Optimal SQ Lower Bounds for Robustly Learning Discrete Product Distributions and Ising Models

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin, Madison

Daniel M. Kane DAKANE@CS.UCSD.EDU
University of California, San Diego

Yuxin Sun YXSUN@CS.WISC.EDU
University of Wisconsin, Madison

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We establish optimal Statistical Query (SQ) lower bounds for robustly learning certain families of discrete high-dimensional distributions. In particular, we show that no efficient SQ algorithm with access to an ϵ-corrupted binary product distribution can learn its mean within ℓ_2-error $o(\epsilon\sqrt{\log(1/\epsilon)})$. Similarly, we show that no efficient SQ algorithm with access to an ϵ-corrupted ferromagnetic high-temperature Ising model can learn the model to total variation distance $o(\epsilon\log(1/\epsilon))$. Our SQ lower bounds match the error guarantees of known algorithms for these problems, providing evidence that current upper bounds for these tasks are best possible. At the technical level, we develop a generic SQ lower bound for discrete high-dimensional distributions starting from low-dimensional moment matching constructions that we believe will find other applications. Additionally, we introduce new ideas to analyze these moment-matching constructions for discrete univariate distributions.

Keywords: robust statistics, statistical query model, discrete distributions, Ising model

1. Introduction

1.1. Background and Motivation

Robust Statistics and Information-Computation Tradeoffs We study high-dimensional learning in the presence of a constant fraction of arbitrary outliers. Robust learning in high dimensions has its roots in robust statistics, a branch of statistics initiated in the 60s with the pioneering works of Tukey and Huber (Tukey, 1960; Huber, 1964). Early work developed minimax optimal estimators for various robust estimation tasks, albeit with runtimes exponential in the dimension. A recent line of work in computer science, starting with Diakonikolas et al. (2016); Lai et al. (2016), developed polynomial time robust estimators for a range of high-dimensional statistical tasks. Algorithmic high-dimensional robust statistics is by now a relatively mature field, see, e.g., Diakonikolas and Kane (2019) for a survey.

The line of work in algorithmic robust statistics established the existence of computationally efficient algorithms with dimension-independent error guarantees for a range of high-dimensional robust estimation tasks. In some instances, these algorithms achieve the information-theoretically optimal error (within constant factors). Alas, in several interesting settings, there is a super-constant gap between the information-theoretic optimum and what known efficient algorithms achieve. This raises the following natural question: *For a given high-dimensional robust estimation task, is the information-theoretically optimal error achievable in polynomial time?*

In several high-dimensional statistical settings, there is strong evidence that inherent resource tradeoffs exist. In robust statistics, the study of such *information-computation tradeoffs* was initiated in Diakonikolas et al. (2017), which established the first such lower bounds in the Statistical Query (SQ) model (Kearns, 1998). The methodology for proving such lower bounds introduced in Diakonikolas et al. (2017) applies to "Gaussian-like" distributions. In particular, the general problem underlying that work — known as Non-Gaussian Component Analysis (NGCA) (Blanchard et al., 2006; Tan and Vershynin, 2018; Goyal and Shetty, 2019) — considers distributions that are distributed as a standard Gaussian in all but one hidden direction.

Here we are interested in exploring information-computation tradeoffs for robustly learning *discrete* high-dimensional distributions. The two concrete examples — that were the main motivation for this work — are (1) the class of binary product distributions and (2) the (more general) class of Ising models. For both of these distribution classes, there are gaps between the information-theoretically optimal error and the error that known polynomial-time algorithms can achieve. Given the aforementioned prior work for Gaussian-like distributions (Diakonikolas et al., 2017), it would be tempting to conjecture that these gaps are in fact inherent. In this work, we develop the necessary methodology that allows us to prove such statements for discrete distributions, and in particular for the aforementioned families.

Before we proceed, we give the necessary background on the SQ model and robust statistics.

Statistical Query (SQ) Model SQ algorithms are the class of algorithms that are only allowed to query expectations of bounded functions of the underlying distribution rather than directly access samples. The SQ model was introduced by Kearns (1998) in the context of supervised learning as a natural restriction of the PAC model (Valiant, 1984) and has been extensively studied in learning theory. A recent line of work (Feldman et al., 2013, 2015, 2017; Feldman, 2017) generalized the SQ framework for search problems over distributions.

The class of SQ algorithms is fairly broad: a wide range of known algorithmic techniques in machine learning are known to be implementable in the SQ model. These include spectral techniques, moment and tensor methods, local search, and many others (see, e.g., Chu et al. (2006); Feldman et al. (2013, 2017)). A notable exception are learning algorithms using Gaussian elimination (in particular for learning parities, see, e.g., Blum et al. (2003)), Brennan et al. (2020) recently established a connection between the SQ model and low-degree polynomial tests under certain assumptions.

Contamination Model We focus on the following contamination model, where the adversary can corrupt the true distribution in total variation distance.

Definition 1 (TV-contamination) *Given a parameter $0 < \epsilon < 1/2$ and a distribution class $\mathcal{D}$, we say that a distribution D' is an ϵ-*corrupted *version of a distribution $D \in \mathcal{D}$ if $d_{\mathrm{TV}}(D, D') \leq \epsilon$.*

We will study algorithms robust against this kind of contamination. In particular, we want algorithms that given sample access to a distribution D' which is an ϵ-corrupted version of some unknown distribution $D \in \mathcal{D}$, can approximate relevant parameters of the "true" distribution D. For such algorithms, one may want to consider the distribution D' to be adversarially selected, perhaps in a way designed to fool the particular algorithm in question. We also note that several algorithms in robust statistics can be made to function in the presence of even stronger contamination models, such as the strong contamination model, where the adversary can inspect the samples drawn and adaptively choose which samples to corrupt and how. However, these stronger models are harder to formalize for SQ algorithms where our lower bounds will apply.

1.2. Problems of Interest and Our Results

With this background, we are ready to summarize prior algorithmic work on the two problems of interest and informally state our contributions.

Robust Mean Estimation for a Binary Product Distribution A binary product distribution is a distribution over $\{0,1\}^M$ whose coordinates are independent. We consider the algorithmic problem of computing an approximation to the mean vector μ_P of a binary product distribution P, in ℓ_2-norm, given access to a set of samples from an ϵ-corrupted version of P. Diakonikolas et al. (2016) gave the first efficient algorithm for this problem that outputs an estimate $\widehat{\mu}$ such that with high probability $\|\widehat{\mu} - \mu_P\| = O(\epsilon\sqrt{\log(1/\epsilon)})$. Information-theoretically, it is possible to approximate μ_P within ℓ_2 error $\Theta(\epsilon)$. Our first main result shows that this gap is inherent for SQ algorithms (see Theorem 24 for a detailed statement).

Theorem 2 (SQ Lower Bound for Binary Products, Informal) *Any SQ algorithm that robustly learns the mean of a binary product distribution over $\{0,1\}^M$, given access to an ϵ-corruption, within ℓ_2-error $o(\epsilon\sqrt{\log(1/\epsilon)})$ either requires at least $2^{M^{\Omega(1)}}$ many statistical queries or must make a query of accuracy inverse super-polynomial in M.*

Theorem 2 shows that no SQ algorithm can robustly approximate the mean of a binary product distribution to error $o(\epsilon\sqrt{\log(1/\epsilon)})$ with a sub-exponential in $M^{\Omega(1)}$ queries, unless using queries of very small tolerance – that would require super-polynomially many samples in M to simulate. In that sense, Theorem 2 is an information-computation tradeoff for robust mean estimation of a binary product distribution within the class of SQ algorithms.

Robustly Learning a Ferromagnetic High-Temperature Ising Model Given a symmetric matrix $(\theta_{ij})_{i,j\in[M]} \in \mathbb{R}_+^{M\times M}$ with zero diagonal, a ferromagnetic Ising model is a distribution over $\{\pm 1\}^M$ with mass function $P_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \exp\left((1/2)\sum_{i,j\in[M]} \theta_{ij}x_ix_j\right)$, where $Z(\theta)$ is a normalizing constant. We say that an Ising model lies in the high-temperature regime if there is a universal constant $0 < \eta < 1$ such that $\max_{i\in[M]} \sum_{j\neq i} |\theta_{ij}| \leq 1 - \eta$. Here we would like an algorithm that given samples from an ϵ-corrupted version of an unknown ferromagnetic high-temperature Ising model P, approximates P in total variational distance. Diakonikolas et al. (2021e) gave the first efficient algorithm for this problem that outputs an estimate $\widehat{P}$ such that with high probability $d_{\mathrm{TV}}(\widehat{P}, P) = O(\epsilon\log(1/\epsilon))$. On the other hand, the information-theoretically optimal error in total variation distance is $\Theta(\epsilon)$. Our second main result shows that this gap is inherent for SQ algorithms (see Theorem 29 for a detailed statement).

Theorem 3 (SQ Lower Bound for Ising Models, Informal) *Any SQ algorithm that robustly learns a ferromagnetic high-temperature Ising Model over $\{\pm 1\}^M$, given access to an ϵ-corruption, within total variation distance $o(\epsilon\log(1/\epsilon))$ either requires at least $2^{M^{\Omega(1)}}$ many statistical queries or must make a query of accuracy inverse super-polynomial in M.*

Similarly, Theorem 3 is an information-computation tradeoff for robust learning of an Ising model within the class of SQ algorithms. In summary, for both of these problems, we show that known algorithms are essentially optimal within the class of Statistical Query (SQ) algorithms.

1.3. Technical Overview

Here we provide an outline of our approach and techniques. To prove our SQ lower bound in the discrete setting, we need to develop a novel generic discrete SQ lower bound machinery for distributions over $\{0,1\}^M$. At a high level, our construction resembles the lower bound construction of Dachman-Soled et al. (2015) (in the context of supervised learning), adapting several ideas of Diakonikolas et al. (2017) from a Gaussian version of this problem.

In particular, showing an SQ lower bound for learning functions in some class essentially boils down to proving lower bounds for the corresponding SQ dimension. In our case, this amounts to finding large families of ϵ-corrupted binary product distributions or ϵ-corrupted Ising models that have pairwise small chi-squared inner product with respect to some given base distribution. We will select as a base distribution the uniform distribution over the hypercube.

To construct these distributions, we will adapt and generalize the techniques of Dachman-Soled et al. (2015). In particular, we aim to find a single distribution D_0 of the appropriate type over $\{0,1\}^m$, for some m substantially smaller than M, so that D_0's low-degree Fourier coefficients vanish. One can then use D_0 to obtain many different distributions over $\{0,1\}^M$ by embedding it over some subset (chosen in one of many different ways) of the coordinates, and using the uniform distribution over the remaining coordinates. One can show (see Lemma 19) that this allows one to produce many nearly orthogonal distributions.

This leaves us with the task of producing an appropriate distribution D_0. To achieve this, we take inspiration from the Gaussian regime (Diakonikolas et al., 2017). In particular, we simplify matters by considering only *symmetric* distributions D_0. This means that D_0 is determined by a *one-dimensional* distribution A — specifically, the distribution over the sum of the coordinates of D_0. This distribution A must be close in total variation distance to an appropriate one-dimensional version of either a binary product distribution or an Ising model, and must match several of its low-degree moments with the Binomial distribution.

In order to construct these one-dimensional distributions, we again borrow ideas from Diakonikolas et al. (2017). We want to obtain a distribution A close to some other distribution B that matches its low-degree moments with the binomial. We will achieve this by starting with the distribution B and modifying its probability mass function (pmf) over some appropriately chosen interval I. In particular, if we modify it by a degree-k polynomial p over I, there will be a unique choice of this polynomial that gives us some specified first k moments. To establish correctness, we need to verify that the resulting polynomial p is not too large (both to ensure that the resulting pmf is non-negative and to ensure that A and B are close in total variation distance). This can be shown via an explicit analysis involving Legendre polynomials (as is done by Diakonikolas et al. (2017), in the continuous case) along with additional technical work required to show that the change to the discrete setting does not significantly affect things.

2. Preliminaries

Notation For $n \in \mathbb{Z}_+$, we denote $[n] \stackrel{\mathrm{def}}{=} \{1, \ldots, n\}$. For two distributions p, q over a probability space Ω, let $d_{\mathrm{TV}}(p, q) = \sup_{S \subseteq \Omega} |p(S) - q(S)|$ denote the total variation distance between p and q. We use $\mathbf{Pr}[\mathcal{E}]$ and $\mathbb{I}[\mathcal{E}]$ for the probability and the indicator of event $\mathcal{E}$. For a real random variable X, we use $\mathbf{E}[X]$, $\mathbf{Var}[X]$ to denote the expectation and variance of X, respectively. For $n \in \mathbb{Z}_+$ and $0 \leq p \leq 1$, we use $\mathrm{Bin}(n, p)$ to denote the Binomial distribution with parameters n and p.

Properties of Legendre Polynomials We record some properties of Legendre polynomials that we will need throughout this paper.

Fact 4 (Szegö (1989)) *The Legendre polynomials, $P_i(x)$, for $i \in \mathbb{Z}_+$, satisfy the following properties: (i) $P_i(x)$ is a degree i-polynomial with $P_0(x) = 1$ and $P_1(x) = x$. (ii) $\int_{-1}^{1} P_i(x)P_j(x)dx = \frac{2\delta_{ij}}{2i+1}$ for all $i, j \geq 0$. (iii) $|P_i(x)| \leq 1$ for all $|x| \leq 1$. (iv) $P_i(x) = (-1)^i P_i(-x)$. (v) $P_i(x) = 2^{-i} \sum_{j=0}^{\lfloor i/2 \rfloor} (-1)^j \binom{i}{j}\binom{2i-2j}{i} x^{i-2j}$. (vi) $|P_i(x)| \leq (4|x|)^i$ for all $|x| \geq 1$. (vii) $\int_{-1}^{1} |P_i(x)|dx \leq O(1/\sqrt{i})$. (viii) $|P_i'(x)| \leq O(i^2)$ for all $|x| \leq 1$.*

Ising Models We recall basic facts about Ising models, which will be used throughout this paper.

Definition 5 (Ising Model) *Given a symmetric matrix $(\theta_{ij})_{i,j\in[d]} \in \mathbb{R}^{d\times d}$ with zero diagonal, the Ising model distribution P_θ is defined as: $P_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \exp\left((1/2)\sum_{i,j\in[d]} \theta_{ij}x_i x_j\right), \forall \mathbf{x} \in \{\pm 1\}^d$, where the normalizing factor $Z(\theta)$ is called the partition function. We call the matrix $(\theta_{ij})_{i,j\in[d]} \in \mathbb{R}^{d\times d}$ the interaction matrix. In addition, we say that P_θ is ferromagnetic if $\theta_{ij} \geq 0, \forall i, j \in [d]$.*

The following Dobrushin's condition for Ising models is a classical assumption needed to rule out certain pathological behaviors. This condition is standard in various areas, including statistical physics, computational biology, machine learning, and theoretical CS (Külske, 2003; Götze et al., 2019; Dagan et al., 2020; Adamczak et al., 2019; Gheissari et al., 2018; Marton, 2015).

Definition 6 (Dobrushin's Condition) *Given an Ising model P_θ with interaction matrix $(\theta_{ij})_{i,j\in[d]}$, we say that it satisfies Dobrushin's condition, or lies in the high temperature regime, if there is a constant $0 < \eta < 1$ such that $\max_{i\in[d]} \sum_{j\neq i} |\theta_{ij}| \leq 1 - \eta$.*

Statistical Query Algorithms We will use the framework of Statistical Query (SQ) algorithms for problems over distributions introduced in Feldman et al. (2013). Before we get into the formal statement of our generic discrete SQ lower bound, we formulate it as a decision problem as follows:

Definition 7 (Decision/Testing Problem over Distributions) *Let D be a distribution and $\mathcal{D}$ be a family of distributions over $\mathbb{R}^M$. We denote by $\mathcal{B}(\mathcal{D}, D)$ the decision (or hypothesis testing) problem in which the input distribution D' is promised to satisfy either (a) $D' = D$ or (b) $D' \in \mathcal{D}$, and the goal of the algorithm is to distinguish between these two cases.*

We define SQ algorithms as algorithms that do not have direct access to samples from the distribution, but instead have access to an SQ oracle. We consider the following standard oracle.

Definition 8 (STAT Oracle) *Let D be a distribution on $\mathbb{R}^M$. A* Statistical Query (SQ) *is a bounded function $f : \mathbb{R}^M \to [-1, 1]$. For $\tau > 0$, the $\mathrm{STAT}(\tau)$ oracle responds to the query f with a value v such that $|v - \mathbf{E}_{X\sim D}[f(X)]| \leq \tau$. We call τ the* tolerance *of the statistical query. A* Statistical Query (SQ) algorithm *is an algorithm whose objective is to learn some information about an unknown distribution D by making adaptive calls to the corresponding $\mathrm{STAT}(\tau)$ oracle.*

To define the SQ dimension, we need the following definition.

Definition 9 *The pairwise correlation of two distributions with probability density functions $D_1, D_2 : \{0,1\}^M \to \mathbb{R}_+$ with respect to a distribution with density $D : \{0,1\}^M \to \mathbb{R}_+$, where the support of D contains the supports of D_1 and D_2, is defined as $\chi_D(D_1, D_2) + 1 \stackrel{\text{def}}{=} \sum_{\mathbf{x}\in\{0,1\}^M} D_1(\mathbf{x})D_2(\mathbf{x})/D(\mathbf{x})$. We say that a set of s distributions $\mathcal{D} = \{D_1, \ldots, D_s\}$ over $\{0,1\}^M$ is (γ, β)-correlated relative to a distribution D if $|\chi_D(D_i, D_j)| \leq \gamma$ for all $i \neq j$, and $|\chi_D(D_i, D_j)| \leq \beta$ for $i = j$.*

Definition 10 (SQ Dimension) *For $\gamma, \beta > 0$, a decision problem $\mathcal{B}(\mathcal{D}, D)$, where D is fixed and $\mathcal{D}$ is a family of distributions over $\{0,1\}^M$, let s be the maximum integer such that there exists $\mathcal{D}_D \subseteq \mathcal{D}$ such that $\mathcal{D}_D$ is (γ, β)-correlated relative to D and $|\mathcal{D}_D| \geq s$. We define the* Statistical Query dimension *with pairwise correlations (γ, β) of $\mathcal{B}$ to be s and denote it by $\mathrm{SD}(\mathcal{B}, \gamma, \beta)$.*

The connection between SQ dimension and lower bounds is captured by the following lemma.

Lemma 11 (Feldman et al. (2013)) *Let $\mathcal{B}(\mathcal{D}, D)$ be a decision problem, where D is the reference distribution and $\mathcal{D}$ is a class of distributions over $\mathbb{R}^M$. For $\gamma, \beta > 0$, let $s = \mathrm{SD}(\mathcal{B}, \gamma, \beta)$. Any SQ algorithm that solves $\mathcal{B}$ with probability at least $2/3$ requires at least $s \cdot \gamma/\beta$ queries to the $\mathrm{STAT}(\sqrt{2\gamma})$ oracles.*

We note that the hypothesis testing problem of Definition 7 may in general be information theoretically hard. In particular, if some distribution $D' \in \mathcal{D}$ is very close to the reference distribution D, it will be hard to distinguish between D' and D. On the other hand, if D' is far from the reference distribution D in total variation distance for any $D' \in \mathcal{D}$, then one can straightforwardly reduce the hypothesis testing problem to the problem of learning an unknown $D' \in \mathcal{D}$ to small accuracy. For completeness, we defer the formal statement and proof to Appendix A.6.

3. Generic Discrete SQ Lower Bound Construction

We start with some basic definitions.

Definition 12 (Characters) *For $\mathbf{x} \in \{0,1\}^M$, we denote $\chi_T(\mathbf{x}) = (-1)^{\sum_{i \in T} x_i}$. For a distribution $\mathbf{P}$ over $\{0,1\}^M$, let $\widehat{\mathbf{P}}(T) = \mathbf{E}_{\mathbf{X} \sim \mathbf{P}}[\chi_T(\mathbf{X})]$.*

We will denote by U_M the uniform distribution over $\{0,1\}^M$. By Plancherel's identity, we have the following fact about the chi-squared inner product in the discrete setting.

Fact 13 *For distributions $\mathbf{P}, \mathbf{Q}$ over $\{0,1\}^M$, we have that $1 + \chi_{U_M}(\mathbf{P}, \mathbf{Q}) = \sum_{T \subseteq [M]} \widehat{\mathbf{P}}(T) \widehat{\mathbf{Q}}(T)$.*

We will require the orthogonal polynomials under the Binomial distribution.

Definition 14 (Kravchuk Polynomial Szegö (1989)) *For $k, m, x \in \mathbb{Z}_+$ with $0 \leq k, x \leq m$, the Kravchuk polynomial $\mathcal{K}_k(x; m)$ is the univariate degree-k polynomial in x defined by $\mathcal{K}_k(x; m) := \sum_{T \subseteq [m], |T| = k} \chi_T(\mathbf{y}) = \sum_{j=0}^{k} (-1)^j \binom{x}{j} \binom{m-x}{k-j}$, where $\mathbf{y}$ has x 1's and $m - x$ 0's.*

Fact 15 (Orthogonality Szegö (1989)) *Let $j, k, m \in \mathbb{Z}_+$. Then, $\mathbf{E}_{X \sim \mathrm{Bin}(m,1/2)}[\mathcal{K}_j(X; m) \mathcal{K}_k(X; m)] = \mathbb{I}[j = k] \binom{m}{k}$. In particular, if $k \geq 1$, then $\mathbf{E}_{X \sim \mathrm{Bin}(m,1/2)}[\mathcal{K}_k(X; m)] = 0$.*

Our basic technique for producing near-orthogonal distributions over the hypercube takes inspiration from Dachman-Soled et al. (2015). They show that if one can construct a distribution D over a small number of coordinates whose degree up-to-k Fourier coefficients agree with the uniform distribution, then by taking embeddings of D into the hypercube as a junta can provide many orthogonal distributions. This leaves us with finding our moment-matching distribution D. Our basic idea will be to make D a symmetric distribution, as this will simplify things substantially due to the added symmetry. Essentially, D will be defined by some distribution A on $\sum x_i$. This distribution A will need to nearly match the first k moments with the Binomial distribution $\mathrm{Bin}(m, 1/2)$.

We now formally define the high-dimensional distribution family that is the basis of our discrete SQ lower bound construction.

Definition 16 (High-Dimensional Hidden Junta Distribution) *Let $m, M \in \mathbb{Z}_+$ with $m < M$. For a distribution A on $[m] \cup \{0\}$ with probability mass function (pmf) $A(x)$ and a subset $S \subseteq [M]$ with $|S| = m$, consider the probability distribution over $\{0,1\}^M$, denoted by $\mathbf{P}_S^A$, such that for $\mathbf{X} \sim \mathbf{P}_S^A$ the distribution $(X_i)_{i \notin S}$ is the uniform distribution on its support and the distribution $(X_i)_{i \in S}$ is symmetric with $\sum_{i \in S} X_i$ distributed according to A. Specifically, $\mathbf{P}_S^A$ is given by the pmf $\mathbf{P}_S^A(\mathbf{x}) = 2^{-M+m} A\left(\sum_{i \in S} x_i\right) \binom{m}{\sum_{i \in S} x_i}^{-1}$.*

We now define the hypothesis testing problem which will be used throughout this paper:

Definition 17 (Hidden Junta Testing Problem) *Let $m, M \in \mathbb{Z}_+$ with $M > m$ and A be a one-dimensional distribution over $[m] \cup \{0\}$. In the (A, M)-Hidden Junta Testing Problem, one is given access to a distribution D so that either H_0: $D = U_M$, H_1: D is given by $\mathbf{P}_S^A$ for some subset $S \subseteq [M]$ with $|S| = m$, where $\mathbf{P}_S^A$ denotes the hidden junta distribution corresponding to A. One is then asked to distinguish between H_0 and H_1.*

Note that this is just the hypothesis testing problem $\mathcal{B}(\mathcal{D}, D)$ with $D = U_M$ and $\mathcal{D} = \{\mathbf{P}_S^A\}$. The following condition describes the approximate moment-matching property of the desired distribution A with the Binomial distribution.

Condition 18 (Approximate Moment-Matching) *Let $\nu > 0$ and $k, m \in \mathbb{Z}_+$ with $k \leq m$. The distribution A on $[m] \cup \{0\}$ satisfies $|\mathbf{E}_{X \sim A}[\mathcal{K}_t(X; m)]| \leq \nu$, for all $1 \leq t \leq k$.*

In particular, if A exactly matches the first k moments with $\mathrm{Bin}(m, 1/2)$, then we will have that $\mathbf{E}_{X \sim A}[\mathcal{K}_t(X; m)] = \mathbf{E}_{X \sim \mathrm{Bin}(m,1/2)}[\mathcal{K}_t(X; m)] = 0$, for all $1 \leq t \leq k$.

In order to prove SQ lower bounds for the above testing problem, one needs to find many sets S for which the corresponding $\mathbf{P}_S^A$ are nearly orthogonal. For this, we show that it suffices to find many subsets S whose intersections are pairwise much smaller than m. In particular, we prove that if $|S \cap S'| = o(m)$, then the corresponding inner product will be sufficiently small. This makes our technique somewhat reminiscent of Diakonikolas et al. (2017), which proves lower bounds in the Gaussian setting, where their hard distributions are equal to some moment-matching distribution A in a hidden direction v and are standard Gaussian in the orthogonal directions. Diakonikolas et al. (2017) shows that if two such distributions have hidden-directions u and v, then the chi-squared inner product of these distributions is on the order of $|u^T v|^d$, where d is the number of matching moments. A significant difference with the Gaussian case here is in the way we embed the one-dimensional distribution A as a higher dimensional one. Our main structural lemma for the discrete setting is the following:

Lemma 19 (Correlation Lemma) *Let $k, m, M \in \mathbb{Z}_+$ with $k \leq m \leq M$. If the distribution A on $[m] \cup \{0\}$ satisfies Condition 18, then for all $S, S' \subseteq [M]$ with $|S| = |S'| = m$, we have that $|\chi_{U_M}(\mathbf{P}_S^A, \mathbf{P}_{S'}^A)| \leq (|S \cap S'|/m)^{k+1} \chi^2(A, \mathrm{Bin}(m, 1/2)) + k\nu^2$.*

Proof By definition, we have that

$$1 + \chi^2(\mathbf{P}_S^A, U_M) = 2^M \textstyle\sum_{\mathbf{x} \in \{0,1\}^M} \left(2^{-M+m} A(\sum_{i \in S} x_i) \binom{m}{\sum_{i \in S} x_i}^{-1}\right)^2$$
$$= 2^m \textstyle\sum_{j=0}^m A(j)^2 / \binom{m}{j} = 1 + \chi^2(A, \mathrm{Bin}(m, 1/2)) ,$$

where we let j denote $\sum_{i\in S} x_i$ in the second equality.

Now we proceed via discrete Fourier analysis. Note that by Definition 12, $\widehat{\mathbf{P}_S^A}(T) = \mathbf{E}_{\mathbf{X}\sim\mathbf{P}_S^A}[\chi_T(\mathbf{X})]$, which is: (i) 0 if $T \not\subseteq S$, (ii) $a_j/\binom{m}{j}$ if $T \subseteq S$ and $|T| = j$, where $a_j = \mathbf{E}_{t\sim A}[\mathcal{K}_j(t;m)]$. This is because symmetry implies that for each $T \subseteq S$ with $|T| = j$ we have that $\widehat{\mathbf{P}_S^A}(T)$ is the same. Furthermore, $\mathbf{E}_{t\sim A}[\mathcal{K}_j(t;m)] = \mathbf{E}_{\mathbf{X}\sim\mathbf{P}_S^A}\left[\sum_{T\subseteq S,|T|=j}\chi_T(\mathbf{X})\right]$. From this, by Fact 13, we have

$$\begin{aligned}1+\chi^2(A,\mathrm{Bin}(m,1/2)) &= 1+\chi^2(\mathbf{P}_S^A,U_M) = 1+\chi_{U_M}(\mathbf{P}_S^A,\mathbf{P}_S^A) = \textstyle\sum_{T\subseteq S}\widehat{\mathbf{P}_S^A}(T)^2\\ &= \textstyle\sum_{t=0}^m\sum_{T\subseteq S,|T|=t}\left(a_t/\binom{m}{t}\right)^2 = 1+\sum_{t=1}^m a_t^2/\binom{m}{t}\ ,\end{aligned}$$

where the last equality follows from the fact $\mathcal{K}_0(t;m) = 1$. In addition, by Fact 13, we have that

$$\begin{aligned}1+\chi_{U_M}(\mathbf{P}_S^A,\mathbf{P}_{S'}^A) &= \textstyle\sum_{T\subseteq S\cap S'}\widehat{\mathbf{P}_S^A}(T)\widehat{\mathbf{P}_{S'}^A}(T) = \sum_{t=0}^m\left|\{T : |T|=t, T\subseteq S\cap S'\}\right| a_t^2/\binom{m}{t}^2\\ &= 1+\textstyle\sum_{t=1}^m\binom{|S\cap S'|}{t}a_t^2/\binom{m}{t}^2\ ,\end{aligned}$$

where the last equality follows from the fact $\mathcal{K}_0(t;m) = 1$. By Condition 18, we have that

$$\textstyle\sum_{t=1}^k\binom{|S\cap S'|}{t}a_t^2/\binom{m}{t}^2 \le \sum_{t=1}^k a_t^2 \le k\nu^2\ .$$

The sum over terms with $t \ge k$ is at most

$$\begin{aligned}&\textstyle\sum_{t=k+1}^m\binom{|S\cap S'|}{t}a_t^2/\binom{m}{t}^2 \le (\sum_{t=1}^m a_t^2/\binom{m}{t})\max_{t>k}\binom{|S\cap S'|}{t}/\binom{m}{t}\\ &= \left(\textstyle\sum_{t=1}^m a_t^2/\binom{m}{t}\right)\max_{t>k}\left(\frac{|S\cap S'|(|S\cap S'|-1)\cdots(|S\cap S'|-t+1)}{m(m-1)\cdots(m-t+1)}\right)\\ &\le \left(\textstyle\sum_{t=1}^m a_t^2/\binom{m}{t}\right)\left(\frac{|S\cap S'|}{m}\right)\left(\frac{|S\cap S'|-1}{m-1}\right)\cdots\left(\frac{|S\cap S'|-k}{m-k}\right)\\ &\le \chi^2(A,\mathrm{Bin}(m,1/2))\left(|S\cap S'|/m\right)^{k+1}\ .\end{aligned}$$

This completes the proof. ■

We will additionally require the following (see Appendix A.5 for the proof).

Claim 20 *Let $m, M \in \mathbb{Z}_+$ with $m < M$. For any $0 < c < 1/2$ and $M > 2m^{1+c}$, there exists a collection $\mathcal{C}$ of $2^{m^{1-2c}/4}$ subsets $S \subseteq [M]$ with $|S| = m$ such that any pair $S, S' \in \mathcal{C}$, with $S \neq S'$, satisfies $|S\cap S'| < m^{1-c}$.*

Proposition 21 (Generic Discrete SQ Hardness) *Let $m, M \in \mathbb{Z}_+$ with $M > 2m^{5/4}$. Let A be a distribution on $[m]\cup\{0\}$ satisfying Condition 18. Let $\tau \ge m^{-(k+1)/4}\chi^2(A,\mathrm{Bin}(m,1/2)) + k\nu^2$. Any SQ algorithm that solves the testing problem of Definition 17 with probability at least $2/3$ either makes queries of accuracy better than $\sqrt{2\tau}$ or makes at least $\frac{2^{\Omega(\sqrt{m})}\tau}{\chi^2(A,\mathrm{Bin}(m,1/2))}$ statistical queries.*

Proof Let $\mathcal{C}$ be a collection of $s = 2^{\Omega(\sqrt{m})}$ subsets $S \subseteq [M]$ with $|S| = m$ whose pairwise intersections are all less than $m^{3/4}$. By Claim 20 (taking the local parameter $c = 1/4$), such a set is guaranteed to exist. By Lemma 19, we have that for $S, S' \in \mathcal{C}$ with $S \neq S'$, it holds that

$$|\chi_{U_M}(\mathbf{P}^A_S, \mathbf{P}^A_{S'})| \leq m^{-(k+1)/4}\chi^2(A, \mathrm{Bin}(m, 1/2)) + k\nu^2 \leq \tau \ .$$

If $S = S'$, then $\chi_{U_M}(\mathbf{P}^A_S, \mathbf{P}^A_S) = \chi^2(\mathbf{P}^A_S, U_M) = \chi^2(A, \mathrm{Bin}(m, 1/2))$. Let $\gamma = \tau$ and $\beta = \chi^2(A, \mathrm{Bin}(m, 1/2))$. We have that the statistical query dimension of this testing problem with correlations (γ, β) is at least s. Then applying Lemma 11 with (γ, β) completes the proof. ■

4. SQ Lower Bound for Robustly Learning a Binary Product Distribution

In this section, we use the framework of Section 3 to prove our super-polynomial SQ lower bound for robustly learning a binary product distribution.

Definition 22 (Hard Instances) *Let $0 \leq \delta \leq 1/2$ and $M, m \in \mathbb{Z}_+$ with $M > m$. For any subset $S \subseteq [M]$ with $|S| = m$, define $U_M^{S,\delta}$ to be the product distribution over $\{0,1\}^M$, where each coordinate has mean $1/2 + \delta$ if it belongs to set S, and has mean $1/2$ otherwise. We let $\mu_M^{S,\delta}$ denote the mean vector of $U_M^{S,\delta}$.*

The following lemma states that the distributions in our hardness family are far from the uniform distribution U_M in total variation distance. We defer the proof to Appendix B.2.

Lemma 23 *Let $m, M \in \mathbb{Z}_+$ with $M > m$. Let $S \subseteq [M]$ with $|S| \leq m$. Then for any sufficiently small $\delta > 0$, $d_{\mathrm{TV}}\left(U_M, U_M^{S, \frac{\delta}{\sqrt{m}}}\right) \geq \Omega(\delta)$, where U_M is the uniform distribution over $\{0,1\}^M$.*

The main result of this section is the following theorem:

Theorem 24 (SQ Lower Bound for Robustly Testing a Binary Product) *Fix $0 < c < 1/2$ and k to be a sufficiently large integer. Let $m, M \in \mathbb{Z}_+$ with $M = 3m^{5/4}$. Let $0 < \epsilon < 1/2$ and $m > C'(\log(1/\epsilon))^3$ for some sufficiently large constant $C' > 0$. Let δ be a sufficiently small constant multiple of $\epsilon\sqrt{\log(1/\epsilon)}/k^2$ and $\tau = \Theta(M^{-(k+1)/5}\delta)$. Then any SQ algorithm which is given access to a distribution $\mathbf{P}$ over $\{0,1\}^M$ so that either H_0: $\mathbf{P} = U_M$, or H_1: $d_{\mathrm{TV}}\left(\mathbf{P}, U_M^{S, \frac{\delta}{\sqrt{m}}}\right) \leq \epsilon$ for some unknown subset $S \subseteq [M]$ with $|S| = m$, and correctly distinguishes between these two cases with probability at least $2/3$, must either make queries of accuracy better than $\sqrt{2\tau}$ or must make at least $2^{\Omega(M^{2/5})}M^{-(k+1)/5}$ statistical queries.*

Theorem 24 will follow by applying our generic discrete SQ lower bound construction in Section 3 along with the following proposition.

Proposition 25 *Fix $\delta > 0$ to be sufficiently small and k to be a sufficiently large integer. Let integer $m \geq \max\left(C_0(\log(1/\delta))^3, \frac{k^2}{\log(1/\delta)}\right)$ for some universal constant $C_0 > 0$ sufficiently large. Then there exists a distribution A over $[m] \cup \{0\}$ satisfying the following conditions: (i) $\mathbf{E}_{X\sim A}[X^i] = \mathbf{E}_{X\sim \mathrm{Bin}(m,1/2)}[X^i]$ for all $1 \leq i \leq k$, (ii) $d_{\mathrm{TV}}(A, \mathrm{Bin}(m, 1/2 + \delta/\sqrt{m})) \leq O\left(\frac{\delta k^2}{\sqrt{\log(1/\delta)}}\right)$, and (iii) $\chi^2(A, \mathrm{Bin}(m, 1/2)) = O(\delta)$.*

In Section 6, we give a technical overview of the proof. The detailed proof of Proposition 25 is deferred to Appendix B.1.

Proof [Theorem 24] We can assume without loss of generality that $\epsilon > 0$ is smaller than a sufficiently small universal constant. Let δ be a sufficiently small constant multiple of $\epsilon\sqrt{\log(1/\epsilon)}/k^2$. From Proposition 25, there is a distribution A over $[m] \cup \{0\}$ such that (i) A and $\mathrm{Bin}(m, 1/2)$ agree on the first k moments. (ii) $d_{\mathrm{TV}}(A, \mathrm{Bin}(m, 1/2 + \delta/\sqrt{m})) \leq O(\delta k^2/\sqrt{\log(1/\delta)})$. (iii) $\chi^2(A, \mathrm{Bin}(m, 1/2)) = O(\delta)$. In this way, for any subset $S \subseteq [M]$ with $|S| = m$, it holds that

$$d_{\mathrm{TV}}\left(\mathbf{P}_S^A, U_M^{S,\frac{\delta}{\sqrt{m}}}\right) = (1/2) \textstyle\sum_{\mathbf{x}\in\{0,1\}^M} \left|\mathbf{P}_S^A(\mathbf{x}) - U_M^{S,\frac{\delta}{\sqrt{m}}}(\mathbf{x})\right|$$

$$= (1/2) \textstyle\sum_{\mathbf{x}\in\{0,1\}^M} \Big|2^{-M+m} A\left(\sum_{i\in S} x_i\right) \binom{m}{\sum_{i\in S} x_i}^{-1}$$

$$- 2^{-M+m} \left(1/2 + \delta/\sqrt{m}\right)^{\sum_{i\in S} x_i} \left(1/2 - \delta/\sqrt{m}\right)^{m-\sum_{i\in S} x_i} \Big|$$

$$= (1/2) \textstyle\sum_{j=0}^m \left|A(j) - \binom{m}{j} \left(1/2 + \delta/\sqrt{m}\right)^j \left(1/2 - \delta/\sqrt{m}\right)^{m-j} \right|$$

$$= d_{\mathrm{TV}}(A, \mathrm{Bin}(m, 1/2 + \delta/\sqrt{m})) = O(\delta k^2/\sqrt{\log(1/\delta)}) \leq \epsilon \ .$$

By Claim 20, there exists a collection $\mathcal{C}$ of $2^{\Omega(m)}$ subsets $S \subseteq [M]$ with $|S| = m$ such that for any pair $S, S' \in \mathcal{C}$, with $S \neq S'$, satisfies $|S \cap S'| < m^{3/4}$. Applying Proposition 21, we determine that any SQ algorithm which, given access to a distribution $\mathbf{P}$ so that either $\mathbf{P} = U_M$, or $\mathbf{P}$ is given by $\mathbf{P}_S^A$ for some unknown subset $S \subseteq [M]$ with $|S| = m$, correctly distinguish between these two cases with probability at least $2/3$ must either make queries of accuracy better than $\sqrt{2\tau}$ or must make at least $\frac{2^{\Omega(\sqrt{m})}\tau}{\chi^2(A,\mathrm{Bin}(m,1/2))} \geq 2^{\Omega(M^{2/5})} M^{-(k+1)/5}$ statistical queries, since $m^{-(k+1)/4}\chi^2(A, \mathrm{Bin}(m, 1/2)) \leq O(M^{-(k+1)/5}\delta) \leq \tau$. This completes the proof of Theorem 24.
■

5. SQ Lower Bound for Robustly Learning a Ferromagnetic High-Temperature Ising Model

In this section, we prove our super-polynomial SQ lower bound for robustly learning a ferromagnetic high-temperature Ising model. We start by transforming the support of Ising models to $\{0, 1\}^M$:

Definition 26 *Given a real symmetric matrix $(\theta_{ij})_{i,j\in[M]}$ with zero diagonal, the Ising model distribution P_θ is defined as follows: For any $\mathbf{x} \in \{0, 1\}^M$, $P_\theta(\mathbf{x}) = \frac{1}{Z(\theta)} \exp\left((1/2) \sum_{i,j\in[M]} (-1)^{x_i+x_j}\theta_{ij}\right)$, where the normalizing factor $Z(\theta)$ is called the partition function. We call the matrix $(\theta_{ij})_{i,j\in[M]} \in \mathbb{R}^{M\times M}$ the interaction matrix.*

Definition 27 (Hard Instances) *Let $m, M \in \mathbb{Z}_+$ with $M > m$. Let $0 \leq \delta \leq \frac{1}{2m}$. For every subset $S \subseteq [M]$ with $|S| = m$, define $Q_M^{S,\delta}$ to be the Ising model with parameter θ, where for every pair $i \neq j \in [M]$ we have that $\theta_{ij} = \delta, \forall i, j \in S$ and $\theta_{ij} = 0$ otherwise. Note that by our choice of parameter δ, the Ising models $Q_M^{S,\delta}$ are both high-temperature and ferromagnetic.*

The following lemma states that the distributions in our hardness family are far from the uniform distribution U_M in total variation distance. We defer the proof to Appendix C.2.

Lemma 28 *Let $m, M \in \mathbb{Z}_+$ with $M > m$. Let $S \subseteq [M]$ with $|S| \leq m$. Then, for any sufficiently small $\delta > 0$, we have that $d_{\mathrm{TV}}\left(U_M, Q_M^{S,\frac{\delta}{m}}\right) \geq \Omega(\delta)$.*

The main result of this section is the following theorem:

Theorem 29 (SQ Lower Bound for Robustly Testing Ising Models) *Fix $0 < c < 1$ and k to be a sufficiently large integer. Let $m, M \in \mathbb{Z}_+$ with $M = 3m^{5/4}$. Let $0 < \epsilon < 1/2$ and $m > C'(\log(1/\epsilon))^3$ for some sufficiently large constant $C' > 0$. Let δ be a sufficiently small multiple of $\epsilon \log(1/\epsilon)/k^3$ and $\tau = \Theta(M^{-(k+1)/5}\delta)$. Then any SQ algorithm which is given access to a distribution $\mathbf{P}$ over $\{0,1\}^M$ so that either H_0: $\mathbf{P} = U_M$, or H_1: $d_{\mathrm{TV}}\left(\mathbf{P}, Q_M^{S,\frac{\delta}{m}}\right) \leq \epsilon$ for some unknown subset $S \subseteq [M]$ with $|S| = m$, and correctly distinguishes between these two cases with probability at least $2/3$ must either make queries of accuracy better than $\sqrt{2\tau}$ or must make at least $2^{\Omega(M^{2/5})}M^{-(k+1)/5}$ statistical queries.*

Definition 30 *Fix n to be a positive integer. Let $\mathrm{IS}(n, \delta)$ be the distribution over $[n] \cup \{0\}$ with $\mathrm{IS}(n,\delta)(x) = \binom{n}{x} \exp\left(h(n,x)\delta\right)/Z_n(\delta)$ for some parameter $-1/n < \delta < 1/n$, where $h(n,x) = 2x^2 - 2nx + \frac{n(n-1)}{2}$ and $Z_n(\delta) = \sum_{x=0}^{n} \binom{n}{x} \exp(h(n,x)\delta)$.*

By Definition 26, $Z_n(\delta)$ is the partition function of the Ising model over $\{0,1\}^n$, where every entry outside of the diagonal of the interaction matrix is δ. Intuitively, $\mathrm{IS}(n,\delta)(x)$ denotes the contribution of the configurations containing x 1's in the Ising model.

Theorem 29 will follow by applying our generic discrete SQ lower bound construction of Section 3 along with the following proposition.

Proposition 31 *Fix $\delta > 0$ to be sufficiently small and k to be an arbitrary positive integer. Let integer $m \geq \max\left(C_0(\log(1/\delta))^3, \frac{k^2}{\log(1/\delta)}\right)$ for some universal constant $C_0 > 0$ sufficiently large. Then there exists a distribution A over $[m] \cup \{0\}$ satisfying the following conditions: (i) $\mathbf{E}_{X\sim A}[X^i] = \mathbf{E}_{X\sim\mathrm{Bin}(m,1/2)}[X^i]$ for all $1 \leq i \leq k$, (ii) $d_{\mathrm{TV}}(A, \mathrm{IS}(m, \delta/m)) \leq O\left(\frac{\delta k^3}{\log(1/\delta)}\right)$, in addition, (iii) $\chi^2(A, \mathrm{Bin}(m, 1/2)) = O(\delta)$.*

In Section 6, we give a technical overview of the proof. The detailed proof of Proposition 31 is deferred to Appendix C.1.

Proof [Theorem 29] We can assume without loss of generality that $\epsilon > 0$ is smaller than a sufficiently small universal constant. Let δ be a sufficiently small constant multiple of $\epsilon\log(1/\epsilon)/k^3$. From Proposition 31, there is a distribution A over $[m]\cup\{0\}$ such that (i) $\mathbf{E}_{X\sim A}[X^i] = \mathbf{E}_{X\sim\mathrm{Bin}(m,1/2)}[X^i]$ for all $0 \leq i \leq k$, (ii) $d_{\mathrm{TV}}(A, \mathrm{IS}(m,\delta/m)) \leq O\left(\frac{\delta k^3}{\log(1/\delta)}\right) \leq O(\epsilon)$, and (iii) $\chi^2(A, \mathrm{Bin}(m, 1/2)) = O(\delta)$. Note that for any subset $S \subseteq [M]$ with $|S| = m$, it holds that

$$d_{\mathrm{TV}}\left(\mathbf{P}_S^A, Q_M^{S,\frac{\delta}{m}}\right) = (1/2)\textstyle\sum_{\mathbf{x}\in\{0,1\}^M} \left|\mathbf{P}_S^A(\mathbf{x}) - Q_M^{S,\frac{\delta}{m}}(\mathbf{x})\right|$$

$$= (1/2)\textstyle\sum_{\mathbf{x}\in\{0,1\}^M} \left|2^{-M+m} A\left(\sum_{i\in S} x_i\right)\binom{m}{\sum_{i\in S} x_i}^{-1} - 2^{-M+m}\left(\exp(h(m, \sum\limits_{i\in S} x_i)\delta/m)/Z_n(\delta/m)\right)\right|$$

$$= (1/2)\sum_{j=0}^{m} \left|A(j) - \binom{m}{j}\exp(h(m,j)\delta/m)/Z_n(\delta/m)\right| = d_{\mathrm{TV}}(A, \mathrm{IS}(m,\delta/m)) = O(\delta k^3/\log(1/\delta)) \leq \epsilon\ .$$

By Claim 20, there exists a collection $\mathcal{C}$ of $2^{\Omega(m)}$ subsets $S \subseteq [M]$ with $|S| = m$ such that for any pair $S, S' \in \mathcal{C}$, with $S \neq S'$, satisfies $|S \cap S'| < m^{3/4}$. Applying Proposition 21, we determine that any SQ algorithm which, given access to a distribution $\mathbf{P}$ so that either $\mathbf{P} = U_M$, or $\mathbf{P}$ is given by $\mathbf{P}_S^A$ for some unknown subset $S \subseteq [M]$ with $|S| = m$, correctly distinguish between these two cases with probability at least $2/3$ must either make queries of accuracy better than $\sqrt{2\tau}$ or must make at least $\frac{2^{\Omega(\sqrt{m})}\tau}{\chi^2(A,\mathrm{Bin}(m,1/2))} \geq 2^{\Omega(M^{2/5})} M^{-(k+1)/5}$ statistical queries, since $m^{-(k+1)/4}\chi^2(A, \mathrm{Bin}(m, 1/2)) \leq O(M^{-(k+1)/5}\delta) \leq \tau$. This completes the proof of Theorem 29. ■

6. Proof Sketch of Proposition 25 and Proposition 31

The construction of the distribution A in both cases is similar in spirit to the technique in Diakonikolas et al. (2017) for constructing a distribution that matches moments with $\mathcal{N}(0, 1)$ but is close in total variation distance to $\mathcal{N}(\delta, 1)$, for an appropriate $\delta > 0$. Specifically, we start from some appropriate one-dimensional version of either a binary product distribution or Ising model, $H(x)$, over $[m] \cup \{0\}$, and then modify it in order to match the first k moments with $\mathrm{Bin}(m, 1/2)$. We achieve this by modifying the probability mass function of A by adding a polynomial q over some appropriately chosen interval $I = [(1/2 - C)m, (1/2 + C)m]$, for some carefully selected $C = \Theta(\sqrt{(\log(1/\delta)/m})$. In particular, for any integer point $x \in I$, we let $q(x) = \int_x^{x+1} p(t)dt$, for some real polynomial p of degree-k and then modify the probability mass function by adding $q(x)$ to $H(x)$. The moment-matching condition amounts to a system of linear equations on the coefficients of p. We show that this system has a unique solution. Then the rest of our analysis focuses on showing that this modification leaves the probability mass function of A still non-negative and sufficiently close to H in total variation distance.

In particular, we express the polynomial p as a linear combination of appropriately scaled Legendre polynomials, i.e., $p(t) = \sum_{i=0}^k a_i P_i\left(\frac{t-m/2}{Cm}\right)$, where P_i denotes the i-th Legendre polynomial and $a_i \in \mathbb{R}$ is a coefficient. Then we show, by analogy to the proof in Diakonikolas et al. (2017), that the L_1 and L_∞ norms of p within the interval I are sufficiently small. In particular, the Diakonikolas et al. (2017) result on the hardness of robustly learning unknown-mean or covariance Gaussians essentially solves the limiting version of this problem (that is achieved as $m \to \infty$). As their analysis shows that this limiting case works, we need to show that when m is sufficiently large, we are sufficiently close to that limiting case that our construction will also succeed. To achieve this, we require some new proof ideas in order to show that with sufficiently large but finite m, our analysis will be close enough to that of the limiting case, so that the results of Diakonikolas et al. (2017) can still be applied. In more detail, by our construction of the polynomial p and the moment-matching condition, we are able to bound from above the coefficients a_i as follows:

$$|a_i| = \left(\frac{2i+1}{2Cm}\right)\left|\int_{t\in I} p(t)P_i\left(\frac{t-m/2}{Cm}\right)dt\right| \leq \frac{(2i+1)(\gamma_i+\beta_i)}{2Cm}, \qquad \forall 1 \leq i \leq k\ ,$$

where we have that $\gamma_i = \left|\sum_{x\in\mathbb{Z}\cap I} P_i\left(\frac{x-m/2}{Cm}\right)\int_x^{x+1} p(t)dt - \int_{t\in I} p(t)P_i\left(\frac{t-m/2}{Cm}\right)dt\right|$ and $\beta_i = \left|\sum_{x=0}^m (H(x) - \mathrm{Bin}(m, 1/2)(x))P_i\left(\frac{x-m/2}{Cm}\right)\right|$. Intuitively speaking, the quantity β_i represents the answer to the continuous version of the problem and the quantity γ_i inherently captures the error between the discrete and limiting continuous versions of our problems. Since the absolute value of the derivative of the i-th Legendre polynomial is at most $O(i^2)$ in the interval $[-1,1]$, by the last property of Fact 4, we are able to apply the mean-value theorem to obtain an upper bound for γ_i in terms of the L_1-norm of p within the interval I. For the quantity β_i, we borrow ideas from Diakonikolas et al. (2017) to view $H(x)-\mathrm{Bin}(m,1/2)(x)$ as a function of some appropriately chosen parameter, and apply Taylor's theorem to expand this difference up to second order terms. Then we can show that both the first order and second order terms are sufficiently small. In summary, we prove the following technical result.

Theorem 32 *Fix $\delta>0$, $0<C<1/2$ and $k\in\mathbb{Z}_+$. Let integer $m>k/(2C)$ such that both $(1/2-C)m$ and $(1/2+C)m$ are integers. Consider the interval $I_{C,m}=[(1/2-C)m,(1/2+C)m-1]$. Let $\{H_{m,x}(t)\}_{x\in[m]\cup\{0\}}$ be a family of real functions. Then there is a unique real polynomial p of degree at most k such that*

$$\sum\nolimits_{x\in\mathbb{Z}\cap I_{C,m}} x^i\int_x^{x+1}p(t)dt=\sum_{x=0}^{m}(H_{m,x}(0)-H_{m,x}(\delta))x^i:=b_i,\qquad \forall 0\le i\le k\ . \tag{1}$$

In addition, we can write $p(t)=\sum_{i=1}^{k}a_iP_i\left(\frac{t-m/2}{Cm}\right)$, where

$$|a_i|\le\left(\frac{2i+1}{2Cm}\right)\left(\beta_i+O\left(\frac{i^2}{Cm}\right)\int_{(1/2-C)m}^{(1/2+C)m}|p(t)|dt\right)\ , \tag{2}$$

for all $1\le i\le k$, where $\beta_i=\left|\sum_{x=0}^{m}(H_{m,x}(0)-H_{m,x}(\delta))P_i\left(\frac{x-m/2}{Cm}\right)\right|$.

Proof We first show that there is a unique real polynomial p of degree at most k satisfying (1). Let $p(t)=\sum_{i=0}^{k}p_it^i$ and $q(x)=\int_x^{x+1}p(t)dt$. We note that each value of i implies a single linear condition on q. This suggests that as long as the support domain $I_{C,m}$ is sufficiently large, we can simply solve a system of linear equations to find it. In more detail, we start by establishing the relationship between p and q. By definition, we have that

$$\begin{aligned}q(x)&=\sum_{i=0}^{k}\frac{p_i((x+1)^{i+1}-x^{i+1})}{i+1}=\sum_{i=0}^{k}\left(\frac{p_i}{i+1}\right)\sum_{j=0}^{i}\binom{i+1}{j}x^j\\&=\sum_{j=0}^{k}x^j\sum_{i=j}^{k}\left(\frac{p_i}{i+1}\right)\binom{i+1}{j}=\sum_{j=0}^{k}q_jx^j\ ,\end{aligned}$$

where $q_j=\sum_{i=j}^{k}\left(\frac{p_i}{i+1}\right)\binom{i+1}{j}$.

This gives us a linear equation to solve for p_i in terms of q_j that is upper triangular and thus has a unique solution. For any two polynomials $r_1(x),r_2(x)$ of degree at most k, we consider the inner product $\langle r_1,r_2\rangle\in\mathbb{R}$ given by $\langle r_1,r_2\rangle:=\sum_{x\in\mathbb{Z}\cap I_{C,m}}r_1(x)r_2(x)$. To show that this inner product is non-degenerate as long as m is sufficiently large, we need to show that for any polynomial $r(x)$ of degree at most k, it holds that $\sum_{x\in\mathbb{Z}\cap I_{C,m}}r^2(x)=0\Rightarrow r=0$. By our assumptions of C,k,m, $\sum_{x\in\mathbb{Z}\cap I_{C,m}}r^2(x)=0$ will imply that the polynomial $r(x)$ of degree at most k has at least $2Cm>k$ different roots, which implies $r=0$. Therefore, we can write the LHS of (1) as

$$\langle x^i,q(x)\rangle=b_i,\qquad 0\le i\le k\ . \tag{3}$$

Since $1, x, \cdots, x^k$ are linearly independent polynomials of degree at most k, there exists a unique polynomial q of degree at most k satisfying the system of equations (3).

To show inequality (2), we first express p as a linear combination of scaled Legendre polynomials whose coefficients are explicitly given by integrals. In particular, since p has degree at most k and the set of polynomials $\left\{P_i\left(\frac{t-m/2}{Cm}\right)\right\}_{0\leq i\leq k}$ contains a polynomial of each degree from 0 to k, there exist $a_i \in \mathbb{R}$ such that $p(t) = \sum_{i=0}^{k} a_i P_i\left(\frac{t-m/2}{Cm}\right)$. It follows from Fact 4 (ii) that

$$\int_{(1/2-C)m}^{(1/2+C)m} p(t)P_i\left(\frac{t-m/2}{Cm}\right)dt = \sum_{j=0}^{k} a_j \int_{(1/2-C)m}^{(1/2+C)m} P_i\left(\frac{t-m/2}{Cm}\right)P_j\left(\frac{t-m/2}{Cm}\right)dt$$
$$= Cm\textstyle\sum_{j=0}^{k} a_j \int_{-1}^{1} P_i(t)P_j(t)dt = (2Cma_i)/(2i+1)\ ,$$

which implies that $a_i = \left(\frac{2i+1}{2Cm}\right)\int_{(1/2-C)m}^{(1/2+C)m} p(t)P_i\left(\frac{t-m/2}{Cm}\right)dt$, for all $0 \leq i \leq k$. In addition, by Equation (1), we have that $a_0 = \frac{1}{2Cm}\int_{(1/2-C)m}^{(1/2+C)m} p(t)dt = 0$. We now bound $|a_i|$ as follows. Let

$$\gamma_i = \left|\sum\nolimits_{x\in\mathbb{Z}\cap I_{C,m}} P_i\Big(\frac{x-m/2}{Cm}\Big)\int_x^{x+1} p(t)dt - \int_{(1/2-C)m}^{(1/2+C)m} p(t)P_i\Big(\frac{t-m/2}{Cm}\Big)dt\right|\ .$$

By Fact 4 (viii) and the mean-value theorem, we have that

$$\begin{aligned}\gamma_i &= \Big|\sum\nolimits_{x\in\mathbb{Z}\cap I_{C,m}}\int_x^{x+1}\Big(P_i\Big(\frac{x-m/2}{Cm}\Big) - P_i\Big(\frac{t-m/2}{Cm}\Big)\Big)p(t)dt\Big| \\ &= \Big(\frac{1}{Cm}\Big)\Big|\sum\nolimits_{x\in\mathbb{Z}\cap I_{C,m}}\int_x^{x+1}(x-t)P_i'\Big(\frac{\xi_t-m/2}{Cm}\Big)p(t)dt\Big| \\ &\leq O\Big(\frac{i^2}{Cm}\Big)\sum\nolimits_{x\in\mathbb{Z}\cap I_{C,m}}\int_x^{x+1}|p(t)|dt = O\Big(\frac{i^2}{Cm}\Big)\int_{(1/2-C)m}^{(1/2+C)m}|p(t)|dt\ ,\end{aligned}$$

where ξ_t is some real number between x and t for each $t \in [x, x+1)$. Therefore, by Equation (1), we have that

$$\begin{aligned}|a_i| &= \Big|\Big(\frac{2i+1}{2Cm}\Big)\int_{(1/2-C)m}^{(1/2+C)m} p(t)P_i\Big(\frac{t-m/2}{Cm}\Big)dt\Big| \\ &\leq \Big(\frac{2i+1}{2Cm}\Big)\Big(\gamma_i + \Big|\sum\nolimits_{x\in\mathbb{Z}\cap I_{C,m}} P_i\Big(\frac{x-m/2}{Cm}\Big)\int_x^{x+1} p(t)dt\Big|\Big) \\ &= \Big(\frac{2i+1}{2Cm}\Big)(\gamma_i+\beta_i) \leq \Big(\frac{2i+1}{2Cm}\Big)\Big(\beta_i + O\Big(\frac{i^2}{Cm}\Big)\int_{(1/2-C)m}^{(1/2+C)m}|p(t)|dt\Big)\ ,\end{aligned}$$

where the last equality follows from Equation (1) and

$$\sum\nolimits_{x\in\mathbb{Z}\cap I_{C,m}} P_i\Big(\frac{x-m/2}{Cm}\Big)\int_x^{x+1} p(t)dt = \sum_{x=0}^{m}(H_{m,x}(0) - H_{m,x}(\delta))P_i\Big(\frac{x-m/2}{Cm}\Big),$$

since $P_i\left(\frac{x-m/2}{Cm}\right)$ is a polynomial in x of degree i. This completes the proof. ■

References

R. Adamczak, M. Kotowski, B. Polaczyk, and M. Strzelecki. A note on concentration for polynomials in the ising model. *Electronic Journal of Probability*, 24, 2019.

G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, and K.-R. Müller. In search of non-gaussian components of a high-dimensional distribution. *Journal of Machine Learning Research*, 7(9):247–282, 2006. URL http://jmlr.org/papers/v7/blanchard06a.html.

A. Blum, A. Kalai, and H. Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM*, 50(4):506–519, 2003.

M. Brennan, G. Bresler, S. B. Hopkins, J. Li, and T.Schramm. Statistical query algorithms and low-degree tests are almost equivalent. *CoRR*, abs/2009.06107, 2020. URL https://arxiv.org/abs/2009.06107.

S. Bubeck, E. Price, and I. P. Razenshteyn. Adversarial examples from computational constraints. *CoRR*, abs/1805.10204, 2018. URL http://arxiv.org/abs/1805.10204.

S. Chatterjee. *Concentration inequalities with exchangeable pairs*. PhD thesis, Stanford University, 2005.

C.-T. Chu, S. K. Kim, Y. A. Lin, Y. Yu, G. Bradski, A. Y. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 281–288, Cambridge, MA, USA, 2006. MIT Press.

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

D. Dachman-Soled, V. Feldman, L.-Y. Tan, A. Wan, and K. Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. In Piotr Indyk, editor, *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 498–511. SIAM, 2015.

Y. Dagan, C. Daskalakis, N. Dikkala, and A. V. Kandiros. Estimating ising models from one sample. *arXiv preprint arXiv:2004.09370*, 2020.

C. Daskalakis, N. Dikkala, and G. Kamath. Concentration of multilinear functions of the ising model with applications to network data. *Advances in Neural Information Processing Systems*, 30:12–23, 2017.

I. Diakonikolas and D. M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *CoRR*, abs/1911.05911, 2019. URL http://arxiv.org/abs/1911.05911.

I. Diakonikolas and D. M. Kane. Near-optimal statistical query hardness of learning halfspaces with massart noise. *CoRR*, abs/2012.09720, 2020. URL https://arxiv.org/abs/2012.09720.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Proceedings of FOCS'16*, pages 655–664, 2016.

I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pages 73–84, 2017. Full version at http://arxiv.org/abs/1611.03473.

I. Diakonikolas, D. M. Kane, and A. Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 1047–1060, 2018. Full version available at https://arxiv.org/abs/1711.07211.

I. Diakonikolas, W. Kong, and A. Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, pages 2745–2754, 2019.

I. Diakonikolas, D. M. Kane, V. Kontonis, and N. Zarifis. Algorithms and SQ lower bounds for PAC learning one-hidden-layer relu networks. In *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pages 1514–1539. PMLR, 2020a.

I. Diakonikolas, D. M. Kane, and N. Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and relus under gaussian marginals. *CoRR*, abs/2006.16200, 2020b. URL https://arxiv.org/abs/2006.16200. Conference version in NeurIPS'20.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robustness meets algorithms. *Commun. ACM*, 64(5):107–115, 2021a.

I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Learning general halfspaces with general massart noise under the gaussian distribution. *CoRR*, abs/2108.08767, 2021b. URL https://arxiv.org/abs/2108.08767.

I. Diakonikolas, D. M. Kane, A. Pensia, T. Pittas, and A. Stewart. Statistical query lower bounds for list-decodable linear regression. *CoRR*, abs/2106.09689, 2021c. URL https://arxiv.org/abs/2106.09689.

I. Diakonikolas, D. M. Kane, T. Pittas, and N. Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the SQ model. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 1552–1584. PMLR, 2021d.

I. Diakonikolas, D. M. Kane, A. Stewart, and Y. Sun. Outlier-robust learning of Ising models under Dobrushin's condition. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 1645–1682. PMLR, 2021e. URL http://proceedings.mlr.press/v134/diakonikolas21e.html.

V. Feldman. A general characterization of the statistical query complexity. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 785–830. PMLR, 2017.

V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of STOC'13*, pages 655–664, 2013. Full version in Journal of the ACM, 2017.

V. Feldman, W. Perkins, and S. Vempala. On the complexity of random satisfiability problems with planted solutions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC, 2015*, pages 77–86, 2015.

V. Feldman, C. Guzman, and S. S. Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In Philip N. Klein, editor, *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017*, pages 1265–1277. SIAM, 2017.

R. Gheissari, E. Lubetzky, and Y. Peres. Concentration inequalities for polynomials of contracting ising models. *Electronic Communications in Probability*, 23, 2018.

F. Götze, H. Sambale, and A. Sinulis. Higher order concentration for functions of weakly dependent random variables. *Electronic Journal of Probability*, 24, 2019.

N. Goyal and A. Shetty. Non-gaussian component analysis using entropy methods. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019*, pages 840–851. ACM, 2019.

P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

M. J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6): 983–1006, 1998.

C. Külske. Concentration inequalities for functions of gibbs fields with application to diffraction and random gibbs measures. *Communications in mathematical physics*, 239(1-2):29–51, 2003.

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *Proceedings of FOCS'16*, 2016.

K. Marton. Logarithmic sobolev inequalities in discrete product spaces: a proof by a transportation cost distance. *arXiv preprint arXiv:1507.02803*, 2015.

G. Szegö. *Orthogonal Polynomials*, volume XXIII of *American Mathematical Society Colloquium Publications*. A.M.S, Providence, 1989.

Y. S. Tan and R. Vershynin. Polynomial time and sample complexity for non-gaussian component analysis: Spectral methods. In *Conference On Learning Theory, COLT 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 498–534. PMLR, 2018.

J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

L. G. Valiant. A theory of the learnable. In *Proc. 16th Annual ACM Symposium on Theory of Computing (STOC)*, pages 436–445. ACM Press, 1984.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Appendix

Appendix A. Omitted Technical Preliminaries

In this section, we record the required definitions and technical facts.

A.1. Basic Facts

Fact 33 $\binom{2n}{n} \le \frac{2^{2n}}{\sqrt{2n}}, \forall n \in \mathbb{Z}_+ \setminus \{0\}$ *and* $\binom{2n+1}{n} \le \frac{2^{2n+1}}{\sqrt{2n+1}}, \forall n \in \mathbb{Z}_+$.

Fact 34 (Cover and Thomas (1991)) *Let $n, k \in \mathbb{Z}$. Then, we have that*

$$\sqrt{\frac{n}{8k(n-k)}}2^{nH(k/n)} \le \binom{n}{k} \le \sqrt{\frac{n}{\pi k(n-k)}}2^{nH(k/n)},$$

where $H(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function.

We will use the following fact to bound from below the expectation of a real random variable.

Fact 35 *Let X be a real random variable with $\mathbf{E}[X^4] > 0$. Then, we have that $\mathbf{E}[|X|] \ge \frac{\mathbf{E}[X^2]^{3/2}}{\mathbf{E}[X^4]^{1/2}}$.*

A.2. Sub-Gaussian and Sub-Exponential Distributions

Here we present basic facts about sub-Gaussian and sub-exponential distributions. The reader is referred to Vershynin (2018).

Definition 36 (Sub-Gaussian Distribution) *A random variable X over $\mathbb{R}$ is sub-Gaussian if $\|X\|_{\psi_2} := \inf\{t > 0 : \mathbf{E}[\exp(X^2/t^2)] \le 2\}$ is finite.*

Definition 37 (Sub-Exponential Distribution) *A random variable X over $\mathbb{R}$ is sub-exponential if $\|X\|_{\psi_1} := \inf\{t > 0 : \mathbf{E}[\exp(|X|/t)] \le 2\}$ is finite.*

Fact 38 *Let X be a real random variable. Suppose there is a real number $K > 0$ such that $\mathbf{Pr}[|X| > t] \le 2\exp(-t^2/K^2)$. Then X is sub-Gaussian with $\|X\|_{\psi_2} \le cK$ for some universal constant $c > 0$. In addition, we have that $\mathbf{E}[|X|^p] \le \min\left(pK^p \left\lfloor \frac{p-1}{2} \right\rfloor!, (c'K\sqrt{p})^p\right)$, where $c' > 0$ is a universal constant.*

Fact 39 *Let X be a real random variable. Suppose there is a real number $K > 0$ such that $\mathbf{Pr}[|X| > t] \le 2\exp(-t/K)$. Then X is sub-exponential with $\|X\|_{\psi_1} \le cK$ for some universal constant $c > 0$. In addition, we have that $\mathbf{E}[|X|^p] \le 2K^p p! \le 2(Kp)^p$.*

Fact 40 *$\|\cdot\|_{\psi_2}$ is a norm on the space of sub-Gaussian random variables. $\|\cdot\|_{\psi_1}$ is a norm on the space of sub-exponential random variables.*

Fact 41 *Let X be sub-Gaussian and Y be sub-exponential with $\mathbf{E}[X] = \mathbf{E}[Y] = 0$. Then there exists universal constants $C_1, C_2 > 0$ such that $\mathbf{E}[\exp(\lambda X)] \le \exp(C_1^2\lambda^2\|X\|_{\psi_2}^2), \forall \lambda \in \mathbb{R}$ and $\mathbf{E}[\exp(\lambda Y)] \le \exp(C_2^2\lambda^2\|Y\|_{\psi_1}^2), |\lambda| \le \frac{1}{C_2\|Y\|_{\psi_1}}$.*

Fact 42 *Let X be sub-Gaussian and Y be sub-exponential. Then, there exists universal constants $c_1, c_2 > 0$ such that $\|X - \mathbf{E}[X]\|_{\psi_2} \le c_1\|X\|_{\psi_2}$ and $\|Y - \mathbf{E}[Y]\|_{\psi_1} \le c_2\|Y\|_{\psi_1}$.*

A.3. Dobrushin's Uniqueness Condition

Here we introduce the original definition of Dobrushin's condition through the influence between points in general graphical model.

Definition 43 (Influence in Graphical Models) *Let D be a distribution over some set of points V. Let S_j denote the set of state pairs (X, Y) which differ only at point j. Then the influence of point $j \in V$ on point $i \in V$ is defined as*

$$I(j,i) = \max_{(X,Y)\in S_j} d_{\mathrm{TV}}(D_i(\cdot \mid X_{-i}), D_i(\cdot \mid Y_{-i})) ,$$

where $D_i(\cdot \mid X_{-i}), D_i(\cdot \mid Y_{-i})$ denote the marginal distribution of point i conditioning on X_{-i} and Y_{-i} respectively.

Definition 44 (Dobrushin's Uniqueness Condition) *Let D be a distribution over some set of points V. Then D is said to satisfy Dobrushin's uniqueness condition if $\max_{i\in V} \sum_{j\in V} I(j,i) < 1$.*

For Ising models, Chatterjee (2005) proves that $\max_{i\in V} \sum_{j\neq i} |\theta_{ij}| < 1$ implies the Dobrushin's uniqueness condition.

A.4. Concentration of Ising Models

Several recent works have studied the concentration and anti-concentration of functions of Ising models (Gheissari et al., 2018; Götze et al., 2019; Daskalakis et al., 2017; Adamczak et al., 2019). Here we record some results which will be used throughout this article.

The following two facts state that for an Ising model satisfying Dobrushin's condition, for some constant $\eta > 0$, the linear form and the quadratic form of Ising models are sub-Gaussian and sub-exponential respectively.

Fact 45 (Götze et al. (2019)) *Let P_θ be an Ising model satisfying Dobrushin's condition, i.e., $\max_{i\in[d]} \sum_{j\neq i} |\theta_{ij}| \leq 1 - \eta$, for some constant $0 < \eta < 1$. Then there is a constant $c(\eta) > 0$ such that for any $b \in \mathbb{R}^d$ and any $t > 0$, we have that $\mathbf{Pr}_{X\sim P_\theta}\left[\left|b^T X - \mathbf{E}_{X\sim P_\theta}\left[b^T X\right]\right| > t\right] \leq 2\exp\left(-\frac{t^2}{c(\eta)\|b\|_2^2}\right)$. This implies that $\left\|b^T X - \mathbf{E}\left[b^T X\right]\right\|_{\psi_2} \leq c'(\eta)\|b\|_2$ for some constant $c'(\eta) > 0$.*

Fact 46 (Götze et al. (2019)) *Let P_θ be an Ising model satisfying Dobrushin's condition, i.e., $\max_{i\in[d]} \sum_{j\neq i} |\theta_{ij}| \leq 1 - \eta$, for some constant $0 < \eta < 1$. Then there is a constant $c(\eta) > 0$ such that for any symmetric matrix $A \in \mathbb{R}^{d\times d}$ with zero diagonal and any $t > 0$, we have that $\mathbf{Pr}_{X\sim P_\theta}\left[\left|X^T AX - \mathbf{E}_{X\sim P_\theta}\left[X^T AX\right]\right| > t\right] \leq 2\exp\left(-\frac{t}{c(\eta)\|A\|_F}\right)$. This implies that $\left\|X^T AX - \mathbf{E}_{X\sim P_\theta}\left[X^T AX\right]\right\|_{\psi_1} \leq c'(\eta)\|A\|_F$ for some constant $c'(\eta) > 0$.*

A.5. Basic Facts about the Hypergeometric Distribution

Let $k, n, N \in \mathbb{Z}_+$. Consider an urn consisting of N balls in total among which k are red, and $N - k$ are blue. Let X denote the number of red balls obtained by sampling n balls from the urn *without* replacements. In this way, we say that $X \sim \mathrm{Hypergeom}(K, N, n)$. We will also use the following standard fact:

Fact 47 *Let $X \sim \mathrm{Hypergeom}(K, N, n)$ and $p = K/N$. Then for any $t > 0$, we have that*

$$\mathbf{Pr}\left[X > np + t\right] \le \exp\left(-2t^2/n\right) .$$

A.5.1. Proof of Claim 20

Let S and S' be independent uniformly random subsets from $[M]$ with $|S| = |S'| = m$. Note that $|S \cap S'| \sim \mathrm{Hypergeom}(m, M, m)$, by Fact 47, we know that

$$\mathbf{Pr}[|S \cap S'| \ge m^{1-c}] \le \mathbf{Pr}\left[|S \cap S'| \ge m\left(\frac{m}{M} + \frac{m^{-c}}{2}\right)\right] \le \exp\left(-\frac{m^{1-2c}}{2}\right) .$$

Therefore, by the union bound,

$$\mathbf{Pr}[\exists |S| = |S'| = m : |S \cap S'| \ge m^{1-c}] \le 2^{\frac{m^{1-2c}}{2}} \cdot \exp\left(-\frac{m^{1-2c}}{2}\right) < 1.$$

A.6. Reduction of Testing to Learning

We have the following simple claim:

Claim 48 *Suppose there exists an SQ algorithm to learn an unknown distribution in $\mathcal{D}$ to total variation distance ϵ using at most N statistical queries of tolerance τ. Suppose furthermore that for each $D' \in \mathcal{D}$ we have that $d_{\mathrm{TV}}(D, D') > 2(\tau + \epsilon)$. Then there exists an SQ algorithm that solves the testing problem $\mathcal{B}(\mathcal{D}, D)$ using at most $n + 1$ queries of tolerance τ.*

Proof We begin by running the learning algorithm under the assumption that the unknown distribution in question is $D_0 \in \mathcal{D}$ to get a hypothesis distribution D'. We let S be a subset so that $d_{\mathrm{TV}}(D, D') = |D(S) - D'(S)|$, and use an additional statistical query to get an estimate v of the expectation of $\mathbb{I}[S]$, the indicator function of S. If the original distribution was D, we have that $|v - D(S)| \le \tau$. If the original distribution was D_0, we have that $|v - D'(S)| \le |v - D_0(S)| + |D_0(S) - D'(S)| \le \tau + \epsilon$. However, we have that

$$|D(S) - D'(S)| = d_{\mathrm{TV}}(D, D') \ge d_{\mathrm{TV}}(D, D_0) - d_{\mathrm{TV}}(D_0, D') > 2(\tau + \epsilon) - \epsilon = 2\tau + \epsilon .$$

Therefore, our distribution is in $\mathcal{D}$ if and only if the expectation of $\mathbb{I}[S]$ is within $\tau + \epsilon$ of $D'(S)$. Thus, determining which of these cases holds will solve our decision problem. ■

Appendix B. Omitted Statements and Proofs from Section 4

Definition 49 *Fix $0 < c < 1/2$ to be a constant. We say that a binary product distribution is c-balanced if every coordinate of the mean vector is in $[c, 1 - c]$.*

For c-balanced binary product distributions, we have the following lemma.

Lemma 50 *Let P and Q be c-balanced binary product distributions with mean vectors μ_p and μ_q. Then, $d_{\mathrm{TV}}(P, Q) \le O(\|\mu_p - \mu_q\|_2/\sqrt{c})$.*

We provide the result for hardness of robust learning of an unknown binary product distribution here. In order to make the distributions in our family far from the reference distribution U_M in total variation distance, we need higher dimension m, M compared with the hardness result for robust hypothesis testing.

Theorem 51 (SQ Lower Bound for Robust Learning of a Binary Product Distribution) *Fix $0 < c < 1$ and k to be a sufficiently large integer. Let $m, M \in \mathbb{Z}_+$ with $M = 3m^{5/4}$. Let $0 < \epsilon < 1/2$ and δ be a sufficiently small multiple of $\epsilon\sqrt{\log(1/\epsilon)}/k^2$. Let $\tau = \Theta(M^{-(k+1)/5}\delta)$. Assume that $m > \max\left(C'/\epsilon, \frac{k^2}{\log(1/\delta)}\right)$ for some sufficiently large constant $C' > 0$. Then any* SQ *algorithm which is given access to a distribution $\mathbf{P}$ over $\{0,1\}^M$ which satisfies $d_{\mathrm{TV}}\left(\mathbf{P}, U_M^{S,\frac{\delta}{\sqrt{m}}}\right) \leq \epsilon$ for some unknown subset $S \subseteq [M]$ with $|S| = m$, outputs a hypothesis $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{P}) \leq O(\delta)$ with probability at least $2/3$ must either make queries of accuracy better than $\sqrt{2\tau}$ or must make at least $2^{\Omega(M^{2/5})}M^{-(k+1)/5}$ statistical queries.*

Proof We need to show that for any subset $S \subseteq [M]$ with $|S| = m$, $\mathbf{P}_S^A$ is far from U_M in total variation distance. In particular, by Lemma 23, we have that

$$d_{\mathrm{TV}}(U_M, \mathbf{P}_S^A) \geq d_{\mathrm{TV}}\left(U_M, U_M^{S,\frac{\delta}{\sqrt{m}}}\right) - d_{\mathrm{TV}}\left(\mathbf{P}_S^A, U_M^{S,\frac{\delta}{\sqrt{m}}}\right) \geq \Omega(\delta) - O(\epsilon) = \Omega(\delta).$$

In addition, by our choice of m, we have that $\sqrt{2\tau} \leq O(\delta)$. Therefore, we have that $d_{\mathrm{TV}}\left(U_M, \mathbf{P}_S^A\right) \geq 2\sqrt{2\tau} + \Omega(\delta)$. Applying Claim 48 and Theorem 24 yields Theorem 51. ■

Theorem 52 (SQ Lower Bound for Robust Mean Estimation of a Binary Product Distribution) *Fix $0 < c < 1$ and k to be a sufficiently large integer. Let $m, M \in \mathbb{Z}_+$ with $M = 3m^{5/4}$. Let $0 < \epsilon < 1/2$ and δ be a sufficiently small multiple of $\epsilon\sqrt{\log(1/\epsilon)}/k^2$. Let $\tau = \Theta(M^{-(k+1)/5}\delta)$. Assume that $m > \max\left(C'/\epsilon, \frac{k^2}{\log(1/\delta)}\right)$ for some sufficiently large constant $C' > 0$. Then any* SQ *algorithm which is given access to a distribution $\mathbf{P}$ over $\{0,1\}^M$ which satisfies $d_{\mathrm{TV}}\left(\mathbf{P}, U_M^{S,\frac{\delta}{\sqrt{m}}}\right) \leq \epsilon$ for some unknown subset $S \subseteq [M]$ with $|S| = m$, outputs an estimate $\widehat{\mu}$ with $\left\|\widehat{\mu} - \mu_M^{S,\frac{\delta}{\sqrt{m}}}\right\|_2 \leq O(\delta)$ with probability at least $2/3$ must either make queries of accuracy better than $\sqrt{2\tau}$ or must make at least $2^{\Omega(M^{2/5})}M^{-(k+1)/5}$ statistical queries.*

Proof Assume there is an algorithm that outputs an estimate $\widehat{\mu}$ such that $\left\|\widehat{\mu} - \mu_M^{S,\frac{\delta}{\sqrt{m}}}\right\|_2 \leq O(\delta)$ for some unknown subset $S \subseteq [M]$ with $|S| = m$. Let $\mathbf{Q}$ be the corresponding binary product distribution with mean vector $\widehat{\mu}$. Note that by our construction, both $\mathbf{Q}$ and $U_M^{S,\frac{\delta}{\sqrt{m}}}$ are c'-balanced binary product distributions for some universal constant $c' > 0$. Therefore, by Lemma 50, we have that $d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{P}_S^A) \leq d_{\mathrm{TV}}\left(\mathbf{Q}, U_M^{S,\frac{\delta}{\sqrt{m}}}\right) + d_{\mathrm{TV}}\left(U_M^{S,\frac{\delta}{\sqrt{m}}}, \mathbf{P}_S^A\right) \leq O\left(\left\|\widehat{\mu} - \mu_M^{S,\frac{\delta}{\sqrt{m}}}\right\|_2\right) + O(\epsilon) \leq O(\delta)$. Applying Theorem 51 yields the result. ■

B.1. Proof of Proposition 25

In this section, we prove Proposition 25. We first introduce the following notation which will be used throughout this section. For some fixed positive integer n and $x \in [n] \cup \{0\}$, we consider the function $F_{n,x}(\delta) = \binom{n}{x}(1/2+\delta)^x(1/2-\delta)^{n-x}$, $-1/2 < \delta < 1/2$. The first and second derivatives of $F_{n,x}(\delta)$ are given by the following fact:

Fact 53 *For any positive integer n and $x \in [n] \cup \{0\}$, we have that*

$$F'_{n,x}(\delta) = \binom{n}{x}(1/2+\delta)^{x-1}(1/2-\delta)^{n-x-1}(x-(1/2+\delta)n) = \frac{F_{n,x}(\delta)(x-(1/2+\delta)n)}{1/4-\delta^2},$$

$$F''_{n,x}(\delta) = \binom{n}{x}(1/2+\delta)^{x-2}(1/2-\delta)^{n-x-2}(x^2-((2\delta+1)n-2\delta)x+(1/2+\delta)^2(n^2-n))$$

$$= \frac{F_{n,x}(\delta)\left((x-(\delta+1/2)n)^2+2\delta(x-(\delta+1/2)n)+n(\delta^2-1/4)\right)}{(1/4-\delta^2)^2}.$$

We now pick $C = \Theta(\sqrt{(\log(1/\delta)/m)})$, where the hidden constant is sufficiently small. Consider the interval $I_{C,m} = [(1/2-C)m, (1/2+C)m-1]$. Without loss of generality, we assume that the two endpoints of $I_{C,m}$ are integers. We define the one-dimensional distribution A to be:

- For $x \notin I_{C,m}$, we define $A(x) = \mathrm{Bin}(m, 1/2+\delta/\sqrt{m})(x)$.
- For $x \in I_{C,m}$, we define $A(x) = \mathrm{Bin}(m, 1/2+\delta/\sqrt{m})(x) + \int_x^{x+1} p(t)dt$, where p is a polynomial of degree at most k satisfying

$$\sum_{x \in \mathbb{Z} \cap I_{C,m}} x^i \int_x^{x+1} p(t)dt = \sum_{x=0}^{m} (\mathrm{Bin}(m,1/2)(x) - \mathrm{Bin}(m,1/2+\delta/\sqrt{m})(x))x^i, \tag{4}$$

  for $0 \le i \le k$.

Applying Theorem 32 with the family of functions $\{F_{m,x}(\delta)\}_{x \in [m] \cup \{0\}}$, we know that there is a unique polynomial p of degree at most k satisfying the above properties. Then we need to show that with sufficiently large m (depending on δ), both the L_1 and L_∞ norms of p on $[(1/2-C)m, (1/2+C)m]$ are sufficiently small in order to make $A(x)$ non-negative and close to $\mathrm{Bin}(m, 1/2+\delta/\sqrt{m})$. The main technical result of this section is the following lemma, which provides upper bounds on the L_1 and L_∞ norms of p on the interval $[(1/2-C)m, (1/2+C)m]$.

Lemma 54 *Let $k, m \in \mathbb{Z}_+$. Suppose $1 \le k^2 \le C_0 C^2 m$ for some universal constant $C_0 > 0$ sufficiently small and $m \ge C_1(\log(1/\delta))^3$ for some universal constant $C_1 > 0$ sufficiently large. Then $\int_{(1/2-C)m}^{(1/2+C)m} |p(t)|dt \le O\left(\frac{\delta k^2}{C\sqrt{m}}\right)$ and $|p(t^*)| \le O\left(\frac{\delta k^{5/2}}{C^2 m^{3/2}}\right)$, where $t^* = \arg\max_{t:|t-m/2| \le Cm} |p(t)|$.*

Before we prove Lemma 54, we first use it to prove our main Proposition 25. The following claim gives the upper bound of the ratio between the mass of $\mathrm{Bin}(m, 1/2)$ and $\mathrm{Bin}(m, 1/2 + \delta/\sqrt{m})$.

Claim 55 *Let $m \in \mathbb{Z}_+$ and $x \in [m] \cup \{0\}$. For any $\delta > 0$, we have that*

$$\frac{F_{m,x}(\delta/\sqrt{m})}{F_{m,x}(0)} \le \exp\left(\frac{4\delta x}{\sqrt{m}} - 2\delta\sqrt{m}\right).$$

Proof By the fact $1 + x \le e^x, \forall x \in \mathbb{R}$, we have that

$$\frac{F_{m,x}(\delta/\sqrt{m})}{F_{m,x}(0)} = \left(1 + \frac{2\delta}{\sqrt{m}}\right)^x \left(1 - \frac{2\delta}{\sqrt{m}}\right)^{m-x} \le \exp\left(\frac{4\delta x}{\sqrt{m}} - 2\delta\sqrt{m}\right) .$$

■

We now bound from above the desired χ^2-divergence:

Lemma 56 *We have that*

$$\chi^2(A, \mathrm{Bin}(m, 1/2)) \le O\left(\delta^2 + \frac{\delta k^2 \exp(4\delta C\sqrt{m})}{C\sqrt{m}} + \left(\frac{\delta k^2}{C\sqrt{m}}\right) \cdot \max_{x\in\mathbb{Z}\cap I_{C,m}} \frac{\int_x^{x+1} |p(t)|dt}{\mathrm{Bin}(m, 1/2)(x)}\right) .$$

Proof Recalling $F_{m,x}(\delta) = \binom{m}{x}(1/2+\delta)^x(1/2-\delta)^{m-x} = \mathrm{Bin}(m, 1/2+\delta)(x)$, $-1/2 < \delta < 1/2$, we have the following:

$$1 + \chi^2(A, \mathrm{Bin}(m, 1/2)) = \sum_{x=0}^{m} \frac{A(x)^2}{F_{m,x}(0)} = \sum_{x=0}^{m} \frac{\left(F_{m,x}(\delta/\sqrt{m}) + \mathbb{I}[x \in I_{C,m}] \int_x^{x+1} p(t)dt\right)^2}{F_{m,x}(0)}$$

$$= \sum_{x=0}^{m} \frac{F^2_{m,x}(\delta/\sqrt{m})}{F_{m,x}(0)} + 2 \sum_{x\in\mathbb{Z}\cap I_{C,m}} \frac{F_{m,x}(\delta/\sqrt{m}) \int_x^{x+1} p(t)dt}{F_{m,x}(0)} + \sum_{x\in\mathbb{Z}\cap I_{C,m}} \frac{\left(\int_x^{x+1} p(t)dt\right)^2}{F_{m,x}(0)} .$$

For the first term, we have that

$$\sum_{x=0}^{m} \frac{F^2_{m,x}(\delta/\sqrt{m})}{F_{m,x}(0)} = 2^{-m} \sum_{x=0}^{m} \binom{m}{x} \left(1 + \frac{2\delta}{\sqrt{m}}\right)^{2x} \left(1 - \frac{2\delta}{\sqrt{m}}\right)^{2(m-x)}$$

$$= 2^{-m} \left(\left(1 + \frac{2\delta}{\sqrt{m}}\right)^2 + \left(1 - \frac{2\delta}{\sqrt{m}}\right)^2\right)^m = \left(1 + \frac{4\delta^2}{m}\right)^m \le 1 + O(\delta^2) .$$

For the second term, by Claim 55, we have that

$$\begin{aligned}\sum_{x\in\mathbb{Z}\cap I_{C,m}} \frac{F_{m,x}(\delta/\sqrt{m}) \int_x^{x+1} p(t)dt}{F_{m,x}(0)} &\le \sum_{x\in\mathbb{Z}\cap I_{C,m}} \exp\left(\frac{4\delta x}{\sqrt{m}} - 2\delta\sqrt{m}\right) \left|\int_x^{x+1} p(t)dt\right| \\ &\le \exp\left(4\delta C\sqrt{m}\right) \sum_{x\in\mathbb{Z}\cap I_{C,m}} \left|\int_x^{x+1} p(t)dt\right| \\ &\le \exp(4\delta C\sqrt{m}) \int_{(1/2-C)m}^{(1/2+C)m} |p(t)|dt \\ &\le O\left(\frac{\delta k^2 \exp(4\delta C\sqrt{m})}{C\sqrt{m}}\right) ,\end{aligned}$$

where the last inequality follows from Lemma 54. Finally for the third term, we have that

$$\sum_{x\in\mathbb{Z}\cap I_{C,m}} \frac{\left(\int_x^{x+1} p(t)dt\right)^2}{F_{m,x}(0)} \leq \int_{(1/2-C)m}^{(1/2+C)m} |p(x)|dx \cdot \max_{x\in\mathbb{Z}\cap I_{C,m}} \frac{\int_x^{x+1}|p(t)|dt}{F_{m,x}(0)}$$
$$\leq O\left(\frac{\delta k^2}{C\sqrt{m}}\right)\cdot \max_{x\in\mathbb{Z}\cap I_{C,m}} \frac{\int_x^{x+1}|p(t)|dt}{F_{m,x}(0)},$$

where the last inequality follows from Lemma 54. Combining the above results together completes the proof. ■

We are now ready to prove Proposition 25. We need to pick C appropriately and check the bounds on k needed for $A(x)$ to satisfy the necessary properties.

Proof [Proof of Proposition 25] Let $C = \Theta(\sqrt{\log(1/\delta)/m})$ with the hidden constant sufficiently small. If $k^2 \geq C\sqrt{m}$, we pick $A = \mathrm{Bin}(m, 1/2)$ and obtain $d_{\mathrm{TV}}(A, \mathrm{Bin}(m, 1/2+\delta/\sqrt{m})) \leq O(\delta) \leq O\left(\frac{k^2\delta}{\sqrt{\log(1/\delta)}}\right)$. Thus, we assume that $k^2 \leq C\sqrt{m}$. In this way, to apply Lemma 54, we need $k^2 \leq C_0 C^2 m$ for some universal constant C_0 sufficiently small, which will be satisfied as long as $\delta \leq \exp(-1/C_0^2)$.

We first show that $A(x)$ is indeed a distribution over $[m]\cup\{0\}$. By definition, $A(x)$ is nonnegative outside the interval $I_{C,m}$. For $x \in \mathbb{Z}\cap I_{C,m}$, we apply Lemma 54 to obtain

$$\begin{aligned}
A(x) &= F_{m,x}(\delta/\sqrt{m}) + \int_x^{x+1} p(t)dt \geq F_{m,x}(\delta/\sqrt{m}) - |p(t^*)| \\
&= 2^{-m}\binom{m}{x}\left(1+\frac{2\delta}{\sqrt{m}}\right)^x\left(1-\frac{2\delta}{\sqrt{m}}\right)^{m-x} - |p(t^*)| \\
&\geq 2^{-m}\binom{m}{(1/2-C)m}\left(1+\frac{2\delta}{\sqrt{m}}\right)^{(\frac{1}{2}-C)m}\left(1-\frac{2\delta}{\sqrt{m}}\right)^{(\frac{1}{2}+C)m} - O\left(\frac{\delta k^{5/2}}{C^2 m^{3/2}}\right) \\
&= 2^{-m}\binom{m}{(1/2-C)m}\left(1-\frac{4\delta^2}{m}\right)^{(\frac{1}{2}-C)m}\left(1-\frac{2\delta}{\sqrt{m}}\right)^{2Cm} - O\left(\frac{\delta k^{5/2}}{C^2 m^{3/2}}\right).
\end{aligned}$$

Let $H(x) = -x\log x - (1-x)\log(1-x)$ denote the binary entropy function. Now applying Fact 34 and the fact $e^{-2x} \leq 1-x, \forall x \in \left[0, \frac{\ln 2}{2}\right]$ yields

$$\begin{aligned}
A(x) &\geq \frac{2^{m(H(1/2+C)-1)}}{\sqrt{(2-8C^2)m}}\cdot\exp(-8\delta^2(1/2-C))\cdot\exp(-8\delta C\sqrt{m}) - O\left(\frac{\delta k^{5/2}}{C^2 m^{3/2}}\right) \\
&\geq \frac{2^{m(H(1/2+C)-1)}}{\sqrt{(2-8C^2)m}}\cdot\exp\left(-4\delta^2 - 8\delta C\sqrt{m}\right) - O\left(\frac{\delta k^{5/2}}{C^2 m^{3/2}}\right) \\
&\geq \frac{\exp\left(-O(C^2 m) - 8\delta C\sqrt{m}\right)}{\sqrt{m}} - O\left(\frac{\delta k^{5/2}}{C^2 m^{3/2}}\right) \\
&\geq \frac{\exp(-O((C\sqrt{m}+\delta)^2))}{\sqrt{m}} - O\left(\frac{\delta k^{5/2}}{C^2 m^{3/2}}\right),
\end{aligned}$$

where the third inequality follows from the Taylor expansion of $H(1/2+C)-H(1/2)$ up to second order terms. Note that $k^2 \le C\sqrt{m}$, where $C = \Theta(\sqrt{\log(1/\delta)/m})$ for some sufficiently small hidden constant in Θ, we have that $\frac{\delta k^{5/2}}{C^2 m} \le O(\delta(\log(1/\delta))^{-3/8})$ and $\exp(-O((C\sqrt{m}+\delta)^2)) \ge \exp\left(-\left(\sqrt{\log(1/\delta)}/2+\delta\right)^2\right) \ge \delta$. Therefore, we have that

$$A(x) \ge \frac{\exp(-O((C\sqrt{m}+\delta)^2))}{\sqrt{m}} - O\left(\frac{\delta k^{5/2}}{C^2 m^{3/2}}\right) \ge 0, \forall x \in \mathbb{Z} \cap I_{C,m}.$$

In addition, by equation (4), we know that

$$\begin{aligned}\sum_{x=0}^{m} A(x) &= \sum_{x=0}^{m}\left(F_{m,x}(\delta/\sqrt{m}) + \mathbb{I}[x \in I_{C,m}]\int_x^{x+1} p(t)dt\right)\\ &= \sum_{x=0}^{m} F_{m,x}(\delta/\sqrt{m}) + \int_{(1/2-C)m}^{(1/2+C)m} p(t)dt = 1,\end{aligned}$$

which implies that the distribution A is well-defined. Furthermore, by Equation (4), we can show that A matches the first k moments of $\mathrm{Bin}(m,1/2)$ as follows:

$$\begin{aligned}\mathbf{E}_{X\sim A}[X^i] &= \sum_{x=0}^{m} A(x)x^i = \sum_{x=0}^{m}\left(F_{m,x}(\delta/\sqrt{m}) + \mathbb{I}[x \in I_{C,m}]\int_x^{x+1} p(t)dt\right)x^i\\ &= \sum_{x=0}^{m} F_{m,x}(\delta/\sqrt{m})x^i + \sum_{x\in\mathbb{Z}\cap I_{C,m}} x^i \int_x^{x+1} p(t)dt\\ &= \sum_{x=0}^{m} F_{m,x}(0)x^i = \mathbf{E}_{X\sim\mathrm{Bin}(m,1/2)}[X^i].\end{aligned}$$

From previous calculation, we have that $A(x) \ge F_{m,x}(\delta/\sqrt{m}) - |p(t^*)| \ge 0, \forall x \in \mathbb{Z}\cap I_{C,m}$, which implies that for every $x \in \mathbb{Z}\cap I_{C,m}$,

$$|p(t^*)| \le F_{m,x}(\delta/\sqrt{m}) \le \exp\left(\frac{4\delta x}{\sqrt{m}} - 2\delta\sqrt{m}\right)F_{m,x}(0) \le \exp\left(4\delta C\sqrt{m}\right)F_{m,x}(0),$$

where the second inequality follows from Claim 55. Therefore, by Lemma 56, we have that

$$\begin{aligned}\chi^2(A,\mathrm{Bin}(m,1/2)) &\le O\left(\delta^2 + \frac{\delta k^2\exp(4\delta C\sqrt{m})}{C\sqrt{m}} + \left(\frac{\delta k^2}{C\sqrt{m}}\right)\cdot \max_{x\in\mathbb{Z}\cap I_{C,m}} \frac{\int_x^{x+1}|p(t)|dt}{\mathrm{Bin}(m,1/2)(x)}\right)\\ &\le O\left(\delta^2 + \delta\left(\exp(4\delta C\sqrt{m}) + \frac{|p(t^*)|}{F_{m,x}(0)}\right)\right) \le O\left(\delta^2 + 2\delta\exp\left(4\delta C\sqrt{m}\right)\right)\\ &\le O\left(\delta^2 + \delta\left(1 + O(\delta\sqrt{\log(1/\delta)})\right)\right) = O(\delta),\end{aligned}$$

where we apply the fact $e^x \le 1+2x, \forall x \in [0,\ln 2]$.

To bound the total variation distance $d_{\mathrm{TV}}(A, \mathrm{Bin}(1/2+\delta/\sqrt{m}))$, we apply Lemma 54 to obtain

$$d_{\mathrm{TV}}(A, \mathrm{Bin}(1/2+\delta/\sqrt{m})) = \sum_{x \in I_{C,m}} \left| \int_x^{x+1} p(t)dt \right| \leq \int_{(1/2-C)m}^{(1/2+C)m} |p(t)|dt$$
$$\leq O\left(\frac{\delta k^2}{C\sqrt{m}}\right) = O\left(\frac{\delta k^2}{\sqrt{\log(1/\delta)}}\right).$$

This completes the proof of Proposition 25. ■

Proof of Lemma 54 By Theorem 32, we have that

$$|a_i| \leq \left(\frac{2i+1}{2Cm}\right)\left(\beta_i + O\left(\frac{i^2}{Cm}\right)\int_{(1/2-C)m}^{(1/2+C)m} |p(t)|dt\right),$$

for all $1 \leq i \leq k$, where $\beta_i = \left|\sum_{x=0}^m (F_{m,x}(0) - F_{m,x}(\delta/\sqrt{m}))P_i\left(\frac{x-m/2}{Cm}\right)\right|$. To get an upper bound for the L_1 and L_∞ norms of the polynomial p over $I_{C,m}$, we only need to upper bound the quantity β_i.

Lemma 57 *If $k^2 \leq C_0 C^2 m$ for some universal constant $C_0 > 0$ sufficiently small, then $\beta_i \leq O\left(\frac{\delta}{C}\sqrt{\frac{i}{m}}\right), \forall 1 \leq i \leq k$.*

We assume $k^2 \leq C_0 C^2 m$ for some universal constant $C_0 > 0$ sufficiently small. First, we apply Taylor's theorem to expand $F_{m,x}(\delta/\sqrt{m}) - F_{m,x}(0)$ up to second order terms:

$$\begin{aligned}\beta_i &= \left|\sum_{x=0}^m \left(F_{m,x}(0) - F_{m,x}(\delta/\sqrt{m})\right) P_i\left(\frac{x-m/2}{Cm}\right)\right| \\ &= \left|\sum_{x=0}^m \left(\frac{\delta F'_{m,x}(0)}{\sqrt{m}} + \frac{F''_{m,x}(\delta_x)\delta^2}{2m}\right) P_i\left(\frac{x-m/2}{Cm}\right)\right| \\ &\leq \underbrace{\left|\sum_{x=0}^m \left(\frac{\delta F'_{m,x}(0)}{\sqrt{m}}\right) P_i\left(\frac{x-m/2}{Cm}\right)\right|}_{\beta'_i} + \underbrace{\left|\sum_{x=0}^m \left(\frac{F''_{m,x}(\delta_x)\delta^2}{2m}\right) P_i\left(\frac{x-m/2}{Cm}\right)\right|}_{\beta''_i},\end{aligned} \tag{5}$$

where for any $x \in [m] \cup \{0\}$, $\delta_x = \widetilde{\delta_x}/\sqrt{m}$ for some $\widetilde{\delta_x} \in [0, \delta]$.

Hence, in order to bound β_i, it suffices to bound the terms $\beta'_i := \frac{\delta}{\sqrt{m}}\left|\sum_{x=0}^m F'_{m,x}(0)P_i\left(\frac{x-m/2}{Cm}\right)\right|$ and $\beta''_i := \frac{\delta^2}{2m}\left|\sum_{x=0}^m F''_{m,x}(\delta_x)P_i\left(\frac{x-m/2}{Cm}\right)\right|$. This is done in the following lemmas.

Lemma 58 *We have that $\beta'_i \leq O\left(\max\left(\frac{\delta}{C}\sqrt{\frac{i}{m}}, \frac{\delta i^{3/2}}{C^2 m}\right)\right), \forall 1 \leq i \leq k.$*

Proof If i is odd, we can rewrite Fact 4 (v) in ascending order of terms, by using change of variables to obtain

$$\begin{aligned}\left|P_i\left(\frac{x-m/2}{Cm}\right)\right| &\le 2^{-i}\sum_{j=1}^{(i+1)/2}\binom{i}{(i-1)/2+j}\binom{i+2j-1}{2j-1}\left|\frac{x-m/2}{Cm}\right|^{2j-1}\\ &\le O\left(\frac{1}{\sqrt{i}}\right)\sum_{j=1}^{(i+1)/2}\frac{(i+2j-1)^{2j-1}}{(2j-1)!}\left|\frac{x-m/2}{Cm}\right|^{2j-1},\end{aligned}$$

where the second inequality follows from Fact 33. Note that $\|X-m/2\|_{\psi_2}\le O(\sqrt{m})$ for $X\sim \mathrm{Bin}(m,1/2)$, applying Fact 38 and Fact 53 yields

$$\begin{aligned}\beta_i' &= \frac{\delta}{\sqrt{m}}\left|\sum_{x=0}^{m}F'_{m,x}(0)P_i\left(\frac{x-m/2}{Cm}\right)\right| = \frac{\delta 2^{2-m}}{\sqrt{m}}\sum_{x=0}^{m}\left|\binom{m}{x}(x-m/2)P_i\left(\frac{x-m/2}{Cm}\right)\right|\\ &\le \frac{\delta 2^{2-m}}{\sqrt{m}}\sum_{j=1}^{(i+1)/2}O\left(\frac{1}{\sqrt{i}}\right)\left(\frac{(i+2j-1)^{2j-1}}{(2j-1)!(Cm)^{2j-1}}\right)\sum_{x=0}^{m}\binom{m}{x}(x-m/2)^{2j}\\ &= \frac{4\delta}{\sqrt{m}}\sum_{j=1}^{(i+1)/2}O\left(\frac{1}{\sqrt{i}}\right)\left(\frac{(i+2j-1)^{2j-1}\mathbf{E}_{X\sim\mathrm{Bin}(m,1/2)}[(X-m/2)^{2j}]}{(2j-1)!(Cm)^{2j-1}}\right)\\ &\le O\left(\frac{\delta}{\sqrt{mi}}\right)\sum_{j=1}^{(i+1)/2}\frac{(i+2j-1)^{2j-1}(2j!)(O(m))^j}{(2j-1)!(Cm)^{2j-1}}\\ &\le O\left(\frac{\delta}{\sqrt{i}}\right)\sum_{j=1}^{\infty}\left(O\left(\frac{i}{C\sqrt{m}}\right)\right)^{2j-1}\le O\left(\frac{\delta}{C}\sqrt{\frac{i}{m}}\right).\end{aligned}$$

If i is even, applying Fact 53 and Fact 4 (v) by using change of variables yields

$$\begin{aligned}
\beta_i' &= \frac{\delta}{\sqrt{m}} \left| \sum_{x=0}^{m} F'_{m,x}(0) P_i\left(\frac{x-m/2}{Cm}\right) \right| = \frac{\delta 2^{2-m}}{\sqrt{m}} \left| \sum_{x=0}^{m} \binom{m}{x} (x-m/2) P_i\left(\frac{x-m/2}{Cm}\right) \right| \\
&= \frac{\delta 2^{2-m}}{\sqrt{m}} \left| \sum_{x=0}^{m} \binom{m}{x} (x-m/2) 2^{-i} \sum_{j=0}^{i/2} (-1)^j \binom{i}{j} \binom{2i-2j}{i} \left(\frac{x-m/2}{Cm}\right)^{i-2j} \right| \\
&= \frac{\delta 2^{2-m}}{\sqrt{m}} \left| \sum_{x=0}^{m} \binom{m}{x} (x-m/2) 2^{-i} \sum_{j=0}^{i/2-1} (-1)^j \binom{i}{j} \binom{2i-2j}{i} \left(\frac{x-m/2}{Cm}\right)^{i-2j} \right| \\
&\leq \frac{\delta 2^{2-m}}{\sqrt{m}} \sum_{x=0}^{m} \binom{m}{x} |x-m/2| 2^{-i} \sum_{j=1}^{i/2} \binom{i}{i/2+j} \binom{i+2j}{2j} \left|\frac{x-m/2}{Cm}\right|^{2j} \\
&\leq \frac{\delta 2^{2-m}}{\sqrt{m}} \sum_{j=1}^{i/2} \sum_{x=0}^{m} \binom{m}{x} |x-m/2|^{2j+1} O\left(\frac{1}{\sqrt{i}}\right) \left(\frac{(i+2j)^{2j}}{(2j)!(Cm)^{2j}}\right) \\
&= \frac{4\delta}{\sqrt{m}} \sum_{j=1}^{i/2} O\left(\frac{1}{\sqrt{i}}\right) \left(\frac{(i+2j)^{2j} \mathbf{E}_{X\sim \mathrm{Bin}(m,1/2)}[|X-m/2|^{2j+1}]}{(2j)!(Cm)^{2j}}\right) \\
&\leq O\left(\frac{\delta}{\sqrt{mi}}\right) \sum_{j=1}^{i/2} \frac{(i+2j)^{2j}(2j+1)(O(m))^{j+1/2} j!}{(2j)!(Cm)^{2j}} \\
&\leq O\left(\frac{\delta}{\sqrt{i}}\right) \sum_{j=1}^{\infty} \left(O\left(\frac{i}{C\sqrt{m}}\right)\right)^{2j} \leq O\left(\frac{\delta i^{3/2}}{C^2 m}\right),
\end{aligned}$$

where the second inequality follows from Fact 33 and the third inequality follows from Fact 38 and the fact that $\|X-m/2\|_{\psi_2} \leq O(\sqrt{m})$ for $X \sim \mathrm{Bin}(m,1/2)$. ■

For the quantity β_i'', we have the following lemma.

Lemma 59 *We have that $\beta_i'' = O(\delta^2), \forall 1 \leq i \leq k$.*

Proof By Fact 53, we have that

$$\begin{aligned}
\beta_i'' &= \left| \sum_{x=0}^{m} F''_{m,x}(\delta_x) P_i\left(\frac{x-m/2}{Cm}\right) \right| \\
&= \frac{1}{(1/4-\delta_x^2)^2} \left| \sum_{x=0}^{m} F_{m,x}(\delta_x) \left((x-(\delta_x+1/2)m)^2 + 2\delta_x(x-(\delta_x+1/2)m) + m(\delta_x^2-1/4)\right) P_i\left(\frac{x-m/2}{Cm}\right) \right| \\
&= O\left(\left| \sum_{x=0}^{m} F_{m,x}(\delta_x) \left((x-(\delta_x+1/2)m)^2 + 2\delta_x(x-(\delta_x+1/2)m) + m(\delta_x^2-1/4)\right) P_i\left(\frac{x-m/2}{Cm}\right) \right| \right).
\end{aligned}$$

We separate the above sum into $x \in \mathbb{Z} \cap I_{C,m}$ and $x \in [m] \cup \{0\} \setminus I_{C,m}$. We are able to use Fact 4 (iii) to bound the sum for $x \in \mathbb{Z} \cap I_{C,m}$, as follows:

$$\begin{aligned}
&\left| \sum_{x \in \mathbb{Z} \cap I_{C,m}} F_{m,x}(\delta_x) \left((x - (\delta_x + 1/2)m)^2 + 2\delta_x (x - (\delta_x + 1/2)m) + m(\delta_x^2 - 1/4) \right) P_i \left(\frac{x - m/2}{Cm} \right) \right| \\
&\leq \sum_{x=0}^{m} F_{m,x}(\delta_x) \left| (x - (\delta_x + 1/2)m)^2 - \mathbf{Var}_{X \sim \mathrm{Bin}(m,1/2+\delta_x)}[X] + 2\delta_x (x - (\delta_x + 1/2)m) \right| \\
&\leq 2\mathbf{Var}_{X \sim \mathrm{Bin}(m,1/2+\delta_x)}[X] + 2\delta_x \mathbf{E}_{X \sim \mathrm{Bin}(m,1/2+\delta_x)} \left[\left| X - \mathbf{E}_{X \sim \mathrm{Bin}(m,1/2+\delta_x)}[X] \right| \right] \\
&\leq O\left(\mathbf{Var}_{X \sim \mathrm{Bin}(m,1/2+\delta_x)}[X] + \delta_x \sqrt{\mathbf{Var}_{X \sim \mathrm{Bin}(m,1/2+\delta_x)}[X]} \right) \\
&\leq O\left(m + \delta \right) \leq O(m).
\end{aligned}$$

Now we bound the sum over $x \in \overline{I_{C,m}}$, where $\overline{I_{C,m}} = [m] \cup \{0\} \setminus I_{C,m}$. Note that for $X \sim \mathrm{Bin}(m, 1/2 + \delta_x)$, by Fact 40, we have that

$$\begin{aligned}
\left\| \frac{4(X - m/2)}{Cm} \right\|_{\psi_2} &= O\left(\left\| \frac{X - (1/2 + \delta_x)m}{Cm} \right\|_{\psi_2} + \|\delta_x / C\|_{\psi_2} \right) = O\left(\left\| \frac{X - (1/2 + \delta_x)m}{Cm} \right\|_{\psi_2} + \delta_x / C \right) \\
&\leq O\left(\frac{1}{C\sqrt{m}} + \frac{\delta}{C\sqrt{m}} \right) = O\left(\frac{1}{C\sqrt{m}} \right).
\end{aligned}$$

Therefore, applying Fact 4 (vi) yields

$$\begin{aligned}
&\left|\sum_{x\in\overline{I_{C,m}}} F_{m,x}(\delta_x)(x-(\delta_x+1/2)m)^2+2\delta_x(x-(\delta_x+1/2)m)+m(\delta_x^2-1/4)P_i\left(\frac{x-m/2}{Cm}\right)\right|\\
&\le \sum_{x=0}^{m} F_{m,x}(\delta_x)\cdot\left|(x-(\delta_x+1/2)m)^2+2\delta_x(x-(\delta_x+1/2)m)+m(\delta_x^2-1/4)\right|\cdot\left|\frac{4(x-m/2)}{Cm}\right|^i\\
&\le O(m)\cdot \mathop{\mathbf{E}}_{X\sim\mathrm{Bin}(m,1/2+\delta_x)}\left[\left|\frac{4(X-m/2)}{Cm}\right|^i\right]+\mathop{\mathbf{E}}_{X\sim\mathrm{Bin}(m,1/2+\delta_x)}\left[(X-(\delta_x+1/2)m)^2\cdot\left|\frac{4(X-m/2)}{Cm}\right|^i\right]\\
&\quad+2\delta_x\mathop{\mathbf{E}}_{X\sim\mathrm{Bin}(m,1/2+\delta_x)}\left[|X-(\delta_x+1/2)m|\cdot\left|\frac{4(X-m/2)}{Cm}\right|^i\right]\\
&\le O(m)\mathop{\mathbf{E}}_{X\sim\mathrm{Bin}(m,1/2+\delta_x)}\left[\left|\frac{4(X-m/2)}{Cm}\right|^i\right]\\
&\quad+\sqrt{\mathop{\mathbf{E}}_{X\sim\mathrm{Bin}(m,1/2+\delta_x)}[(X-(\delta_x+1/2)m)^4]\cdot\mathop{\mathbf{E}}_{X\sim\mathrm{Bin}(m,1/2+\delta_x)}\left[\left|\frac{4(X-m/2)}{Cm}\right|^{2i}\right]}\\
&\quad+2\delta_x\sqrt{\mathop{\mathbf{E}}_{X\sim\mathrm{Bin}(m,1/2+\delta_x)}[(X-(\delta_x+1/2)m)^2]\cdot\mathop{\mathbf{E}}_{X\sim\mathrm{Bin}(m,1/2+\delta_x)}\left[\left|\frac{4(X-m/2)}{Cm}\right|^{2i}\right]}\\
&\le O(m)\cdot\left(O\left(\frac{1}{C}\sqrt{\frac{i}{m}}\right)\right)^i+\sqrt{O(m^2)\cdot\left(O\left(\frac{i}{C^2m}\right)\right)^i}+2\delta_x\sqrt{O(m)\cdot\left(O\left(\frac{i}{C^2m}\right)\right)^i}\\
&\le O(m),
\end{aligned}$$

where the third inequality follows from Cauchy-Schwarz and the fourth inequality follows from Fact 38. Combine the above results together, we have that

$$\beta_i''=\frac{\delta^2}{2m}\left|\sum_{x=0}^{m}F''_{m,x}(\delta_x)P_i\left(\frac{x-m/2}{Cm}\right)\right|\le\left(\frac{\delta^2}{2m}\right)\cdot O(m)=O(\delta^2).$$

■

Proof [Proof of Lemma 57] Since $k^2\le C^2m$, by Lemma 58, Lemma 59 and equation (5), we have that $\beta_i\le\beta_i'+\beta_i''\le O\left(\frac{\delta}{C}\sqrt{\frac{i}{m}}+\delta^2\right)=O\left(\frac{\delta}{C}\sqrt{\frac{i}{m}}\right)$. ■

We are now ready to prove Lemma 54.

Proof [Proof of Lemma 54] By Theorem 32, we have that

$$\begin{aligned}
|p(t^*)| &\le \sum_{i=1}^{k} |a_i| \le \sum_{i=1}^{k} \left(\frac{2i+1}{2Cm}\right)\beta_i + \sum_{i=1}^{k}\left(\frac{2i+1}{2Cm}\right) O\left(\frac{i^2}{Cm}\right)\int_{(1/2-C)m}^{(1/2+C)m} |p(t)|dt \\
&\le \sum_{i=1}^{k}\left(\frac{2i+1}{2Cm}\right)\beta_i + |p(t^*)| \sum_{i=1}^{k}\left(\frac{2i+1}{2}\right) O\left(\frac{i^2}{Cm}\right) \\
&\le \sum_{i=1}^{k}\left(\frac{2i+1}{2Cm}\right)\beta_i + O\left(\frac{k^4}{Cm}\right)|p(t^*)|,
\end{aligned}$$

where the first inequality follows from Fact 4 (iii). Similarly, by Theorem 32, we have that

$$\begin{aligned}
\int_{(1/2-C)m}^{(1/2+C)m} |p(t)|dt &\le \sum_{i=1}^{k} |a_i| \int_{(1/2-C)m}^{(1/2+C)m} \left| P_i\left(\frac{t-m/2}{Cm}\right)\right| dt \\
&\le Cm \sum_{i=1}^{k}\left(\frac{2i+1}{2Cm}\right)\left(\beta_i + O\left(\frac{i^2}{Cm}\right)\int_{(1/2-C)m}^{(1/2+C)m}|p(t)|dt\right)\int_{-1}^{1}|P_i(y)|dy \\
&\le \sum_{i=1}^{k} O(\sqrt{i})\beta_i + \sum_{i=1}^{k} O(\sqrt{i}) O\left(\frac{i^2}{Cm}\right)\int_{(1/2-C)m}^{(1/2+C)m}|p(t)|dt \\
&\le \sum_{i=1}^{k} O(\sqrt{i})\beta_i + O\left(\frac{k^{7/2}}{Cm}\right)\int_{(1/2-C)m}^{(1/2+C)m}|p(t)|dt,
\end{aligned}$$

where the third inequality follows from Fact 4 (vii). By our assumption on k, C, m, δ, we know that $\frac{k^{7/2}}{Cm} \le \frac{k^4}{Cm} \le 1/2$. Therefore, by Lemma 57, we have that

$$\begin{aligned}
|p(t^*)| &\le 2\sum_{i=1}^{k}\left(\frac{2i+1}{2Cm}\right)\beta_i \le \sum_{i=1}^{k}\left(\frac{2i+1}{2Cm}\right) O\left(\frac{\delta}{C}\sqrt{\frac{i}{m}}\right) \le O\left(\frac{\delta k^{5/2}}{C^2 m^{3/2}}\right), \\
\int_{(1/2-C)m}^{(1/2+C)m}|p(t)|dt &\le 2\sum_{i=1}^{k} O(\sqrt{i})\beta_i \le 2\sum_{i=1}^{k} O(\sqrt{i}) O\left(\frac{\delta}{C}\sqrt{\frac{i}{m}}\right) \le O\left(\frac{\delta k^2}{C\sqrt{m}}\right).
\end{aligned}$$

This completes the proof. ■

B.2. Proof of Lemma 23

Let $n = |S| \le m$. Define $f(\mathbf{x}) = \sum_{i \in S} x_i, \forall \mathbf{x} \in \{0,1\}^M$. For $\mathbf{X} \sim U_M$ and $\mathbf{Y} \sim U_M^{S,\frac{\delta}{\sqrt{m}}}$, by the data processing inequality, we have that

$$d_{\mathrm{TV}}\left(U_M, U_M^{S,\frac{\delta}{\sqrt{m}}}\right) \ge d_{\mathrm{TV}}(f(\mathbf{X}), f(\mathbf{Y})) = d_{\mathrm{TV}}(\mathrm{Bin}(n,1/2), \mathrm{Bin}(n, 1/2+\delta/\sqrt{m})).$$

Recalling that $F_{n,x}(\delta) = \binom{n}{x}(1/2+\delta)^x(1/2-\delta)^{n-x}$, we can write

$$d_{\mathrm{TV}}(\mathrm{Bin}(n,1/2), \mathrm{Bin}(n,1/2+\delta/\sqrt{m})) = \frac{1}{2}\sum_{x=0}^{n}|F_{n,x}(\delta/\sqrt{m}) - F_{n,x}(0)| = \frac{1}{2}\sum_{x=0}^{n}\left|\frac{\delta F'_{n,x}(0)}{\sqrt{m}} + \frac{F''_{n,x}(\delta_x)\delta^2}{2m}\right|,$$

where for any $x \in [m] \cup \{0\}$, $\delta_x = \widetilde{\delta_x}/\sqrt{m}$ for some $\widetilde{\delta_x} \in (0, \delta)$. Applying Fact 53 and Fact 35 yields

$$\begin{aligned}\sum_{x=0}^{n} |F'_{n,x}(0)| = 4\sum_{x=0}^{n} F_{n,x}(0)|x - n/2| &= 4\mathbf{E}_{X\sim \mathrm{Bin}(n,1/2)}[|X - n/2|] \\ &\geq \frac{4\mathbf{E}_{X\sim \mathrm{Bin}(n,1/2)}[(X-n/2)^2]^{3/2}}{\mathbf{E}_{X\sim \mathrm{Bin}(n,1/2)}[(X-n/2)^4]^{1/2}} \\ &= \frac{4(n/4)^{3/2}}{\sqrt{(n/4)(1+(3n-6)/4)}} \\ &= \Theta(\sqrt{n}).\end{aligned}$$

In addition, by Fact 53, we have that

$$\begin{aligned}\sum_{x=0}^{n} |F''_{n,x}(\delta_x)| &= \frac{1}{(1/4-\delta_x^2)^2}\sum_{x=0}^{n} |F_{n,x}(\delta_x)\left((x-(\delta_x+1/2)n)^2 + 2\delta_x(x-(\delta_x+1/2)n) + n(\delta_x^2 - 1/4)\right)| \\ &\leq \frac{1}{(1/4-\delta_x^2)^2}\left(2\mathbf{Var}_{X\sim \mathrm{Bin}\left(n,\frac{1}{2}+\delta_x\right)}[X] + 2\delta_x \mathbf{E}_{X\sim \mathrm{Bin}\left(n,\frac{1}{2}+\delta_x\right)}\left[\left|X - \mathbf{E}_{X\sim \mathrm{Bin}\left(n,\frac{1}{2}+\delta_x\right)}[X]\right|\right]\right) \\ &\leq \frac{1}{(1/4-\delta_x^2)^2}\left(2\mathbf{Var}_{X\sim \mathrm{Bin}\left(n,\frac{1}{2}+\delta_x\right)}[X] + 2\delta_x\sqrt{\mathbf{Var}_{X\sim \mathrm{Bin}\left(n,\frac{1}{2}+\delta_x\right)}[X]}\right) \\ &\leq O\left(n + \delta\sqrt{\frac{n}{m}}\right) \leq O(n).\end{aligned}$$

Therefore, we have that

$$\begin{aligned}d_{\mathrm{TV}}(\mathrm{Bin}(n,1/2), \mathrm{Bin}(n, 1/2+\delta/\sqrt{m})) &= \frac{1}{2}\sum_{x=0}^{n}\left|\frac{\delta F'_{n,x}(0)}{\sqrt{m}} + \frac{F_{n,x}(\delta_x)\delta^2}{2m}\right| \\ \geq \frac{\delta}{2\sqrt{m}}\sum_{x=0}^{n}|F'_{n,x}(0)| - \frac{\delta^2}{4m}\sum_{x=0}^{n}|F''_{n,x}(\delta_x)| &\geq \Omega\left(\delta\sqrt{\frac{n}{m}}\right) - O\left(\frac{\delta^2 n}{m}\right) = \Omega(\delta) - O(\delta^2) = \Omega(\delta).\end{aligned}$$

Appendix C. Omitted Statements and Proofs from Section 5

We provide the hardness result for robust learning of an unknown ferromagnetic high temperature Ising model here. In order to make the distributions in our family far from the reference distribution U_M in total variation distance, we need higher dimension m, M compared with the hardness result for robust hypothesis testing.

Theorem 60 (SQ Lower Bound for Robust Learning of an Unknown Ising Model) *Fix $0 < c < 1$ and k to be a sufficiently large integer. Let $m, M \in \mathbb{Z}_+$ with $M = 3m^{5/4}$. Let $0 < \epsilon < 1/2$ and δ be a sufficiently small multiple of $\epsilon\log(1/\epsilon)/k^3$. Let $\tau = \Theta(M^{-(k+1)/5}\delta)$. Assume that $m > \max\left(C'/\epsilon, \frac{k^2}{\log(1/\delta)}\right)$ for some sufficiently large constant $C' > 0$. Then any SQ algorithm which is given access to a distribution $\mathbf{P}$ over $\{0,1\}^M$ which satisfies $d_{\mathrm{TV}}\left(\mathbf{P}, Q_M^{S,\frac{\delta}{m}}\right) \leq \epsilon$ for some unknown subset $S \subseteq [M]$ with $|S| = m$, outputs a hypothesis $\mathbf{Q}$ with $d_{\mathrm{TV}}(\mathbf{Q}, \mathbf{P}) \leq O(\delta)$ with probability at least $2/3$ must either make queries of accuracy better than $\sqrt{2\tau}$ or must make at least $2^{\Omega(M^{2/5})}M^{-(k+1)/5}$ statistical queries.*

Proof We need to show that for any subset $S \subseteq [M]$ with $|S| = m$, $\mathbf{P}_S^A$ is far from U_M in total variation distance. In particular, by Lemma 28, we have that

$$d_{\mathrm{TV}}(U_M, \mathbf{P}_S^A) \geq d_{\mathrm{TV}}\left(U_M, Q_M^{S,\frac{\delta}{m}}\right) - d_{\mathrm{TV}}\left(\mathbf{P}_S^A, Q_M^{S,\frac{\delta}{m}}\right) \geq \Omega(\delta) - O(\epsilon) = \Omega(\delta) \ .$$

In addition, by our choice of m, we have that $\sqrt{2\tau} \leq O(\delta)$. Therefore, we have that $d_{\mathrm{TV}}\left(U_M, \mathbf{P}_S^A\right) \geq 2\sqrt{2\tau} + \Omega(\delta)$. Applying Claim 48 and Theorem 29 yields Theorem 60. ■

C.1. Proof of Proposition 31

In this section, we prove Proposition 31. We first introduce the following notations which will be used throughout this section. For some fixed positive integer n and $x \in [n] \cup \{0\}$, we consider the function $G_{n,x}(\delta) = \mathrm{IS}(n,\delta)(x), -1/n < \delta < 1/n$. By definition, we have that $\mathrm{IS}(n,\delta)(x) = \mathrm{IS}(n,\delta)(n-x)$ and $G_{n,x}(\delta) = G_{n,n-x}(\delta), x \in [n] \cup \{0\}, -1/n < \delta < 1/n$. In particular, $\mathrm{IS}(n,0)$ is exactly the binomial distribution $\mathrm{Bin}(n,1/2)$.

Claim 61 *Let $n \in \mathbb{Z}_+$ and $X \sim \mathrm{IS}(n,0)$. Then, $\mathbf{E}_{X\sim \mathrm{IS}(n,0)}[h(n,X)] = 0$.*

Proof By definition, we have that

$$\begin{aligned}
&\mathbf{E}_{X\sim \mathrm{IS}(n,0)}[h(n,X)] = \mathbf{E}_{X\sim \mathrm{Bin}(n,1/2)}\left[2X^2 - 2nX + n(n-1)/2\right] \\
&= 2\left(\mathbf{E}_{X\sim \mathrm{Bin}(n,1/2)}[X]^2 + \mathbf{Var}_{X\sim \mathrm{Bin}(n,1/2)}[X]\right) - 2n\mathbf{E}_{X\sim \mathrm{Bin}(n,1/2)}[X] + n(n-1)/2 \\
&= 2\left(n^2/4 + n/4\right) - n^2 + n(n-1)/2 = 0.
\end{aligned}$$

■

The first and second derivatives of $G_{n,x}(\delta)$ are given by the following claim:

Claim 62 *Let $n \in \mathbb{Z}_+$ and $x \in [n] \cup \{0\}$. For any $-1/n < \delta < 1/n$, we have that*

$$\begin{aligned}
G'_{n,x}(\delta) &= G_{n,x}(\delta)\left(h(n,x) - \mathbf{E}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)]\right), \\
G''_{n,x}(\delta) &= G_{n,x}(\delta)\left((h(n,x) - \mathbf{E}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)])^2 - \mathbf{Var}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)]\right).
\end{aligned}$$

Proof By definition, we have that $Z_n(\delta) = \sum_{x=0}^n \binom{n}{x}\exp(h(n,x)\delta)$ and $Z'_n(\delta) = \sum_{x=0}^n \binom{n}{x} h(n,x)\exp(h(n,x)\delta)$. Therefore,

$$\begin{aligned}
G'_{n,x}(\delta) &= \binom{n}{x}\left(\frac{h(n,x)\exp(h(n,x))}{Z_n(\delta)} - \frac{\exp(h(n,x)\delta)Z'_n(\delta)}{Z_n(\delta)^2}\right) \\
&= \binom{n}{x}\left(\frac{h(n,x)\exp(h(n,x))}{Z_n(\delta)} - \left(\frac{\exp(h(n,x)\delta)}{Z_n(\delta)}\right)\left(\frac{\sum_{y=0}^n \binom{n}{y} h(n,y)\exp(h(n,y)\delta)}{Z_n(\delta)}\right)\right) \\
&= G_{n,x}(\delta)\left(h(n,x) - \mathbf{E}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)]\right).
\end{aligned}$$

For the second derivative, applying the above result for the first derivative yields

$$\begin{aligned}
G''_{n,x}(\delta) &= G'_{n,x}(\delta)\left(h(n,x) - \mathbf{E}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)]\right) + G_{n,x}(\delta)\frac{d}{d\delta}\left(\mathbf{E}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)]\right) \\
&= G_{n,x}(\delta)\left(\left(h(n,x) - \mathbf{E}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)]\right)^2 + \sum_{y=0}^{n} h(n,y)G'_{n,y}(\delta)\right) \\
&= G_{n,x}(\delta)\left(\left(h(n,x) - \mathbf{E}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)]\right)^2 + \sum_{y=0}^{n} h(n,y)G_{n,y}(\delta)\left(h(n,y) - \mathbf{E}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)]\right)\right) \\
&= G_{n,x}(\delta)\left(\left(h(n,x) - \mathbf{E}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)]\right)^2 + \mathbf{E}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)^2] - \mathbf{E}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)]^2\right) \\
&= G_{n,x}(\delta)\left((h(n,x) - \mathbf{E}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)])^2 - \mathbf{Var}_{Y\sim \mathrm{IS}(n,\delta)}[h(n,Y)]\right) .
\end{aligned}$$

■

The following claim states that for any sufficiently small parameter $\delta \geq 0$, $X - n/2$ and $h(n,X)$ have sharp sub-Gaussian and sub-exponential tail, respectively.

Claim 63 *Let $n \in \mathbb{Z}_+$. There exists universal constants $C_1, C_2 > 0$ such that, for any $0 \leq \delta \leq \frac{1}{2n}$, we have that $\|X - n/2\|_{\psi_2} \leq C_1\sqrt{n}$ and $\|h(n,X)\|_{\psi_1} \leq C_2 n$, where $X \sim \mathrm{IS}(n,\delta)$.*

Proof We consider the Ising model P_θ, where $\theta_{ij} = \delta, \forall i,j \in [n], i \neq j$. Since $0 \leq \delta \leq \frac{1}{2n}$, we have that $\sum_{j\in[n],j\neq i} |\theta_{ij}| \leq 1/2, \forall i \in [n]$. Let $X \sim \mathrm{IS}(n,\delta)$. By definition, we know that X denotes the number of 1's in the random vector of P_θ. Therefore, applying Fact 45 by taking b to be the all-ones vector, we have that $\|X - n/2\|_{\psi_2} \leq C_1\sqrt{n}$ for some universal constant $C_1 > 0$. Similarly, applying Fact 46 by taking A to be the all-ones matrix, we have that $\|h(n,X)\|_{\psi_1} \leq C_2 n$ for some universal constant $C_2 > 0$. ■

We pick $C = \Theta(\sqrt{(\log(1/\delta)/m)})$, where the hidden constant is sufficiently small and consider the interval $I_{C,m} = [(1/2 - C)m, (1/2 + C)m - 1]$. Without loss of generality, we assume that the two endpoints of $I_{C,m}$ are integers. We define the one-dimensional distribution A to be:

- For $x \notin I_{C,m}$, we define $A(x) = \mathrm{IS}(m, \delta/m)(x)$.
- For $x \in I_{C,m}$, we define $A(x) = \mathrm{IS}(m, \delta/m)(x) + \int_x^{x+1} p(t)dt$, where p is a polynomial of degree at most k satisfying

$$\sum_{x\in\mathbb{Z}\cap I_{C,m}} x^i \int_x^{x+1} p(t)dt = \sum_{x=0}^{m} (\mathrm{IS}(m,0)(x) - \mathrm{IS}(m,\delta/m)(x))x^i, \tag{6}$$

  for $0 \leq i \leq k$.

Applying Theorem 32 with the family of functions $\{G_{m,x}(\delta)\}_{x\in[m]\cup\{0\}}$, we know that there is a unique real polynomial p of degree at most k satisfying the above properties. Then we need to show that with sufficiently large m (depending on δ), both the L_1 and L_∞ norms of p on $[(1/2 - C)m, (1/2 + C)m]$ are sufficiently small in order to make $A(x)$ non-negative. The main technical result of this section is the following lemma, which provides upper bounds on the L_1 and L_∞ norms of p on the interval $[(1/2 - C)m, (1/2 + C)m]$.

Lemma 64 *Let $1 \le k^2 \le C_0 C^2 m$ for some universal constant $C_0 > 0$ sufficiently small and $m \ge C_1(\log(1/\delta))^3$ for some universal constant $C_1 > 0$ sufficiently large. Then $\int_{(1/2-C)m}^{(1/2+C)m} |p(t)| dt \le O\left(\frac{\delta k^3}{C^2 m}\right)$ and $|p(t^*)| \le O\left(\frac{\delta k^{7/2}}{C^3 m^2}\right)$, where $t^* = \arg\max_{|t-m/2| \le Cm} |p(t)|$.*

Before we prove Lemma 64, we first use it to prove our main Proposition 31. The following lemma gives both the lower and upper bound of the ratio between the mass of $\mathrm{IS}(m,0)$ and $\mathrm{IS}(m,\delta/m)$.

Lemma 65 *Let $m \in \mathbb{Z}_+$ and $x \in [m] \cup \{0\}$. There is a universal constant $\delta_0 > 0$ such that for any $0 \le \delta \le \delta_0$, we have that*

$$e^{-\delta^2/\delta_0^2} \cdot \exp(h(m,x)\delta/m) \le \frac{G_{m,x}(\delta/m)}{G_{m,x}(0)} \le \exp(h(m,x)\delta/m).$$

Proof By definition, we have that

$$\frac{G_{m,x}(\delta/m)}{G_{m,x}(0)} = \frac{\exp(h(m,x)\delta/m)}{2^{-m} Z_m(\delta/m)} = \frac{\exp(h(m,x)\delta/m)}{2^{-m}\sum_{y=0}^{m}\binom{m}{y}\exp(h(m,y)\delta/m)}$$
$$= \frac{\exp(h(m,x)\delta/m)}{\mathbf{E}_{Y\sim \mathrm{IS}(m,0)}[\exp(h(m,Y)\delta/m)]}.$$

The upper bound is due to Claim 61 and Jenson's inequality that

$$\mathbf{E}_{Y\sim \mathrm{IS}(m,0)}[\exp(h(m,Y)\delta/m)] \ge \exp\left(\mathbf{E}_{Y\sim \mathrm{IS}(m,0)}[h(m,Y)\delta/m]\right) = 1.$$

To prove the lower bound, by Claim 63, we have that $\|h(m,Y)\|_{\psi_1} \le O(m)$. Therefore, by Fact 41, there is a universal constant $\delta_0 > 0$ such that for every $0 \le \delta \le \delta_0$, we have that $\mathbf{E}_{Y\sim \mathrm{IS}(m,0)}[\exp(h(m,Y)\delta/m)] \le \exp((m^2/\delta_0^2)(\delta^2/m^2)) = \exp(\delta^2/\delta_0^2)$. ■

We now bound from above the desired χ^2-divergence:

Lemma 66 *We have that*

$$\chi^2(A, \mathrm{Bin}(m,1/2)) \le O\left(\delta^2 + \frac{\delta k^3 \exp(2\delta C^2 m)}{C^2 m} + \left(\frac{\delta k^3}{C^2 m}\right)\cdot \max_{x\in\mathbb{Z}\cap I_{C,m}} \frac{\int_x^{x+1}|p(t)|dt}{G_{m,x}(0)}\right).$$

Proof We have the following:

$$1 + \chi^2(A, \mathrm{Bin}(m,1/2)) = \sum_{x=0}^{m}\frac{A(x)^2}{G_{m,x}(0)} = \sum_{x=0}^{m}\frac{\left(G_{m,x}(\delta/m) + \mathbb{I}[x\in I_{C,m}]\int_x^{x+1}p(t)dt\right)^2}{G_{m,x}(0)}$$
$$= \sum_{x=0}^{m}\frac{G_{m,x}(\delta/m)^2}{G_{m,x}(0)} + 2\sum_{x\in\mathbb{Z}\cap I_{C,m}}\frac{G_{m,x}(\delta/m)\int_x^{x+1}p(t)dt}{G_{m,x}(0)} + \sum_{x\in\mathbb{Z}\cap I_{C,m}}\frac{\left(\int_x^{x+1}p(t)dt\right)^2}{G_{m,x}(0)}.$$

For the first term, by Lemma 65, we have that

$$\sum_{x=0}^{m} \frac{G_{m,x}(\delta/m)^2}{G_{m,x}(0)} \le \sum_{x=0}^{m} G_{m,x}(0)\exp(2h(m,x)\delta/m) = \mathbf{E}_{X\sim \mathrm{IS}(m,0)}[\exp(2h(m,X)\delta/m)]$$
$$\le \exp(O(\delta^2)) \le 1+O(\delta^2).$$

For the second term, by Lemma 65 and Lemma 64, we have that

$$\sum_{x\in\mathbb{Z}\cap I_{C,m}} \frac{G_{m,x}(\delta/m)\int_x^{x+1}p(t)dt}{G_{m,x}(0)} \le \sum_{x\in\mathbb{Z}\cap I_{C,m}} \exp(h(m,x)\delta/m)\cdot\left|\int_x^{x+1}p(t)dt\right|$$
$$\le \exp\left((2C^2m-1/2)\delta\right)\sum_{x\in\mathbb{Z}\cap I_{C,m}}\left|\int_x^{x+1}p(t)dt\right| \le \exp(2\delta C^2 m)\int_{(1/2-C)m}^{(1/2+C)m}|p(t)|dt$$
$$\le O\left(\frac{\delta k^3\exp(2\delta C^2m)}{C^2m}\right),$$

where the second inequality follows from the fact that $h(m,x)=2x^2-2mx+\frac{m(m-1)}{2}$ attains its maximum at $x=(1/2-C)m$ over the interval $I_{C,m}$. For the third term, by Lemma 64, we have that

$$\sum_{x\in\mathbb{Z}\cap I_{C,m}} \frac{\left(\int_x^{x+1}p(t)dt\right)^2}{G_{m,x}(0)} \le \int_{(1/2-C)m}^{(1/2+C)m}|p(t)|dt\cdot \max_{x\in\mathbb{Z}\cap I_{C,m}}\frac{\int_x^{x+1}|p(t)|dt}{G_{m,x}(0)}$$
$$\le O\left(\frac{\delta k^3}{C^2m}\right)\cdot\max_{x\in\mathbb{Z}\cap I_{C,m}}\frac{\int_x^{x+1}|p(t)|dt}{G_{m,x}(0)}.$$

Combining the above results together completes the proof. ■

We are now ready to prove Proposition 31. We need to pick C appropriately and check the bounds on k needed for $A(x)$ to satisfy the necessary properties.

Proof [Proof of Proposition 31] Let $C=\Theta(\sqrt{\log(1/\delta)/m})$. If $k^3\ge C^2m$, we pick $A=\mathrm{Bin}(m,1/2)$ and obtain $d_{\mathrm{TV}}(A,G_{m,\delta/m})\le O(\delta)\le O\left(\frac{\delta k^3}{\log(1/\delta)}\right)$. Thus, we assume that $k^3\le C^2m$. In this way, to apply Lemma 54, we need $k^2\le C_0C^2m$ for some universal constant C_0 sufficiently small, which will be satisfied as long as $\delta\le\exp(-1/C_0^3)$.

We first show that $A(x)$ is indeed a distribution over $[m]\cup\{0\}$. By definition, $A(x)$ is nonnegative outside the interval $I_{C,m}$. For $x\in\mathbb{Z}\cap I_{C,m}$, we apply Fact 34, Lemma 54 and Lemma 65 to obtain

$$
\begin{aligned}
A(x) &= G_{m,x}(\delta/m) + \mathbb{I}[x \in I_{C,m}] \int_x^{x+1} p(t)dt \\
&\geq e^{-O(\delta^2)} \exp(h(m,x)\delta/m) G_{m,x}(0) - |p(t^*)| = \frac{\binom{m}{x} \exp\left(h(m,x)\delta/m\right)}{2^m \exp(O(\delta^2))} - |p(t^*)| \\
&\geq \sqrt{\frac{m}{8x(m-x)}} 2^{mH(x/m)} \cdot \frac{\exp\left((2x^2/m - 2x + (m-1)/2)\delta\right)}{2^m \exp(O(\delta^2))} - O\left(\frac{\delta k^{7/2}}{C^3 m^2}\right) \\
&\geq \sqrt{\frac{1}{2m}} \cdot 2^{mH(q)} \cdot \frac{\exp\left((2mq^2 - 2mq + (m-1)/2)\delta\right)}{2^m \exp(O(\delta^2))} - O\left(\frac{\delta k^{7/2}}{C^3 m^2}\right) \\
&\geq \Omega\left(\sqrt{\frac{1}{m}}\right) \cdot 2^{m\left(H(q)+2\delta(q^2-q)-1\right)} \cdot \exp\left((m-1)\delta/2\right) - O\left(\frac{\delta k^{7/2}}{C^3 m^2}\right),
\end{aligned}
$$

where we let $q = x/m$ and apply the fact $x(m-x) \leq m^2/4$ in the third inequality. Let $f(q) = H(q) + \lambda(q^2 - q) - 1$, where $\lambda = 2\delta$. We have that $f'(q) = \log_2\left(\frac{1-q}{q}\right) + (2q-1)\lambda$ and $f''(q) = 2\lambda - \frac{1}{q(1-q)} \leq 2\lambda - 4 < 0$ as long as $\lambda < 2$, which implies that $f(q)$ is strictly concave over $[0,1]$ and attains its maximum at $q = 1/2$. Therefore, we have that

$$
\begin{aligned}
A(x) &\geq \Omega\left(\sqrt{\frac{1}{m}}\right) \cdot 2^{mf(x/m)} \cdot \exp\left((m-1)\delta/2\right) - O\left(\frac{\delta k^{7/2}}{C^3 m^2}\right) \\
&\geq \Omega\left(\sqrt{\frac{1}{m}}\right) \cdot 2^{mf(1/2+C)} \cdot \exp\left((m-1)\delta/2\right) - O\left(\frac{\delta k^{7/2}}{C^3 m^2}\right) \\
&\geq \Omega\left(\sqrt{\frac{1}{m}}\right) \cdot 2^{m(H(1/2+C)-1)} \cdot 2^{(2mC^2-1/2)\delta} - O\left(\frac{\delta k^{7/2}}{C^3 m^2}\right) \\
&\geq \sqrt{\frac{1}{m}} \cdot \exp\left(-O(C^2 m)\right) - O\left(\frac{\delta k^{7/2}}{C^3 m^2}\right),
\end{aligned}
$$

where the last inequality follows from the Taylor expansion of $H(1/2+C) - H(1/2)$ up to second order terms. Note that $k^3 \leq C^2 m$, by our choice of C, where $C = \Theta(\sqrt{\log(1/\delta)/m})$ for some sufficiently small hidden constant in Θ, we have that $\frac{\delta k^{7/2}}{C^3 m^{3/2}} \leq O(\delta(\log(1/\delta))^{-1/3})$ and $\exp(-O(C^2 m)) \geq \delta$. Therefore, we have that

$$
A(x) \geq \sqrt{\frac{1}{m}} \cdot \exp\left(-O(C^2 m)\right) - O\left(\frac{\delta k^{7/2}}{C^3 m^2}\right) \geq 0, \forall x \in \mathbb{Z} \cap I_{C,m}.
$$

In addition, by Equation (6), we know that

$$
\begin{aligned}
\sum_{x=0}^{m} A(x) &= \sum_{x=0}^{m} \left(G_{m,x}(\delta/\sqrt{m}) + \mathbb{I}[x \in I_{C,m}] \int_x^{x+1} p(t)dt \right) \\
&= \sum_{x=0}^{m} G_{m,x}(\delta/\sqrt{m}) + \int_{(1/2-C)m}^{(1/2+C)m} p(t)dt = 1,
\end{aligned}
$$

which implies that the distribution A is well-defined. Furthermore, by Equation (6), we can show that A matches the first k moments of $\mathrm{Bin}(m, 1/2)$ as follows:

$$\begin{aligned}\mathbf{E}_{X\sim A}[X^i] &= \sum_{x=0}^{m} A(x)x^i = \sum_{x=0}^{m}\left(G_{m,x}(\delta/\sqrt{m}) + \mathbb{I}[x\in I_{C,m}]\int_x^{x+1} p(t)dt\right)x^i \\ &= \sum_{x=0}^{m} G_{m,x}(\delta/\sqrt{m})x^i + \sum_{x\in\mathbb{Z}\cap I_{C,m}} x^i \int_x^{x+1} p(t)dt \\ &= \sum_{x=0}^{m} G_{m,x}(0)x^i = \mathbf{E}_{X\sim\mathrm{Bin}(m,1/2)}[X^i].\end{aligned}$$

From the previous calculations, we have that $A(x) \geq e^{-O(\delta^2)}\exp(h(m,x)\delta/m)G_{m,x}(0) - |p(t^*)| \geq 0, \forall x \in \mathbb{Z}\cap I_{C,m}$, which implies that for every $x \in \mathbb{Z}\cap I_{C,m}$,

$$|p(t^*)| \leq e^{-O(\delta^2)}\exp(h(m,x)\delta/m)G_{m,x}(0) \leq \exp(2\delta C^2 m)G_{m,x}(0).$$

Therefore, by Lemma 66, we have that

$$\begin{aligned}\chi^2(A, \mathrm{Bin}(m,1/2)) &\leq O\left(\delta^2 + \frac{\delta k^3 \exp(2\delta C^2 m)}{C^2 m} + \left(\frac{\delta k^3}{C^2 m}\right)\cdot \max_{x\in\mathbb{Z}\cap I_{C,m}} \frac{\int_x^{x+1}|p(t)|dt}{G_{m,x}(0)}\right) \\ &\leq O\left(\delta^2 + \delta\left(\exp(2\delta C^2 m) + \frac{|p(t^*)|}{G_{m,x}(0)}\right)\right) \leq O\left(\delta^2 + 2\delta\exp(2\delta C^2 m)\right) \\ &\leq O\left(\delta^2 + \delta(1 + O(\delta\log(1/\delta)))\right) = O(\delta),\end{aligned}$$

where the last inequality follows from the fact $e^x \leq 1 + 2x, \forall x \in [0, \ln 2]$.

Finally, to bound the total variation distance $d_{\mathrm{TV}}(A, \mathrm{IS}(m, \delta/m))$, we apply Lemma 64 to obtain

$$\begin{aligned}d_{\mathrm{TV}}(A, \mathrm{IS}(m,\delta/m))) &= \sum_{x\in\mathbb{Z}\cap I_{C,m}}\left|\int_x^{x+1} p(t)dt\right| \leq \int_{(1/2-C)m}^{(1/2+C)m} |p(t)|dt \\ &\leq O\left(\frac{\delta k^3}{C^2 m}\right) = O\left(\frac{\delta k^3}{\log(1/\delta)}\right).\end{aligned}$$

■

Proof of Lemma 64 By Theorem 32, we have that

$$|a_i| \leq \left(\frac{2i+1}{2Cm}\right)\left(\beta_i + O\left(\frac{i^2}{Cm}\right)\int_{(1/2-C)m}^{(1/2+C)m}|p(t)|dt\right),$$

for all $1 \leq i \leq k$, where $\beta_i = \left|\sum_{x=0}^{m}(G_{m,x}(0) - G_{m,x}(\delta/m))P_i\left(\frac{x-m/2}{Cm}\right)\right|$. To get an upper bound for the L_1 and L_∞ norms of the polynomial p over $[(1/2-C)m, (1/2+C)m]$, we only need to upper bound the quantity β_i.

Lemma 67 *If $k^2 \le C_0 C^2 m$ for some universal constant $C_0 > 0$ sufficiently small, then $\beta_i \le O\left(\frac{\delta i^{3/2}}{C^2 m}\right), \forall 1 \le i \le k$.*

We assume $k^2 \le C_0 C^2 m$ for some universal constant $C_0 > 0$ sufficiently small. Note that by our definition $G_{m,x}(\delta) = G_{m,m-x}(\delta), \forall -1/2 < \delta < 1/2$, by Fact 4 (iv), we have that $\beta_i = \left|\sum_{x=0}^{m}(G_{m,x}(0) - G_{m,x}(\delta/m))P_i\left(\frac{x-m/2}{Cm}\right)\right| = 0$ for any odd i. Hence, we only need to bound β_i for every even i. We apply Taylor's theorem to expand $G_{m,x}(\delta/m) - G_{m,x}(0)$ up to second order terms:

$$\begin{aligned}
\beta_i &= \left|\sum_{x=0}^{m}\left(G_{m,x}(0) - G_{m,x}(\delta/\sqrt{m})\right)P_i\left(\frac{x-m/2}{Cm}\right)\right| \\
&= \left|\sum_{x=0}^{m}\left(\frac{\delta G'_{m,x}(0)}{\sqrt{m}} + \frac{G''_{m,x}(\delta_x)\delta^2}{2m}\right)P_i\left(\frac{x-m/2}{Cm}\right)\right| \\
&\le \underbrace{\left|\sum_{x=0}^{m}\left(\frac{\delta G'_{m,x}(0)}{\sqrt{m}}\right)P_i\left(\frac{x-m/2}{Cm}\right)\right|}_{\beta'_i} + \underbrace{\left|\sum_{x=0}^{m}\left(\frac{G''_{m,x}(\delta_x)\delta^2}{2m}\right)P_i\left(\frac{x-m/2}{Cm}\right)\right|}_{\beta''_i}, \qquad (7)
\end{aligned}$$

where for any $x \in [m] \cup \{0\}$, $\delta_x = \widetilde{\delta_x}/\sqrt{m}$ for some $\widetilde{\delta_x} \in [0, \delta]$.

Hence, in order to bound β_i, it suffices to bound the terms $\beta'_i := \frac{\delta}{\sqrt{m}}\left|\sum_{x=0}^{m} G'_{m,x}(0)P_i\left(\frac{x-m/2}{Cm}\right)\right|$ and $\beta''_i := \frac{\delta^2}{2m}\left|\sum_{x=0}^{m} G''_{m,x}(\delta_x)P_i\left(\frac{x-m/2}{Cm}\right)\right|$. This is done in the following lemmas.

Lemma 68 *For every even i, we have that $\beta'_i \le O\left(\frac{\delta i^{3/2}}{C^2 m}\right)$.*

Proof By Claim 61 and Claim 62, we have that $G'_{m,x}(0) = G_{m,x}(0)h(m,x)$. Applying Claim 61 and Fact 4 (v) by using the change of variables yields

$$\begin{aligned}
\beta'_i &= \left|\sum_{x=0}^{m}\left(\frac{\delta G'_{m,x}(0)}{m}\right)P_i\left(\frac{x-m/2}{Cm}\right)\right| = \frac{\delta}{m}\left|\sum_{x=0}^{m} G_{m,x}(0)h(m,x)P_i\left(\frac{x-m/2}{Cm}\right)\right| \\
&= \frac{\delta}{m}\left|\sum_{x=0}^{m} G_{m,x}(0)h(m,x)2^{-i}\sum_{j=0}^{i/2}(-1)^j\binom{i}{j}\binom{2i-2j}{i}\left(\frac{x-m/2}{Cm}\right)^{i-2j}\right| \\
&= \frac{\delta}{m}\left|\sum_{x=0}^{m} G_{m,x}(0)h(m,x)2^{-i}\sum_{j=0}^{i/2-1}(-1)^j\binom{i}{j}\binom{2i-2j}{i}\left(\frac{x-m/2}{Cm}\right)^{i-2j}\right| \\
&\le \frac{\delta}{m}\sum_{x=0}^{m} G_{m,x}(0)\left|h(m,x)\right|2^{-i}\sum_{j=1}^{i/2}\binom{i}{i/2+j}\binom{i+2j}{2j}\left|\frac{x-m/2}{Cm}\right|^{2j} \\
&\le \frac{\delta}{m}\sum_{j=1}^{i/2}\sum_{x=0}^{m} G_{m,x}(0)\left|h(m,x)(x-m/2)^{2j}\right| O\left(\frac{1}{\sqrt{i}}\right)\frac{(i+2j)^{2j}}{(2j)!(Cm)^{2j}}.
\end{aligned}$$

Since $\|X - m/2\|_{\psi_2} \leq O(\sqrt{m})$ for $X \sim \mathrm{Bin}(m, 1/2)$, applying Fact 38 and Fact 33 yields

$$\begin{aligned}
\beta_i &\leq \frac{\delta}{m} \sum_{j=1}^{i/2} \sum_{x=0}^{m} G_{m,x}(0) \left| h(m,x)(x-m/2)^{2j} \right| O\left(\frac{1}{\sqrt{i}}\right) \left(\frac{(i+2j)^{2j}}{(2j)!(Cm)^{2j}} \right) \\
&= \frac{\delta}{m} \sum_{j=1}^{i/2} \sum_{x=0}^{m} G_{m,x}(0) \left| (2x^2 - 2mx + m(m-1)/2)(x-m/2)^{2j} \right| O\left(\frac{1}{\sqrt{i}}\right) \left(\frac{(i+2j)^{2j}}{(2j)!(Cm)^{2j}} \right) \\
&\leq \frac{\delta}{m} \sum_{j=1}^{i/2} \sum_{x=0}^{m} G_{m,x}(0) \left(2(x-m/2)^{2j+2} + m(x-m/2)^{2j}/2 \right) O\left(\frac{1}{\sqrt{i}}\right) \left(\frac{(i+2j)^{2j}}{(2j)!(Cm)^{2j}} \right) \\
&= \frac{\delta}{m} \sum_{j=1}^{i/2} O\left(\frac{1}{\sqrt{i}}\right) \left(\frac{(i+2j)^{2j} \mathbf{E}_{X \sim \mathrm{Bin}(m,1/2)}[2(X-m/2)^{2j+2} + m(X-m/2)^{2j}/2]}{(2j)!(Cm)^{2j}} \right) \\
&\leq O\left(\frac{\delta}{m\sqrt{i}}\right) \sum_{j=1}^{i/2} \frac{(i+2j)^{2j}}{(2j)!(Cm)^{2j}} \left((2j+2)(j!)(O(m))^{j+1} + (j!)(O(m))^{j+1} \right) \\
&\leq O\left(\frac{\delta}{\sqrt{i}}\right) \sum_{j=1}^{\infty} \left(O\left(\frac{i}{C\sqrt{m}}\right) \right)^{2j} \leq O\left(\frac{\delta i^{3/2}}{C^2 m} \right).
\end{aligned}$$

■

Lemma 69 *For every even i, we have that $\beta_i'' \leq O(\delta^2)$.*

Proof By Claim 62, we have that

$$\begin{aligned}
\beta_i'' &= \left| \sum_{x=0}^{m} \left(\frac{G_{m,x}''(\delta_x)\delta^2}{2m^2} \right) P_i\left(\frac{x-m/2}{Cm} \right) \right| \\
&= \frac{\delta^2}{2m^2} \left| \sum_{x=0}^{m} G_{m,x}(\delta_x) \left(\left(h(m,x) - \mathop{\mathbf{E}}_{Y \sim \mathrm{IS}(m,\delta_x)}[h(m,Y)] \right)^2 - \mathop{\mathbf{Var}}_{Y \sim \mathrm{IS}(m,\delta_x)}[h(m,Y)] \right) P_i\left(\frac{x-m/2}{Cm} \right) \right|.
\end{aligned}$$

We separate the above sum into $x \in \mathbb{Z} \cap I_{C,m}$ and $x \in [m] \cup \{0\} \setminus I_{C,m}$. We are able to use Fact 4 (iii) to bound the sum for $x \in \mathbb{Z} \cap I_{C,m}$, as follows:

$$\begin{aligned}
&\sum_{x \in \mathbb{Z} \cap I_{C,m}} G_{m,x}(\delta_x) \left| \left(h(m,x) - \mathop{\mathbf{E}}_{Y \sim \mathrm{IS}(m,\delta_x)}[h(m,Y)] \right)^2 - \mathop{\mathbf{Var}}_{Y \sim \mathrm{IS}(m,\delta_x)}[h(m,Y)] \right| \left| P_i\left(\frac{x-m/2}{Cm} \right) \right| \\
&\leq \sum_{x=0}^{m} G_{m,x}(\delta_x) \left| \left(h(m,x) - \mathop{\mathbf{E}}_{Y \sim \mathrm{IS}(m,\delta_x)}[h(m,Y)] \right)^2 - \mathop{\mathbf{Var}}_{Y \sim \mathrm{IS}(m,\delta_x)}[h(m,Y)] \right| \\
&\leq 2\mathbf{Var}_{Y \sim \mathrm{IS}(m,\delta_x)}[h(m,Y)] \leq O(m^2),
\end{aligned}$$

where the last inequality follows from Fact 39, Fact 42 and Claim 63.

Now we bound the sum over $x \in \overline{I_{C,m}}$, where $\overline{I_{C,m}} = [m] \cup \{0\} \setminus I_{C,m}$. Note that for $X \sim \mathrm{IS}(m, \delta_x)$, we have that $\left\| \frac{4(X-m/2)}{Cm} \right\|_{\psi_2} \leq O\left(\frac{1}{C\sqrt{m}}\right)$ and $\|h(m,X)\|_{\psi_1} \leq O(m)$. Therefore, applying Fact 4 (vi) yields

$$
\begin{aligned}
&\sum_{x \in \overline{I_{C,m}}} G_{m,x}(\delta_x) \left| \left(h(m,x) - \mathop{\mathbf{E}}_{Y \sim G_{m,\delta_x}}[h(m,Y)]\right)^2 - \mathop{\mathbf{Var}}_{Y \sim \mathrm{IS}(m,\delta_x)}[h(m,Y)] \right| \left| P_i\left(\frac{x-m/2}{Cm}\right) \right| \\
&\leq \sum_{x=0}^{m} G_{m,x}(\delta_x) \left| \left(h(m,x) - \mathop{\mathbf{E}}_{Y \sim \mathrm{IS}(m,\delta_x)}[h(m,Y)]\right)^2 - \mathop{\mathbf{Var}}_{Y \sim \mathrm{IS}(m,\delta_x)}[h(m,Y)] \right| \left(\frac{4|x-m/2|}{Cm}\right)^i \\
&\leq \mathop{\mathbf{E}}_{X \sim \mathrm{IS}(m,\delta_x)} \left[\left(h(m,X) - \mathop{\mathbf{E}}_{Y \sim \mathrm{IS}(m,\delta_x)}[h(m,Y)]\right)^2 \left(\frac{4|X-m/2|}{Cm}\right)^i \right] \\
&\quad + O(m^2) \cdot \mathop{\mathbf{E}}_{X \sim \mathrm{IS}(m,\delta_x)} \left[\left(\frac{4|X-m/2|}{Cm}\right)^i \right] \\
&\leq \sqrt{ \mathop{\mathbf{E}}_{X \sim \mathrm{IS}(m,\delta_x)} \left[\left(h(m,X) - \mathop{\mathbf{E}}_{Y \sim \mathrm{IS}(m,\delta_x)}[h(m,Y)]\right)^4 \right] \cdot \mathop{\mathbf{E}}_{X \sim \mathrm{IS}(m,\delta_x)} \left[\left(\frac{4|X-m/2|}{Cm}\right)^{2i} \right] } \\
&\quad + O(m^2) \cdot \mathop{\mathbf{E}}_{X \sim \mathrm{IS}(m,\delta_x)} \left[\left(\frac{4|X-m/2|}{Cm}\right)^i \right] \\
&\leq \sqrt{ O(m^4) \left(O\left(\frac{i}{C^2 m}\right)\right)^i } + O(m^2) \left(O\left(\frac{1}{C}\sqrt{\frac{i}{m}}\right)\right)^i \leq O(m^2),
\end{aligned}
$$

where the third inequality follows from Cauchy-Schwarz and the fourth inequality follows from Fact 38, Fact 39, Fact 42 and Claim 63. Combine the above results together, we have that

$$
\beta_i'' = \frac{\delta^2}{2m^2} \left| \sum_{x=0}^{m} G_{m,x}''(\delta_x) P_i\left(\frac{x-m/2}{Cm}\right) \right| \leq \left(\frac{\delta^2}{2m^2}\right) \cdot O(m^2) = O(\delta^2).
$$

■

Now we are ready to prove Lemma 64.

Proof [Proof of Lemma 64] By Theorem 32, we have that

$$
\begin{aligned}
|p(t^*)| &\leq \sum_{i=1}^{k} |a_i| \leq \sum_{i=1}^{k} \left(\frac{2i+1}{2Cm}\right) \beta_i + \sum_{i=1}^{k} \left(\frac{2i+1}{2Cm}\right) O\left(\frac{i^2}{Cm}\right) \int_{(1/2-C)m}^{(1/2+C)m} |p(t)| dt \\
&\leq \sum_{i=1}^{k} \left(\frac{2i+1}{2Cm}\right) \beta_i + |p(t^*)| \sum_{i=1}^{k} \left(\frac{2i+1}{2}\right) O\left(\frac{i^2}{Cm}\right) \\
&\leq \sum_{i=1}^{k} \left(\frac{2i+1}{2Cm}\right) \beta_i + O\left(\frac{k^4}{Cm}\right) |p(t^*)|,
\end{aligned}
$$

where the first inequality follows from Fact 4 (iii). Similarly, by Theorem 32, we have that

$$\begin{aligned}\int_{(1/2-C)m}^{(1/2+C)m} |p(t)|dt &\le \sum_{i=1}^{k} |a_i| \int_{(1/2-C)m}^{(1/2+C)m} \left|P_i\left(\frac{t-m/2}{Cm}\right)\right| dt \\ &\le Cm \sum_{i=1}^{k} \left(\frac{2i+1}{2Cm}\right)\left(\beta_i + O\left(\frac{i^2}{Cm}\right)\int_{(1/2-C)m}^{(1/2+C)m} |p(t)|dt\right)\int_{-1}^{1} |P_i(y)|dy \\ &\le \sum_{i=1}^{k} O(\sqrt{i})\beta_i + \sum_{i=1}^{k} O(\sqrt{i}) O\left(\frac{i^2}{Cm}\right)\int_{(1/2-C)m}^{(1/2+C)m} |p(t)|dt \\ &\le \sum_{i=1}^{k} O(\sqrt{i})\beta_i + O\left(\frac{k^{7/2}}{Cm}\right)\int_{(1/2-C)m}^{(1/2+C)m} |p(t)|dt,\end{aligned}$$

where the third inequality follows from Fact 4 (vii). By our assumption on k, C, m, δ, we know that $\frac{k^{7/2}}{Cm} \le \frac{k^4}{Cm} \le 1/2$. Therefore, by Lemma 57, we have that

$$|p(t^*)| \le 2\sum_{i=1}^{k}\left(\frac{2i+1}{2Cm}\right)\beta_i \le \sum_{i=1}^{k}\left(\frac{2i+1}{2Cm}\right) O\left(\frac{\delta i^{3/2}}{C^2 m}\right) \le O\left(\frac{\delta k^{7/2}}{C^3 m^2}\right),$$

$$\int_{(1/2-C)m}^{(1/2+C)m} |p(x)|dx \le 2\sum_{i=1}^{k} O(\sqrt{i})\beta_i \le 2\sum_{i=1}^{k} O(\sqrt{i}) O\left(\frac{\delta i^{3/2}}{C^2 m}\right) \le O\left(\frac{\delta k^3}{C^2 m}\right).$$

This completes the proof. ■

C.2. Proof of Lemma 28

Let $n = |S|$. Recalling that $G_{n,x}(\delta) = \binom{n}{x} \exp\left(h(n,x)\delta\right)/Z_n(\delta)$, where $h(n,x) = 2x^2 - 2nx + \frac{n(n-1)}{2}$ and $Z_n(\delta) = \sum_{x=0}^{n} \binom{n}{x} \exp(h(n,x)\delta)$. By Claim 63, for any $0 \le \delta \le \frac{1}{2n}$, we have that $\|h(n,X)\|_{\psi_1} \le O(n)$ and $\|X - n/2\|_{\psi_2} \le O(\sqrt{n})$ for $X \sim \mathrm{IS}(n,\delta)$.

Define $f(\mathbf{x}) = \sum_{i\in S} x_i, \forall \mathbf{x} \in \{0,1\}^M$. For $\mathbf{X} \sim U_M$ and $\mathbf{Y} \sim Q_M^{S,\frac{\delta}{m}}$, by the data processing inequality, we have that

$$d_{\mathrm{TV}}\left(U_M, Q_M^{S,\frac{\delta}{m}}\right) \ge d_{\mathrm{TV}}(f(\mathbf{X}), f(\mathbf{Y})) = d_{\mathrm{TV}}(\mathrm{IS}(n,0), \mathrm{IS}(n,\delta/m)).$$

By the mean value theorem, we have that

$$d_{\mathrm{TV}}(\mathrm{IS}(n,0), \mathrm{IS}(n,\delta/m)) = \frac{1}{2}\sum_{x=0}^{n} |G_{n,x}(\delta/m) - G_{n,x}(0)| = \frac{1}{2}\sum_{x=0}^{n}\left|G'_{n,x}(0)(\delta/m) + \frac{G''_{n,x}(\delta_x)\delta^2}{2m^2}\right|,$$

where for any $x \in [m] \cup \{0\}$, $\delta_x = \widetilde{\delta_x}/m$ for some $\widetilde{\delta_x} \in (0,\delta)$.

By elementary calculation, we have that

$$\begin{aligned}\mathbf{E}_{X\sim \mathrm{IS}(n,0)}[h(n,X)^2] &= 4\mathbf{E}_{X\sim \mathrm{Bin}(n,1/2)}\left[\left((X-n/2)^2 - \mathbf{E}_{X\sim\mathrm{Bin}(n,1/2)}[(X-n/2)^2]\right)^2\right] \\ &= 4\left(\mathbf{E}_{X\sim\mathrm{Bin}(n,1/2)}\left[(X-n/2)^4\right] - \mathbf{E}_{X\sim\mathrm{Bin}(n,1/2)}\left[(X-n/2)^2\right]^2\right) \\ &= \left(\frac{3n^2}{4} - \frac{n}{2}\right) - \frac{n^2}{4} = \frac{n^2-n}{2}.\end{aligned}$$

By Fact 39, we have that $\mathbf{E}_{X\sim\mathrm{IS}(n,0)}[h(n,X)^4] \le O(n^4)$. Therefore, by Fact 35, we have that

$$\sum_{x=0}^{n} |G'_{n,x}(0)| = \sum_{x=0}^{n} G_{n,x}(0)|h(n,x)| = \mathbf{E}_{X\sim\mathrm{Bin}(n,1/2)}[|h(n,X)|] \ge \frac{\mathbf{E}_{X\sim\mathrm{Bin}(n,1/2)}[h(n,X)^2]^{3/2}}{\mathbf{E}_{X\sim\mathrm{Bin}(n,1/2)}[h(n,X)^4]^{1/2}} \ge \Omega(n).$$

In addition, by Fact 39, we have that

$$\begin{aligned}\sum_{x=0}^{n} |G''_{n,x}(\delta_x)| &= \sum_{x=0}^{n} G_{n,x}(\delta_x) \left|(h(n,x) - \mathbf{E}_{Y\sim\mathrm{IS}(n,\delta_x)}[h(n,Y)])^2 - \mathbf{Var}_{Y\sim\mathrm{IS}(n,\delta_x)}[h(n,Y)]\right| \\ &\le 2\mathbf{Var}_{X\sim\mathrm{IS}(n,\delta_x)}[h(n,X)] \le O\left(n^2\right).\end{aligned}$$

Therefore, we have that

$$d_{\mathrm{TV}}\left(U_M, Q_M^{S,\frac{\delta}{m}}\right) \ge \frac{1}{2}\sum_{x=0}^{n} \left| G'_{n,x}(0)(\delta/m) + \frac{G''_{n,x}(\delta_x)\delta^2}{2m^2} \right| \ge \Omega\left(\frac{\delta n}{m}\right) - O\left(\frac{n^2\delta^2}{m^2}\right) \ge \Omega(\delta) - O(\delta^2) = \Omega(\delta).$$