_\gamma$$

because ψ has compact support [see e.g. Katznelson, 2004, Chapter VI].

All of the hypotheses of Proposition 3.1 are satisfied, so we apply the result with

$$\gamma = \frac{d}{2\varepsilon}$$

to derive Theorem 3.2.

Proof of Corollary 3.1

Recall from the proof of Theorem 3.2 that k_s is a kernel of all orders. By a standard bias-variance trade-off [see e.g. Tsybakov, 2009, Section 1.2], it holds for the KDE $\hat{f}$ with bandwidth h built on the entire dataset that

$$\mathbb{E}_f \| \hat{f} - f \|_2 \leq c_{\beta,d,L} \left(h^\beta + \frac{1}{\sqrt{n h^d}} \right). \tag{3.24}$$

Moreover, from (3.23) applied to k_s , setting $T = c_d m^{1/d}$, we get

$$\| \hat{g}_S - \hat{f} \|_2 \leq c_{\beta,d}\, h^\beta + c_{d,\gamma}\, m^{-\gamma/d} h^{-d/2-\gamma}. \tag{3.25}$$

Choosing

$$\gamma = (\beta + \frac{d}{2})(\frac{\beta}{d\varepsilon} - 1), \quad h = m^{-\frac{1}{d}+\frac{\varepsilon}{\beta}}$$

(assuming without loss of generality that $\varepsilon > 0$ is sufficiently small so that $\gamma > 0$), then the triangle inequality, (3.24), (3.25), and the upper bound on m yield the conclusion of Corollary 3.1.

Proof of Theorem 3.4

Let $\lambda = \lambda_1, \ldots, \lambda_m$ and let $\tilde{\lambda} = \tilde{\lambda}_1, \ldots, \tilde{\lambda}_m$. Observe that

$$\begin{aligned} \|\sum_{j\in S} \lambda_j k_h(X_j - y) - \sum_{j\in S} \tilde{\lambda}_j k_h(X_j - y)\|_2 &\leq \sum_{j\in S} \left|\lambda_j - \tilde{\lambda}_j\right| \|k_h(X_j - y)\|_2 \\ &\leq \left|\lambda - \tilde{\lambda}\right|_\infty n^2 h^{-d/2}. \end{aligned} \tag{3.26}$$

Using this we develop a decorated coreset-based estimator $\hat{f}_S$ (see Section 3.6.1) that approximates $\hat{g}_S$ well. Set $\delta = c_{\beta,d,L} n^{-4} h^{d/2}$ for $c_{\beta,d,L}$ sufficiently small and to be chosen later. Order the points of the coreset X_S according to their first coordinate. This gives rise to an ordering $\preceq$ so that

$$X'_1 \preceq X'_2 \preceq \cdots \preceq X'_m$$

denote the elements of X_S. Let $\lambda \in \mathbb{R}^m$ denote the correspondingly reordered collection of weights so that

$$\hat{g}_S(y) = \sum_{j=1}^{m} \lambda_j k_h(X'_j - y).$$

Construct a δ-net $\mathcal{N}_\delta$ with respect to the sup-norm $|\cdot|_\infty$ on the set $\{\nu \in \mathbb{R}^m : |\nu|_\infty \leq n^B\}$. Observe that

$$\log |\mathcal{N}_\delta| = \log(n^B \delta^{-1})^m = c_{\beta,d,L}\,(B + A) m \log n \tag{3.27}$$

Define R to be the smallest integer larger than the right-hand-side above. Then we can construct a surjection $\phi : \{0,1\}^R \to \mathcal{N}_\delta$. Note that ϕ is constructed before observing any data: it simply labels the elements of the δ-net $\mathcal{N}_\delta$ by strings of length R.

Given $\hat{g}_S(y) = \sum_{j\in S} \lambda_j k_h(X_j - y)$, define $\hat{f}_S$ as follows:

1. Let $\tilde{\lambda} \in \mathbb{R}^m$ denote the closest element in $\mathcal{N}_\delta$ to $\lambda \in \mathbb{R}^m$.

2. Choose $\sigma \in \{0,1\}^R$ such that $\phi(\sigma) = \tilde{\lambda}$.

3. Define the decorated coreset $Y_S = (X_S, \sigma)$.

4. Order the points of X_S by their first coordinate. Pair the i-th element of $\tilde{\lambda}$ with

the i-th element X'_i of X_S, and define

$$\hat{f}_S(y) = \sum_{j=1}^{m} \tilde{\lambda}_j k_h(X'_j - y)$$

We see that $\hat{f}_S$ is a decorated-coreset based estimator because in step 4 this estimator is constructed only by looking at the coreset X_S and the bit string σ. Moreover, by (3.26) and the setting of δ,

$$\|\hat{f}_S - \hat{g}_S\|_2 \leq c_{\beta,d,L}\, n^{-2}. \tag{3.28}$$

By Proposition 3.2 and our choice of R,

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f \|\hat{f}_S - f\|_2 \geq c_{\beta,d,L} \left((A+B)^{-\frac{\beta}{d}} (m \log n)^{-\frac{\beta}{d}} + n^{-\frac{\beta}{2\beta+d}} \right).$$

Applying the triangle inequality and (3.28) yields Theorem 3.4.

3.6.3 Proofs from Section 3.4

Notation: Given a set of points $X = x_1, \ldots, x_m \in [-1/2, 1/2]$ (not necessarily a sample), we let

$$\hat{f}_X(y) = \frac{1}{m} \sum_{i=1}^{m} k_h(X_i - y)$$

denote the uniformly weighted KDE on X.

Proof of Theorem 3.5

The proof of Theorem 3.5 follows directly from Propositions 3.3 and 3.4, which are presented in Sections 3.6.3 and 3.6.3, respectively.

Small bandwidth

First we show that uniformly weighted coreset KDEs on m points poorly approximate densities that are very close to 0 everywhere.

Lemma 3.6. *Let $\hat{f}_X$ denote a uniformly weighted coreset KDE built from an even kernel $k : \mathbb{R} \to \mathbb{R}$ with bandwidth h on m points $X = x_1, \ldots, x_m \in \mathbb{R}$. Suppose that quantiles $0 \leq q_1 \leq q_2$ satisfy*

$$\int_{-q_1}^{q_1} k(t)\mathrm{d}t \geq 0.9, \qquad \textit{and} \tag{3.29}$$

$$\int_{-q_2}^{q_2} k(t)\mathrm{d}t \geq 1 - \gamma. \tag{3.30}$$

Let U denote an interval $[0, u]$ where

$$u \geq 8q_2 h, \tag{3.31}$$

and suppose that $f : U \to \mathbb{R}$ satisfies

$$\frac{1}{100q_1 mh} \leq f(x) \leq \frac{45}{44} \cdot \frac{1}{100q_1 mh} \tag{3.32}$$

for all $x \in U$.

Then

$$\inf_{X:|X|=m} \|(\hat{f}_X - f)\mathbb{1}_U\|_1 \geq \frac{u}{440q_1 mh} - \gamma.$$

Proof. Let N denote the number of $x_i \in X$ such that $[x_i - q_1 h, x_i + q_1 h] \subset [0, u]$. The argument proceeds in two cases. With foresight, we set $\alpha = 1/(44q_1)$. Also let $C_1 = 1/(100q_1)$ and $C_2 = 45/(4400q_1)$.

Case 1: $N \geq \frac{\alpha u}{h}$. Then by (3.29) and the nonnegativity of k,

$$\|\hat{f}_X \mathbb{1}_U\|_1 \geq \frac{0.9N}{m} \geq \frac{0.9\alpha u}{mh}.$$

By (3.32),

$$\|f\|_1 \leq \frac{C_2 u}{mh}.$$

Hence,

$$\|(\hat{f}_X - f)\mathbb{1}_U\|_1 \geq \frac{u}{mh}(0.9\alpha - C_2) = C_2 \frac{u}{mh} = \frac{45}{4400} \cdot \frac{u}{q_1 mh}.$$

Thus Lemma 3.6 holds in Case 1 where $N \geq \alpha u/h$.

Case 2: $N \leq \frac{\alpha u}{h}$. Let

$$V = [2hq_2, u - 2hq_2] \setminus \bigcup_{j \in T} [x_j - q_1 h, x_j + q_1 h]$$

where T is the set of indices j so that $[x_j - q_1 h, x_j + q_1 h] \subset U$. Observe that if $j \notin T$, then by (3.30),

$$\int_V \frac{1}{h} k\left(\frac{x_j - t}{h}\right) \mathrm{d}t \leq \gamma.$$

If $j \in T$, then by (3.29),

$$\int_V \frac{1}{h} k\left(\frac{x_j - t}{h}\right) \mathrm{d}t \leq 0.1.$$

Thus,

$$\|\hat{f}_X \mathbb{1}_V\|_1 \leq \frac{0.1N}{m} + \gamma \leq \frac{\alpha 0.1u}{mh} + \gamma.$$

By the union bound, observe that the Lebesgue measure of V is at least

$$u - 4hq_2 - 2Nhq_1 \geq \frac{u}{2} - 2Nhq_1 \geq u(\frac{1}{2} - 2\alpha q_1).$$

Next, by (3.32),

$$\|f\mathbb{1}_V\|_1 \geq C_1 \frac{u}{mh}(\frac{1}{2} - 2\alpha q_1).$$

Therefore,

$$\|(\hat{f}_X - f)\mathbb{1}_U\|_1 \geq \frac{u}{mh}(C_1(1/2 - 2\alpha q_1) - 0.1\alpha) - \gamma = \frac{u}{440 q_1 mh} - \gamma. \tag{3.33}$$

□

Proposition 3.3. *Let $L > 2$. Let $0 < \delta < 1/3$ denote an absolute constant. Let $\hat{f}_X$ denote a uniformly weighted coreset KDE with bandwidth h built from a kernel k on $X = x_1, \ldots, x_m$. Suppose that $k(t) \leq \Delta|t|^{-(k+1)}$ for some absolute constants $\Delta > 0, k \geq 1$. If $h \leq n^{-1/3+\delta}$, then for*

$$m \leq \frac{n^{2/3-2\delta}}{\log n}$$

it holds that

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(1,L)} \inf_{X:|X|=m} \|\hat{f}_X - f\|_2 = \Omega\left(\frac{n^{-1/3+\delta}}{\log n}\right). \tag{3.34}$$

Proof. Let

$$f(t) = \lambda\left(e^{-1/t}\mathbb{1}(t \in [-1/2, 0]) + e^{-1/(1-t)}\mathbb{1}(t \in [0, 1/2])\right),$$

where λ is a normalizing constant so that $\int f = 1$. Observe that $f \in \mathcal{P}_{\mathcal{H}}(1, L)$. Our first goal is to show that

$$\|\hat{f}_X - f\|_1 = \Omega\left(\frac{1}{mh\log^2(mh)}\right)$$

holds for all $\tau/h \leq m \leq h^{-2}$ and for all $h \leq n^{-1/3+\delta}$, where τ is an absolute constant to be determined.

We apply Lemma 3.6 to the density f. Let q_1 be defined as in Lemma 3.6, and set $C_1 = 1/(100q_1)$ and $C_2 = 45/(4400q_1)$. Set $\tau = 10C_2/\lambda$. Let

$$U = [t_1, t_2] := \left[\frac{1}{\log(\lambda mh/C_1)}, \frac{1}{\log(\lambda mh/C_2)}\right].$$

The function $f|_U$ satisfies the bounds (3.32) from Lemma 3.6. Observe that the length of U is

$$u := t_2 - t_1 = \Omega(\frac{1}{\log^2(mh)}).$$

We set the parameter γ in Lemma 3.6 to be

$$\gamma = \frac{1}{800 q_1 mh \log^2(mh)}.$$

By the decay assumption on k, we may set

$$q_2 := \left(\frac{2\Delta}{k\gamma}\right)^{1/k}.$$

Therefore,

$$\begin{aligned} u - 8q_2 h &= \Omega(\frac{1}{\log^2(mh)}) - 8h\left(\frac{2\Delta}{k\gamma}\right)^{1/k} && (3.35)\\ &= \Omega(\frac{1}{\log^2(mh)}) - O(h(mh\log^2(mh))^{1/k}) && (3.36)\\ &= \Omega(\frac{1}{\log^2(h^{-1})}) - O(h^{1-1/k}\log^2(h^{-1})) > 0 && (3.37)\end{aligned}$$

for n sufficiently large, because we assume $\tau/h \le m \le h^{-2}$, $h \le n^{-1/3+\delta}$, and $k > 1$. Hence, condition (3.31) is satisfied for m, h in the specified range, so we apply Cauchy–Schwarz and Lemma 3.6 to conclude that for all $\tau/h \le m \le h^{-2}$ and $h \le n^{-1/3+\delta}$,

$$\|\hat{f}_X - f\|_2 \ge \|\hat{f}_X - f\|_1 = \Omega\left(\frac{1}{mh\log^2(mh)}\right) = \Omega\left(\frac{1}{mh\log^2(h^{-1})}\right). \tag{3.38}$$

Suppose first that $\log^2(1/h) \ge n^{1/3-\delta}$. Then clearly the right-hand side of (3.38) is $\Omega(1)$ for $m \le n$. Otherwise, we have for all $h \le n^{-1/3+\delta}$ that if m is in the range

$$\frac{\tau}{h} \le m \le \min\left(\frac{n^{1/3-\delta}\log n}{h\log^2(1/h)},\ h^{-2}\right) =: N_h,$$

then (3.38) implies

$$\|\hat{f}_X - f\|_2 = \Omega\left(\frac{n^{-1/3+\delta}}{\log n}\right). \tag{3.39}$$

Moreover, a uniformly weighted coreset KDE on $m = O(1/h)$ points can be expressed as a uniformly weighted coreset KDE on $\Omega(1/h)$ points by setting some of the x_i's to be duplicates. Hence (3.39) holds for all $1 \le m \le N_h$. Since N_h is a decreasing function of h, it follows that (3.39) holds for all $m \le n^{2/3-2\delta}/\log n$ and $h \le n^{-1/3+\delta}$, as desired.

□

Large bandwidth

Lemma 3.7. *Let $\varepsilon = \varepsilon(n) > 0$, and let $\hat{f}_X$ denote the uniformly weighted coreset KDE on X with bandwidth h. Suppose that $\phi : \mathbb{R} \to \mathbb{R}$ is an odd $\mathcal{C}^\infty$ function supported on $[-1/4, 1/4]$. Let $f(t) : [-1/2, 1/2] \to \mathbb{R}_{\geq 0}$ denote the density*

$$f(t) = \frac{12}{11}(1 - t^2) + \varepsilon\phi(t)\cos\left(\frac{t}{\varepsilon}\right).$$

Then

$$\|\hat{f}_X - f\|_2^2 \geq \frac{1}{2}\varepsilon^2\left(\|\phi\|_2^2 - \left|\mathcal{F}[\phi^2](2\varepsilon^{-1})\right|\right) - \|\phi\|_1 \sup_{|\omega| \geq h\varepsilon^{-1}/2} |\mathcal{F}[k](\omega)| - 2\varepsilon \int_{|\omega| \geq \varepsilon^{-1}/2} |\mathcal{F}[\phi](\omega)|\,\mathrm{d}\omega. \quad (3.40)$$

Proof. Let $g(t) = (12/11)(1 - t^2)$ and $\psi(t) = \varepsilon\phi(t)\cos(t/\varepsilon)$. Observe that

$$\begin{aligned}\|\hat{f}_X - f\|_2^2 &\geq \|g - f\|_2^2 - 2\langle \hat{f}_X, g - f\rangle + 2\langle g, \psi(t)\rangle \\ &= \|g - f\|_2^2 - 2\langle \hat{f}_X, g - f\rangle \end{aligned} \quad (3.41)$$

because $g(t)\psi(t)$ is an odd function. Next, using $\cos^2(\theta) = (1/2)(\cos(2\theta) + 1)$,

$$\begin{aligned}\|g - f\|_2^2 &= \varepsilon^2 \int_{-1/2}^{1/2} \cos^2(t/\varepsilon)\phi^2(t)\mathrm{d}t \\ &\geq \frac{\varepsilon^2}{2}\|\phi\|_2^2 - \frac{\varepsilon^2}{2}\left|\mathcal{F}[\phi^2](2\varepsilon^{-1})\right|. \end{aligned} \quad (3.42)$$

By the triangle inequality and Parseval's formula,

$$\frac{\left|\langle \hat{f}_X, g - f\rangle\right|}{\varepsilon} \leq \Bigg(\underbrace{\int_{|\omega| \leq h\varepsilon^{-1}/2}}_{=:A} + \underbrace{\int_{|\omega| \geq h\varepsilon^{-1}/2}}_{=:B}\Bigg)\left|\mathcal{F}[k]\left(-\frac{h}{\varepsilon} - \omega\right)\frac{1}{h}\mathcal{F}[\phi]\left(-\frac{\omega}{h}\right)\right|\mathrm{d}\omega.$$

Moreover,

$$A \leq \frac{1}{2\varepsilon}\|\phi\|_1 \cdot \sup_{|\omega| \geq h\varepsilon^{-1}/2} |\mathcal{F}[k](\omega)|, \quad (3.43)$$

$$B \leq \|k\|_1 \cdot \int_{|\omega| > \varepsilon^{-1}/2} |\mathcal{F}[\phi](\omega)|\,\mathrm{d}\omega. \quad (3.44)$$

Then (3.40) follows from $\|k\|_1 = 1$ and equations (3.41), (3.42), (3.43), and (3.44). □

Proposition 3.4. *Let $\varepsilon = n^{-1/3+\gamma}$ for some absolute constant $\gamma > 0$. Let $\hat{f}_X$ denote a uniformly weighted coreset KDE with bandwidth h built from a kernel k on $X = x_1, \ldots, x_m$. Suppose that $|\mathcal{F}[k](\omega)| \leq |\omega|^{-\ell}$. If $h \geq c\varepsilon^{1-2/\ell} = cn^{(-1/3+\gamma)(1-2/\ell)}$ for c*

sufficiently large, then for all m it holds that

$$\sup_{f\in\mathcal{P}_{\mathcal{H}}(\beta,L)} \inf_{X:|X|=m} \|\hat{f}_X - f\|_2 = \Omega(\varepsilon) = \Omega\left(n^{-1/3+\gamma}\right) \tag{3.45}$$

Proof. The proof is a direct application of Lemma 3.7. Let $f(t) = g(t)+\varepsilon\phi(t)\cos(t/\varepsilon)$, where we set

$$\phi(t) = -e^{\frac{1}{x(x+1/4)}}\mathbb{1}(x\in[-1/4,0]) + e^{-\frac{1}{x(x-1/4)}}\mathbb{1}(x\in[0,1/4]).$$

Observe that ϕ is odd and $\phi\in\mathcal{C}^\infty$. Thus, $\phi^2\in\mathcal{C}^\infty$, so by the Riemann–Lebesgue lemma [see e.g. Katznelson, 2004, Chapter VI], $\mathcal{F}[\phi^2](\varepsilon^{-1}) \le 10\varepsilon$. Using a similar argument and noting that $\mathcal{F}[\phi](\omega) = \omega^{-2}\mathcal{F}[\phi'](\omega) \le 10\omega^{-3}$, we obtain

$$\int_{|\omega|\ge 2\varepsilon^{-1}} |\mathcal{F}[\phi](\omega)|\,\mathrm{d}\omega \le 100\varepsilon^2.$$

Also $\|\phi\|_2 \ge c'$ for a small absolute constant, and $\|\phi\|_1 \le 2$.

Thus Lemma 3.7, the hypothesis on k, and $h \ge c'\varepsilon^{1-2/\ell}$ imply that

$$\|\hat{f}_X - f\|_2^2 \ge \frac{c^2}{2}\varepsilon^2 - 2\left(\frac{\varepsilon}{h}\right)^\ell - 200\varepsilon^3 = \Omega(\varepsilon^2).$$

Since $f\in\mathcal{P}_{\mathcal{H}}(1,L)$, the statement of the lemma follows. □

Proof of Theorem 3.6

We follow a similar strategy to the proof of Theorem 3.5 by handling the cases of small and large bandwidth separately.

Let $q_1 = q_1(k) > 0$ be the minimum number such that $\int_{|t|>q_1} k(t)dt \le 0.1$. By the assumption in the theorem, there exists $a > 0$ such that

$$\int_{|t|>s} k(t)dt \le \frac{1}{a}\exp(-as), \quad \forall s \ge 0.$$

Note that we can set $L_\beta^{(1)}$ large such that for any $\delta\in[0,1]$, there exists $f\in\mathcal{P}_{\mathcal{H}}(\beta,L_\beta^{(1)})$ such that $f(x) = \delta$ for $x\in[0,1/2]$. We first show that for any given m and h, we have

$$\inf_{S:|S|\le m} \sup_{f\in\mathcal{P}_{\mathcal{H}}(\beta,L_\beta^{(1)})} \mathbb{E}\|\hat{f}_S^{\mathsf{unif}} - f\|_1 \ge 0.2\left(1\wedge\frac{1}{100q_1mh}\right)1\left\{h\le\frac{0.02a}{\log\left(\frac{mq_1}{0.001a}\vee\frac{10}{a}\right)}\wedge 1\right\}. \tag{3.46}$$

Let f be an arbitrary function in $f\in\mathcal{P}_{\mathcal{H}}(\beta,L_\beta^{(1)})$ such that

$$f(x) = 1\wedge\frac{1}{100q_1mh}, \quad \forall x\in[0,1/2].$$

Let T be the set of $i \in S$ for which $x_i \in [q_1 h, 1/2 - q_1 h]$.

Case 1: $|T| \geq m\left(1 \wedge \frac{1}{100 q_1 m h}\right)$. Since $k \geq 0$, we have

$$\|\hat{f}_X 1_{[0,1/2]}\|_1 \geq \frac{0.9|T|}{m} \geq 0.9\left(1 \wedge \frac{1}{100 q_1 m h}\right).$$

On the other hand,

$$\|f 1_{[0,1/2]}\|_1 \leq \frac{1}{2}\left(1 \wedge \frac{1}{100 q_1 m h}\right),$$

therefore,

$$\|(\hat{f}_X - f) 1_{[0,1/2]}\|_1 \geq 0.4\left(1 \wedge \frac{1}{100 q_1 m h}\right).$$

Case 2: $|T| < m\left(1 \wedge \frac{1}{100 q_1 m h}\right)$. Define

$$\gamma := 0.1\left(1 \wedge \frac{1}{100 q_1 m h}\right)$$

and

$$q_2 := \frac{0.02}{h}.$$

Note that to verify (3.46) we only need to consider the event of $h \leq \frac{0.02a}{\log\left(\frac{m q_1}{0.001a} \vee \frac{10}{a}\right)} \wedge 1$, in which case

$$\begin{aligned}
\int_{|t|>q_2} k(t)dt &\leq \frac{1}{a}\exp(-a q_2) \\
&\leq \frac{1}{a} \cdot \left(\frac{0.001a}{m q_1} \wedge 0.1a\right) \\
&\leq \frac{1}{a} \cdot \left(\frac{0.001a}{q_1 m h} \wedge 0.1a\right) \\
&= 0.1(1 \wedge \frac{1}{100 q_1 m h}) \\
&= \gamma.
\end{aligned}$$

Moreover since $\gamma \leq 0.1$ we see that $q_2 \geq q_1$. Now define

$$V := [2hq_2, 1/2 - 2hq_2] \setminus \bigcup_{j \in T} [x_j - q_1 h, x_j - q_1 h].$$

Then for $j \notin T$, we have

$$\int_V \frac{1}{h} k\left(\frac{x_j - t}{h}\right) dt \leq \gamma$$

while for $j \in T$ we have

$$\int_V \frac{1}{h} k\left(\frac{x_j - t}{h}\right) dt \leq 0.1.$$

Thus,

$$\|\hat{f}_X 1_V\|_1 \leq \frac{0.1|T|}{m} + \gamma \leq 0.2\left(1 \wedge \frac{1}{100 q_1 m h}\right).$$

On the other hand, by the union bound we see that the Lebesgue measure of V is at least

$$\frac{1}{2} - 4q_2 h - 2q_1 h|T| \geq 0.5 - 4q_2 h - 0.02 \geq 0.4$$

where we used the fact that $q_2 h = 0.02$. Then

$$\|f 1_V\|_1 \geq 0.4\left(1 \wedge \frac{1}{100 q_1 m h}\right)$$

and hence

$$\|(\hat{f}_X - f)1_{[0,1/2]}\|_1 \geq \|(\hat{f}_X - f)1_V\|_1 \geq 0.2\left(1 \wedge \frac{1}{100 q_1 m h}\right).$$

This concludes the proof of (3.46).

The second step is to show that for given m and h, we have

$$\inf_{S:|S|\leq m} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta, L)} \mathbb{E}\|\hat{f}_S^{\mathsf{unif}} - f\|_1 \geq \frac{1}{4}\left(\frac{b(h \wedge 1)}{\log m}\right)^{\beta} - \frac{1}{bm^2} \tag{3.47}$$

sufficiently large m and L to be determined later, and $0 < b < \infty$ is such that

$$\mathcal{F}[k](\omega) \leq \frac{1}{b}\exp(-b\omega), \quad \forall \omega \in \mathbb{R}$$

whose existence is guaranteed by the assumption of the theorem. Let ϕ be a smooth, even, nonnegative function supported on $[-1/2, 1/2]$ satisfying $\int_{[-1/2,1/2]} \phi = 1$. Define

$$f_\epsilon(t) := \phi(t)\left(c_\epsilon + \epsilon^\beta \sin\frac{t}{\epsilon}\right)$$

where $c_\epsilon > 0$ is chosen so that $\int_{[-1/2,1/2]} f_\epsilon = 1$. Then $\lim_{\epsilon \to 0} c_\epsilon = 1$, and in particular $f_\epsilon \geq 0$ when $\epsilon < \epsilon(\phi, \beta)$ for some $\epsilon(\phi, \beta)$. Moreover we can find $L_\beta^{(2)} < \infty$ such that

$f_\epsilon \in \mathcal{P}_{\mathcal{H}}(\beta, L_\beta^{(2)})$ for all $\epsilon < \epsilon(\phi, \beta)$. Now

$$\begin{aligned}
\|f_\epsilon - \hat{f}_X\|_1 &\geq |\mathcal{F}[f_\epsilon](1/\epsilon) - \mathcal{F}[\hat{f}_X](1/\epsilon)| \\
&\geq \left| \int_{[-1/2,1/2]} f_\epsilon(t) e^{-it/\epsilon} dt \right| - \left| \mathcal{F}[k](\frac{h}{\epsilon}) \right| \\
&\geq \left| \int_{[-1/2,1/2]} f_\epsilon(t) \sin \frac{t}{\epsilon} dt \right| - \left| \mathcal{F}[k](\frac{h}{\epsilon}) \right| \\
&= \epsilon^\beta \left| \int_{[-1/2,1/2]} \phi(t) \sin^2 \frac{t}{\epsilon} dt \right| - \left| \mathcal{F}[k](\frac{h}{\epsilon}) \right| \qquad (3.48)
\end{aligned}$$

where (3.48) used the fact that ϕ is even. Since $\lim_{\epsilon \to 0} \int_{[-1/2,1/2]} \phi(t) \sin^2 \frac{t}{\epsilon} dt = \frac{1}{2}$, there exists $\epsilon'(\phi)$ such that

$$\int_{[-1/2,1/2]} \phi(t) \sin^2 \frac{t}{\epsilon} dt \geq \frac{1}{4}$$

for any $\epsilon \leq \epsilon'(\phi)$. Now define

$$\epsilon''(h, m) = \frac{b(h \wedge 1)}{2 \log m}.$$

There exists $m(\phi, \beta, b) < \infty$ such that $\sup_{h>0} \epsilon''(h, m) < \epsilon(\phi, \beta) \wedge \epsilon'(\phi)$ whenever $m \geq m(\phi, \beta, b)$. With the choice of $\epsilon = \epsilon''(h, m)$, we can continue lower bounding (3.48) as (for $m \geq m(\phi, \beta, b)$):

$$\frac{1}{4} \left(\frac{b(h \wedge 1)}{\log m} \right)^\beta - \frac{1}{bm^2}.$$

Finally, we collect the results for step 1 and step 2. First observe that the main term in the risk in step 1 can be simplified as

$$\begin{aligned}
&\left(1 \wedge \frac{1}{100 q_1 m h} \right) 1 \left\{ h \leq \frac{0.02a}{\log \left(\frac{m q_1}{0.001a} \vee \frac{10}{a} \right)} \wedge 1 \right\} \\
&= \frac{1}{100 q_1 m h} \wedge 1 \{\mathcal{A}\} \qquad (3.49)
\end{aligned}$$

where $\mathcal{A}$ denotes the event in the left side of (3.49).

Thus up to multiplicative constant depending on k, β, we can lower bound the risk by taking the max of the risks in the two steps:

$$\left(\frac{1}{mh} \wedge 1\{\mathcal{A}\} \right) \vee \left(\left(\frac{b(h \wedge 1)}{\log m} \right)^\beta - \frac{1}{bm^2} \right) \qquad (3.50)$$

whenever $L \geq L_\beta := L_\beta^{(1)} \vee L_\beta^{(2)}$. We can use the distributive law to open up the

parentheses in (3.50). By checking the $h > m^{-\frac{1}{\beta}}$ and $h \le m^{-\frac{1}{\beta}}$ cases respectively, it is easy to verify that

$$\frac{1}{mh} \vee \left(\left(\frac{b(h \wedge 1)}{\log m} \right)^{\beta} - \frac{1}{bm^2} \right) = \Omega \left(\frac{m^{-\frac{\beta}{\beta+1}}}{\log^{\beta} m} \right).$$

Next, if $\mathcal{A}$ is true, we evidently have

$$1\{\mathcal{A}\} \vee \left(\left(\frac{b(h \wedge 1)}{\log m} \right)^{\beta} - \frac{1}{bm^2} \right) = 1 = \Omega \left(\frac{m^{-\frac{\beta}{\beta+1}}}{\log^{\beta} m} \right).$$

If $\mathcal{A}$ is not true, then $h > \frac{0.02a}{\log\left(\frac{mq_1}{0.001a} \vee \frac{10}{a}\right)} \wedge 1$, and we have

$$\begin{aligned} 1\{\mathcal{A}\} \vee \left(\left(\frac{b(h \wedge 1)}{\log m} \right)^{\beta} - \frac{1}{bm^2} \right) &= \left(\left(\frac{b(h \wedge 1)}{\log m} \right)^{\beta} - \frac{1}{bm^2} \right) \\ &= \Omega \left(\log^{-2\beta} m \right) \\ &= \Omega \left(\frac{m^{-\frac{\beta}{\beta+1}}}{\log^{\beta} m} \right). \end{aligned}$$

In either case the risk with respect to L_1 is $\Omega \left(\frac{m^{-\frac{\beta}{\beta+1}}}{\log^{\beta} m} \right)$. It remains to convert this to a lower bound in L^2. We consider two cases. First note that by the fast decay condition on the Fourier transform, $k \in \mathcal{C}^1$. Let $B = B_k$ denote a constant such that

$$\sup_{x \in [-1/2, 1/2]} |k'(x)| \le B. \tag{3.51}$$

Set $\Delta = B^{1/2} \vee k(0) \vee 1$.

Case 1: $h \le \Delta$. Let $U = \{|y| \ge \frac{1}{2} + c_{\beta,\Delta,a} \log m\}$, and let $U^c = \mathbb{R} \backslash U$. If $h \le \Delta$, then because $X_i \in [-1/2, 1/2]$ and by the exponential decay of k,

$$\|\hat{f}_X(y) \mathbb{1}_U\|_1 \le m^{-2}$$

for $c_{\beta,\Delta,a}$ sufficiently large. Thus by Cauchy–Schwarz,

$$\begin{aligned} \|(\hat{f}_X - f)\mathbb{1}_{U^c}\|_2 &\ge c'_{\beta,\Delta,a} (\log m)^{-1/2} \|(\hat{f}_X - f)\mathbb{1}_{U^c}\|_2 \\ &= c'_{\beta,\Delta,a} (\log m)^{-1/2} \left(\|(\hat{f}_X - f)\|_1 - \|(\hat{f}_X - f)\mathbb{1}_U\|_1 \right) \\ &\ge c'_{\beta,\Delta,a} (\log m)^{-1/2} \left(c_{\beta,k} \left(\frac{m^{-\frac{\beta}{\beta+1}}}{\log^{\beta} m} \right) - m^{-2} \right) \\ &= \Omega \left(\frac{m^{-\frac{\beta}{\beta+1}}}{\log^{\beta+\frac{1}{2}} m} \right) \end{aligned}$$

Case 2: $h \geq \Delta$. In this case, $k(X_i - y)$ is nearly constant for all i. By (3.51) and Taylor's theorem,

$$\left| k(0) - k\left(\frac{X_i - y}{h}\right) \right| \leq 2B$$

for all $y \in [-1/2, 1/2]$ and for all i. Hence, for all $y \in [-1/2, 1/2]$, using $h \geq \Delta$,

$$\hat{f}_X(y) = \frac{1}{mh} \sum_{i=1}^{m} k\left(\frac{X_i - y}{h}\right) \leq \frac{1}{h}(k(0) + 2B) \leq 3.$$

For L_β large enough, we see that for the function $f \in \mathcal{P}_{\mathcal{H}}(\beta, L_\beta)$ with $f|_{[0, \frac{1}{100}]} \equiv 4$,

$$\|\hat{f}_X - f\|_2 \geq \|(\hat{f}_X - f)\mathbb{1}_{[0, \frac{1}{100}]}\|_1 = \Omega(1).$$

3.6.4 Proofs from Section 3.5

Proof of Theorem 3.7

Here we adapt the results in Section 2 of Phillips and Tai [2018b] to our setting where the bandwidth $h = n^{-1/(2\beta+d)}$ is shrinking. Using their notation, we define $K_s(x, y) = k_s\left(\frac{x-y}{h}\right)$ and study the kernel discrepancy of the kernel K_s. First we verify the assumptions on the kernel (bounded influence, Lipschitz, and positive semidefiniteness) needed to apply their results.

First, the kernel K_s is *bounded influence* [see Phillips and Tai, 2018b, Section 2] with constant $c_K = 2$ and $\delta = n^{-1}$, which means that

$$|K_s(x, y)| \leq \frac{1}{n}$$

if $|x - y|_\infty \geq n^2$. This follows from the fast decay of κ_s.

Note that if x and y differ on a single coordinate i, then

$$|k_s(x) - k_s(y)| \leq \left| c(x_i - y_i) \prod_{j \neq i} \kappa_s(x_j) \right| \leq c\,|x_i - y_i|$$

because $|\kappa_s(x)| \leq \|\psi\|_1$ for all x and the function κ_s is c-Lipschitz for some constant c. Hence by the triangle and Cauchy–Schwarz inequalities, the function k_s is Lipschitz:

$$|k_s(x) - k_s(y)| \leq dc_k\,|x - y|_1 \leq d^{3/2} c_\kappa\,|x - y|_2\,.$$

Therefore the kernel $K_s(x, y)$ is *Lipschitz* [see Phillips and Tai, 2018b] with constant $C_K = d^{3/2} c_\kappa h^{-1}$. Moreover, the kernel K_s is *positive semidefinite* because the Fourier transform of κ_s is nonnegative.

Given the shrinking bandwidth $h = n^{-1/(2\beta+d)}$, we slightly modify the lattice used in Phillips and Tai [2018b, Lemma 1]. Define the lattice

$$\mathcal{L} = \{\, (i_1\delta, i_2\delta, \ldots, i_d\delta) \mid i_j \in \mathbb{Z} \},$$

where

$$\delta = \frac{1}{c_\kappa d^2 n h^{-1}}.$$

The calculation at the top of page 6 of Phillips and Tai [2018b, Lemma 1] yields

$$\mathsf{disc}(X, \chi, y) := \left| \sum_{i=1}^{n} \chi(X_i) K_s(X_i, y) \right| \leq \left| \sum_{i=1}^{n} \chi(X_i) K_s(X_i, y_0) \right| + 1$$

where y_0 is the closest point to y in the lattice $\mathcal{L}$, and χ assigns either $+1$ or -1 to each element of $X = X_1, \ldots, X_n$. Moreover, with the bounded influence of K_s, if

$$\min_i |y - X_i|_\infty \geq n^2,$$

then

$$\mathsf{disc}(X, \chi, y) = \left| \sum_{i=1}^{n} \chi(X_i) K_s(X_i, y) \right| \leq 1.$$

On defining

$$\mathcal{L}_X = \mathcal{L} \cap \{y : \min_i |y - X_i|_\infty \leq n^2\},$$

we see that

$$\max_{y \in \mathbb{R}^d} \mathsf{disc}(X, \chi, y) \leq \max_{y \in \mathcal{L}_X} \mathsf{disc}(X, \chi, y) + 1$$

for all signings $\chi : X \to \{-1, +1\}$. This is precisely the conclusion of Phillips and Tai [2018b, Lemma 1].

This established, the positive definiteness and bounded diagonal entries of K_s and Phillips and Tai [2018b, Lemmas 2 and 3] imply that

$$\mathsf{disc}_{K_s} = O(\sqrt{d \log n}).$$

Given $\varepsilon > 0$, the halving algorithm can be applied to K_s as in Phillips and Tai [2018b, Corollary 5] to yield a coreset X_S of size $m = O(\varepsilon^{-1}\sqrt{d \log \varepsilon^{-1}})$ such that

$$\| \frac{1}{n} \sum_{j=1}^{n} K_s(X_j, y) - \frac{1}{m} \sum_{j \in S} K_s(X_j, y) \|_\infty \leq \varepsilon.$$

Rescaling by h^{-d}, we have

$$\| \hat{f} - \hat{f}_S^{\mathsf{unif}} \|_\infty = \| \frac{1}{n} \sum_{j=1}^{n} k_s(X_j, y) - \frac{1}{m} \sum_{j \in S} k_s(X_j, y) \|_\infty \leq \varepsilon h^{-d}.$$

Recall from Section 3.6.2 that $\hat{f}$ attains the minimax rate of estimation on $\mathcal{P}_{\mathcal{H}}(\beta, L)$. Thus setting $\varepsilon = h^d n^{-\beta/(2\beta+d)}$ we get a coreset of size $\tilde{O}_d(n^{\frac{\beta+d}{2\beta+d}})$ that attains the minimax rate $c_{\beta,d,L}\, n^{-\beta/(2\beta+d)}$, as desired. Moreover, by the results of Phillips and Tai [2018b], this coreset can be constructed in polynomial time.

Chapter 4

Interpolation of density estimators

4.1 Introduction

Density estimation is the task of estimating an unknown density f given an i.i.d. sample $X_1, \ldots, X_n \sim \mathbb{P}_f$, where $\mathbb{P}_f$ is the probability distribution associated to f. A popular choice of density estimator is the kernel density estimator (KDE)

$$\hat{f}(y) := \frac{1}{nh^d} \sum_{j=1}^{n} K\left(\frac{X_i - y}{h}\right). \tag{4.1}$$

With proper setting of the bandwidth parameter h and choice of kernel K, the KDE $\hat{f}$ is a minimax optimal estimator over the L-Hölder smooth densities $\mathcal{P}_{\mathcal{H}}(\beta, L)$ of order β [see e.g. Tsybakov, 2009, Theorem 1.2]:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f \|\hat{f} - f\|_2 = \Theta_{\beta,d,L}(n^{-\frac{\beta}{2\beta+d}}). \tag{4.2}$$

Despite its statistical utility, the KDE (4.1) has the computational drawback that it naively requires $\Omega(n)$ time to evaluate a query. The problem of improving on these computational aspects has thus received a lot of attention.

The fast evaluation of kernel density estimators has been well-studied including approaches based on the fast Gauss transform [Greengard and Strain, 1991], hierarchical space decompositions [Greengard and Rokhlin, 1987], locality sensitive hashing [Charikar and Siminelakis, 2017, Backurs et al., 2018, Siminelakis et al., 2019, Backurs et al., 2019], and binning [Scott and Sheather, 1985], as well as interpolation [Jones, 1989, Kogure, 1998], our main technique in this work. Typically these techniques carefully leverage the structure of the kernel under consideration, and many of them operate in a worst-case framework over the dataset. In this work, we consider the problem of fast evaluation of a density estimator $\hat{f}$ in a statistical setting where $\hat{f}$ gives a good pointwise approximation to an unknown density $f : [0,1]^d \to \mathbb{R}$ that lies in a Hölder class of smooth functions. We show that a pointwise approximation guarantee alone, without assuming any specific structure of the estimator $\hat{f}$, is enough to construct a new estimator $\tilde{f}$ that can be stored and queried cheaply, and whose approximation error is similar to that of the original estimator. Our approach

is based on a multivariate polynomial interpolation scheme of Nicolaides [1972] [see also Chung and Yao, 1977] and provides an explicit formula for $\tilde{f}$ in terms of some judiciously chosen queries of the original estimator.

4.1.1 Related work

Motivated by multi-body problems, Greengard and Strain [1991] developed the fast Gauss transform to rapidly evaluate sums of the form (4.1) when $K(x) = \exp(-|x|_2^2)$ is the Gaussian kernel. Their work is posed a worst-case batch setting where $\hat{f}$ is to be evaluated at m points $y_1, \ldots, y_m$ specified in advance and the locations $X_1, \ldots, X_n$ lie in a box. Their techniques use hierarchical space decompositions and series expansions to show that (4.1) may be evaluated at $y_1, \ldots, y_m$ with precision ε in time $h^{-d}(\log \frac{1}{\varepsilon})^d(n+m)$. These results apply to any kernel that has a rapidly converging Hermite expansion [see also Greengard and Rokhlin, 1987]. There are also follow up works on the improved fast Gauss transform and tree-based methods that use related ideas [Yang et al., 2003, Lee et al., 2006].

More recently, several works [Charikar and Siminelakis, 2017, Backurs et al., 2018, Siminelakis et al., 2019, Backurs et al., 2019, Coleman and Shrivastava, 2020] are devoted to the problem of fast evaluation of (4.1) in high dimension using locality sensitive hashing. In these works, the dataset is carefully reweighted for importance sampling such that a randomly drawn datapoint X_r's corresponding kernel value $K(X_r - y)$ gives a good approximation to $\hat{f}(y)$. This sampling procedure can be executed efficiently using hashing-based methods. For example, Backurs et al. [2019] show that for the Laplace and Exponential kernels with bandwidth $h = 1$, e.g., the value $\hat{f}(y)$ can be computed with multiplicative $1 \pm \varepsilon$ error in time $O(\frac{d}{\sqrt{\tau}\varepsilon^2})$ even in worst case over the dataset, where τ is a uniform lower bound on the KDE.

Another effective approach to this problem in high dimensions is through coresets [Agarwal et al., 2005, Clarkson, 2010, Phillips and Tai, 2018a,b]. A coreset is a representative subset $\{X_i\}_{i \in S}$ of a dataset such that

$$\hat{f}(y) \approx \frac{1}{nh^d} \sum_{i \in S} K\left(\frac{X_i - y}{h}\right).$$

When $h = O(1)$, for example, the results of Phillips and Tai [2018b] give a polynomial time algorithm in n, d such that the coreset KDE yields an additive ε approximation to $\hat{f}$ using a coreset of size $\tilde{O}(\frac{\sqrt{d}}{\varepsilon})$. Their results hold in worst case over the dataset and for a variety of popular kernels. The methods of Phillips and Tai [2018b] are powered by state-of-the-art algorithms from discrepancy theory [Bansal et al., 2018] [see Matoušek, 1999, Chazelle, 2000, for a comprehensive exposition on discrepancy].

Our approach is most closely related to prior work on the interpolation of kernel density estimators due to Jones [1989] and Kogure [1998]. Motivated by visualization and computational aspects, Jones [1989] studies binned and piecewise linearly interpolated univariate kernel density estimators and provides precise bounds on the mean-integrated squared error. Kogure [1998] extends this work and constructs higher order piecewise polynomial interpolants of multivariate kernel density estimators, and

shows that for very smooth densities, this procedure improves the mean-integrated squared error. In addition, we note the recent work of Belkin et al. [2019], Liang et al. [2020] demonstrating the perhaps surprising effectiveness of interpolation in nonparametric regression. We also remark that nonparametric estimators based on multivariate piecewise polynomials are well-studied in statistics [see e.g. Györfi et al., 2006, Chapter 10], and there is a line of related literature in computer science on fast estimation of univariate densities that are well-approximated by piecewise polynomials [Chan et al., 2014, Acharya et al., 2017, Hao et al., 2020].

Our work differs from Kogure [1998] in a few important respects. We do not assume $\hat{f}$ to be a KDE in the first place, but rather give a general method for effectively interpolating a minimax density estimator. Also, our results hold for the entire range of the smoothness parameter β and dimension d, while Kogure [1998] requires the density to be at least qd times differentiable when interpolating KDEs with kernels of order q [Tsybakov, 2009, Definition 1.3]. On the other hand, our method increases the mean squared by a multiplicative factor $\tilde{O}(c_{\beta,d})$, while Kogure's approach improves the mean squared error (though our focus here is the L^∞ norm). Finally, we use a different interpolation scheme as detailed in Section 4.2.1.

4.1.2 Results

We seek to impose minimal requirements on a density estimator $\hat{f}$ of an unknown smooth density f so that it can be converted to a new estimator $\tilde{f}$ that performs well on the following criteria.

1. (**Minimax**) $\tilde{f}$ is a minimax estimator for f
2. (**Space-efficient**) $\tilde{f}$ can be stored efficiently
3. (**Fast querying**) $\tilde{f}$ can be evaluated efficiently
4. (**Fast preprocessing**) $\tilde{f}$ can be constructed efficiently

In this work, we focus on near-minimax estimation in the L^∞ norm, motivated by the aforementioned works on efficient evaluation of kernel density estimators. Since we impose that the unknown density f is supported on $[0,1]^d$, such a guarantee also implies upper bounds on the L^p error for all $p \geq 1$.

In the statistical setup where typically $\beta, d = O(1)$, by *efficient* we mean requiring only polynomial time or space in the sample size n. In particular for fixed β, by (4.2) consistent estimation is only possible when $d \ll \log n$. In what follows we indicate dependencies on the parameters β and d for clarity.

The requirement that we place on the estimator $\hat{f}$ to be converted is the following assumption.

Assumption 4.1. *For all $y \in [0,1]^d$ and $1 \geq t \geq \varepsilon$, we have*

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta, L)} \mathbb{P}_f \left[\left| \hat{f}(y) - f(y) \right| > t \right] \leq 2 \exp\left(-\frac{t^2}{\varepsilon^2} \right),$$

where $\varepsilon := c^* n^{-\beta/(2\beta+d)}$ *is the minimax rate of estimating L-Hölder smooth densities* $\mathcal{P}_{\mathcal{H}}(\beta, L)$ *of order* β *and* $c^* = c_{\beta,d,L} > 0$.

The formal definition of the Hölder class $\mathcal{P}_{\mathcal{H}}(\beta, L)$ we consider is given in Section 4.1.3. In particular Assumption 4.1 is satisfied if the pointwise error is a sub-Gaussian random variable with parameter ε that captures the minimax rate of estimation. For the KDE built from a kernel K of order $\ell := \lfloor \beta \rfloor$ [Tsybakov, 2009, Definition 1.3] and bandwidth $h = n^{-\frac{1}{2\beta+d}}$, this assumption follows from a standard bias-variance trade-off and an application of Bernstein's inequality (see Appendix 4.4.1).

Under Assumption 4.1, we have our main result.

Theorem 4.1. *Let* $f : [0,1]^d \to \mathbb{R}$ *denote a probability density function, and let* $\hat{f}$ *denote an estimator satisfying Assumption 4.1 for some* $\beta > 0$ *and* $d \geq 1$. *Let* Q *denote the amount of time it takes to query* $\hat{f}$. *Set* $\ell = \lfloor \beta \rfloor$. *Then there exists an estimator* $\tilde{f}$ *that can be constructed in time* $c_{\mathsf{con}} Q\, n^{\frac{d}{2\beta+d}}$, *that requires at most* $c_{\mathsf{sto}}\, n^{\frac{d}{2\beta+d}} \log n$ *bits to store, that can be queried in time* $c_{\mathsf{que}} \log n$, *and that satisfies*

$$\mathbb{E}_f \|\tilde{f} - f\|_\infty < c_{\mathsf{err}} (\log n)^{1/2} n^{-\frac{\beta}{2\beta+d}}.$$

In Theorem 4.1, we may take

$$\begin{aligned}
c_{\mathsf{con}} &= \binom{\ell+d}{\ell}, \\
c_{\mathsf{sto}} &= 5d(\ell+1)(\log L)\binom{\ell+d}{\ell}, \\
c_{\mathsf{que}} &= 14(d+\ell)^2\binom{\ell+d}{d}, \text{ and} \\
c_{\mathsf{err}} &= 8c^* L d^{\frac{3}{2}\ell+2}\ell^\ell \binom{\ell+d}{\ell}\sqrt{\log 2\binom{\ell+d}{\ell}}.
\end{aligned}$$

In particular, for $\beta, d = O(1)$, we can evaluate queries to $\tilde{f}$ in nearly constant time, and the estimator $\tilde{f}$ can be stored using sublinear space. Moreover, $\tilde{f}$ can be preprocessed in subquadratic time, assuming that the evaluation time of the original estimator $\hat{f}$ is $O_d(n)$, which holds for the KDE (4.1). We also note that $\tilde{f}$ is a near-minimax estimator in the sup norm, up to logarithmic factors, and thus by our domain assumption is also near-minimax in the L^p norms, again up to logarithmic factors. Finally, our construction in Section 4.2.1 yields an explicit formula for $\tilde{f}$ in terms of a sublinear number of initial queries of $\hat{f}$ on a judiciously chosen mesh. Specifically, the estimator $\tilde{f}$ is a piecewise multivariate interpolation of the estimator $\hat{f}$ on this mesh.

Though our focus is on density estimation, our method is not limited to this setting. The next result holds under a modified version of Assumption 4.1 and is derived by following the proof of Theorem 4.1. We omit the argument as it is very similar.

Theorem 4.2. *Let $f : [0,1]^d \to \mathbb{R}$ denote an L-smooth Hölder function of order β, and suppose that one has query access to a function $\hat{f}$ where $\|\hat{f} - f\|_\infty \le \varepsilon$. Then by first computing $c_{\mathsf{con}}\, \varepsilon^{-\frac{d}{\beta}}$ initial queries of $\hat{f}$, one can construct a new function $\tilde{f}$ that satisfies $\|f - \tilde{f}\|_\infty \le c_{\mathsf{err}}\, \varepsilon$, that can be stored using $c_{\mathsf{sto}}\, \varepsilon^{-\frac{d}{\beta}} \log \varepsilon^{-1}$ bits, and that can be queried in time $c_{\mathsf{que}} \log \varepsilon^{-1}$.*

Theorem 4.2 is useful when it is possible to design a procedure for estimating a smooth function f pointwise, but that procedure cannot necessarily be carried out efficiently per query. For example in nonparametric regression, Nadaraya–Watson estimators are known to be accurate pointwise [Tsybakov, 2009] but naively require evaluation time that is linear in the number of data points. One can also imagine a numerical or experimental setting where it is only possible to gather a limited number of accurate measurements of a smooth response, and one wants to graph the underlying function efficiently and accurately over the entire domain.

4.1.3 Setup and notation

Fix an integer $d \ge 1$. For any multi-index $s = (s_1, \dots, s_d) \in \mathbb{Z}^d_{\ge 0}$, let $|s| = s_1 + \cdots + s_d$ and for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, define $s! = s_1! \cdots s_d!$ and $x^s = x_1^{s_1} \cdots x_d^{s_d}$. Let D^s denote the differential operator

$$D^s = \frac{\partial^{|s|}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}\,.$$

Fix a positive real number β, and let $\lfloor \beta \rfloor$ denote the maximal integer *strictly* less than β. We reserve the notation $\|\cdot\|_p$ for the L^p norm and $|\cdot|_p$ for the ℓ^p norm.

Given $L > 0$ we let $\mathcal{H}(\beta, L)$ denote the space of Hölder functions $f : \mathbb{R}^d \to \mathbb{R}$ that are supported on the cube $[0,1]^d$, are $\lfloor \beta \rfloor$ times differentiable, and satisfy

$$|D^s f(x) - D^s f(y)| \le L\, |x - y|_2^{\beta - \lfloor \beta \rfloor}\,,$$

for all $x, y \in \mathbb{R}^d$ and for all multi-indices s such that $|s| = \lfloor \beta \rfloor$.

Let $\mathcal{P}_{\mathcal{H}}(\beta, L)$ denote the set of probability density functions contained in $\mathcal{H}(\beta, L)$. For $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$, let $\mathbb{P}_f$ (resp. $\mathbb{E}_f$) denote the probability distribution (resp. expectation) associated to f.

The parameter L will be fixed in what follows, so typically we write $\mathcal{P}_{\mathcal{H}}(\beta) := \mathcal{P}_{\mathcal{H}}(\beta, L)$. The constants $c, c_{\beta,d}, c_L$, etc. vary from line to line and their subscripts indicate parameter dependences.

4.2 Efficient interpolation of density estimators

The important implication of Assumption 4.1 is that we can query $\hat{f}$ at a polynomial number of data points such that for each query y, $\hat{f}(y) \approx f(y)$, where f is the unknown density.

Lemma 4.1. *Let $A > 0$ and set $N = \Delta n^A$ with $\Delta \geq 1$. Let $y_1, \ldots, y_N \subset [0,1]^d$ denote a set of points. Then with probability at least $1 - n^{-2}$,*

$$\left|\hat{f}(y_i) - f(y_i)\right| \leq \sqrt{\log(2\Delta n^{A+2})}\, \varepsilon$$

for all $1 \leq i \leq N$, where $\varepsilon = c^* n^{-\beta/(2\beta+d)}$ is the minimax rate.

Proof. Set $t = \sqrt{\log(2\Delta n^{A+2})}\, \varepsilon \geq \varepsilon$ and apply Assumption 4.1 to y_i. Then by the union bound,

$$\mathbb{P}\left[\exists y_i \,:\, \left|\hat{f}(y_i) - f(y_i)\right| > t\right] \leq 2\Delta n^A e^{-\frac{t^2}{\varepsilon^2}} \leq n^{-2}.$$

□

We now describe our construction of $\tilde{f}$. Define $\ell := \lfloor \beta \rfloor$ and $M = \binom{\ell+d}{\ell}$.

Construction of $\tilde{f}$ (*informal*):

1. PARTITION: Divide $[0,1]^d$ into h^{-d} sub-cubes $\{I_{\vec{j}}\} \subset [0,1]^d$ of side-length $h = n^{-1/(2\beta+d)}$ where $\vec{j} \in \mathbb{Z}^d_{\geq 0}$ and $I_{\vec{j}} := [0,h]^d + h\vec{j}$.

2. MESH: For each $\vec{j}$, construct a mesh consisting of $M = \binom{\ell+d}{\ell}$ points $U^{\vec{j}}_1, \ldots, U^{\vec{j}}_M \in I_{\vec{j}}$.

3. INTERPOLATE: In each sub-cube $I_{\vec{j}}$, construct a multivariate polynomial interpolant $\hat{q}_{\vec{j}}$ on the M points $(U^{\vec{j}}_1, \hat{f}(U^{\vec{j}}_1)), \ldots, (U^{\vec{j}}_M, \hat{f}(U^{\vec{j}}_M))$.

Return: $\tilde{f} : [0,1]^d \to \mathbb{R}$ defined by

$$\tilde{f}(y) = \sum_{\vec{j}} \hat{q}_{\vec{j}}(y) \mathbb{1}(y \in I_{\vec{j}}).$$

We first give some intuition for why $\tilde{f}$ is an accurate estimator. On each sub-cube $I_{\vec{j}}$, the true density $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$ is approximated up to the minimax error by a polynomial $q_{\vec{j}}$ of degree at most ℓ by the properties of Hölder functions. Upon setting $\Delta = M$ and $A = d/(2\beta + d)$ in Lemma 4.1, this guarantees that for all points $U^{\vec{j}}_k$ in the mesh, $\hat{f}(U^{\vec{j}}_k) \approx f(U^{\vec{j}}_k) \approx q_{\vec{j}}(U^{\vec{j}}_k)$ with high probability. By studying the stability of the resulting polynomial system of equations, we can show that this construction yields a good approximation to the 'true' interpolation polynomial $q_{\vec{j}}$ on the sub-cube $I_{\vec{j}}$. This argument, carried out formally later in this section, yields the estimation bound of Theorem 4.1.

Next, we comment on the remaining guarantees of Theorem 4.1. As we show later, there is an explicit formula for $\hat{q}_{\vec{j}}$, so the main preprocessing bottleneck is the evaluation of $\hat{f}$ on the $Mn^{d/(2\beta+d)}$ points in the mesh, which naively takes $QMn^{d/(2\beta+d)}$ time. For the space requirement, it suffices to store the values $\{\hat{f}(U^{\vec{j}}_k)\}$ up to polynomial precision as well as the elements of the mesh. Querying $\hat{f}$ at a point $y \in [0,1]^d$

requires checking which sub-cube y belongs to by scanning its d coordinates and then evaluating $\hat{q}_{\vec{j}}(y)$, which is a d-variate polynomial of degree $\lfloor \beta \rfloor$. By a careful consideration of the numerical precision required to perform these steps in Section 4.2.2, we obtain the computational guarantees of Theorem 4.1.

4.2.1 Interpolation on the principal lattice

To construct our interpolant, we refer to the next definition and theorem which are classical in finite element analysis [Nicolaides, 1972, Chung and Yao, 1977]. The lattice $\mathcal{P}_\ell \subset [0,1]^d$, dubbed the ℓ-th principal lattice, has the special property that every function defined on $\mathcal{P}_\ell$ admits a unique polynomial interpolant of degree at most ℓ. This property is known to be equivalent to a combinatorial geometric condition referred to as GC in Chung and Yao [1977]. A set of points $\mathcal{P}$ is called GC if every point $x \in \mathcal{P}$ has an associated set $\mathcal{H}_x$ consisting of ℓ affine hyperplanes whose union contains $\mathcal{P} \backslash x$ and such that none of these hyperplanes contain x.

Definition 4.1 (ℓ-th principal lattice of Δ_d)**.** *Let $\Delta_d \subset [0,1]^d$ denote the simplex on the points $\{0\} \cup \{e_i\}_{i=1}^d \subset \mathbb{R}^d$, where e_i denotes the i-th standard basis vector in $\mathbb{R}^d$. Label the vertices of Δ_d to be $v_0 = 0, v_i = e_i$ for $1 \leq i \leq d$. For all $x \in \mathbb{R}^d$, there exists a unique vector $(\lambda_0(x), \ldots, \lambda_d(x))$ with entries summing to one such that*

$$x = \sum_{i=0}^{d} \lambda_i(x) v_i \,.$$

Let $\Lambda : \mathbb{R}^d \to \mathbb{R}^{d+1}$ denote the function such that $\Lambda(x) = (\lambda_0(x), \ldots, \lambda_d(x))$. For $\ell \geq 1$, the ℓ-th principal lattice $\mathcal{P}_\ell$ of Δ_d is defined to be

$$\mathcal{P}_\ell = \left\{ x \in \Delta_d : \, \ell \Lambda(x) \in \mathbb{Z}_{\geq 0}^{d+1} \right\} . \tag{4.3}$$

We also define $\mathcal{P}_0 = 0 \in \mathbb{R}^d$.

Given a point $x \in \mathcal{P}_\ell$, the associated set of affine hyperplanes satisfying the GC condition is

$$\mathcal{H}_x = \bigcup_{\substack{t=0 \\ \lambda_t(x)>0}}^{d} \bigcup_{r=0}^{\ell\lambda_t(x)-1} \left\{ \sum_{i=0}^{d} \alpha_i v_i \Big| \; \ell\alpha_t = r, \, \sum_{i=0}^{d} \alpha_i = 1 \right\} .$$

Given a set of hyperplanes satisfying this combinatorial condition, it is straightforward to write down a Lagrangian-type interpolation formula, as was first computed for the principal lattice by Nicolaides [1972].

Theorem 4.3 (Nicolaides [1972],Chung and Yao [1977])**.** *Write $\mathcal{P}_\ell = \{U_1, \ldots, U_M\} \subset \Delta_d$ and let $g : \mathcal{P}_\ell \to \mathbb{R}$ denote a function defined on this lattice. For $\ell \geq 1$, define the polynomial*

$$p_i(x) = \prod_{\substack{t=0 \\ \lambda_t(U_i)>0}}^{d} \prod_{r=0}^{\ell\lambda_t(U_i)-1} \frac{\lambda_t(x) - \frac{r}{\ell}}{\lambda_t(U_i) - \frac{r}{\ell}} \,, \tag{4.4}$$

where we recall that $\lambda_t(x)$ is from Definition 4.1. If $\ell = 0$, then $M = 1$, and we simply define $p_1(x) \equiv 1$. Then

$$p(x) := \sum_{i=1}^{M} p_i(x) g(U_i)$$

satisfies $p(U_i) = g(U_i)$ for all $U_i \in \mathcal{P}_\ell$. Moreover, this is the unique polynomial of degree at most ℓ with this property.

Since $\lambda_t(x)$ is linear in $x \in \mathbb{R}^d$, it is easy to see that $p_i(x)$ is a polynomial of degree ℓ, and moreover $p_i(U_j) = 1$ if $i = j$ and zero otherwise.

We are now ready to give a precise description of the construction of $\tilde{f}$. The idea is to generate the mesh for interpolation using a shifted and rescaled version of the ℓ-th principal lattice on $\Delta_d \subset [0,1]^d$. Recall that $\hat{f}$ is a density estimator that satisfies Assumption 4.1.

Construction of $\tilde{f}$ (*formal version*):

1. PARTITION: Divide $[0,1]^d$ into h^{-d} sub-cubes $\{I_{\vec{j}}\} \subset [0,1]^d$ of side-length $h = n^{-1/(2\beta+d)}$ where $\vec{j} \in \mathbb{Z}^d_{\geq 0}$ and $I_{\vec{j}} := [0,h]^d + h\vec{j}$.

2. MESH: For each $\vec{j}$, construct a mesh on $I_{\vec{j}}$ consisting of $M = \binom{\ell+d}{\ell}$ points given by the shifted and rescaled principal lattice $\mathcal{P}_\ell^{\vec{j}} := \{h(x + \vec{j}) : x \in \mathcal{P}_\ell\} \subset I_{\vec{j}}$. Let $U_1^{\vec{j}}, \ldots, U_M^{\vec{j}}$ denote the points in $\mathcal{P}_\ell^{\vec{j}}$.

3. INTERPOLATE: In each sub-cube $I_{\vec{j}}$, construct a multivariate polynomial interpolant $\hat{q}_{\vec{j}}$ through the M points $(U_1^{\vec{j}}, \hat{f}(U_1^{\vec{j}})), \ldots, (U_M^{\vec{j}}, \hat{f}(U_M^{\vec{j}}))$ given by $\hat{q}_{\vec{j}}(y) = p_{\vec{j}}(y/h - \vec{j})$, where p is the polynomial interpolant from Theorem 4.3 given by

$$p_{\vec{j}}(x) = \sum_{k=1}^{M} p_k(x) \hat{f}(U_k^{\vec{j}}).$$

Return: $\tilde{f} : [0,1]^d \to \mathbb{R}$ defined by

$$\tilde{f}(y) = \sum_{\vec{j}} \hat{q}_{\vec{j}}(y) \mathbb{1}(y \in I_{\vec{j}}).$$

4.2.2 Proof of Theorem 4.1

We prove Theorem 4.1 in two parts, first by studying the estimation error $\|\tilde{f} - f\|_\infty$ in Section 4.2.2 and second by proving the storage and time complexity upper bounds in Section 4.2.2.

Estimation error

First, we quantify the error in the approximation of the values of $q_{\vec{j}}$ on the mesh points. Let $f_{z,\ell}$ denote the degree ℓ polynomial given by the Taylor expansion of $f \in \mathcal{P}_{\mathcal{H}}(\beta)$ at z. Since $f \in \mathcal{P}_{\mathcal{H}}(\beta)$, by a standard fact (see Lemma 4.5) it holds that

$$|f(y) - f_{z,\ell}(y)| \le \frac{Ld^{\ell/2}}{\ell!} |y - z|_2^{\beta},$$

where $f_{z,\ell}$ is the degree-ℓ Taylor expansion of the function f at $z \in \mathbb{R}^d$.

For $\vec{j} \in \{0, \dots, h^{-1} - 1\}^d$, define $q_{\vec{j}} := f_{z_{\vec{j}},\ell}$, where $z_{\vec{j}}$ is the vertex of $I_{\vec{j}}$ closest to the origin. Then for all $y \in I_{\vec{j}}$, it holds that

$$\begin{aligned} |f(y) - q_{\vec{j}}(y)| &\le \left(\frac{Ld^{\beta}}{\ell!}\right) h^{\beta} \\ &= \left(\frac{Ld^{\beta}}{\ell!}\right) n^{-\beta/(2\beta+d)} \\ &=: \hat{c} n^{-\beta/(2\beta+d)} \end{aligned} \tag{4.5}$$

Note that the right-hand side is the minimax rate of estimation in (4.2) up to constant factors.

Next, by Lemma 4.1 (setting $\Delta = M$ and $A = \frac{d}{2\beta+d}$) and (4.5) it holds with probability at least $1 - n^{-2}$ that

$$\begin{aligned} \left| q_{\vec{j}}(U_k^{\vec{j}}) - \hat{f}(U_k^{\vec{j}}) \right| &\le (c^* \sqrt{4 \log 2M} + \hat{c})(\log n)^{\frac{1}{2}} n^{-\frac{\beta}{2\beta+d}} \\ &=: \breve{c} (\log n)^{\frac{1}{2}} n^{-\frac{\beta}{2\beta+d}} \end{aligned} \tag{4.6}$$

for all $\vec{j} \in \{0, \dots, h^{-1} - 1\}^d$ and $k \in [M]$. Using this fact, we can show that the polynomial interpolant built on $\{(U_k^{\vec{j}}, \hat{f}(U_k^{\vec{j}}))\}_{k=1}^M$ provides a good approximation for $q_{\vec{j}}$ on the interval $I_{\vec{j}}$, which is our next task. The following lemma establishes stability of the polynomial approximation.

Lemma 4.2. *Let $\hat{q}_{\vec{j}}$ denote the unique polynomial of degree at most ℓ that passes through the points $\{(U_k^{\vec{j}}, \hat{f}(U_k^{\vec{j}}))\}_{k=1}^M$. Then with probability at least $1 - n^{-2}$, for all $\vec{j}$ and all $x \in I_{\vec{j}}$,*

$$\left| q_{\vec{j}}(x) - \hat{q}_{\vec{j}}(x) \right| \le c_{\beta,d,L} (\log n)^{\frac{1}{2}} n^{-\frac{\beta}{2\beta+d}}. \tag{4.7}$$

Proof. Define $\hat{g}_{\vec{j}}(x) = \hat{q}_{\vec{j}}(h(x + \vec{j}))$ and $g_{\vec{j}}(x) = q_{\vec{j}}(h(x + \vec{j}))$ to be polynomials restricted to the domain $[0, 1]^d$. Recall that $\hat{g}$ and g are given by formulas as in Theorem 4.3. It holds by (4.6) that for all $1 \le k \le M$,

$$\left| \hat{g}(U_k^{\vec{j}}) - g(U_k^{\vec{j}}) \right| \le \breve{c} (\log n)^{1/2} n^{-\frac{\beta}{2\beta+d}}.$$

Let $y \in [0,1]^d$, and observe that by Theorem 4.3 and the triangle inequality,

$$|\hat{g}(y) - g(y)| \leq M \sup_{\substack{x \in [0,1]^d \\ 1 \leq k \leq M}} \left| p_k(x) \left(\hat{g}(U_k^{\vec{j}}) - g(U_k^{\vec{j}}) \right) \right|$$
$$\leq M \, \check{c} \, (\log n)^{1/2} n^{-\frac{\beta}{2\beta+d}} \sup_{\substack{x \in [0,1]^d \\ 1 \leq k \leq M}} |p_k(x)| \,. \tag{4.8}$$

Observe that for $x \in [0,1]^d$, we have $|\lambda_0(x)| = |1 - \sum x_i| \leq d$, and for $1 \leq t \leq d$, we have $|\lambda_t(x)| = |x_i| \leq 1$. Therefore, by the definition of p_k and $U_k^{\vec{j}}$,

$$|p_k(x)| \leq \ell^\ell d.$$

By this bound, (4.8), and translation and scale invariance of $\|\cdot\|_\infty$, Lemma 4.2 follows with $c_{\beta,d,L} = \check{c} M d \ell^\ell$. □

Define $\tilde{f}(x) = \sum_{\vec{j}} \hat{q}_{\vec{j}}(x) \mathbb{1}(x \in I_{\vec{j}})$, and observe that Theorem 4.1 follows from (4.5), Lemma 4.2, and the triangle inequality. Though we have derived a high probability bound, the expectation claimed in Theorem 4.1 follows using the uniform boundedness of Hölder functions as stated in Lemma 4.4. Tracing constants above yields the expression for c_{err}.

Time and space requirements

Recall that $M = \binom{\ell+d}{\ell}$ where $\ell = \lfloor \beta \rfloor$. For the space requirement, we store the principal lattices and the values of $\hat{f}$ on these lattice points, and note that each query is at most $Ld^{O(\beta+1)}$ in magnitude by Lemma 4.4. The queries per sub-cube can thus be stored with $M(\log Ld^{O(\beta+1)} + \log n)$ bits. The extra $\log n$ bits are required so that the interpolating polynomials can be queried with sufficient precision. The lattices are composed of rational points in $\mathbb{R}^d$, so we need at most $Md \log(\beta+1)$ bits per sub-cube to store them. Since there are $n^{\frac{d}{2\beta+d}}$ sub-cubes, the space requirement of Theorem 4.1 follows and is a conservative estimate for simplicity.

Next we characterize the time complexity. Assume first that $\ell \geq 1$. For $1 \leq k \leq M$, it holds that

$$|p_k(y) - p_k(y')| \leq (d+1) 2^\ell \ell^{\ell+1} |y - y'|_\infty$$

because by expanding the product in the formula in Theorem 4.3, p_i is a sum of at most $2^\ell(d+1)$ terms, each having coefficients of size at most ℓ^ℓ, and moreover for $|\alpha| \leq \ell$, the monomial y^α is ℓ-Lipschitz with respect to $|\cdot|_\infty$ over the cube. Therefore, it also holds that

$$\left| \hat{q}_{\vec{j}}(y) - \hat{q}_{\vec{j}}(y') \right| \leq ML d^{\frac{3}{2}\beta + \frac{1}{2}} (d+1) 2^\ell \ell^{\ell+1} |y - y'|_\infty$$

by the formula in the interpolation step of $\tilde{f}$, noting that without loss of generality,

$\left|\hat{f}(U_k^{\vec{j}})\right| = Ld^{O(\beta+1)}$ by Lemma 4.4. By the form of c_{err}, given a query y it suffices to round its coordinates to $B := \ell + \log d + \log n$ bits to compute $\hat{q}_{\vec{j}}(y)$ with the required level of accuracy.

Next, the number of arithmetic operations needed to evaluate $\hat{q}_{\vec{j}}(y)$ is bounded conservatively by $6(d+\ell)M$. To identify which sub-cube contains y requires time at most $2d \log n$. Hence, the total complexity is upper bounded by

$$6(d+\ell)MB + 2d \log n \leq 16(d+\ell)^2 M \log n =: c_{\mathsf{que}}$$

This bound also holds conservatively when $\ell = 0$ since in that case, to evaluate $\hat{f}(y)$, we just need to match the given query y to the sub-cube $I_{\vec{j}}$ containing it and output $\hat{f}(U_1^{\vec{j}})$.

4.3 A result of Kolmogorov and Tikhomirov

Given a function class $\mathcal{F}$, let $N(\mathcal{F}, \delta)$ denote the minimal number of L^∞ balls of radius δ that cover $\mathcal{F}$, and define $H(\mathcal{F}, \delta) = \log N(\mathcal{F}, \delta)$ to be the metric entropy. Let $\mathcal{H}(\beta) = \mathcal{H}(\beta, L)$ denote the class of Hölder functions supported on $[0,1]^d$ as defined in Section 4.1.3. A classical result of Kolmogorov and Tikhomirov [1993] shows that

$$H(\mathcal{H}(\beta), \delta) \leq c_{\beta,d,L}\, \delta^{-\frac{d}{\beta}}. \tag{4.9}$$

Their proof strategy is conceptually similar to our piecewise multivariate polynomial approximation scheme in that they subdivide the cube as we do here, approximate f by its Taylor polynomial in each cube, and then discretize the coefficients. We show now that our techniques imply a slightly weaker version of the bound (4.9).

Define a mesh as in steps 1 and 2 of our formal construction of $\tilde{f}$ as in Section 4.2.1, but now for a general parameter $h > 0$ to be set later. This mesh has Mh^{-d} points that we denote by $\{U_k^{\vec{j}}\}_{\vec{j},k}$. Let $f, g \in \mathcal{H}(\beta)$ be such that for all $\vec{j}, k$ it holds that

$$\left|f(U_k^{\vec{j}}) - g(U_k^{\vec{j}})\right| \leq h^\beta.$$

By the Hölder condition and Lemma 4.5, there exists a degree $\ell = \lfloor \beta \rfloor$ polynomial $q_{\vec{j}}$ approximating f in $I_{\vec{j}}$ and a degree $\ell = \lfloor \beta \rfloor$ polynomial $r_{\vec{j}}$ approximating g in $I_{\vec{j}}$, each with error h^β pointwise. We conclude that

$$\left|q_{\vec{j}}(U_k^{\vec{j}}) - r_{\vec{j}}(U_k^{\vec{j}})\right| \leq c_{\beta,d,L}\, h^\beta$$

for all $\vec{j}, k$. Following the proof of Lemma 4.2, this implies that for all $x \in I_{\vec{j}}$,

$$\left|q_{\vec{j}}(x) - r_{\vec{j}}(x)\right| \leq c_{\beta,d,L}\, h^\beta.$$

Hence we conclude that for all $x \in [0,1]^d$,

$$|f(x) - g(x)| \leq c_{\beta,d,L}\, h^\beta.$$

The Hölder functions are uniformly bounded by some constant $c_{\beta,d,L}$ (see Lemma 4.4). Hence setting $\delta = c_{\beta,d,L}\, h^\beta$ and rounding the values of each function at each point $U_k^{\vec{j}}$ to multiples of h^β, we see that there exists a δ-net of size at most

$$\left(\frac{c_{\beta,d,L}}{\delta}\right)^{M c'_{\beta,d,L} \delta^{-d/\beta}}.$$

Therefore

$$H(\mathcal{H}(\beta), \delta) \leq c_{\beta,d,L}\, \delta^{-\frac{d}{\beta}} \log \frac{1}{\delta},$$

a mildly weaker bound than (4.9).

4.4 Appendix

4.4.1 KDEs satisfy Assumption 1

In this section, for completeness we verify that for appropriate kernels, the standard KDE satisfies Assumption 4.1.

Proposition 4.1. *Let $K(\cdot)$ denote a kernel of order $\lfloor \beta \rfloor$ satisfying*

$$\|K\|_\infty < \infty, \ \int K^2(x)\mathrm{d}x < \infty, \ \int |x^\alpha K(x)|\,\mathrm{d}x < \infty$$

for all multi-indices $\alpha \in \mathbb{R}^d_{\geq 0}$ with $|\alpha| = \beta$. Then Assumption 4.1 is satisfied for the KDE $\hat{f}$ with bandwidth $h = cn^{-1/(2\beta+d)}$.

Proof. For brevity, c denotes a constant that varies from line to line and can depend on β, d, L and K. Fix $y \in [0,1]^d$. It is well-known that under the conditions of Proposition 4.1 [see e.g. Tsybakov, 2009],

$$b = b(y) := \left|\mathbb{E} f(y) - \hat{f}(y)\right| \leq ch^\beta,$$

and for a data point $X_i \sim \mathbb{P}_f$,

$$\tau^2 = \tau^2(y) := \mathsf{Var}\, K_h(X_i - y) \leq \frac{c}{h^d}.$$

By the triangle inequality and Bernstein's inequality for bounded random variables

[Vershynin, 2018],

$$\Pr\left(\left|\hat{f}(y) - f(y)\right| > t\right) \leq \exp\left(-\frac{n(t-b)^2}{2\tau^2 + 2\|K_h\|_\infty (t-b)/3}\right). \quad (4.10)$$

Let $h = cn^{-1/(2\beta+d)}$. Note that $\|K_h\|_\infty = h^{-d}\|K\|_\infty$ and $(nh^d)^{-1} = cn^{-\beta/(2\beta+d)}$. Then we recover Assumption 4.1 by setting $t \geq cn^{-\beta/(2\beta+d)}$ in (4.10).

□

4.4.2 Properties of Hölder functions

For completeness, we provide proofs of standard facts about the class of Hölder functions.

Lemma 4.3 (Inclusion)**.** *Let $\mathcal{H}(\beta, d, L)$ denote the class of Hölder functions supported on $[0,1]^d$ in dimension d. If $\beta > 1$, then it holds that $\mathcal{H}(\lfloor\beta\rfloor, d, L) \subset \mathcal{H}(\lfloor\beta\rfloor - 1, d, d^{3/2}L)$.*

Proof. Let $f \in \mathcal{H}(\beta, d, L)$. Since f is supported on $[0,1]^d$ and smooth on $\mathbb{R}^d$, we have that

$$|D^s f(x)| \leq L\,|x|_2 \leq L\sqrt{d} \quad (4.11)$$

for all $|s| = \lfloor\beta\rfloor$.

Fix $x, y \in [0,1]^d$, and define for $1 \leq i \leq d+1$ the point $z^i \in [0,1]^d$ to be

$$z^i_j = \begin{cases} x_j & \text{if } j \geq i \\ y_j & \text{if } j < i. \end{cases}$$

Observe that $z^1 = x$ and $z^{d+1} = y$.

Let t denote a multi-index with $|t| = \lfloor\beta\rfloor - 1$. By the fundamental theorem of calculus and the Hölder condition,

$$|D^t f(x) - D^t f(y)| \leq \sum_{i=1}^{d} \left|D^t f(z^i) - D^t f(z^{i+1})\right| = \sum_{i=1}^{d} \left|\int_{x_i}^{y_i} \frac{\partial}{\partial x_i} D^t f(x_1, \ldots, z, y_{i+1}, ..., y_d)\,\mathrm{d}z\right|.$$

Using (4.11), the expression in the second line is bounded above by $Ld^{3/2}$, which proves the lemma.

□

Lemma 4.4 (Uniform boundedness)**.** *The class $\mathcal{H}(\beta)$ is uniformly bounded. In particular,*

$$\sup_{f \in \mathcal{H}(\beta)} \|f\|_\infty \leq d^{3\lfloor\beta\rfloor/2 + 1/2}\, L.$$

Proof. Suppose first that $f \in \mathcal{H}(\beta)$ for $\beta > 1$. By repeated application of Lemma 4.3, f is $(d^{3\lfloor\beta\rfloor/2}L)$-Lipschitz. Since f is supported on $[0,1]^d$,

$$|f(x)| = |f(x) - f(0)| \leq d^{3\lfloor\beta\rfloor/2} L \, |x|_2 \leq d^{3\lfloor\beta\rfloor/2+1/2} L.$$

If $\beta \leq 1$, then arguing as in the previous display, we see that $|f(x)| \leq L\sqrt{d}$ for all $x \in \mathbb{R}^d$. □

Lemma 4.5 (Taylor approximation)**.** *Given $f \in \mathcal{H}(\beta)$, let $f_{x,\lfloor\beta\rfloor}$ denote its Taylor polynomial of degree $\lfloor\beta\rfloor$ at a point $x \in \mathbb{R}^d$,*

$$f_{x,\lfloor\beta\rfloor}(y) = \sum_{|s| \leq \lfloor\beta\rfloor} \frac{(y-x)^s}{s!} D^s f(x) \,, \quad y \in \mathbb{R}^d \,.$$

Then it holds that

$$\left| f(y) - f_{x,\lfloor\beta\rfloor}(y) \right| \leq \frac{L d^{\lfloor\beta\rfloor/2}}{\lfloor\beta\rfloor!} \, |x-y|_2^\beta \,, \quad x, y \in \mathbb{R}^d \,.$$

Proof. By Taylor's theorem with remainder [see, eg., Folland, 1999]

$$\left| f(y) - f_{x,\lfloor\beta\rfloor}(y) \right| = \left| \sum_{|s| = \lfloor\beta\rfloor} \frac{1}{s!} \left[D^s f(x + c(y-x)) - D^s f(x) \right] (y-x)^s \right|$$

for some constant $c \in (0,1)$. By the triangle inequality and the Hölder condition, the expression in the second line is bounded above by

$$\sum_{|s| = \lfloor\beta\rfloor} \frac{L \, |x-y|_2^{\beta - \lfloor\beta\rfloor}}{s!} \, |(y-x)^s| = \frac{L \, |x-y|_2^{\beta - \lfloor\beta\rfloor}}{\lfloor\beta\rfloor!} \left(\sum_{i=1}^d |x_i - y_i| \right)^{\lfloor\beta\rfloor},$$

where the equality is by the multinomial theorem. In turn, this last expression is bounded above by

$$\frac{L d^{\lfloor\beta\rfloor/2}}{\lfloor\beta\rfloor!} \, |x-y|_2^\beta$$

using Cauchy–Schwarz. □

Chapter 5

Pedigree reconstruction

5.1 Introduction

5.1.1 Motivation

The decreased costs of sequencing technologies have enabled large-scale, data-driven analyses of genomes Institute [2019]. Recent science and news articles feature stories only possible due to this plethora of data, such as the recent identification and capture of a high-profile criminal Kolata and Murphy [2018] predicated on DNA evidence. In this effort, an individual's genetic information was compared to a large, curated database called GEDMatch consisting of over one million individual genomes. In comparison, there exist databases which are of several orders of magnitude larger in size such as MyHeritage (∼3.7 million MyHeritage), 23andMe (∼10 million 23andMe), and Ancestry (∼15 million Ancestry.com).

This raises the question: how much kinship information can be learned from DNA? Current databases already contain a considerable amount of this information. Indeed, it is estimated that a given US individual of European ancestry, on average, has a third cousin or closer who is already in the MyHeritage database Erlich et al. [2018]. However, such databases are still far from complete. This calls into question the ability to detect missing kinships based on individuals already present in the database.

This discussion also highlights the issue of *genomic privacy*. Indeed, it becomes much easier to identify and locate individuals by combining the genetic and genealogical information with outside information (addresses, e-mails, family photos, etc.). This potential, having already been demonstrated by the resolution of the aforementioned criminal case, was brought to attention by Erlich et al. [2018]. From this point of view, the ability to reconstruct genealogies from collected genetic data is of concern for individuals whose information is revealed, even if one has *never* been sequenced. Since our work establishes a positive result in a pessimistic scenario where we start with no ground truth information, we believe that our work brings to attention this critical issue via a theoretical framework.

5.1.2 Our contributions

Without any prior knowledge about the ground truth, can we learn *everyone's* genealogy using their genetic information? In this paper, we study the inference problem of recovering ancestral kinship relationships of a population of *extant* (present-day) individuals, using only their genetic data. Our goal is to use this extant genetic data to recover the *pedigree* of the extant population, under an idealized model. A pedigree is a graph whose nodes (individuals) have edges that encode parent-sibling relationships. The topology and reconstruction of pedigrees are well-studied in bioinformatics from both a theoretical and empirical perspective, and in general the study of pedigrees poses formidable computational and statistical challenges.

In this paper, we introduce a novel recursive algorithm Rec-Gen for pedigree reconstruction. To demonstrate the effectiveness of our approach, we give a mathematical proof that for an idealized generative model on pedigrees, our algorithm is able to approximately recover the true, unknown pedigree only using the genetic data of the extant population. In terms of *sample complexity*, which for our purposes refers to the common gene sequence length of an extant individual, our algorithm greatly outperforms the naive reconstruction method (estimate pairwise distances between the extant individuals, then construct the pedigree that produces these distances). We propose our approach in this work as a prototype for the future study of more general pedigrees, including those involving real-life genetic data, from both a theoretical and empirical perspective. For further discussion on our model of pedigree generation, as well as its features and limitations, see Section 5.1.4 and Section 5.1.6.

5.1.3 Related works

A common method in theoretical evolutionary biology is to model lineages and inheritance via a family of directed acyclic graphs. One line of work is that of *phylogenetics* (refer to Semple and Steel [2003] for an overview) which uses trees to model the occurrence of large-scale *speciation events* in evolutionary biology. Another line of work is *coalescent theory*, which focuses on variable-height inheritance trees between genes as its main statistic to infer large-scale *population sizes*, as in e.g. Kim et al. [2019]. In contrast, pedigrees capture small-scale *individual genealogies* that encode familial relationships. Specifically, most pedigree models are for human genealogies, where we designate exactly two parents to each individual. By construction, such graphs are no longer trees and warrant different strategies for inference.

Steel and Hein [2006] posed the formal definition of pedigrees using graph-theoretic language. In that work, the authors gave combinatorial arguments proving that one can reconstruct complete pedigrees, assuming the correct ancestral history is provided as an input for each extant individual. Our definition of pedigrees is essentially the same as the one outlined by these authors, though we make the simplification that we do not identify the vertex set bipartition (corresponding to the biological sex of the individuals).

To tie in more closely with real-world applications, one must consider the challenge of estimating these histories from data. Along these lines, Thatte and Steel [2008]

studied stochastic processes that one can associate with the pedigree, in such a way that one can prove negative results (information-theoretic impossibility) or positive results (an algorithm) for the reconstruction of the pedigree from extant data. The stochastic process used to show their positive result was based on a very specific family of Markov chains which allows for inference but is quite different from our model.

For the problem of performing pedigree reconstruction on real data, there is a wealth of literature Thompson [2000], Kirkpatrick et al. [2011], He et al. [2013], Thompson [2013], He et al. [2014], Shem-Tov and Halperin [2014], Huisman [2017], Wang [2019]. Such studies apply heuristics that take into account various complications and phenomena observed in human genomes, such as varying levels of correlations between different sites and the presence of mutations that are not inherited from parents.

One line of work particularly relevant to this paper is He et al. [2013, 2014] in which the authors also tackle the problem of pedigree reconstruction from real extant genetic data. Assuming answers to queries of the form, "how much DNA did i and j simultaneously inherit from their ancestors?", they design a statistical test that distinguishes between siblings, half-siblings and cousins. Their method leverages this information with a maximal-clique finding algorithm to iteratively reconstruct the parents, layer-by-layer. There is no proof of correctness provided, but they provide benchmarks on real and simulated data to provide experimental justification. Our contributions have a slightly different flavor: using a similar iterative strategy but with a different statistical test (the novel part of our algorithm) and for a more optimistic set of assumptions, one can actually *provably* reconstruct the pedigree correctly in a sample-efficient way, in an asymptotic sense.

The authors of He et al. [2014] specifically emphasize their method's ability to reconstruct half-siblings. Technically speaking, this is not allowed in our model and therefore it may appear to the reader that there is something too restrictive or suboptimal about our analysis. One major difference between our model and the aforementioned work is that we model *haploid* individuals (one copy of DNA), while in reality humans are *diploids* (two copies of DNA). Furthermore, in our proof, we guarantee reconstruction of monogamous *couples* of haploid individuals – in other words, up to permutation of the two individuals within each couple. It can be observed that given a monogamous pedigree with a haploid model, one can construct a natural, non-monogamous pedigree with a diploid model such that the total variation of the extant data of the two pedigrees is zero. Therefore, we think that our results should also hold for a diploid model with minor modifications and have correctness guarantees to match the empirical results of the aforementioned work He et al. [2014], for example by interpreting Fig. 5-1(a) as a pair of diploid half-siblings.

Our work is also closely related to the problem of phylogenetic reconstruction Erdős et al. [1999], Mossel [2004], Mossel and Roch [2005], Daskalakis et al. [2006]. In this setting, symbols are passed from the root of a phylogenetic tree to descendants via a Markov process such as in the Cavender–Farris–Neyman model, a basic model for mutations. Similar to our inference problem in this work, in phylogenetic reconstruction, one is tasked with reconstructing the tree given only the symbols at the leaves.

The main result of Erdős et al. [1999] characterizes the *sample complexity*—the minimal string length of the data at the leaves such that reconstruction is possible—as logarithmic in the depth of the tree, a phenomenon that our results suggest also holds for the pedigree reconstruction problem. The work Mossel and Roch [2005] provides theoretical guarantees for the problem of learning the phylogenetic generative model (*i.e.*, the topology of the tree as well as the transition matrices), which includes hidden Markov models as a special case, from the extant data under a spectral assumption on the transition matrices (see also later work of Hsu et al. [2012]). Most closely related to our approach in this paper is the work Mossel [2004], which shows how to recursively reconstruct phylogenies using techniques from the theory of broadcast processes on trees (see also Daskalakis et al. [2006]). This approach provides inspiration for our main algorithm Rec-Gen, which uses similar techniques to recursively reconstruct pedigrees. We direct the reader to Evans et al. [2000] and Mossel [2001] for studies of broadcast processes on trees with binary and large alphabet respectively, and Makur et al. [2018] for a generalization to directed acyclic graphs.

5.1.4 Model description and results

We now give an informal, detailed description of our framework for pedigree reconstruction, with a more detailed treatment of the generative model in Section 5.3. Our generative model on pedigrees consists of two parts: a parametric model for generating the network structure on the set of ancestors and extant individuals, and an inheritance procedure for transmitting genetic data from the *founders*, the oldest individuals in the pedigree, to the extant population.

To generate the pedigree network structure, we begin with a large founding population of size N_T. The founders randomly mate monogamously, and each couple gives birth to a random number of children, so that the average number of offspring per couple is a constant[1] α. This procedure of random monogamous mating continues for T subsequent generations, eventually yielding the extant nodes and a pedigree $\mathcal{P}$ formed by the individuals in generations $0, 1, \ldots, T$, with N_i nodes at each level i.

Next we describe how genetic data transmits from the founding population to the extant. Every individual in the pedigree has a gene sequence consisting of B symbols placed in B distinct blocks. Each individual in the founding population is initialized with independent uniformly random draws from a very large alphabet Σ. Now we state how parents pass down genes to their children. In a given block, a child inherits, with equal probability, either its mother's or its father's symbol in the corresponding block. This procedure repeats for all couples in a given generation and then continues over subsequent generations so that genetic data is iteratively transferred through the pedigree, eventually giving rise to the gene sequences of the extant individuals.

Our main result is summarized in the following theorem. See Theorem 5.3 for a formal statement.

[1]More precisely, each couple has a random number of children distributed as a Poisson random variable with expectation α.

Theorem 5.1 (Main result, informal)**.** *Let α and β denote sufficiently large absolute constants independent of N_T, the size of the founding population. Let ε denote a sufficiently small absolute constant independent of N_T. Assume that the alphabet size $|\Sigma|$ is very large with respect to N_T.*

Then given extant genetic data produced from the generative model with alphabet Σ, growth rate α, gene sequence length $B = \beta \log N_T$, and number of generations $T = \varepsilon \log N_T$ as described above, the algorithm Rec-Gen *recovers 90% of the true pedigree in every generation, with high probability. Moreover, this algorithm runs in polynomial time in the size of the pedigree and the number of blocks per extant individual.*

Let $\mathcal{P}$ denote the true, unknown pedigree. Our formal version of Theorem 5.1 (see Theorem 5.3) implies that with high probability Rec-Gen outputs a reconstructed pedigree $\hat{\mathcal{P}}$ whose size is at least $0.9N_i$ in each generation $i \in \{0, \dots, T\}$, such that every node $\hat{u} \in \hat{\mathcal{P}}$ can be identified with exactly one node $u \in \mathcal{P}$, and this identification preserves relationships in the sense that $\hat{u}$ is a child of $\hat{v}$ in $\hat{\mathcal{P}}$ if and only if u is a child of v in $\mathcal{P}$. In graph-theoretic terminology, our reconstruction $\hat{\mathcal{P}}$ is a (very large) induced subgraph of the truth $\mathcal{P}$.

We note that the stipulation that we recover 90% of the nodes at each level is actually a simplification; in fact, we can make the fraction of reconstructed nodes in each generation *arbitrarily large* by taking α to be large enough. We refer the reader to Theorem 5.3 for details.

5.1.5 The Rec-Gen algorithm

The algorithm Rec-Gen consists of a recursive procedure that uses only the genetic information from the extant population to construct a good approximation for the true pedigree $\mathcal{P}$ of depth T that generated the observations. In the first phase of recursion, the algorithm reconstructs the parents of the extant nodes, which we label as the 1^{st} generation. In the t^{th} phase, the algorithm adds a t^{th} generation to the partially reconstructed version of the true pedigree given by the output of the previous phase. The algorithm terminates after T phases of recursion, producing a pedigree $\hat{\mathcal{P}}$ with T generations that well-approximates the true, unknown pedigree $\mathcal{P}$.

We next give a simplified version of our recursive procedure that serves to illustrate the main ideas. See Section 5.6 for a detailed description of Rec-Gen. Suppose that we have constructed a pedigree $\hat{\mathcal{P}}_t$ of depth t, and recall that B refers to the length of the gene sequence of an individual. Also recall that a *couple* refers to a pair of mated individuals.

Note that the first step of our recursive procedure equips each couple with an empirical gene sequence of length B where each block can contain *two* distinct symbols. This empirical gene sequence is constructed based on extant data and should be thought of as determining which symbols belong to at least one of the individuals from the couple in a given block. Also, we say that three gene sequences $\sigma, \sigma', \sigma''$ *overlap* in a block if all three sequences have some symbol in common in that block.

Perform the following steps to output a pedigree $\hat{\mathcal{P}}_{t+1}$ of depth $t + 1$.

(1) Collect-Symbols For each couple c in generation t of $\hat{\mathcal{P}}_t$, use the extant genetic data to recover symbols that belong to c as follows.

- Recover a symbol σ in block $b \in [B]$ of c if c has three extant descendants descended from distinct children of c that all share symbol σ in block b.
- Repeat this procedure to recover at most one other symbol $\sigma' \neq \sigma$ for c in block b.

(2) Test-Siblinghood For every triple of couples $c, c', c'' \in \hat{\mathcal{P}}_t$ in generation t, determine c, c', c'' to be (mutually) 'siblings' if and only if at least $0.21B$ of their recovered symbols mutually overlap.

(3) Assign-Parents For every maximal collection $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ of couples in generation t such that every triple in $\mathcal{C}$ consists of mutual siblings, construct a pair of parents in generation $t+1$ that have as children precisely one individual from each couple in $\mathcal{C}$.[2]

After T iterations of the above recursive procedure, we output a pedigree $\hat{\mathcal{P}}_T$ that gives a good approximation to the underlying pedigree that generated the extant genetic data as described in Theorem 5.1. We remark that working with triples as above greatly simplifies our analysis, as discussed in Section 5.2.3.

5.1.6 Model discussion and future directions

Our generative model imposes various constraints on the typical pedigrees that we consider. We discuss these modeling assumptions here and also consider the problem of investigating more general models that could more accurately capture properties of real-world data.

First, we consider the assumption that the size of the alphabet Σ is very large with respect to the size N_T of the founding population. Since a "block" represents the unit of inheritance from a parent[3], this implies that with very high probability all of the founders have distinct symbols in their gene sequences, and no two founders share a common symbol.[4] Our large alphabet assumption is equivalent to the assertion that the founders are unrelated.

Second, the stochastic process describing inheritance in our model has the following biological interpretation. A standard concept in population genetics refers to long-running sequence matches as being *identical by descent* (IBD) if they arose due to inheritance from a common ancestor Thompson [2013]. In contrast, the term *identity by state* refers to the event that two identical tracts in the genome arose by coincidence – via mutations – in two unrelated individuals. Our inheritance model

[2]We perform this step in such a way that every child is assigned at most 2 parents.

[3]Using biology terminology, each block can be considered as an idealized abstraction of a collection of *single-nucleotide polymorphisms* (sites of variation) with high *linkage disequilibrium* (empirical measure of correlation) that are passed from parent to child.

[4]Mathematically, this can be thought of as an improper prior on a countably infinite alphabet Σ.

contains the assertion that each block corresponds to true IBD sequences: if two individuals have the same symbol, we can always identify a common ancestor that gave rise to these symbols.

Third, we recall the hypothesis that every couple has on average α children, where α is a sufficiently large absolute constant independent of the size N_T of the founding population. This ensures that, roughly speaking, every new generation is a factor $\alpha/2$ larger than the previous one. Assuming roughly uniform growth of generations, it is necessary that $\alpha > 0$ — otherwise the population would die out and there would be no extant nodes after T generations. More subtly, it is necessary that $\alpha \geq 2$ — otherwise, via standard results from the theory of branching processes (see, *e.g.* Kimmel and Axelrod [2015]) a founding node has a very low probability of passing on its symbols to the extant. In this situation, even *detection* of such an ancestor from extant genetic data alone is information-theoretically impossible. On the other hand, our assumption that α is a large constant essentially amplifies the signal sent from a founder to the extant, and this simplifies our mathematical analysis.

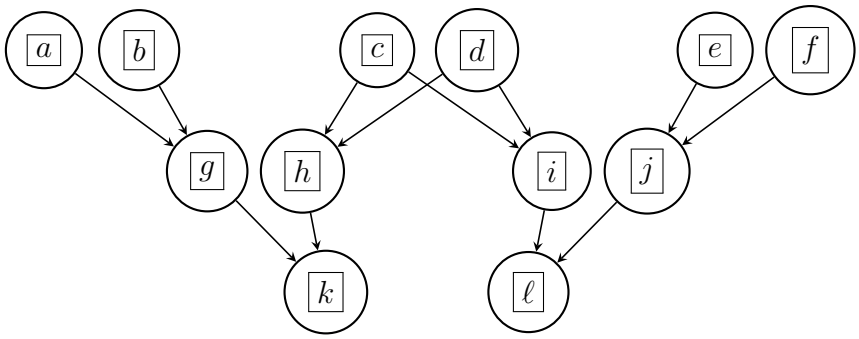
Our first open question considers relaxing the previously discussed assumptions.

Question 5.1. *What theoretical guarantees can be established for pedigree reconstruction in the context of our generative model when α is very close to 2? What about when the size of the alphabet Σ is finite? Can we analyze more generic models of inheritance where blocks are not inherited i.i.d. from parents?*

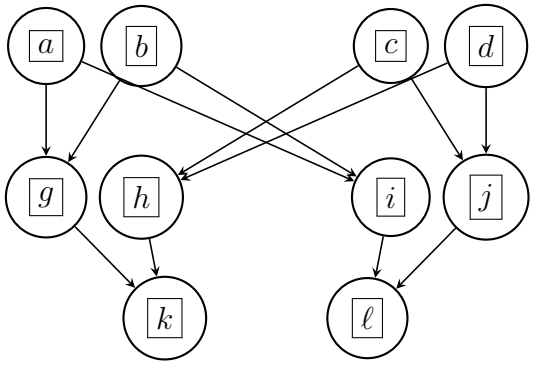
A more subtle consequence of our generative model is *inbreeding*, a term we use to refer to the following phenomena: (1) the presence of multiple lowest common ancestors for a pair of extant nodes, and (2) the presence of mated couples such that the two individuals in the couple have a lowest common ancestor (LCA) (see Definition 5.5 for the formal definition of an LCA). The *degree* of inbreeding qualitatively refers to the frequency of such structures in the pedigree. Moreover, inbreeding as in (2) is mathematically equivalent to having cycles in the pedigree. In general, a higher degree of inbreeding makes the pedigree reconstruction problem more difficult and in some cases information-theoretically impossible (see Section 5.2.1 for detailed examples). Our choice of model allows for some degree of inbreeding, and our algorithm and analysis are carefully tailored to circumvent this obstacle.

Other assumptions inherent in our model include that the pedigree is *graded*, *i.e.*, couples are formed from individuals in the same generation, and *monogamous*: a given individual only mates with one other individual. Furthermore, *mutations* — errors in the transmission of genetic data from parents to offspring — are a central component in biological applications that our current model does not incorporate.

Question 5.2. *What theoretical guarantees can be established for reconstruction of pedigrees in generative models with some combination of (i) a higher degree of inbreeding, (ii) mutations, (iii) non-monogamous mating, and (iv) inter-generational mating?*



(a) three sets of grandparents (cousins, one way)



(b) two sets of grandparents (cousins, two ways)

Figure 5-1: *Simple examples of depth-3 complete pedigrees with a single block. The letters inside the boxes represents the block data. 5-1(a): The overlap probability is* $\Pr(k = \ell) = \frac{1}{8}$. *5-1(b): An altered version of 5-1(a) with only two sets of grandparents, which yields* $\Pr(k = \ell) = \frac{1}{4}$.

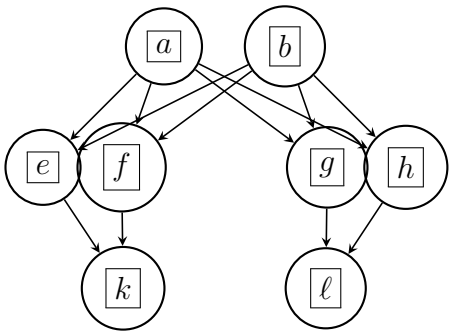
5.2 Inference challenges and techniques

In this section, we detail some of the challenges posed by the reconstruction of pedigrees constructed from our generative model as well as our techniques and analysis for handling them. To develop some intuition for our strategy, we first illustrate some of the properties of pedigrees using concrete examples.

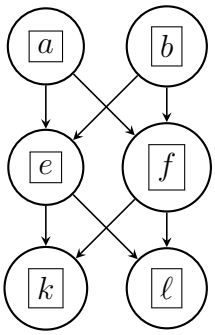
5.2.1 Examples: complications from inbreeding

Recall that two individuals u, v that share the same set of parents are *siblings*. If two individuals share a common subset of grandparents (but not parents), we refer to them as *cousins*.

First consider the pedigrees displayed in Fig. 5-1(a). An important statistic for determining relationships is the correlation between symbols of nodes at the same level. Consider the event E that the left extant shares the same symbol as the right extant. Note that these two extant nodes are cousins sharing a single set of grandparents. The grandparents are the founders in this example, so we assign to each of them a unique symbol ($a \neq b \neq c \neq d \neq e \neq f$). The occurrence of E implies



(a) four siblings begetting cousins.



(b) two siblings begetting siblings.

Figure 5-2: *Two examples of complete pedigrees with inbreeding. The extants in 5-2(a) are cousins, yet they have a coincidence of $\frac{1}{2}$ as if they were generic siblings from unrelated parents. In comparison, 5-2(b) yields $\frac{3}{4}$ which exceeds the coincidence of siblings.*

that $k = c$ or $k = d$ via the left extant receiving a symbol from its right parent; this occurs with probability $\frac{1}{2}$. Conditioned on this occurring, the right extant block ℓ is the same as k with probability $\frac{1}{4}$, so the overall probability that both receive the same symbol is $\frac{1}{8}$.

Compare this to the example shown in Fig. 5-1(b), where the two extant are cousins in two ways (*siblings marrying siblings*). Note that whichever symbol (out of a, b, c, d) that k is, the right grandchild receives the same independently with probability $\frac{1}{4}$. This is an example of a type of inbreeding where two extant nodes have more than one LCA.

The examples in Fig. 5-2 demonstrate how the correlation between extant nodes is boosted due to the presence of inbreeding. Note that in the *generic* case where extant siblings have an ancestral pedigree that is a tree, these individuals have a $\frac{1}{2}$ fraction overlap in their blocks. For comparison, let us compute the probability of coincidence for the two extant nodes in Fig. 5-2(a). The probability that $k = a$, for example, is

$$\Pr(k = a) = \Pr(e = f = a) + \frac{1}{2}\Pr\left(\{e, f\} = \{a, b\}\right) = \frac{1}{4} + \left(\frac{1}{2}\right)^2 = \frac{1}{2}.$$

Since k and ℓ inherit symbols independently from their grandparents, the overall probability is

$$\Pr(k = \ell) = \Pr(k = \ell = a) + \Pr(k = \ell = b) = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \frac{1}{2},$$

which is precisely the probability that two generic siblings inherit the same symbol.

The situation in Fig. 5-2(b) is even more pronounced. The two parents share the same symbol (either a or b) with probability $\frac{1}{2}$ and have different symbols with probability $\frac{1}{2}$. This means that the coincidence probability is now $\frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{3}{4}$:

their correlation between overlaps is much stronger than that of siblings in the generic case.

From the example in Fig. 5-2(a), we conclude that the statistical model of extant data parametrized by pedigrees is unidentifiable. Stated another way, it is information-theoretically impossible to distinguish between siblings and inbred cousins using only extant data. Thus, in order for any algorithm to succeed in reconstructing a large fraction of the pedigree using only extant data, it is necessary to bound the amount of inbreeding in the ensemble of pedigrees of interest. We accomplish this using a careful analysis of our generative model.

5.2.2 Informal analysis of Rec-Gen

In this section, we present a high-level analysis of the Rec-Gen algorithm. Theorem 5.1 states that Rec-Gen yields an accurate reconstruction on 90% of nodes for typical pedigrees from our generative model[5]. Note that a formal statement of this theorem, our main result, is given by Theorem 5.3, and a complete proof is contained in the upcoming sections.

Suppose we construct a pedigree $\hat{\mathcal{P}}_t$ on t generations that, for simplicity of the discussion, *exactly* matches the true, unknown pedigree $\mathcal{P}$ up to generation t. We show that Collect-Symbols, Test-Siblings, and Assign-Parents applied to $\hat{\mathcal{P}}_t$ provide an accurate reconstruction of 90% of the nodes at generation $t+1$. In the remainder of this section we give a high-level argument that the output $\hat{\mathcal{P}}_{t+1}$ satisfies the following conditions:

(i) every individual $\hat{u}$ in $\hat{\mathcal{P}}_{t+1}$ can be identified with a unique individual u in $\mathcal{P}$ at generation $t+1$,

(ii) at most 10% of the nodes in generation $t+1$ of $\mathcal{P}$ are not identified with an individual in $\hat{\mathcal{P}}_{t+1}$, and

(iii) if v is a child of $\hat{u}$ in $\hat{\mathcal{P}}_{t+1}$, then v is a child of u in $\mathcal{P}$.

Recall that for the purposes of reconstruction, we only have access to the genetic data of the extant.

In this discussion, we refer to three couples $c, c', c'' \in \mathcal{P}$ as (mutual) siblings if there exist individuals $u \in c, u' \in c'$, and $u'' \in c''$ such that u, u', and u'' are mutually siblings. A *clique* refers to a collection of couples $\mathcal{C} = \{c_1, \ldots, c_k\}$ such that every triple from $\mathcal{C}$ consists of mutual siblings.

The next two facts are essential to the argument.

(A) If Collect-Symbols recovers symbol σ in block b for a couple c in generation t, then c also has the symbol σ in block b in $\mathcal{P}$ (Claim 5.5).

(B) Collect-Symbols recovers at least 99% of the symbols for at least 99% of the couples in generation t (Lemma 5.17).

[5]We note again that the 90% is for simplicity of exposition, and in reality we can recover an arbitrarily large fraction of nodes. This is made precise in Theorem 5.3.

Together, (A) and (B) imply that for 99% of the couples in generation t, our algorithm gets all of the siblings relationships between these couples correct. To see why, we can use a similar calculation as in the first example of Section 5.2.1 to conclude that the average overlap between the symbols of three individuals that are mutually siblings is 25%. By concentration of binomial random variables about their means, it follows that with high probability, all triples of individuals that are mutually siblings in $\mathcal{P}$ have at least 24.9% mutual overlap between their symbols. A simple union bound combined with (A) and (B) implies that for most triples of individuals in generation t that are mutually siblings, the recovered symbols from Collect-Symbols in those individuals' corresponding couples have overlap at least 21%. Hence, Test-Siblinghood infers correct siblinghood relationships for a majority of triples.

Moreover, our siblings test on the recovered symbols does not have any false-positives:

(C) Test-Siblinghood never misclassifies non-siblings as siblings (Lemma 5.20).

The next and last key fact argues that our naive assignment of parents to individuals in cliques as in Assign Parents is in fact the correct assignment in a typical pedigree. This property holds with very high probability over our generative model.

(D) Let $\mathcal{C} \subset \mathcal{P}$ denote a clique at generation t in the true pedigree. Then there exists a couple $\tilde{c}$, which we refer to as the *parents of* $\mathcal{C}$, in generation $t+1$ of $\mathcal{P}$ that has exactly one child in every couple of $\mathcal{C}$, and no other couple has more than 1 child in $\mathcal{C}$ (Lemma 5.6).

Together, (A), (B), (C), and (D) imply that our reconstruction criteria (i), (ii), and (iii) from the beginning of this section hold, as we now justify. Recall that we already showed (A) and (B) imply that we classify a large fraction of the couples at generation t correctly as siblings. Moreover, part (C) and the transitivity of siblinghood in $\mathcal{P}$ imply that cliques in our reconstruction really correspond to cliques in the truth. By part (D) such cliques have unique parents. Thus, for (i), we identify newly constructed couples $\hat{u} \in \mathcal{P}_{t+1}$ with the *unique* parents $u \in \mathcal{P}$ of the clique formed by the children of $\hat{u}$, further pairing the two individuals in u with those in $\hat{u}$ arbitrarily. With this identification, (iii) follows immediately. To show part (ii), later in the paper we give a sufficient condition for a couple at generation t to have 99% of its symbols collected by Collect-Symbols as in (B) (see Lemma 5.17). Then we show that 90% of individuals in generation $t+1$ have children in such couples (see Proposition 5.5), which proves part (ii). Essentially, this sufficient condition amounts to saying that a couple c at generation t has no inbreeding (cycles) above or below it (*i.e.* among its ancestors or descendants, respectively) and that the pedigree of descendants of c contains a $\alpha/4$-ary tree (see Definition 5.18).

5.2.3 Motivation for using triples

It is tempting to employ a seemingly simpler recursive scheme than the one described in Section 5.1.5 that operates on pairs instead of triples. As an example, consider an alternative recursive procedure such that:

1. Collect-Symbols only uses **pairs** of extant descendants to recover symbols of a couple c,

2. Test-Siblinghood considers only **pairs** of couples at generation t and detects them to be siblings if their strings overlap by at least 49%, and

3. Assign-Parents assigns parents to individuals in maximal collections $\mathcal{C}$ such that every **pair** of couples is (tested as) siblings.

Unfortunately, this simpler approach encounters two major technical complications.

First, working with a pairwise siblings test introduces a problem for the step of assigning parents. Define a *pairwise clique* to be a collection of couples so that every pair of couples passes the pairwise siblings test. With high probability, it turns out in every generation there exist a constant number of pairwise cliques that are not explained in the naive way of assigning to this clique parents that have precisely one child per couple. In particular, in the true pedigree $\mathcal{P}$ it is possible to have three couples that mutually pass the pairwise siblings test, yet there are **three** distinct parent couples each having precisely two children among these three couples. See Fig. 5-3 for an illustration. This type of structure, though rare, occurs a constant number of times in each generation, and thus introduces inherent errors in our reconstruction that accumulate at every step of iteration.

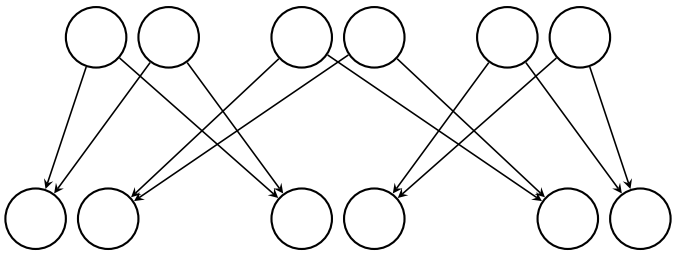


Figure 5-3: *An undesirable subpedigree, where three child couples have mutual siblingship, but they do not mutually share a parent couple.*

A second problem caused by working with pairs arises in the step of collecting symbols. The pairwise version of our algorithm assigns a symbol to a couple if that symbol occurs in two extant descendants that are descended from distinct children of that couple. In our generative model, it turns out that with high probability there are a *logarithmic* number of pairs of extant nodes that have at least two LCA's. For such pairs, the pairwise algorithm does not accurately assign symbols to their reconstructed ancestors. Similar to the previous issue, these errors snowball and make the analysis for proving Theorem 5.1 very difficult.

On the other hand, working with an algorithm using triples as described in Section 5.1.5 makes for a much cleaner analysis and nicer reconstruction guarantee. This innovation circumvents the technical complications of the pairwise version because every clique (recall that this is a collection of couples where every triple consists of mutual siblings) can be explained in a naive way (Lemma 5.6), and in our generative model every triple of extant individuals descended from distinct children of a

given ancestor have that ancestor as their *unique* LCA with very high probability (Lemma 5.8).

5.2.4 Outline of technical arguments

The remainder of the paper, which provides a formal proof of Theorem 5.1, is divided into four parts.

- Section 5.3 provides preliminary definitions and a formal definition of our generative model.
- Section 5.4 proves important properties about the typical network structure of pedigrees from our generative model.
- Section 5.5 proves important properties about the block statistics of the extant nodes in a typical pedigree from our generative model.
- Section 5.6 gives a precise description of Rec-Gen and provides a formal statement and proof of Theorem 5.1.

Specifically, in Section 5.4 we rigorously quantify the degree of inbreeding in typical pedigrees from our model by counting the number of *collisions* (see Definition 5.12 and Lemma 5.4). This has several useful consequences, including that every clique has a unique parent (fact (D) from Section 5.2.2, also see Lemma 5.6) and that the extant individuals used in Collect-Symbols have a unique LCA (see Lemma 5.8). In particular, the latter is key to showing fact (A) from Section 5.2.2.

In Section 5.5, we provide a definition (see Definition 5.18) that essentially characterizes the individuals in $\mathcal{P}$ that are reconstructible via Rec-Gen. We show that couples involving such individuals, referred to as *awesome couples*, transmit many of their symbols to the extant, with high probability (see Lemma 5.17). In particular, awesome couples have at least 99% of their symbols recovered by Collect-Symbols (fact (B) from Section 5.2.2). We also prove an important result for our siblings test: triples of individuals that are not mutually siblings have mutually overlap at most 19% (see Lemma 5.11). This combined with fact (A) from Section 5.2.2 essentially shows that Test-Siblinghood never classifies non-siblings as siblings (fact (C) from Section 5.2.2, see also Lemma 5.20).

Our final section, Section 5.6 ties everything together, following fairly closely the high-level argument presented in Section 5.2.2 to prove the formal version of Theorem 5.1.

5.3 Formal setup and technical preliminaries

5.3.1 Key definitions and terms

Definition 5.1. *A **pedigree** $\mathcal{P} = (V, E)$ is a directed acyclic graph (DAG) with vertices V and edges E where every vertex has indegree at most 2. The collection of*

vertices of indegree zero are referred to as the ***founders****, and the collection of vertices of outdegree zero are referred to as the* ***extant****.*

Definition 5.2. *If the indegree of each vertex in the underlying DAG is either* 2 *or* 0*, then* $\mathcal{P}$ *is called a* ***complete*** *pedigree.*

In this work, we focus on a special family of complete pedigrees that are both *graded* and *monogamous*.

Definition 5.3. $\mathcal{P}$ *is said to be* ***graded*** *if the vertices* $V(\mathcal{P})$ *can be partitioned into* $\bigcup_{i=0}^{T} V_i(\mathcal{P})$ *such that* $V_T(\mathcal{P})$ *are the founders,* $V_0(\mathcal{P})$ *are the extant, and all directed paths* $e_T, \ldots, e_1$ *from* $V_T(\mathcal{P})$ *to* $V_0(\mathcal{P})$ *can be written as a sequence of edges* $e_t = (v_t \to v_{t-1})$ *where* $v_t \in V_t(\mathcal{P})$ *and* $v_{t-1} \in V_{t-1}(\mathcal{P})$ *for each* t*. The founders' index* T *is the* ***depth*** *of the pedigree.*

$\mathcal{P}$ *is said to be* ***monogamous*** *if for every vertex* u *of outdegree* > 0*, there exists a unique vertex* u' *such that* $(u \to v) \in E \iff (u' \to v) \in E$*. The unordered pair* $\{u, u'\}$ *is referred to as a* ***couple****.*

We assume that every non-extant individual in the pedigree is in a couple, and so the number of vertices at each non-extant level is even. This assumption is effectively without loss of generality—if an individual is not in a couple, then it has no descendants, and so we cannot recover information about this individual or even its existence.

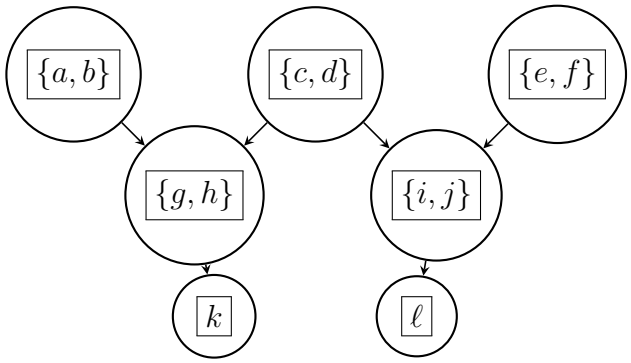
An example of a complete, graded, monogamous pedigree is shown in Fig. 5-1(a). In our model, symbols are passed down from parents to children in a completely symmetric way. Thus, given the data of the children, it is impossible to distinguish the owner of each symbol from amongst the two parents. The goal of this paper is to show how one can provably infer the structure of a complete pedigree from extant genetic data via the reconstruction of the ancestral symbols, modulo *block phasing* (determining which symbol belongs to which parent for each block). Therefore, we introduce the following version of a pedigree which condenses this information.

Definition 5.4. *A* ***coupled*** *pedigree* $\mathcal{Q} = (V_\mathcal{Q}, E_\mathcal{Q})$ *induced by a complete, monogamous pedigree* $\mathcal{P} = (V_\mathcal{P}, E_\mathcal{P})$ *is defined as follows:*

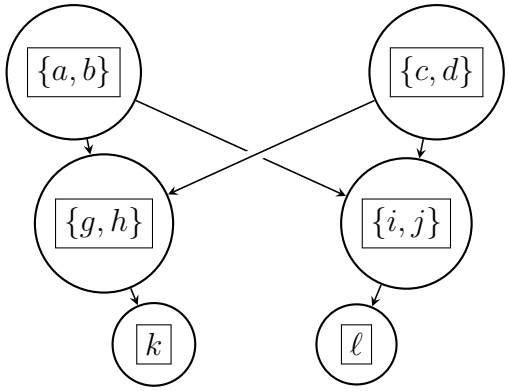
- $V_\mathcal{Q} \subset \binom{V_\mathcal{P}}{2}$ *is obtained by merging couples* $c = \{u, u'\} \subset V_\mathcal{P}$ *into a single node (extant individuals remain singletons), introducing edge multiplicity.*
- $E_\mathcal{Q}$ *is the result of halving the number of resulting copies of each edge after merging couples.*

In particular, a coupled pedigree is also a pedigree. Examples are drawn in Fig. 5-4 in relation to Fig. 5-1, where the complete pedigree 5-1(a) induces a coupled pedigree 5-4(a) and 5-1(b) induces 5-4(b).

The only information that is lost after transforming a complete, monogamous pedigree into a coupled pedigree is the block phasing. Indeed, observe that given the coupled structure $\mathcal{Q} = (V_\mathcal{Q}, E_\mathcal{Q})$, one can easily obtain the individual structure



(a) coupled version of 5-1(a)



(b) coupled version of 5-1(b)

Figure 5-4: *5-1(a) induces coupled pedigree 5-4(a), while 5-1(b) induces 5-4(b).*

$\mathcal{P} = (V_\mathcal{P}, E_\mathcal{P})$ up to block phasing as follows: (1) add the extant individuals in $V_0 \subset V_\mathcal{Q}$ to $V_\mathcal{P}$, (2) for every non-extant node $c \in V_\mathcal{Q}$ add individuals u_c, u'_c to $V_\mathcal{P}$, and (3) given parents c_1 and c_2 of c in $\mathcal{Q}$, add the four edges $u_{c_1} \to u_c, u'_{c_1} \to u_c, u_{c_2} \to u'_c, u'_{c_2} \to u'_c$ to $E_\mathcal{P}$. In addition, if $\mathcal{P}$ is graded, $\mathcal{Q}$ retains a graded structure $V_\mathcal{Q} = V_0(\mathcal{Q}) \cup \cdots \cup V_T(\mathcal{Q})$ so that $V_0(\mathcal{Q})$ are the extant nodes and $V_1(\mathcal{Q}), \ldots, V_T(\mathcal{Q})$ are depth-graded couple nodes. In particular, the graph structure of an individuals pedigree $\mathcal{P}$ uniquely determines the graph structure of its associated coupled pedigree $\mathcal{Q}$ and vice versa.

Given the previous discussion, since our goal is to recover the graph structure of an underlying true pedigree $\mathcal{P}$ given gene sequences of a large number of extant individuals, it suffices to reconstruct the associated coupled pedigree $\mathcal{Q}$.

Furthermore, since the graph underlying a pedigree is a DAG, given a subset S of the pedigree, it is natural to consider the notion of "ancestors" (nodes $\mathsf{anc}(S)$ **from** which there is a directed path to S) and "descendants" (nodes $\mathsf{desc}(S)$ **to** which there is a directed path from S). Also for simplicity, we stipulate that every node v is both a descendant and an ancestor of itself, *i.e.*, $v \in \mathsf{anc}(v)$ and $v \in \mathsf{desc}(v)$. Since the indegree of each node can be more than one, it is possible for two nodes to have more than one "lowest common ancestor". We define this now.

Definition 5.5 (Lowest Common Ancestors)**.** *Let S denote a set of nodes in a pedigree $\mathcal{P}$. The set of* ***lowest common ancestors*** *of S, denoted $\mathsf{LCA}(S)$, consists of all nodes $u \in \mathcal{P}$ such that u is an ancestor of every node in S, and moreover, no descendant of u is an ancestor of every node in S.*

During our analysis, we often restrict our attention to the information that the pedigree contains about the ancestors or descendants of a particular collection of nodes. In particular, we want to exploit (sub)structures that are not too intertwined. The following definitions make these ideas precise:

Definition 5.6 (Subpedigrees)**.** *Let $W \subset V_\mathcal{P}$ denote a subset of nodes of pedigree $\mathcal{P}$. The subgraph of $(V_\mathcal{P}, E_\mathcal{P})$ induced by W is itself a pedigree, which we call the* ***subpedigree*** *of $\mathcal{P}$* ***induced by*** *W.*

Definition 5.7 (Ancestral pedigrees)**.** *Let $W_k \subset V_k(\mathcal{P})$ denote a subset of vertices at level k of a graded pedigree $\mathcal{P}$. The subpedigree induced by $W_k \cup \mathsf{anc}(W_k)$ is the (level k)* ***ancestral subpedigree*** *of $\mathcal{P}$ induced by W_k.*

Definition 5.8 (Descendant pedigrees)**.** *Let $W_k \subset V_k(\mathcal{P})$ denote a subset of vertices at level k of a graded pedigree $\mathcal{P}$. The subpedigree induced by $W_k \cup \mathsf{desc}(W_k)$ is the (level k)* ***descendant subpedigree*** *of $\mathcal{P}$ induced by W_k.*

Definition 5.9 (Tree pedigrees)**.** *A pedigree $\mathcal{P}$ that has no undirected cycles (when the directions of the edges in $E_\mathcal{P}$ are ignored) is called a* ***tree*** *pedigree.*

Note that coupled pedigrees can have edges of multiplicity two, though only in the case where two siblings form a coupled node, which a rare structure in our generative model. In coupled pedigrees, we consider a double edge to be an undirected cycle of length two. Hence, a tree pedigree consists entirely of simple or multiplicity 1 edges.

As we demonstrate (*e.g.* Lemma 5.11), coupled tree pedigrees exhibit a type of correlation decay between blocks that enable us to perform inference on the structure. In contrast, non-tree coupled pedigrees correspond to pedigrees with inbreeding, which can arise in nature and appear in our probabilistic model as well. Section 5.2.1 illustrates examples of such structures. These types of structures introduce challenges for performing inference under our generative model.

5.3.2 Siblings in a pedigree

Note that siblinghood is a transitive relationship: if u, v are siblings and v, w are siblings, then so are u, w. As alluded to in Section 5.2.3, it is important to look at these relationships in *triplets*. We now detail how one can encode this information as a *3-uniform hypergraph*.

Definition 5.10. *A **3-uniform hypergraph** is a pair (V, E) of vertices and a multiset of edges, so that each edge is an unordered triple $\{u, v, w\}$ of vertices in V.*

Definition 5.11. *Let $\mathcal{P}$ be a coupled pedigree of depth T (each non-extant node is a set of a pair of individuals). The **siblinghood hypergraph** G_k of $\mathcal{P}$ at level $k > 0$ is the 3-uniform hypergraph that describes the three-way sibling relationships of its level-k members. For every triple $e = \{c_1, c_2, c_3\}$, the edge multiplicity $n(e; G_k)$ is*

$$n(e; G_k) = \begin{cases} 0 & \text{if } \nexists\ (u_1, u_2, u_3) \in c_1 \times c_2 \times c_3 \text{ such that } u_1, u_2, u_3 \text{ are siblings} \\ 1 & \text{if } \exists \text{ unique } (u_1, u_2, u_3) \in c_1 \times c_2 \times c_3 \text{ such that } u_1, u_2, u_3 \text{ are siblings} \\ 2 & \text{else} \end{cases}$$

The siblinghood hypergraph G_0 is defined similarly, by considering each extant individual u as a degenerate (cardinality 1) couple $c_u = \{u\}$ and applying the above definition (Each hyperedge appears zero or once, never twice).

Recall that a **clique** in a 3-uniform hypergraph is a collection of vertices such that all possible triplets form an edge. The next statement is an observation that follows from the definition of G_k and the transitivity of siblinghood.

Proposition 5.1. *If $c_1, \ldots, c_m$ are level-k couples that respectively contain individuals $u_1, \ldots, u_m$ which are siblings, then $c_1, \ldots, c_m$ form a clique in G_k.*

5.3.3 Probability Tools

We denote a Poisson distribution with mean λ as $\mathsf{Pois}(\lambda)$. We use some basic tools from probability theory in our proof. The first is referred to in literature as *Poisson thinning*, see *e.g.* Lalley [2016].

Proposition 5.2 (Poisson Thinning)**.** *Let $N \sim \mathsf{Pois}(\lambda)$, and let $X_1, X_2, \ldots$ be iid Ber(p) random variables that are independent of N. Then $X = \sum_{i=1}^{N} X_i$ is $\mathsf{Pois}(\lambda p)$-distributed.*

Second, we recall that sums of Poisson random variables are themselves Poissons:

Proposition 5.3. *Fix $N > 0$ and let $X_1, X_2, \ldots, X_N$ be iid $\mathsf{Pois}(\lambda)$ random variables. Then $X = \sum_{i=1}^{N} X_i$ is $\mathsf{Pois}(\lambda N)$-distributed.*

Third, we will invoke the following standard variants of the Chernoff–Hoeffding bounds for sums of Bernoulli random variables:

Theorem 5.2 (Bernoulli tail probability (Chernoff–Hoeffding Bounds))**.** *Let $X = \sum_{i=1}^{N} X_i$ be a sum of independent Bernoulli(p_i) random variables. Let $\mu = \mathbb{E}[X]$. Then*

$$\Pr(X > (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2}{2+\delta}\mu\right)$$
$$\Pr(X < (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2}{2}\mu\right)$$
$$\Pr(|X-\mu| > \gamma N) \leq 2\exp(-2N\gamma^2).$$

Lastly and in the same spirit as the Chernoff–Hoeffding bound, we will use the fact that Poisson distributions also have sub-exponential tails.

Proposition 5.4 (Poisson tail probability)**.** *Let $X \sim \mathsf{Pois}(\lambda)$. Then for any $x > 0$, we have*

$$\Pr(|X-\lambda| \geq x) \leq 2\exp\left(-\frac{x^2}{2(\lambda+x)}\right)$$

For a proof, refer to Chapter 2 of Pollard [2015].

5.4 Structure of Poisson Pedigrees

5.4.1 Model Description

We now describe our simple model for generating a population and its genetic data. The model is best viewed in two stages. In the first stage, we generate the population as well as the pedigree topology $\mathcal{P}_{\mathsf{indiv}}$ on these individuals, and in the second stage, we generate the genetic data given this pedigree structure. Note that the random individual pedigree $\mathcal{P}_{\mathsf{indiv}}$ constructed below is **graded, monogamous,** and **complete**.

Part I: Pedigree topology

1. To generate $\mathcal{P}_{\mathsf{indiv}}$, start with $N_T = N$ founding individuals in V_T and make an arbitrary maximum matching of these individuals to create a set of mated couples. For each couple, generate an independent $\mathrm{Pois}(\alpha)$ number of children, where $\alpha > 0$ is a fixed parameter throughout the entire pedigree. These newly generated individuals form the nodes in V_{T-1}.

2. Repeat the above process to generate the individuals in $V_{T-2}, \ldots, V_0$.

Once we have the population and pedigree structure as above, we generate the genetic data in the following manner.

Part II: Inheritance procedure

1. Each individual u in $\mathcal{P}_{\mathsf{indiv}}$ has a length-B string σ_u (u's **gene sequence**). The string's indices are referred to as **blocks**.

2. For each founding individual u in V_T and for each block $b \in [B]$, each $\sigma_u(b)$ is drawn i.i.d. uniformly from an alphabet Σ. For our model, Σ is an infinite-sized alphabet: we simply require that each block of each founder has a unique symbol.

3. Every other individual v in the population has exactly two parents f and m. Conditioned on σ_f and σ_m, independently over $[B]$, the ith block of v copies $\sigma_f(i)$ with probability 0.5 and $\sigma_m(i)$ with probability 0.5.

Remark 5.1. *We adopt the following conventions in the remainder of the paper.*

1. *We let $\mathcal{P}$ denote the **coupled pedigree** induced (see Definition 5.4) by the randomly generated individual pedigree $\mathcal{P}_{\mathsf{indiv}}$ constructed in Part I above.*

2. *We use the term **coupled node**, or simply **node** when the context is clear, to refer to a vertex of $\mathcal{P}$. We use the term **individual** to refer to an element of $\mathcal{P}_{\mathsf{indiv}}$ contained in a coupled node of $\mathcal{P}$. Unless otherwise explicitly noted, parent-child relationships are taken according to the structure of the coupled pedigree $\mathcal{P}$. That is, given $u, v \in \mathcal{P}$ we use the phrase, "u **is a child of** v," to mean that the couple u contains an individual who is an offspring of the mated couple v. Finally, we say that coupled nodes $u, v \in \mathcal{P}$ are **siblings** if u and v contain individuals who are siblings in $\mathcal{P}_{\mathsf{indiv}}$.*

3. Pr *denotes the probability measure over the randomly generated pedigree $\mathcal{P}$ as well as the random inheritance procedure.*

To given an example of our terminology, there are two individuals in a non-extant coupled node. Each individual is a vertex of $\mathcal{P}_{\mathsf{indiv}}$, and together they form a coupled node, which is a vertex of $\mathcal{P}$. Note that as an artifact of our definitions, extant individuals are both coupled nodes *and* individuals in $\mathcal{P}$. Moreover extant nodes have exactly one parent in $\mathcal{P}$ given by the coupled node containing the individuals comprising that extant individuals biological parents, as determined by our generative model.

To further emphasize the previous remark, recall that by the discussion in Section 5.3.1, there is a unique correspondence between coupled pedigrees and individual pedigrees. Hence, it suffices to give a (partial) reconstruction $\hat{\mathcal{P}}$ of $\mathcal{P}$ to (partially) reconstruct the original individual pedigree $\mathcal{P}_{\mathsf{indiv}}$. Thus the content of our main result Theorem 5.3 and the remainder of this paper primarily work with the coupled pedigree $\mathcal{P}$.

Parameters: For convenience, we collect the various parameters of interest here.

Parameter	Description	Value
N	Size of founding population	
B	Number of blocks for each individual	$\Theta(\log(N))$
α	Expected # of children per couple	$\Theta(1)$
T	Number of generations in population	$\varepsilon \log(N)$, $\varepsilon = O(1/\log(\alpha))$
$\lvert\Sigma\rvert$	Size of block alphabet	∞

We set $B = O(\log(N))$ for a sufficiently large constant. The expected number of children per couple, α, will be set to a sufficiently large constant that is at least 3. Finally, the number of generations T will be set to $\varepsilon \log(N)$, where $\varepsilon > 0$ is sufficiently small with respect to $1/\log(\alpha)$.

5.4.2 Concentration bounds and upper bounds on inbreeding

In this section we quantify the degree of inbreeding in $\mathcal{P}$. To do so, we first describe an alternative description of our generative model. An equivalent procedure for constructing the coupled pedigree structure $\mathcal{P}$ is to (1) sample the generation sizes according to Poisson random variables with appropriate parameters, (2) pair up individuals in each generation at random into coupled nodes, and (3) have coupled nodes choose two parent coupled nodes at random from the previous generation. This is described formally below.

Lemma 5.1. *The (coupled) pedigree $\mathcal{P}$ described in Section 5.4.1 can be equivalently viewed as follows:*

1. *Let $N_T := N$ be the size of the founding population. For i from T to 1: Let $N_i' \stackrel{def}{=} \lfloor N_i/2 \rfloor \cdot 2$ be the number of individuals in couples, and sample $N_{i-1} \sim \mathsf{Pois}(\alpha N_i'/2)$.*

2. *For each level i, match the individuals at level i randomly, leaving out a single individual if N_i was odd.*

3. *For each level i, sample a vector $\vec{v} \in [N_i'/2]^{N_{i-1}}$ from a Multinomial distribution with parameters*

$$(N_{i-1}, (2/N_i', \ldots, 2/N_i')).$$

   For any $k \in [N_i'/2]$, the set of coordinates $\{j : v_j = k\}$ are interpreted as children of the k^{th} couple at level i (and are therefore siblings at level $i-1$).

4. *Convert the resulting pedigree on individuals from steps 1–3 to a coupled pedigree $\mathcal{P}$.*

Proof. The number of vertices at each level in the statement of Lemma 5.1 is the same as the model in Section 5.4.1. This follows by induction. The number of founding vertices N is the same in both models. In the model in Section 5.4.1, the

number of individuals at level $i-1$ is distributed as $\sum_{j=1}^{N_i'/2} X_j$, where the X_j are iid $\mathsf{Pois}(\alpha)$ and N_i' is the number of individuals at level i that are matched. The value of this sum is distributed as $\mathsf{Pois}(\alpha N_i'/2)$ (due to Proposition 5.3), the same as in the statement Lemma 5.1.

The random matching in Step 2 of Lemma 5.1 is the same as the matching in Section 5.4.1.

The final step in the process above assigns individuals in V_{i-1} to parents in V_i by sampling a vector $\vec{v}$ of length N_{i-1} with entries in $[N_i'/2]$ from a multinomial distribution and assigning individuals to parents based on these labels. Indeed, if we look at the number of children of a fixed couple (say, the j^{th} couple in V_i), this is distributed as $\mathsf{Bin}(X, 2/N_i')$, where $X \sim \mathsf{Pois}(\alpha N_i'/2)$. By Poisson thinning (Proposition 5.2), this distribution is simply $\mathsf{Pois}(\alpha)$, which is exactly the distribution of the number of children of the j^{th} couple in Section 5.4.1. □

Next we use tail bounds on Poisson random variables to show that the sizes of each level are well-concentrated with high probability, assuming a sufficiently large size of the initial population. Recall that N_i denotes the number of *individuals* in generation i.

Lemma 5.2 (Concentration of generations)**.** *Fix δ such that $0 < \delta < \alpha/2 - 1$, and suppose that the founding population size N is at least $\alpha/\delta + 1$. Then, for some constant $C_1 = C_1(\delta)$, with probability at least $1 - T\exp(-C_1\alpha N)$ we have that, for all $i \in \{0, \ldots, T-1\}$*

$$(\alpha/2 - \delta)N_{i+1} \leq N_i \leq (\alpha/2 + \delta) \cdot N_{i+1}. \tag{5.1}$$

Remark 5.2. *An immediate corollary of this result is that*

$$(\alpha/2 - \delta)^i \cdot N \leq N_{T-i} \leq (\alpha/2 + \delta)^i \cdot N \tag{5.2}$$

for each $i \leq T$ with high probability.

Proof of Lemma 5.2. Our goal is to upper bound the right-hand-side of

$$\Pr[\text{some } N_j \text{ fails Eq. (5.1)}] \leq \sum_{i=0}^{T-1} \Pr[N_i \text{ fails Eq. (5.1)} \mid N_{i+1} \text{ satisfies Eq. (5.2)}]$$

and so it suffices to show

$$\Pr[N_i \text{ fails Eq. (5.1)} \mid N_{i+1} \text{ satisfies Eq. (5.2)}] \leq 2\exp(-\Theta(\alpha^2(N-1)/(\alpha+\delta))).$$

Consider fixing the number of individuals at level $i+1$ to be an arbitrary number N_{i+1} satisfying Eq. (5.2). We know that the number of individuals at level i is distributed as $N_i \sim \mathsf{Pois}(\alpha N_{i+1}'/2)$. By applying the Poisson tail bound Proposition 5.4, we see that

$$\Pr\left[|N_i - \alpha N'_{i+1}/2| > (\delta/2)N'_{i+1} \mid N_{i+1} \text{ satisfies Eq. (5.2)}\right] \tag{5.3}$$
$$< 2\exp\left(\frac{-(\alpha N'_{i+1}/2)^2}{2(\alpha/2+\delta/2)N'_{i+1}}\right)$$
$$< 2\exp\left(-\alpha\frac{-N'_{i+1}}{4(1+\delta)}\right) \tag{5.4}$$

We now claim that $|N_i - \alpha N_{i+1}/2| > \delta N_{i+1}$ implies that $|N_i - \alpha N'_{i+1}/2| > (\delta/2)N'_{i+1}$, which follows from the facts that $|N_{i+1} - N'_{i+1}| \leq 1$ and that $N_{i+1} \geq N$ (Eq. (5.2)). Namely, assume that $N_i > (\alpha/2+\delta)N_{i+1}$. Then $N_i > (\alpha/2+\delta/2)N'_{i+1}$, since $N_{i+1} \geq N'_{i+1}$. Now assume instead that $N_i < (\alpha/2-\delta)N_{i+1}$. Then

$$\begin{aligned} N_i &< (\alpha/2-\delta)N_{i+1} \\ &\leq (\alpha/2-\delta)(N'_{i+1}+1) \\ &\leq (\alpha/2-\delta/2)(N'_{i+1}) \end{aligned}$$

where in the last line we use the fact that $(\delta/2)N'_{i+1} \geq (\delta/2)(N-1) \geq \alpha/2$.

Hence, we get that

$$\Pr[|N_i - \alpha N_{i+1}/2| > \delta N_{i+1} \mid N_{i+1} \text{ satisfies Eq. (5.2)}] \leq 2\exp\left(-\alpha\left[\frac{N-1}{4(1+\delta)}\right]\right)$$

where we use the fact that $N_{i+1} \geq N$ since N_{i+1} satisfies Eq. (5.2). □

Remark 5.3 (Dependence on δ)**.** *The strategy from this point onwards is to condition on the event from Eq.* (5.1)*. Since this event fails with probability that is exponentially small in N, we lose only an additive* $\exp(-c_\delta \alpha N)$ *probability.*

As mentioned in Section 5.2.1, two nodes may have significantly higher amounts of symbol overlap caused by inbreeding in their ancestral pedigree than would be expected given their distance in the pedigree. This can cause us to reconstruct an incorrect pedigree if we attempt to explain the symbol overlap without accounting for inbreeding; for instance, we may see two nodes and think they are siblings, when in reality they are cousins with inbreeding in their family tree (see Section 5.2.1 for a detailed example). To formally connect different patterns of inbreeding with the amount of spurious symbol overlap they cause, we introduce the notion of *collisions* in an ancestral pedigree. Roughly speaking, triples of coupled nodes with relatively few collisions in their ancestral pedigree do not have many spurious overlaps, which we prove in Section 5.5. We first define collisions and then bound the number that occur under our probabilistic assumptions in Lemma 5.4. We also give an alternative characterization of collisions in Lemma 5.3 that is useful later.

Definition 5.12 (Collisions)**.** *Let $\mathcal{P}$ denote a coupled pedigree. Fix a subset of nodes $A \subset V_k(\mathcal{P})$, where $k \neq T$. If $k > 0$, we say that this collection has z* collisions at level

$k+1$ if the set of parents of A in $\mathcal{P}$ has size $2|A| - z$. If $k = 0$, we say that it has z collisions at level 1 if the set of parents in $\mathcal{P}$ has size $|A| - z$. Write

$$\mathsf{coll}_{k+1}(A) := (\#\ \textit{collisions at level } k+1 \textit{ in } A)$$

Extend the notion of collisions to ancestral subgraphs as follows. If we have nodes $u_1, \ldots, u_J \in V_k(\mathcal{P})$, the number of collisions between the ancestral subpedigrees $\mathsf{anc}(u_j)$ for $j = 1, \ldots, J$ is equal to

$$\mathsf{coll}(u_1, \ldots, u_J) := \sum_{i=0}^{T-k-1} \mathsf{coll}_{i+1}(\mathsf{anc}_i(u_1) \cup \cdots \cup \mathsf{anc}_i(u_J))$$

where $\mathsf{anc}_i(u_j)$ denotes the set of ancestors i levels above u_j.

Lemma 5.3 (Ancestral collisions, alternate characterization)**.** *Let $u_1, \ldots, u_J$ denote a set of nodes that are all at the same level. Consider the subpedigree $\mathcal{T} = anc(u_1, \ldots, u_J)$. Let k_j denote the number of nodes in $\mathcal{T}$ that have outdegree j in the subpedigree $\mathcal{T}$. Then*

$$\mathsf{coll}(u_1, \ldots, u_J) = \sum_{j \geq 2} (j-1)k_j.$$

Proof. Let S denote a set of nodes at level i. Let $k_{ij}(S)$ denote the set of parents of S that have outdegree j in the subpedigree $\mathsf{anc}(S)$. Let $\mathsf{coll}_{i+1}(S)$ denote the number of collisions that S has at level $i+1$. Then we claim that

$$\mathsf{coll}_{i+1}(S) = \sum_j (j-1)k_{ij}(S). \tag{5.5}$$

This is true by induction on the cardinality of S, as we now demonstrate. We prove this assuming that S is a set of non-extant coupled nodes; the case for extant nodes is extremely similar. The base case $|S| = 1$ follows because the unique node $u \in S$ either has two distinct parents, in which case there are no collisions and each has outdegree 1, or u has a single parent, in which case the number of collisions is 1 and the parent has outdegree 2. In both cases Eq. (5.5) holds.

For the inductive step, suppose that Eq. (5.5) is valid for all S with $|S| \leq s$. Now consider S with $|S| = s+1$. Choose an arbitrary $u \in S$ and consider $S' = S \backslash \{u\}$. Observe that by Definition 5.12 and induction:

$$\begin{aligned}
\mathsf{coll}_{i+1}(S) &= 2|S| - |par(S)| \\
&= 2|S'| - |par(S')| + 2|\{u\}| - |par(u) \backslash par(S')| \\
&= \mathsf{coll}(S') + 2 - |par(u) \backslash par(S')| \\
&= \sum_j (j-1)k_{ij}(S') + 2 - |par(u) \backslash par(S')|.
\end{aligned}$$

Therefore, if u has $\ell \in \{0, 1, 2\}$ parents contained in $par(S')$, then

$$\mathsf{coll}_{i+1}(S) = \ell + \sum_j (j-1)k_{ij}(S') = \sum_j (j-1)k_{ij}(S),$$

because each parent of u contained in $par(S')$ increases the degree of some node in S' by 1.

Applying this argument over all levels i to the sets $\cup_{\ell=1}^{J} \mathsf{anc}_i(u_\ell)$, we see by Definition 5.12 and summing over all levels i that Lemma 5.3 holds for coupled nodes. □

In our model and in light of Lemma 5.1, a collision between sets A and B intuitively corresponds to a node in B "choosing" a parent couple that was already chosen by another node in $A \cup B$. This observation lets us bound the number of collisions between the ancestors of 3 nodes with high probability.

Lemma 5.4 (Exponential tail of collisions)**.** *Fix three nodes $u, v, w \in \mathcal{P}$ in the same level k, and let c be a positive integer. Then*

$$\Pr[\mathsf{coll}(u, v, w) \geq c] = O\left(\frac{72^c \cdot 2^{2cT}}{N^c}\right) \tag{5.6}$$

Proof. We show that the probability on the left-hand-side of Eq. (5.6) can be upper bounded by the probability that a binomial random variable with sufficiently small mean is at least c, from which the result follows.

We assume that each level has at least N individuals. This is a high probability event by Lemma 5.2 (which actually describes a much stronger situation). Since we just want an upper bound, we condition such an event and this assumption is made without loss of generality.

Let $S_i := \mathsf{anc}_i(u) \cup \mathsf{anc}_i(v) \cup \mathsf{anc}_i(w)$. Note that $|S_i| \leq 3 \cdot 2^i$, regardless of how many collisions have happened underneath it. The distribution of $\mathsf{coll}(\mathsf{anc}_i(u), \mathsf{anc}_i(v), \mathsf{anc}_i(w))$ is equal to a sum of at most $3 \cdot 2^{i+1}$ Bernoulli random variables, two for each node in S_i, which are indicator random variables that a parent coupled node selected by some node in $u \in S_i$ is the same as a parent coupled node previously selected by $v \in S_i$ (Lemma 5.1). Furthermore, each of these indicator random variables is 1 with probability at most $3 \cdot 2^{T+2}/N$, even conditioned on the previously set random variables—indeed, there are only $3 \cdot 2^{i+1} \leq 3 \cdot 2^T$ parents selected in total, so there are only this many nodes that can be selected from to cause a collision, and there are at least $\lfloor N/2 \rfloor \geq N/4$ coupled nodes at level $i+1$. Therefore, the random variable $\mathsf{coll}(S_i)$ is stochastically dominated by $\mathsf{Bin}(3 \cdot 2^{i+1}, 3 \cdot 2^{T+2}/N)$. Let $X_i \sim \mathsf{Bin}(3 \cdot 2^{i+1}, 3 \cdot 2^{T+2}/N)$. Then we get that

$$\begin{aligned}
\Pr[\mathsf{coll}(u, v, w) \geq c] &= \Pr[\sum_i \mathsf{coll}_{k+i}(S_i) \geq c] \\
&\leq \Pr[\sum_{i=k}^{T-1} X_i \geq c] \\
&\leq \Pr[X \geq c] \qquad (5.7)
\end{aligned}$$

where $X \sim \mathsf{Bin}(3 \cdot 2^{T+1}, 3 \cdot 2^{T+2}/N)$. By bounding the binomial tail and noting that

we take $N > 144 \cdot 2^{2T}$, (Eq. (5.7)) can be bounded by

$$\begin{aligned}\Pr[X \geq c] &\leq \sum_{i=c}^{3\cdot 2^{T+1}} \binom{3 \cdot 2^{T+1}}{i} \left(\frac{3 \cdot 2^{T+2}}{N}\right)^i \\ &\leq \sum_{i=c}^{3\cdot 2^{T+1}} (3 \cdot 2^{T+1})^i \left(\frac{3 \cdot 2^{T+2}}{N}\right)^i \\ &\leq 2 \cdot 72^c \cdot \frac{2^{2cT}}{N^c}\end{aligned}$$

□

In particular, by union bounding over all triples of nodes in the coupled pedigree $\mathcal{P}$, we get the following corollary. Note that there are most $(\alpha/2 + \delta)^T \cdot N$ nodes in the pedigree when we condition on the high-probability event from Lemma 5.2.

Corollary 5.1.

$$\Pr[\exists u, v, w : \mathsf{coll}(u, v, w) \geq 4] = O\left(\frac{(\alpha/2 + \delta)^{3T} 2^{8T}}{N}\right)$$

Since we take the ratio $T/\log(N)$ to be sufficiently small (Section 5.4.1), the probability of the above event is negligible. Hence, we can assume without loss of generality for the rest of the document that the number of collisions in the ancestral trees of any three nodes is at most 3.

Additionally, by applying Lemma 5.4 to a single node (repeated three times) and applying linearity of expectation, we can bound the probability that there are many coupled nodes u with collisions in their ancestral pedigrees $\mathsf{anc}(u)$ using Markov's inequality. We state this as a corollary.

Corollary 5.2. *For any $C > 0$,*

$$\Pr\left[\left|\{u : \mathsf{coll}(u) \geq 1\}\right| \geq C(2\alpha + 4\delta)^T\right] \leq 72/C$$

as long as N is sufficiently large.

Definition 5.13 (d-Richness)**.** *Fix a pedigree $\mathcal{P}$, and let $d \geq 3$ be an integer. All extant nodes in $\mathcal{P}$ are d-rich. For all $k > 0$, a level k-node is d-**rich** if it has at least d children that are d-rich.*

Lemma 5.5 (Most nodes are d-rich)**.** *Fix a constant $0 < \tau < 1$, and let $\delta > 0$ as in Lemma 5.2. As long as N and α are sufficiently large, there exists a constant $C_2 = C_2(\tau, \delta)$ such that with probability $1 - T\exp(-C_2\alpha N)$, at least $(1 - \tau)$ fraction of level-k coupled nodes in $\mathcal{P}$ are d-rich for all k.*

Proof of Lemma 5.5. Let the term "d-poor node" refer to coupled nodes that are not d-rich. Let M_k denote the number of coupled nodes at level k in $\mathcal{P}$. Our goal is to prove an upper bound on the event that there are at least τM_{k+1} d-poor nodes at

level $k+1$, conditioned on the event that there are at least $(1-\tau)M_k$ d-rich nodes at level k.

Let R_k denote the event that there are at least $(1-\tau)M_k$ d-rich nodes at level k. Let E denote the event $(\alpha/2-\delta)M_{k+1} \le M_k \le (\alpha/2+\delta)M_{k+1}$ for all k, which occurs with probability $1-\exp(-C_1\alpha N)$ by Lemma 5.2. We also condition on the sizes of $M_0, \dots, M_T$, abbreviating this conditioning as $M_{0:T}$.

Let S be an arbitrary subset of nodes at level $k+1$ of size $\tau M_{k+1}+1$, and consider the event where S only consists of d-poor nodes. This implies that the number of d-rich children of S is at most $(d-1)(\tau M_{k+1}+1)$. Let X_i be iid Bernoulli RVs, which represent indicators for the event where the ith d-rich child chooses at least one of its parents to be in S. Note that $\Pr(X_i = 1) = \left(1 - \left(1 - \frac{|S|}{M_{k+1}}\right)^2\right) > \frac{|S|}{M_{k+1}}$.

$$\begin{aligned}
&\Pr(S \text{ only has } d\text{-poor nodes} \mid R_k, E, M_{0:T}) \\
&\quad\le \Pr\left[\sum_{i=1}^{(1-\tau)M_k} X_i \le (d-1)|S| \;\middle|\; M_{0:T}\right] \\
&\quad\le \exp\left[-\frac{(1-\tau)M_k|S|}{2M_{k+1}}\left(1 - \frac{(d-1)M_{k+1}}{(1-\tau)M_k}\right)^2\right] \quad \textit{(Chernoff–Hoeffding Bound)}
\end{aligned}$$

Observe that there are $\binom{M_{k+1}}{|S|} \le \left(\frac{e}{\tau}\right)^{\tau M_{k+1}+1}$ many choices for S. To apply a union bound, it suffices for α to be large enough so that $\frac{(1-\tau)M_k}{M_{k+1}}\left(1 - \frac{(d-1)M_{k+1}}{(1-\tau)M_k}\right)^2 \approx (1-\tau)\alpha(1-\frac{d-1}{(1-\tau)\alpha})^2$ looks linear in α. In that case, we obtain a bound of the form

$$\begin{aligned}
&\Pr(\text{at least } \tau M_{k+1} \; d\text{-poor nodes at level } k+1 \mid R_k, E, M_{0:T}) \\
&\qquad\le \exp\left(-CM_{k+1}\alpha\right).
\end{aligned}$$

Therefore, we may write

$$\begin{aligned}
&\Pr(\text{at least } (1-\tau) \text{ fraction of } d\text{-rich at all levels}) \\
&\qquad\ge (1 - e^{-C_1\alpha N}) \prod_{k=1}^{T} (1 - \exp(-CM_{k+1}\alpha)) \\
&\qquad\ge 1 - \exp(-C_1\alpha N) - \sum_{k=0}^{T-1} \exp(-CN(\alpha/2-\delta)^k\alpha) \\
&\qquad\ge 1 - T\exp(-C_2\alpha N)
\end{aligned}$$

for an appropriate constant C_2 depending only on τ and δ.

□

Lemma 5.6 (Cliques have unique parents)**.** *Let G_k denote the siblinghood hypergraph at level k. Let $\delta > 0$ be as in Lemma 5.2. For a constant $C_3 = C_3(\delta)$, with probability at least $1 - \frac{1}{N}e^{C_3 T \log\alpha}$, for all hypercliques $\mathcal{C} \subset G_k$ with at least one hyperedge, there is a unique node at level $k+1$ that is a parent of every node in $\mathcal{C}$. We refer this node*

*as the **parent** of $\mathcal{C}$.*

Proof. By Proposition 5.1, a hyperclique corresponds to a set of coupled nodes that contain a set of mutual siblings, where each couple has at least one of the siblings in it. This establishes that there is a coupled node at level $k+1$ that is at least one parent of every node in $\mathcal{C}$. In the case where $\mathcal{C}$ is a hyperclique of extant nodes, we are done: every node in $\mathcal{C}$ is an individual and has exactly one parent coupled node.

If $\mathcal{C}$ is at a higher level, note that there can be at most two parents for $\mathcal{C}$, as defined above. The reason is that any individual has exactly one parent couple, and since there are only two individuals in a couple, there cannot be three parent couples each with one child in each couple in $\mathcal{C}$.

Next we show that if there are two coupled nodes, both of which are parents of $\mathcal{C}$, then there must be many collisions among the ancestors of $\mathcal{C}$, and therefore we can rule this out as a low-probability event. Since $\mathcal{C}$ has at least one hyperedge, we know that $|\mathcal{C}| \geq 3$. This means that any arbitrary set of three nodes from $\mathcal{C}$ must have at least $6 - 2 = 4$ collisions by Definition 5.12—but Corollary 5.1 shows that with probability $O\left(\frac{(\alpha/2+\delta)^{3T}2^{8T}}{N}\right)$, this does not occur anywhere in the pedigree. □

Lemma 5.7 (Disjointness of maximal cliques)**.** *Let G_k denote the siblinghood hypergraph at level k. For $k = 0$, each extant node is contained in a unique maximal clique, and moreover, the maximal cliques in G_0 are vertex disjoint (and thus, also edge-disjoint). For $k > 0$, each node is contained in at most two maximal cliques. Moreover, with probability $1 - \frac{1}{N}e^{C_3 T \log \alpha}$, the maximal cliques in G_k are edge-disjoint.*

Proof. Note that maximal cliques in the siblinghood hypergraph correspond to maximal sets of siblings. The claim for extant nodes is relatively trivial - extants are individuals, and so the maximal sets of siblings partition the set of extant nodes.

For $k > 0$, since each individual in a coupled node has one pair of parents, a coupled node can have at most two parents. Thus it can be part of at most two sets of siblings. Hence, it is part of at most two maximal cliques.

Finally, we need to establish that the maximal cliques in G_k are edge-disjoint. To do this, it suffices to show that the intersection between any two maximal cliques is less than 3, so there can be no hyper-edge. Indeed, if three nodes that are simultaneously in two maximal cliques, these three nodes would themselves form a clique with two different parents in level $k+1$, which occurs with probability at most $1 - \frac{1}{N}e^{C_3 T \log \alpha}$ by Lemma 5.6. □

5.4.3 The joint LCA and its uniqueness

The next two lemmas are crucial in Section 5.6 to show that we can accurately collect symbols for accurately reconstructed coupled nodes. Here we define the *joint lowest common ancestor*, which is a special type of LCA for a triple of coupled nodes.

Definition 5.14. *Let u, v, w denote coupled nodes in $\mathcal{P}$. We say that u, v, w have a **joint LCA** z if it holds that $z \in \mathsf{LCA}(u, v, w)$ and there exist distinct children c_u, c_v, c_w of z so that for all $x \in \{u, v, w\}$, c_x is an ancestor of x.*

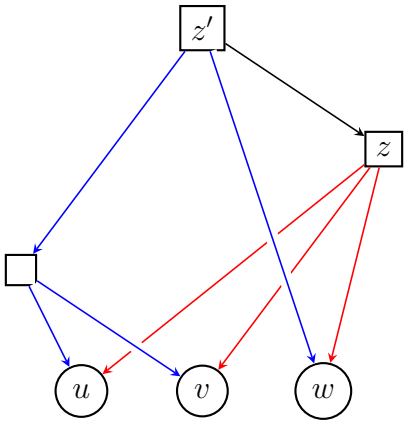


Figure 5-5: *"Proof-by-picture" of Lemma 5.9.*

Lemma 5.8 (Joint LCA is unique)**.** *Suppose that each triple of coupled nodes in $\mathcal{P}$ has at most 3 collisions. Further suppose that u, v, w have a joint LCA $z \in \mathsf{LCA}(u, v, w)$. Then z is the unique LCA of u, v, w.*

Proof. For the sake of contradiction, suppose that u, v, w have another LCA $z' \neq z$. By the definition of LCA, z' is neither an ancestor nor a descendant of z.

If z' is a joint LCA of u, v, w, then both z and z' have outdegree 3 in $\mathsf{anc}(u, v, w)$, which by Lemma 5.3 implies that $\mathsf{anc}(u, v, w)$ has at least $2 \times (3 - 1) = 4$ collisions.

If z' is not a joint LCA, then z' has outdegree 2 in $\mathsf{anc}(u, v, w)$. Moreover, there exists a unique lowest node $y \in \mathsf{desc}(z') \cap \mathsf{anc}(u, v, w)$ that is an ancestor of precisely two nodes in $\{u, v, w\}$. In particular, y has outdegree at least 2 in $\mathsf{anc}(u, v, w)$. Observe that the nodes y, z, z' are all distinct. Hence by Lemma 5.3, the number of collisions is at least $2 \times (2 - 1) + 1 \times (3 - 1) = 4$.

In either case, $\mathsf{anc}(u, v, w)$ has at least 4 collisions, which is a contradiction. □

Lemma 5.9 (Inheritance paths go through LCA)**.** *Suppose that each triple of coupled nodes in $\mathcal{P}$ has at most 3 collisions. Further suppose that $u, v, w \in \mathcal{P}$ have an LCA z. Let z' denote a strict ancestor of z. Then for some $x \in \{u, v, w\}$, all paths from z' to x in $\mathcal{P}$ pass through z.*

Proof. To draw a contradiction, suppose that for all $x \in \{u, v, w\}$ that z' has a path to x that does not go through z. Suppose further, without loss of generality, that z' is the lowest node in $\mathcal{P}$ that is an ancestor of z and has this property.

Let $\mathcal{T}$ denote a spanning tree on $\mathsf{desc}(z) \cap \mathsf{anc}(u, v, w)$ (red edges in Fig. 5-5). Also select a spanning tree $\mathcal{T}'$ on the union of all paths from z' to u, v, w that do not go through z (blue edges in Fig. 5-5). Observe that z' has outdegree at least 2 in $\mathcal{T}'$. Since z' also has a path to z, then z' has outdegree at least 3 in $\mathsf{anc}(u, v, w)$. Moreover, $\mathcal{T}$ has 2 collisions. Since z' is not contained in $\mathcal{T}$, we conclude by Lemma 5.3 that $\mathsf{anc}(u, v, w)$ has at least $2 + 1 \times (3 - 1) = 4$ collisions. The first terms accounts for the collisions in $\mathcal{T}$, and the second applies Lemma 5.3 to z'. This is a contradiction. □

Note that by Corollary 5.1, Lemmas 5.8 and 5.9 hold for all triples $u, v, w \in \mathcal{P}$ with high probability.

5.5 Lemmas that enable reconstruction

In this section, we prove bounds on "overlap statistics" previously explored in Section 5.2. Since we now have switched to talking about coupled pedigrees, we re-define its notion now.

Definition 5.15 (Diploid blocks)**.** *Let $\mathcal{P}_{\mathsf{indiv}}$ induce the coupled pedigree $\mathcal{P}$. Given (haploid) gene sequences $(\sigma_u)_{u \in V(\mathcal{P}_{\mathsf{indiv}})}$, we associate with each non-extant couple $v = \{v_1, v_2\}$ node a **diploid sequence** σ_v defined in terms of each block b as a multiset $\sigma_v(b) := \sigma_{v_1}(b) \cup \sigma_{v_2}(b)$. Each extant node's block is thought of as a singleton set.*

Definition 5.16 (Diploid overlap)**.** *Three diploid sequences $\sigma, \sigma', \sigma''$ **overlap** in block b if*

$$\sigma(b) \cap \sigma'(b) \cap \sigma''(b) \neq \emptyset.$$

*The term **fraction of mutual overlaps** between coupled nodes u, v, w in refers to the statistic*

$$\frac{\#\text{ overlapping blocks of } \sigma_u, \sigma_v, \sigma_w}{B} = \frac{|\{b \in [B] : \sigma_u(b) \cap \sigma_v(b) \cap \sigma_w(b) \neq \emptyset\}|}{B}.$$

5.5.1 Distinguishing siblings from non-siblings

In this section, we establish the following high-probability separation condition for triples of coupled nodes at the same level:

- if u, v, w are mutually siblings, they overlap in at least $1/4$ fraction of blocks.
- if u, v, w are not mutually siblings, they overlap in at most $3/16$ fraction of blocks.

In order to reconstruct the pedigree, we perform inference on the underlying pedigree structure from the symbols at the extant level. The key step of our reconstruction algorithm is to infer which triples of nodes are mutually siblings based on the overlap between their reconstructed symbols. The conditions stated above justify using the number of overlapping symbols in triples as a statistic for determining siblinghood. The first fact (Lemma 5.10) is easy to prove. In contrast, the second fact (Lemma 5.11) is rather non-trivial; we prove it using casework.

Lemma 5.10 (Symbol overlap in siblings)**.** *With probability $1 - O(\alpha^{3T} N^3 \exp(-\gamma^2 B))$, the fraction of mutual overlap in symbols between any triple of coupled nodes $u, v, w \in \mathcal{P}$ that are mutually siblings is at least $\frac{1}{4} - \gamma$ for any arbitrarily small $\gamma > 0$.*

Proof. It suffices to consider the overlap of the individuals u_1, v_1, w_1 in u, v, w, respectively, that are siblings, *i.e.*, u_1, v_1, w_1 have a common parent in $\mathcal{P}_{\mathsf{indiv}}$. We claim that the expected fraction of overlap for u_1, v_1, w_1 is at least $1/4$. Indeed, any individual symbol at the parent (couple) node survives to all three children with probability $1/8$, and there are $2B$ symbols at the parent (one per block per member of the couple).

The Chernoff–Hoeffding bound gives that for any fixed triple (u, v, w) of siblings, the probability that it has less than $1/4 - \gamma$ mutual overlap is at most $\exp(-\gamma^2 B)$. To be explicit, let X_i denote the indicator of an overlap between u, v, w in block b.

$$\begin{aligned}
\Pr(\text{average overlap} < 1/4 - \gamma) &= \Pr\left(\frac{1}{B}\sum_{i=1}^{B} X_i < 1/4 + \gamma\right) \\
&= \Pr\left(\frac{1}{B}\sum_{i=1}^{B} (X_i - \mathbb{E}[X_i]) < 1/4 - \mathbb{E}[X_1] + \gamma\right) \\
&\le \Pr\left(\frac{1}{B}\sum_{i=1}^{B} (X_i - \mathbb{E}[X_i]) < -\gamma\right) \\
&\le 2\exp(-2B\gamma^2).
\end{aligned}$$

In the second line we use that X_i are i.i.d., in the third line we use that the expectation is at least $1/4$, and to finish we apply Chernoff–Hoeffding. A union bound over all $O((\alpha^T N)^3)$ triples of siblings yields the result. □

Lemma 5.11 (Symbol overlap in non-siblings)**.** *Fix $\gamma > 0$. With probability $1 - O(1/N_T) - O(\alpha^{3T} N^3 \exp(-\gamma^2 B))$, every triple of coupled nodes u, v, and w that are at the same level but are **not** mutual siblings share overlap in less than $\frac{3}{16} + \gamma$ fraction of their symbols.*

Proof of Lemma 5.11

Remark 5.4. *In this proof, we condition on the high probability event from Corollary 5.1 that all triples u, v, w of coupled nodes have at most 3 collisions in their ancestral subpedigree $\mathsf{anc}(u, v, w)$.*

It is clear that if u, v, w are completely unrelated, then their mutual overlap is zero, since we assume an infinite alphabet. If u, v, w have a common ancestor, then typically their ancestral pedigree has two collisions, and all triples have at most three collisions in their ancestral pedigree by our conditioning in Remark 5.4. We refer to triples with three collisions as being *inbred* and think of the extra collision as the *site* of inbreeding, a notion that we later formalize in this section.

Recall the definition of tree subpedigree (Definition 5.9), which we refer to simply as a *tree* in what follows. Also recall that an edge of multiplicity 2 in a pedigree is considered to be an undirected cycle of length 2. Thus, a tree subpedigree consists only of simple (multiplicity 1) edges. Our strategy for proving Lemma 5.11 follows the recipe below for casework.

1. u, v, w have exactly two LCAs, and the ancestral pedigree of u, v, w is a tree.
2. u, v, w have exactly one LCA, and the LCA has a cycle above it.
3. u, v, w have exactly one LCA, and the ancestral pedigree of u, v, w is a tree.

4. u, v, w have exactly one LCA, and the ancestral pedigree of u, v, w contains a cycle that is not completely above the LCA.

We now assert that the above cases cover all possibilities; this is proven in the next two claims.

Claim 5.1. *For u, v, and w to have a single LCA, their ancestors must have at least 2 collisions.*

Proof. All three nodes need a common ancestor, which means there are at least 2 collisions are present in $\mathsf{anc}(u, v, w)$. □

Claim 5.2. *The nodes u, v, and w have at most two LCAs, with two LCAs only if $\mathsf{anc}(u, v, w)$ has three collisions. Furthermore, if there are two LCAs, then $\mathsf{anc}(u, v, w)$ is a tree pedigree.*

Proof. By the previous claim, creating a single LCA for three nodes requires 2 collisions in $\mathsf{anc}(u, v, w)$. By definition, one LCA cannot be an ancestor of another LCA. This means there must be at least one more collision in $\mathsf{anc}(u, v, w)$ to create the second LCA, bringing the total number of collisions required in $\mathsf{anc}(u, v, w)$ to three. This immediately yields the final part of the claim by Remark 5.4.

To establish that there are at most two LCAs, suppose we add a third LCA. Then by the same argument, this LCA cannot be an ancestor of either of the two other LCAs, and so there must be another collision to explain it. This leads to four collisions among the ancestors, which we have ruled out. □

We now upper bound the expected overlap between u, v and w by doing the above casework on the structure of their ancestral pedigrees. We simply upper bound the expected overlap, relying on the independence of inheritance in the different blocks so that we can apply a Chernoff–Hoeffding bound.

Lemma 5.12 (Case 1: exactly two LCAs)**.** *Suppose that u, v, and w have exactly two LCAs. Then the expected fraction of mutual overlap is at most $1/8$.*

Proof. Fig. 5-6 illustrates the topology of interest. First we note that neither of the LCAs can have repeated symbols, since their ancestral pedigrees contain no collisions. Consider the ancestral pedigree from u, v, and w up to any one particular LCA, noting that this pedigree is a tree by Claim 5.2. Any configuration containing u, v, w and their ancestors leading up to that LCA has at least 5 edges, since u, v, w are not mutual siblings. Therefore, the probability that a single symbol propagates from that LCA to all of u, v, and w is $\leq (1/2)^5 = 1/32$, which yields an expected $1/16$ fraction of overlap since there are $2|B|$ symbols at the LCA (since it is a coupled node). Since there are two such LCAs, the expectation is at most $1/8$. □

In the remaining cases, we assume there is exactly one LCA. Note that any common symbols across u, v, and w must be present in this LCA—if u, v, and w inherit a symbol that is not present in this LCA, then by tracing their paths of inheritance for the symbol we can find another LCA. However, this does not guarantee that *all*

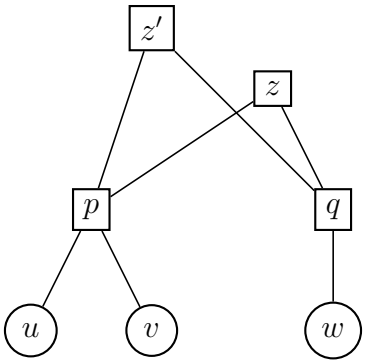


Figure 5-6: *The topologies of Lemma 5.12 with two LCAs. Others are obtained by swapping the roles of u, v, w.*

common symbols in u, v, and w can be traced back to inheritance from the LCA— if there is inbreeding, some nodes in $\{u, v, w\}$ can potentially inherit a symbol via an ancestor of the LCA through a path does not go through the LCA, while the rest inherit it from the LCA.

Lemma 5.13 (Case 2: one LCA with cycle above)**.** *Suppose that u, v, and w have exactly one LCA z. Furthermore, this LCA has at least one collision in its ancestral pedigree. Then the fraction of mutual overlap is at most* $1/8$ *in expectation.*

Proof. We know that u, v, and w, must have at least two distinct parents between them that are connected to z (else z would be their parent). This means there are at least two edges in the graph between z and the parents of u, v, and w, and at least three edges between u, v, and w and their respective parents.

Since we know there are at most three collisions among the ancestors of u, v, and w, there can be only one collision in the ancestral pedigree of z, and the presence of this collision means there are no other collisions in $\mathsf{anc}(u, v, w)$. Therefore, each of the parent couples of u, v, and w have an individual that is unrelated to z, and so there are no repeated symbols within any of the parent couples. So even if the parents were to get 100% overlap in the blocks due to inheritance from z, it holds that u, v, and w inherit at most $1/8$ fraction of these blocks on expectation.

Finally, all common symbols between u, v, and w must have been inherited from z— if a common symbol was instead inherited by some $x \in \{u, v, w\}$ from some ancestor of z, this would create a fourth collision in $\mathsf{anc}(u, v, w)$. □

Lemma 5.14 (Case 3: one LCA and $\mathsf{anc}(u, v, w)$ is a tree)**.** *Suppose u, v, and w have exactly one LCA and $\mathsf{anc}(u, v, w)$ is a tree. Then the fraction of mutual overlap is at most* $1/16$ *in expectation.*

Proof. The lack of any cycles in $\mathsf{anc}(u, v, w)$ means that all inheritance of common symbols comes from the lone LCA z. Any such union of paths from z to u, v and w forms a directed tree with at least five edges; see Fig. 5-7. In addition, z has two distinct symbols in every block. Therefore, for any particular symbol the probability

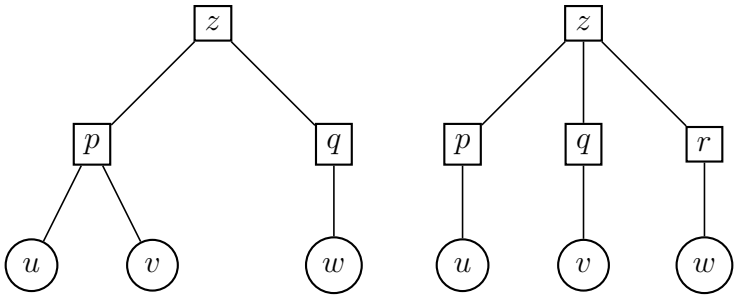


Figure 5-7: *Exhaustive list of topologies from Lemma 5.14, up to re-labelling of u, v, w. Each edge represents a path of length > 1.*

that all three of u, v, w inherit it is $\leq (1/2)^5 = 1/32$, which yields an expected fraction of at most $1/16$ overlapping blocks.

□

The final case is the most complicated one to analyze.

Lemma 5.15 (Case 4: one LCA with cycle not completely above)**.** *Suppose u, v, and w have exactly one LCA and $\mathsf{anc}(u, v, w)$ contains a cycle that does not lie completely above $z = \mathsf{LCA}(u, v, w)$. Then the fraction of mutual overlap is at most $3/16$ in expectation.*

As an aid in proving Lemma 5.15, it is helpful to first identify the **"most recent" inbred** node. We make this notion precise now.

Definition 5.17 (Witness)**.** *We call a node $g \in \mathsf{anc}(u, v, w)$ a* witness to inbreeding *or simply a* witness *if g is the lowest node in $\mathsf{anc}(u, v, w)$ that is part of an undirected cycle.*

Lemma 5.16 (Unique witness)**.** *Under the conditions of Lemma 5.15, there exists a unique witness in $\mathsf{anc}(u, v, w)$. Moreover, this witness lies strictly below the LCA z.*

Proof. We know that $\mathcal{T} := \mathsf{anc}(u, v, w)$ is not a tree, so there exists a cycle in $\mathcal{T}$. We show that there can only be one cycle. Suppose that there exist two cycles $\mathcal{C}, \mathcal{C}'$ in $\mathcal{T}$. Then we claim that $\mathsf{coll}(u, v, w) \geq 4$.

Consider a spanning tree $\mathcal{T}'$ of $\mathcal{T}$. Then $\mathcal{T}'$ has two collisions. Moreover, $\mathcal{T}' \cup \mathcal{C}$ contains a single cycle, so we conclude that there exists a node in $\mathcal{T}'$ whose outdegree is increased by one upon adding the edges from $\mathcal{C}$ to $\mathcal{T}'$ (Otherwise, $\mathcal{T}' \cup \mathcal{C}$ would still be a tree). Therefore, by Lemma 5.3, $\mathcal{T}' \cup \mathcal{C}$ has three collisions. By similar reasoning and using that $\mathcal{C} \neq \mathcal{C}'$, we conclude that $\mathcal{T}' \cup \mathcal{C} \cup \mathcal{C}'$ has 4 collisions. Since $\mathcal{T}' \cup \mathcal{C} \cup \mathcal{C}' \subset \mathcal{T}$, we conclude that $\mathcal{T}$ has at least 4 collisions. But under our conditioning, no subpedigree has 4 or more collisions. It follows that in Lemma 5.15 there is exactly one cycle in $\mathcal{T}$, and thus, exactly one witness.

To prove the final statement, note that if the witness is located above z in $\mathsf{anc}(u, v, w)$, then the cycle lies completely above z. □

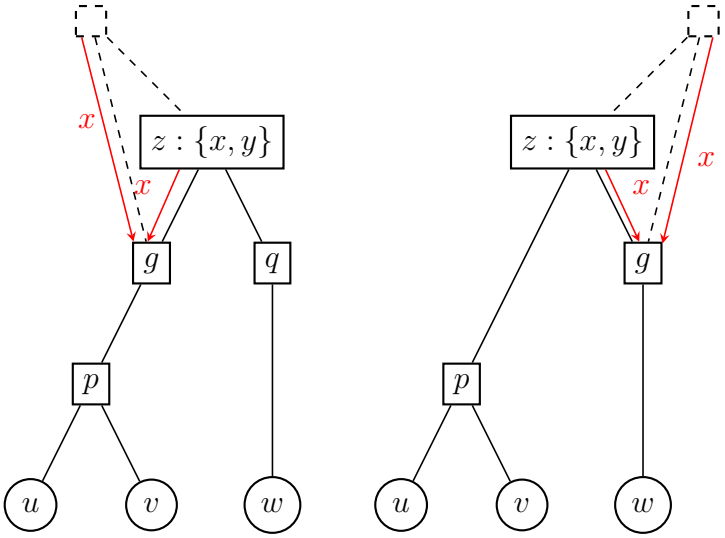


Figure 5-8: *Example of structures being analyzed in the proof of Lemma 5.15, Subcase 2. Here $\{x, y\}$ depict the symbols of the LCA z in a specific block. The red edges delineate the inheritance events (possibly occurring simultaneously) of a common symbol x.*

Proof of Lemma 5.15. Consider u, v, w and the subpedigree $\mathcal{T} = \mathsf{anc}(u, v, w)$ consisting of the ancestors of u, v, w. Recall that z is the unique LCA of u, v, w. By Lemma 5.16, there is a unique witness $g \in \mathcal{T}$, which is the lowest node in the unique cycle occurring in $\mathcal{T}$.

Subcase 1: $\mathsf{LCA}(u, v) = \mathsf{LCA}(v, w) = \mathsf{LCA}(u, w) = \mathsf{LCA}(u, v, w)$.

Without further loss of generality, suppose that the witness g lies along the path from u to z. Then it follows that there is a unique path from v to z in $\mathcal{T}$. Otherwise, there would exist two cycles in $\mathcal{T}$, which is a contradiction as this would lead to 4 collisions in $\mathcal{T}$. Similarly, there is a unique path from w to z in $\mathcal{T}$. Moreover, $\mathsf{anc}(z)$ is a tree. It follows that the subpedigree $\mathsf{anc}(v, w)$ of the ancestors of v and w is a tree. Observe that z is at least two levels above v, w, and by the topology of this subcase, there are at least 4 edges in the tree subpedigree from z to v and w. This implies that the expected overlap between v and w is at most $2 \cdot (1/2)^4 = 1/8$. Thus the expected overlap between u, v, w is at most the expected overlap between u and v, which is bounded by $1/8$.

Subcase 2: Without loss of generality, $\mathsf{LCA}(u, v) \neq \mathsf{LCA}(u, v, w)$.

Let $p = \mathsf{LCA}(u, v)$. Either g is on the branch that leads to u and v, or it is on the branch that leads to w. First, suppose that g is on the branch that leads to u and v. Then we may further assume g is on the path from z to p. For if, say, g is on the path from p to u, then $\mathsf{anc}(v, w)$ is a tree, in which case we can argue as in Subcase 1 that the mutual expected overlap between u, v, w is at most $1/8$.

Therefore, it suffices to consider the cases g is on the path from z to p or g is on the path from z to w (Fig. 5-8). In the first case, the descendants of g form a tree with at least two edges. Moreover, there is a unique node q at the same level

as g in $\mathcal{T}$, and this individual is located on the path from z to w. Let $\sigma(z) = \{x, y\}$ denote the (distinct) symbols of z in a given block. By these facts, symmetry, and conditional independence of inheritance,

$$\begin{aligned}
&\Pr[\sigma(u) \cap \sigma(v) \cap \sigma(w) \neq \emptyset] \\
&\leq 2 \Pr[\sigma(g) = \{x, x\},\, x \in \sigma(u) \cap \sigma(v)] \Pr[x \in \sigma(q), x \in \sigma(w)] \\
&\quad + 2 \Pr[\sigma(g) = \{x, y\}, x \in \sigma(u) \cap \sigma(v)] \Pr[x \in \sigma(q), x \in \sigma(w)] \\
&\leq 2 \times \left(\frac{1}{4} \times 1\right) \times \left(\frac{1}{2} \times \frac{1}{2}\right) + 2 \times \left(\frac{1}{2} \times \frac{1}{4}\right) \times \left(\frac{1}{2} \times \frac{1}{2}\right) \\
&= \frac{3}{16}.
\end{aligned}$$

The second line includes a factor of 2 to account for either x or y being passed down to u, v, w. The terms in the third line are ordered to correspond to the events in the two lines above. In particular, we have by conditional independence of inheritance that

$$\Pr[\sigma(g) = \{x, x\}] \leq 1/4$$

because there are at most 2 paths from z to g, and each has probability at most $1/2$ of passing down x. The bound

$$\Pr[\sigma(g) = \{x, y\}] \leq 1/2$$

holds similarly.

Now suppose that g is on the path from z to w. Then

$$\begin{aligned}
\Pr[\sigma(u) \cap \sigma(v) \cap \sigma(w) \neq \emptyset] &\leq 2 \Pr[x \in \sigma(u) \cap \sigma(v)] \Pr[x \in \sigma(w)] \\
&\leq 2 \cdot \frac{1}{8} \cdot \frac{3}{4} = \frac{3}{16}.
\end{aligned}$$

Above, we used the fact that tree pedigree from z to u, v has at least 3 edges. We also used the fact

$$\Pr[x \in \sigma(w)] \leq \frac{3}{4},$$

which holds because there are at most two paths to w from z, each path has probability at least $1/2$ of not passing down x, and so by conditional independence of inheritance, the probability that both paths do not pass down x is at least $1/4$. □

Finally, to finish the proof of Lemma 5.11 using Lemmas 5.12, 5.13, 5.14, and 5.15, note that in all four cases the expected overlap between coupled nodes u, v, w is at most $3/16$. Thus, the probability that u, v, w mutually share more than $3/16 + \gamma$ fraction of symbols in all cases is at most $2\exp(-2B\gamma^2)$ by Chernoff–Hoeffding, similar to the analysis of Lemma 5.10. Union bounding over all $O((\alpha^T N)^3)$ possible triples gives an $O(\alpha^{3T} N^3 \exp(-B\gamma^2))$ upper bound of the chance that there is some triple with at least $3/16 + \gamma$ overlap. By also ruling out the bad event in Corollary 5.1 (which occurs with probability $O(1/N_T)$), we obtain the desired upper bound.

5.5.2 Which ancestors are reconstructible?

In this section, we characterize nodes that are of importance in our analysis: couples whose history *lacks inbreeding (e.g. graph structure is reconstructible using blocks)* and *have ample extant information (e.g. blocks are recoverable).* We present this in two parts respectively in Definition 5.18 and Definition 5.19.

Definition 5.18 (Awesome Node)**.** *Call a node in the pedigree $\mathcal{P}$* awesome *if:*

1. *It is d-rich.*
2. *It is not an ancestor of any extant node that has a collision within its own ancestral pedigree (including itself).*

Definition 5.19 (b-goodness)**.** *Let $b \in [B]$ be a specific block. Say that a coupled node v in a pedigree $\mathcal{P}$ is b-**good** if v has at least two sets of three extant descendants x_1, y_1, z_1 and x_2, y_2, z_2 in $\mathcal{P}$ such that:*

1. *v is a joint LCA of x_1, y_1, z_1 and is a joint LCA of x_2, y_2, z_2.*
2. *x_1, y_1, and z_1 all have the same symbol σ_1 in block b, and x_2, y_2, and z_2 all have the same symbol σ_2 in block b.*
3. *$\sigma_1 \neq \sigma_2$.*

We furthermore define every extant node to be b-good, for all $b \in [B]$.

We now deliver the main message of this section: *most nodes have these properties*, given the assumptions of our model (Proposition 5.5 and Lemma 5.17). Therefore, this characterization enables a natural reconstruction algorithm (Section 5.6).

Proposition 5.5 (Many awesome nodes)**.** *Let $d > 0$ (as in Definition 5.18) be a constant, let α be a sufficiently large constant with respect to d, and let N be sufficiently large with respect to both d and α. With probability at least $1 - \alpha^{-\Omega(T)}$, in every layer of the pedigree at least $1 - 1/d$ fraction of the nodes are awesome.*

Proof. Since α and N are sufficiently large with respect to d, we can apply Lemma 5.5 with $\tau = 1/(2d)$ and $\delta = d$. This tells us that at least $1 - 1/(2d)$ fraction of nodes in each layer are d-rich with probability $1 - T\exp(-C_2\alpha N)$, where the constant $C_2 = C_2(1/(2d), d)$ depends only on d.

Applying Corollary 5.2 with $C = \alpha^T$, there are at most $\alpha^{O(T)}$ nodes at the extant level with collisions in their ancestral pedigree, with probability $1 - \alpha^{-\Omega(T)}$. This means there are at most $2^T \cdot \alpha^{O(T)}$ *ancestors* of these nodes. It follows that the number of nodes that are d-rich but not awesome is at most $2^T \cdot \alpha^{O(T)}$. This is at most $\frac{N}{2d}$, provided N is sufficiently large with respect to d and α and we take $\varepsilon = T/\log N$ to be small with respect to $1/\log(\alpha)$.

The first probability $1 - T\exp(-C_2\alpha N)$ is exponentially small in N, while the second probability $1 - \alpha^{-\Omega(T)}$ is exponentially small in $T = \varepsilon \log N$. Therefore, the probability of both events occurring simultaneously can be lower bounded by $1 - \alpha^{-\Omega(T)}$, by taking the constant hidden in the Ω to be slightly smaller than what is found in the previous paragraph. □

Lemma 5.17 (Awesome implies b-good)**.** *Let $d > 0$ (as in Definition 5.18) be a sufficiently large constant. With probability $1 - \exp(-\Omega(B))$ over the symbol inheritance process, every awesome coupled node in $\mathcal{P}$ is b-good for at least 99% fraction of blocks $b \in [B]$.*

The figure "99%" is an arbitrary choice for simplification. It can be replaced by anything arbitrarily close to 1, which changes the constant factor of $\Omega(B)$ found in the lemma above. To prove Lemma 5.17, first we need a structural claim about awesome nodes:

Claim 5.3. *For any awesome coupled node, the subpedigree formed by it and its awesome descendants contains an induced d-ary tree that goes down to the extant level.*

Proof of Claim 5.3. First, we show that this subpedigree has no undirected cycles within it, which establishes the tree structure. Then, we argue that each node has d children within this subpedigree.

Suppose that there an undirected cycle within this subpedigree. We show that this implies the presence of a collision within the subpedigree, contradicting the awesomeness of all nodes in the subpedigree. Note that there must be a node within this subpedigree with a cycle in its ancestral pedigree - for instance, take the node at the lowest level within the cycle. Applying Lemma 5.3 to this awesome node, we see it has a collision among its ancestors, which contradicts condition 2) of Definition 5.18.

Now we establish that each node has at least d children in the subpedigree. An awesome coupled node v has at least d children that are d-rich, since it is d-rich itself. Furthermore, none of these children have descendants with collisions in their ancestral pedigree, so they are all awesome, which finishes the proof. □

Proof of Lemma 5.17. Every awesome coupled node in $\mathcal{P}$ has exactly 2 distinct symbols in each block. Indeed, assume for contradiction that there is an awesome coupled node v with a block in which it only has one distinct symbol. Due to the infinite alphabet assumption, we know that we can trace any symbol in a block back to a unique founder. Hence, there must be a collision in the ancestral pedigree of v, which is a contradiction with condition 2) of (Definition 5.18).

Now we can proceed with showing that every awesome coupled node is b-good for 99% fraction of blocks $b \in [B]$. Fix an awesome node v and a block $b \in [B]$.

We use condition (1) of awesomeness to show that, with probability tending to 1 as $d \to \infty$, there exist two sets of three extant nodes that both have v as a joint LCA, where the first set has a symbol σ_1 in block b, and the second set has a symbol $\sigma_2 \neq \sigma_1$.

Towards this end, let us follow the inheritance of σ_1 among an induced d-ary tree of awesome descendants, as guaranteed by Claim 5.3. The inheritance follows a broadcast process with copy probability $1/2$ on this d-ary tree. The probability that the symbol makes it to at least three distinct children of v, and this symbol in turn survives to the extant nodes can be expressed as

$$\left(1 - (1/2)^d \left(1 + d + \binom{d}{2}\right)\right) \cdot c_{d,1/2} \tag{5.8}$$

where $c_{d,1/2}$ refers to the survival probability of percolation on the d-ary tree with copy probability $1/2$. The first term refers to the probability that the symbol is inherited by at least 3 of the d awesome children of v. Additionally, these three extant nodes have v as an LCA, as they have paths of inheritance from v that do not all intersect at any other node.

Naturally, Eq. (5.8) also gives the probability that σ_2 is similarly inherited. Furthermore, from standard results about Galton-Watson processes (see e.g. Kimmel and Axelrod [2015]), we know that as $d \to \infty$, $c_{d,1/2} \to 1$. Hence, we conclude that Eq. (5.8) tends to 1 as $d \to \infty$. Thus it follows from the union bound the probability that there exist two sets of three extant nodes that both have v as a lowest common ancestor, the first set has σ_1 in block b, and the second set has σ_2, also tends to 1 as $d \to \infty$.

Hence, given a specific block b, the probability that an awesome coupled node is b-good is at least 0.995. The high probability of this occurring for all blocks follows from a standard Chernoff–Hoeffding bound. □

5.6 Reconstructing the Pedigree

On the following page, we provide pseudocode for Rec-Gen which is the proposed reconstruction procedure, with details of the inner procedures following it (Collect-Symbols, Test-Siblinghood, and Assign-Parents). Note that for the first iteration of Rec-Gen, we do not need to collect symbols as the extant genetic data is given to us. Thus we simply test siblinghood at iteration $k = 1$ by using the true gene sequences.

The goal of the rest of this section is to prove the correctness of Rec-Gen. We now formally state our guarantee:

Theorem 5.3 (Main theorem, formal)**.** *Let $\hat{\mathcal{P}}$ be the depth-T coupled pedigree output by the algorithm Rec-Gen, applied to the gene sequences in $V_0(\mathcal{P})$. With probability tending towards 1 as $N \to \infty$, $\hat{\mathcal{P}}$ is an induced subpedigree of $\mathcal{P}$ such that $|V_i(\hat{\mathcal{P}})| \geq \eta(\alpha)|V_i(\mathcal{P})|$ for all levels $i \in \{0, \ldots, T\}$, where $\eta(\alpha) \to 1$ as $\alpha \to \infty$. The probability is over the randomness of the coupled pedigree $\mathcal{P}$ and the inheritance procedure with parameters set as in Section 5.4.1.*

We define $\eta(\alpha) := 1 - (1/d(\alpha))$ where, for a given value of α, $d(\alpha)$ is defined to be the largest value of d such that Proposition 5.5 holds. Observe that $d(\alpha) \to \infty$ as $\alpha \to \infty$ because Proposition 5.5 holds for arbitrarily large values d. Therefore, $\eta(\alpha) \to 1$ as $\alpha \to \infty$.

We make use of the following high-probability events, provided α is a large enough constant so that $d = d(\alpha)$ satisfies the hypothesis of Lemma 5.17, N is sufficiently large with respect to α, the total number of generations is $T = \varepsilon \log N$, where $\varepsilon = O(1/\log \alpha)$, and the gene sequence length is $B = \Omega(\log N)$.

Algorithm 1 Reconstruct a depth-T coupled pedigree, given extant individuals V_0.

procedure REC-GEN(T, V_0)
  $\hat{\mathcal{P}} \leftarrow (V = V_0, E = \varnothing)$ ▷ Extant Pedigree with no edges
  for $k = 1$ to T **do**
    if $k > 1$ **then**
      for all vertices v in level $k-1$ of $\hat{\mathcal{P}}$ **do**
        COLLECT-SYMBOLS($v, \hat{\mathcal{P}}$)
    $\hat{G} \leftarrow$ TEST-SIBLINGHOOD($\hat{\mathcal{P}}$)
    ASSIGN-PARENTS($\hat{\mathcal{P}}, \hat{G}$)
  return $\hat{\mathcal{P}}$

Algorithm 2 Empirically reconstruct the symbols of top-level node v in $\mathcal{P}$.

procedure COLLECT-SYMBOLS($v, \hat{\mathcal{P}}$)
  for all blocks $b \in [B]$ **do**
    repeat
      Find extant triple (x, y, z) such that:
      1) v is a joint LCA of x, y, z,
      2) x, y, and z all have the same symbol σ in b, and
      3) σ is not yet recorded for block b in v.
      Record the symbol σ for block b in v.
    until two distinct symbols are recorded for block b, or no such triple exists.

Algorithm 3 Perform statistical tests to detect siblinghood

procedure TEST-SIBLINGHOOD(depth $(k-1)$ pedigree $\hat{\mathcal{P}}$)
  $V \leftarrow \{v \in V_{k-1}(\hat{\mathcal{P}}) : (\#\text{ fully recovered blocks of } v) \geq 0.99|B|\}$
  $E \leftarrow \varnothing$
  for all distinct triples $\{u, v, w\} \subset 2^V$ at level $k-1$ **do**
    if $\geq 0.21|B|$ blocks b such that $\hat{s}_u(b) \cap \hat{s}_v(b) \cap \hat{s}_w(b) \neq \varnothing$ **then**
      $E \leftarrow E \cup \{u, v, w\}$
  return $\hat{G} = (V, E)$ ▷ 3-wise sibling hypergraph

Algorithm 4 Construct ancestors, given top-level 3-way sibling relationship.

procedure ASSIGN-PARENTS($\mathcal{P}, G$)
  repeat
    $\mathcal{C} \leftarrow$ ANY-MAXIMAL-CLIQUE(G)
    Remove one copy of all hyper-edges in $\mathcal{C}$ from G.
    If $|\mathcal{C}| \geq d$, attach a level-k parent in $\mathcal{P}$ for all nodes from $\mathcal{C}$.
  until no maximal cliques of size $\geq d$ remain in G.

Proposition 5.6 (Key Reductions)**.** *With probability tending towards* 1 *as* $N \to \infty$*, the pedigree* $\mathcal{P}$ *satisfies:*

1. *For each level* k*, each clique of* G_k *has a single parent (Lemma 5.6).*
2. *For each level* k*, the maximal cliques of* G_k *are edge-disjoint, in such a way that each* $v \in V_k(\mathcal{P})$ *is contained in at most two maximal cliques (Lemma 5.7).*
3. *Each triple* u, v, w *of nodes, has at most 3 collisions (Corollary 5.1), implying*
   (a) *their joint LCA is unique (Lemma 5.8), and*
   (b) *all inheritance paths for some node* $x \in \{u, v, w\}$ *go through the unique LCA (Lemma 5.9).*
4. *The fraction of overlap is at least* 24.9% *for siblings in* $\mathcal{P}$ *while for non-mutual siblings it is at most* 18.85% *(Lemmas 5.10 and 5.11).*
5. *For each level* k*, at least* $\eta(\alpha)$ *fraction of nodes in* $V_k(\mathcal{P})$ *are awesome (Proposition 5.5).*
6. *If* $u \in V(\mathcal{P})$ *is awesome, then it is* b*-good for* 99% *of blocks* $b \in [B]$ *(Lemma 5.17).*

The "probability tending towards 1" portion of Theorem 5.3 can be quantified via a union bound on the probability of failure of any of the events in Proposition 5.6, while the "$|V(\hat{\mathcal{P}})| \geq \eta(\alpha)|V(\mathcal{P})|$" guarantee comes from the fact that we recover 100% of the awesome nodes in conjunction with Condition 5. With this as a simplification, we proceed with the proof of Theorem 5.3.

The upcoming lemma (Lemma 5.18) proves the correctness of the very first iteration (depth 1 from depth 0), and therefore serves as the base case. The inductive step (Lemma 5.19) is presented immediately afterwards. For the remainder of this section, we write $\hat{\mathcal{P}}_k$ to denote the depth-k reconstructed pedigree after the kth iteration of Rec-Gen, ($\hat{\mathcal{P}}_0$ is the depth-0 pedigree of all the extant nodes). In contrast, let $\mathcal{P}_k$ denote the subpedigree of $\mathcal{P}$ (the ground truth) induced by graded levels V_0 up to V_k.

Lemma 5.18. *Let* $\hat{G}_0$ *denote the estimated* 3*-regular siblinghood hypergraph for the extant nodes (line 7 of Test-Siblinghood). Consider the pedigree* $\hat{\mathcal{P}}_1$ *created by Assign-Parents applied to* $(\hat{\mathcal{P}}_0, \hat{G}_0)$*. Then there exists an injective homomorphism* $\phi : \hat{\mathcal{P}}_1 \to \mathcal{P}_1$ *so that the induced subgraph on* $\phi(\hat{\mathcal{P}}_1)$ *is isomorphic to* $\hat{\mathcal{P}}_1$*. Moreover,* $\phi(\hat{\mathcal{P}}_1)$ *contains* $A_{\leq 1}$*, where* $A_{\leq 1}$ *is the set of awesome nodes at levels* ≤ 1 *in* $\mathcal{P}$*.*

Proof. Let G_0 denote the true siblinghood hypergraph on extant nodes with at least two siblings. By Condition 4, we have that $\hat{G}_0 \cong G_0$. Since both graphs have the same set of vertices, we simply write $\hat{G}_0 = G_0$.

This gives a natural, explicit characterization of ϕ. For an extant node $v \in V_0(\mathcal{P}_1)$, define $\phi(v) = v$ so that it is the identity map on the extant. Given couple $\hat{u} \in V_1(\hat{\mathcal{P}}_1)$, define $\phi(\hat{u})$ to be the parent couple $u \in V_1(\mathcal{P}_1)$ of the children of $\hat{u}$. The

condition $\hat{G}_0 \cong G_0$ implies that at least one such choice for u exists, and moreover by Condition 1, u is the unique parent.

ϕ is injective: Let $\hat{u}, \hat{v} \in V_1(\hat{\mathcal{P}}_1)$ with $\hat{u} \neq \hat{v}$. At the extant level, the maximal cliques in G_0 are vertex disjoint by Condition 2. Hence, the children of $\hat{u}$ and the children of $\hat{v}$ have empty intersection. Moreover in $\mathcal{P}_1$, vertex-disjoint maximal cliques have distinct parents. Therefore, $\phi(\hat{u}) \neq \phi(\hat{v})$, as desired.

ϕ respects edges: We already know that $(\hat{u}, v) \in E(\hat{\mathcal{P}}_1) \implies (\phi(\hat{u}), v) \in E(\mathcal{P}_1)$. Now suppose that $(\phi(\hat{u}), v)$ is an edge in $\mathcal{P}_1$ for $\hat{u} \in V_1(\hat{\mathcal{P}}_1)$ and $v \in V_0(\hat{\mathcal{P}}_1)$. Since u is in the image of ϕ, it follows that u has at least 3 children w, x, y that passed the siblings test in our algorithm. If v is one of w, x, y, we're done, so suppose not. By Condition 5.10, the extant triples $\{v, w, x\}$, $\{v, x, y\}$, and $\{v, w, y\}$ all have at least 24% overlap. Therefore, v, w, x, y form a clique in $\hat{G}_0$, and line 5 of Assign-Parents states that $\hat{u}$ is a parent of all four, so $(\hat{u}, v)$ is an edge in $\hat{\mathcal{P}}_1$.

The image of ϕ contains the awesome nodes in $\mathcal{P}_1$: This part is trivially true for the extant nodes, so consider only the awesome nodes $A_1 \subset V_1(\mathcal{P}_1)$. By definition, any awesome node $u \in V_1(\mathcal{P}_1)$ is d-rich. Since $d \geq 3$, the children of u form a maximal clique of size at least 3 in G_0. Therefore, Assign-Parents creates a parent $\hat{u}$ for these children in $\hat{\mathcal{P}}_1$, which gives the pre-image of u. □

Lemma 5.19. *Let $k \geq 2$ and suppose that we are given $\hat{\mathcal{P}}_{k-1}$. Assume that there exists an injective homomorphism $\phi : \hat{\mathcal{P}}_{k-1} \to \mathcal{P}_{k-1}$ which satisfies*

1. *$\phi|_{\hat{\mathcal{P}}_0} \equiv Id$,*
2. *$\phi(\hat{\mathcal{P}}_{k-1}) \subset \mathcal{P}_{k-1}$ induces a subgraph isomorphic to $\hat{\mathcal{P}}_{k-1}$, and*
3. *$\phi(\hat{\mathcal{P}}_{k-1})$ contains the awesome nodes in sets $A_0, A_1, \ldots, A_{k-1}$.*

Let $\hat{\mathcal{P}}_k$ be the level-k extension of $\hat{\mathcal{P}}_{k-1}$, via lines 4 through 7 of Rec-Gen. Then there exists a level-k extension of the map $\phi : \hat{\mathcal{P}}_k \to \mathcal{P}_k$ with the same properties.

We prove this in two stages. The first part (Lemma 5.20) asserts that we reconstruct the sibling relationships correctly, while the latter (Lemma 5.21) assures that the cliques of this estimated siblinghood hypergraph are actually the faithful, "largest possible" groupings of siblings.

Lemma 5.20. *Assume the hypotheses of Lemma 5.19, and let $\hat{G}_{k-1}$ be the estimated siblinghood hypergraph constructed by Test-Siblinghood, line 7, on input $\hat{\mathcal{P}}_{k-1}$. Then the subgraph of G_{k-1} induced by $\phi(\hat{G}_{k-1})$ is isomorphic to $\hat{G}_{k-1}$, and moreover $\phi(\hat{G}_{k-1})$ contains all of the awesome nodes A_{k-1} at level $k-1$.*

The upcoming statements (Claim 5.4, Claim 5.5 and Claim 5.6) are pivotal for the proof of Lemma 5.20.

Definition 5.20. *For an awesome node $u \in \mathcal{P}_k$, its* awesome subtree *is the subgraph of $\mathcal{P}_k$ that is the union of all paths from u to extant nodes that consist entirely of awesome nodes.*

Claim 5.4. *Suppose that there is a reconstruction map $\phi : \hat{\mathcal{P}}_{k-1} \to \mathcal{P}_{k-1}$ satisfying the hypotheses in Lemma 5.19. Then for any awesome node $u = \phi(\hat{u}) \in V_{k-1}(\mathcal{P}_{k-1})$, its awesome subtree S_u satisfies $\phi^{-1}(S_u) = \mathsf{desc}(\hat{u})$.*

Proof of Claim 5.4. Note that Line 5 of Assign-Parents ensures that every node in $\hat{\mathcal{P}}_{k-1}$ is d-rich. Since ϕ is an injective homomorphism, it follows that every node in $\phi(\mathsf{desc}(\hat{u}))$ is also d-rich in $\mathcal{P}$. Furthermore, u being awesome implies that all of its descendants are awesome in $\mathcal{P}$, since none of its descendants can have collisions in its ancestral pedigree (Definition 5.18). By the definition of the awesome subtree (Definition 5.20), it holds that $\phi(\mathsf{desc}(\hat{u})) \subseteq S_u$.

For the other direction ($\phi(\mathsf{desc}(\hat{u})) \supseteq S_u$), let $v \in V_0(\mathcal{P})$ be an extant node so that there is a path from u to v consisting only of awesome nodes. By condition 3 of Lemma 5.19, all of the nodes along this path are in the image of ϕ. □

Claim 5.5. *Let ϕ be as in Lemma 5.19, and let $u = \phi(\hat{u})$ for some $\hat{u} \in V_{k-1}(\hat{\mathcal{P}}_{k-1})$. Suppose that in block b the symbols $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are recovered for $\hat{u}$ by applying Algorithm 1 to $\hat{u}$. Then it holds that u also has symbols $\hat{\sigma}_1, \hat{\sigma}_2$ in block b.*

Proof of Claim 5.5. For $i = 1, 2$, suppose that nodes $x_i, y_i, z_i \in V_0(\hat{\mathcal{P}}_0) = V_0(\mathcal{P})$ have the symbol $\hat{\sigma}_i$ in block b and are used by Collect-Symbols to recover $\hat{\sigma}_i$ in block b of $\hat{u}$. Recall that x_i, y_i, z_i are all descended from distinct children of $\hat{u}$. Let $\phi(\hat{\mathcal{P}}_{k-1})$ induce subpedigree $\mathcal{Q}$ in $\mathcal{P}$.

By the hypotheses of Lemma 5.19, $\mathcal{Q} \cong \hat{\mathcal{P}}_{k-1}$ and so u must be a common ancestor of x_i, y_i, z_i in $\mathcal{Q}$. By line 4 of Collect-Symbols and because $\mathcal{Q} \cong \hat{\mathcal{P}}_{k-1}$, $\hat{u}$ – and therefore u – is their joint LCA. With respect to $\mathcal{P}$, Conditions 3a and 3b tell us the much stronger condition that u is their only LCA, and that all paths in $\mathcal{P}$ from any common ancestor of x_i, y_i, z_i to x_i (without loss of generality) must pass through u. Therefore, if x_i, y_i, z_i all inherit symbols $\hat{\sigma}_i$ in block b, the symbol $\hat{\sigma}_i$ must have passed through block b of u via the infinite symbols assumption. □

Claim 5.6. *Let ϕ be as in Lemma 5.19, and let $u = \phi(\hat{u})$ for some $\hat{u} \in V_{k-1}(\hat{\mathcal{P}}_{k-1})$. Suppose that u is awesome in $\mathcal{P}$. If u is b-good and has symbols σ_1, σ_2 in block b, then Collect-Symbols recovers the symbols σ_1 and σ_2 for $\hat{u}$ in block b.*

Proof of Claim 5.6. By Claim 5.5, we only need to show that at least two symbols in block b are reconstructed by Collect-Symbols applied to $\hat{u}$. Note that b-goodness implies $\sigma_1 \neq \sigma_2$.

By b-goodness of u, as in the proof of Lemma 5.17, there is a witnessing triple for each of the σ_i contained in the extant of the awesome subtree S_u. By Claim 5.4, $\mathsf{desc}(\hat{u})$ also contains these witnesses. Since extant nodes are the exact same in $\mathcal{P}$ compared to $\hat{\mathcal{P}}_{k-1}$ by hypothesis 1 of Lemma 5.19, Collect-Symbols applied to $\hat{u}$ recovers σ_1, σ_2 in block b. □

Proof of Lemma 5.20. By assumption, $\phi : \hat{G}_{k-1} \to G_{k-1}$ is injective. To first see that ϕ is a hypergraph homomorphism, let $\hat{u}, \hat{v}, \hat{w} \in V_{k-1}(\hat{\mathcal{P}}_{k-1})$ be distinct nodes satisfying line 2 of Test-Siblinghood, and let $u = \phi(\hat{u}), v = \phi(\hat{v})$, and $w = \phi(\hat{w})$ denote their counterparts in $\mathcal{P}$.

Suppose that u, v, w are not mutually siblings. By Condition 4, u, v, w have at most $0.1885|B|$ mutually overlapping blocks. By Claim 5.5, for all $\hat{x} \in \{\hat{u}, \hat{v}, \hat{w}\}$, the symbols reconstructed for $\hat{x}$ in block b using Collect-Symbols are a subset of the symbols in block b of $x := \phi(\hat{x}) \in \{u, v, w\}$. Therefore, $\hat{u}, \hat{v}, \hat{w}$ have mutually overlapping symbols in at most $0.1885|B|$ blocks. Since $0.1885 < 0.21$, Test-Siblinghood does not place a hyperedge between $\hat{u}, \hat{v}, \hat{w}$ in $\hat{G}_1$.

To conclude that the induced subgraph $\phi(\hat{G}_{k-1})$ is isomorphic to $\hat{G}_{k-1}$, it remains to show that if u, v, w are mutual siblings in $\mathcal{P}$, then $\{\hat{u}, \hat{v}, \hat{w}\}$ is a hyperedge in $\hat{G}_{k-1}$. Note that 99% of the blocks of $\hat{u}, \hat{v}, \hat{w}$ were recovered by Collect-Symbols by the definition of $\hat{G}_{k-1}$, and by Claim 5.5, the symbols of $\hat{u}, \hat{v}, \hat{w}$ in block b are a subset of the symbols of u, v, w, respectively, in block b. By Condition 4, the mutual overlap between the siblings u, v, w is at least $0.249|B|$. Thus, by a union bound on the occurrence of 1%-fraction of unrecovered blocks, the mutual overlap between $\hat{u}, \hat{v}, \hat{w}$ is at least $(0.249 - 0.03)|B| \geq 0.21|B|$. Therefore, Test-Siblinghood constructs a hyperedge on $\hat{u}, \hat{v}, \hat{w}$, as desired. It follows that the induced subgraph $\phi(\hat{G}_{k-1})$ is isomorphic to $\hat{G}_{k-1}$.

Finally, we show that the awesome nodes A_{k-1} are fully contained in $\phi(\hat{G}_{k-1})$. By Condition 6, awesome nodes are b-good. Now apply Claim 5.6, to conclude that Collect-Symbols reconstructs 99% of the blocks in each awesome node u, so $u \in \hat{G}_{k-1}$ according to Line 2 of Test-Siblinghood. □

Lemma 5.21. *Let $\mathscr{C}$ denote the maximal (hyper)cliques in the subgraph of G_{k-1} induced by $\phi(\hat{G}_{k-1})$, and let $\mathscr{C}_{algo}$ denote the (hyper)cliques probed by Assign-Parents applied to $\hat{G}_{k-1}$. Given $\mathcal{C} \in \mathscr{C}_{algo}$, define $\phi(\mathcal{C})$ to be the set given by the image of $\mathcal{C}$ under ϕ. Then ϕ is a bijection between $\mathscr{C}_{algo}$ and $\mathscr{C}$.*

Proof. By Lemma 5.20, the subgraph H induced by $\phi(\hat{G}_{k-1})$ is isomorphic to $\hat{G}_{k-1}$. Hence, it suffices to show that the cliques probed by Assign-Parents applied to H are precisely the maximal cliques of H. Recall that by Condition 2, the maximal cliques in H are edge-disjoint, and every node of H is involved in at most 2 cliques.

It is helpful to imagine the cliques $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_M \in \mathscr{C}_{\text{algo}}$ as being listed out in the same order that they are probed by Assign-Parents, indexed by timesteps $m = 1, 2, \ldots, M$. Let $H^{(0)} = H$, and let $H^{(m)}$ denote the result of removing the edges of the clique $\mathcal{C}_t$ from $H^{(m-1)}$.

We argue that for all m, the graph $H^{(m)}$ is a union of edge-disjoint maximal cliques, and any two maximal cliques intersect in at most a single vertex. The base case $m = 1$ is true by Condition 2. This holds for $m > 1$ because the above property is preserved when all of the edges are removed from a single maximal clique in $H^{(m-1)}$. Moreover, for all m, the maximal cliques in $H^{(m)}$ are the same as those of $H^{(m-1)}$ but with a single maximal clique $\mathcal{C}_m$ in $H^{(m-1)}$ removed. Hence, it also follows by induction that for all m, the maximal clique $\mathcal{C}_m$ in $H^{(m-1)}$ is also a maximal clique in H.

Since Assign-Parents terminates at the first time M when $H^{(M)}$ has no hyperedges, we conclude that $\mathcal{C}_1, \ldots, \mathcal{C}_M$ are *all* of the maximal cliques in H, as desired. □

Proof of Lemma 5.19. We first extend the definition of ϕ to level k. For $\hat{u} \in V_k(\hat{\mathcal{P}}_k)$,

we define $\phi(\hat{u}) \in V_k(\mathcal{P}_k)$ as follows. Let $\hat{\mathcal{C}} \subset V_{k-1}(\hat{\mathcal{P}}_k)$ denote the children of $\hat{u}$. By Lemmas 5.20 and 5.21, $\phi(\hat{\mathcal{C}})$ is a clique in G_{k-1}. Define $\phi(\hat{u}) \in V_k(\mathcal{P}_k)$ to be the parent of the children of the clique $\phi(\hat{\mathcal{C}})$ in $\mathcal{P}$. The map ϕ is well-defined at level k because of Condition 1. It remains to show that ϕ is an isomorphism onto its image, and moreover that its image contains all of the awesome nodes at level k.

The map ϕ is injective: We know this is true for $\phi\big|_{\hat{\mathcal{P}}_{k-1}}$, so it suffices to consider injectivity of ϕ when restricted to the nodes at level k in $\hat{\mathcal{P}}_k$. Let $\hat{u}, \hat{v} \in V_k(\hat{\mathcal{P}}_k)$ with $\hat{u} \neq \hat{v}$. Let $\mathcal{C}$ (resp., $\mathcal{C}'$) denote the maximal clique in $\hat{G}_{k-1}$ that consists of the children of $\hat{u}$ (resp., $\hat{v}$). By Lemma 5.21, $\phi(\mathcal{C})$ and $\phi(\mathcal{C}')$ are distinct maximal cliques in the induced subgraph $\phi(\hat{G}_{k-1})$, and therefore, are contained in distinct maximal cliques in G_{k-1}. Distinct maximal cliques in G_{k-1} have distinct parents, so by the definition of ϕ, we conclude that $\phi(\hat{u}) \neq \phi(\hat{v})$, as desired.

The map ϕ is edge-preserving: Suppose that $(\hat{u}, \hat{v})$ is an edge in $\hat{\mathcal{P}}_k$ with $\hat{u} \in V_k(\hat{\mathcal{P}}_k)$ and $\hat{v} \in V_{k-1}(\hat{\mathcal{P}}_k)$. Consider the maximal clique $\hat{\mathcal{C}}$ containing $\hat{v}$ in $\hat{G}_{k-1}$. By Lemma 5.21, $\phi(\hat{\mathcal{C}})$ is a maximal clique in the induced subgraph $\phi(\hat{G}_{k-1}) \subset G_{k-1}$, and by construction of ϕ, the parent of $\phi(\hat{\mathcal{C}})$ is $\phi(\hat{u})$. Therefore, the edge $(\phi(\hat{u}), \phi(\hat{v}))$ is in the pedigree $\mathcal{P}_k$.

Suppose now that the edge $(u, v) = (\phi(\hat{u}), \phi(\hat{v}))$ is in the pedigree $\mathcal{P}_k$. Consider the maximal clique $\mathcal{C}' \subset G_{k-1}$ containing v. By Lemma 5.21, $\mathcal{C} := \phi^{-1}(\mathcal{C}') = \{x \in \hat{\mathcal{P}}_k : \phi(x) \in \mathcal{C}'\}$ is a maximal clique in $\hat{G}_{k-1}$. By Lemma 5.21 and the construction in Assign-Parents, we conclude that the parent of $\hat{v}$ in $\hat{\mathcal{P}}_k$ is mapped to u under ϕ. By injectivity of ϕ, this parent is precisely $\phi^{-1}(u) = \hat{u}$. Therefore, $(\hat{u}, \hat{v})$ is an edge in $\hat{\mathcal{P}}_k$.

The image of ϕ contains the awesome nodes in $\mathcal{P}_k$: It suffices to prove the statement for the awesome nodes at level k, which we denote by A_k. Suppose that u is an awesome node at level k of $\mathcal{P}$. By awesomeness, u has at least d awesome children. Let $\mathcal{C}'$ denote the clique in G_{k-1} given by the awesome children of u. By Lemmas 5.20 and 5.21, $\mathcal{C} := \phi^{-1}(\mathcal{C}')$ satisfies $|\mathcal{C}| = |\mathcal{C}'| \geq d$ because all of the awesome children up to level $k-1$ are in the image of ϕ, by the inductive hypotheses. By Lemma 5.21, the maximal clique $\tilde{\mathcal{C}}$ containing $\mathcal{C}$ in $\hat{G}_{k-1}$ satisfies that $\phi(\tilde{\mathcal{C}})$ are all children of u. By the definition of Assign-Parents and ϕ at level k, we conclude that a parent $\hat{u}$ is constructed for $\tilde{\mathcal{C}} \supset \mathcal{C}$ and $\phi(\hat{u}) = u$, as desired. □

Bibliography

23andMe. About us. https://mediacenter.23andme.com/company/about-us/.

Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1278–1289. SIAM, 2017.

D. Achiloptas and C. Moore. The asymptotic order of the random k-sat threshold. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings*, Vancouver, BC, Canada, November 2002. IEEE.

Pankaj K Agarwal, Sariel Har-Peled, and Kasturi R Varadarajan. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52:1–30, 2005.

N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley and Sons, Inc., New Jersey, 3 edition, 2008.

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 1961–1971, 2017.

Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Weed. Massively scalable Sinkhorn distances via the Nyström method. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 12 2019. To appear.

Ancestry.com. Ancestry continues to lead the industry with world's largest consumer dna network. https://www.ancestry.com/corporate/newsroom/press-releases/ancestry%C2%AE-surpasses-15-million-members-its-dna-network-powering-unparalleled.

Hassan Ashtiani, Shai Ben-David, Nicholas JA Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, 67(6):1–42, 2020.

Benjamin Aubin, Will Perkins, and Lenka Zdeborova. Storage capacity in symmetric binary perceptrons. *Journal of Physics A: Mathematical and Theoretical*, 2019.

Francis R Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *ICML*, 2012.

Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.

Olivier Bachem, Mario Lucic, and Andreas Krause. Scalable k-means clustering via lightweight coresets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1119–1127, 2018.

Arturs Backurs, Moses Charikar, Piotr Indyk, and Paris Siminelakis. Efficient density evaluation for smooth kernels. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 615–626. IEEE, 2018.

Arturs Backurs, Piotr Indyk, and Tal Wagner. Space and time efficient kernel density estimation in high dimensions. In *Advances in Neural Information Processing Systems*, pages 15799–15808, 2019.

Mihai Bādoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via coresets. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 250–257, 2002.

Afonso S. Bandeira, Amelia Perry, and Alexander S. Wein. Notes on computational-to-statistical gaps: predictions using statistical physics. *arXiv:1803.11132*, 2018.

Nikhil Bansal. Constructive algorithms for discrepancy minimization. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 3–10, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4244-7. doi: 10.1109/FOCS.2010.7. URL `http://dx.doi.org/10.1109/FOCS.2010.7`.

Nikhil Bansal and Raghu Meka. On the discrepancy of random low degree set systems. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 2557–2564, 2019. doi: 10.1137/1.9781611975482.157. URL `https://doi.org/10.1137/1.9781611975482.157`.

Nikhil Bansal, Daniel Dadush, Shashwat Garg, and Shachar Lovett. The gram-schmidt walk: a cure for the banaszczyk blues. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pages 587–597, 2018. doi: 10.1145/3188745.3188850. URL `https://doi.org/10.1145/3188745.3188850`.

József Beck and Tibor Fiala. "integer-making" theorems. *Discrete Applied Mathematics*, 3(1):1 – 8, 1981. ISSN 0166-218X. doi: https://doi.org/10.1016/0166-218X(81)90022-6. URL `http://www.sciencedirect.com/science/article/pii/0166218X81900226`.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *Ann. Statist.*, 41(1):1780–1815, 2013a.

Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *COLT 2013 - The 26th Conference on Learning Theory, Princeton, NJ, June 12-14, 2013*, volume 30 of *JMLR W&CP*, pages 1046–1066, 2013b.

Quentin Berthet, Philippe Rigollet, and Piyush Srivastava. Exact recovery in the Ising blockmodel. *Annals of Statistics (to appear).*, 2018.

Dimitris Bertsimas, Mac Johnson, and Nathan Kallus. The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, 63(4):868–876, 2015.

Stefan Boettcher and Stephan Mertens. Analysis of the karmarkar-karp differencing algorithm. *CoRR*, abs/0802.4040, 2008. URL `http://arxiv.org/abs/0802.4040`.

C. Borgs, J. Chayes, and B. Pittel. Phase Transition and Finite-Size Scaling for the Integer Partitioning Problem. *Random Structures and Algorithms*, 19:247–288, 2001.

C. Borgs, J. Chayes, S. Mertens, and C. Nair. Proof of the local REM conjecture for number partitioning I: Constant energy scales. *Random Structures and Algorithms*, 34:217–240, December 2008a.

C. Borgs, J. Chayes, S. Mertens, and C. Nair. Proof of the local REM conjecture for number partitioning II: Growing energy scales. *Random Structures and Algorithms*, 34:241–284, December 2008b.

Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR, 2020.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 48–166. PMLR, 06–09 Jul 2018.

Sébastien Bubeck. *Convex optimization: algorithms and complexity*. Now Publishers Inc., 2015.

Trevor Campbell and Tamara Broderick. Automated scalable bayesian inference via hilbert coresets. *The Journal of Machine Learning Research*, 20(1):551–588, 2019.

C. Carathéodory. Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen, March 1907. URL https://doi.org/10.1007/bf01449883.

Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 604–613, 2014.

Karthekeyan Chandrasekaran and Santosh S. Vempala. Integer feasibility of random polytopes: Random integer programs. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, ITCS '14, pages 449–458, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2698-8. doi: 10.1145/2554797.2554838. URL http://doi.acm.org/10.1145/2554797.2554838.

Moses Charikar and Paris Siminelakis. Hashing-based-estimators for kernel density in high dimensions. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1032–1043. IEEE, 2017.

Chazelle and Matoušek. On linear-time deterministic algorithms for optimization problems in fixed dimension. *Journal of Algorithms*, 21(3):579–597, 1996.

B. Chazelle. *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press, Cambridge, 2000.

K. C. Chung and T. H. Yao. On lattices admitting unique lagrange interpolations. *SIAM Journal on Numerical Analysis*, 14(4):735–743, 1977. ISSN 00361429. URL http://www.jstor.org/stable/2156491.

Sebastian Claici, Aude Genevay, and Justin Solomon. Wasserstein measure coresets, 2020.

Kenneth L Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):1–30, 2010.

Michael B Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1758–1777. SIAM, 2017.

Benjamin Coleman and Anshumali Shrivastava. Sub-linear race sketches for approximate kernel density estimation on streaming data. In *Proceedings of The Web Conference 2020*, pages 1739–1749, 2020.

K. Costello. Balancing Gaussian Vectors. *Israeli Journal of Math*, 172:145–156, 2009.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006.

Marco Cuturi and Gabriel Peyré. Computational optimal transport. *ArXiv:1803.00567*, 2018.

Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Optimal phylogenetic reconstruction. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 159–168, 2006.

Abhimanyu Dubey, Moitreya Chatterjee, and Narendra Ahuja. Coreset-based neural network compression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 454–470, 2018.

Ronen Eldan and Mohit Singh. Efficient algorithms for discrepancy minimization in convex sets. *Random Structures & Algorithms*, 53(2):289–307, 2018.

Péter L Erdős, Michael A Steel, László A Székely, and Tandy J Warnow. A few logs suffice to build (almost) all trees (i). *Random Structures & Algorithms*, 14(2): 153–184, 1999.

Yaniv Erlich, Tal Shor, Itsik Pe'er, and Shai Carmi. Identity inference of genomic data using long-range familial searches. *Science*, 362(6415):690–694, 2018. ISSN 0036-8075. doi: 10.1126/science.aau4832.

William Evans, Claire Kenyon, Yuval Peres, and Leonard J. Schulman. Broadcasting on trees and the ising model. *Ann. Appl. Probab.*, 10(2):410–433, 05 2000. doi: 10.1214/aoap/1019487349. URL `https://doi.org/10.1214/aoap/1019487349`.

Esther Ezra and Shachar Lovett. On the Beck-Fiala Conjecture for Random Set Systems. In K. Jansen, C. Mathieu, J. Rolim, and C. Umans, editors, *APPROX/RANDOM*, volume 60 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 29:1–29:10, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi: 10.4230/LIPIcs.APPROX-RANDOM.2016.29.

Dan Feldman, Matthew Faulkner, and Andreas Krause. Scalable training of mixture models via coresets. In *Advances in neural information processing systems*, pages 2142–2150, 2011.

Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1434–1453. SIAM, 2013.

R.A. Fisher. *The design of experiments.* Oliver and Boyd, Edinburgh, 1935.

Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

Gereon Frahling and Christian Sohler. Coresets in dynamic geometric data streams. In *Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '05, pages 209–217, New York, NY, USA, 2005. Association for

Computing Machinery. ISBN 1581139608. doi: 10.1145/1060590.1060622. URL https://doi.org/10.1145/1060590.1060622.

Marguerite Frank, Philip Wolfe, et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

C. Franks and M. Saks. On the discrepancy of random matrices with many columns. *arXiv*, pages 1–24, July 2018.

David Gamarnik and Eren C Kızıldağ. Algorithmic obstructions in the random number partitioning problem. *arXiv preprint arXiv:2103.01369*, 2021.

Bernd Gärtner and Martin Jaggi. Coresets for polytope distance. In *Proceedings of the twenty-fifth annual symposium on Computational geometry*, pages 33–42, 2009.

Leslie Greengard and Vladimir Rokhlin. A fast algorithm for particle simulations. *Journal of computational physics*, 73(2):325–348, 1987.

Leslie Greengard and John Strain. The fast gauss transform. *SIAM Journal on Scientific and Statistical Computing*, 12(1):79–94, 1991.

Robert Greevy, Bo Lu, Jeffrey H. Silber, and Paul Rosenbaum. Optimal multivariate matching before randomization. *Biostatistics*, 5(2):263–275, 2004.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Steve Hanneke, Aryeh Kontorovich, and Menachem Sadigurschi. Sample compression for real-valued learners. In *Algorithmic Learning Theory*, pages 466–488, 2019.

Yi Hao, Ayush Jain, Alon Orlitsky, and Vaishakh Ravindrakumar. Surf: A simple, universal, robust, fast distribution learning algorithm. *arXiv preprint arXiv:2002.09589*, 2020.

Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. *Discrete & Computational Geometry*, 37(1):3–19, 2007.

Christopher Harshaw, Fredrik Sävje, Daniel Spielman, and Peng Zhang. Balancing covariates in randomized experiments using the gram–schmidt walk. *CoRR*, abs/1911.03071, 2019. URL https://arxiv.org/abs/1911.03071.

Nicholas JA Harvey, Roy Schwartz, and Mohit Singh. Discrepancy without partial colorings. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2014.

Nick Harvey and Samira Samadi. Near-optimal herding. In *Conference on Learning Theory*, pages 1165–1182, 2014.

Dan He, Zhanyong Wang, Buhm Han, Laxmi Parida, and Eleazar Eskin. Iped: inheritance path-based pedigree reconstruction algorithm using genotype data. *Journal of Computational Biology*, 20(10):780–791, 2013.

Dan He, Zhanyong Wang, Laxmi Parida, and Eleazar Eskin. Iped2: Inheritance path based pedigree reconstruction algorithm for complicated pedigrees. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, page 202–210, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450328944. doi: 10.1145/2649387.2649438. URL https://doi.org/10.1145/2649387.2649438.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

R. Hoberg and T. Rothvoss. A Fourier-analytic approach for the discrepancy of random set systems. *arXiv*, pages 1–19, July 2018.

Rebecca Hoberg, Harishchandra Ramadas, Thomas Rothvoss, and Xin Yang. Number balancing is as hard as minkowski's theorem and shortest vector. In *Integer Programming and Combinatorial Optimization - 19th International Conference, IPCO 2017, Waterloo, ON, Canada, June 26-28, 2017, Proceedings*, pages 254–266, 2017. doi: 10.1007/978-3-319-59250-3_21. URL https://doi.org/10.1007/978-3-319-59250-3_21.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Jonathan Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. In *Advances in Neural Information Processing Systems*, pages 4080–4088, 2016.

Jisca Huisman. Pedigree reconstruction from snp data: parentage assignment, sibship clustering and beyond. *Molecular ecology resources*, 17(5):1009–1024, 2017.

National Human Genome Research Institute. The cost of sequencing a human genome. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost, 2019.

M Chris Jones. Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, 84(407):733–741, 1989.

Sarang Joshi, Raj Varma Kommaraji, Jeff M. Phillips, and Suresh Venkatasubramanian. Comparing distributions and shapes using the kernel distance. In *Proceedings of the Twenty-Seventh Annual Symposium on Computational Geometry*, SoCG '11, page 47–56, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306829. doi: 10.1145/1998196.1998204. URL https://doi.org/10.1145/1998196.1998204.

Nathan Kallus. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1): 85–112, 2018.

N. Karmarkar, R. Karp, G. Lueker, and A. Odlyzko. Probabilistic Analysis of Optimum Partitioning. *Journal of Applied Probability*, 23:626–645, September 1986.

Narendra Karmarkar and Richard Karp. The differencing method of set partitioning. Technical report, University of California, Berkeley, 12 1982.

Zohar Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1975–1993, Phoenix, USA, 25–28 Jun 2019. PMLR. URL http://proceedings.mlr.press/v99/karnin19a.html.

Yitzhak Katznelson. *An Introduction to Harmonic Analysis*. Cambridge Mathematical Library. Cambridge University Press, 3 edition, 2004. doi: 10.1017/CBO9781139165372.

Younhun Kim, Frederic Koehler, Ankur Moitra, Elchanan Mossel, and Govind Ramnarayan. How many subpopulations is too many? exponential lower bounds for inferring population histories. *Journal of Computational Biology*, 2019.

Younhun Kim, Elchanan Mossel, Govind Ramnarayan, and Paxton Turner. Efficient reconstruction of stochastic pedigrees. *arXiv preprint arXiv:2005.03810*, 2020.

Marek Kimmel and David Axelrod. *Branching Processes in Biology*. Springer Publishing Company, Incorporated, 2nd edition, 2015. ISBN 1493915584.

Bonnie Kirkpatrick, Shuai Cheng Li, Richard M Karp, and Eran Halperin. Pedigree reconstruction using identity by descent. *Journal of Computational Biology*, 18 (11):1481–1493, 2011.

Atsuyuki Kogure. Effective interpolations for kernel density estimators. *Journal of Nonparametric Statistics*, 9(2):165–195, 1998.

Gina Kolata and Heather Murphy. The golden state killer is tracked through a thicket of dna, and experts shudder. *The New York Times*, 4 2018.

A. N. Kolmogorov and V. M. Tikhomirov. *ε-Entropy and ε-Capacity of Sets In Functional Spaces*, pages 86–170. Springer Netherlands, Dordrecht, 1993. ISBN 978-94-017-2973-4. doi: 10.1007/978-94-017-2973-4_7. URL https://doi.org/10.1007/978-94-017-2973-4_7.

A M Krieger, D Azriel, and A Kapelner. Nearly random designs with greatly improved balance. *Biometrika*, 106(3):695–701, 05 2019. ISSN 0006-3444. doi: 10.1093/biomet/asz026. URL https://doi.org/10.1093/biomet/asz026.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. *arXiv preprint arXiv:1907.11636*, 2019.

Steven Lalley. Poisson processes. *Statistics 312: Stochastic Processes*, 2016.

Michel Ledoux and Jie-Xiang Zhu. On optimal matching of Gaussian samples III. Available on the first author's webpage, 2019.

Dongryeol Lee, Andrew W Moore, and Alexander G Gray. Dual-tree fast gauss transforms. In *Advances in Neural Information Processing Systems*, pages 747–754, 2006.

Avi Levy, Harishchandra Ramadas, and Thomas Rothvoss. Deterministic discrepancy minimization via the multiplicative weight update method. In *Integer Programming and Combinatorial Optimization - 19th International Conference, IPCO 2017, Waterloo, ON, Canada, June 26-28, 2017, Proceedings*, pages 380–391, 2017. doi: 10.1007/978-3-319-59250-3_31. URL https://doi.org/10.1007/978-3-319-59250-3_31.

Xinran Li, Peng Ding, and Donald B. Rubin. Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162, 2018.

Tengyuan Liang, Alexander Rakhlin, et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. *Unpublished manuscript*, 1986.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1452–1461, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/lopez-paz15.html.

Shachar Lovett and Raghu Meka. Constructive discrepancy minimization by walking on the edges. In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, FOCS '12, pages 61–67, Washington, DC, USA, 2012.

IEEE Computer Society. ISBN 978-0-7695-4874-6. doi: 10.1109/FOCS.2012.23. URL https://doi.org/10.1109/FOCS.2012.23.

Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Fast and accurate least-mean-squares solvers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/475fbefa9ebfba9233364533aafd02a3-Paper.pdf.

Anuran Makur, Elchanan Mossel, and Yury Polyanskiy. Broadcasting on bounded degree dags. *arXiv preprint arXiv:1803.07527*, 2018.

J. Matoušek. *Geometric Discrepancy: an Illustrated Guide*. Springer, New York, 1999.

Daniel McDonald. Minimax Density Estimation for Growing Dimension. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 194–203, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL http://proceedings.mlr.press/v54/mcdonald17a.html.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020.

Kari Lock Morgan and Donald B. Rubin. Rerandomization to improve covariate balance in experiments. *Ann. Statist.*, 40(2):1263–1282, 04 2012.

Elchanan Mossel. Reconstruction on trees: beating the second eigenvalue. *Annals of Applied Probability*, pages 285–300, 2001.

Elchanan Mossel. Phase transitions in phylogeny. *Transactions of the American Mathematical Society*, 356(6):2379–2404, 2004.

Elchanan Mossel and Sébastien Roch. Learning nonsingular phylogenies and hidden markov models. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 366–375, 2005.

Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David P. Woodruff. On coresets for logistic regression. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 6562–6571, Red Hook, NY, USA, 2018. Curran Associates Inc.

J.D. Murray. *Asymptotic Aanlysis*, volume 48. Springer, New York, 1984.

MyHeritage. Myheritage end-of-year infographic. https://blog.myheritage.com/2019/12/wrapping-up-a-fantastic-2019.

R. A. Nicolaides. On a class of finite elements generated by lagrange interpolation. *SIAM Journal on Numerical Analysis*, 9(3):435–445, 1972. ISSN 00361429. URL http://www.jstor.org/stable/2156141.

Jeff M Phillips. ε-samples for kernels. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1622–1632. SIAM, 2013.

Jeff M Phillips and Wai Ming Tai. Improved coresets for kernel density estimates. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2718–2727. SIAM, 2018a.

Jeff M. Phillips and Wai Ming Tai. Near-optimal coresets of kernel density estimates. In *34th International Symposium on Computational Geometry, SoCG 2018, June 11-14, 2018, Budapest, Hungary*, pages 66:1–66:13, 2018b. doi: 10.4230/LIPIcs.SoCG.2018.66. URL https://doi.org/10.4230/LIPIcs.SoCG.2018.66.

D Pollard. Mini empirical. Manuscript. http://www.stat.yale.edu/~pollard/Books/Mini/, 2015.

Aditya Potukuchi. Discrepancy in random hypergraph models. *CoRR*, abs/1811.01491, 2018. URL http://arxiv.org/abs/1811.01491.

Thomas Rothvoss. Constructive discrepancy minimization for convex sets. *SIAM J. Comput.*, 46(1):224–234, 2017. doi: 10.1137/141000282. URL https://doi.org/10.1137/141000282.

Erich Schubert, Arthur Zimek, and Hans-Peter Kriegel. Generalized outlier detection with flexible kernel density estimates. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 542–550. SIAM, 2014.

David W Scott and Simon J Sheather. Kernel density estimation with binned data. *Communications in Statistics-Theory and Methods*, 14(6):1353–1359, 1985.

Charles Semple and Mike Steel. *Phylogenetics*, volume 24. Oxford University Press on Demand, 2003.

Doron Shem-Tov and Eran Halperin. Historical pedigree reconstruction from extant populations using partitioning of relatives (prepare). *PLoS computational biology*, 10(6), 2014.

Paris Siminelakis, Kexin Rong, Peter Bailis, Moses Charikar, and Philip Levis. Rehashing kernel evaluation in high dimensions. In *International Conference on Machine Learning*, pages 5789–5798, 2019.

J. Spencer. Six Standard Deviations Suffice. *Transactions of the American Mathematical Society*, 289:679–706, 1985.

Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

Mike Steel and Jotun Hein. Reconstructing pedigrees: a combinatorial perspective. *Journal of theoretical biology*, 240(3):360–367, 2006.

J Michael Steele. Euclidean semi-matchings of random samples. *Mathematical Programming*, 53(1-3):127–146, 1992.

Student. Comparison between balanced and random arrangements of field plots. *Biometrika*, 29(3-4):363–379, 1938.

Bhalchandra D Thatte and Mike Steel. Reconstructing pedigrees: a stochastic perspective. *Journal of theoretical biology*, 251(3):440–449, 2008.

Elizabeth A. Thompson. Statistical inference from genetic data on pedigrees. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 6:i–169, 2000. ISSN 19355920, 23290978. URL `http://www.jstor.org/stable/4153187`.

Elizabeth A Thompson. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326, 2013.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794. URL `https://doi.org/10.1007/b13794`.

Paxton Turner, Raghu Meka, and Philippe Rigollet. Balancing gaussian vectors in high dimension. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3455–3486. PMLR, 09–12 Jul 2020. URL `http://proceedings.mlr.press/v125/turner20a.html`.

Paxton Turner, Jingbo Liu, and Philippe Rigollet. Efficient interpolation of density estimators. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2503–2511. PMLR, 13–15 Apr 2021a. URL `http://proceedings.mlr.press/v130/turner21a.html`.

Paxton Turner, Jingbo Liu, and Philippe Rigollet. A statistical perspective on coreset density estimation. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2512–2520. PMLR, 13–15 Apr 2021b. URL `http://proceedings.mlr.press/v130/turner21b.html`.

VN Vapnik and AJ Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Prob Appl 16 (2): 264–280*, 16:264–280, 1971.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Jinliang Wang. Pedigree reconstruction from poor quality genotype data. *Heredity*, 122(6):719–728, 2019.

Benjamin Yakir. The differencing algorithm ldm for partitioning: A proof of a conjecture of karmarkar and karp. *Math. Oper. Res.*, 21(1):85–99, February 1996. ISSN 0364-765X. doi: 10.1287/moor.21.1.85. URL `http://dx.doi.org/10.1287/moor.21.1.85`.

Changjiang Yang, Ramani Duraiswami, Nail A Gumerov, and Larry Davis. Improved fast gauss transform and efficient kernel density estimation. In *Proceedings of the Ninth IEEE International Conference on Computer Vision-Volume 2*, page 464, 2003.

Jiyan Yang, Yin-Lam Chow, Christopher Ré, and Michael W Mahoney. Weighted sgd for lp regression with randomized preconditioning. *The Journal of Machine Learning Research*, 18(1):7811–7853, 2017.

Yan Zheng and Jeff M Phillips. Coresets for kernel regression. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 645–654, 2017.