# Robust Estimators in High-Dimensions Without the Computational Intractability

SIAM J. COMPUT.
Vol. 48, No. 2, pp. 742–864


# ROBUST ESTIMATORS IN HIGH-DIMENSIONS WITHOUT THE COMPUTATIONAL INTRACTABILITY*

ILIAS DIAKONIKOLAS†, GAUTAM KAMATH‡, DANIEL KANE§, JERRY LI¶, ANKUR MOITRA‖, AND ALISTAIR STEWART#

**Abstract.** We study high-dimensional distribution learning in an agnostic setting where an adversary is allowed to arbitrarily corrupt an $\varepsilon$-fraction of the samples. Such questions have a rich history spanning statistics, machine learning, and theoretical computer science. Even in the most basic settings, the only known approaches are either computationally inefficient or lose dimension-dependent factors in their error guarantees. This raises the following question: Is high-dimensional agnostic distribution learning even possible, algorithmically? In this work, we obtain the first computationally efficient algorithms with dimension-independent error guarantees for agnostically learning several fundamental classes of high-dimensional distributions: (1) a single Gaussian, (2) a product distribution on the hypercube, (3) mixtures of two product distributions (under a natural balancedness condition), and (4) mixtures of spherical Gaussians. Our algorithms achieve error that is independent of the dimension, and in many cases scales nearly linearly with the fraction of adversarially corrupted samples. Moreover, we develop a general recipe for detecting and correcting corruptions in high-dimensions that may be applicable to many other problems.

**Key words.** robust learning, high-dimensions, Gaussian distribution, mixture models, product distributions

**AMS subject classifications.** 68Q25, 68Q32

**DOI.** 10.1137/17M1126680

## 1. Introduction.

**1.1. Background.** A central goal of machine learning is to design efficient algorithms for fitting a model to a collection of observations. In recent years, there has been considerable progress on a variety of problems in this domain, including algorithms with provable guarantees for learning mixture models [FOS08, KMV10, MV10,

*Received by the editors April 21, 2017; accepted for publication (in revised form) January 15, 2019; published electronically April 30, 2019.
http://www.siam.org/journals/sicomp/48-2/M112668.html
**Funding:** The first author's work was supported by NSF Award CCF-1652862 (CAREER) and a Sloan Research Fellowship. Part of this author's work was performed while at the University of Edinburgh, supported in part by a Marie Curie Career Integration grant. The first and sixth authors were supported in part by EPSRC grant EP/L021749/1. The second author was supported by NSF Award CCF-0953960 (CAREER) and ONR grant N00014-12-1-0999. This author's work was done in part while an intern at Microsoft Research Cambridge, visiting the Simons Institute for the Theory of Computing, and a graduate student at MIT. Part of the third author's work was performed while visiting the University of Edinburgh. The fourth author was supported by an NSF Fellowship. This author's work was done in part while an intern at Microsoft Research Cambridge and a graduate student at MIT. The fifth author was supported by a grant from the MIT NEC Corporation. The fourth and fifth authors were supposed by NSF CAREER Award CCF-1453261 and a Google Faculty Research Award. Part of the sixth author's work was performed while at the University of Edinburgh and the University of Southern California.
†Department of Computer Science, University of Southern California, Los Angeles, CA 90089 (diakonik@usc.edu).
‡Simons Institute for the Theory of Computing, University of California, Berkeley, Berkeley, CA 94720 (g@csail.mit.edu.)
§Department of Computer Science and Engineering, University of California, San Diego, San Diego, CA 92093 (dakane@ucsd.edu).
¶Microsoft Research AI, Redmond, WA 98052 (jerrl@microsoft.com).
‖Massachusetts Institute for Technology, Cambridge, MA 02139 (moitra@mit.edu).
#Web3 Foundation, 6300 Zug, Switzerland (stewart.al@gmail.com).

BS10, HK13], phylogenetic trees [CGG02, MR05], hidden Markov models [AHK12], topic models [AGM12, AGHK13], and independent component analysis [AGMS15]. These algorithms crucially rely on the assumption that the observations were actually generated by a model in the family. However, this simplifying assumption is not meant to be exactly true, and it is an important direction to explore what happens when it holds only in an approximate sense. In this work, we study the following family of questions.

QUESTION 1.1. *Let $\mathcal{D}$ be a family of distributions on $\mathbb{R}^d$. Suppose we are given samples generated from the following process: First, $m$ samples are drawn from some unknown distribution $P$ in $\mathcal{D}$. Then, an adversary is allowed to arbitrarily corrupt an $\varepsilon$-fraction of the samples. Can we efficiently find a distribution $P'$ in $\mathcal{D}$ that is $f(\varepsilon, d)$-close, in total variation distance, to $P$?*

This is a natural formalization of the problem of designing robust and efficient algorithms for distribution estimation. We refer to it as *(proper) agnostic distribution learning*, and we refer to the samples as being *$\varepsilon$-corrupted*. This family of problems has its roots in many fields, including statistics, machine learning, and theoretical computer science. Within computational learning theory, it is related to the agnostic learning model of Haussler [Hau92] and Kearns, Schapire, and Sellie [KSS94], where the goal is to learn a labeling function whose agreement with some underlying target function is close to the best possible, among all functions in some given class. In the even more challenging malicious noise model [Val85, KL93], an adversary is allowed to corrupt both the labels and the samples. A major difference with our setting is that these models apply to supervised learning problems, while here we will work in an unsupervised setting.

Within statistics and machine learning, inference problems like Question 1.1 are often termed "estimation under model misspecification." The usual prescription is to use the maximum likelihood estimator [Hub67, Whi82], which is unfortunately hard to compute in general. Even when ignoring computational considerations, the maximum likelihood estimator is only guaranteed to converge to the distribution $P'$ in $\mathcal{D}$ that is closest (in Kullback–Leibler divergence) to the distribution from which the observations are generated. This is problematic because such a distribution is not necessarily close to $P$ at all.

A branch of statistics—called robust statistics [HR09, HRRS86]—aims to tackle questions like the one above. The usual formalization is in terms of *breakdown point*, which (informally) is the fraction of observations that an adversary would need to control to be able to completely corrupt an estimator. In low-dimensions, this leads to the prescription that one should use the empirical median instead of the empirical mean to robustly estimate the mean of a distribution, and interquartile range for robust estimates of the variance. In high-dimensions, the Tukey median [Tuk75] is a high-dimensional analogue of the median that, although provably robust, is hard to compute [JP78]. Similar hardness results have been shown [Ber06, HM13] for essentially all known estimators in robust statistics.

*Is high-dimensional agnostic distribution learning even possible, algorithmically?* The difficulty is that corruptions are often hard to detect in high-dimensions and could bias the natural estimator by dimension-dependent factors. In this work, we study agnostic distribution learning for a number of fundamental classes of distributions: (1) a single Gaussian, (2) a product distribution on the hypercube $\{0,1\}^d$, (3) mixtures of two product distributions (under a natural balancedness condition), and (4) mixtures of $k$ Gaussians with spherical covariances. Prior to our work, all known efficient algorithms (e.g., [LT15, BD15]) for these classes required the error guarantee, $f(\varepsilon, d)$, to depend *polynomially* on the dimension $d$. Hence, previous efficient estimators could only tolerate at most a $1/\operatorname{poly}(d)$-fraction of errors. In this work, we obtain *the first* efficient algorithms for the aforementioned problems, where $f(\varepsilon, d)$ is *completely independent of* $d$ and depends polynomially (often, nearly linearly) on the fraction $\varepsilon$ of corrupted samples. Our work is just a first step in this direction, and there are many exciting questions left to explore.

**1.2. Our techniques.** All of our algorithms are based on a common recipe. The first question to address is the following: Even if we were given a candidate hypothesis $P'$, how could we test if it is $\varepsilon$-close in total variation distance to $P$? The usual way to certify closeness is to exhibit a coupling between $P$ and $P'$ that marginally samples from both distributions, where the samples produced from each agree with probability $1-\varepsilon$. However, we have no control over the process by which samples are generated from $P$, in order to produce such a coupling. And even then, the way that an adversary decides to corrupt samples can introduce complex statistical dependencies.

We circumvent this issue by working with an appropriate notion of parameter distance, which we use as a proxy for the total variation distance between two distributions in the class $\mathcal{D}$. Various notions of parameter distance underly several efficient algorithms for distribution learning in the following sense. If $\theta$ and $\theta'$ are two sets of parameters that define distributions $P_\theta$ and $P_{\theta'}$ in a given class $\mathcal{D}$, a learning algorithm often relies on establishing the following type of relation[1] between $d_{\mathrm{TV}}(P_\theta, P_{\theta'})$ and the parameter distance $d_p(\theta, \theta')$:

$$\operatorname{poly}(d_p(\theta, \theta'), 1/d) \le d_{\mathrm{TV}}(P_\theta, P_{\theta'}) \le \operatorname{poly}(d_p(\theta, \theta'), d)\ . \tag{1}$$

Unfortunately, in our agnostic setting, we cannot afford for (1) to depend on the dimension $d$ at all. Any such dependence would appear in the error guarantee of our algorithm. Instead, the starting point of our algorithms is a notion of parameter distance that satisfies

$$\operatorname{poly}(d_p(\theta, \theta')) \le d_{\mathrm{TV}}(P_\theta, P_{\theta'}) \le \operatorname{poly}(d_p(\theta, \theta'))\ , \tag{2}$$

which allows us to reformulate our goal of designing robust estimators, with distribution-independent error guarantees, as the goal of robustly estimating $\theta$ according to $d_p$. In several settings, the choice of the parameter distance is rather straightforward. It is often the case that some variant of the $\ell_2$-distance between the parameters works.[2]

Given our notion of parameter distance satisfying (2), our main ingredient is an efficient method for robustly estimating the parameters. We provide two algorithmic approaches, which are based on similar principles. Our first approach is faster, requiring only approximate eigenvalue computations. Our second approach relies on convex programming and achieves slightly better sample complexity, in some cases matching

[1] For example, the work of Kalai, Moitra, and Valiant [KMV10] can be reformulated as showing that for any pair of mixtures of two Gaussians (with suitably bounded parameters), the following quantities are polynomially related: (1) discrepancy in their low-order moments, (2) their parameter distance, and (3) their total variation distance. This ensures that any candidate set of parameters that produce almost identical moments must itself result in a distribution that is close in total variation distance.

[2] This discussion already points to why it may be challenging to design agnostic algorithms for mixtures of arbitrary Gaussians or arbitrary product distributions: It is not clear what notion of parameter distance is polynomially related to the total variation distance between two such mixtures, without any dependence on $d$.

the information-theoretic limit. Notably, either approach can be used to give all of our concrete learning applications with nearly identical sample complexity and error guarantees. In what follows, we specialize to the problem of robustly learning the mean $\mu$ of a Gaussian whose covariance is promised to be the identity, which we will use to illustrate how both approaches operate. We emphasize that what is needed to learn the parameters in more general settings requires many additional ideas.

Our first algorithmic approach is an iterative greedy method that, in each iteration, filters out some of the corrupted samples. Given a set of samples $S'$ that contains a set $S$ of uncorrupted samples, an iteration of our algorithm either returns the sample mean of $S'$ or finds a *filter* that allows us to efficiently compute a set $S'' \subset S'$ that is much closer to $S$. Note the sample mean $\widehat{\mu} = \sum_{i=1}^{N}(1/N)X_i$ (even after we remove points that are obviously outliers) can be $\Omega(\varepsilon\sqrt{d})$-far from the true mean in $\ell_2$-distance. The filter approach shows that either the sample mean is already a good estimate for $\mu$ or else there is an elementary spectral test that rejects some of the corrupted points and almost none of the uncorrupted ones. The crucial observation is that if a small number of corrupted points are responsible for a large change in the sample mean, it must be the case that many of the error points are very far from the mean in some particular direction. Thus, we obtain our filter by computing the top absolute eigenvalue of a modified sample covariance matrix.

Our second algorithmic approach relies on convex programming. Here, instead of rejecting corrupted samples, we compute appropriate *weights* $w_i$ for the samples $X_i$, such that the weighted empirical average $\widehat{\mu}_w = \sum_{i=1}^{N} w_i X_i$ is close to $\mu$. We work with the convex set

$$\mathcal{C}_\delta = \left\{ w_i \mid 0 \le w_i \le 1/((1-\varepsilon)N), \sum_{i=1}^{N} w_i = 1, \left\| \sum_{i=1}^{N} w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 \le \delta \right\}.$$

We prove that *any* set of weights in $\mathcal{C}_\delta$ yields a good estimate $\widehat{\mu}_w = \sum_{i=1}^{N} w_i X_i$ in the obvious way. The catch is that the set $\mathcal{C}_\delta$ is defined based on $\mu$, *which is unknown*. Nevertheless, it turns out that we can use the same type of spectral arguments that underlie the filtering approach to design an approximate separation oracle for $\mathcal{C}_\delta$. Combined with standard results in convex optimization, this yields an algorithm for robustly estimating $\mu$.

The third and final ingredient is some new concentration bounds. In both of the approaches above, at best we are hoping that we can remove all of the corrupted points and be left with only the uncorrupted ones, and then use standard estimators (e.g., the empirical average) on them. However, an adversary could have removed an $\varepsilon$-fraction of the samples in a way that biases the empirical average of the remaining uncorrupted samples. What we need are concentration bounds that show for sufficiently large $N$, for samples $X_1, X_2, \ldots, X_N$ from a Gaussian with mean $\mu$ and identity covariance, that every set of $(1-\varepsilon)N$ samples produces a good estimate for $\mu$. In some cases, we can derive such concentration bounds by appealing to known concentration inequalities and taking a union bound. However, in other cases (e.g., concentration bounds for degree-two polynomials of Gaussian random variables) the existing concentration bounds are not strong enough, and we need other arguments to prove that every set of $(1-\varepsilon)N$ samples produces a good estimate.

**1.3. Our results.** We give the first efficient algorithms for agnostically learning several important distribution classes with dimension-independent error guarantees. Our first main result is for a single arbitrary Gaussian with mean $\mu$ and covariance $\Sigma$, which we denote by $\mathcal{N}(\mu, \Sigma)$. In the previous subsection, we described our convex programming approach for learning the mean vector when the covariance is promised to be the identity. A technically more involved version of the technique can handle the case of zero mean and unknown covariance. More specifically, consider the following convex set, where $\Sigma$ is the unknown covariance matrix and $\|\cdot\|_F$ is the Frobenius norm:

$$\mathcal{C}_\delta = \Bigg\{ w_i \mid 0 \le w_i \le 1/((1-\varepsilon)N), \sum_{i=1}^N w_i = 1, \\ \left\| \Sigma^{-1/2} \left( \sum_{i=1}^N w_i X_i X_i^T \right) \Sigma^{-1/2} - I \right\|_F \le \delta \Bigg\} .$$

We design an approximate separation oracle for this unknown convex set, by analyzing the spectral properties of the fourth moment tensor of a Gaussian. Combining these two intermediate results, we obtain our first main result (below). Throughout this paper, we will abuse notation and write $N \ge \widetilde{\Omega}(f(d,\varepsilon,\tau))$ when referring to our sample complexity, to signify that our algorithm works if $N \ge Cf(d,\varepsilon,\tau)\mathrm{polylog}(f(d,\varepsilon,\tau))$ for a large enough universal constant $C$.

THEOREM 1.2. *Let $\mu, \Sigma$ be arbitrary and unknown, and let $\varepsilon, \tau > 0$. There is a polynomial-time algorithm which, given $\varepsilon, \tau$, and an $\varepsilon$-corrupted set of $N$ samples from $\mathcal{N}(\mu,\Sigma)$ with $N \ge \widetilde{\Omega}\left(\frac{d^2 \log^5(1/\tau)}{\varepsilon^2}\right)$, produces $\widehat{\mu}$ and $\widehat{\Sigma}$ such that with probability $1-\tau$ we have $d_{\mathrm{TV}}(\mathcal{N}(\mu,\Sigma), \mathcal{N}(\widehat{\mu},\widehat{\Sigma})) \le O(\varepsilon \log^{3/2}(1/\varepsilon))$.*

We can alternatively establish Theorem 1.2 via our filtering technique. See section 5. In the first version of our paper, our analysis required $N \gtrsim d^3 \log^2(1/\tau)/\varepsilon^2$ samples. In [DKK+17], we showed that a simple adaptation of our algorithm and analysis achieves the improved sample complexity above, which is information-theoretically optimal up to logarithmic factors. We have incorporated this modification (along with the analysis) into this version of the paper, for the sake of completeness.

Our second agnostic learning result is for a product distribution on the hypercube—arguably the most fundamental discrete high-dimensional distribution. We solve this problem using our filter technique, though our convex programming approach would also yield similar results. We start by analyzing the balanced case, when no coordinate is very close to being deterministic. This special case is interesting in its own right and captures the essential ideas of our more involved analysis for the general case. The reason is that, for two balanced product distributions, the $\ell_2$-distance between their means is equivalent to their total variation distance (up to a constant factor). This leads to a clean and elegant presentation of our spectral arguments. For an arbitrary product distribution, we handle the coordinates that are essentially deterministic separately. Moreover, we use the $\chi^2$-distance between the means as the parameter distance and, as a consequence, we need to apply the appropriate corrections to the covariance matrix. Formally, we prove the following theorem.

THEOREM 1.3. *Let $\Pi$ be an unknown binary product distribution, and let $\varepsilon, \tau > 0$. There is a polynomial-time algorithm which, given $\varepsilon, \tau$, and an $\varepsilon$-corrupted set of $N$ samples from $\Pi$ with $N \ge \Omega\left(\frac{d^6 \log(1/\tau)}{\varepsilon^3}\right)$, produces a binary product distribution $\widetilde{\Pi}$ such that with probability $1-\tau$, we have $d_{\mathrm{TV}}(\Pi, \widetilde{\Pi}) \le O(\sqrt{\varepsilon \log(1/\varepsilon)})$.*

For the sake of simplicity in the presentation, we did not make an effort to optimize the sample complexity of our robust estimators in the above setting. We note that methods similar to the analysis of the Gaussian setting can lead to near-optimal sample complexity in this setting as well. We also remark that for the case of balanced binary product distributions, our algorithm achieves an error of $O(\varepsilon\sqrt{\log(1/\varepsilon)})$.

Interestingly enough, the above two distribution classes are trivial to learn in the noiseless case, but in the agnostic setting the learning problem turns out to be surprisingly challenging. Using additional ideas, we are able to generalize our agnostic learning algorithms to *mixtures* of the above classes under some natural conditions. We note that even in the noiseless case, learning mixtures of the above families is nontrivial. First, we study 2-mixtures of $c$-balanced products, which stipulate that the coordinates of the mean vector of each component are in the range $(c, 1-c)$. We prove the following theorem.

THEOREM 1.4 (informal). *Let $\Pi$ be an unknown mixture of two $c$-balanced binary product distributions, and let $\varepsilon, \tau > 0$. There is a polynomial-time algorithm which, given $\varepsilon, \tau$, and an $\varepsilon$-corrupted set of $N$ samples from $\Pi$ with $N \geq \tilde{\Omega}\left(\frac{d^4 \log(1/\tau)}{\varepsilon^{13/6}}\right)$, produces a mixture of two binary product distributions $\widetilde{\Pi}$ such that with probability $1-\tau$, we have $d_{\mathrm{TV}}(\Pi, \widetilde{\Pi}) \leq O_c(\varepsilon^{1/6})$, where the notation $O_c(\cdot)$ suppresses dependence on $c$.*

This generalizes the algorithm of Freund and Mansour [FM99] to the agnostic setting. An interesting open question is to improve the $\varepsilon$-dependence in the above bound to (nearly) linear, or to remove the assumption of balancedness and obtain an agnostic algorithm for mixtures of two arbitrary product distributions.

Finally, we give an agnostic learning algorithm for mixtures of spherical Gaussians.

THEOREM 1.5 (informal). *Let $k$ be a positive integer, and let $\varepsilon, \tau > 0$ be constants. Let $\mathcal{M}$ be a mixture of $k$ Gaussians with spherical covariances. There is a polynomial-time algorithm which, given $\varepsilon, \tau$, and an $\varepsilon$-corrupted set of $N$ samples from $\mathcal{M}$ with $N \geq \mathrm{poly}(k, d, 1/\varepsilon, \log(1/\tau))$, outputs an $\mathcal{M}'$ such that with probability $1-\tau$, we have $d_{\mathrm{TV}}(\mathcal{M}, \mathcal{M}') \leq \widetilde{O}(\mathrm{poly}(k) \cdot \sqrt{\varepsilon})$.*

Our agnostic algorithms for (mixtures of) balanced product distributions and for (mixtures of) spherical Gaussians are conceptually related, since in both cases the goal is to robustly learn the means of each component with respect to $\ell_2$-distance.

In total, these results give new robust and computationally efficient estimators for several well-studied distribution learning problems that can tolerate a constant fraction of errors independent of the dimension. This points to an interesting new direction of making robust statistics algorithmic. The general recipe we have developed here gives us reason to be optimistic about many other problems in this domain.

**1.4. Discussion and related work.** Our results fit in the framework of *density estimation* and *parameter learning* which are both classical problems in statistics with a rich history (see, e.g., [BBBB72, DG85, Sil86, Sco92, DL01]). While these problems have been studied for several decades by different communities, the computational complexity of learning is still not well understood, even for some surprisingly simple distribution families. Most textbook estimators are hard to compute in general, especially in high-dimensional settings. In the past few decades, a rich body of work within theoretical computer science has focused on designing computationally efficient distribution learning algorithms. In a seminal work, Kearns et al. [KMR+94] initiated a systematic investigation of the computational complexity of distribution learning. Since then, efficient learning algorithms have been developed for a wide range of distributions in both low- and high-dimensions [Das99, FM99, AK01, VW02, CGG02, MR05, BV08, KMV10, MV10, BS10, DDS12, CDSS13, DDO+13, CDSS14a, CDSS14b, HP15, ADLS17, DDS15b, DDKT16, DKS16b, DKS16a].

We will be particularly interested in efficient learning algorithms for mixtures of high-dimensional Gaussians and mixtures of product distributions, as this is the focus of our algorithmic results in the agnostic setting. In a pioneering work, Dasgupta [Das99] introduced the problem of parameter estimation of a Gaussian mixture to theoretical computer science, and gave the first provably efficient algorithms under the assumption that the components are suitably well-separated. Subsequently, a number of works improved these separation conditions [AK01, VW02, BV08] and ultimately removed them entirely [KMV10, MV10, BS10]. In another line of work, Freund and Mansour [FM99] gave the first polynomial-time algorithm for properly learning mixtures of two binary product distributions. This algorithm was substantially generalized to phylogenetic trees [CGG02] and to mixtures of any constant number of discrete product distributions [FOS08]. Given the vast body of work on high-dimensional distribution learning, there are a plethora of problems where one could hope to reconcile robustness and computational efficiency. Thus far, the only setting where robust and efficient algorithms are known is in one-dimensional distribution families, where brute-force search or some form of polynomial regression often works. In contrast, essentially nothing is known about efficient agnostic distribution learning in the high-dimensional setting that we study here.

Question 1.1 also resembles learning in the presence of malicious errors [Val85, KL93]. There, an algorithm is given samples from a distribution along with their labels according to an unknown target function. The adversary is allowed to corrupt an $\varepsilon$-fraction of both the samples and their labels. A sequence of works studied the problem of learning a homogeneous halfspace with malicious noise in the setting where the underlying distribution is a Gaussian [Ser01, Ser03, KLS09], culminating in the work of Awasthi, Balcan, and Long [ABL17], who gave an efficient algorithm that finds a halfspace with agreement $O(\varepsilon)$. There is no direct connection between their problem and ours, especially since one is a supervised learning problem and the other is unsupervised. We note, however, that there is an interesting technical parallel in that the work [KLS09] also uses spectral methods to detect outliers. Both their work and our algorithm for agnostically learning the mean are based on the intuition that an adversary can only substantially bias the empirical mean if the corruptions are correlated along some direction. More specifically, the authors of [KLS09] produce a "hard" filter which leads to errors that scale logarithmically with the dimension, even in a weaker corruption model than ours. Our algorithms need to handle many significant conceptual and technical complications that arise when working with higher moments or other distribution families.

Another connection is to the work on robust principal component analysis (PCA). PCA is a transformation that (among other things) is often justified as being able to find the affine transformation $Y = \Sigma^{-1/2}(X - \mu)$ that would place a collection of Gaussian random variables in isotropic position. One can think of our results on agnostically learning a Gaussian as a type of robust PCA that tolerates gross corruptions, where entire samples are corrupted. This is different from other variants of the problem where random sets of coordinates of the points are corrupted [CLMW11], or where the uncorrupted points were assumed to lie in a low-dimensional subspace to begin with [ZL14, LMTZ15]. Finally, Brubaker [Bru09] studied the problem of clustering samples from a *well-separated* mixture of Gaussians in the presence of adversarial noise. The goal of [Bru09] was to separate the Gaussian components from each other, while the adversarial points are allowed to end up in any of clusters. Our work is orthogonal to [Bru09], since even if such a clustering is given, the problem still remains to estimate the parameters of each component.

**1.5. Concurrent and subsequent work.** In concurrent and independent work, Lai, Rao, and Vempala [LRV16] also study high-dimensional agnostic learning. Their results were shown to apply for more general types of distributions, but our guarantees are stronger when learning a Gaussian. Our results are qualitatively similar when the mean is unknown and the covariance is promised to be the identity. But when the covariance is also unknown, their algorithm estimates the mean and covariance to within error $O(\sqrt{\varepsilon\|\Sigma\|_2 \log d})$ and $O(\sqrt{\varepsilon \log d}\|\Sigma\|_2)$, measured in $\ell_2$-norm and Frobenius norm, respectively. However, such guarantees do not directly imply bounds on the total variation distance (which is our main focus), because one needs to estimate the parameters with respect to Mahalanobis distance. In contrast, by virtue of being close in total variation distance, our estimates for the mean and covariance are within $\widetilde{O}(\varepsilon\sqrt{\|\Sigma\|_2})$ and $\widetilde{O}(\varepsilon\|\Sigma\|_2)$ of the true values, again measured in $\ell_2$ norm and Frobenius norm, respectively. An interesting open question is to bridge these two works—what are the most general families of distributions for which one can obtain nearly optimal agnostic learning guarantees?

After the initial publication of our results [DKK+16], there has been a flurry of recent work on robust high-dimensional estimation. Diakonikolas, Kane, and Stewart [DKS16c] studied the problem of learning the parameters of a graphical model in the presence of noise, when given its graph theoretic structure. Charikar, Steinhardt, and Valiant [CSV17] developed algorithms that can tolerate a fraction of corruptions greater than a half, under the weaker goal of outputting a small list of candidate hypotheses that contains a parameter set close to the true values. Balakrishnan, Du, Li, and Singh (see [Li17, DBS17, BDLS17]) studied sparse mean and covariance estimation in the presence of noise obtaining computationally efficient robust algorithms with sample complexity sublinear in the dimension. Diakonikolas, Kane, and Stewart [DKS17] proved statistical query lower bounds providing evidence that the error guarantees of our robust mean and covariance estimation algorithms are best possible, within constant factors, for efficient algorithms. In a subsequent paper [DKK+17], we obtained improved bounds on the sample complexity of our algorithms, which are optimal up to polylogarithmic factors. For the sake of completeness, we include these improved sample bounds in the present version of this paper. In the same work [DKK+17], we showed that our algorithmic approach easily extends to obtain dimension-independent robustness guarantees under much weaker distributional assumptions, and gave a practical demonstration of the efficacy of our robust algorithms on both real and synthetic data.

Since the initial submission of the journal version of this paper, there has been a substantial amount of work on robust high-dimensional estimation in a variety of settings. Diakonikolas, Kane, and Stewart [DKS18a] studied probably approximately correct (PAC) learning of geometric concept classes (including low-degree polynomial threshold functions and intersections of halfspaces) in the same corruption model as ours, obtaining the first dimension-independent error guarantees for these classes. Steinhardt, Charikar, and Valiant [SCV18] focused on deterministic conditions of a dataset which allow robust estimation to be possible. In our initial publication, we gave explicit deterministic conditions in various settings; by focusing directly on this goal, [SCV18] somewhat relaxed some of these assumptions. Meister and Valiant [MV17] studied learning in a crowdsourcing model, where the fraction of honest workers may be very small (similar to [CSV17]). Qiao and Valiant [QV18] considered robust estimation of discrete distributions in a setting where we have several sources (a fraction of which are adversarial) who each provide a batch of samples. A number of simultaneous works [KSS18, HL18, DKS18b] investigated robust mean estimation in even more general settings, and we apply their techniques to learning mixtures of spherical Gaussians under minimal separation conditions. Finally, several concurrent results studied robustness in supervised learning tasks [PSBR18, KKM18, DKK+18], including regression and support vector machine (SVM) problems. Despite all of this rapid progress, there are still many interesting theoretical and practical questions left to explore.

**1.6. Organization.** The structure of this paper is as follows: In section 2, we introduce basic notation and a number of useful facts that will be required throughout the paper, as well as the formal definition of our adversary model. In section 3, we discuss several natural approaches to high-dimensional agnostic learning, all of which lose polynomial factors that depend on the dimension, in terms of their error guarantee.

The main body of the paper is in sections 4–8. Sections 4 and 6 illustrate our convex programming framework, while sections 5, 7, and 8 illustrate our filter framework. More specifically, in sections 4 and 5, we analyze the setting of a single Gaussian with unknown mean and unknown covariance, using our convex programming and filter frameworks, respectively. In section 6, we generalize the convex programming method to obtain an agnostic algorithm for mixtures of spherical Gaussians with unknown means. In section 7, we apply our filter techniques to a binary product distribution, and generalize these in section 8 to obtain an agnostic learning algorithm for a mixture of two balanced binary product distributions.

We note that for some of the more advanced applications of our frameworks, the technical details can get in the way of the fundamental ideas. For the reader who is interested in seeing the details of our most basic application of the convex programming framework, we recommend reading the case of a Gaussian with unknown mean in section 4.3. Similarly, for the filter framework, we suggest either the Gaussian with unknown mean in section 5.1 or the balanced product distribution in section 7.1.

## 2. Preliminaries.

**2.1. Basic notation.** Throughout this paper, if $v$ is a vector, we will let $\|v\|_2$ denote its Euclidean norm. If $M$ is a matrix, we will let $\|M\|_2$ denote its spectral norm, and $\|M\|_F$ denote its Frobenius norm. We will also let $\preceq$ and $\succeq$ denote the positive semidefinite (PSD) ordering on matrices. For a discrete distribution $P$, we will denote by $P(x)$ the probability mass at point $x$. For a continuous distribution, let it denote the probability density function at $x$. Let $S$ be a multiset over $\{0,1\}^d$ . We will write $X \in_u S$ to denote that $X$ is drawn from the empirical distribution defined by $S$. Throughout the paper, we let $\otimes$ denote the Kronecker product of matrices.

As a measure of distance between distributions, we will use the notion of total variation distance.

Definition 2.1. *Let $P, Q$ be two probability distributions on $\mathbb{R}^d$. Then the* total variation distance *between $P$ and $Q$, denoted $d_{\mathrm{TV}}(P,Q)$, is defined as*

$$d_{\mathrm{TV}}(P,Q) = \sup_{A \subseteq \mathbb{R}^d} |P(A) - Q(A)| \,.$$

**2.2. Types of adversaries.** In this paper, we will consider a powerful model for agnostic distribution learning that generalizes many other existing models. The standard setup involves an *oblivious adversary* who chooses a distribution that is close in total variation distance to an unknown distribution in some class $\mathcal{D}$.

Definition 2.2. *Given $\varepsilon > 0$ and a class of distributions $\mathcal{D}$, the* oblivious adversary *chooses a distribution $P$ such that there is an unknown distribution $D \in \mathcal{D}$ with $d_{\mathrm{TV}}(P, D) \leq \varepsilon$. An algorithm is then given $m$ independent samples $X_1, X_2, \ldots, X_m$ from $P$.*

The goal of the algorithm is to return the parameters of a distribution $\widehat{D}$ in $\mathcal{D}$, where $d_{\mathrm{TV}}(D, \widehat{D})$ is small. We refer to the above adversary as oblivious because it fixes the model for noise before seeing any of the samples. In contrast, a more powerful adversary is allowed to inspect the samples before corrupting them, both by adding corrupted points and deleting uncorrupted points. We refer to this as the *full adversary.*

Definition 2.3. *Given $\varepsilon > 0$ and a class of distributions $\mathcal{D}$, the* full adversary *operates as follows: The algorithm specifies some number of samples $m$. The adversary generates $m$ samples $X_1, X_2, \ldots, X_m$ from some (unknown) distribution $D \in \mathcal{D}$. It then draws $m'$ from an appropriate distribution. This distribution is allowed to depend on $X_1, X_2, \ldots, X_m$, but when marginalized over the $m$ samples satisfies $m' \sim Bin(m, \varepsilon)$. The adversary is allowed to inspect the samples, removes $m'$ of them, and replaces them with arbitrary points. The set of $m$ points is given (in any order) to the algorithm.*

We remark that there are no computational restrictions on the adversary. As before, the goal is to return the parameters of a distribution $\widehat{D}$ in $\mathcal{D}$, where $d_{\mathrm{TV}}(D, \widehat{D})$ is small. The reason we allow the draw $m'$ to depend on the samples $X_1, X_2, \ldots, X_m$ is because our algorithms will tolerate this extra generality, and it will allow us to show that the full adversary is at least as strong as the oblivious adversary (this would not necessarily be true if $m'$ were sampled independently from $\mathrm{Bin}(m, \varepsilon)$).

We rely on the following well-known fact.

Fact 2.4. *Let $P, D$ be two distributions such that $d_{\mathrm{TV}}(P, D) = \varepsilon$. Then there are distributions $N_1$ and $N_2$ such that $(1 - \varepsilon_1)P + \varepsilon_1 N_1 = (1 - \varepsilon_2)D + \varepsilon_2 N_2$, where $\varepsilon_1 + \varepsilon_2 = \varepsilon$.*

Now we can describe how the full adversary can corrupt samples from $D$ to get samples distributed according to $P$.

Claim 2.5. *The full adversary can simulate any oblivious adversary.*

*Proof.* We draw $m$ samples $X_1, X_2, \ldots, X_m$ from $D$. We delete each sample $X_i$ independently with probability $\varepsilon_2$ and replace it with an independent sample from $N_2$. This gives a set of samples $Y_1, Y_2, \ldots, Y_m$ that are independently sampled from $(1 - \varepsilon_2)D + \varepsilon_2 N_2$. Since the distributions $(1 - \varepsilon_1)P + \varepsilon_1 N_1$ and $(1 - \varepsilon_2)D + \varepsilon_2 N_2$ are identical, we can couple them to independent samples $Z_1, Z_2, \ldots, Z_m$ from $(1 - \varepsilon_1)P + \varepsilon_1 N_1$. Now we can delete and replace each sample $Z_i$ that came from $N_1$ with an independent sample from $P$. The result is a set of samples that are independently sampled from $P$ where we have made $m'$ edits and marginally $m' \sim \mathrm{Bin}(m, \varepsilon_1 + \varepsilon_2)$, although $m'$ has and needs to have some dependence on the original samples from $D$. ∎

The challenge in working with the full adversary is that even the samples that came from $D$ can have biases. The adversary can now choose how to remove uncorrupted points in a careful way so as to compensate for certain other biases that he introduces using the corrupted points.

Throughout this paper, we will make use of the following notation and terminology.

Definition 2.6. *We say a set of samples $X_1, X_2, \dots, X_m$ is an* $\varepsilon$-corrupted *set of samples generated by the oblivious (resp., full) adversary if it is generated by the process described above in the definition of the oblivious (resp., full) adversary. If it was generated by the full adversary, we let $G \subseteq [m]$ denote the indices of the uncorrupted samples, and we let $E \subseteq [m]$ denote the indices of the corrupted samples.*

In this paper, we will give a number of algorithms for agnostic distribution learning that work in the full adversary model. In our analysis, we will identify a set of events that ensure the algorithm succeeds and will bound the probability that any of these events does not occur when $m$ is suitably large. We will often explicitly invoke the assumption that $|E| \leq 2\varepsilon m$. We can do this even though the number of points that are corrupted is itself a random variable, because by the Chernoff bound, as long as $m \geq O\left(\frac{\log 1/\tau}{\varepsilon}\right)$, we know that $|E| \leq 2\varepsilon m$ holds with probability at least $1 - O(\tau)$. Thus, making the assumption that $|E| \leq 2\varepsilon m$ costs us an additional additive $O(\tau)$ term in our union bound, when bounding the failure probability of our algorithms.

**2.3. Distributions of interest.** One object of study in this paper is the Gaussian (or normal) distribution.

Definition 2.7. *A* Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ *with mean $\mu$ and covariance $\Sigma$ is the distribution with probability density function*

$$f(x) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right).$$

We will also be interested in binary product distributions.

Definition 2.8. *A* (binary) product distribution *is a probability distribution over $\{0,1\}^d$ whose coordinate random variables are independent. Note that a binary product distribution is completely determined by its mean vector.*

We will also be interested in mixtures of such distributions.

Definition 2.9. *A* mixture *$P$ of distributions $P_1, \dots, P_k$ with mixing weights $\alpha_1, \dots, \alpha_k$ is the distribution defined by*

$$P(x) = \sum_{j \in [k]} \alpha_j P_k(x),$$

*where $\alpha_j \geq 0$ for all $j$ and $\sum_{j \in [k]} \alpha_j = 1$.*

**2.4. Bounds on TV distance.** The *Kullback–Leibler (KL) divergence* (also known as *relative entropy*, *information gain*, or *information divergence*) is a well-known measure of distance between two distributions.

Definition 2.10. *Let $P, Q$ be two probability distributions on $\mathbb{R}^d$. Then the* KL divergence *between $P$ and $Q$, denoted $d_{\mathrm{KL}}(P \| Q)$, is defined as*

$$d_{\mathrm{KL}}(P \| Q) = \int_{\mathbb{R}^d} \log \frac{dP}{dQ} dP .$$

The primary interest we have in this quantity is the fact that (1) the KL divergence between two Gaussians has a closed form expression, and (2) it can be related (often with little loss) to the total variation distance between the Gaussians. The first statement is expressed in the fact below.

FACT 2.11. *Let* $\mathcal{N}(\mu_1, \Sigma_1)$ *and* $\mathcal{N}(\mu_2, \Sigma_2)$ *be two Gaussians such that* $\det(\Sigma_1)$, $\det(\Sigma_2) \neq 0$. *Then*

$$(3)\qquad \begin{aligned} &d_{\mathrm{KL}}\left(\mathcal{N}(\mu_1, \Sigma_1) \| \mathcal{N}(\mu_2, \Sigma_2)\right) \\ &= \frac{1}{2}\left(\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1) - d - \ln\left(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}\right)\right). \end{aligned}$$

The second statement is encapsulated in the well-known *Pinsker's inequality.*

THEOREM 2.12 (Pinsker's inequality). *Let* $P, Q$ *be two probability distributions over* $\mathbb{R}^d$. *Then*

$$d_{\mathrm{TV}}(P, Q) \leq \sqrt{\frac{1}{2} d_{\mathrm{KL}}(P \| Q)}\ .$$

With this we can show the following two useful corollaries, which allow us to relate parameter distance between two Gaussians to their total variation distance. The first corollary bounds the total variation distance between two Gaussians with identity covariance in terms of the Euclidean distance between the means.

COROLLARY 2.13. *Let* $\mu_1, \mu_2 \in \mathbb{R}^d$ *be arbitrary. Then* $d_{\mathrm{TV}}\left(\mathcal{N}(\mu_1, I), \mathcal{N}(\mu_2, I)\right) \leq \frac{1}{\sqrt{2}}\|\mu_2 - \mu_1\|_2$.

*Proof.* In the case where $\Sigma_1 = \Sigma_2 = I$, (3) simplifies to

$$d_{\mathrm{KL}}\left(\mathcal{N}(\mu_1, I) \| \mathcal{N}(\mu_2, I)\right) = \frac{1}{2}\|\mu_2 - \mu_1\|_2^2.$$

Pinsker's inequality (Theorem 2.12) then implies that

$$d_{\mathrm{TV}}\left(\mathcal{N}(\mu_1, I), \mathcal{N}(\mu_2, I)\right) \leq \sqrt{\frac{1}{2} d_{\mathrm{KL}}\left(\mathcal{N}(\mu_1, I) \| \mathcal{N}(\mu_2, I)\right)} = \frac{1}{\sqrt{2}}\|\mu_2 - \mu_1\|_2,$$

as desired. ∎

The second corollary bounds the total variation distance between two mean 0 Gaussians in terms of the Frobenius norm of the difference between their covariance matrices.

COROLLARY 2.14. *Let* $\delta > 0$ *be sufficiently small. Let* $\Sigma_1, \Sigma_2$ *such that* $\|I - \Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2}\|_F = \delta$. *Then,*

$$d_{\mathrm{TV}}(\mathcal{N}(0, \Sigma_1) || \mathcal{N}(0, \Sigma_2)) \leq O(\delta)\ .$$

*Proof.* Let $M = \Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2}$. Then (3) simplifies to

$$d_{\mathrm{KL}}\left(\mathcal{N}(\mu_1, \Sigma_1) \| \mathcal{N}(\mu_2, \Sigma_2)\right) = \frac{1}{2}\left(\mathrm{tr}(M) - d - \ln\det(M)\right)\ .$$

Since both terms in the last line are rotationally invariant, we may assume without loss of generality that $M$ is diagonal. Let $M = \mathrm{diag}(1 + \lambda_1, \ldots, 1 + \lambda_d)$. Thus, the KL divergence between the two distributions is given exactly by $\frac{1}{2}\sum_{i=1}^d (\lambda_i - \log(1 + \lambda_i))$, where we are guaranteed that $(\sum_{i=1}^d \lambda_i^2)^{1/2} = \delta$. By the second order Taylor approximation to $\ln(1 + x)$, for $x$ small we have that for $\delta$ sufficiently small,

$$\sum_{i=1}^d \lambda_i - \log(1 + \lambda_i) = \Theta\left(\sum_{i=1}^d \lambda_i^2\right) = \Theta(\delta^2)\ .$$

Thus, we have shown that for $\delta$ sufficiently small, $d_{\mathrm{KL}}\left(\mathcal{N}(\mu_1,\Sigma_1)\|\mathcal{N}(\mu_2,\Sigma_2)\right) \leq O(\delta^2)$. The result now follows by an application of Pinsker's inequality (Theorem 2.12). ∎

Our algorithm for agnostically learning an arbitrary Gaussian will be based on solving two intermediate problems: (1) We are given samples from $\mathcal{N}(\mu, I)$ and our goal is to learn $\mu$. (2) We are given samples from $\mathcal{N}(0,\Sigma)$ and our goal is to learn $\Sigma$. The above bounds on total variation distance will allow us to conclude that our estimate is close in total variation distance to the unknown Gaussian distribution in each of the two settings.

We note the following folklore sample complexity bounds for learning a Gaussian in the nonagnostic setting.

THEOREM 2.15. $N = \Theta\left(\frac{d+\log(1/\tau)}{\varepsilon^2}\right)$ *samples are both necessary and sufficient to learn a d-dimensional Gaussian with unknown mean and known covariance to total variation distance* $\varepsilon$ *with probability* $1-\tau$.

THEOREM 2.16. $N = \Theta\left(\frac{d^2+\log(1/\tau)}{\varepsilon^2}\right)$ *samples are both necessary and sufficient to learn a d-dimensional Gaussian with unknown mean and covariance to total variation distance* $\varepsilon$ *with probability* $1-\tau$.

We will also need the following lemma bounding the total variation distance between two product distributions.

LEMMA 2.17. *Let* $P, Q$ *be binary product distributions with mean vectors* $p, q \in (0,1)^d$. *We have that*

$$d_{\mathrm{TV}}^2(P,Q) \leq 2\sum_{i=1}^{d} \frac{(p_i-q_i)^2}{(p_i+q_i)(2-p_i-q_i)} .$$

*Proof.* We include the simple proof for completeness. By Kraft's inequality (see, e.g., Theorem 5.2.1 in [CT06]), for any pair of distributions, we have that $d_{\mathrm{TV}}^2(P,Q) \leq 2H^2(P,Q)$, where $H(P,Q)$ denotes the Hellinger distance between $P, Q$. Since $P, Q$ are product measures, we have that

$$1 - H^2(P,Q) = \prod_{i=1}^{d}(1-H^2(P_i,Q_i)) = \prod_{i=1}^{d}\left(\sqrt{p_iq_i} + \sqrt{(1-p_i)(1-q_i)}\right) .$$

The elementary inequality $2\sqrt{ab} = a+b-(\sqrt{a}-\sqrt{b})^2$, $a,b>0$, gives that

$$\sqrt{p_iq_i} + \sqrt{(1-p_i)(1-q_i)} \geq 1 - \frac{(p_i-q_i)^2}{(p_i+q_i)(2-p_i-q_i)} .$$

Let

$$z_i = \frac{(p_i-q_i)^2}{(p_i+q_i)(2-p_i-q_i)} .$$

We have

$$d_{\mathrm{TV}}^2(P,Q) \leq 2\cdot\left(1-\prod_{i=1}^{d}(1-z_i)\right) \leq 2\sum_{i=1}^{d} z_i ,$$

where the last inequality follows from the union bound. ∎

**2.5. Additional concentration lemmas.** In this section, we list a number of standard concentration inequalities for nice random variables which we will frequently use throughout this paper. The proofs of these results are standard and omitted; see, e.g., [Ver10] for a more thorough treatment of these results. The first is a Chernoff bound for bounded random variables.

THEOREM 2.18. *Let $Z_1, \ldots, Z_d$ be independent random variables with $Z_i$ supported on $[a_i, b_i]$. Let $Z = \sum_{i=1}^d Z_i$. Then for any $T > 0$,*

$$\Pr(|Z - \mathbb{E}[Z]| > T) \leq 2 \exp\left( \frac{-2T^2}{\sum_{i=1}^d (b_i - a_i)^2} \right).$$

We will also require the following tail bounds for Gaussians and quadratic forms of Gaussians.

LEMMA 2.19. *Let $n$ be a positive integer. Let $D$ be a sub-Gaussian distribution with mean $0$ and covariance $I$. Let $Y_i \sim D$ be independent, for $i = 1, \ldots, n$. Let $v \in \mathbb{R}^d$ be an arbitrary unit vector. Then, there exists a universal constant $B > 0$ so that for all $T > 0$, we have*

$$\Pr\left[ \left| \frac{1}{n} \sum_{i=1}^n \langle v, Y_i \rangle \right| > T \right] \leq 4 \exp\left( -BnT^2 \right) .$$

LEMMA 2.20 (Hanson–Wright). *Let $n$ be a positive integer. Let $D$ be a sub-Gaussian distribution with mean $0$ and covariance $\Sigma \preceq I$. Let $Y_i \sim D$ be independent, for $i = 1, \ldots, n$. Let $U \in \mathbb{R}^{d \times d}$ satisfy $U \succeq 0$ and $\|U\|_F = 1$. Then, there exists a universal constant $B > 0$ so that for all $T > 0$, we have*

$$\Pr\left[ \left| \frac{1}{n} \sum_{i=1}^n \mathrm{tr}(X_i X_i^\top U) - \mathrm{tr}(U) \right| > T \right] \leq 4 \exp\left( -Bn \min(T, T^2) \right) .$$

By standard union bound arguments (see, e.g., [Ver10]), we obtain the following concentration results for the empirical mean and covariance of a set of Gaussian vectors.

LEMMA 2.21. *Let $n$ be a positive integer. Let $D$ be a sub-Gaussian distribution with mean $0$ and covariance $I$. Let $Y_i \sim D$ be independent, for $i = 1, \ldots, n$. Then, there exist universal constants $A, B > 0$ so that for all $t > 0$ we have*

$$\Pr\left[ \left\| \frac{1}{n} \sum_{i=1}^n Y_i \right\|_2 > t \right] \leq 4 \exp\left( Ad - Bnt^2 \right) .$$

LEMMA 2.22. *With the same setup as in Lemma 2.21, there exist universal constants $A, B > 0$ so that for all $t > 0$ we have*

$$\Pr\left[ \left\| \frac{1}{n} \sum_{i=1}^n Y_i Y_i^\top - I \right\|_2 > t \right] \leq 4 \exp\left( Ad - Bn \min(t, t^2) \right) .$$

**2.6. Agnostic hypothesis selection.** Several of our algorithms will return a polynomial-sized list of hypotheses at least one of which is guaranteed to be close to the target distribution. Usually (e.g., in a nonagnostic setting), one could use a polynomial number of additional samples to run a tournament to identify the candidate hypothesis that is (roughly) the closest to the target distribution. In the discussion that follows, we will refer to these additional samples as test samples. Such *hypothesis selection* algorithms have been extensively studied [Yat85, DL96, DL97, DL01, DK14, AJOS14, SOAJ14, DDS15a, DDS15b]. Unfortunately, against a strong adversary we run into a serious technical complication: the training samples and test samples are not necessarily independent. Moreover, even if we randomly partition our samples into training and test, a priori there are an unbounded set of possible hypotheses that the training phase could output, and when we analyze the tournament we cannot condition on the list of hypotheses and assume that the test samples are sampled anew. Our approach is to require our original algorithm to return only hypotheses from some finite set of possibilities; we will see this mitigates the problem.

LEMMA 2.23. *Let $\mathcal{C}$ be a class of probability distributions. Suppose that for some $N, \varepsilon, \tau > 0$ there exists a polynomial-time algorithm that, given $N$ independent samples from some $\Pi \in \mathcal{C}$, of which up to a $2\varepsilon$-fraction have been arbitrarily corrupted, returns a list $\mathcal{L}$ of $M$ distributions whose probability density functions are explicitly computable and which can be effectively sampled from such that with $1-\tau/2$ probability there exists a $\Pi' \in \mathcal{L}$ with $d_{\mathrm{TV}}(\Pi', \Pi) < \delta$. Suppose, furthermore, that the distributions returned by this algorithm are all in some fixed set $\mathcal{M}$. Then there exists another polynomial-time algorithm, which, given $O(N + (\log(|\mathcal{M}|) + \log(1/\tau))/\varepsilon^2)$ samples from $\Pi$, an $\varepsilon$-fraction of which have been arbitrarily corrupted, returns a single distribution $\Pi'$ such that with $1-\tau$ probability $d_{\mathrm{TV}}(\Pi', \Pi) < O(\delta + \varepsilon)$.*

*Remark* 2.24. As a simple corollary of the agnostic tournament, observe that this allows us to do agnostic learning without knowing the precise error rate $\varepsilon$. Throughout the paper, we assume the algorithm knows $\varepsilon$, and guarantees that the output will have error which is at most $O(f(\varepsilon))$. However, if the algorithm is not given this information, and instead is given an $\eta$ and asked to return something with error at most $O(f(\varepsilon + \eta))$, we may simply grid over $\{\eta, (1+\gamma)\eta, (1+\gamma)^2\eta, \ldots, 1\}$ (here $\gamma$ is some arbitrary constant that governs a tradeoff between runtime and accuracy), run our algorithm with $\varepsilon$ set to each element in this set, and perform hypothesis selection via TOURNAMENT. Then it is not hard to see that we are guaranteed to output something which has error at most $O(f(\varepsilon + (1+\gamma)\eta))$.

*Proof.* First, we randomly choose a subset of $N$ of our samples and a disjoint subset of $C(\log(|\mathcal{M}|) + \log(1/\tau))/\varepsilon^2$ of our samples for some sufficiently large $C$. Note that with high probability over our randomization, at most a $2\varepsilon$-fraction of samples from each subset are corrupted. Thus, we may instead consider the stronger adversary who sees a set $S_1$ of $N$ independent samples from $\Pi$ and another set, $S_2$, of $C(\log(|\mathcal{M}|) + \log(1/\tau))/\varepsilon^2$ samples from $\Pi$ and can arbitrary corrupt a $2\varepsilon$-fraction of each, giving sets $S_1'$,$S_2'$.

With probability at least $1-\tau/2$ over $S_1$, the original algorithm run on $S_1'$ returns a set $\mathcal{L}$ satisfying the desired properties.

For two distributions $P$ and $Q$ in $\mathcal{M}$ we let $A_{PQ}$ be the set of inputs $x$ where $\Pr_P(x) > \Pr_Q(x)$. We note that we can test membership in $A_{PQ}$ as, by assumption, the probability density functions are computable. We also note that $d_{\mathrm{TV}}(P, Q) = \Pr_P(A_{PQ}) - \Pr_Q(A_{PQ})$. Our tournament will depend on the fact that if $P$ is close to the target and $Q$ is far away, then many samples will necessarily lie in $A_{PQ}$.

We claim that with probability at least $1-\tau/2$ over the choice of $S_2$, we have for any $P, Q \in \mathcal{M}$,

$$\Pr_{x \in_u S_2}(x \in A_{PQ}) = \Pr_{x \sim \Pi}(x \in A_{PQ}) + O(\varepsilon).$$

This follows by Chernoff bounds and a union bound over the $|\mathcal{M}|^2$ possibilities for $P$ and $Q$. Since the total variation distance between the uniform distributions over $S_2$ and $S_2'$ is at most $2\varepsilon$, we also have for $S_2'$ that

$$\Pr_{x\in_u S_2'}(x \in A_{PQ}) = \Pr_{x\sim\Pi}(x \in A_{PQ}) + O(\varepsilon).$$

Suppose that $d_{\mathrm{TV}}(P,\Pi) < \delta$ and $d_{\mathrm{TV}}(Q,\Pi) > 5\delta + C\varepsilon$. We then have that

$$\begin{aligned}\Pr_{x\in_u S_2'}(x \in A_{PQ}) = \Pr_{x\sim\Pi}(x \in A_{PQ}) + O(\varepsilon) &\geq \Pr_{x\sim P}(x \in A_{PQ}) + O(\varepsilon) - \delta \\ &\geq \Pr_{x\sim Q}(x \in A_{PQ}) + \delta + C\varepsilon/5.\end{aligned}$$

On the other hand, if $d_{\mathrm{TV}}(\Pi, Q) < \delta$, then

$$\Pr_{x\in_u S_2'}(x \in A_{PQ}) = \Pr_{x\sim\Pi}(x \in A_{PQ}) + O(\varepsilon) < \Pr_{x\sim Q}(x \in A_{PQ}) + \delta + C\varepsilon/5.$$

Therefore, if we throw away any $Q$ in our list for which there is a $P$ in our list such that

$$\Pr_{x\in_u S_2'}(x \in A_{PQ}) \geq \Pr_{x\sim Q}(x \in A_{PQ}) + \delta + C\varepsilon/5,$$

we have thrown away all the $Q$ with $d_{\mathrm{TV}}(Q,\Pi) > 5\delta + C\varepsilon$, but none of the $Q$ with $d_{\mathrm{TV}}(Q,\Pi) < \delta$. Therefore, there will be a $Q$ remaining, and returning it will yield an appropriate $\Pi'$. ◻

**3. Some natural approaches, and why they fail.** Many of the agnostic distribution learning problems that we study are so natural that one would immediately wonder why simpler approaches do not work. Here we detail some other plausible approaches, and what causes them to lose dimension-dependent factors (if they have any guarantees at all). For the discussion that follows, we note that by Corollary 2.13 in order to achieve an estimate that is $O(\varepsilon)$-close in total variation distance (for a Gaussian when $\mu$ is unknown and $\Sigma = I$), it is necessary and sufficient that $\|\hat{\mu} - \mu\|_2 = O(\varepsilon)$.

**Learn each coordinate separately.** One plausible approach for robust mean estimation in high-dimensions is to agnostically learn along each coordinate separately. For instance, if our goal is to agnostically learn the mean of a Gaussian with known covariance $I$, we could try to learn each coordinate of the mean separately. But since an $\varepsilon$-fraction of the samples are corrupted, our estimate can be off by $\varepsilon$ in each coordinate and would be off by $\varepsilon\sqrt{d}$ in high-dimensions.

**Maximum likelihood.** Given a set of samples $X_1, \ldots, X_N$ and a class of distributions $\mathcal{D}$, the maximum likelihood estimator (MLE) is the distribution $F \in \mathcal{D}$ that maximizes $\prod_{i=1}^N F(X_i)$. Equivalently, $F$ minimizes the negative log likelihood (NLL), which is given by

$$\mathrm{NLL}(F, X_1, \ldots, X_N) = -\sum_{i=1}^N \log F(X_i)\ .$$

In particular, if $\mathcal{D} = \{\mathcal{N}(\mu, I) : \mu \in \mathbb{R}^d\}$ is the set of Gaussians with unknown mean and identity covariance, we see that for any $\mu \in \mathbb{R}^d$, the NLL of the set of samples is given by

$$\begin{aligned}\mathrm{NLL}(\mathcal{N}(\mu, I), X_1, \ldots, X_N) &= -\sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}} e^{-\|X_i-\mu\|_2^2/2}\right) \\ &= N \log\sqrt{2\pi} + \frac{1}{2}\sum_{i=1}^{N} \|X_i - \mu\|_2^2 \,,\end{aligned}$$

and so the $\mu$ which minimizes $\mathrm{NLL}(\mathcal{N}(\mu, I), X_1, \ldots, X_N)$ is the mean of the samples $X_i$, since for any set of vectors $v_1, \ldots, v_N$, the average of the $v_i$'s is the minimizer of the function $h(x) = \sum_{i=1}^{N} \|v_i - x\|_2^2$. Hence, if an adversary places an $\varepsilon$-fraction of the points at some very large distance, then the estimate for the mean would need to move considerably in that direction. By placing the corruptions further and further away, the MLE can be an arbitrarily bad estimate. That is, even though it is well known [Hub67, Whi82] that the MLE converges to the distribution $F \in \mathcal{D}$ that is closest in KL divergence to the distribution from which our samples were generated (i.e., after the adversary has added corruptions), $F$ is not necessarily close to the uncorrupted distribution.

**Geometric median.** In one-dimension, it is well-known that the median provides a provably robust estimate for the mean in a number of settings. The mean of a set of points $a_1, \ldots, a_N$ is the minimizer of the function $f(x) = \sum_{i=1}^{N}(a_i - x)^2$, and in contrast the median is the minimizer of the function $f(x) = \sum_{i=1}^{N} |a_i - x|$. In higher-dimensions, there are many natural definitions for the median that generalize the one-dimensional case. The *Tukey median* is one such notion, but as we discussed it is hard to compute [JP78], and the best-known algorithms run in time exponential in $d$. Motivated by this, the geometric median is another high-dimensional notion of a median. It often achieves better robustness than the mean and can be computed quickly [CLM+16]. The formal definition is

$$\mathrm{geomed}(S) \triangleq \min_{v} \sum_{x \in S} \|x - v\|_2 \,.$$

Unfortunately, this notion of median still incurs an error containing a factor of $O(\sqrt{d})$.

PROPOSITION 3.1 (Proposition 2.1 of [LRV16]). *Given a set $S$ of $N = \Omega\left(\frac{d+\log(1/\tau)}{\varepsilon^2}\right)$ samples from $\mathcal{N}(0, I)$, then with probability at least $1 - \tau$, there exists a corruption $S'$ of $S$, such that*

$$\mathrm{geomed}(S') = \Omega(\varepsilon\sqrt{d}).$$

**4. Agnostically learning a Gaussian, via convex programming.** In this section we give a polynomial-time algorithm to agnostically learn a single Gaussian up to error $\widetilde{O}(\varepsilon)$. Our approach is based on the following ingredients: First, in section 4.1, we define the set $S_{N,\varepsilon}$, which will be a key algorithmic object in our framework. In section 4.2 we give key, new concentration bounds on certain statistics of Gaussians. We will make crucial use of these concentration bounds throughout this section. In section 4.3 we give an algorithm to agnostically learn a Gaussian with unknown mean and whose covariance is promised to be the identity via convex programming. This will be an important subroutine in our overall algorithm, and it also helps to illustrate our algorithmic approach without many of the additional complications that arise in our later applications. In section 4.4 we show how to robustly learn a Gaussian with mean 0 and unknown covariance again via convex programming. Finally, in section 4.5 we show how to combine these two intermediate results to get our overall algorithm.

**4.1. The set $S_{N,\varepsilon}$.** An important algorithmic object for us will be the following set.

DEFINITION 4.1. *For any $\frac{1}{2} > \varepsilon > 0$ and any integer $N$, let*

$$S_{N,\varepsilon} = \left\{ (w_1, \dots, w_N) : \sum_{i=1}^{N} w_i = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-2\varepsilon)N} \ \forall i \right\} .$$

Next, we motivate this definition. For any $J \subseteq [N]$, let $w^J \in \mathbb{R}^N$ be the vector which is given by $w^J_i = \frac{1}{|J|}$ for $i \in J$ and $w^J_i = 0$ otherwise. Then, observe that

$$S_{N,\varepsilon} = \operatorname{conv}\left\{ w^J : |J| = (1-2\varepsilon)N \right\} ,$$

and so we see that this set is designed to capture the notion of selecting a set of $(1-2\varepsilon)N$ samples from $N$ samples.

Given $w \in S_{N,\varepsilon}$ we will use the notation

$$w_g = \sum_{i \in G} w_i \quad \text{and} \quad w_b = \sum_{i \in E} w_i$$

to denote the total weight on good and bad points, respectively. The following facts are immediate from $|E| \le 2\varepsilon N$ and the properties of $S_{N,\varepsilon}$.

FACT 4.2. *If $w \in S_{N,\varepsilon}$ and $|E| \le 2\varepsilon N$, then $w_b \le \frac{2\varepsilon}{1-2\varepsilon}$. Moreover, the renormalized weights $w'$ on good points, given by $w'_i = \frac{w_i}{w_g}$ for all $i \in G$ and $w'_i = 0$ otherwise, satisfy $w' \in S_{N,4\varepsilon}$.*

**4.2. Concentration inequalities.** Throughout this section and in section 6, we will make use of various concentration bounds on low moments of Gaussian random variables. Some are well-known, and others are new but follow from known bounds and appropriate union bound arguments.

**4.2.1. Empirical estimates of first and second moments of large subsets.** We will also be interested in how well various statistics of Gaussians concentrate around their expectation, when we take the worst-case set of weights in $S_{N,\varepsilon}$. This is more subtle than standard settings such as Lemma 2.21 or 2.22 because as we take more samples, any fixed statistic (e.g., taking the uniform distribution over the samples) concentrates better, but the size of $S_{N,\varepsilon}$ (e.g., the number of sets of $(1-2\varepsilon)N$ samples) grows, too. We defer the proofs to Appendix A. The first concerns the behavior of the empirical covariance.

LEMMA 4.3. *Fix $\varepsilon \le 1/2$ and $\tau \le 1$. There is a $\delta_1 = O(\varepsilon \log 1/\varepsilon)$ such that if $Y_1, \dots, Y_N$ are independent samples from $\mathcal{N}(0, I)$ and $N = \Omega\left(\frac{d+\log(1/\tau)}{\delta_1^2}\right)$, then*

$$\Pr\left[ \exists w \in S_{N,\varepsilon} : \left\| \sum_{i=1}^{N} w_i Y_i Y_i^T - I \right\|_2 \ge \delta_1 \right] \le \tau . \tag{4}$$

A nearly identical argument (using Hoeffding instead of Bernstein in the proof of Theorem 5.50 in [Ver10]) yields the following lemma.

LEMMA 4.4. *Fix $\varepsilon$ and $\tau$ as above. There is a $\delta_2 = O(\varepsilon\sqrt{\log 1/\varepsilon})$ such that if $Y_1, \ldots, Y_N$ are independent samples from $\mathcal{N}(0, I)$ and $N = \Omega\left(\frac{d+\log(1/\tau)}{\delta_2^2}\right)$, then*

$$\Pr\left[\exists w \in S_{N,\varepsilon} : \left\|\sum_{i=1}^N w_i Y_i\right\|_2 \geq \delta_2\right] \leq \tau \,. \tag{5}$$

Note that by Cauchy–Schwarz, this implies the following corollary.

COROLLARY 4.5. *Fix $\varepsilon$ and $\tau$ as above. There is a $\delta_2 = O(\varepsilon\sqrt{\log 1/\varepsilon})$ such that if $Y_1, \ldots, Y_N$ are independent samples from $\mathcal{N}(0, I)$ and $N = \Omega\left(\frac{d+\log(1/\tau)}{\delta_2^2}\right)$, then*

$$\Pr\left[\exists v \in \mathbb{R}^d, \exists w \in S_{N,\varepsilon} : \left\|\left(\sum_{i=1}^N w_i Y_i\right) v^T\right\|_2 \geq \delta_2\|v\|_2\right] \leq \tau \,. \tag{6}$$

We will also require the following, well-known concentration, which says that no sample from a Gaussian deviates too far from its mean in $\ell_2$-distance.

FACT 4.6. *Fix $\tau > 0$. Let $X_1, \ldots, X_N \sim \mathcal{N}(0, I)$. Then, with probability $1 - \tau$, we have that $\|X_i\|_2 \leq O\left(\sqrt{d\log(N/\tau)}\right)$ for all $i = 1, \ldots, N$.*

**4.2.2. Estimation error in the Frobenius norm.** Let $X_1, \ldots, X_N$ be $N$ i.i.d. samples from $\mathcal{N}(0, I)$. In this section we demonstrate a tight bound on how many samples are necessary such that the sample covariance is close to $I$ in Frobenius norm. Let $\widehat{\Sigma}$ denote the empirical covariance, defined to be

$$\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^N X_i X_i^T \,.$$

By self-duality of the Frobenius norm, we know that

$$\begin{aligned}\|\widehat{\Sigma} - I\|_F &= \sup_{\|U\|_F=1}\left|\left\langle \widehat{\Sigma} - I, U\right\rangle\right| \\ &= \sup_{\|U\|_F=1}\left|\frac{1}{N}\sum_{i=1}^N \operatorname{tr}(X_i X_i^T U) - \operatorname{tr}(U)\right| \,.\end{aligned}$$

Since there is a 1/4-net over all PSD matrices with Frobenius norm 1 of size $9^{d^2}$ (see, e.g., Lemma 1.18 in [RH17]), the Vershynin-type union bound argument combined with Lemma 2.20 immediately gives us the following corollary.

COROLLARY 4.7. *There exist universal constants $A, B > 0$ so that for all $t > 0$, we have*

$$\Pr\left[\left\|\frac{1}{N}\sum_{i=1}^N X_i X_i^\top - I\right\|_F > t\right] \leq 4\exp\left(Ad^2 - BN\min(t, t^2)\right) \,.$$

By the argument as used in the proof of Lemma 4.3, we obtain the following corollary.

COROLLARY 4.8. *Fix $\varepsilon, \tau > 0$. There is a $\delta_1 = O(\varepsilon\log 1/\varepsilon)$ such that if $X_1, \ldots, X_N$ are independent samples from $\mathcal{N}(0, I)$, with*

$$N = \Omega\left(\frac{d^2 + \log 1/\tau}{\delta_1^2}\right) \,,$$

*then*

$$\Pr\left[\exists w \in S_{N,\varepsilon} : \left\|\sum_{i=1}^{N} w_i X_i X_i^\top - I\right\|_F \geq \delta_1\right] \leq \tau \ .$$

Since the proof is essentially identical to the proof of Lemma 4.3, we omit the proof. However, we note that, in fact, the proof technique there can be used to show something slightly stronger, which we will require later. The technique actually shows that if we take any set of size at most $\varepsilon N$, and take the uniform weights over that set, then the empirical covariance is not too far away from the truth.

COROLLARY 4.9. *Fix $\varepsilon, \tau > 0$. There is a $\delta_2 = O(\varepsilon \log 1/\varepsilon)$ such that if $X_1, \ldots, X_N$ are independent samples from $\mathcal{N}(0, I)$, with*

$$N = \Omega\left(\frac{d^2 + \log 1/\tau}{\delta_2^2}\right) \ ,$$

*then*

$$\Pr\left[\exists T \subseteq [N] : |T| \leq \varepsilon N \ \text{and} \ \left\|\sum_{i \in T} \frac{1}{|T|} X_i X_i^\top - I\right\|_F \geq O\left(\delta_2 \frac{N}{|T|}\right)\right] \leq \tau \ .$$

We prove this corollary in Appendix A.

**4.2.3. Understanding the fourth moment tensor.** Our algorithms will be based on understanding the behavior of the fourth moment tensor of a Gaussian when restricted to various subspaces. Let $\otimes$ denote the Kronecker product on matrices. We will make crucial use of the following definition.

DEFINITION 4.10. *For any matrix $M \in \mathbb{R}^{d \times d}$, let $M^\flat \in \mathbb{R}^{d^2}$ denote its canonical flattening into a vector in $\mathbb{R}^{d^2}$, and for any vector $v \in \mathbb{R}^{d^2}$, let $v^\sharp$ denote the unique matrix $M \in \mathbb{R}^{d \times d}$ such that $M^\flat = v$.*

We will also require the following definitions.

DEFINITION 4.11. *Let $\mathcal{S}_{\text{sym}} = \{M^\flat \in \mathbb{R}^{d^2} : M \text{ is symmetric}\}$, let $\mathcal{S} \subseteq \mathcal{S}_{\text{sym}}$ be the subspace given by*

$$\mathcal{S} = \{v \in \mathcal{S}_{\text{sym}} : \text{tr}(v^\sharp) = 0\} \ ,$$

*and let $\Pi_S$ and $\Pi_{\mathcal{S}^\perp}$ denote the projection operators onto $\mathcal{S}$ and $\mathcal{S}^\perp$, respectively. Finally, let*

$$\|v\|_{\mathcal{S}} = \|\Pi_{\mathcal{S}} v\|_2 \quad \text{and} \quad \|v\|_{\mathcal{S}^\perp} = \|\Pi_{\mathcal{S}^\perp} v\|_2 \ .$$

*Moreover, for any $M \in \mathbb{R}^{d^2 \times d^2}$, let*

$$\|M\|_{\mathcal{S}} = \sup_{v \in \mathcal{S} - \{0\}} \frac{v^T M v}{\|v\|_2^2} \ .$$

In fact, the projection of $v = M^\flat$ onto $\mathcal{S}$ where $M$ is symmetric can be written out explicitly. Namely, it is given by

$$M = \left(M - \frac{\text{tr}(M)}{d} I\right) + \frac{\text{tr}(M)}{d} I \ .$$

By construction, the flattening of the first term is in $\mathcal{S}$ and the flattening of the second term is in $\mathcal{S}^\perp$. The expression above immediately implies that $\|v\|_{\mathcal{S}^\perp} = \frac{|\mathrm{tr}(M)|}{\sqrt{d}}$.

The key result in this section is the following theorem.

THEOREM 4.12. *Let $X \sim \mathcal{N}(0, \Sigma)$. Let $M$ be the $d^2 \times d^2$ matrix given by $M = \mathbb{E}[(X \otimes X)(X \otimes X)^T]$. Then, as an operator on $\mathcal{S}_{\mathrm{sym}}$, we have*

$$M = 2\Sigma^{\otimes 2} + \left(\Sigma^\flat\right)\left(\Sigma^\flat\right)^T \ .$$

It is important to note that the two terms above are *not* the same; the first term is high rank, but the second term is rank one. The proof of this theorem will require Isserlis' theorem, and is deferred to Appendix A.

**4.2.4. Concentration of the fourth moment tensor.** We also need to show that the fourth moment tensor concentrates.

THEOREM 4.13. *Fix $\varepsilon, \tau > 0$. Let $Y_i \sim \mathcal{N}(0, I)$ be independent, for $i = 1, \ldots, N$, where we set*

$$N = \widetilde{\Omega}\left(\frac{d^2 \log^5 1/\tau}{\delta_3^2}\right) \ .$$

*Let $Z_i = Y_i^{\otimes 2}$. Let $M_4 = \mathbb{E}[Z_i Z_i^T]$ be the canonical flattening of the true fourth moment tensor. There is a $\delta_3 = O(\varepsilon \log^2 1/\varepsilon)$ such that if $Y_1, \ldots, Y_N$, and $Z_1, \ldots, Z_m$ are as above, then we have*

$$\Pr\left[\exists w \in S_{N,\varepsilon} : \left\|\sum_{i=1}^N w_i Z_i Z_i^T - M_4\right\|_{\mathcal{S}} \geq \delta_3\right] \leq \tau \ .$$

To do so will require somewhat more sophisticated techniques than the ones used so far to bound spectral deviations. At a high level, this is because fourth moments of Gaussians have a sufficiently larger variance that the union bound techniques used so far are insufficient. However, we will show that the tails of degree four polynomials of Gaussians still sufficiently concentrate such that removing points cannot change the mean by too much. The proof requires slightly fancy machinery and appears in Appendix B.

**4.3. Finding the mean, using a separation oracle.** In this section, we consider the problem of approximating $\mu$ given $N$ samples from $\mathcal{N}(\mu, I)$ in the full adversary model. Our algorithm will be based on working with the following convex set:

$$\mathcal{C}_\delta = \left\{ w \in S_{N,\varepsilon} : \left\|\sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I\right\|_2 \leq \delta \right\} \ .$$

It is not hard to show that $\mathcal{C}_\delta$ is nonempty for reasonable values of $\delta$ (and we will show this later). Moreover, we will show that for any set of weights $w$ in $\mathcal{C}_\delta$, the empirical average

$$\widehat{\mu} = \sum_{i=1}^N w_i X_i$$

will be a good estimate for $\mu$. The challenge is that since $\mu$ itself is unknown, there is not an obvious way to design a separation oracle for $\mathcal{C}_\delta$ even though it is convex. Our algorithm will run in two basic steps. First, it will run a very naive outlier detection to remove any points which are more than $O(\sqrt{d})$ away from the good points. These points are sufficiently far away that a very basic test can detect them. Then, with the remaining points, it will use the approximate separation oracle given below to approximately optimize with respect to $C_\delta$. It will then take the outputted set of weights and output the empirical mean with these weights. We will explain these steps in detail below.

Our results will hold under the following deterministic conditions:

$$\|X_i - \mu\|_2 \le O\left(\sqrt{d \log(N/\tau)}\right) \;\; \forall i \in G\,, \tag{7}$$

$$\left\| \sum_{i \in G} w_i (X_i - \mu)(X_i - \mu)^T - w_g I \right\|_2 \le \delta_1 \;\; \forall w \in S_{N,4\varepsilon}, \text{ and} \tag{8}$$

$$\left\| \sum_{i \in G} w_i (X_i - \mu) \right\|_2 \le \delta_2 \;\; \forall w \in S_{N,4\varepsilon}\;\;. \tag{9}$$

The concentration bounds we gave earlier were exactly bounds on the failure probability of either of these conditions, albeit for $S_{N,\varepsilon}$ instead of $S_{N,4\varepsilon}$.

**4.3.1. Naive pruning.** The first step of our algorithm will be to remove points which have distance which is much larger than $O(\sqrt{d})$ from the mean. Our algorithm is very naive; it computes all pairwise distances between points, and throws away all points which have distance more than $O(\sqrt{d})$ from more than a $2\varepsilon$-fraction of the remaining points.

**Algorithm 1** Naive pruning.

1: **function** NAIVEPRUNE($X_1, \ldots, X_N$)
2: For $i, j = 1, \ldots, N$, define $\delta_{i,j} = \|X_i - X_j\|_2$.
3: **for** $i = 1, \ldots, j$ **do**
4: Let $A_i = \{j \in [N] : \delta_{i,j} > \Omega(\sqrt{d \log(N/\tau)})\}$.
5: **if** $|A_i| > 2\varepsilon N$ **then**
6: Remove $X_i$ from the set.
7: **return** the pruned set of samples.

Then we have the following fact.

FACT 4.14. *Suppose that* (7) *holds. Then* NAIVEPRUNE *removes no uncorrupted points, and moreover, if $X_i$ is not removed by* NAIVEPRUNE, *we have* $\|X_i - \mu\|_2 \le O\left(\sqrt{d \log(N/\tau)}\right)$.

*Proof.* That no uncorrupted point is removed follows directly from (7) and the fact that there can be at most $2\varepsilon N$ corrupted points. Similarly, if $X_i$ is not removed by NAIVEPRUNE, that means there must be an uncorrupted $X_j$ such that $\|X_i - X_j\|_2 \le O(\sqrt{d \log(N/\tau)})$. Then the desired property follows from (7) and a triangle inequality. ∎

Henceforth, for simplicity we shall assume that no point was removed by NAIVEPRUNE, and that for all $i = 1, \ldots, N$, we have $\|X_i - \mu\|_2 < O(\sqrt{d \log(N/\tau)})$. Otherwise, we can simply work with the pruned set, and it is evident that nothing changes.

**4.3.2. The separation oracle.** Our main result in this section is an approximate separation oracle for $\mathcal{C}_\delta$. Throughout this section, let $w \in S_{N,\varepsilon}$ and set $\widehat{\mu} = \sum_{i=1}^N w_i X_i$. Moreover, let $\Delta = \mu - \widehat{\mu}$. Our first step is to show that any set of weights that does not yield a good estimate for $\mu$ cannot be in the set $\mathcal{C}_\delta$.

LEMMA 4.15. *Suppose that* (8)–(9) *holds. Suppose that* $\|\Delta\|_2 = \Omega(\sqrt{\varepsilon\delta_1}) = \Omega(\varepsilon \log 1/\varepsilon)$. *Then*

$$\left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 \geq \Omega\left( \frac{\|\Delta\|_2^2}{\varepsilon} \right) .$$

*Proof.* By Fact 4.2 and (9) we have $\| \sum_{i \in G} \frac{w_i}{w_g} X_i - \mu\|_2 \leq \delta_2$. Now by the triangle inequality we have

$$\left\| \sum_{i \in E} w_i (X_i - \mu) \right\|_2 \geq \|\Delta\|_2 - \left\| \sum_{i \in G} w_i (X_i - \mu) - w_g \mu \right\|_2 \geq \Omega(\|\Delta\|_2) .$$

Using the fact that the variance is nonnegative, we have

$$\sum_{i \in E} \frac{w_i}{w_b} (X_i - \mu)(X_i - \mu)^T \succeq \left( \sum_{i \in E} \frac{w_i}{w_b} (X_i - \mu) \right) \left( \sum_{i \in E} \frac{w_i}{w_b} (X_i - \mu) \right)^T ,$$

and therefore,

$$\left\| \sum_{i \in E} w_i (X_i - \mu)(X_i - \mu)^T \right\|_2 \geq \Omega\left( \frac{\|\Delta\|_2^2}{w_b} \right) \geq \Omega\left( \frac{\|\Delta\|_2^2}{\varepsilon} \right) .$$

On the other hand,

$$\begin{aligned} \left\| \sum_{i \in G} w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 &\leq \left\| \sum_{i \in G} w_i (X_i - \mu)(X_i - \mu)^T - w_g I \right\|_2 + w_b \\ &\leq \delta_1 + w_b, \end{aligned}$$

where in the last inequality we have used Fact 4.2 and (8). Hence altogether this implies that

$$\left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 \geq \Omega\left( \frac{\|\Delta\|_2^2}{\varepsilon} \right) - w_b - \delta_1 \geq \Omega\left( \frac{\|\Delta\|_2^2}{\varepsilon} \right) ,$$

as claimed. ∎

As a corollary, we find that *any* set of weights in $\mathcal{C}_\delta$ immediately yields a good estimate for $\mu$.

COROLLARY 4.16. *Suppose that* (8) *and* (9) *hold. Let* $w \in \mathcal{C}_\delta$ *for* $\delta = O(\varepsilon \log 1/\varepsilon)$. *Then*

$$\|\Delta\|_2 \leq O(\varepsilon \sqrt{\log 1/\varepsilon}).$$

Our main result in this section is an approximate separation oracle for $\mathcal{C}_\delta$ with $\delta = O(\varepsilon \log 1/\varepsilon)$.

THEOREM 4.17. *Fix $\varepsilon > 0$, and let $\delta = O(\varepsilon \log 1/\varepsilon)$. Suppose that (8) and (9) hold. Let $w^*$ denote the weights which are uniform on the uncorrupted points. Then there is a constant $c$ and an algorithm such that the following hold:*

1. *(Completeness) If $w = w^*$, then it outputs "YES".*
2. *(Soundness) If $w \notin \mathcal{C}_{c\delta}$, the algorithm outputs a hyperplane $\ell : \mathbb{R}^N \to \mathbb{R}$ such that $\ell(w) \geq 0$ but $\ell(w^*) < 0$. Moreover, if the algorithm ever outputs a hyperplane $\ell$, then $\ell(w^*) < 0$.*

We remark that these two facts imply that for any $\tau > 0$, the ellipsoid method with this separation oracle will output a $w'$ such that $\|w - w'\|_\infty < \varepsilon/(N\sqrt{d \log(N/\tau)})$, for some $w \in \mathcal{C}_{c\delta}$ in $\text{poly}(d, 1/\varepsilon, \log 1/\tau)$ steps.

*Remark* 4.18. The conditions satisfied by the separation oracle given here are slightly weaker than the traditional guarantees given, for instance, in [GLS88]. However, the correctness of the ellipsoid algorithm with this separation oracle follows because outside $C_{c\delta}$, the separation oracle acts exactly as a separation oracle for $w^*$. Thus, as long as the algorithm continues to query points outside of $C_{c\delta}$, the action of the algorithm is equivalent to one with a separation oracle for $w^*$. Moreover, the behavior of the algorithm is such that it will never exclude $w^*$, even if queries are made within $C_{c\delta}$. From these two conditions, it is clear from the classical theory presented in [GLS88] that the ellipsoid method satisfies the guarantees given above.

The separation oracle is given in Algorithm 2. Next, we prove correctness for our approximate separation oracle:

**Algorithm 2** Separation oracle subprocedure for agnostically learning the mean.

**function** SEPARATIONORACLEUNKNOWNMEAN($w, \varepsilon, X_1, \ldots, X_N$)
  Let $\widehat{\mu} = \sum_{i=1}^N w_i X_i$.
  Let $\delta = O(\varepsilon \log 1/\varepsilon)$.
  For $i = 1, \ldots, N$, define $Y_i = X_i - \widehat{\mu}$.
  Let $\lambda$ be the eigenvalue of largest magnitude of $M = \sum_{i=1}^N w_i Y_i Y_i^T - I$.
  Let $v$ be its associated eigenvector.
  **if** $|\lambda| \leq \frac{c}{2}\delta$ **then**
    **return** "YES".
  **else if** $\lambda > \frac{c}{2}\delta$ **then**
    **return** the hyperplane $\ell(u) = \left(\sum_{i=1}^N u_i \langle Y_i, v\rangle^2 - 1\right) - \lambda$.
  **else**
    **return** the hyperplane $\ell(u) = \lambda - \left(\sum_{i=1}^N u_i \langle Y_i, v\rangle^2 - 1\right)$.

*Proof of Theorem* 4.17. Again, let $\Delta = \mu - \widehat{\mu}$, and let $M = \sum_{i=1}^N w_i Y_i Y_i^T - I$.

By expanding out the formula for $M$, we get

$$\begin{aligned}\sum_{i=1}^N w_i Y_i Y_i^T - I &= \sum_{i=1}^N w_i (X_i - \mu + \Delta)(X_i - \mu + \Delta)^T - I \\ &= \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I + \sum_{i=1}^N w_i (X_i - \mu)\Delta^T \\ &\qquad + \Delta \sum_{i=1}^N w_i (X_i - \mu)^T + \Delta\Delta^T \\ &= \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I - \Delta\Delta^T \ .\end{aligned}$$

Let us now prove completeness.

CLAIM 4.19. *Suppose $w = w^*$. Then $\|M\|_2 < \frac{c}{2}\delta$.*

*Proof.* Recall that $w^*$ are the weights that are uniform on the uncorrupted points. Because $|E| \leq 2\varepsilon N$ we have that $w^* \in S_{N,\varepsilon}$. We can now use (8) to conclude that $w^* \in \mathcal{C}_{\delta_1}$. Now by Corollary 4.16 we have that $\|\Delta\|_2 \leq O(\varepsilon\sqrt{\log 1/\varepsilon})$. Thus

$$\begin{aligned}&\left\| \sum_{i=1}^N w_i^* (X_i - \mu)(X_i - \mu)^T - I - \Delta\Delta^T \right\|_2 \\ &\leq \left\| \sum_{i=1}^N w_i^* (X_i - \mu)(X_i - \mu)^T - I \right\|_2 + \|\Delta\Delta^T\|_2 \\ &\leq \delta_1 + O(\varepsilon^2 \log 1/\varepsilon) < \frac{c\delta}{2} \ .\end{aligned}$$ ∎

We now turn our attention to soundness.

CLAIM 4.20. *Suppose that $w \notin C_{c\delta}$. Then $|\lambda| > \frac{c}{2}\delta$.*

*Proof.* By the triangle inequality, we have

$$\left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I - \Delta\Delta^T \right\|_2 \geq \left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 - \left\| \Delta\Delta^T \right\|_2 \ .$$

Let us now split into two cases. If $\|\Delta\|_2 \leq \sqrt{c\delta/10}$, then the first term above is at least $c\delta$ by definition and we can conclude that $|\lambda| > c\delta/2$. On the other hand, if $\|\Delta\|_2 \geq \sqrt{c\delta/10}$, by Lemma 4.15, we have that

$$\left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I - \Delta\Delta^T \right\|_2 \geq \Omega\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right) - \|\Delta\|_2^2 = \Omega\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right) , \tag{10}$$

which for sufficiently small $\varepsilon$ also yields $|\lambda| > c\delta/2$. ∎

Now by construction $\ell(w) \geq 0$. All that remains is to show that $\ell(w^*) < 0$ always holds. We will only consider the case where the top eigenvalue $\lambda$ of $M$ is positive.

The other case (when $\lambda < -\frac{c}{2}\delta$) is symmetric. We will split the analysis into two parts:

$$\begin{aligned}\left\| \frac{1}{|G|} \sum_{i \in G} (X_i - \widehat{\mu})(X_i - \widehat{\mu})^T - I \right\|_2 &= \left\| \frac{1}{|G|} \sum_{i \in G} (X_i - \mu + \Delta)(X_i - \mu + \Delta)^T - I \right\|_2 \\ &\leq \underbrace{\left\| \frac{1}{|G|} \sum_{i \in G} (X_i - \mu)(X_i - \mu)^T - I \right\|_2}_{\leq \delta_1} + 2\|\Delta\|_2 \underbrace{\left\| \frac{1}{|G|} \sum_{i \in G} (X_i - \mu) \right\|_2}_{\leq 2\delta_2 \|\Delta\|_2 \text{ since } w^* \in C_{\delta_2}} \\ &\quad + \|\Delta\|_2^2 \, . \end{aligned} \tag{11}$$

Suppose $\|\Delta\|_2 \leq \sqrt{c\delta/10}$. By (11) we immediately have

$$\ell(w^*) \leq \delta_1 + 2\delta_2 \|\Delta\|_2 + \|\Delta\|_2^2 - \lambda \leq \frac{c\delta}{5} - \lambda < 0 \, ,$$

since $\lambda > c\delta/2$. On the other hand, if $\|\Delta\|_2 \geq \sqrt{c\delta/10}$, then by (10) we have $\lambda = \Omega\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right)$. Putting it all together we have

$$\ell(w^*) \leq \underbrace{\left\| \frac{1}{|G|} \sum_{i \in G} (X_i - \widehat{\mu})(X_i - \widehat{\mu})^T - I \right\|_2}_{\leq \delta_1 + 2\delta_2 \|\Delta\|_2 + \|\Delta\|_2^2} - \lambda \, ,$$

where in the last line we used the fact that $\lambda > \Omega\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right)$, and $\|\Delta\|_2^2 \geq \Omega(\varepsilon^2 \log 1/\varepsilon)$. This now completes the proof. ∎

**4.3.3. The full algorithm.** This separation oracle, along with the classical theory of convex optimization [GLS88], implies that we have shown the following corollary.

COROLLARY 4.21. *Fix $\varepsilon, \tau > 0$, and let $\delta = O(\varepsilon\sqrt{\log 1/\varepsilon})$. Let $X_1, \ldots, X_N$ be an $\varepsilon$-corrupted set of points satisfying (8)–(9), for $\delta_1 \leq \delta$ and $\delta_2 \leq \delta\sqrt{\log 1/\varepsilon}$. Let $c$ be a sufficiently large constant. Then, there is an algorithm* LEARNAPPROXMEAN$(\varepsilon, \tau, X_1, \ldots, X_N)$ *which runs in time* $\text{poly}(N, d, 1/\varepsilon, \log 1/\tau)$*, and outputs a set of weights $w' \in S_{N,\varepsilon}$ such that there is a $w \in C_{c\delta}$ such that $\|w - w'\|_\infty \leq \varepsilon/(N\sqrt{d \log(N/\tau)})$.*

This algorithm, while an extremely powerful primitive, is technically not sufficient. However, given this, the full algorithm is not too difficult to state: simply run NAIVEPRUNE, then optimize over $C_{c\delta}$ using this separation oracle, and get some $w$ which is approximately in $C_{c\delta}$. Then, output $\sum_{i=1}^N w_i X_i$. For completeness, the pseudocode for the algorithm is given below. In the pseudocode, we assume that ELLIPSOID(SEPARATIONORACLEUNKNOWNMEAN, $\varepsilon'$) is a convex optimization routine, which given the SEPARATIONORACLEUNKNOWNMEAN separation oracle and a target error $\varepsilon'$, outputs a $w'$ such that $\|w - w'\|_\infty \leq \varepsilon'$. From the classical theory of optimization, we know such a routine exists and runs in polynomial time.

THEOREM 4.22. *Fix $\varepsilon, \tau > 0$, and let $\delta = O(\varepsilon\sqrt{\log 1/\varepsilon})$. Let $X_1, \ldots, X_N$ be an $\varepsilon$-corrupted set of samples, where*

$$N = \Omega\left(\frac{d + \log 1/\tau}{\delta^2}\right) \, .$$

**Algorithm 3** Convex programming algorithm for agnostically learning the mean.

```
function LearnMean(ε, τ, X_1, ..., X_N)
    Run NaivePrune(X_1, ..., X_N). Let {X_i}_{i∈I} be the pruned set of samples.
      /* For simplicity assume I = [N] */
    Let w' ← LearnApproxMean(ε, τ, X_1, ..., X_N).
    return Σ_{i=1}^N w'_i X_i.
```

*Let* $\widehat{\mu}$ *be the output of* LearnMean$(\varepsilon, \tau, X_1, \dots, X_N)$. *Then with probability* $1-\tau$, *we have* $\|\widehat{\mu} - \mu\|_2 \le \delta$.

*Proof.* By Fact 4.6 and Lemmas 4.3 and 4.4, we know that (7)–(9) hold with probability $1-\tau$, with $\delta_1, \delta_2 \le \delta$. Condition on the event that (7)–(9) simultaneously hold. After NaivePrune, by Fact 4.14 we may assume that no uncorrupted points are removed, and all points satisfy $\|X_i - \mu\|_2 \le O(\sqrt{d \log(N/\tau)})$. Let $w'$ be the output of the algorithm, and let $w \in C_{c\delta}$ be such that $\|w - w'\|_\infty < \varepsilon/(N\sqrt{d\log(N/\tau)})$. By Corollary 4.16, we know that $\| \sum_{i=1}^N w_i X_i - \mu\|_2 \le O(\delta)$. Hence, we have

$$\left\| \sum_{i=1}^N w'_i X_i - \mu \right\|_2 \le \left\| \sum_{i=1}^N w_i X_i - \mu \right\|_2 + \sum_{i=1}^N |w_i - w'_i| \cdot \|X_i - \mu\|_2 \le O(\delta) + \varepsilon \ ,$$

so the entire error is at most $O(\delta)$, as claimed. ∎

**4.4. Finding the covariance, using a separation oracle.** In this section, we consider the problem of approximating $\Sigma$ given $N$ samples from $\mathcal{N}(0, \Sigma)$ in the full adversary model. Let $U_i = \Sigma^{-1/2} X_i$ such that if $X_i \sim \mathcal{N}(0,\Sigma)$, then $U_i \sim \mathcal{N}(0, I)$. Moreover let $Z_i = U_i^{\otimes 2}$. Our approach will parallel the one given earlier in section 4.3. Again, we will work with a convex set

$$C_\delta = \left\{ w \in S_{N,\varepsilon} : \left\| \Sigma^{-1/2} \left( \sum_{i=1}^m w_i X_i X_i^T \right) \Sigma^{-1/2} - I \right\|_F \le \delta \right\} ,$$

and our goal is to design an approximate separation oracle. Our results in this section will rely on the following deterministic conditions:

$$\|U_i\|_2^2 \le O\left(d \log(N/\tau)\right) \quad \forall i \in G \tag{12}$$

$$\left\| \sum_{i \in G} w_i U_i U_i^T - w_g I \right\|_F \le \delta_1 \ , \tag{13}$$

$$\left\| \sum_{i \in T} \frac{1}{|T|} U_i U_i^T - I \right\|_F \le O\left( \delta_2 \frac{N}{|T|} \right) \ , \text{ and} \tag{14}$$

$$\left\| \sum_{i \in G} w_i Z_i Z_i^T - w_g M_4 \right\|_{\mathcal{S}} \le \delta_3 \tag{15}$$

for all $w \in S_{N,\varepsilon}$, and all sets $T \subseteq G$ of size $|T| \le 2\varepsilon N$. As before, by Fact 4.2, the renormalized weights over the uncorrupted points are in $S_{N,4\varepsilon}$. Hence, we can appeal to Fact 4.6, Corollaries 4.8 and 4.9, and Theorem 4.13 with $S_{N,4\varepsilon}$ instead of $S_{N,\varepsilon}$ to bound the probability that this event does not hold. Let $w^*$ be the set of weights which are uniform over the uncorrupted points; by (13) for $\delta \ge \Omega(\varepsilon\sqrt{\log 1/\varepsilon})$ we have that $w^* \in C_\delta$.

THEOREM 4.23. *Let $\delta = O(\varepsilon \log 1/\varepsilon)$. Suppose that* (13), (14), *and* (15) *hold for $\delta_1, \delta_2 \le O(\delta)$ and $\delta_3 \le O(\delta \log 1/\varepsilon)$. Then, there are a constant $c$ and an algorithm such that, given any input $w \in S_{N,\varepsilon}$, we have the following:*

1. *(Completeness) If $w = w^*$, the algorithm outputs "YES".*
2. *(Soundness) If $w \notin C_{c\delta}$, the algorithm outputs a hyperplane $\ell : \mathbb{R}^m \to \mathbb{R}$ such that $\ell(w) \ge 0$ but we have $\ell(w^*) < 0$. Moreover, if the algorithm ever outputs a hyperplane $\ell$, then $\ell(w^*) < 0$.*

As in the case of learning an unknown mean, by the classical theory of convex optimization this implies that we will find a point $w$ such that $\|w - w'\|_\infty \le \frac{\varepsilon}{\mathrm{poly}(N)}$ for some $w' \in C_{c\delta}$, using polynomially many calls to this oracle. We make this more precise in the following subsubsection.

The pseudocode for the (approximate) separation oracle is given in Algorithm 4. Observe briefly that this algorithm does indeed run in polynomial time. Lines 2–6 require only taking top eigenvalues and eigenvectors, and so can be done in polynomial time. For any $\xi \in \{-1, +1\}$, line 7 can be run by sorting the samples by $w_i \big( \frac{\|Y_i\|^2}{\sqrt{d}} - \sqrt{d} \big)$ and seeing if there is a subset of the top $2\varepsilon N$ samples satisfying the desired condition, and line 8 can be executed similarly.

---

**Algorithm 4** Convex programming algorithm for agnostically learning the covariance.

---

1: **function** SEPARATIONORACLEUNKNOWNCOVARIANCE($w$)

2: Let $\widehat{\Sigma} = \sum_{i=1}^N w_i X_i X_i^T$.

3: For $i = 1, \ldots, N$, let $Y_i = \widehat{\Sigma}^{-1/2} X_i$ and let $Z_i = (Y_i)^{\otimes 2}$.

4: Let $v$ be the top eigenvector of $M = \sum_{i=1}^N w_i Z_i Z_i^T - 2I$ restricted to $\mathcal{S}$, and let $\lambda$ be its associated eigenvalue.

5: **if** $|\lambda| > \Omega(\varepsilon \log^2 1/\varepsilon)$ **then**

6: Let $\xi = \mathrm{sgn}(\lambda)$ and **return** the hyperplane

$$\ell(u) = \xi \left( \sum_{i=1}^N u_i \langle v, Z_i \rangle^2 - 2 - \lambda \right) .$$

7: **else if** there exist a sign $\xi \in \{-1, 1\}$ and a set $T$ of samples of size at most $2\varepsilon N$ such that

$$\alpha = \xi \sum_{i \in T} w_i \left( \frac{\|Y_i\|_2^2}{\sqrt{d}} - \sqrt{d} \right) > \frac{(1-\varepsilon)\alpha\delta}{2}$$

**then**

8: **return** the hyperplane

$$\ell(u) = \xi \sum_{i \in T} u_i \left( \frac{\|Y_i\|_2^2}{\sqrt{d}} - \sqrt{d} \right) - \alpha .$$

9: **return** "YES".

---

We now turn our attention to proving the correctness of this separation oracle. We require the following technical lemmas.

CLAIM 4.24. *Let $w_i$ for $i = 1, \ldots, N$ be a set of nonnegative weights such that $\sum_{i=1}^N w_i = 1$, and let $a_i \in \mathbb{R}$ be arbitrary. Then*

$$\sum_{i=1}^N a_i^2 w_i \geq \left( \sum_{i=1}^N a_i w_i \right)^2 .$$

*Proof.* Let $P$ be the distribution where $a_i$ is chosen with probability $w_i$. Then $\mathbb{E}_{X\sim P}[X] = \sum_{i=1}^N a_i w_i$ and $\mathbb{E}_{X\sim P}[X^2] = \sum_{i=1}^N a_i w_i^2$. Since $\mathrm{Var}_{X\sim P}[X] = \mathbb{E}_{X\sim P}[X^2] - \mathbb{E}_{X\sim P}[X]^2$ is always a nonnegative quantity, by rearranging the desired conclusion follows. ∎

LEMMA 4.25. *Fix $\delta < 1$ and suppose that $M$ is symmetric. If $\|M - I\|_F \geq \delta$, then $\|M^{-1} - I\|_F \geq \frac{\delta}{2}$.*

*Proof.* We will prove this lemma in the contrapositive, by showing that if $\|M^{-1} - I\|_F < \frac{\delta}{2}$, then $\|M - I\|_F < \delta$. Since the Frobenius norm is rotationally invariant, we may assume that $M^{-1} = \mathrm{diag}(1 + \nu_1, \ldots, 1 + \nu_d)$, where by assumption $\sum \nu_i^2 < \delta^2/4$. By our assumption that $\delta < 1$, we have $|\nu_i| \leq 1/2$ for all $i$. Thus

$$\sum_{i=1}^d \left( 1 - \frac{1}{1 + \nu_i} \right)^2 \leq \sum_{i=1}^d 4\nu_i^2 < \delta ,$$

where we have used the inequality $|1 - \frac{1}{1+x}| \leq |2x|$ which holds for all $|x| \leq 1/2$. This completes the proof. ∎

LEMMA 4.26. *Let $M, N \in \mathbb{R}^{d\times d}$ be arbitrary matrices. Then $\|MN\|_F \leq \|M\|_2 \|N\|_F$.*

*Proof.* Let $N_1, \ldots, N_d$ be the columns of $N$. Then

$$\|MN\|_F^2 = \sum_{i=1}^d \|MN\|_2^2 \leq \|M\|_2^2 \sum_{i=1}^d \|N_i\|_2^2 = \|M\|_2^2 \|N\|_F^2 ,$$

so the desired result follows by taking square roots of both sides. ∎

LEMMA 4.27. *Let $M \in \mathbb{R}^{d\times d}$. Then, $\left\| \left(M^\flat\right) \left(M^\flat\right)^T \right\|_{\mathcal{S}} \leq \|M - I\|_F^2$.*

*Proof.* By the definition of $\|\cdot\|_{\mathcal{S}}$, we have

$$\left\| \left(M^\flat\right) \left(M^\flat\right)^T \right\|_{\mathcal{S}} = \sup_{\substack{A^\flat \in \mathcal{S} \\ \|A\|_F = 1}} \left(A^\flat\right)^T \left(M^\flat\right) \left(M^\flat\right)^T A^\flat = \sup_{\substack{A \in \mathcal{S} \\ \|A\|_F = 1}} \langle A, M \rangle^2 .$$

By self-duality of the Frobenius norm, we know that

$$\langle A, M \rangle = \langle A, M - I \rangle \leq \|M - I\|_F ,$$

since $I^\flat \in \mathcal{S}^\perp$. The result now follows. ∎

*Proof of Theorem* 4.23. Let us first prove completeness. Observe that by Theorem 4.12, we know that restricted to $\mathcal{S}$, we have that $M_4 = 2I$. Therefore, by (15) we will not output a hyperplane in line 6 of Algorithm 4. Moreover, by (14), we will not output a hyperplane in line 7 of Algorithm 4. This proves completeness.

Thus it suffices to show soundness. Suppose that $w \notin \mathcal{C}_{c\delta}$. We will make use of the following elementary fact.

FACT 4.28. *Let* $A = \Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}$ *and* $B = \widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2}$. *Then*

$$\|A^{-1} - I\|_F = \|B - I\|_F \ .$$

*Proof.* In particular, $A^{-1} = \Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}$. Using this expression and the fact that all the matrices involved are symmetric, we can write

$$\begin{aligned}\|A^{-1} - I\|_F^2 &= \operatorname{tr}\left((A^{-1} - I)^T(A^{-1} - I)\right) \\ &= \operatorname{tr}\left(\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}\Sigma^{1/2} - 2\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2} - I\right) \\ &= \operatorname{tr}\left(\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1/2} - 2\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2} - I\right) \\ &= \operatorname{tr}\left((B - I)^T(B - I)\right) = \|B - I\|_F^2 \ ,\end{aligned}$$

where in the third line we have used the fact that the trace of a product of matrices is preserved under cyclic shifts. ∎

This allows us to show the following claim.

CLAIM 4.29. *Assume* (13) *holds with* $\delta_1 \le O(\delta)$ *and assume furthermore that* $\|A - I\|_F \ge c\delta$. *Then, if we let* $\delta' = \frac{(1-\varepsilon)c}{2}\delta = \Theta(\delta)$, *we have*

$$\left\|\sum_{i\in E} w_i Z_i - w_b I^\flat\right\|_{\mathcal{S}} + \left\|\sum_{i\in E} w_i Z_i - w_b I^\flat\right\|_{\mathcal{S}^\perp} \ge \delta' \ . \tag{16}$$

*Proof.* Let $A, B$ be as in Fact 4.28. Combining Lemma 4.25 and Fact 4.28 we have

$$\|A - I\|_F \ge c\delta \Rightarrow \|B - I\|_F \ge \frac{c\delta}{2} \ . \tag{17}$$

We can rewrite (13) as the expression $\sum_{i\in G} w_i X_i X_i^T = w_g \Sigma^{1/2}(I + R)\Sigma^{1/2}$, where $R$ is symmetric and satisfies $\|R\|_F \le \delta_1$. By the definition of $\widehat{\Sigma}$ we have that $\sum_{i=1}^N w_i Y_i Y_i^T = I$, and so

$$\begin{aligned}\left\|\sum_{i\in E} w_i Y_i Y_i^T - w_b I\right\|_F &= \left\|\sum_{i\in G} w_i Y_i Y_i^T - w_g I\right\|_F \\ &= w_g \left\|\widehat{\Sigma}^{-1/2}\Sigma^{1/2}(I + R)\Sigma^{1/2}\widehat{\Sigma}^{-1/2} - I\right\|_F \ .\end{aligned}$$

Furthermore, we have

$$\left\|\widehat{\Sigma}^{-1/2}\Sigma^{1/2}R\Sigma^{1/2}\widehat{\Sigma}^{-1/2}\right\|_F \le \delta_1 \left\|\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2}\right\|_2$$

by applying Lemma 4.26. And putting it all together we have

$$\left\|\sum_{i\in E} w_i Y_i Y_i^T - w_b I\right\|_F \ge w_g \left(\left\|\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2} - I\right\|_F - \delta_1 \left\|\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2}\right\|_2\right) \ .$$

It is easily verified that for $c > 10$, we have that for all $\delta$, if $\|\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2} - I\|_F \ge c\delta$, then

$$\|\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2} - I\|_F \ge 2\delta\|\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2}\|_2 \ .$$

Hence all this implies that

$$\left\| \sum_{i \in E} w_i Y_i Y_i^T - w_b I \right\|_F \geq \delta' ,$$

where $\delta' = \frac{c(1-\varepsilon)}{2}\delta = \Theta(\delta)$. The desired result then follows from the Pythagorean theorem. ∎

Claim 4.29 tells us that if $w \notin C_{c\delta}$, we know that one of the terms in (17) must be at least $\frac{1}{2}\delta'$. We first show that if the first term is large, then the algorithm outputs a separating hyperplane.

Claim 4.30. *Assume that* (13)–(15) *hold with* $\delta_1, \delta_2 \leq O(\delta)$ *and* $\delta_3 \leq O(\delta \log 1/\varepsilon)$. *Moreover, suppose that*

$$\left\| \sum_{i \in E} w_i Z_i - w_b I^\flat \right\|_{\mathcal{S}} \geq \frac{1}{2}\delta' .$$

*Then the algorithm outputs a hyperplane in line* 6*, and moreover, it is a separating hyperplane.*

*Proof.* Let us first show that given these conditions, then the algorithm indeed outputs a hyperplane in line 6. Since $I^\flat \in S^\perp$, the first term is just equal to $\left\|\sum_{i\in E} w_i Z_i\right\|_S$. But this implies that there is some $M^\flat \in S$ such that $\|M^\flat\|_2 = \|M\|_F = 1$ and such that

$$\sum_{i \in E} w_i \langle M^\flat, Z_i \rangle \geq \frac{1}{2}\delta' ,$$

which implies that

$$\sum_{i \in E} \frac{w_i}{w_b} \langle M^\flat, Z_i \rangle \geq \frac{1}{2}\frac{\delta'}{w_b} .$$

The $w_i/w_b$ are a set of weights satisfying the conditions of Claim 4.24 and so this implies that

$$\begin{aligned} \sum_{i \in E} w_i \langle M^\flat, Z_i \rangle^2 &\geq O\left(\frac{\delta'^2}{w_b}\right) \\ &\geq O\left(\frac{\delta'^2}{\varepsilon}\right) . \end{aligned} \tag{18}$$

Let $\widetilde{\Sigma} = \widehat{\Sigma}^{-1}\Sigma$. By Theorem 4.12 and (15), we have that

$$\sum_{i \in G} w_i Z_i Z_i^T = w_g \left( \left(\widetilde{\Sigma}^\flat\right)\left(\widetilde{\Sigma}^\flat\right)^T + 2\widetilde{\Sigma}^{\otimes 2} + \left(\widetilde{\Sigma}^{1/2}\right)^{\otimes 2} R \left(\widetilde{\Sigma}^{1/2}\right)^{\otimes 2} \right) ,$$

where $\|R\|_2 \leq \delta_3$. Hence,

$$\begin{aligned}\left\| \sum_{i \in G} w_i Z_i Z_i^T - 2I \right\|_S &= w_g \left\| \left(\widetilde{\Sigma}^\flat\right) \left(\widetilde{\Sigma}^\flat\right)^T + 2\left(\widetilde{\Sigma}^{\otimes 2} - I\right) \right. \\ &\qquad \left. + (1-w_g)I + \left(\widetilde{\Sigma}^{1/2}\right)^{\otimes 2} R \left(\widetilde{\Sigma}^{1/2}\right)^{\otimes 2} \right\|_S \\ &\leq \|\widetilde{\Sigma} - I\|_F^2 + 2\|\widetilde{\Sigma} - I\|_2 + (1-w_g) + \|R\|\|\widetilde{\Sigma}\|^2 \\ &\leq 3\|\widetilde{\Sigma} - I\|_F^2 + \delta\|\widetilde{\Sigma}\|^2 + O(\varepsilon)\ . \\ &\leq O\left({\delta'}^2 + \delta'\right)\ , \end{aligned} \tag{19}$$

since it is easily verified that $\delta\|\widetilde{\Sigma}\|^2 \leq O(\|\widetilde{\Sigma} - I\|_F)$ as long as $\|\widetilde{\Sigma} - I\|_F \geq \Omega(\delta)$, which it is by (17).

Equations (18) and (19) then together imply that

$$\sum_{i=1}^N w_i (M^\flat)^T Z_i Z_i^T (M^\flat) - (M^\flat)^T I M^\flat \geq O\left(\frac{\delta^2}{\varepsilon}\right)\ ,$$

and so the top eigenvalue of $M$ is greater in magnitude than $\lambda$, and so Algorithm 4 will output a hyperplane in line 6. Letting $\ell$ denote the hyperplane output by the algorithm, by the same calculation as for (19), we must have $\ell(w^*) < 0$, so this is indeed a separating hyperplane. Hence in this case, the algorithm correctly operates. ∎

Moreover, observe that from the calculations in (19), we know that if we ever output a hyperplane in line 6, which implies that $\lambda \geq \Omega(\varepsilon \log^2 1/\varepsilon)$, then we must have that $\ell(w^*) < 0$.

Now let us assume that the first term on the left-hand side (LHS) is less than $\frac{1}{2}\delta'$, such that the algorithm does not necessarily output a hyperplane in line 6. Thus, the second term on the LHS of (16) is at least $\frac{1}{2}\delta'$. We now show that this implies that the algorithm will output a separating hyperplane in line 8.

CLAIM 4.31. *Assume that* (13)–(15) *hold. Moreover, suppose that*

$$\left\| \sum_{i \in E} w_i Z_i - w_b I^\flat \right\|_{\mathcal{S}^\perp} \geq \frac{1}{2}\delta'\ .$$

*Then the algorithm outputs a hyperplane in line* 8*, and moreover, it is a separating hyperplane.*

*Proof.* By the definition of $\mathcal{S}^\perp$, the assumption implies that

$$\left| \sum_{i \in E} w_i \frac{\mathrm{tr}(Z_i^\sharp)}{\sqrt{d}} - M_b\sqrt{d} \right| \geq \frac{1}{2}\delta'\ ,$$

which is equivalent to the condition that

$$\xi \sum_{i \in E} w_i \left(\frac{\|Y_i\|_2^2}{\sqrt{d}} - \sqrt{d}\right) \geq \frac{(1-\varepsilon)\delta'}{2}\ ,$$

for some $\xi \in \{-1, 1\}$. In particular, the algorithm will output a hyperplane

$$\ell(w) = \xi \sum_{i \in S} w_i \left( \frac{\|Y_i\|_2^2}{\sqrt{d}} - \sqrt{d} \right) - \lambda$$

in step 8 of Algorithm 4, where $S$ is some set of size at most $\varepsilon N$, and $\lambda = O(\delta')$. Since it will not affect anything, without loss of generality let us assume that $\xi = 1$. The other case is symmetrical.

It now suffices to show that $\ell(w^*) < 0$ always. Let $T = S \cap G$. By (14), we know that

$$\sum_{i \in T} \frac{1}{|T|} Y_i Y_i^T - I = \widetilde{\Sigma}^{1/2} (I + A) \widetilde{\Sigma}^{1/2} - I ,$$

where $\|A\|_F = O\left(\delta \frac{N}{|T|}\right)$. Hence,

$$\begin{aligned} \left\| \sum_{i \in T} \frac{1}{(1-\varepsilon)N} Y_i Y_i^T - \frac{|T|}{(1-\varepsilon)N} I \right\|_F &= \frac{|T|}{(1-\varepsilon)N} \left\| \widetilde{\Sigma}^{1/2} (I + A) \widetilde{\Sigma}^{1/2} - I \right\|_F \\ &\le \frac{|T|}{(1-\varepsilon)N} \left( \|\widetilde{\Sigma} - I\|_F + \|A\|_F \|\widetilde{\Sigma}\|_2 \right) \\ &\le \frac{|T|}{(1-\varepsilon)N} \|\widetilde{\Sigma} - I\|_F + O(\delta) \|\widetilde{\Sigma}\|_2 \\ &\le O(\delta\delta' + \delta) , \end{aligned}$$

as long as $\delta' \ge O(\delta)$. By self-duality of the Frobenius norm, using the test matrix $\frac{1}{\sqrt{d}} I$, this implies that

$$\left| \sum_{i \in T} \frac{1}{(1-\varepsilon)N} \left( \|Y_i\|^2 - \sqrt{d} \right) \right| \le O(\delta\delta' + \delta) < \alpha$$

and hence $\ell(w^*) < 0$, as claimed. ∎

These two claims in conjunction directly imply the correctness of Theorem 4.23. ∎

**4.4.1. The full algorithm.** As before, this separation oracle and the classical theory of convex optimization [GLS88] show that we have demonstrated an algorithm FindApproxCovariance with the following properties.

Theorem 4.32. *Fix $\varepsilon, \tau > 0$, and let $\delta = O(\varepsilon \log 1/\varepsilon)$. Let $c > 0$ be a universal constant which is sufficiently large. Let $X_1, \ldots, X_N$ be an $\varepsilon$-corrupted set of points satisfying* (13)–(15)*, for $\delta_1, \delta_2 \le O(\delta)$ and $\delta_3 \le O(\delta \log 1/\varepsilon)$. Then we see that* FindApproxCovariance$(\varepsilon, \tau, X_1, \ldots, X_N)$ *runs in time* $\mathrm{poly}(N, d, 1/\varepsilon, \log 1/\tau)$*, and outputs a $u$ such that there is some $w \in C_{c\delta}$ such that $\|w - u\|_\infty \le \varepsilon/(Nd \log(N/\tau))$.*

As before, this is not quite sufficient to actually recover the covariance robustly. Naively, we would just like to output $\sum_{i=1}^N u_i X_i X_i^T$. However, this can run into issues if there are points $X_i$ such that $\|\Sigma^{-1/2} X_i\|_2$ is extremely large. We show here that we can postprocess the $u$ such that we can weed out these points. First, observe that we have the following lemma.

Lemma 4.33. *Assume $X_1, \ldots, X_N$ satisfy* (13)*. Let $w \in S_{N,\varepsilon}$. Then*

$$\sum_{i=1}^N w_i X_i X_i^T \succeq (1 - O(\delta_1)) \Sigma .$$

*Proof.* This follows since by (13), we have that $\sum_{i \in G} w_i X_i X_i^T \succeq w_g(1-\delta_1)\Sigma \succeq (1-O(\delta_1))\Sigma$. The lemma then follows since $\sum_{i \in E} w_i X_i X_i^T \succeq 0$ always. ∎

Now, for any set of weights $w \in S_{N,\varepsilon}$, let $\widetilde{w}^- \in \mathbb{R}^N$ be the vector given by $\widetilde{w}_i^- = \max(0, w_i - \varepsilon/(Nd\log(N/\tau)))$, and let $w^-$ be the set of weights given by renormalizing $\widetilde{w}^-$. It is a straightforward calculation that for any $w \in S_{N,\varepsilon}$, we have $w^- \in S_{N,2\varepsilon}$. In particular, this implies the following lemma.

LEMMA 4.34. *Let $u$ be such that there is $w \in C_{c\delta}$ such that $\|u-w\|_\infty \le \varepsilon/(Nd\log(N/\tau))$. Then, $\sum_{i=1}^N u_i^- X_i X_i^T \preceq (1+O(\delta))\Sigma$.*

*Proof.* By the definition of $C_{c\delta}$, we must have that $\sum_{i=1}^N w_i X_i X_i^T \preceq (1+c\delta)\Sigma$. Moreover, we must have $\widetilde{u}_i^- \le w_i$ for every index $i \in [N]$. Thus we have that $\sum_{i=1}^N \widetilde{u}_i^- w_i X_i X_i^T \preceq (1+c\delta)\Sigma$, and hence $\sum_{i=1}^N u_i^- w_i X_i X_i^T \preceq (1+c\delta)\Sigma$, since $\sum_{i=1}^N u_i^- w_i X_i X_i^T \preceq (1+O(\varepsilon))\sum_{i=1}^N \widetilde{u}_i^- w_i X_i X_i^T$. ∎

We now give the full algorithm. The algorithm proceeds as follows: first run FINDAPPROXCOVARIANCE to get some set of weights $u$ which is close to some element of $C_{c\delta}$. We then compute the empirical covariance $\Sigma_1 = \sum_{i=1}^N u_i X_i X_i^T$ with the weights $u$ and remove any points which have $\|\Sigma_1^{-1/2} X_i\|_2^2$ which are too large. We shall show that this removes no good points, and removes all corrupted points which have $\|\Sigma^{-1/2} X_i\|_2^2$ which are absurdly large. We then rerun FINDAPPROXCOVARIANCE with this pruned set of points, and output the empirical covariance with the output of this second run. Formally, we give the pseudocode for the algorithm in Algorithm 5.

**Algorithm 5** Full algorithm for learning the covariance agnostically.

**function** LEARNCOVARIANCE$(\varepsilon, \tau, X_1, \ldots, X_N)$
  Let $u \leftarrow$ FINDAPPROXCOVARIANCE$(\varepsilon, \tau, X_1, \ldots, X_N)$.
  Let $\Sigma_1 = \sum_{i=1}^N u_i^- X_i X_i^T$.
  **for** $i = 1, \ldots, N$ **do**
    **if** $\|\Sigma_1^{-1/2} X_i\|_2^2 \ge \Omega(d\log N/\tau)$ **then**
      Remove $X_i$ from the set of samples.
  Let $S'$ be the set of pruned samples.
  Let $u' \leftarrow$ FINDAPPROXCOVARIANCE$(\varepsilon, \tau, \{X_i\}_{i \in S'})$.
  **return** $\sum_{i=1}^N u_i' X_i X_i^T$.

We now show that this algorithm is correct.

THEOREM 4.35. *Let $1/2 \ge \varepsilon > 0$, and let $\tau > 0$. Let $\delta = O(\varepsilon \log 1/\varepsilon)$. Let $X_1, \ldots, X_N$ be a $\varepsilon$-corrupted set of samples from $\mathcal{N}(0,\Sigma)$ where*

$$N = \widetilde{\Omega}\left(\frac{d^2 \log^5 1/\tau}{\varepsilon^2}\right).$$

*Let $\widehat{\Sigma}$ be the output of* LEARNCOVARIANCE$(\varepsilon, \tau, X_1, \ldots, X_N)$. *Then with probability $1-\tau$, $\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I\|_F \le O(\delta)$.*

*Proof.* We first condition on the event that we satisfy (12)–(15) with $\delta_1, \delta_2 \le O(\delta)$ and $\delta_3 \le O(\delta \log 1/\varepsilon)$. By our choice of $N$, Fact 4.6, Corollaries 4.7 and 4.9, and Theorem 4.13, and a union bound, we know that this event happens with probability $1-\tau$.

By Theorem 4.32 and Lemmas 4.33 and 4.34, we have that since $\varepsilon$ is sufficiently small,

$$\frac{1}{2}\Sigma \preceq \Sigma_1 \preceq 2\Sigma \ .$$

In particular, this implies that for every vector $X_i$, we have

$$\frac{1}{2}\|\Sigma^{-1/2}X_i\|_2^2 \leq \|\Sigma_1^{-1/2}X_i\|_2^2 \leq 2\|\Sigma^{-1/2}X_i\|_2^2 \ .$$

Therefore, by (12), we know that in line 6 of Algorithm 5, we never throw out any uncorrupted points, and moreover, if $X_i$ is corrupted with $\|\Sigma^{-1/2}X_i\|_2^2 \geq \Omega(d\log N/\tau)$, then it is thrown out. Thus, let $S'$ be the set of pruned points. Because no uncorrupted point is thrown out, we have that $|S'| \geq (1-2\varepsilon)N$, and moreover, this set of points still satisfies (13)–(15),[3] and moreover, for every $i \in S'$, we have $\|\Sigma^{-1/2}X_i\|_2^2 \leq O(d\log N/\tau)$. Therefore, by Theorem 4.32, we have that there is some $u'' \in C_{c|I|}$ such that $\|u'-u''\|_\infty < \varepsilon/(Nd\log(N/\tau))$. But now if $\widehat{\Sigma} = \sum_{i\in|I|} u_i' X_i X_i^T$, we have

$$\begin{aligned}\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I\|_F &\leq \left\|\sum_{i\in I} u_i''\Sigma^{-1/2}X_iX_i^T\Sigma^{-1/2} - I\right\|_F + \sum_{i\in I}|u_i'-u_i'|\|\Sigma^{-1/2}X_i\|_2^2 \\ &\leq c\delta + O(\varepsilon) \leq O(\delta)\ ,\end{aligned}$$

which completes the proof. ∎

**4.5. Learning an arbitrary Gaussian agnostically.** We have shown how to agnostically learn the mean of a Gaussian with known covariance, and we have shown how to agnostically learn the covariance of a mean 0 Gaussian. In this section, we show how to use these two in conjunction to agnostically learn an arbitrary Gaussian. Throughout, let $X_1,\ldots,X_N$ be an $\varepsilon$-corrupted set of samples from $\mathcal{N}(\mu,\Sigma)$, where both $\mu$ and $\Sigma$ are unknown. We will set

$$\widetilde{\Omega}\left(\frac{d^2\log^5 1/\tau}{\varepsilon^2}\right) \ .$$

**4.5.1. From unknown mean, unknown covariance, to zero mean, unknown covariance.** We first show a simple trick which, at the price of doubling the amount of error, allows us to assume that the mean is zero, without changing the covariance. We do so as follows: for each $i=1,\ldots,N/2$, let $X_i' = (X_i - X_{N/2+i})/\sqrt{2}$. Observe that if both $X_i$ and $X_{N/2+i}$ are uncorrupted, then $X_i' \sim \mathcal{N}(0,\Sigma)$. Moreover, observe that $X_i'$ is corrupted only if either $X_i$ or $X_{N/2+i}$ is corrupted. Then we see that if $X_1,\ldots,X_N$ is $\varepsilon$-corrupted, then the $X_1',\ldots,X_{N/2}'$ is an $N/2$-sized set of samples which is $2\varepsilon$-corrupted. Thus, by using the results from section 4.4, with probability $1-\tau$, we can recover a $\widehat{\Sigma}$ such that

$$\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I\|_F \leq O(\varepsilon\log 1/\varepsilon)\ , \tag{20}$$

which, in particular by Corollary 2.14, implies that

$$d_{\mathrm{TV}}(\mathcal{N}(0,\widehat{\Sigma}),\mathcal{N}(0,\Sigma)) \leq O(\varepsilon\log 1/\varepsilon)\ . \tag{21}$$

[3]Technically, the samples satisfy a slightly different set of conditions since we may have thrown out some corrupted points, and so, in particular, the number of samples may have changed, but the meaning should be clear.

**4.5.2. From unknown mean, approximate covariance, to approximate recovery.** For each $X_i$, let $X_i'' = \widehat{\Sigma}^{-1/2} X_i$. Then, for $X_i$ which is not corrupted, we have that $X_i'' \sim \mathcal{N}(\widehat{\Sigma}^{-1/2}\mu, \Sigma_1)$, where $\Sigma_1 = \widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2}$. By Corollary 2.14 and Lemma 4.25, if (20) holds, then we have

$$d_{\mathrm{TV}}(\mathcal{N}(\widehat{\Sigma}^{-1/2}\mu, \Sigma_1), \mathcal{N}(\widehat{\Sigma}^{-1/2}\mu, I)) \leq O(\varepsilon \log 1/\varepsilon) \ .$$

By Claim 2.5, this means that if (20) holds, the uncorrupted set of $X_i''$ can be treated as an $O(\varepsilon \log 1/\varepsilon)$-corrupted set of samples from $\mathcal{N}(\widehat{\Sigma}^{-1/2}\mu, I)$. Thus, if (20) holds, the entire set of samples $X_1'', \ldots, X_m''$ is an $O(\varepsilon \log 1/\varepsilon)$-corrupted set of samples from $\mathcal{N}(\widehat{\Sigma}^{-1/2}\mu, I)$. Then, by using results from section 4.3, with probability $1-\tau$, assuming that (20) holds, we can recover a $\widehat{\mu}$ such that $\|\widehat{\mu} - \widehat{\Sigma}^{-1/2}\mu\|_2 \leq O(\varepsilon \log^{3/2}(1/\varepsilon))$. Thus, by Corollary 2.13, this implies that

$$d_{\mathrm{TV}}(\mathcal{N}(\widehat{\mu}, I), \mathcal{N}(\widehat{\Sigma}^{-1/2}\mu, I)) \leq O(\varepsilon \log^{3/2}(1/\varepsilon)) \ ,$$

or equivalently,

$$d_{\mathrm{TV}}(\mathcal{N}(\widehat{\Sigma}^{1/2}\widehat{\mu}, \widehat{\Sigma}), \mathcal{N}(\mu, \widehat{\Sigma})) \leq O(\varepsilon \log^{3/2}(1/\varepsilon)) \ ,$$

which in conjunction with (21), implies that

$$d_{\mathrm{TV}}(\mathcal{N}(\widehat{\Sigma}^{1/2}\widehat{\mu}, \widehat{\Sigma}), \mathcal{N}(\mu, \Sigma)) \leq O(\varepsilon \log^{3/2}(1/\varepsilon)) \ ,$$

and thus by following this procedure, whose formal pseudocode is given in Algorithm 6, we have shown in Theorem 4.36.

**Algorithm 6** Algorithm for learning an arbitrary Gaussian robustly.

```
function RecoverRobustGuassian(ε, τ, X_1, ..., X_N)
    For i = 1, ..., N/2, let X'_i = (X_i − X_{N/2+i})/√2.
    Let Σ̂ ← LearnCovariance(ε, τ, X'_1, ..., X'_{N/2}).
    For i = 1, ..., N, let X''_i = Σ̂^{−1/2} X_i.
    Let μ̂ ← LearnMean(ε, τ, X''_1, ..., X''_N).
    return the Gaussian with mean Σ̂^{1/2} μ̂, and covariance Σ̂.
```

THEOREM 4.36. *Fix $\varepsilon, \tau > 0$. Let $X_1, \ldots, X_N$ be an $\varepsilon$-corrupted set of samples from $\mathcal{N}(\mu, \Sigma)$, where $\mu, \Sigma$ are both unknown, and*

$$N = \widetilde{\Omega}\left(\frac{d^2 \log^5 1/\tau}{\varepsilon^2}\right) \ .$$

*There is a polynomial-time algorithm* RecoverRobustGaussian$(\varepsilon, \tau, X_1, \ldots, X_N)$ *which with probability $1-\tau$, outputs a $\widehat{\Sigma}, \widehat{\mu}$ such that*

$$d_{\mathrm{TV}}(\mathcal{N}(\widehat{\Sigma}^{1/2}\widehat{\mu}, \widehat{\Sigma}), \mathcal{N}(\mu, \Sigma)) \leq O(\varepsilon \log^{3/2}(1/\varepsilon)) \ .$$

## 5. Agnostically learning a Gaussian, via filters.

**5.1. Learning a Gaussian with unknown mean.** In this section, we use our filter technique to give an agnostic learning algorithm for an unknown mean Gaussian with known covariance matrix. More specifically, we prove the following theorem.

THEOREM 5.1. *Let $G$ be a Gaussian distribution on $\mathbb{R}^d$ with mean $\mu^G$, covariance matrix $I$, and $\varepsilon, \tau > 0$. Let $S'$ be an $\varepsilon$-corrupted set of samples from $G$ of size $\Omega((d/\varepsilon^2) \operatorname{poly}\log(d/\varepsilon\tau))$. There exists an efficient algorithm that, on input $S'$ and $\varepsilon > 0$, returns a mean vector $\widehat{\mu}$ such that with probability at least $1-\tau$ we have $\|\widehat{\mu} - \mu^G\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$.*

**Notation.** We will denote $\mu^S = \frac{1}{|S|}\sum_{X\in S} X$ and $M_S = \frac{1}{|S|}\sum_{X\in S}(X-\mu^G)(X-\mu^G)^T$ for the sample mean and modified sample covariance matrix of the set $S$.

We start by defining our notion of good sample, i.e., a set of conditions on the uncorrupted set of samples under which our algorithm will succeed.

DEFINITION 5.2. *Let $G$ be an identity covariance Gaussian in $d$ dimensions with mean $\mu^G$ and covariance matrix $I$, and $\varepsilon, \tau > 0$. We say that a multiset $S$ of elements in $\mathbb{R}^d$ is* $(\varepsilon, \tau)$-good with respect to $G$ *if the following conditions are satisfied:*

(i) *For all $x \in S$ we have $\|x - \mu^G\|_2 \le O(\sqrt{d\log(|S|/\tau)})$.*
(ii) *For every affine function $L : \mathbb{R}^d \to \mathbb{R}$ such that $L(x) = v\cdot(x-\mu^G) - T$, $\|v\|_2 = 1$, we have that $|\Pr_{X\in_u S}[L(X)\ge 0] - \Pr_{X\sim G}[L(X)\ge 0]| \le \frac{\varepsilon}{T^2\log\left(d\log(\frac{d}{\varepsilon\tau})\right)}$.*
(iii) *We have that $\|\mu^S - \mu^G\|_2 \le \varepsilon$.*
(iv) *We have that $\|M_S - I\|_2 \le \varepsilon$.*

We show in Appendix B that a sufficiently large set of independent samples from $G$ is $(\varepsilon, \tau)$-good (with respect to $G$) with high probability. Specifically, we prove the following lemma.

LEMMA 5.3. *Let $G$ be a Gaussian distribution with identity covariance, and $\varepsilon, \tau > 0$. If the multiset $S$ is obtained by taking $\Omega((d/\varepsilon^2) \operatorname{poly}\log(d/\varepsilon\tau))$ independent samples from $G$, it is $(\varepsilon, \tau)$-good with respect to $G$ with probability at least $1-\tau$.*

We require the following definition that quantifies the extent to which a multiset has been corrupted.

DEFINITION 5.4. *Given finite multisets $S$ and $S'$, we let $\Delta(S, S')$ be the size of the symmetric difference of $S$ and $S'$ divided by the cardinality of $S$.*

As in the convex program case, we will first use NAIVEPRUNE to remove points which are far from the mean. Then, we iterate the algorithm whose performance guarantee is given by the following proposition.

PROPOSITION 5.5. *Let $G$ be a Gaussian distribution on $\mathbb{R}^d$ with mean $\mu^G$, covariance matrix $I$, $\varepsilon > 0$ sufficiently small, and $\tau > 0$. Let $S$ be an $(\varepsilon, \tau)$-good set with respect to $G$. Let $S'$ be any multiset with $\Delta(S, S') \le 2\varepsilon$ and for any $x, y \in S'$, $\|x - y\|_2 \le O(\sqrt{d\log(d/\varepsilon\tau)})$. There exists a polynomial-time algorithm* FILTER-GAUSSIAN-UNKNOWN-MEAN *that, given $S'$ and $\varepsilon > 0$, returns one of the following:*

(i) *a mean vector $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu^G\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$,*
(ii) *a multiset $S'' \subseteq S'$ such that $\Delta(S, S'') \le \Delta(S, S') - \varepsilon/\alpha$, where $\alpha \stackrel{\text{def}}{=} d \log\left(\frac{d}{\varepsilon\tau}\right)\log\left(d\log(\frac{d}{\varepsilon\tau})\right)$.*

We start by showing how Theorem 5.1 follows easily from Proposition 5.5.

*Proof of Theorem* 5.1. By the definition of $\Delta(S, S')$, since $S'$ has been obtained from $S$ by corrupting an $\varepsilon$-fraction of the points in $S$, we have that $\Delta(S, S') \le 2\varepsilon$. By

Lemma 5.3, the set $S$ of uncorrupted samples is $(\varepsilon, \tau)$-good with respect to $G$ with probability at least $1-\tau$. We henceforth condition on this event.

Since $S$ is $(\varepsilon, \tau)$-good, all $x \in S$ have $\|x-\mu^G\|_2 \le O(\sqrt{d \log |S|/\tau})$. Thus, the NaivePrune procedure does not remove from $S'$ any member of $S$. Hence, its output, $S''$, has $\Delta(S, S'') \le \Delta(S, S')$ and for any $x \in S''$, there is a $y \in S$ with $\|x-y\|_2 \le O(\sqrt{d \log |S|/\tau})$. By the triangle inequality, for any $x, z \in S''$, $\|x-z\|_2 \le O(\sqrt{d \log |S|/\tau}) = O(\sqrt{d \log (d/\varepsilon\tau)})$.

Then, we iteratively apply the Filter-Gaussian-Unknown-Mean procedure of Proposition 5.5 until it terminates returning a mean vector $\mu$ with $\|\widehat{\mu}-\mu^G\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$. We claim that we need at most $O(\alpha)$ iterations for this to happen. Indeed, the sequence of iterations results in a sequence of sets $S_i'$, such that $\Delta(S, S_i') \le \Delta(S, S') - i \cdot \varepsilon/\alpha$. Thus, if we do not output the empirical mean in the first $2\alpha$ iterations, in the next iteration there are no outliers left. Hence in the next iteration it is impossible for the algorithm to output a subset satisfying condition (ii) of Proposition 5.5, so it must output a mean vector satisfying (i), as desired. ∎

**5.1.1. Algorithm Filter-Gaussian-Unknown-Mean: Proof of Proposition 5.5.** In this subsection, we describe the efficient algorithm establishing Proposition 5.5 and prove its correctness. Our algorithm calculates the empirical mean vector $\mu^{S'}$ and empirical covariance matrix $\Sigma$. If the matrix $\Sigma$ has no large eigenvalues, it returns $\mu^{S'}$. Otherwise, it uses the eigenvector $v^*$ corresponding to the maximum magnitude eigenvalue of $\Sigma$ and the mean vector $\mu^{S'}$ to define a filter. Our efficient filtering procedure is presented in detailed pseudocode below.

**Algorithm 7** Filter algorithm for a Gaussian with unknown mean and identity covariance.

1: **procedure** Filter-Gaussian-Unknown-Mean$(S', \varepsilon, \tau)$
**input:** A multiset $S'$ such that there exists an $(\varepsilon, \tau)$-good $S$ with $\Delta(S, S') \le 2\varepsilon$
**output:** Multiset $S''$ or mean vector $\widehat{\mu}$ satisfying Proposition 5.5
2: Compute the sample mean $\mu^{S'} = \mathbb{E}_{X \in_u S'}[X]$ and the sample covariance matrix $\Sigma$ , i.e., $\Sigma = (\Sigma_{i,j})_{1 \le i,j \le d}$ with $\Sigma_{i,j} = \mathbb{E}_{X \in_u S'}[(X_i - \mu_i^{S'})(X_j - \mu_j^{S'})]$.
3: Compute approximations for the largest absolute eigenvalue of $\Sigma - I$, $\lambda^* := \|\Sigma - I\|_2$, and the associated unit eigenvector $v^*$.
4: **if** $\|\Sigma - I\|_2 \le O(\varepsilon \log(1/\varepsilon))$ **then return** $\mu^{S'}$.
5: Let $\delta := 3\sqrt{\varepsilon\|\Sigma - I\|_2}$. Find $T > 0$ such that

$$\Pr_{X \in_u S'}\left[|v^* \cdot (X - \mu^{S'})| > T + \delta\right] > 8\exp(-T^2/2) + 8\frac{\varepsilon}{T^2 \log\left(d \log(\frac{d}{\varepsilon\tau})\right)}.$$

6: **return** the multiset $S'' = \{x \in S' : |v^* \cdot (x - \mu^{S'})| \le T + \delta\}$.

By definition, there exist disjoint multisets $L, E$, of points in $\mathbb{R}^d$, where $L \subset S$, such that $S' = (S \setminus L) \cup E$. With this notation, we can write $\Delta(S, S') = \frac{|L|+|E|}{|S|}$. Our assumption $\Delta(S, S') \le 2\varepsilon$ is equivalent to $|L| + |E| \le 2\varepsilon \cdot |S|$, and the definition of $S'$ directly implies that $(1-2\varepsilon)|S| \le |S'| \le (1+2\varepsilon)|S|$. Throughout the proof, we assume that $\varepsilon$ is a sufficiently small constant.

We define $\mu^G, \mu^S, \mu^{S'}, \mu^L$, and $\mu^E$ to be the means of $G, S, S', L$, and $E$, respectively.

Our analysis will make essential use of the following matrices:

- $M_{S'}$ denotes $\mathbb{E}_{X \in_u S'}[(X - \mu^G)(X - \mu^G)^T]$,

- $M_S$ denotes $\mathbb{E}_{X\in_u S}[(X-\mu^G)(X-\mu^G)^T]$,
- $M_L$ denotes $\mathbb{E}_{X\in_u L}[(X-\mu^G)(X-\mu^G)^T]$, and
- $M_E$ denotes $\mathbb{E}_{X\in_u E}[(X-\mu^G)(X-\mu^G)^T]$.

Our analysis will hinge on proving the important claim that $\Sigma - I$ is approximately $(|E|/|S'|)M_E$. This means two things for us. First, it means that if the positive errors align in some direction (causing $M_E$ to have a large eigenvalue), there will be a large eigenvalue in $\Sigma - I$. Second, it says that any large eigenvalue of $\Sigma - I$ will correspond to an eigenvalue of $M_E$, which will give an explicit direction in which many error points are far from the empirical mean.

**Useful structural lemmas.** We will use the following simple fact about the concentration of Gaussian random variables.

FACT 5.6. *If $G$ is Gaussian on $\mathbb{R}^d$ with mean vector $\mu$, then for any unit vector $v \in \mathbb{R}^d$ we have that* $\Pr_{X\sim G}[|v\cdot(X-\mu)| \geq T] \leq \exp(-t^2/2)$.

We begin by noting that we have concentration bounds on $G$ and therefore on $S$, due to its goodness.

FACT 5.7. *Let $w \in \mathbb{R}^d$ be any unit vector. Then for any $T > 0$,*

$$\Pr_{X\sim G}\left[|w\cdot(X-\mu^G)| > T\right] \leq 2\exp(-T^2/2)$$

*and*

$$\Pr_{X\in_u S}\left[|w\cdot(X-\mu^G)| > T\right] \leq 2\exp(-T^2/2) + \frac{\varepsilon}{T^2\log\left(d\log(\frac{d}{\varepsilon\tau})\right)}.$$

*Proof.* The first line is Fact 5.6, and the second follows from it using the goodness of $S$. ∎

By using the above fact, we obtain the following simple claim.

CLAIM 5.8. *Let $w \in \mathbb{R}^d$ be any unit vector. Then for any $T > 0$, we have that*

$$\Pr_{X\sim G}[|w\cdot(X-\mu^{S'})| > T + \|\mu^{S'}-\mu^G\|_2] \leq 2\exp(-T^2/2)$$

*and*

$$\Pr_{X\in_u S}[|w\cdot(X-\mu^{S'})| > T + \|\mu^{S'}-\mu^G\|_2] \leq 2\exp(-T^2/2) + \frac{\varepsilon}{T^2\log\left(d\log(\frac{d}{\varepsilon\tau})\right)}.$$

*Proof.* This follows from Fact 5.7 upon noting that $|w\cdot(X-\mu^{S'})| > T + \|\mu^{S'}-\mu^G\|_2$ only if $|w\cdot(X-\mu^G)| > T$. ∎

We can use the above facts to prove concentration bounds for $L$. In particular, we have the following lemma.

LEMMA 5.9. *We have that* $\|M_L\|_2 = O\left(\log(|S|/|L|) + \varepsilon|S|/|L|\right)$.

*Proof.* Since $L \subseteq S$, for any $x \in \mathbb{R}^d$, we have that

$$|S| \cdot \Pr_{X\in_u S}(X = x) \geq |L| \cdot \Pr_{X\in_u L}(X = x)\,. \tag{22}$$

Since $M_L$ is a symmetric matrix, we have $\|M_L\|_2 = \max_{\|v\|_2=1}|v^T M_L v|$. So, to bound $\|M_L\|_2$ it suffices to bound $|v^T M_L v|$ for unit vectors $v$. By definition of $M_L$, for any $v \in \mathbb{R}^d$ we have that

$$|v^T M_L v| = \mathop{\mathbb{E}}_{X\in_u L}[|v\cdot(X-\mu^G)|^2].$$

For unit vectors $v$, the right-hand side (RHS) is bounded from above as follows:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{X\in_u L}\left[|v\cdot(X-\mu^G)|^2\right]\\
&= 2\int_0^\infty \Pr_{X\in_u L}\left[|v\cdot(X-\mu^G)|>T\right]TdT\\
&= 2\int_0^{O(\sqrt{d\log(d/\varepsilon\tau)})} \Pr_{X\in_u L}[|v\cdot(X-\mu^G)|>T]TdT\\
&\le 2\int_0^{O(\sqrt{d\log(d/\varepsilon\tau)})} \min\left\{1,\frac{|S|}{|L|}\cdot\Pr_{X\in_u S}\left[|v\cdot(X-\mu^G)|>T\right]\right\}TdT\\
&\ll \int_0^{4\sqrt{\log(|S|/|L|)}} TdT\\
&+(|S|/|L|)\int_{4\sqrt{\log(|S|/|L|)}}^{O(\sqrt{d\log(d/\varepsilon\tau)})}\Big(\exp(-T^2/2)+\frac{\varepsilon}{T^2\log\big(d\log(\frac{d}{\varepsilon\tau})\big)}\Big)TdT\\
&\ll \log(|S|/|L|)+\varepsilon\cdot|S|/|L|\ ,
\end{aligned}
$$

where the third line follows from the fact that $\|v\|_2=1$, $L\subset S$, and $S$ satisfies condition (i) of Definition 5.2; the fourth line follows from (22); and the fifth line follows from Fact 5.7. ∎

As a corollary, we can relate the matrices $M_{S'}$ and $M_E$, in spectral norm.

COROLLARY 5.10. *We have that* $M_{S'}-I=(|E|/|S'|)M_E+O(\varepsilon\log(1/\varepsilon))$, *where the* $O(\varepsilon\log(1/\varepsilon))$ *term denotes a matrix of spectral norm* $O(\varepsilon\log(1/\varepsilon))$.

*Proof.* By definition, we have that $|S'|M_{S'}=|S|M_S-|L|M_L+|E|M_E$. Thus, we can write

$$
\begin{aligned}
M_{S'}&=(|S|/|S'|)M_S-(|L|/|S'|)M_L+(|E|/|S'|)M_E\\
&=I+O(\varepsilon)+O(\varepsilon\log(1/\varepsilon))+(|E|/|S'|)M_E\ ,
\end{aligned}
$$

where the second line uses the fact that $1-2\varepsilon\le|S|/|S'|\le1+2\varepsilon$, the goodness of $S$ (condition (iv) in Definition 5.2), and Lemma 5.9. Specifically, Lemma 5.9 implies that $(|L|/|S'|)\|M_L\|_2=O(\varepsilon\log(1/\varepsilon))$. Therefore, we have that

$$M_{S'}=I+(|E|/|S'|)M_E+O(\varepsilon\log(1/\varepsilon))\ ,$$

as desired. ∎

We now establish a similarly useful bound on the difference between the mean vectors.

LEMMA 5.11. *We have that* $\mu^{S'}-\mu^G=(|E|/|S'|)(\mu^E-\mu^G)+O(\varepsilon\sqrt{\log(1/\varepsilon)})$, *where the* $O(\varepsilon\sqrt{\log(1/\varepsilon)})$ *term denotes a vector with* $\ell_2$*-norm at most* $O(\varepsilon\sqrt{\log(1/\varepsilon)})$.

*Proof.* By definition, we have that

$$|S'|(\mu^{S'}-\mu^G)=|S|(\mu^S-\mu^G)-|L|(\mu^L-\mu^G)+|E|(\mu^E-\mu^G).$$

Since $S$ is a good set, by condition (iii) of Definition 5.2, we have $\|\mu^S-\mu^G\|_2=O(\varepsilon)$. Since $1-2\varepsilon\le|S|/|S'|\le1+2\varepsilon$, it follows that $(|S|/|S'|)\|\mu^S-\mu^G\|_2=O(\varepsilon)$. Using the valid inequality $\|M_L\|_2 \geq \|\mu^L - \mu^G\|_2^2$ and Lemma 5.9, we obtain that $\|\mu^L - \mu^G\|_2 \leq O\left(\sqrt{\log(|S|/|L|)} + \sqrt{\varepsilon|S|/|L|}\right)$. Therefore,

$$(|L|/|S'|)\|\mu^L - \mu^G\|_2 \leq O\left((|L|/|S|)\sqrt{\log(|S|/|L|)} + \sqrt{\varepsilon|L|/|S|}\right) = O(\varepsilon\sqrt{\log(1/\varepsilon)}) .$$

In summary,

$$\mu^{S'} - \mu^G = (|E|/|S'|)(\mu^E - \mu^G) + O(\varepsilon\sqrt{\log(1/\varepsilon)}) ,$$

as desired. This completes the proof of the lemma. ∎

By combining the above, we can conclude that $\Sigma - I$ is approximately proportional to $M_E$. More formally, we obtain the following corollary.

COROLLARY 5.12. *We have* $\Sigma - I = (|E|/|S'|)M_E + O(\varepsilon\log(1/\varepsilon)) + O(|E|/|S'|)^2 \|M_E\|_2$, *where the additive terms denote matrices of appropriately bounded spectral norm.*

*Proof.* By definition, we can write $\Sigma - I = M_{S'} - I - (\mu^{S'} - \mu^G)(\mu^{S'} - \mu^G)^T$. Using Corollary 5.10 and Lemma 5.11, we obtain

$$\begin{aligned}\Sigma - I &= (|E|/|S'|)M_E + O(\varepsilon\log(1/\varepsilon)) + O((|E|/|S'|)^2\|\mu^E - \mu^G\|_2^2) + O(\varepsilon^2\log(1/\varepsilon)) \\ &= (|E|/|S'|)M_E + O(\varepsilon\log(1/\varepsilon)) + O(|E|/|S'|)^2\|M_E\|_2 ,\end{aligned}$$

where the second line follows from the valid inequality $\|M_E\|_2 \geq \|\mu^E - \mu^G\|_2^2$. This completes the proof. ∎

**Case of small spectral norm.** We are now ready to analyze the case that the mean vector $\mu^{S'}$ is returned by the algorithm in step 4 of Algorithm 7. In this case, we have that $\lambda^* \stackrel{\text{def}}{=} \|\Sigma - I\|_2 = O(\varepsilon\log(1/\varepsilon))$. Hence, Corollary 5.12 yields that

$$(|E|/|S'|)\|M_E\|_2 \leq \lambda^* + O(\varepsilon\log(1/\varepsilon)) + O(|E|/|S'|)^2\|M_E\|_2 ,$$

which, in turn, implies that

$$(|E|/|S'|)\|M_E\|_2 = O(\varepsilon\log(1/\varepsilon)) .$$

On the other hand, since $\|M_E\|_2 \geq \|\mu^E - \mu^G\|_2^2$, Lemma 5.11 gives that

$$\|\mu^{S'} - \mu^G\|_2 \leq (|E|/|S'|)\sqrt{\|M_E\|_2} + O(\varepsilon\sqrt{\log(1/\varepsilon)}) = O(\varepsilon\sqrt{\log(1/\varepsilon)}).$$

This proves part (i) of Proposition 5.5.

**Case of large spectral norm.** We next show the correctness of the algorithm when it returns a filter in step 5 of Algorithm 7.

We start by proving that if $\lambda^* \stackrel{\text{def}}{=} \|\Sigma - I\|_2 > C\varepsilon\log(1/\varepsilon)$, for a sufficiently large universal constant $C$, then a value $T$ satisfying the condition in step 5 of Algorithm 7 exists. We first note that $\|M_E\|_2$ is appropriately large. Indeed, by Corollary 5.12 and the assumption that $\lambda^* > C\varepsilon\log(1/\varepsilon)$ we deduce that

$$(|E|/|S'|)\|M_E\|_2 = \Omega(\lambda^*) . \tag{23}$$

Moreover, using the inequality $\|M_E\|_2 \geq \|\mu^E - \mu^G\|_2^2$ and Lemma 5.11 as above, we get that

$$\|\mu^{S'} - \mu^G\|_2 \leq (|E|/|S'|)\sqrt{\|M_E\|_2} + O(\varepsilon\sqrt{\log(1/\varepsilon)}) \leq \delta/2 , \tag{24}$$

where we used the fact that $\delta \stackrel{\text{def}}{=} \sqrt{\varepsilon\lambda^*} > C'\varepsilon\sqrt{\log(1/\varepsilon)}$.

Suppose for the sake of contradiction that for all $T > 0$ we have that

$$\Pr_{X\in_u S'}\left[|v^* \cdot (X - \mu^{S'})| > T + \delta\right] \leq 8\exp(-T^2/2) + 8\frac{\varepsilon}{T^2 \log\left(d\log(\frac{d}{\varepsilon\tau})\right)} .$$

Using (24), we obtain that for all $T > 0$ we have that

$$\Pr_{X\in_u S'}\left[|v^* \cdot (X - \mu^{G})| > T + \delta/2\right] \leq 8\exp(-T^2/2) + 8\frac{\varepsilon}{T^2 \log\left(d\log(\frac{d}{\varepsilon\tau})\right)} . \tag{25}$$

Since $E \subseteq S'$, for all $x \in \mathbb{R}^d$ we have that $|S'|\Pr_{X\in_u S'}[X = x] \geq |E|\Pr_{Y\in_u E}[Y = x]$. This fact, combined with (25), implies that for all $T > 0$

$$\Pr_{X\in_u E}\left[|v^* \cdot (X - \mu^{G})| > T + \delta/2\right] \leq C(|S'|/|E|)\left(\exp(-T^2/2) + \frac{\varepsilon}{T^2 \log\left(d\log(\frac{d}{\varepsilon\tau})\right)}\right) , \tag{26}$$

for some universal constant $C''$.

We now have the following sequence of inequalities:

$$\begin{aligned}
\|M_E\|_2 &= \mathop{\mathbb{E}}_{X\in_u E}\left[|v^* \cdot (X - \mu^{G})|^2\right] = 2\int_0^\infty \Pr_{X\in_u E}\left[|v^* \cdot (X - \mu^{G})| > T\right] TdT \\
&= 2\int_0^{O(\sqrt{d\log(d/\varepsilon\tau)})} \Pr_{X\in_u E}\left[|v^* \cdot (X - \mu^{G})| > T\right] TdT \\
&\leq 2\int_0^{O(\sqrt{d\log(d/\varepsilon\tau)})} \min\left\{1, \frac{|S'|}{|E|} \cdot \Pr_{X\in_u S'}\left[|v^* \cdot (X - \mu^{G})| > T\right]\right\} TdT \\
&\leq \int_0^{4\sqrt{\log(|S'|/|E|)}+\delta} TdT + C''\frac{|S'|}{|E|}\int_{4\sqrt{\log(|S'|/|E|)}+\delta}^{O(\sqrt{d\log(d/\varepsilon\tau)})} \Bigg(\exp(-T^2/2) \\
&\qquad + \frac{\varepsilon}{T^2 \log\left(d\log(\frac{d}{\varepsilon\tau})\right)}\Bigg) TdT \\
&\leq \int_0^{4\sqrt{\log(|S'|/|E|)}+\delta} TdT \\
&\qquad + C''\frac{|S'|}{|E|}\left(\int_{4\sqrt{\log(|S'|/|E|)}+\delta}^{\infty} \Big(\exp(-T^2/2)\Big) TdT + O(\varepsilon)\right) \\
&\leq \log(|S'|/|E|) + \delta^2 + O(1) + O(\varepsilon) \cdot |S'|/|E| \\
&\leq \log(|S'|/|E|) + \varepsilon\lambda^* + O(\varepsilon) \cdot |S'|/|E| \; .
\end{aligned}$$

Rearranging the above, we get that

$$(|E|/|S'|)\|M_E\|_2 \ll (|E|/|S'|)\log(|S'|/|E|) + (|E|/|S'|)\varepsilon\lambda^* + O(\varepsilon) = O(\varepsilon\log(1/\varepsilon) + \varepsilon^2\lambda^*).$$

Combining this with (23), we obtain $\lambda^* = O(\varepsilon\log(1/\varepsilon))$, which is a contradiction if $C$ is sufficiently large. Therefore, it must be the case that for some value of $T$ the condition in step 5 of Algorithm 7 is satisfied.

The following claim completes the proof of Proposition 5.5.

CLAIM 5.13. *Fix* $\alpha \stackrel{\text{def}}{=} d\log(d/\varepsilon\tau)\log(d\log(\frac{d}{\varepsilon\tau}))$. *We have that* $\Delta(S, S'') \leq \Delta(S, S') - 2\varepsilon/\alpha$ .

*Proof.* Recall that $S' = (S \setminus L) \cup E$, with $E$ and $L$ disjoint multisets such that $L \subset S$. We can similarly write $S'' = (S \setminus L') \cup E'$, with $L' \supseteq L$ and $E' \subset E$. Since

$$\Delta(S, S') - \Delta(S, S'') = \frac{|E \setminus E'| - |L' \setminus L|}{|S|},$$

it suffices to show that $|E \setminus E'| \geq |L' \setminus L| + \varepsilon|S|/\alpha$. Note that $|L' \setminus L|$ is the number of points rejected by the filter that lie in $S \cap S'$. Note that the fraction of elements of $S$ that are removed to produce $S''$ (i.e., satisfy $|v^* \cdot (x - \mu^{S'})| > T + \delta$) is at most $2\exp(-T^2/2) + \varepsilon/\alpha$. This follows from Claim 5.8 and the fact that $T = O(\sqrt{d \log(d/\varepsilon\tau)})$.

Hence, it holds that $|L' \setminus L| \leq (2\exp(-T^2/2) + \varepsilon/\alpha)|S|$. On the other hand, step 5 of Algorithm 7 ensures that the fraction of elements of $S'$ that are rejected by the filter is at least $8\exp(-T^2/2) + 8\varepsilon/\alpha$. Note that $|E \setminus E'|$ is the number of points rejected by the filter that lie in $S' \setminus S$. Therefore, we can write

$$\begin{aligned} |E \setminus E'| &\geq (8\exp(-T^2/2) + 8\varepsilon/\alpha)|S'| - (2\exp(-T^2/2) + \varepsilon/\alpha)|S| \\ &\geq (8\exp(-T^2/2) + 8\varepsilon/\alpha)|S|/2 - (2\exp(-T^2/2) + \varepsilon/\alpha)|S| \\ &\geq (2\exp(-T^2/2) + 3\varepsilon/\alpha)|S| \\ &\geq |L' \setminus L| + 2\varepsilon|S|/\alpha\ , \end{aligned}$$

where the second line uses the fact that $|S'| \geq |S|/2$ and the last line uses the fact that $|L' \setminus L|/|S| \leq 2\exp(-T^2/2) + \varepsilon/\alpha$. Noting that $\log(d/\varepsilon\tau) \geq 1$, this completes the proof of the claim. ∎

**5.2. Learning a Gaussian with unknown covariance.** In this subsection, we use our filter technique to agnostically learn a Gaussian with zero mean vector and unknown covariance. By combining the algorithms of the current and the previous subsections, as in our convex programming approach (section 4.5), we obtain a filter-based algorithm to agnostically learn an arbitrary unknown Gaussian.

The main result of this subsection is the following theorem.

THEOREM 5.14. *Let $G \sim \mathcal{N}(0, \Sigma)$ be a Gaussian in $d$ dimensions with mean 0 and unknown covariance, and let $\varepsilon, \tau > 0$. Let $S$ be an $\varepsilon$-corrupted set of samples from $G$ of size $\Omega((d^2/\varepsilon^2)\,\mathrm{poly}\log(d/\varepsilon\tau))$. There exists an efficient algorithm that, given $S$ and $\varepsilon$, returns the parameters of a Gaussian distribution $G' \sim \mathcal{N}(0, \widehat{\Sigma})$ such that with probability at least $1 - \tau$, it holds that $\|I - \Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}\|_F = O(\varepsilon \log(1/\varepsilon))$.*

As in the previous subsection, we will need a condition on $S$ under which our algorithm will succeed.

DEFINITION 5.15. *Let $G$ be a Gaussian in $\mathbb{R}^d$ with mean 0 and covariance $\Sigma$. Let $\varepsilon > 0$ be sufficiently small. We say that a multiset $S$ of points in $\mathbb{R}^d$ is* $(\varepsilon, \tau)$-good with respect to $G$ *if the following hold:*

1. *For all $x \in S$, $x^T \Sigma^{-1} x < O(d \log(|S|/\tau))$.*
2. *We have that $\|\Sigma^{-1/2}\, Cov(S) \Sigma^{-1/2} - I\|_F = O(\varepsilon)$.*
3. *For all even degree-2 polynomials $p$, we have that $\mathrm{Var}(p(S)) = \mathrm{Var}(p(G))(1 + O(\varepsilon))$.*
4. *For $p$ an even degree-2 polynomial with $\mathbb{E}[p(G)] = 0$ and $\mathrm{Var}(p(G)) = 1$, and for any $T > 10\ln(1/\varepsilon)$ we have that*

$$\Pr_{x \in_u S}(|p(x)| > T) \leq \varepsilon/(T^2 \log^2(T)).$$

Let us first note some basic properties of such polynomials on a normal distribution. The proof of this lemma is deferred to Appendix B.

LEMMA 5.16. *For any even degree-2 polynomial $p : \mathbb{R}^d \to \mathbb{R}$, we can write $p(x) = (\Sigma^{-1/2}x)^T P_2(\Sigma^{-1/2}x) + p_0$, for a $d \times d$ symmetric matrix $P_2$ and $p_0 \in \mathbb{R}$. Then, for $X \sim G$, we have*

1. $\mathbb{E}[p(X)] = p_0 + \mathrm{tr}(P_2)$,
2. $\mathrm{Var}[p(X)] = 2\|P_2\|_F^2$, *and*
3. *for all $T > 1$, $\Pr(|p(X) - \mathbb{E}[p(X)]| \geq T) \leq 2e^{1/3-2T/3\,\mathrm{Var}[p(X)]}$.*
4. *For all $\delta > 0$, $\Pr(|p(X)| \leq \delta^2) \leq O(\delta)$.*

We note that if $S$ is obtained by taking random samples from $G$, then $S$ is good with high probability. The proof of this lemma is also deferred to Appendix B.

LEMMA 5.17. *Let $G$ be a $d$-dimensional Gaussian with mean 0, and let $\varepsilon, \tau > 0$. Let $N$ be a sufficiently large constant multiple of $d^2 \log^5(d/\varepsilon\tau)/\varepsilon^2$. Then a set $S$ of $N$ independent samples from $G$ is $(\varepsilon, \tau)$-good with respect to $G$ with probability at least $1 - \tau$.*

As in Definition 5.4, $\Delta(S, S')$ is the size of the symmetric difference of $S$ and $S'$ divided by $|S|$.

The basic thrust of our algorithm is as follows: By Lemma 5.17, with high probability we have that $S$ is $(\varepsilon, \tau)$-good with respect to $G$. The algorithm is then handed a new set $S'$ such that $\Delta(S, S') \leq 2\varepsilon|S|$. The algorithm will run in stages. In each stage, the algorithm will either output $G'$ or will return a new set $S''$ such that $\Delta(S, S'') < \Delta(S, S')$. In the latter case, the algorithm will recurse on $S''$. We formalize this idea below.

PROPOSITION 5.18. *There is an algorithm that, given a finite set $S' \subset \mathbb{R}^d$, such that there is a mean 0 Gaussian $G$ and a set $S$ that is $(\varepsilon, \tau)$-good with respect to $G$ with $\Delta(S, S') \leq 2\varepsilon|S|$, runs in time $\mathrm{poly}(d \log(1/\tau)/\varepsilon)$ and returns either the parameters of a Gaussian $G'$ with $d_{\mathrm{TV}}(G, G') \leq O(\varepsilon \log(1/\varepsilon))$ or a subset $S''$ of $\mathbb{R}^d$ with $\Delta(S, S'') < \Delta(S, S')$.*

Given Proposition 5.18, the proof of Theorem 5.14 is straightforward. By Lemma 5.17 the original set $S$ is $(\varepsilon, \tau)$-good with respect to $G$ with probability at least $1 - \tau$. Then, $S'$ satisfies the hypotheses of Proposition 5.18. We then repeatedly iterate the algorithm from Proposition 5.18 until it outputs a distribution $G'$ close to $G$. This must eventually happen because at every step the distance between $S$ and the set returned by the algorithm decreases by at least 1.

**5.2.1. Analysis of filter-based algorithm: Proof of Proposition 5.18.** We now turn our attention to the proof of Proposition 5.18. We first define the matrix $\Sigma'$ to be $\mathbb{E}_{X \in S'}[XX^T]$, and let $G'$ be the mean 0 Gaussian with covariance matrix $\Sigma'$. Our goal will be to either obtain a certificate that $G'$ is close to $G$ or to devise a filter that allows us to clean up $S'$ by removing some elements, most of which are not in $S$. The idea here is the following: We know by Corollary 2.14 that $G$ and $G'$ are close unless $I - \Sigma^{-1/2}\Sigma'\Sigma^{-1/2}$ has large Frobenius norm. This happens if and only if there is some matrix $M$ with $\|M\|_F = 1$ such that

$$\mathrm{tr}(M\Sigma^{-1/2}\Sigma'\Sigma^{-1/2} - M) = \mathop{\mathbb{E}}_{X \in_u S'}[(\Sigma^{-1/2}X)^T M(\Sigma^{-1/2}X) - \mathrm{tr}(M)]$$

is far from 0. On the other hand, we know that the distribution of $p(X) = (\Sigma^{-1/2}X)^T M(\Sigma^{-1/2}X) - \mathrm{tr}(M)$ for $X \in_u S$ is approximately that of $p(G)$, which is a variance

$O(1)$ polynomial of Gaussians with mean 0. In order to substantially change the mean of this function, while only changing $S$ at a few points, one must have several points in $S'$ for which $p(X)$ is abnormally large. This, in turn, will imply that the variance of $p(X)$ for $X$ from $S'$ must be large. This phenomenon will be detectable as a large eigenvalue of the matrix of fourth moments of $X \in S'$ (thought of as a matrix over the space of second moments). If such a large eigenvalue is detected, we will have a $p$ with $p(X)$ having large variance. By throwing away from $S'$ elements for which $|p|$ is too large, we will return a cleaner version of $S'$. The algorithm is as follows.

**Algorithm 8** Filter algorithm for a Gaussian with unknown covariance matrix.

**procedure** FILTER-GAUSSIAN-UNKNOWN-COVARIANCE$(S', \varepsilon, \tau)$

**input:** A multiset $S'$ such that there exists an $(\varepsilon, \tau)$-good $S$ with $\Delta(S, S') \leq 2\varepsilon$

**output:** Either a set $S''$ with $\Delta(S, S'') < \Delta(S, S')$ or the parameters of a Gaussian $G'$ with $d_{\mathrm{TV}}(G, G') = O(\varepsilon \log(1/\varepsilon))$

  Let $C > 0$ be a sufficiently large universal constant.

  Let $\Sigma'$ be the matrix $\mathbb{E}_{X \in_u S'}[XX^T]$, and let $G'$ be the mean 0 Gaussian with covariance matrix $\Sigma'$.

  **if** there is any $x \in S'$ such that $x^T (\Sigma')^{-1} x \geq Cd \log(|S'|/\tau)$ **then**

    **return** $S'' = S' \setminus \{x : x^T (\Sigma')^{-1} x \geq Cd \log(|S'|/\tau)\}$.

  Let $L$ be the space of even degree-2 polynomials $p$ such that $\mathbb{E}_{X \sim G'}[p(X)] = 0$.

  Define two quadratic forms on $L$:

  (i) $Q_{G'}(p) = \mathbb{E}[p^2(G')]$ ,

  (ii) $Q_{S'}(p) = \mathbb{E}_{X \in_u S'}[p^2(X)]$ .

  Computing $\max_{p \in L \setminus \{0\}} Q_{S'}(p)/Q_{G'}(p)$ and the associated polynomial $p^*(x)$ normalized such that $Q_{G'}(p^*) = 1$ using FIND-MAX-POLY below.

  **if** $Q_{S'}(p^*) \leq (1 + C\varepsilon \log^2(1/\varepsilon)) Q_{G'}(p^*)$ **then**

    **return** $G'$

  Let $\mu$ be the median value of $p^*(X)$ over $X \in S'$.

  Find a $T \geq C'$ such that

$$\Pr_{X \in_u S'}(|p^*(X) - \mu| \geq T + 3) \geq \mathrm{Tail}(T, d, \varepsilon, \tau) ,$$

  where $\mathrm{Tail}(T, d, \varepsilon, \tau) = 3\varepsilon/(T^2 \log^2(T))$ when $T \geq 10 \ln(1/\varepsilon)$, and $\mathrm{Tail}(T, d, \varepsilon, \tau) = 1$ when $T < 10 \log(1/\varepsilon)$.

  **return** $S'' = \{X \in S' : |p^*(X) - \mu| < T\}$.

The function FIND-MAX-POLY uses similar notation to SEPARATIONORACLE-UNKNOWNCOVARIANCE, such that FILTER-GAUSSIAN-UNKNOWN-COVARIANCE and SEPARATIONORACLEUNKNOWNCOVARIANCE can be more easily compared.

Let us first show that FIND-MAX-POLY is correct.

CLAIM 5.19. *Algorithm* FIND-MAX-POLY *is correct and* FILTER-GAUSSIAN-UNKNOWN-COVARIANCE *runs in time* $\mathrm{poly}(d \log \tau/\varepsilon)$.

*Proof.* First, assume that we can compute all eigenvalues and eigenvectors exactly. By Lemma 5.16 all even polynomials with degree-2 that have $\mathbb{E}_{X \sim G}[p(X)] = 0$ can be written as $p(x) = (\Sigma'^{-1/2}x)^T P_2 (\Sigma'^{-1/2}x) - \mathrm{tr}(P_2)$ for a symmetric matrix $P_2$. If we take $P_2 = v^\sharp/\sqrt{2}$ for a unit vector $v$ such that $v^\sharp$ is symmetric, then $\mathrm{Var}_{X \sim G'}[p(X)] = 2\|P_2\|_F = \|v_2\| = 1$.

**Algorithm 9** Algorithm for maximizing $Q_{S'}(p)/Q_{G'}(p)$.

**function** FIND-MAX-POLY$(S', \Sigma')$
**input:** A multiset $S'$ and a Gaussian $G' = \mathcal{N}(0, \Sigma')$
**output:** The even degree-2 polynomial $p^*(x)$ with $\mathbb{E}_{X\sim G'}[p^*(X)] \approx 0$ and $Q_{G'}(p^*) \approx 1$ that approximately maximizes $Q_{S'}(p^*)$ and this maximum $\lambda^* = Q_{S'}(p^*)$
    Compute an approximate eigendecomposition of $\Sigma'$ and use it to compute $\Sigma'^{-1/2}$.
    Let $x_{(1)}, \ldots, x_{(|S'|)}$ be the elements of $S'$.
    For $i = 1, \ldots, |S'|$, let $y_{(i)} = \Sigma'^{-1/2} x_{(i)}$ and $z_{(i)} = y_{(i)}^{\otimes 2}$.
    Let $T_{S'} = -I^\flat I^{\flat T} + (1/|S'|) \sum_{i=1}^{|S'|} z_{(i)} z_{(i)}^T$.
    Approximate the top eigenvalue $\lambda^*$ and corresponding unit eigenvector $v^*$ of $T_{S'}$.
    Let $p^*(x) = \frac{1}{\sqrt{2}}((\Sigma'^{-1/2}x)^T v^{*\sharp}(\Sigma'^{-1/2}x) - \mathrm{tr}(v^{*\sharp}))$.
    **return** $p^*$ and $\lambda^*/2$.

Note that since the covariance matrix of $S'$ is $\Sigma'$ , we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{X\sim S'}[p(X)] &= \mathop{\mathbb{E}}_{X\sim S'}[(\Sigma'^{-1/2}X)^T P_2 (\Sigma'^{-1/2}X) - \mathrm{tr}(P_2)] \\
&= \mathop{\mathbb{E}}_{X\sim S'}[\mathrm{tr}((XX^T)\Sigma'^{-1/2} P_2 \Sigma'^{-1/2})] - \mathrm{tr}(P_2) \\
&= \mathrm{tr}(\mathop{\mathbb{E}}_{X\sim S'}[(XX^T)]\Sigma'^{-1/2} P_2 \Sigma'^{-1/2}) - \mathrm{tr}(P_2) \\
&= \mathrm{tr}(\Sigma' \Sigma'^{-1/2} P_2 \Sigma'^{-1/2}) - \mathrm{tr}(P_2) = 0\ .
\end{aligned}$$

We let $T'$ be the multiset of $y = \Sigma^{-1/2} x$ for $x \in S'$ and $U'$ the multiset of $z = y^{\otimes 2}$ for $y$ in $T'$. Recall that $P_2^\flat = \sqrt{2} v$. We thus have

$$\begin{aligned}
Q_{S'}(p) := \mathop{\mathbb{E}}_{X\in_u S'}[p(X)^2] &= \mathop{\mathbb{E}}_{Y\sim T'}[(Y^T P_2 Y - \mathrm{tr}(P_2))^2] \\
&= \mathop{\mathbb{E}}_{Y\in_u T'}[(Y^T P_2 Y)^2] + \mathrm{tr}(P_2)^2 - 2\mathrm{tr}(P_2)^2 \\
&= \mathop{\mathbb{E}}_{Y\in_u T'}[\mathrm{tr}((YY^T)P_2)^2] - \mathrm{tr}(P_2 I)^2 - 0 \\
&= \mathop{\mathbb{E}}_{Z\in_u U'}[(Z^T v)^2/2] - (v^T I^\flat)^2/2 \\
&= \mathop{\mathbb{E}}_{Z\in_u U'}[v^T (ZZ^T) v/2] - 2v^T (I^\flat I^{\flat T}) v/2 \\
&= v^T T_{S'} v/2\ .
\end{aligned}$$

Thus, the $p(x)$ that maximizes $Q_{S'}(p)$ is given by the unit vector $v$ that maximizes $v^T T_{S'} v$ subject to $v^\sharp$ being symmetric.

Let $v' = v^{\sharp T \flat}$. Note that $v^T T_{S'} v = v'^T T_{S'} v'$ by symmetries of $T_{S'}$. Thus, by linearity, $v'' = v/2 + v'/2$ also has $v''^T T_{S'} v'' = v^T T_{S'} v$. However, if $v^\sharp$ is not symmetric, $v''$ has $\|v''\|_2 < 1$. Thus, the unit vector $v''/\|v''\|_2$ achieves a higher value of the bilinear form. Consequently, $v^{*\sharp}$ is symmetric.

Now we have that $p^*(x)$ that maximizes $Q_{S'}(p)$ is given by the unit vector $v$ that maximizes $v^T T_{S'} v$. Since $Q_{G'}(p) := \mathbb{E}_{X\sim G'}[p(X)^2] = 2\|P_2\|_F = \|v\|_2 = 1$, this also maximizes $Q_{S'}(p)/Q_{G'}(p)$.

We note that we can achieve $\mathbb{E}_{X\sim G'}[p^*(X)] = O(\varepsilon^2)$ and $\mathbb{E}_{X\sim G'}[(p^*(X))^2] = 1 + O(\varepsilon^2)$ in time $\mathrm{poly}(\varepsilon/d)$ using standard algorithms to compute the eigendecomposition of a symmetric matrix. This suffices for the correctness of the remaining part of Filter-Gaussian-Unknown-Covariance. The other steps in Filter-Gaussian-Unknown-Covariance can be easily done in $\mathrm{poly}(|S'|d\log(1\tau)/\varepsilon)$ time. □

In order to analyze algorithm Filter-Gaussian-Unknown-Covariance, we note that we can write $S' = (S \setminus L) \cup E$, where $L = S \setminus S'$ and $L = S' \setminus S$. It is then the case that $\Delta(S, S') = (|L| + |E|)/|S|$. Since this is small we have that $|L|, |E| = O(\varepsilon|S'|)$. We can also write $\Sigma'$ and $\Sigma_{S\setminus L}((|S|-|L|)/|S'|) + \Sigma_E(|E|/|S'|) = \Sigma_{S\setminus L} + O(\varepsilon)(\Sigma_E - \Sigma_{S\setminus L})$, where $\Sigma_{S\setminus L} = \mathbb{E}_{X\in_u S\setminus L}[XX^T], \Sigma_E = \mathbb{E}_{X\in_u E}[XX^T]$. A critical part of our analysis will be to note that $\Sigma_{S\setminus L}$ is very close to $\Sigma$, and thus that either $\Sigma'$ is very close to $\Sigma$ or else $\Sigma_E$ is very large in some direction.

Lemma 5.20. *We have that*

$$\|I - \Sigma^{-1/2}\Sigma_{S\setminus L}\Sigma^{-1/2}\|_F = O(\varepsilon\log(1/\varepsilon)).$$

To prove Lemma 5.20, we will require the following lemma.

Lemma 5.21. *Let $p(x)$ be an even degree-2 polynomial with $\mathbb{E}_{X\sim G}[p(X)] = 0$ and $\mathrm{Var}_{X\sim G}[p(X)] = 1$. Then, we have that $|L|\,\mathbb{E}_{X\in_u L}[p(X)^2] = O(\varepsilon\log^2(1/\varepsilon)|S|)$ and $|L||\,\mathbb{E}_{X\in_u L}[p(X)]| = O(\varepsilon\log(1/\varepsilon)|S|)$.*

*Proof.* This holds essentially because the distribution of $p(X)$ for $X \in S$ is close to that for $p(G)$, which has rapidly decaying tails. Therefore, throwing away an $\varepsilon$-fraction of the mass cannot change the value of the variance by very much. In particular, we have that

$$\begin{aligned}
|L| \mathop{\mathbb{E}}_{X\in_u L}[p(X)^2] &\le \int_0^\infty |L| \Pr_{X\in_u L}(|p(X)| > T)2T dT \\
&\le \int_0^\infty |S| \min\left(2\varepsilon, \Pr_{X\in_u S}(|p(X)| > T)\right) 2T dT \\
&\le \int_0^{10\ln(1/\varepsilon)} 4\varepsilon|S|T dT + \int_{10\ln(1/\varepsilon)}^\infty 6|S|\varepsilon T/(T^2\log^2(T)) dT \\
&\le O(\varepsilon|S|\log^2(1/\varepsilon)) + \int_{10\ln(1/\varepsilon)}^\infty 6|S|\varepsilon/(T\log^2(T)) dT \\
&= O(\varepsilon|S|\log^2(1/\varepsilon)) + 6\varepsilon|S|/\ln(10\ln(1/\varepsilon)) \\
&= O(\varepsilon\log^2(1/\varepsilon)|S|)\ .
\end{aligned}$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
(|L|/|S|)|\mathop{\mathbb{E}}_{x\in_u L}[p(X)]| &\le (|L|/|S|)\sqrt{\mathop{\mathbb{E}}_{x\in_u L}[p(X)^2]} \le \sqrt{|L|/|S|}\cdot\sqrt{O(\varepsilon\log^2(1/\varepsilon))} \\
&= O(\varepsilon\log(1/\varepsilon)\ .
\end{aligned}$$ □

Now we can prove Lemma 5.20.

*Proof of Lemma* 5.20. Note that, since the matrix inner product is an inner product,

$$\|I - \Sigma^{-1/2}\Sigma_{S\setminus L}\Sigma^{-1/2}\|_F = \sup_{\|M\|_F=1}\left(\mathrm{tr}(M\Sigma^{-1/2}\Sigma_{S\setminus L}\Sigma^{-1/2}) - \mathrm{tr}(M)\right).$$

We need to show that for any $M$ with $\|M\|_F = 1$ that $\mathrm{tr}(M\Sigma^{-1/2}\Sigma_{S\setminus L}\Sigma^{-1/2}) - \mathrm{tr}(M)$ is small.

Since

$$\mathrm{tr}(M\Sigma^{-1/2}\Sigma_{S\setminus L}\Sigma^{-1/2}) = \mathrm{tr}(M^T\Sigma^{-1/2}\Sigma_{S\setminus L}\Sigma^{-1/2}) = \mathrm{tr}(\tfrac{1}{2}(M+M^T)\Sigma^{-1/2}\Sigma_{S\setminus L}\Sigma^{-1/2})$$

and $\|\frac{1}{2}(M+M^T)\|_F \leq \frac{1}{2}(\|M\|_F + \|M^T\|_F) = 1$, we may assume without loss of generality that $M$ is symmetric.

Consider such an $M$. We note that

$$\begin{aligned}\mathrm{tr}(M\Sigma^{-1/2}\Sigma_{S\setminus L}\Sigma^{-1/2}) &= \mathop{\mathbb{E}}_{X\in_u S\setminus L}[\mathrm{tr}(M\Sigma^{-1/2}XX^T\Sigma^{-1/2})] \\ &= \mathop{\mathbb{E}}_{X\in_u S\setminus L}[(\Sigma^{-1/2}X)^T M(\Sigma^{-1/2}X)].\end{aligned}$$

Let $p(x)$ denote the quadratic polynomial

$$p(x) = (\Sigma^{-1/2}x)^T M(\Sigma^{-1/2}x) - \mathrm{tr}(M).$$

By Lemma 5.16, $\mathbb{E}_{X\sim G}[p(X)] = 0$ and $\mathrm{Var}_{X\sim G}[p(X)] = 2\|M\|_F^2 = 2$.

Since $S$ is $(\varepsilon,\tau)$-good with respect to $G$, we have that $\mathbb{E}_{X\in S}[p(X)] = \varepsilon\sqrt{\mathbb{E}_{X\sim G}[p^2(X)]} = O(\varepsilon)$. Therefore, it suffices to show that the contribution from $L$ is small. In particular, it will be enough to show that $(|L|/|S|)|\,\mathbb{E}_{x\in_u L}[p(X)]| \leq O(\varepsilon\log(1/\varepsilon))$. This follows from Lemma 5.21, which completes the proof. ∎

As a corollary of this we note that $\Sigma'$ cannot be too much smaller than $\Sigma$.

Corollary 5.22. *We present*

$$\Sigma' \succeq (1 - O(\varepsilon\log(1/\varepsilon)))\Sigma.$$

*Proof.* Lemma 5.20 implies that $\Sigma^{-1/2}\Sigma_{S\setminus L}\Sigma^{1/2}$ has all eigenvalues in the range $1\pm O(\varepsilon\log(1/\varepsilon))$. Therefore, $\Sigma_{S\setminus L} \succeq (1+O(\varepsilon\log(1/\varepsilon)))\Sigma$. Our result now follows from noting that $\Sigma' = \Sigma_{S\setminus L}((|S|-|L|)/|S'|) + \Sigma_E(|E|/|S'|)$, and $\Sigma_E = \mathbb{E}_{X\in_u E}[XX^T] \geq 0$. ∎

The first step in verifying correctness is to note that if our algorithm returns on step 5 of Algorithm 8 that it does so correctly.

Claim 5.23. *If our algorithm returns on step* 5 *of Algorithm* 8*, then* $\Delta(S,S'') < \Delta(S,S')$.

*Proof.* This is clearly true if we can show that all $x$ removed have $x \notin S$. However, this follows because $(\Sigma')^{-1} \leq 2\Sigma^{-1}$, and therefore, by $(\varepsilon,\tau)$-goodness, all $x \in S$ satisfy

$$x^T(\Sigma')^{-1}x \leq 2x^T\Sigma^{-1}x < Cd\log(N/\tau)$$

for $C$ sufficiently large. ∎

Next, we need to show that if our algorithm returns a $G'$ in step 10 of Algorithm 8, then $d_{\mathrm{TV}}(G,G')$ is small.

Claim 5.24. *If our algorithm returns in step* 10 *of Algorithm* 8*, then* $d_{\mathrm{TV}}(G,G') = O(\varepsilon\log(1/\varepsilon))$.

*Proof.* By Corollary 2.14, it suffices to show that

$$\|I - \Sigma^{-1/2}\Sigma'\Sigma^{-1/2}\|_F = O(\varepsilon\log(1/\varepsilon)).$$

However, we note that

$$\|I - \Sigma^{-1/2}\Sigma'\Sigma^{-1/2}\|_F \le \|I - \Sigma^{-1/2}\Sigma_{S\setminus L}\Sigma^{-1/2}\|_F + (|E|/|S'|)\|I - \Sigma^{-1/2}\Sigma_E\Sigma^{-1/2}\|_F$$
$$\le O(\varepsilon\log(1/\varepsilon)) + (|E|/|S'|)\|I - \Sigma^{-1/2}\Sigma_E\Sigma^{-1/2}\|_F.$$

Therefore, we will have an appropriate bound unless $\|I - \Sigma^{-1/2}\Sigma_E\Sigma^{-1/2}\|_F = \Omega(\log(1/\varepsilon))$.

Next, note that there is a matrix $M$ with $\|M\|_F = 1$ such that

$$\|I - \Sigma^{-1/2}\Sigma_E\Sigma^{-1/2}\|_F = \mathrm{tr}(M\Sigma^{-1/2}\Sigma_E\Sigma^{-1/2} - M)$$
$$= \mathop{\mathbb{E}}_{X\in_u E}[(\Sigma^{-1/2}X)^T M(\Sigma^{-1/2}X) - \mathrm{tr}(M)].$$

Indeed we can take $M = (I - \Sigma^{-1/2}\Sigma_E\Sigma^{-1/2})/\|I - \Sigma^{-1/2}\Sigma_E\Sigma^{-1/2}\|_F$. Thus, there is a symmetric $M$ such that this holds.

We let $p(X)$ be the polynomial

$$p(X) = (\Sigma^{-1/2}X)^T M(\Sigma^{-1/2}X) - \mathrm{tr}(M).$$

Using Lemma 5.16, $\mathbb{E}_{X\sim G}[p(X)] = 0$ and $\mathrm{Var}_{X\sim G}[p(X)] = 2$. Therefore, $p \in L$ and $Q_{G'}(p) = 2$. We now compare this to the size of $Q_{S'}(p)$. On the one hand, we note that using methodology similar to that used in Lemma 5.20 we can show that $\mathbb{E}_{X\in_u S\setminus L}[p^2(X)]$ is not much less than 2. In particular,

$$\mathop{\mathbb{E}}_{X\in_u S\setminus L}[p^2(X)] \ge \left(\mathop{\mathbb{E}}_{X\in_u S}[p^2(X)] - \frac{\sum_{X\in L} p^2(X)}{|S|}\right).$$

On the one hand, we have that

$$\mathop{\mathbb{E}}_{X\in_u S}[p^2(X)] \le \mathbb{E}[p^2(G)](1+\varepsilon) = 2 + O(\varepsilon)\ ,$$

by assumption. On the other hand, by Lemma 5.21, we have $|L|\,\mathbb{E}_{X\in_u L}[p^2(X)]/|S| \le O(\varepsilon\log^2(1/\varepsilon))$.

Therefore, we have that $\mathbb{E}_{X\in_u S\setminus L}[p^2(X)] = 2 + O(\varepsilon\log^2(1/\varepsilon))$. Since, by assumption $Q_{S'}(p) \le 2 + O(\varepsilon\log^2(1/\varepsilon))$, this implies that $(|E|/|S'|)\,\mathbb{E}_{X\in_u E}[p^2(X)] = O(\varepsilon\log^2(1/\varepsilon))$. By Cauchy–Schwarz, this implies that

$$(|E|/|S'|)\mathop{\mathbb{E}}_{X\in_u E}[p(X)] \le \sqrt{(|E|/|S'|)}\sqrt{(|E|/|S'|)\mathop{\mathbb{E}}_{X\in_u E}[p^2(X)]} = O(\varepsilon\log(1/\varepsilon)).$$

Thus,

$$(|E|/|S'|)\|I - \Sigma^{-1/2}\Sigma_E\Sigma^{-1/2}\|_F = O(\varepsilon\log(1/\varepsilon)).$$

This shows that if Algorithm 8 returns in this step, it does so correctly. ∎

Next, we need to show that if Algorithm 8 reaches step 12, then such a $T$ exists.

CLAIM 5.25. *If Algorithm* 8 *reaches step* 12*, then there exists a* $T > 1$ *such that*

$$\Pr_{X\in_u S'}(|p(X) - \mu| \ge T) \ge 12\exp(-(T-1)/3) + 3\varepsilon/(d\log(N/\tau))^2.$$

*Proof.* Before we begin, we will need the following critical lemma.

LEMMA 5.26. *We present*

$$\mathop{\mathrm{Var}}_{X\sim G}[p(X)] \le 1 + O(\varepsilon\log(1/\varepsilon)).$$

*Proof.* We note that since $\mathrm{Var}_{X\sim G'}(p(G')) = Q_{G'}(p) = 1$, we just need to show that the variance with respect to $G$ instead of $G'$ is not too much larger. This will essentially be because the covariance matrix of $G$ cannot be much bigger than the covariance matrix of $G'$ by Corollary 5.22.

Using Lemma 5.16, we can write

$$p(x) = (\Sigma'^{-1/2}x)^T P_2(\Sigma'^{-1/2}x) + p_0 \ ,$$

where $\|P_2\|_F = \frac{1}{2}\mathrm{Var}_{X\sim G'}(p(G')) = \frac{1}{2}$ and $p_0 = \mu + \mathrm{tr}(P_2)$. We can also express $p(x)$ in terms of $G$ as $p(x) = (\Sigma^{-1/2}x)^T M(\Sigma^{-1/2}x) + p_0$, and have $\mathrm{Var}_{X\sim G}[p(X)] = \|M\|_F$. Here, $M$ is the matrix $\Sigma^{1/2}\Sigma'^{-1/2}P_2\Sigma'^{-1/2}\Sigma^{1/2}$. By Corollary 5.22, it holds that $\Sigma' \geq (1 - O(\varepsilon\log(1/\varepsilon)))\Sigma$. Consequently, $\Sigma^{1/2}\Sigma'^{-1/2} \leq (1 + O(\varepsilon\log(1/\varepsilon)))I$, and so $\|\Sigma^{1/2}\Sigma'^{-1/2}\|_2 \leq 1 + O(\varepsilon\log(1/\varepsilon))$. Similarly, $\|\Sigma'^{-1/2}\Sigma^{1/2}\|_2 \leq 1 + O(\varepsilon\log(1/\varepsilon))$.

We claim that if $A, B$ are matrices, then $\|AB\|_F \leq \|A\|_2\|B\|_F$. If $B_j$ are the columns of $B$, then we have $\|AB\|_F^2 = \sum_j \|AB_j\|_2^2 \leq \|A\|_2^2 \sum_j \|B_j\|_2^2 = (\|A\|_2\|B\|_F)^2$. Similarly for rows, we have $\|AB\|_F \leq \|A\|_F\|B\|_2$.

Thus, we have

$$\mathop{\mathrm{Var}}_{X\sim G}[p(X)] = 2\|M\|_F \leq 2\|\Sigma^{1/2}\Sigma'^{-1/2}\|_2\|P_2\|_F\|\Sigma'^{-1/2}\Sigma^{1/2}\|_2 \leq 1 + O(\varepsilon\log(1/\varepsilon)) \ .$$

Next, we need to consider $\mu$. In particular, we note that by the similarity of $S$ and $S'$, $\mu$ must be between the 40 and 60 percentiles of values of $p(X)$ for $X \in S$. However, since $S$ is $(\varepsilon, \tau)$-good, this must be between the 30 and 70 percentiles of $p(G)$. Therefore, by Cantelli's inequality,

$$|\mu - \hat{\mu}| \leq 2\sqrt{\mathop{\mathrm{Var}}_{X\sim G}[p(X)]} \leq 3 \ , \tag{27}$$

where $\hat{\mu} = \mathbb{E}_{X\sim G}[p(X)]$. We are now ready to proceed. Our argument will follow by noting that while $Q_{S'}(p)$ is much larger than expected, very little of this discrepancy can be due to points in $S \setminus L$. Therefore, the points of $E$ must provide a large contribution. Given that there are few points in $E$, much of this contribution must come from there being many points near the tails, and this will guarantee that some valid threshold $T$ exists.

In particular, we have that $\mathrm{Var}_{X\in_u S'}(p(X)) = Q_{S'}(p) \geq 1 + C\varepsilon\ln^2(1/\varepsilon)$, which means that

$$\frac{\sum_{X\in S'} |p(X) - \hat{\mu}|^2}{|S'|} \geq 1 + C\varepsilon\ln^2(1/\varepsilon).$$

Now, because $S$ is good, we know that

$$\begin{aligned}\frac{\sum_{X\in S} |p(X) - \hat{\mu}|^2}{|S|} &= \mathbb{E}[|p(G) - \hat{\mu}|^2](1 + O(\varepsilon)) \\ &= \mathop{\mathrm{Var}}_{X\sim G}[p(X)](1 + O(\varepsilon)) \leq 1 + O(\varepsilon\log(1/\varepsilon)).\end{aligned}$$

Therefore, using (27), we have that

$$\frac{\sum_{X\in S\setminus L} |p(X) - \hat{\mu}|^2}{|S'|} \leq 1 + O(\varepsilon\log(1/\varepsilon)).$$

Hence, for $C$ sufficiently large, it must be the case that

$$\sum_{X\in E} |p(X) - \hat{\mu}|^2 \geq (C/2)\varepsilon\ln^2(1/\varepsilon)|S'| \ ,$$

and therefore,

$$\sum_{X \in E} |p(X) - \mu|^2 \geq (C/3)\varepsilon \ln^2(1/\varepsilon)|S'| \,.$$

On the other hand, we have that

$$\begin{aligned}\sum_{X \in E} |p(X) - \mu|^2 &= \int_0^\infty \{X \in E : |p(X) - \mu| > t\} 2t dt \\ &\leq \int_0^{C^{1/4}\ln(1/\varepsilon)} O(t\varepsilon|S'|)dt + \int_{C^{1/4}\ln(1/\varepsilon)}^\infty \{X \in E : |p(X) - \mu| > t\} 2t dt \\ &\leq O(C^{1/2}\varepsilon \log^2(1/\varepsilon)|S'|) + |S'| \int_{C^{1/4}\ln(1/\varepsilon)}^\infty \Pr_{X \in_u S'}(|p(X) - \mu| > t) 2t dt \,.\end{aligned}$$

Therefore, we have that

$$\int_{C^{1/4}\ln(1/\varepsilon)}^\infty \Pr_{X \in_u S'}(|p(X) - \mu| > t) 2t dt \geq (C/4)\varepsilon \log^2(1/\varepsilon) \,. \tag{28}$$

Assume for the sake of contradiction that

$$\Pr_{X \in_u S'}(|p(X) - \mu| \geq T + 3) \leq \mathrm{Tail}(T, d, \varepsilon, \tau)$$

for all $T > 1$.

Thus, we have that

$$\begin{aligned}\int_{10\ln(1/\varepsilon)+3}^\infty \Pr_{X \in_u S'}(|p(X) - \mu| > T) 2T dT &\leq \int_{10\ln(1/\varepsilon)}^\infty 6(T+3)\varepsilon/(T^2 \log^2 T) dT \\ &= \int_{10\ln(1/\varepsilon)}^\infty 8\varepsilon/(T \log^2 T) dT \\ &= 8\varepsilon / \ln(10\ln(1/\varepsilon)) \,.\end{aligned}$$

For a sufficiently large $C$, this contradicts (28). ∎

Finally, we need to verify that if Algorithm 8 returns output in step 13, then it is correct.

CLAIM 5.27. *If Algorithm* 8 *returns during step* 13, *then* $\Delta(S, S'') \leq \Delta(S, S') - \varepsilon/(d \log(N/\tau))^2$.

*Proof.* We note that it is sufficient to show that $|E \setminus S''| > |(S \setminus L) \setminus S''|$. In particular, it suffices to show that

$$|\{X \in E : |p(X) - \mu| > T + 3\}| > |\{X \in S \setminus L : |p(X) - \mu| > T + 3\}| \,.$$

For this, it suffices to show that

$$|\{X \in S' : |p(X) - \mu| > T + 3\}| > 2|\{X \in S \setminus L : |p(X) - \mu| > T + 3\}| \,,$$

or that

$$|\{X \in S' : |p(X) - \mu| > T + 3\}| > 2|\{X \in S : |p(X) - \mu| > T + 3\}| \,.$$

By assumption, we have that

$$|\{X \in S' : |p(X) - \mu| > T + 3\}| > 3|S'|\varepsilon/(T^2 \log^2 T) \,.$$

On the other hand, using (27) and the $\varepsilon$-goodness of $S$, we have that

$$|\{X \in S : |p(X) - \mu| > T + 3\}| \leq |\{X \in S : |p(X) - \hat{\mu}| > T\}| \leq |S|\varepsilon/(T^2 \log^2 T) .$$

This completes our proof. ∎

**6. Agnostically learning a mixture of spherical Gaussians, via convex programming.** In this section, we give an algorithm to agnostically learn a mixture of $k$ Gaussians with identical spherical covariance matrices up to error $\widetilde{O}(\mathrm{poly}(k)\cdot\sqrt{\varepsilon})$. Let $\mathcal{M} = \sum_{j\in[k]} \alpha_j \mathcal{N}(\mu_j, \sigma^2 I)$ be the unknown $k$ Gaussian, each of whose components are spherical. Throughout this section we shall refer to such a distribution as a $k$-Gaussian mixture model (or $k$-GMM for short). For $X \sim \mathcal{M}$, we write $X \sim_j \mathcal{M}$ if $X$ was drawn from the $j$th component of $\mathcal{M}$.

Our main result of this section is the following theorem.

THEOREM 6.1. *Fix $\varepsilon, \tau > 0$, and $k \in \mathbb{N}$. Let $X_1, \ldots, X_N$ be an $\varepsilon$-corrupted set of samples from a $k$-GMM $\mathcal{M} = \sum_{j\in[k]} \alpha_j \mathcal{N}(\mu_j, \sigma_j^2 I)$, where all $\alpha_j, \mu_j$, and $\sigma_j^2$ are unknown, and*

$$N = \widetilde{\Omega}\left(\mathrm{poly}\left(d, k, 1/\varepsilon, \log(1/\tau)\right)\right) .$$

*There is an algorithm which, with probability $1 - \tau$, outputs a distribution $\mathcal{M}'$ such that*

$$d_{\mathrm{TV}}(\mathcal{M}, \mathcal{M}') \leq \widetilde{O}(\mathrm{poly}(k) \cdot \sqrt{\varepsilon}) .$$

*The running time of the algorithm is* $\mathrm{poly}(d, 1/\varepsilon, \log(1/\tau))^{k^2}$.

Our overall approach will be a combination of our method for agnostically learning a single Gaussian and recent work on properly learning mixtures of multivariate spherical Gaussians [SOAJ14, LS17]. At a high level, this recent work relies upon the empirical covariance matrix giving an accurate estimate of the overall covariance matrix in order to locate the subspace in which the component means lie. However, as we have already observed, the empirical moments do not necessarily give good approximations of the true moments in the agnostic setting. Therefore, we will use our separation oracle framework to approximate the covariance matrix, and the rest of the arguments follow similarly to those of previous methods.

The organization of this section will be as follows. We define some of the notation we will be using and the Schatten top-$k$ norm in section 6.1. Section 6.2 states the various concentration inequalities we require. In section 6.3, we go over our overall algorithm in more detail. Section 6.4 describes a first naive clustering step, which deals with components which are very well separated. Section 6.5 contains details on our separation oracle approach, allowing us to approximate the true covariance. Section 6.6 describes our spectral clustering approach to cluster components with means separated more than $\Omega_k(\log 1/\varepsilon)$. In section 6.7, we describe how to exhaustively search over a particular subspace to obtain a good estimate for the component means. In section 6.8, we go over how to limit the set of hypotheses in order to satisfy the conditions of Lemma 2.23. For clarity of exposition, all of the above describe the algorithm assuming all $\sigma_j^2$ are equal. In section 6.9, we discuss the changes to the algorithm which are required to handle unequal variances.

For conciseness, many of the proofs are deferred to Appendix C.

**6.1. Notation and norms.** Recall the definition of $S_{N,\varepsilon}$ from section 4.1, which we will use extensively in this section. We will use the notation $\mu = \sum_{j\in[k]} \alpha_j \mu_j$ to denote the mean of the unknown GMM. Also, we define parameters $\gamma_j = \alpha_j \|\mu_j - \mu\|_2^2$ and let $\gamma = \max_j \gamma_j$. And for ease of notation, let

$$f(k,\gamma,\varepsilon) = k^{1/2}\varepsilon + k\gamma^{1/2}\varepsilon + k\varepsilon^2$$

and

$$h(k,\gamma,\varepsilon) = k^{1/2}\varepsilon + k\gamma^{1/2}\varepsilon + k\gamma\varepsilon + k\varepsilon^2 = f(k,\gamma,\varepsilon) + k\gamma\varepsilon.$$

Finally, we use the notation

$$Q = \sum_{j\in[k]} \alpha_j(\mu_j - \mu)(\mu_j - \mu)^T \tag{29}$$

to denote the covariance of the unknown GMM. Our algorithm for learning spherical $k$-GMMs will rely heavily on the following, nonstandard norm.

Definition 6.2. *For any symmetric matrix $M \in \mathbb{R}^{d\times d}$ with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$, let the* Schatten top-$k$ norm *be defined as*

$$\|M\|_{T_k} = \sum_{i=1}^{k} \sigma_i \ ;$$

*i.e., it is the sum of the top-$k$ singular values of $M$.*

It is easily verified that $\|\cdot\|_{T_k}$ has a dual characterization

$$\|M\|_{T_k} = \max_{X\in\mathbb{R}^{d\times k}} \mathrm{Tr}(X^T \sqrt{M^T M} X) \ ,$$

where the maxima is taken over all $X$ with orthonormal columns. From this, it is easy to see that the Schatten top-$k$ norm is indeed a norm, as its name suggests.

Fact 6.3. *$\|M\|_{T_k}$ is a norm on symmetric matrices.*

**6.2. Concentration inequalities.** In this section, we will establish some concentration inequalities that we will need for our algorithm for agnostically learning mixtures of spherical Gaussians. Recall the notation as described in section 6.1. The following two concentration lemmas follow from the same proofs as for Lemmas 42 and 44 in [LS17].

Lemma 6.4. *Fix $\varepsilon, \delta > 0$. If $Y_1, \ldots, Y_N$ are independent samples from the GMM with PDF $\sum_{j\in[k]} \alpha_j \mathcal{N}(\mu_j, \Sigma_j)$ where $\alpha_j \geq \Omega(\varepsilon)$ for all $j$, and $N = \Omega\left(\frac{d+\log(k/\delta)}{\varepsilon^2}\right)$, then with probability at least $1 - O(\delta)$,*

$$\left\| \frac{1}{N}\sum_{i=1}^{N} (Y_i - \mu)(Y_i - \mu)^T - I - Q \right\|_2 \leq O\left(f(k,\gamma,\varepsilon)\right),$$

*where $Q$ is defined as in (29).*

Lemma 6.5. *Fix $\varepsilon, \delta > 0$. If $Y_1, \ldots, Y_N$ are independent samples from the GMM with PDF $\sum_{j\in[k]} \alpha_j \mathcal{N}(\mu_j, \Sigma_j)$ where $\alpha_j \geq \Omega(\varepsilon)$ for all $j$, and $N = \Omega\left(\frac{d+\log(k/\delta)}{\varepsilon^2}\right)$, then with probability at least $1 - O(\delta)$,*

$$\left\| \frac{1}{N}\sum_{i=1}^{N} Y_i - \mu \right\|_2 \leq O\left(k^{1/2}\varepsilon\right).$$

From the same techniques as before, we get the same sort of union bounds as usual over the weight vectors.

LEMMA 6.6. *Fix $\varepsilon \leq 1/2$ and $\tau \leq 1$. There is a $\delta = O(\varepsilon\sqrt{\log 1/\varepsilon})$ such that if $Y_1, \ldots, Y_N$ are independent samples from the GMM with PDF $\sum_{j\in[k]} \alpha_j \mathcal{N}(\mu_j, \Sigma_j)$ where $\alpha_j \geq \Omega(\varepsilon)$ for all $j$, and $N = \Omega\left(\frac{d+\log(k/\tau)}{\delta_1^2}\right)$, then*

$$\Pr\left[\exists w \in S_{N,\varepsilon} : \left\| \sum_{i=1}^N w_i (Y_i - \mu)(Y_i - \mu)^T - I - Q \right\|_2 \geq f(k, \gamma, \delta_1)\right] \leq \tau \,, \tag{30}$$

*where $Q$ is defined as in* (29).

LEMMA 6.7. *Fix $\varepsilon \leq 1/2$ and $\tau \leq 1$. There is a $\delta = O(\varepsilon\sqrt{\log 1/\varepsilon})$ such that if $Y_1, \ldots, Y_N$ are independent samples from the GMM with PDF $\sum_{j\in[k]} \alpha_j \mathcal{N}(\mu_j, \Sigma_j)$ where $\alpha_j \geq \Omega(\varepsilon)$ for all $j$, and $N = \Omega\left(\frac{d+\log(k/\tau)}{\delta_2^2}\right)$, then*

$$\Pr\left[\exists w \in S_{N,\varepsilon} : \left\| \sum_{i=1}^N w_i Y_i - \mu \right\|_2 \geq k^{1/2}\delta_2\right] \leq \tau. \tag{31}$$

**6.3. Algorithm.** Our approach is based on a *tournament*, as used in several recent works [DK14, SOAJ14, DDS15a, DDS15b, DKT15, DDKT16]. We will generate a list $\mathcal{S}$ of candidate hypotheses (i.e., of $k$-GMMs) of size $|\mathcal{S}| = \mathrm{poly}(d, 1/\varepsilon, \log(1/\tau))^{k^2}$ with the guarantee that there is some $\mathcal{M}^* \in \mathcal{S}$ such that $d_{\mathrm{TV}}(\mathcal{M}, \mathcal{M}^*) \leq \tilde{O}(\mathrm{poly}(k) \cdot \sqrt{\varepsilon})$. We then find (roughly) the best candidate hypothesis on the list. It is most natural to describe the algorithm as performing several layers of *guessing*. We will focus our discussion on the main steps in our analysis, and defer a discussion of guessing the mixing weights, the variance $\sigma^2$, and performing naive clustering until later. For reasons we justify in section 6.8, we may assume that the mixing weights and the variance are known exactly, and that the variance $\sigma^2 = 1$.

Our algorithm is based on the following deterministic conditions:

$$\frac{|\{X_i \in G, X_i \sim_j \mathcal{M} : \|X_i - \mu_j\|_2^2 \geq \Omega(d \log k/\varepsilon)\}|}{|\{X_i \in G, X_i \sim_j \mathcal{M}\}|} \leq \varepsilon/k \;\forall j = 1, \ldots, N \,, \tag{32}$$

$$\left\| \sum_{i\in G} w_i (X_i - \mu)(X_i - \mu)^T - w_g I - w_g Q \right\|_2 \leq f(k, \gamma, \delta_1) \;\forall w \in S_{N,4\varepsilon}, \text{ and} \tag{33}$$

$$\left\| \sum_{i\in G} w_i (X_i - \mu) \right\|_2 \leq k^{1/2}\delta_2 \;\forall w \in S_{N,4\varepsilon} \;. \tag{34}$$

Equation (32) follows from basic Gaussian concentration, and (33) and (34) follow from the results in section 6.2 for $N$ sufficiently large. Note that these trivially imply similar conditions for the Schatten top-$k$ norm, at the cost of an additional factor of $k$ on the RHS of the inequalities. For the rest of this section, let $\delta = \max(\delta_1, \delta_2)$.

At this point, we are ready to apply our separation oracle framework. In particular, we will find a weight vector $w$ over the points such that

$$\left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I - \sum_{j\in[k]} \alpha_j (\mu_j - \mu)(\mu_j - \mu)^T \right\|_2 \leq \eta,$$

for some choice of $\eta$. The set of such weights is convex, and concentration implies that the true weight vector will have this property. Furthermore, we can describe a separation oracle given any weight vector not contained in this set (as long as $\eta$ is not too small). At this point, we use classical convex programming methods to find a vector which satisfies these conditions. Further details are provided in section 6.5.

After this procedure, Lemma 6.13 shows that the weighted empirical covariance is spectrally close to the true covariance matrix. We are now in the same regime as [SOAJ14], which obtains their results as a consequence of the empirical covariance concentrating on the true covariance matrix. Thus, we will appeal to their analysis, highlighting the differences between our approach and theirs. We note that [LS17] also follows a similar approach and the interested reader may also adapt their arguments instead.

First, if $\gamma$ is sufficiently large (i.e., $\Omega_k(\log(1/\varepsilon))$), this implies a separation condition between some component mean and the mixture's mean. This allows us to cluster the points further, using a spectral method. We take the top eigenvector of the weighted empirical covariance matrix and project in this direction, using the sign of the result as a classifier. In contrast to previous work, which requires that no points are misclassified, we can tolerate $\mathrm{poly}(\varepsilon/k)$ misclassifications, since our algorithms are agnostic. This crucially allows us to avoid a dependence on $d$ in our overall agnostic learning guarantee. Further details are provided in section 6.6.

Finally, if $\gamma$ is sufficiently small, we may perform an exhaustive search. The span of the means is in the span of the top $k-1$ eigenvectors of the true covariance matrix, which we can approximate with our weighted empirical covariance matrix. Since $\gamma$ is small, by trying all points within a sufficiently tight mesh, we can guess a set of candidate means which are sufficiently close to the true means. Combining the approximations to the means with Corollary 2.13 and the triangle inequality, we can guarantee that at least one of our guesses is sufficiently close to the true distribution. Additional details are provided in section 6.7.

To conclude our algorithm, we can apply Lemma 2.23. We note that this hypothesis selection problem has been studied before (see, e.g., [DL01, DK14]), but we must adapt it for our agnostic setting. This allows us to select a hypothesis which is sufficiently close to the true distribution, thus concluding the proof. We note that the statement of Lemma 2.23 requires the hypotheses to come from some fixed finite set, while there are an infinite number of GMM. In section 6.8, we discuss how to limit the number of hypotheses based on the set of uncorrupted samples in order to satisfy the conditions of Lemma 2.23.

**6.4. Naive clustering.** We give a very naive clustering algorithm, the generalization of NaivePrune, which recursively allows us to cluster components if they are extremely far away. The algorithm is very simple: for each $X_i$, add all points within distance $O(d\log(k/\varepsilon))$ to a cluster $S_i$. Let $\mathcal{C}$ be the set of clusters which contain at least $4\varepsilon N$ points, and let the final clustering be $C_1,\dots,C_{k'}$ be formed by merging clusters in $\mathcal{C}$ if they overlap, and stopping if no clusters overlap. We give the pseudocode in Algorithm 10.

We prove here that this process (which may throw away points) throws away only at most an $\varepsilon$-fraction of good points, and moreover, the resulting clustering only misclassifies at most an $O(\varepsilon)$-fraction of the good points, assuming (32).

Theorem 6.8. *Let $X_1,\dots,X_m$ be a set of samples satisfying* (32)*. Let $C_1,\dots,C_{k'}$ be the set of clusters returned. For each component $j$, let $\ell(j)$ be the $\ell$ such that $C_\ell$ contains the most points from $j$. Then the following hold:*

**Algorithm 10** Naive clustering algorithm for spherical GMMs.

```
function NaiveClusterGMM(X_1, ..., X_n)
    for i = 1, ..., N do
        Let S_i = {i' : ||X_i − X_{i'}||_2^2 ≤ Θ(dk log 1/ε)}.
    Let C = {S_i : |S_i| ≥ 4εN}.
    while ∃C, C' ∈ C such that C ≠ C' and C ∪ C' ≠ ∅ do
        Remove C, C' from C.
        Add C ∪ C' to C.
    return the set of clusters C.
```

1. *Then, for each* $\ell$*, there is some* $j$ *such that* $\ell(j) = \ell$*.*
2. *For all* $j$*, we have*

$$|\{X_i \in G, X_i \sim_j \mathcal{M}\}| - |\{X_i \in G, X_i \sim_j \mathcal{M}, X_i \in C_{\ell(j)}\}| \leq O\left(\frac{\varepsilon}{k}|\{X_i \in G, X_i \sim_j \mathcal{M}\}|\right) .$$

3. *For all* $j, j'$*, we have that if* $\ell(j) = \ell(j')$*, then* $\|\mu_j - \mu_{j'}\|_2^2 \leq O(dk \log k/\varepsilon)$*.*
4. *If* $X_i, X_j \in C_\ell$*, then* $\|X_i - X_j\|_2^2 \leq dk \log 1/\varepsilon$*.*

Thus, we have that by applying this algorithm, given an $\varepsilon$-corrupted set of samples from $\mathcal{M}$, we may cluster them in a way which misclassifies at most an $\varepsilon/k$-fraction of the samples from any component, and such that within each cluster, the means of the associated components differ by at most $dk \log k/\varepsilon$. Thus, each separate cluster is simply an $\varepsilon$-corrupted set of samples from the mixture restricted to the components within that cluster; moreover, the number of components in each cluster must be strictly smaller than $k$. Therefore, we may simply recursively apply our algorithm on these clusters to agnostically learn the mixture for each cluster, since if $k = 1$, this is a single Gaussian, which we know how to learn agnostically.

Thus, for the remainder of this section, let us assume that for all $j, j'$, we have $\|\mu_j - \mu_j'\|_2^2 \leq O(dk \log 1/\varepsilon)$. Moreover, we may assume that there are no points $j, j'$ (corrupted or uncorrupted), such that $\|X_j - X_{j'}\|_2^2 \geq \Omega(dk \log 1/\varepsilon)$.

**6.5. Estimating the covariance using convex programming.** In this section, we will apply our separation oracle framework to estimate the covariance matrix. While in the nonagnostic case, the empirical covariance will approximate the actual covariance, this is not necessarily true in our case. As such, we will focus on determining a weight vector over the samples such that the weighted empirical covariance *is* a good estimate for the true covariance.

We first define the convex set for which we want an interior point:

$$\mathcal{C}_\eta = \left\{ w \in S_{N,\varepsilon} : \left\| \sum_{i=1}^{N} w_i (X_i - \mu)(X_i - \mu)^T - I - \sum_{j \in [k]} \alpha_j (\mu_j - \mu)(\mu_j - \mu)^T \right\|_2 \leq \eta \right\} .$$

In section 6.5.1, we prove lemmas indicating important properties of this set. In section 6.5.2, we give a separation oracle for this convex set. We conclude with Lemma 6.13, which shows that we have obtained an accurate estimate of the true covariance.

**6.5.1. Properties of our convex set.** We start by proving the following lemma, which states that for any weight vector which is not in our set, the weighted empirical covariance matrix is noticeably larger than it should be (in Schatten top-$k$ norm).

LEMMA 6.9. *Suppose that* (33) *holds, and* $w \notin \mathcal{C}_{ckh(k,\gamma,\delta)}$. *Then*

$$\left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_{T_k} \geq \sum_{j \in [k]} \gamma_j + \frac{3ckh(k,\gamma,\delta)}{4}.$$

We also require the following lemma, which shows that if a set of weights poorly approximates $\mu$, then it is not in our convex set.

LEMMA 6.10. *Suppose that* (33) *and* (34) *hold. Let* $w \in S_{m,\varepsilon}$ *and set* $\widehat{\mu} = \sum_{i=1}^m w_i X_i$ *and* $\Delta = \mu - \widehat{\mu}$. *Furthermore, suppose that* $\|\Delta\|_2 \geq \Omega(h(k,\gamma,\delta))$. *Then*

$$\left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I - \sum_{j \in [k]} \alpha_j (\mu_j - \mu)(\mu_j - \mu)^T \right\|_2 \geq \Omega\left( \frac{\|\Delta\|_2^2}{\varepsilon} \right).$$

By contraposition, if a set of weights is in our set, then it provides a good approximation for $\mu$.

COROLLARY 6.11. *Suppose that* (33) *and* (34) *hold. Let* $w \in \mathcal{C}_{h(k,\gamma,\delta)}$ *for* $\delta = \Omega(\varepsilon \log 1/\varepsilon)$. *Then*

$$\|\Delta\|_2 \leq O(\varepsilon \sqrt{\log 1/\varepsilon}).$$

**6.5.2. Separation oracle.** In this section, we provide a separation oracle for $\mathcal{C}_\eta$. In particular, we have the following theorem.

THEOREM 6.12. *Fix* $\varepsilon > 0$, *and let* $\delta = \Omega(\varepsilon \log 1/\varepsilon)$. *Suppose that* (33) *and* (34) *hold. Let* $w^*$ *denote the weights which are uniform on the uncorrupted points. Then there is a constant* $c$ *and an algorithm such that the following hold:*

1. *(Completeness) If* $w = w^*$, *then it outputs "YES".*
2. *(Soundness) If* $w \notin \mathcal{C}_{ckh(k,\gamma,\delta)}$, *the algorithm outputs a hyperplane* $\ell : \mathbb{R}^m \to \mathbb{R}$ *such that* $\ell(w) \geq 0$ *but* $\ell(w^*) < 0$.

*These two facts imply that the ellipsoid method with this separation oracle will terminate in* $\mathrm{poly}(d, 1/\varepsilon)$ *steps, and moreover, will with high probability output a* $w'$ *such that* $\|w - w'\|_\infty \leq \varepsilon/(Ndk \log 1/\varepsilon)$ *for some* $w \in \mathcal{C}_{ckh(k,\gamma,\delta)}$. *Moreover, it will do so in polynomially many iterations.*

The proof is deferred to Appendix C.1.

**Algorithm 11** Separation oracle subprocedure for agnostically learning the span of the means of a GMM.

**function** SEPARATIONORACLEGMM($w$)
  Let $\widehat{\mu} = \sum_{i=1}^N w_i X_i$.
  For $i = 1, \ldots, N$, define $Y_i = X_i - \widehat{\mu}$.
  Let $M = \sum_{i=1}^N w_i Y_i Y_i^T - I$.
  **if** $\|M\|_{T_k} < \sum_{j \in [k]} \gamma_j + \frac{ckh(k,\gamma,\delta)}{2}$ **then**
    **return** "YES".
  **else**
    Let $\Lambda = \|M\|_{T_k}$.
    Let $U$ be a $d \times k$ matrix with orthonormal columns which span the top $k$ eigenvectors of $M$.
    **return** the hyperplane $\ell(w) = \mathrm{Tr}\left(U^T \left(\sum_{i=1}^N w_i Y_i Y_i^T - I\right) U\right) - \Lambda > 0$.

After running this procedure, we technically do not have a set of weights in $\mathcal{C}_{ckh(k,\gamma,\delta)}$. But by the same argument as in section 4.3, because the maximum distance between two points within any cluster is bounded, and we have the guarantee that $\|X_i - X_j\|^2 \le O(dk \log 1/\varepsilon)$ for all $i, j$, we may assume we have a set of weights satisfying

$$\left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I - \sum_{j \in [k]} \alpha_j (\mu_j - \mu)(\mu_j - \mu)^T \right\|_2 \le 2ckh(k, \gamma, \delta).$$

We require the following lemma, describing the accuracy of the empirical covariance matrix with the obtained weights.

LEMMA 6.13. *Let $\widehat{\mu} = \sum_{i=1}^N w_i X_i$. After running the algorithm above, we have a vector $w$ such that*

$$\left\| \sum_{i=1}^N w_i (X_i - \widehat{\mu})(X_i - \widehat{\mu})^T - I - \sum_{j \in [k]} \alpha_j (\mu_j - \mu)(\mu_j - \mu)^T \right\|_2 \le 3ckh(k, \gamma, \delta).$$

*Proof.* By the triangle inequality and Corollary 6.11,

$$\begin{aligned} &\left\| \sum_{i=1}^N w_i (X_i - \widehat{\mu})(X_i - \widehat{\mu})^T - I - \sum_{j \in [k]} \alpha_j (\mu_j - \mu)(\mu_j - \mu)^T \right\|_2 \\ &\le \left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I - \sum_{j \in [k]} \alpha_j (\mu_j - \mu)(\mu_j - \mu)^T \right\|_2 + \|\Delta\|_2^2 \\ &\le 2ckh(k, \gamma, \delta) + O(\delta) \le 3ckh(k, \gamma, \delta). \end{aligned}$$ ∎

**6.6. Spectral clustering.** Now that we have a good estimate of the true covariance matrix, we will perform spectral clustering while $\gamma$ is sufficiently large. We will adapt Lemma 6 from [SOAJ14], giving the following lemma.

LEMMA 6.14. *Given a weight vector $w$ as output by Algorithm* 11*, if $\gamma \ge \Omega(\text{poly}(k) \cdot \log 1/\varepsilon)$, there exists an algorithm which produces a unit vector $v$ with the following guarantees:*

- *There exists a nontrival partition of $[k]$ into $S_0$ and $S_1$ such that $v^T \mu_j > 0$ for all $j \in S_0$ and $v^T \mu_j < 0$ for all $j \in S_1$.*
- *The probability of a sample being misclassified is at most $O(\text{poly}(\varepsilon/k))$, where a misclassification is defined as a sample $X$ generated from a component in $S_0$ having $v^T X < 0$, or a sample generated from a component in $S_1$ having $v^T X > 0$.*

The algorithm will be as follows. Let $v$ be the top eigenvector of

$$\sum_{i=1}^N w_i (X_i - \widehat{\mu})(X_i - \widehat{\mu})^T - I.$$

For a sample $X$, cluster it based on the sign of $v^T X$. After performing this clustering, recursively perform our algorithm from the start on the two clusters.

The proof is very similar to that of Lemma 6 in [SOAJ14]. The authors' main concentration lemma is Lemma 30 in [SOAJ14], which states that they obtain a good estimate of the true covariance matrix, akin to our Lemma 6.13. Lemma 31 in [SOAJ14] argues that the largest eigenvector of this estimate is highly correlated with the top eigenvector of the true covariance matrix. Since $\gamma$ is large, this implies there is a large margin between the mean and the hyperplane. However, by standard Gaussian tail bounds, the probability of a sample landing on the opposite side of this hyperplane is small.

We highlight the main difference between our approach and theirs. For their clustering step, they require that no sample is misclustered with high probability. As such, they may perform spectral clustering while $\gamma = \Omega\left(\mathrm{poly}(k) \cdot \log(d/\varepsilon)\right)$. We note that, in the next step of our algorithm, we will perform an exhaustive search. This will result in an approximation which depends on the value of $\gamma$ at the start of the step, and as such, using the same approach as theirs would result in an overall approximation which depends logarithmically on the dimension.

We may avoid paying this cost by noting that our algorithm is agnostic. They require that no sample is misclustered with high probability, while our algorithm tolerates that a $\mathrm{poly}(\varepsilon/k)$-fraction of points are misclustered. As such, we can continue spectral clustering until $\gamma = O\left(\mathrm{poly}(k) \cdot \log(1/\varepsilon)\right)$.

**6.7. Exhaustive search.** The final stage of the algorithm is when we know that all $\gamma_i$'s are sufficiently small. We can directly apply the following lemma.

LEMMA 6.15 (Lemma 7 of [SOAJ14]). *Given a weight vector $w$ as output by Algorithm 11, then the projection of $\frac{\mu_j - \mu}{\|\mu_j - \mu\|_2}$ onto the space orthogonal to the span of the top $k-1$ eigenvectors of*

$$\left\| \sum_{i=1}^{N} w_i (X_i - \widehat{\mu})(X_i - \widehat{\mu})^T - I \right\|_2$$

*has magnitude at most*

$$O\left(\mathrm{poly}(k) \cdot \sqrt{h(k, \gamma, \delta)}/\gamma_i^{1/2}\right) = O\left(\mathrm{poly}(k) \cdot \frac{\sqrt{\varepsilon}\log(1/\varepsilon)}{\gamma_i^{1/2}}\right).$$

At this point, our algorithm is identical to the exhaustive search of [SOAJ14]. We find the span of the top $k-1$ eigenvectors by considering the $(k-1)$-cube with side length $2\gamma$ centered at $\widehat{\mu}$. By taking an $\eta$-mesh over the points in this cube (for $\eta = \mathrm{poly}(\varepsilon/dk)$ sufficiently small), we obtain a set of points $\tilde{M}$. Via identical arguments as in the proof of Theorem 8 of [SOAJ14], for each $j \in [k]$ there exists some point $\tilde{\mu}_j \in \tilde{M}$ such that

$$\|\tilde{\mu}_j - \mu_j\|_2 \le O\left(\mathrm{poly}(k) \cdot \frac{\sqrt{\varepsilon}\log(1/\varepsilon)}{\sqrt{\alpha_j}}\right).$$

By taking a $k$-wise Cartesian product of this set, we are guaranteed to obtain a vector which has this guarantee simultaneously for all $\mu_j$.

**6.8. Applying the tournament lemma.** In this section, we discuss details about how to apply our hypothesis selection algorithm. First, in section 6.8.1, we describe how to guess the mixing weights and the variance of the components. Then in section 6.8.2, we discuss how to ensure our hypotheses come from some fixed finite set, in order to deal with technicalities which arise when performing hypothesis selection with our adversary model.

**6.8.1. Guessing the mixing weights and variance.** The majority of our algorithm is focused on generating guesses for the means of the Gaussians. In this section, we guess the remaining parameters: the mixing weights and the variance. While most of these guessing arguments are standard, we emphasize that we reap an additional benefit because our algorithm is agnostic. In particular, most algorithms must deal with error incurred due to misspecification of the parameters. Since our algorithm is agnostic, we can pretend the misspecified parameter is the true one, at the cost of increasing the value of the agnostic parameter $\varepsilon$. If our misspecified parameters are accurate enough, the agnostic learning guarantee remains unchanged.

Guessing the mixing weights is fairly straightforward. For some $\nu = \text{poly}(\varepsilon/k)$ sufficiently small, our algorithm generates a set of at most $(1/\nu)^k = \text{poly}(k/\varepsilon)^k$ possible mixing weights by guessing the values $\{0, \varepsilon, \varepsilon+\nu, \varepsilon+2\nu, \ldots, 1-\nu, 1\}$ for each $\alpha_j$. Note that we may assume each weight is at least $\varepsilon$, since components with weights less than this can be specified arbitrarily at a total cost of $O(k\varepsilon)$ in total variation distance.

Next, we need to guess the variance $\sigma^2$ of the components. To accomplish this, we will take $k+1$ samples (hoping to find only uncorrupted ones) and compute the minimum distance between any pair of them. Since we assume $k \ll 1/\varepsilon$, we can repeatedly draw $k+1$ samples until we have the guarantee that at least one set is uncorrupted. If none of the $k+1$ samples are corrupted, then at least two of them came from the same component, and in our high-dimensional setting the distance between any pair of samples from the same component concentrates around $\sqrt{2d}\sigma$. After rescaling this distance, we can then multiplicatively enumerate around this value with granularity $\text{poly}(\varepsilon/dk)$ to get an estimate for $\sigma^2$ that is sufficiently good for our purposes. Applying Corollary 2.14 bounds the cost of this misspecification by $O(\varepsilon)$. By rescaling the points, we may assume that $\sigma^2 = 1$.

**6.8.2. Pruning our hypotheses.** In this section, we describe how to prune our set of hypotheses in order to apply Lemma 2.23. Recall that this lemma requires our hypotheses to come from some fixed finite set, rather than the potentially infinite set of GMM hypotheses. We describe how to prune and discretize the set of hypotheses obtained during the rest of the algorithm to satisfy this condition. For the purposes of this section, a hypothesis will be a $k$-tuple of $d$-dimensional points, corresponding only to the means of the components. While the candidate mixing weights already come from a fixed finite set (so no further work is needed), the unknown variance must be handled similarly to the means. The details for handling the variance are similar to (and simpler than) those for handling the means, and are omitted.

More precisely, this section will describe a procedure to generate a set of hypotheses $\mathcal{M}$, which is exponentially large in $k$ and $d$, efficiently searchable, and comes from a finite set of hypotheses which are fixed with respect to the true distribution. Then, given our set of hypotheses generated by the main algorithm (which is exponentially large in $k$ but polynomial in $d$), we iterate over this set, either replacing each hypothesis with a "close" hypothesis from $\mathcal{M}$ (i.e., one which is within $O(\varepsilon)$ total variation distance), or discarding the hypothesis if none exists. Finally, we run the tournament procedure of Lemma 2.23 on the resulting set of hypotheses.

At a high level, the approach will be as follows. We will take a small set of samples, and remove any samples from this set which are clear outliers (due to having too few nearby neighbors). This will give us a set of points, each of which is within a reasonable distance from some component mean. Taking a union of balls around these samples will give us a space that is a subset of a union of (larger) balls centered at the component centers. We take a discrete mesh over this space to obtain a fixed finite set of possible means, and round each hypothesis such that its means are within this set.

We start by taking $N = O(k \log(1/\tau)/\varepsilon^2)$ samples, which is sufficient to ensure that the number of (uncorrupted) samples from component $j$ will be $(w_j \pm \Theta(\varepsilon))N$ for all $j \in [k]$ with probability $1 - O(\tau)$. Recall that we are assuming that $w_j = \Omega(\varepsilon)$ for all $j$, as all other components may be defined arbitrarily at the cost of $O(k\varepsilon)$ in total variation distance. This implies that even after corruption, each component has generated at least $\varepsilon N$ uncorrupted samples.

By standard Gaussian concentration bounds, we know that if $N$ samples are taken from a Gaussian, the maximum distance between a sample and the Gaussian's mean will be at most $\zeta = O(\sqrt{d \log(N/\tau)})$ with probability $1 - \tau$. Assume this condition holds, and thus each component's mean will have at least $\varepsilon N$ points within distance $\zeta$. We prune our set of samples by removing any point with fewer than $\varepsilon N$ other points at distance less than $2\zeta$. This will not remove any uncorrupted points by the above assumption and triangle inequality. However, this will remove any corrupted points at distance at least $3\zeta$ from all component means, due to the fact that the adversary may only move an $\varepsilon$-fraction of the points, and reverse triangle inequality.

Now, we consider the union of the balls of radius $3\zeta$ centered at each of the remaining points. This set contains all of the component means, and is also a subset of the union of the balls of radius $6\zeta$ centered at the component means. We discretize this set by taking its intersection with a lattice of side-length $\frac{\varepsilon}{k\sqrt{d}}$. We note that any two points in this discretization are at distance at most $\varepsilon/k$. By a volume argument, the number of points in the intersection is at most $k\left(\frac{12\zeta k\sqrt{d}}{\varepsilon}\right)^d$. Each hypothesis will be described by the $k$-wise Cartesian product of these points, giving us a set $\mathcal{M}$ of at most $k^k\left(\frac{12\zeta k\sqrt{d}}{\varepsilon}\right)^{kd}$ hypotheses.

Given a set of hypotheses $\mathcal{H}$ from the main algorithm, we prune it using $\mathcal{M}$ as a reference. For each $h \in \mathcal{H}$, we see if there exists some $h' \in \mathcal{M}$ such that the means in $h$ are at distance at most $\varepsilon/k$ from the corresponding means in $h'$.[4] If such an $h'$ exists, we replace $h$ with $h'$; otherwise, $h$ is simply removed. By Corollary 2.13 and the triangle inequality, this replacement incurs a cost of $O(\varepsilon)$ in total variation distance. At this point, the conditions of Lemma 2.23 are satisfied and we may run this procedure to select a sufficiently accurate hypothesis.

**6.9. Handling unequal variances.** In this section, we describe the changes required to allow the algorithm to handle different variances for the Gaussians. The main idea is to find the minimum variance of any component and perform clustering so we only have uncorrupted samples from Gaussians with variances within some known, polynomially-wide interval. This allows us to grid within this interval in order to guess the variances, and the rest of the algorithm proceeds with minor changes.

The first step is to locate the minimum variance of any component. Again using standard Gaussian concentration, in sufficiently high-dimensions, if $N$ samples are taken from a Gaussian with variance $\sigma^2 I$, the distance between any two samples will be concentrated around $\sigma(\sqrt{2d} - \Theta(d^{1/4}))$. With this in hand, we use the following procedure to estimate the minimum variance. For each sample $i$, record the distance to the $(\varepsilon N + 1)$st closest sample. We take the $(\varepsilon N + 1)$st smallest of *these* values,

[4]We observe that the complexity of this step is polynomial in $d$ and $k$, not exponential, if one searches for the nearest lattice point in the sphere surrounding each unpruned sample, rather than performing a naive linear scan over the entire list.

rescale it by $1/\sqrt{2d}$, and similar to before, guess around it using a multiplicative $(1 + \mathrm{poly}(\varepsilon/kd))$ grid, which will give us an estimate $\hat{\sigma}^2_{min}$ for the smallest variance. We note that discarding the smallest $\varepsilon N$-fraction of the points prevents this statistic from being grossly corrupted by the adversary. For the remainder of this section, assume that $\sigma^2_{min}$ is known exactly.

At this point, we partition the points into those that come from components with small variance, and those with large variance. We will rely upon the following concentration inequality from [SOAJ14], which gives us the distance between samples from different components.

LEMMA 6.16 (Lemma 34 from [SOAJ14]). *Given $N$ samples from a collection of Gaussian distributions, with probability $1 - O(\tau)$, the following holds for every pair of samples $X, Y$:*

$$\|X - Y\|_2^2 \in \left(d(\sigma_1^2 + \sigma_2^2) + \|\mu_1 - \mu_2\|_2^2\right)\left(1 \pm 4\sqrt{\frac{\log \frac{N^2}{\tau}}{d}}\right),$$

*where $X \sim \mathcal{N}(\mu_1, \sigma_1^2 I)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2 I)$.*

Assume the event that this condition holds. Now, let $H_\ell$ be the set of all points with at least $\varepsilon N$ points at squared-distance at most $2\left(1 + \frac{1}{k}\right)^{\ell-1} \sigma^2_{\min}\left(1 + 4\sqrt{\frac{\log \frac{N^2}{\tau}}{d}}\right)$, for $\ell \in [k]$. Note that $H_\ell \subseteq H_{\ell+1}$. Let $\ell^*$ be the minimum $\ell$ such that $H_\ell = H_{\ell+1}$, or $k$ if no such $\ell$ exists, and partition the set of samples into $H_{\ell^*}$ and $\overline{H}_{\ell^*}$. This partition will contain all samples from components with variance at most some threshold $t$, where $t \leq e\sigma^2_{\min}$ in $H_{\ell^*}$. All samples from components with variance at least $t$ will fall into $\overline{H}_{\ell^*}$. We continue running the algorithm with $H_{\ell^*}$, and begin the algorithm recursively on $\overline{H}_{\ell^*}$.[5]

This procedure works due to the following argument. When we compute $H_1$, we are guaranteed that it will contain all samples from components with variance $\sigma^2_{\min}$, by the upper bound in Lemma 6.16. However, it may also contain samples from other components—in particular, those with variance at most $\gamma\sigma^2_{\min}$, for

$$\gamma \leq \left(1 + 16\sqrt{\frac{\log \frac{N^2}{\tau}}{d}}\right) \Bigg/ \left(1 - 4\sqrt{\frac{\log \frac{N^2}{\tau}}{d}}\right) \leq 1 + \frac{1}{k},$$

where the second inequality follows for $d$ sufficiently large. Therefore, we compute $H_2$, which contains all samples from such components. This is repeated for at most $k$ iterations, since if a set $H_{\ell+1}$ is distinct from $H_\ell$, it must have added at least one component, and we have only $k$ components. Note that $\left(1 + \frac{1}{k}\right)^k \leq e$, giving the upper bound on variances in $H_{\ell^*}$.

After this clustering step, the algorithm follows similarly to before. The main difference is in the convex programming steps and concentration bounds. For instance,

[5] We require an additional guess of "$k_1$ and $k_2$": the split into how many components are within $H_{\ell^*}$ and $\overline{H}_{\ell^*}$, respectively.

before we considered the set

$$\mathcal{C}_\eta = \left\{ w \in S_{N,\varepsilon} : \left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - \sigma^2 I \right.\right.$$
$$\left.\left. - \sum_{j \in [k]} \alpha_j (\mu_j - \mu)(\mu_j - \mu)^T \right\|_2 \leq \eta \right\}.$$

Now, to reflect the different expression for the covariance of the GMM, we replace $\sigma^2 I$ with $\sum_{j \in [k]} \alpha_j \sigma_j^2 I$; for example,

$$\mathcal{C}_\eta = \left\{ w \in S_{N,\varepsilon} : \left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T \right.\right.$$
$$\left.\left. - \sum_{j \in [k]} \alpha_j \sigma_j^2 I - \sum_{j \in [k]} \alpha_j (\mu_j - \mu)(\mu_j - \mu)^T \right\|_2 \leq \eta \right\}.$$

We note that since all variances in each cluster are off by a factor of at most $e$, this will only affect our concentration and agnostic guarantees by a constant factor.

**7. Agnostically learning binary product distributions, via filters.** In this section, we study the problem of agnostically learning a binary product distribution. Such a distribution is entirely determined by its coordinatewise mean, which we denote by the vector $p$, and our first goal is to estimate $p$ within $\ell_2$-distance $\widetilde{O}(\varepsilon)$. Recall that the approach for robustly learning the mean of an identity covariance Gaussian, sketched in the introduction, was to compute the top absolute eigenvalue of a modified empirical covariance matrix. Our modification was crucially based on the promise that the covariance of the Gaussian is the identity. Here, it turns out that what we should do to modify the empirical covariance matrix is subtract a diagonal matrix whose entries are $p_i^2$. These values seem challenging to directly estimate. Instead, we directly zero out the diagonal entries of the empirical covariance matrix. Then the filtering approach proceeds as before, and allows us to estimate $p$ within $\ell_2$-distance $\widetilde{O}(\varepsilon)$, as we wanted. In the case when $p$ has no coordinates that are too biased towards either zero or one, our estimate is already $\widetilde{O}(\varepsilon)$ close in total variation distance. We give an agnostic learning algorithm for this so-called balanced case (see Definition 7.2) in section 7.1.

However, when $p$ has some very biased coordinates, this need not be the case. Each coordinate that is biased needs to be learned multiplicatively correctly. Nevertheless, we can use our estimate for $p$ that is close in $\ell_2$-distance as a starting point for handling binary product distributions that have imbalanced coordinates. Instead, we control the total variation distance via the $\chi^2$-distance between the mean vectors. Let $P$ and $Q$ be two product distributions whose means are $p$ and $q$, respectively. From Lemma 2.17, it follows that

$$d_{\mathrm{TV}}(P, Q)^2 \leq 4 \sum_i \frac{(p_i - q_i)^2}{q_i(1 - q_i)}.$$

So, if our estimate $q$ is already close in $\ell_2$-distance to $p$, we can interpret the RHS above as giving a renormalization of how we should measure the distance between $p$ and $q$ such that being close (in $\chi^2$-distance) implies that our estimate is close in total variation distance. We can then set up a corrected eigenvalue problem using our initial estimate $q$ as follows. Let $\chi^2(v)_q = \sum_i v_i^2 q_i(1-q_i)$. Then, we compute

$$\max_{\chi^2(v)_q=1} v^T \Sigma v \ ,$$

where $\Sigma$ is the modified empirical covariance. Ultimately, we show that this yields an estimate that is $\widetilde{O}(\sqrt{\varepsilon})$ close in total variation distance. See section 7.2 for further details.

**7.1. The balanced case.** The main result of this section is the following theorem.

THEOREM 7.1. *Let $P$ be a binary product distribution in $d$ dimensions and $\varepsilon, \tau > 0$. Let $S$ be a multiset of $\Theta(d^4 \log(1/\tau)/\varepsilon^2)$ independent samples from $P$, and $S'$ be a multiset obtained by arbitrarily changing an $\varepsilon$-fraction of the points in $S$. There exists a polynomial-time algorithm that returns a product distribution $P'$ such that, with probability at least $1-\tau$, we have $\|p-p'\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$, where $p$ and $p'$ are the mean vectors of $P$ and $P'$, respectively.*

Note that Theorem 7.1 applies to all binary product distributions, and its performance guarantee relates the $\ell_2$-distance between the mean vectors of the hypothesis $P'$ and the target product distribution $P$. If $P$ is balanced, i.e., it does not have coordinates that are too biased towards 0 or 1, this $\ell_2$-guarantee implies a similar total variation guarantee. Formally, we have the following definition.

DEFINITION 7.2. *For $0 < c < 1/2$, we say that a binary product distribution is* $c$-balanced *if the expectation of each coordinate is in $[c, 1-c]$.*

For $c$-balanced binary product distributions, we have the following corollary of Lemma 2.17.

FACT 7.3. *Let $P$ and $Q$ be $c$-balanced binary product distributions with mean vectors $p$ and $q$. Then, we have that $d_{\mathrm{TV}}(P,Q) = O\left(c^{-1/2} \cdot \|p-q\|_2\right)$.*

That is, for two $c$-balanced binary product distributions, where $c$ is a fixed constant, the $\ell_2$-distance between their mean vectors is a good proxy for their total variation distance. Using Fact 7.3, we obtain the following corollary of Theorem 7.1.

COROLLARY 7.4. *Let $P$ be a $c$-balanced binary product distribution in $d$ dimensions, where $c > 0$ is a fixed constant, and $\varepsilon, \tau > 0$. Let $S$ be a multiset of $\Theta(d^4 \log(1/\tau) /\varepsilon^2)$ independent samples from $P$, and $S'$ be a multiset obtained by arbitrarily changing an $\varepsilon$-fraction of the points in $S$. There exists a polynomial-time algorithm that returns a product distribution $P'$ such that with probability at least $1-\tau$, $d_{\mathrm{TV}}(P', P) = O(\varepsilon\sqrt{\log(1/\varepsilon)}/\sqrt{c})$.*

We start by defining a condition on the uncorrupted set of samples $S$, under which our algorithm will succeed.

DEFINITION 7.5 (good set of samples). *Let $P$ be an arbitrary distribution on $\{0,1\}^d$ and $\varepsilon > 0$. We say that a multiset $S$ of elements in $\{0,1\}^d$ is* $\varepsilon$-good with respect to $P$ *if for every affine function $L : \{0,1\}^d \to \mathbb{R}$ we have $|\Pr_{X \in_u S}(L(X) \geq 0) - \Pr_{X \sim P}(L(X) \geq 0)| \leq \varepsilon/d$.*

The following simple lemma shows that a sufficiently large set of independent samples from $P$ is $\varepsilon$-good (with respect to $P$) with high probability.

LEMMA 7.6. *Let $P$ be an arbitrary distribution on $\{0,1\}^d$ and $\varepsilon, \tau > 0$. If the multiset $S$ is obtained by taking $\Omega((d^4 + d^2 \log(1/\tau))/\varepsilon^2)$ independent samples from $P$, it is $\varepsilon$-good with respect to $P$ with probability at least $1-\tau$.*

*Proof.* For a fixed affine function $L : \{0,1\}^d \to \mathbb{R}$, an application of the Chernoff bound yields that after drawing $N$ samples from $P$, we have that $|\Pr_{X \in_u S}(L(X) \geq 0) - \Pr_{X\sim P}(L(X) \geq 0)| > \varepsilon/d$ with probability at most $2\exp(-N\varepsilon^2/d^2)$. Since there are at most $2^{d^2}$ distinct linear threshold functions on $\{0,1\}^d$, by the union bound, the probability that there exists an $L$ satisfying the condition $|\Pr_{X \in_u S}(L(X) \geq 0) - \Pr_{X\sim P}(L(X) \geq 0)| > \varepsilon/d$ is at most $2^{d^2+1}\exp(-N\varepsilon^2/d^2)$, which is at most $\tau$ for $N = \Omega((d^4 + d^2 \log(1/\tau))/\varepsilon^2)$. ∎

Recall (see Definition 5.4) that $\Delta(S, S')$ is the size of the symmetric difference of $S$ and $S'$ divided by the cardinality of $S$.

Our agnostic learning algorithm establishing Theorem 7.1 is obtained by repeated application of the efficient procedure whose performance guarantee is given in the following proposition.

PROPOSITION 7.7. *Let $P$ be a binary product distribution with mean vector $p$ and $\varepsilon > 0$ be sufficiently small. Let $S$ be $\varepsilon$-good with respect to $P$, and $S'$ be any multiset with $\Delta(S, S') \leq 2\varepsilon$. There exists a polynomial-time algorithm* FILTER-BALANCED-PRODUCT *that, given $S'$ and $\varepsilon > 0$, returns one of the following:*

(i) *a mean vector $p'$ such that $\|p - p'\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$,*

(ii) *a multiset $S'' \subseteq S'$ such that $\Delta(S, S'') \leq \Delta(S, S') - 2\varepsilon/d$.*

We start by showing how Theorem 7.1 follows easily from Proposition 7.7.

*Proof of Theorem* 7.1. The proof of Theorem 7.1 is very similar to that of Theorem 5.1; however, we include it here for completeness. By the definition of $\Delta(S, S')$, since $S'$ has been obtained from $S$ by corrupting an $\varepsilon$-fraction of the points in $S$, we have that $\Delta(S, S') \leq 2\varepsilon$. By Lemma 7.6, the set $S$ of uncorrupted samples is $\varepsilon$-good with respect to $P$ with probability at least $1 - \tau$. We henceforth condition on this event.

Our algorithm iteratively applies the FILTER-BALANCED-PRODUCT procedure of Proposition 7.7 until it terminates returning a mean vector $p'$ with $\|p - p'\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$. We claim that we need at most $d + 1$ iterations for this to happen. Indeed, the sequence of iterations results in a sequence of sets $S_0 = S', S'_1, \ldots,$ such that $\Delta(S, S'_i) \leq \Delta(S, S') - i \cdot (2\varepsilon/d)$. Thus, if the algorithm does not terminate in the first $d$ iterations, we have $S'_d = S$, and in the next iteration we output the sample mean of $S$. ∎

**7.1.1. Algorithm filter-balanced-product: Proof of Proposition 7.7.** In this section, we describe the efficient procedure establishing Proposition 7.7 followed by its proof of correctness. Our algorithm FILTER-BALANCED-PRODUCT is very simple: We consider the empirical distribution defined by the (corrupted) sample multiset $S'$. We calculate its mean vector $\mu^{S'}$ and covariance matrix $M$. If the matrix $M$ has no large eigenvalues, we return $\mu^{S'}$. Otherwise, we use the eigenvector $v^*$ corresponding to the maximum magnitude eigenvalue $\lambda^*$ of $M$ and the mean vector $\mu^{S'}$ to define a filter. We zero out the diagonal elements of the covariance matrix for the following reason: The diagonal elements could contribute up to $\Omega(1)$ to the spectral norm, even without noise. This would prevent us from obtaining the desired error of $\widetilde{O}(\varepsilon)$. Our efficient filtering procedure is presented in detailed pseudocode below.

**Algorithm 12** Filter algorithm for a balanced binary product distribution.

**procedure** FILTER-BALANCED-PRODUCT($\varepsilon, S'$)
**input:** A multiset $S'$ such that there exists an $\varepsilon$-good $S$ with $\Delta(S, S') \leq 2\varepsilon$
**output:** Multiset $S''$ or mean vector $p'$ satisfying Proposition 7.7
  Compute the sample mean $\mu^{S'} = \mathbb{E}_{X \in_u S'}[X]$ and the sample covariance $M$ with zeroed diagonal, i.e., $M = (M_{i,j})_{1 \leq i,j \leq d}$ with $M_{i,j} = \mathbb{E}_{X \in_u S'}[(X_i - \mu_i^{S'})(X_j - \mu_j^{S'})]$, $i \neq j$, and $M_{i,i} = 0$.
  Compute approximations for the largest absolute eigenvalue of $M$, $\lambda^* := \|M\|_2$, and the associated unit eigenvector $v^*$.
  **if** $\|M\|_2 \leq O(\varepsilon \log(1/\varepsilon))$ **then return** $\mu^{S'}$.
  Let $\delta := 3\sqrt{\varepsilon \|M\|_2}$.
  Find $T > 0$ such that

$$\Pr_{X \in_u S'}(|v^* \cdot (X - \mu^{S'})| > T + \delta) > 8\exp(-T^2/2) + 8\varepsilon/d.$$

  **return** the multiset $S'' = \{x \in S' : |v^* \cdot (x - \mu^{S'})| \leq T + \delta\}$.

**Tightness of our analysis.** We remark that the analysis of our filter-based algorithm is tight, and more generally our bound of $O(\varepsilon\sqrt{\log(1/\varepsilon)})$ is a bottleneck for filter-based approaches.

More specifically, we note that our algorithm will never successfully add points back to $S$ after they have been removed by the adversary. Therefore, if an $\varepsilon$-fraction of the points in $S$ are changed, our algorithm may be able to remove these outliers from $S'$, but will not be able to replace them with their original values. These changed values can alter the sample mean by as much as $\Omega(\varepsilon\sqrt{\log(1/\varepsilon)})$.

To see this, consider the following example. Let $P$ be the product distribution with mean $p$, where $p_i = 1/2$ for all $i$. Set $\varepsilon = 2^{-(d-1)}$. We draw a $\Theta(d^4 \log(1/\tau)/\varepsilon^2)$ size multiset $S$ which we assume is $\varepsilon$-good. The fraction of times the all-zero vector appears in $S$ is less than $2^{-(d-1)}$. So, the adversary is allowed to corrupt all such zero-vectors. More specifically, the adversary replaces each occurrence of the all-zero vector with fresh samples from $P$, repeating if any all-zero vector is drawn. In effect, this procedure generates samples from the distribution $\widetilde{P}$, defined as $P$ conditioned on not being the all-zero vector. Indeed, with high probability, the set $S'$ is $\varepsilon$-good for $\widetilde{P}$. So, with high probability, the mean of $S'$ in each coordinate is at least $1/2 + 2^{-(d+2)}$. Thus, the $\ell_2$-distance between the mean vectors of $P$ and $\widetilde{P}$ is at least $\sqrt{d}2^{-(d+2)} = \Theta(\varepsilon\sqrt{\log(1/\varepsilon)})$. Note that for any affine function $L$, we have that $\Pr_{X \in_u S'}(L(X) \geq 0) \leq \Pr_{X \in_u S}(L(X) \geq 0)/(1-\varepsilon) + 2\varepsilon/d$, which means that no such function can effectively distinguish between $S' \setminus S$ and $S$, as would be required by a useful filter.

The rest of this section is devoted to the proof of correctness of algorithm FILTER-BALANCED-PRODUCT.

**7.1.2. Setup and basic structural lemmas.** By definition, there exist disjoint multisets $L, E$, of points in $\{0,1\}^d$, where $L \subset S$, such that $S' = (S \setminus L) \cup E$. With this notation, we can write $\Delta(S, S') = \frac{|L|+|E|}{|S|}$. Our assumption $\Delta(S, S') \leq 2\varepsilon$ is equivalent to $|L| + |E| \leq 2\varepsilon \cdot |S|$, and the definition of $S'$ directly implies that $(1-2\varepsilon)|S| \leq |S'| \leq (1+2\varepsilon)|S|$. Throughout the proof, we assume that $\varepsilon$ is a sufficiently small constant. Our analysis will make essential use of the following matrices:

- $M_P$ denotes the matrix with $(i,j)$-entry $\mathbb{E}_{X\sim P}[(X_i-\mu_i^{S'})(X_j-\mu_j^{S'})]$, but 0 on the diagonal.
- $M_S$ denotes the matrix with $(i,j)$-entry $\mathbb{E}_{X\in_u S}[(X_i-\mu_i^{S'})(X_j-\mu_j^{S'})]$, but 0 on the diagonal.
- $M_E$ denotes the matrix with $(i,j)$-entry $\mathbb{E}_{X\in_u E}[(X_i-\mu_i^{S'})(X_j-\mu_j^{S'})]$.
- $M_L$ denotes the matrix with $(i,j)$-entry $\mathbb{E}_{X\in_u L}[(X_i-\mu_i^{S'})(X_j-\mu_j^{S'})]$.

Our first claim follows from the Chernoff bound and the definition of a good set.

CLAIM 7.8. *Let $w\in\mathbb{R}^d$ be any unit vector; then for any $T>0$,*

$$\Pr_{X\in_u S}(|w\cdot(X-\mu^{S'})|>T+\|\mu^{S'}-p\|_2)\leq 2\exp(-T^2/2)+\varepsilon/d$$

*and*

$$\Pr_{X\sim P}(|w\cdot(X-\mu^{S'})|>T+\|\mu^{S'}-p\|_2)\leq 2\exp(-T^2/2).$$

*Proof.* Since $S$ is $\varepsilon$-good, the first inequality follows from the second one. To prove the second inequality, it suffices to bound the probability that $|w\cdot(X-\mu^{S'})-\mathbb{E}[w\cdot(X-\mu^{S'})]|>T$, $X\sim P$, since the expectation in question is $w\cdot(p-\mu^{S'})$, whose absolute value is at most $\|\mu^{S'}-p\|_2$, by Cauchy–Schwarz. Note that $w\cdot(X-\mu^{S'})$ is a sum of independent random variables $w_i(X_i-\mu_i^{S'})$, each supported on an interval of length $2|w_i|$. An application of the Chernoff bound completes the proof. ∎

The following sequence of lemmas is bound from above the spectral norms of the associated matrices. Our first simple lemma says that the (diagonally reduced) empirical covariance matrix $M_S$, where $S$ is the set of uncorrupted samples drawn from the binary product distribution $P$, is a good approximation to the matrix $M_P$, in spectral norm.

LEMMA 7.9. *If $S$ is $\varepsilon$-good, $\|M_P-M_S\|_2\leq O(\varepsilon)$.*

*Proof.* It suffices to show that $|(M_P)_{i,j}-(M_S)_{i,j}|\leq O(\varepsilon/d)$ for all $i\neq j$. Then, we have that

$$\|M_P-M_S\|_2\leq\|M_P-M_S\|_F\leq O(\varepsilon).$$

Let $e_i$ denote the standard basis vector in the $i$th direction in $\mathbb{R}^d$. For $i\neq j$ we have

$$\begin{aligned}(M_P)_{i,j}&=\mathop{\mathbb{E}}_{X\sim P}[(X_i-\mu_i^{S'})(X_j-\mu_j^{S'})]\\&=\mathop{\mathbb{E}}_{X\sim P}[X_iX_j]-\mu_i^{S'}\mathop{\mathbb{E}}_{X\sim P}[X_j]-\mu_j^{S'}\mathop{\mathbb{E}}_{X\sim P}[X_i]+\mu_j^{S'}\mu_i^{S'}\\&=\Pr_{X\sim P}((e_i+e_j)\cdot X\geq 2)-\mu_i^{S'}\Pr_{X\sim P}(e_j\cdot X\geq 1)\\&\quad-\mu_j^{S'}\Pr_{X\sim P}(e_i\cdot X\geq 1)+\mu_j^{S'}\mu_i^{S'}.\end{aligned}$$

A similar expression holds for $M_S$ except with probabilities for $X\in_u S$. Since $S$ is $\varepsilon$-good with respect to $P$, we have $|(M_P)_{i,j}-(M_S)_{i,j}|\leq\varepsilon/d+\mu_i^{S'}\varepsilon/d+\mu_j^{S'}\varepsilon/d\leq 3\varepsilon/d$. This completes the proof. ∎

As a simple consequence of the above lemma, we obtain the following claim.

CLAIM 7.10. *If $S$ is $\varepsilon$-good, $\|M-(1/|S'|)(|S|M_P+|E|M_E-|L|M_L)\|_2=O(\varepsilon)$.*

*Proof.* First note that we can write $|S'|M=|S|M_S+|E|M_E^0-|L|M_L^0$, where $M_E^0$ and $M_L^0$ are obtained from $M_E$ and $M_L$ by zeroing out the diagonal. Observe that $|E|+|L|=O(\varepsilon)|S'|$. This follows from the assumption that $\Delta(S,S')\leq 2\varepsilon$ and the definition of $S'$. Now note that the matrices $M_E-M_E^0$ and $M_L-M_L^0$ are diagonal with entries at most 1, and thus have spectral norm at most 1. The claim now follows from Lemma 7.9. ∎

Recall that if $\mu^{S'} = p$, $M_P$ would equal the (diagonally reduced) covariance matrix of the product distribution $P$, i.e., the identically zero matrix. The following simple lemma bounds from above the spectral norm of $M_P$ by the $\ell_2^2$-norm between the corresponding mean vectors.

LEMMA 7.11. *We have that* $\|M_P\|_2 \leq \|\mu^{S'} - p\|_2^2$.

*Proof.* Note that $(M_P)_{i,j} = (\mu_i^{S'} - p_i)(\mu_j^{S'} - p_j)$ for $i \neq j$ and 0 otherwise. Therefore, $M_P$ is the difference of $(\mu^{S'} - p)(\mu^{S'} - p)^T$ and the diagonal matrix with entries $(\mu_i^{S'} - p_i)^2$. This, in turn, implies that

$$(\mu^{S'} - p)(\mu^{S'} - p)^T \succeq M_P \succeq \mathrm{Diag}(-(\mu_i^{S'} - p_i)^2) \; .$$

Note that both bounding matrices have spectral norm at most $\|\mu^{S'} - p\|_2^2$, hence so does $M_P$. ∎

The following lemma, bounding from above the spectral norm of $M_L$, is the main structural result of this section. This is the core result needed to establish that the subtractive error cannot change the sample mean by much.

LEMMA 7.12. *We have that* $\|M_L\|_2 = O(\log(|S|/|L|) + \|\mu^{S'} - p\|_2^2 + \varepsilon \cdot |S|/|L|)$, *hence*

$$(|L|/|S'|) \cdot \|M_L\|_2 = O(\varepsilon \log(1/\varepsilon) + \varepsilon \|\mu^{S'} - p\|_2^2).$$

*Proof.* Since $L \subseteq S$, for any $x \in \{0,1\}^d$, we have that

$$|S| \cdot \Pr_{X \in_u S}(X = x) \geq |L| \cdot \Pr_{X \in_u L}(X = x) \; . \tag{35}$$

Since $M_L$ is a symmetric matrix, we have $\|M_L\|_2 = \max_{\|v\|_2 = 1} |v^T M_L v|$. So, to bound $\|M_L\|_2$ it suffices to bound $|v^T M_L v|$ for unit vectors $v$. By definition of $M_L$, for any $v \in \mathbb{R}^d$ we have that

$$|v^T M_L v| = \mathop{\mathbb{E}}_{X \in_u L}[|v \cdot (X - \mu^{S'})|^2].$$

The RHS is, in turn, bounded from above as follows:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X \in_u L}[|v \cdot (X - \mu^{S'})|^2] \\
&= 2 \int_0^{\sqrt{d}} \Pr_{X \in_u L}\left(|v \cdot (X - \mu^{S'})| > T\right) \cdot T dT \\
&\leq 2 \int_0^{\sqrt{d}} \min\left\{1, |S|/|L| \cdot \Pr_{X \in_u S}\left(|v \cdot (X - \mu^{S'})| > T\right)\right\} T dT \\
&\ll \int_0^{4\sqrt{\log(|S|/|L|)} + \|\mu^{S'} - p\|_2} T dT \\
&+ (|S|/|L|) \int_{4\sqrt{\log(|S|/|L|)} + \|\mu^{S'} - p\|_2}^{\sqrt{d}} \left(\exp(-(T - \|\mu^{S'} - p\|_2)^2/2) T + \varepsilon T/d\right) dT \\
&\ll \log(|S|/|L|) + \|\mu^{S'} - p\|_2^2 + \varepsilon \cdot |S|/|L| \; ,
\end{aligned}$$

where the second line follows from (35) and the third line follows from Claim 7.8. This establishes the first part of the lemma.

The bound $(|L|/|S|)\|M_L\|_2 = O(\varepsilon \log(1/\varepsilon) + \varepsilon\|\mu^{S'} - p\|_2^2)$ follows from the previously established bound using the monotonicity of the function $x\log(1/x)$, and the fact that $|L|/|S| \le 2\varepsilon$. The observation $|S|/|S'| \le 1 + 2\varepsilon \le 2$ completes the proof of the second part of the lemma. ∎

Claim 7.10 combined with Lemmas 7.11 and 7.12 and the triangle inequality yield the following corollary.

COROLLARY 7.13. *We have that* $\|M - (|E|/|S'|)M_E\|_2 = O(\varepsilon \log(1/\varepsilon) + \|\mu^{S'} - p\|_2^2)$.

We are now ready to analyze the two cases of the algorithm FILTER-BALANCED-PRODUCT.

**7.1.3. The case of small spectral norm.** We start by analyzing the case where the mean vector $\mu^{S'}$ is returned. This corresponds to the case that the spectral norm of $M$ is appropriately small, namely $\|M\|_2 \le O(\varepsilon \log(1/\varepsilon))$. We start with the following simple claim.

CLAIM 7.14. *Let* $\mu^E, \mu^L$ *be the mean vectors of* $E$ *and* $L$*, respectively. Then,* $\|\mu^E - \mu^{S'}\|_2^2 \le \|M_E\|_2$ *and* $\|\mu^L - \mu^{S'}\|_2^2 \le \|M_L\|_2$.

*Proof.* We prove the first inequality, the proof of the second being identical. Note that $M_E$ is a symmetric matrix, so $\|M_E\|_2 = \max_{\|v\|_2=1} |v^T M_E v|$. Moreover, for any vector $v$ we have that

$$v^T M_E v = \mathop{\mathbb{E}}_{X \in_u E}[|v \cdot (X - \mu^{S'})|^2] \ge |v \cdot (\mu^E - \mu^{S'})|^2.$$

Let $w = \mu^E - \mu^{S'}$ and take $v = w/\|w\|_2$. We conclude that $\|M_E\|_2 \ge \|w\|_2^2$, as desired. ∎

The following crucial lemma, bounding from above the distance $\|\mu^{S'} - p\|_2$ as a function of $\varepsilon$ and $\|M\|_2$, will be important for both this and the following subsections.

LEMMA 7.15. *We have that* $\|\mu^{S'} - p\|_2 \le 2\sqrt{\varepsilon\|M\|_2} + O(\varepsilon\sqrt{\log(1/\varepsilon)})$.

*Proof.* First, we observe that the mean vector $\mu^S$ of the uncorrupted sample set $S$ is close to $p$. Since $S$ is $\varepsilon$-good, this follows from the fact that for any $i \in [d]$, we have

$$|\mu_i^S - p_i| = |\Pr_{X \in_u S}[e_i \cdot X \ge 1] - \Pr_{X \sim P}[e_i \cdot X \ge 1]| \le \varepsilon/d.$$

Therefore, we get that $\|\mu^S - p\|_2 \le \varepsilon/\sqrt{d}$.

Consider $\mu^E$ and $\mu^L$, the mean vectors of $E$ and $L$, respectively. By definition, we have that

$$|S'|\mu^{S'} = |S|\mu^S + |E|\mu^E - |L|\mu^L ,$$

and thus by the triangle inequality we obtain

$$\|\mu^{S'} - p\|_2 \le \|(|E|/|S'|)(\mu^E - p) - (|L|/|S'|)(\mu^L - p)\|_2 + \varepsilon/\sqrt{d} .$$

Therefore, we have the following sequence of inequalities:

$$\begin{aligned}
&\|\mu^{S'} - p\|_2 \\
&\le (|E|/|S'|) \cdot \|\mu^E - \mu^{S'}\|_2 + (|L|/|S'|) \cdot \|\mu^L - \mu^{S'}\|_2 + O(\varepsilon) \cdot \|\mu^{S'} - p\|_2 + \varepsilon/\sqrt{d} \\
&\le (|E|/|S'|) \cdot \sqrt{\|M_E\|_2} + (|L|/|S'|) \cdot \sqrt{\|M_L\|_2} + O(\varepsilon) \cdot \|\mu^{S'} - p\|_2 + \varepsilon/\sqrt{d} \\
&\le O(\varepsilon\sqrt{\log(1/\varepsilon)}) + (3/2)\sqrt{\varepsilon\|M\|_2} + O(\sqrt{\varepsilon}) \cdot \|\mu^{S'} - p\|_2 \\
&\le O(\varepsilon\sqrt{\log(1/\varepsilon)}) + (3/2)\sqrt{\varepsilon\|M\|_2} + \|\mu^{S'} - p\|_2/4 ,
\end{aligned}$$

where the first line follows from the triangle inequality, the second uses Claim 7.14, while the third uses Lemma 7.12 and Corollary 7.13. Finally, the last couple of lines assume that $\varepsilon$ is sufficiently small. The proof of Lemma 7.15 is now complete. ∎

We can now deduce the correctness of step 4 of Algorithm 12, FILTER-BALANCED-PRODUCT, since for $\|M\|_2 \leq O(\varepsilon \log(1/\varepsilon))$, Lemma 7.15 directly implies that $\|\mu^{S'} - p\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$.

**7.1.4. The case of large spectral norm.** We next show the correctness of the algorithm FILTER-BALANCED-PRODUCT if it returns a filter (rejecting an appropriate subset of $S'$) in step 5. This corresponds to the case that $\|M\|_2 \geq C\varepsilon\log(1/\varepsilon)$, for a sufficiently large universal constant $C > 0$. We will show that the multiset $S'' \subset S'$ computed in step 5 satisfies $\Delta(S, S'') \leq \Delta(S, S') - 2\varepsilon/d$.

We start by noting that, as a consequence of Lemma 7.15, we have the following claim.

CLAIM 7.16. *We have that* $\|\mu^{S'} - p\|_2 \leq \delta := 3\sqrt{\varepsilon\|M\|_2}$.

*Proof.* By Lemma 7.15, we have that $\|\mu^{S'} - p\|_2 \leq 2\delta/3 + O(\varepsilon\sqrt{\log(1/\varepsilon)})$. Recalling that $\|M\|_2 \geq C\varepsilon\log(1/\varepsilon)$, if $C > 0$ is sufficiently large, the term $O(\varepsilon\sqrt{\log(1/\varepsilon)})$ is at most $\delta/3$. ∎

By construction, $v^*$ is the unit eigenvector corresponding to the maximum magnitude eigenvalue of $M$. Thus, we have $(v^*)^T M v^* = \|M\|_2 = \delta^2/(9\varepsilon)$. We thus obtain that

$$\mathop{\mathbb{E}}_{X\in_u E}[|v^*\cdot(X-\mu^{S'})|^2] = (v^*)^T M_E v^* \geq \frac{\delta^2|S'|}{20\varepsilon|E|}\,, \tag{36}$$

where the equality holds by definition, and the inequality follows from Corollary 7.13 and Claim 7.16 using the fact that $\varepsilon$ is sufficiently small and the constant $C$ is sufficiently large (noting that the constant in the RHS of Corollary 7.13 does not depend on $C$).

We show that (36) implies the existence of a $T > 0$ with the properties specified in step 5 of Algorithm 12, FILTER-BALANCED-PRODUCT. More specifically, we have the following crucial lemma.

LEMMA 7.17. *If* $\|M\|_2 \geq C\varepsilon\log(1/\varepsilon)$, *for a sufficiently large constant* $C > 0$, *there exists a* $T > 0$ *satisfying the property in step* 5 *of Algorithm* 12, FILTER-BALANCED-PRODUCT, *i.e., such that*

$$\Pr_{X\in_u S'}(|v^*\cdot(X-\mu^{S'})| > T+\delta) > 8\exp(-T^2/2) + 8\varepsilon/d\,.$$

*Proof.* Assume for the sake of contradiction that this is not the case, i.e., that for all $T > 0$ we have that

$$\Pr_{X\in_u S'}(|v^*\cdot(X-\mu^{S'})| \geq T+\delta) \leq 8\exp(-T^2/2) + 8\varepsilon/d\,. \tag{37}$$

Since $E \subseteq S'$, for all $x \in \{0,1\}^d$, we have that $|S'|\Pr_{X\in_u S'}[X = x] \geq |E|\Pr_{Y\in_u E}[Y = x]$. This fact, combined with (37), implies that for all $T > 0$

$$\Pr_{Y\in_u E}(|v^*\cdot(Y-\mu^{S'})| \geq T+\delta) \ll (|S'|/|E|)(\exp(-T^2/2) + \varepsilon/d)\,. \tag{38}$$

Using (36) and (38), we have the following sequence of inequalities:

$$\begin{aligned}
\delta^2|S'|/(\varepsilon|E|) &\ll \mathop{\mathbb{E}}_{Y\in_u E}[|v^*\cdot(Y-\mu^{S'})|^2] \\
&= 2\int_0^\infty \Pr_{Y\in_u E}\left(|v^*\cdot(Y-\mu^{S'})|\geq T\right)\cdot TdT \\
&\ll (|S'|/|E|)\int_0^{O(\sqrt{d})} \min\left\{|E|/|S'|, \exp(-(T-\delta)^2/2)+\varepsilon/d\right\}TdT \\
&\ll \int_0^{4\sqrt{\log(|S'|/|E|)}+\delta} Tdt \\
&\quad + \int_{4\sqrt{\log(|S'|/|E|)}+\delta}^\infty (|S'|/|E|)\exp(-(T-\delta)^2/2)TdT + \int_0^{O(\sqrt{d})} \frac{\varepsilon|S'|}{d|E|}TdT \\
&\ll \log(|S'|/|E|) + \delta^2 + \frac{\varepsilon|S'|}{|E|}\ .
\end{aligned}$$

This yields the desired contradiction recalling that the assumption $\|M\|_2 \geq C\varepsilon\log(1/\varepsilon)$ and the definition of $\delta$ imply that $\delta \geq C'\varepsilon\sqrt{\log(1/\varepsilon)}$ for an appropriately large $C' > 0$. ∎

The following simple claim completes the proof of Proposition 7.7.

CLAIM 7.18. *We have that* $\Delta(S,S'') \leq \Delta(S,S') - 2\varepsilon/d$ .

*Proof.* Recall that $S' = (S\setminus L)\cup E$, with $E$ and $L$ disjoint multisets such that $L\subset S$. We can similarly write $S'' = (S\setminus L')\cup E'$, with $L'\supseteq L$ and $E'\subset E$. Since

$$\Delta(S,S') - \Delta(S,S'') = \frac{|E\setminus E'| - |L'\setminus L|}{|S|},$$

it suffices to show that $|E\setminus E'| \geq |L'\setminus L| + 2\varepsilon|S|/d$. Note that $|L'\setminus L|$ is the number of points rejected by the filter that lie in $S\cap S'$. By Claims 7.8 and 7.16, it follows that the fraction of elements $x\in S$ that are removed to produce $S''$ (i.e., satisfy $|v^*\cdot(x-\mu_{S'})| > T+\delta$) is at most $2\exp(-T^2/2)+\varepsilon/d$. Hence, it holds that $|L'\setminus L| \leq (2\exp(-T^2/2)+\varepsilon/d)|S|$. On the other hand, step 5 of Algorithm 12 ensures that the fraction of elements of $S'$ that are rejected by the filter is at least $8\exp(-T^2/2)+8\varepsilon/d$. Note that $|E\setminus E'|$ is the number of points rejected by the filter that lie in $S'\setminus S$. Therefore, we can write

$$\begin{aligned}
|E\setminus E'| &\geq (8\exp(-T^2/2)+8\varepsilon/d)|S'| - (2\exp(-T^2/2)+\varepsilon/d)|S| \\
&\geq (8\exp(-T^2/2)+8\varepsilon/d)|S|/2 - (2\exp(-T^2/2)+\varepsilon/d)|S| \\
&\geq (2\exp(-T^2/2)+3\varepsilon/d)|S| \\
&\geq |L'\setminus L| + 2\varepsilon|S|/d\ ,
\end{aligned}$$

where the second line uses the fact that $|S'| \geq |S|/2$ and the last line uses the fact that $|L'\setminus L|/|S| \leq (2\exp(-T^2/2)+\varepsilon/d)$. This completes the proof of the claim. ∎

**7.2. Agnostically learning arbitrary binary product distributions.** In this subsection, we build on the approach of the previous subsection to show the following theorem.

THEOREM 7.19. *Let* $P$ *be a binary product distribution in* $d$ *dimensions and* $\varepsilon,\tau > 0$. *There is a polynomial-time algorithm that, given* $\varepsilon$ *and a set of* $\Theta(d^6\log(1/\tau)/\varepsilon^3)$

*independent samples from $P$, an $\varepsilon$-fraction of which have been arbitrarily corrupted, outputs (the mean vector of) a binary product distribution $\widetilde{P}$ such that, with probability at least $1-\tau$, $d_{\mathrm{TV}}(P,\widetilde{P}) \leq O(\sqrt{\varepsilon \log(1/\varepsilon)})$.*

By Lemma 2.17, the total variation distance between two binary product distributions can be bounded from above by the square root by the $\chi^2$-distance between the corresponding means. For the case of balanced product distributions, the $\chi^2$-distance and the $\ell_2$-distance are within a constant factor of each other. Unfortunately, this does not hold in general, hence the guarantee of our previous algorithm is not sufficient to get a bound on the total variation distance. Note, however, that the $\chi^2$-distance and the $\ell_2$-distance can be related by rescaling each coordinate by the standard deviation of the corresponding marginal. When we rescale the covariance matrix in this way, we can use the top eigenvalue and eigenvector as before, except that we obtain bounds that involve the $\chi^2$- in place of the $\ell_2$-distance. The concentration bounds we obtain with this rescaling are somewhat weaker, and as a result, our quantitative guarantees for the general case are correspondingly weaker than in the balanced case. As in the filter algorithm for approximating the mean under second moment assumptions in [DKK+17], to handle this weaker concentration, we will choose a threshold at random, weighted towards larger thresholds instead of looking for a violation of a concentration inequality. This gives a filter that rejects more corrupted than uncorrupted samples in expectation and we will show that with high probability we still only throw away an $O(\varepsilon)$-fraction of samples in the course of the algorithm.

Similar to the case of balanced product distributions, we will require a notion of a "good" set for our distribution. For technical reasons, the definition in this setting turns out to be more complicated. Roughly speaking, this is to allow us to ignore coordinates for which the small fraction of errors is sufficient to drastically change the sample mean.

DEFINITION 7.20 (good set of samples). *Let $P$ be a binary product distribution and $\varepsilon, \eta > 0$. We say that a multiset $S$ of elements in $\{0,1\}^d$ is* $(\varepsilon,\eta)$-good *with respect to $P$ if for every affine function $L : \{0,1\}^d \to \mathbb{R}$ and every subset of coordinates $T \subseteq [d]$ satisfying $\sum_{i\in T} p_i(1-p_i) < \eta$ the following holds: Letting $S_T$ be the subset of points in $S$ that have their $i$th coordinate equal to the most common value under $P$ for all $i \in T$, and letting $P_T$ be the conditional distribution of $P$ under this condition, then*

$$\left| \Pr_{X\in_u S_T}(L(X)\geq 0) - \Pr_{X\sim P_T}(L(X)\geq 0)\right| \leq \varepsilon^{3/2}/d^2 .$$

We note that a sufficiently large set of samples from $P$ will satisfy the above properties with high probability.

LEMMA 7.21. *If $S$ is obtained by taking $\Omega(d^6 \log(1/\tau)/\varepsilon^3)$ independent samples from $P$, it is $(\varepsilon, 1/5)$-good with respect to $P$ with probability at least $9/10$.*

The proof of this lemma is deferred to Appendix D.

We will also require a notion of the number of coordinates on which $S$ nontrivially depends.

DEFINITION 7.22. *For $S$ a multiset of elements in $\{0,1\}^d$, let $\operatorname{supp}(S)$ be the subset of $[d]$ consisting of indices $i$ such that the $i$th coordinate of elements of $S$ is not constant.*

Similar to the balanced case, our algorithm is obtained by repeated application of an efficient filter procedure, whose precise guarantee is described below.

PROPOSITION 7.23. *Let $P$ be a binary product distribution in $d$ dimensions and $\varepsilon > 0$. Suppose that $S$ is an $(\varepsilon, \eta)$-good multiset with respect to $P$ with $\eta > 10\varepsilon$ and $S'$ be any multiset with $\Delta(S, S') \leq 20\varepsilon$. There exists a polynomial-time algorithm which, given $\varepsilon$ and $S'$, returns one of the following:*

(i) *The mean vector of a product distribution $P'$ with $d_{\mathrm{TV}}(P, P') = O(\sqrt{\varepsilon \log(1/\varepsilon)})$.*

(ii) *A multiset $S'' \subset S'$ of elements of $\{0,1\}^d$ such that there exists a product distribution $\widetilde{P}$ with mean $\widetilde{p}$ and a multiset $\widetilde{S}$ that is $(\varepsilon, \eta - \|p - \widetilde{p}\|_1)$-good with respect to $\widetilde{P}$ such that*

$$\mathbb{E}[\Delta(\widetilde{S}, S'')] + \|p - \widetilde{p}\|_1/6 \leq \Delta(S, S') .$$

Our agnostic learning algorithm is then obtained by iterating this procedure. We can prove Theorem 7.19 given Proposition 7.23.

*Proof of Theorem* 7.19. We draw $N = \Theta(d^6/\varepsilon^3)$ samples forming a set $S$, which is $(\varepsilon, 1/5)$-good with probability 9/10 by Lemma 7.21. We condition on this event. The adversary corrupts an $\varepsilon$-fraction of $S$ producing a set $S'$ with $\Delta(S, S') \leq 2\varepsilon$. The iterations of the algorithm produce a sequence of sets $S_0 = S, S_1, \ldots, S_k$, where $S_i$ is $(\varepsilon, \eta_i)$-good for some binary product distribution $P_i$ and some sets $S'_i$. We note that $\Delta(S_i, S'_i)$ is monotonically decreasing in expectation. Since $|\mu^{P_i} - \mu^{P_{i+1}}| \leq d_{\mathrm{TV}}(P_i, P_{i+1})$, in the $i$th iteration, we have that $\mathbb{E}[\Delta(S_{i+1}, S'_{i+1}) - d_{\mathrm{TV}}(P_i, P_{i+1})] \leq \Delta(S_i, S'_i)$, as long as $\Delta(S_i, S'_i) \leq 20\varepsilon$.

We need to show that the probability that we will ever have $\Delta(S_i, S'_i) > 20\varepsilon$ is small. Indeed we show that the probability that $\Delta(S_i, S'_i) + \sum_{j=0}^{i-1} d_{\mathrm{TV}}(P_i, P_{i+1})$ is ever large is 1/10.

We analyze the following procedure: We iteratively run FILTER-PRODUCT. We stop if we output an approximation to the mean or if $\Delta(S_i, S'_i) + \sum_{j=0}^{i-1} d_{\mathrm{TV}}(P_i, P_{i+1}) > 20\varepsilon|S|$. Proposition 7.23 gives that $\mathbb{E}[\Delta(S_{i+1}, S'_{i+1}) - d_{\mathrm{TV}}(P_i, P_{i+1})/6] \leq \Delta(S_i.S'_i)$. This expectation is conditioned on the state of the algorithm after previous iterations, which is determined by $S'_i$. Thus, if we consider the random variables $X_i = \Delta(S_i, S'_i) + \sum_{j=0}^{i-1} d_{\mathrm{TV}}(P_i, P_{i+1})/6$, then we have $\mathbb{E}[X_{i+1}|S'_i] \leq X_i$, i.e., the sequence $X_i$ is a submartingale with respect to $S'_i$. Using the convention that $S'_{i+1} = S'_i$, if we stop in fewer than $i$ iterations, since we must terminate $N$ iterations as every iteration removes at least one sample, the algorithm fails if and only if $|X_N| > 20\varepsilon$. By a simple induction or standard results on submartingales, we have $\mathbb{E}[X_N] \leq X_0$. Now $X_0 = \Delta(S_0, S'_0) \leq 2\varepsilon|S'_0|$. Thus, $\mathbb{E}[X_N] \leq 2\varepsilon|S|$. By Markov's inequality, except with probability 1/10, we have $X_N \leq 20\varepsilon|S|$. Therefore, the probability that we will ever have $|X_i| > 20\varepsilon$ is at most 1/10.

By a union bound, using Lemma 7.21, $S_0$ is $(\varepsilon, 1/5)$-good and we have $|X_i| \leq 20\varepsilon$ with probability at least 4/5. We assume that this holds. By induction, $S_i$ is $(\varepsilon, 1/5 - \sum_{j=0}^{i-1} d_{\mathrm{TV}}(P_i, P_{i+1}))$-good, and so is $(\varepsilon, 1/5 - 100\varepsilon)$-good, which suffices since $1/5 - 100\varepsilon \geq 10\varepsilon$.

When it terminates, the algorithm outputs a product distribution $P'$ with $d_{\mathrm{TV}}(P_k, P') = O(\sqrt{\varepsilon \log(1/\varepsilon)})$. By the triangle inequality, we have that

$$d_{\mathrm{TV}}(P, P') \leq d_{\mathrm{TV}}(P_k, P') + \sum_{j=0}^{k-1} d_{\mathrm{TV}}(P_i, P_{i+1}) \leq O(\sqrt{\varepsilon \log(1/\varepsilon)}) + 100\varepsilon \leq O(\sqrt{\varepsilon \log(1/\varepsilon)}) .$$

When $\tau \leq 1/5$, we will need to draw fresh $\varepsilon$-corrupted samples and repeat this procedure $O(\log(1/\tau))$ times, and then one of the resulting output distributions is within total variation distance $O(\sqrt{\varepsilon \log(1/\varepsilon)})$ with probability at least $1-\tau/2$. Then we use the agnostic hypothesis selection procedure of Lemma 2.23. ∎

**7.2.1. Algorithm filter-product: Proof of Proposition 7.23.** In this section, we describe and analyze the efficient routine establishing Proposition 7.23. Our efficient filtering procedure is presented in detailed pseudocode below.

**Algorithm 13** Filter algorithm for an arbitrary binary product distribution.

1: **procedure** FILTER-PRODUCT($\varepsilon, S'$)
**input:** $\varepsilon > 0$ and multiset $S'$ such that there exists an $\varepsilon$-good $S$ with $\Delta(S, S') \le 2\varepsilon$
**output:** Multiset $S''$ or mean vector $p'$ satisfying Proposition 7.23
2: Compute the sample mean $\mu^{S'} = \mathbb{E}_{X \in_u S'}[X]$ and the sample covariance matrix $M$, i.e., $M = (M_{i,j})_{1 \le i,j \le d}$ with $M_{i,j} = \mathbb{E}_{X \in S'}[(X_i - \mu_i^{S'})(X_j - \mu_j^{S'})]$.
3: **if** there exists $i \in [d]$ with $0 < \mu_i^{S'} < \varepsilon/d$ or $0 < 1 - \mu_i^{S'} < \varepsilon/d$ **then**
4: Let $S''$ be the subset of elements of $S'$ in which those coordinates take their most common value.
5: **return** $S''$.
/* *For the later steps, we ignore any coordinates not in* supp($S'$). */
6: Compute approximations for the largest magnitude eigenvalue $\lambda'$ of $DMD$, $\lambda' := \|DMD\|_2$, where $D = \mathrm{Diag}(1/\sqrt{\mu_i^{S'}(1-\mu_i^{S'})})$, and the associated unit eigenvector $v'$.
7: **if** $\|DMD\|_2 < O(\log(1/\varepsilon))$ **then return** $\mu^{S'}$ (reinserting all coordinates affected by step 5).
8: Draw $Z$ from the distribution on $[0,1]$ with probability density function $2x$.
9: Let $T = Z \max\{|v^* \cdot (x - \mu^{S'})| : x \in S'\}$, where $v^* := Dv'$.
10: **return** the multiset $S'' = \{x \in S' : |v^* \cdot (x - \mu^{S'})| < T\}$ .

This completes the description of the algorithm. We now proceed to prove correctness.

**7.2.2. Chi-squared distance and basic reductions.** As previously mentioned, our algorithm will use the $\chi^2$-distance between the mean vectors as a proxy for the total variation distance between two binary product distributions. Since the mean vector of the target distribution is not known to us, we will not be able to use the symmetric definition of the $\chi^2$-distance used in Lemma 2.17. We will instead require the following asymmetric version of the $\chi^2$-distance.

DEFINITION 7.24. *The $\chi^2$-distance of $x, y \in \mathbb{R}^d$ is defined by* $\chi^2(x,y) \stackrel{\mathrm{def}}{=} \sum_{i=1}^d \frac{(x_i-y_i)^2}{x_i(1-x_i)}$.

The next fact follows directly from Lemma 2.17.

FACT 7.25. *Let $P, Q$ be binary product distributions with mean vectors $p, q$, respectively. Then,* $d_{\mathrm{TV}}(P,Q) = O(\sqrt{\chi^2(p,q)})$.

There are two problems with using the $\chi^2$-distance between the mean vectors as a proxy for the total variation distance. The first is that the $\chi^2$-distance between the means is a very loose approximation of the total variational distance when the means are close to 0 or 1 in some coordinate. To circumvent this obstacle, we remove such coordinates via an appropriate preprocessing in step 5 of Algorithm 13. The second is that the above asymmetric notion of the $\chi^2$-distance may be quite far from the symmetric definition. To overcome this issue, it suffices to have that $q_i = O(p_i)$ and

$1 - q_i = O(1 - p_i)$. To ensure this condition is satisfied, we appropriately modify the target product distribution (that we aim to be close to). Next, we will show how we deal with these problems in detail.

Before we embark on a proof of the correctness of algorithm FILTER-PRODUCT, we will make a few reductions that we will apply throughout. First, we note that if some coordinate in step 5 of Algorithm 13 exists, then removing the uncommon values of that coordinate increases $\Delta(S, S')$ by at most $\varepsilon/d$ but decreases $|\text{supp}(S')|$ by at least 1. We also note that if $N$ is the set of coordinates outside of the support of $S'$, the probability that an element in $S'$ has a coordinate in $N$ that does not take its constant value is 0. Note that this is at most $O(\varepsilon)$ away from the probability that an element taken from $P$ has this property, and thus we can assume that $\sum_{i \in N} \min\{p_i, 1-p_i\} = O(\varepsilon)$. Therefore, after step 5 of Algorithm 13, we can assume that all coordinates $i$ have $\varepsilon/d \le p_i \le 1 - \varepsilon/d$.

The next reduction will be slightly more complicated and depends on the following idea: Suppose that there is a new product distribution $\widetilde{P}$ with mean $\widetilde{p}$ and an $(\varepsilon, \eta - \|p - \widetilde{p}\|_1)$-good multiset $\widetilde{S}$ for $\widetilde{P}$ such that

$$\Delta(\widetilde{S}, S') + \|p - \widetilde{p}\|_1/5 \le \Delta(S, S').$$

Then, it suffices to show that our algorithm works for $\widetilde{P}$ and $\widetilde{S}$ instead of $P$ and $S$ (note that the input to the algorithm, $S'$ and $\varepsilon$, is the same in either case). This is because the conditions imposed by the output in this case would be strictly stronger. In particular, we may assume that $\mu_i^{S'} \ge p_i/3$ for all $i$.

LEMMA 7.26. *There is a product distribution $\widetilde{P}$ whose mean vector $\widetilde{p}$ satisfies $\mu_i^{S'} \ge \widetilde{p}_i/3$ and $1 - \mu_i^{S'} \ge (1 - \widetilde{p}_i/3)$ for all $i$, and a set $\widetilde{S} \subseteq S$ that is $(\varepsilon, \eta - \|p - \widetilde{p}\|_1)$-good for $\widetilde{P}$ and satisfies*

$$\Delta(\widetilde{S}, S') + \|p - \widetilde{p}\|_1/5 \le \Delta(S, S').$$

*Proof.* If all coordinates $i$ have $\mu_i^{S'} \ge p_i/3$ and $1 - \mu_i^{S'} \ge (1 - p_i/3)$, then we can take $\widetilde{P} = P$ and $\widetilde{S} = S$.

Suppose that the $i$th coordinate has $\mu_i^{S'} < p_i/3$. Let $\widetilde{P}$ be the product whose mean vector $\widetilde{p}$ has $\widetilde{p}_i = 0$ and $\widetilde{p}_j = p_j$ for $j \ne i$. Let $\widetilde{S}$ be obtained by removing from $S$ all of the entries with 1 in the $i$th coordinate. Then, we claim that $\widetilde{S}$ is $(\varepsilon, \eta - p_i)$-good for $\widetilde{P}$ and has $\Delta(\widetilde{S}, S') + p_i/5 \le \Delta(S, S')$. Note that here we have $\|p - \widetilde{p}\|_1 = p_i$.

First, we show that $\widetilde{S}$ is $(\varepsilon, \eta - p_i)$-good for $\widetilde{P}$. For any affine function $L(x)$ and set $T \subseteq [d]$ with $\sum_{j \in T} \widetilde{p}_j(1 - \widetilde{p}_j) \le \eta - p_i$, we need to show that

$$\left| \Pr_{X \in_u \widetilde{S}_T}(L(X) > 0) - \Pr_{X \sim \widetilde{P}_T}(L(X) > 0) \right| \le \varepsilon^{3/2}/d^2 .$$

Let $\widetilde{T} = T \cup \{i\}$. We may or may not have $i \in T$ but, from the definition of $\widetilde{p}$,

$$\sum_{j \in T} \widetilde{p}_j(1 - \widetilde{p}_j) = \sum_{j \in T \setminus \{i\}} \widetilde{p}_j(1 - \widetilde{p}_j) = \sum_{j \in T \setminus \{i\}} p_j(1 - p_j).$$

Thus,

$$\sum_{j \in \widetilde{T}} p_j(1 - p_j) = p_i(1 - p_i) + \sum_{j \in T} \widetilde{p}_j(1 - \widetilde{p}_j) \le \eta - p_i + p_i(1 - p_i) \le \eta.$$

Since $S$ is good for $P$, we have that

$$\left| \Pr_{X \in_u S_{\widetilde{T}}} (L(X) \geq 0) - \Pr_{X \sim P_{\widetilde{T}}} (L(X) \geq 0) \right| \leq \varepsilon^{3/2}/d^2 \ .$$

Moreover, note that $S_{\widetilde{T}} = \widetilde{S}_T$ and $P_{\widetilde{T}} = \widetilde{P}_T$. Thus, $\widetilde{S}$ is $(\varepsilon, \eta - p_i)$-good for $\widetilde{P}$.

Next, we show that $\Delta(\widetilde{S}, S') + p_i/5 \leq \Delta(S, S')$. We write $S = \widetilde{S} \setminus \widetilde{L} \cup \widetilde{E}$. We write $S_1, L_1, S'_1$ for the subset of $S, L, S'$, respectively, where the $i$th coordinate is 1. Since $S$ is $(\varepsilon, \eta)$-good for $P$, we have that $|\mu_i^S - p_i| \leq \varepsilon^{3/2}/d^2$. Recall that we are already assuming that $\widetilde{p}_i \geq \varepsilon/d$. Thus, $\mu_i^S \geq 29p_i/30$. Therefore, we have that $|S_1| \geq 29p_i|S|/30$. On the other hand, we have that $|S'_1| \leq p_i|S'|/3 \leq 11p_i|S|/30$. Thus, $|L_1| = |S_1 \setminus S'_1| \geq 18p_i|S|/30$. This means that $p_i = O(\Delta(\widetilde{S}, S')) = O(\varepsilon)$. However, $\widetilde{E} = E \cup S'_1$ and $\widetilde{L} = L \setminus L_1$. This gives

$$\begin{aligned}
\Delta(\widetilde{S}, S') = \frac{|\widetilde{E}| + |\widetilde{L}|}{|\widetilde{S}|} &\leq \frac{|E| + |S'_1| + |L| - |L_1|}{|\widetilde{S}|} \\
&\leq \frac{|E| + |L| - 7p_i/30}{|\widetilde{S}|} = \frac{|E| + |L| - 7p_i|S|/30}{|S|(1 - \mu_i^S)} \\
&\leq \frac{\Delta(S, S') - 7p_i/30}{1 - 31p_i/30} = \Delta(S, S') - 7p_i/30 + O(\varepsilon p_i) \\
&\leq \Delta(S, S') - p_i/5 \ .
\end{aligned}$$

Similarly, suppose that instead the $i$th coordinate has $1 - \mu_i^{S'} < (1-p_i)/3$. Let $\widetilde{P}$ be the product whose mean vector $\widetilde{p}$ has $\widetilde{p}_i = 1$ and $\widetilde{p}_j = p_j$ for $j \neq i$. Let $\widetilde{S}$ be obtained by removing from $S$ all of the entries with 0 in the $i$th coordinate. Then, by a similar proof we have that $\widetilde{S}$ is $(\varepsilon, \eta - (1-p_i))$-good for $\widetilde{P}$ and has $\Delta(\widetilde{S}, S') + (1-p_i)/5 \leq \Delta(S, S')$. Note that here we have $\|p - \widetilde{p}\|_1 = 1 - p_i$.

By an easy induction, we can set all coordinates $i$ with $\mu_i^{S'} \geq \widetilde{p}_i/3$ and $1 - \mu_i^{S'} \geq (1 - \widetilde{p}_i/3)$ to 0 or 1, respectively, giving an $\widetilde{S}$ and $\widetilde{P}$ such that $\widetilde{S}$ is $(\varepsilon, \eta - \|p - \widetilde{p}\|_1)$-good for $\widetilde{P}$ and

$$\Delta(\widetilde{S}, S') + \|p - \widetilde{p}\|_1/5 \leq \Delta(S, S') \ ,$$

as desired. ∎

In conclusion, throughout the rest of the proof we may and will assume that for all $i$,

- $\varepsilon/d \leq \mu_i^{S'} \leq 1 - \varepsilon/d$.
- $\mu_i^{S'} \geq p_i/3$ and $1 - \mu_i^{S'} \geq (1 - p_i)/3$.

**7.2.3. Setup and basic structural lemmas.** As in the balanced case, we can write $S' = (S \setminus L) \cup E$ for disjoint multisets $L$ and $E$. Similarly, we define the following matrices:

- $M_P$ to be the matrix with $(i,j)$-entry $\mathbb{E}_{X \sim P}[(X_i - \mu_i^{S'})(X_j - \mu_j^{S'})]$,
- $M_S$ to be the matrix with $(i,j)$-entry $\mathbb{E}_{X \in_u S}[(X_i - \mu_i^{S'})(X_j - \mu_j^{S'})]$,
- $M_E$ to be the matrix with $(i,j)$-entry $\mathbb{E}_{X \in_u E}[(X_i - \mu_i^{S'})(X_j - \mu_j^{S'})]$, and
- $M_L$ to be the matrix with $(i,j)$-entry $\mathbb{E}_{X \in_u L}[(X_i - \mu_i^{S'})(X_j - \mu_j^{S'})]$.

Note that we no longer zero out the diagonals of $M_P$ and $M_S$. This will turn out to allow us to more naturally relate spectral properties of these matrices to the $\chi^2$-distance between the means. We start with the following simple claim.

CLAIM 7.27. *For any $v \in \mathbb{R}^d$ satisfying $\sum_{i=1}^d v_i^2 \mu_i^{S'}(1-\mu_i^{S'}) \le 1$, the following statements hold:*
(i) $\mathrm{Var}_{X\sim P}[v\cdot X] \le 9$ *and* $|v\cdot(p-\mu^{S'})| \le \sqrt{\chi^2(\mu^{S'},p)}$, *and*
(ii) $\mathrm{Pr}_{X\sim P}\left(|v\cdot X - \mu^{S'}| \ge T + \sqrt{\chi^2(\mu^{S'},p)}\right) \le 9/T^2$ .

*Proof.* Recall that $p$ denotes the mean vector of the binary product $P$. To show (i), we use the fact that $X_i \sim \mathrm{Ber}(p_i)$ and the $X_i$'s are independent. This implies that

$$\operatorname*{Var}_{X\sim P}\left[\sum_{i=1}^d v_i X_i\right] = \sum_{i=1}^d v_i^2 \,\mathrm{Var}[X_i] = \sum_{i=1}^d v_i^2 p_i(1-p_i) \le 9\sum_{i=1}^d v_i^2 \mu_i^{S'}(1-\mu_i^{S'}) \le 9 ,$$

where we used that $p_i \le 3\mu_i^{S'}$, $(1-p_i) \le 3(1-\mu_i^{S'})$, and the assumption in the claim statement. For the second part of (i), note that

$$\begin{aligned}|v\cdot(p-\mu^{S'})| &= \left|\sum_{i=1}^d v_i\sqrt{\mu_i^{S'}(1-\mu_i^{S'})}\cdot\frac{p_i-\mu_i^{S'}}{\sqrt{\mu_i^{S'}(1-\mu_i^{S'})}}\right| \\ &\le \sqrt{\sum_{i=1}^d v_i^2\mu_i^{S'}(1-\mu_i^{S'})}\cdot\sqrt{\chi^2(\mu^{S'},p)} \le \sqrt{\chi^2(\mu^{S'},p)} ,\end{aligned}$$

where the first inequality is Cauchy–Schwarz, and the second follows from the assumption in the claim statement that $\sum_{i=1}^d v_i^2\mu_i^{S'}(1-\mu_i^{S'}) \le 1$. This proves (i).

To prove (ii), we note that Chebyshev's inequality gives

$$\operatorname*{Pr}_{X\sim P}(|v\cdot(X-p)| \ge T) \le \operatorname*{Var}_{X\sim P}[v\cdot X]/T^2 \le 9/T^2 ,$$

where the second inequality follows from (i). To complete the proof note the inequality

$$|v\cdot(X-\mu^{S'})| \ge T + \sqrt{\chi^2(\mu^{S'},p)}$$

implies that

$$|v\cdot(X-p)| \ge |v\cdot(X-\mu^{S'})| - |v\cdot(p-\mu^{S'})| \ge T ,$$

where we used the triangle inequality and the second part of (i). ∎

Let Cov$[S]$ denote the sample covariance matrix with respect to $S$, and Cov$[P]$ denote the covariance matrix of $P$. We will need the following lemma.

LEMMA 7.28. *We have the following:*
(i) $\left|\sqrt{\chi^2(\mu^{S'},\mu^S)} - \sqrt{\chi^2(\mu^{S'},p)}\right| \le O(\varepsilon/d)$, *and*
(ii) $\|D\,(Cov[S]-Cov[P])\,D\|_2 \le O(\sqrt{\varepsilon})$ .

*Proof.* For (i): Since $S$ is good, for any $i \in [d]$, we have

$$|\mu_i^S - p_i| = \left|\operatorname*{Pr}_{X\in_u S}(e_i\cdot X \ge 1) - \operatorname*{Pr}_{X\sim P}(e_i\cdot X\ge 1)\right| \le \varepsilon^{3/2}/d^2 .$$

Therefore, by the triangle inequality we get

$$\left|\sqrt{\chi^2(\mu^{S'},\mu^S)} - \sqrt{\chi^2(\mu^{S'},p)}\right| \le \sqrt{\sum_{i=1}^d \frac{(\mu_i^S-p_i)^2}{\mu_i^{S'}(1-\mu_i^{S'})}} \le \sqrt{\frac{d\cdot(\varepsilon^3/d^4)}{\varepsilon/(2d)}} \le O(\varepsilon/d) ,$$

where the second inequality uses the fact that $\mu_i^{S'}(1-\mu_i^{S'}) \geq \varepsilon/(2d)$.

For (ii): Since $S$ is good, for any $i,j \in [d]$, we have

$$\left| \mathop{\mathbb{E}}_{X \in_u S}[X_i X_j - p_i p_j] \right| = \left| \mathop{\Pr}_{X \in_u S}[(e_i + e_j) \cdot X \geq 1] - \mathop{\Pr}_{X \sim P}[(e_i + e_j) \cdot X \geq 1] \right| \leq \varepsilon^{3/2}/d^2 \ .$$

Combined with the bound $|\mu_i^S - p_i| \leq \varepsilon^{3/2}/d^2$ above, this gives

$$|\mathrm{Cov}[S]_{i,j} - \mathrm{Cov}[P]_{i,j}| \leq O(\varepsilon^{3/2}/d^2) \ .$$

We thus obtain

$$\|\mathrm{Cov}[S] - \mathrm{Cov}[P]\|_2 \leq \|(\mathrm{Cov}[S] - \mathrm{Cov}[P])\|_F \leq O(\varepsilon^{3/2}/d) \ .$$

Note that $\|D\|_2 = \max_i \left(1/\sqrt{\mu_i^{S'}(1-\mu_i^{S'})}\right) \leq \sqrt{2d/\varepsilon}$. Therefore,

$$\|D\left(\mathrm{Cov}[S] - \mathrm{Cov}[P]\right)D\|_2 \leq O(\sqrt{\varepsilon}) \ .$$

Combining Claim 7.27 and Lemma 7.28 we obtain the following corollary.

COROLLARY 7.29. *For any $v \in \mathbb{R}^d$ with $\sum_{i=1}^d v_i^2 \mu_i^{S'}(1-\mu_i^{S'}) \leq 1$, we have*
(i) $\mathrm{Var}_{X \in_u S}[v \cdot X] \leq 10$ *and* $|v \cdot (\mu^S - \mu^{S'})| \leq \sqrt{\chi^2(\mu^{S'}, p)} + O(\varepsilon/d)$, *and*
(ii) $\Pr_{X \in_u S}\left(|v \cdot X - \mu^{S'}| \geq T + \sqrt{\chi^2(\mu^{S'}, p)}\right) \leq 9/T^2 + \varepsilon^{3/2}/d^2$ .

*Proof.* We have that

$$\begin{aligned}\left| \mathop{\mathrm{Var}}_{X \in_u S}[v \cdot X] - \mathop{\mathrm{Var}}_{Y \sim P}[v \cdot Y] \right| &= v^T\left(\mathrm{Cov}[S] - \mathrm{Cov}[P]\right)v \\ &\leq \|D^{-1}v\|_2^2 \cdot \|D\left(\mathrm{Cov}[S] - \mathrm{Cov}[P]\right)D\|_2 \\ &\leq O(\sqrt{\varepsilon}) \\ &\leq 1 \ ,\end{aligned}$$

where the second line uses Lemma 7.28 (ii), and the assumption $\|D^{-1}v\|_2^2 = \sum_{i=1}^d v_i^2 \mu_i^{S'}(1-\mu_i^{S'}) \leq 1$, and the third line holds for small enough $\varepsilon$. Thus, using Claim 7.27 (i), we get that

$$\mathop{\mathrm{Var}}_{X \in_u S}[v \cdot X] \leq \mathop{\mathrm{Var}}_{Y \sim P}[v \cdot Y] + 1 \leq 10 \ .$$

By the Cauchy–Schwarz inequality and Lemma 7.28, we get

$$|v \cdot (\mu^S - \mu^{S'})| \leq \sqrt{\chi^2(\mu^{S'}, \mu^S)} \leq \sqrt{\chi^2(\mu^{S'}, p)} + O(\varepsilon/d) \ .$$

This proves (i).

Part (ii) follows directly from Claim 7.27 (ii) and the assumption that $S$ is good for $P$. ∎

LEMMA 7.30. *We have that* $\|D(M_S - M_P)D\|_2 \leq O(\sqrt{\varepsilon})$.

*Proof.* We can show that $|(M_S)_{i,j} - (M_P)_{i,j}| \leq O(\varepsilon^{3/2}/d^2)$ for all $i,j \in [d]$, by expanding the LHS in terms of the differences of linear threshold functions on $S$ and $P$ in the same way as in the proof of Lemma 7.28. Thus,

$$\|M_S - M_P\|_2^2 \leq \|M_S - M_P\|_F^2 \leq \sum_{i,j} |(M_S)_{i,j} - (M_P)_{i,j}|^2 \leq O(\varepsilon^3/d^2) \ .$$

Finally, note that $\|D\|_2 = \max_i \left(1/\sqrt{\mu_i^{S'}(1-\mu_i^{S'})}\right) \leq \sqrt{2d/\varepsilon}$, and so

$$\|D(M_S - M_P)D\|_2 \leq \|D\|_2^2 \|M_S - M_P\|_2 \leq 2d/\varepsilon \cdot O(\varepsilon^{3/2}/d) = O(\sqrt{\varepsilon}) \ .$$

Combining the above we obtain the following corollary.

COROLLARY 7.31. *We have that* $\|D(|S'|M - |S|M_P - |E|M_E + |L|M_L)D\|_2 = O(|S'| \cdot \sqrt{\varepsilon})$ .

*Proof.* The proof follows from Lemma 7.30 combined with the fact that $|S'|M = |S|M_S + |E|M_E - |L|M_L$ and the observation $|S| \le |S'|/(1-2\varepsilon) \le 2|S'|$. ∎

We have the following lemma.

LEMMA 7.32. *We have that* $\|DM_PD\|_2 \le 9 + \chi^2(\mu^{S'}, p)$.

*Proof.* Note that $M_P = (\mu^{S'} - p)(\mu^{S'} - p)^T + \mathrm{Diag}(p_i(1-p_i))$. For any $v'$ with $\|v_2'\| \le 1$, the vector $v = Dv'$ satisfies $\sum_{i=1}^d v_i^2 \mu_i^{S'}(1-\mu_i^{S'}) \le 1$. Therefore, we can write

$$v'^T DM_PDv' = v^T M_P v = (v \cdot (\mu^{S'} - p))^2 + v^T \mathrm{Diag}(p_i(1-p_i))v .$$

Using Claim 7.27 (i), we get

$$(v \cdot (\mu^{S'} - p))^2 \le \chi^2(\mu^{S'}, p)$$

and

$$|v^T \mathrm{Diag}(p_i(1-p_i))v| = |\operatorname*{Var}_{X\sim P}(v \cdot (X-p))| \le 9 .$$

This completes the proof. ∎

The following crucial lemma bounds from above the contribution to the error from $L$.

LEMMA 7.33. *The spectral norm* $\|DM_LD\|_2 = O(|S'|/|L| + \chi^2(\mu^{S'}, p))$.

*Proof.* Similarly, we need to bound from above the quantity $|v'^T DM_LDv'|$ for all $v' \in \mathbb{R}^d$ with $\|v'\|_2 \le 1$. Note that $|v'^T DM_LDv'| = |v^T M_L v| = \mathbb{E}_{X\in_u L}[|v\cdot(X-\mu^{S'})|^2]$, where the vector $v = Dv'$ satisfies $\sum_{i=1}^d v_i^2\mu_i^{S'}(1-\mu_i^{S'}) \le 1$. The latter expectation is bounded from above as follows:

$$\begin{aligned}
\mathop{\mathbb{E}}_{X\in_u L}[(v\cdot(X-\mu^{S'}))^2] &\le 2\mathop{\mathbb{E}}_{X\in_u L}[(v\cdot(X-p))^2] + 2(v\cdot(\mu^{S'}-p))^2 \\
&\le 2\mathop{\mathbb{E}}_{X\in_u L}[(v\cdot(X-p))^2] + 2\chi^2(\mu^{S'}, p) \\
&\le (2|S|/|L|)\cdot \mathop{\mathbb{E}}_{X\in_u S}[(v\cdot(X-p))^2] + 2\chi^2(\mu^{S'}, p) \\
&\le 20|S|/|L| + 2\chi^2(\mu^{S'}, p) \\
&\le 21|S'|/|L| + 2\chi^2(\mu^{S'}, p) ,
\end{aligned}$$

where the first line uses the triangle inequality, the second line uses Claim 7.27 (i), the third line follows from the fact that $L \subseteq S$, the fourth line uses Corollary 7.29 (i), and the last line uses the fact that $\varepsilon$ is small enough. ∎

The above lemmas and the triangle inequality yield the following corollary.

COROLLARY 7.34. *We have that* $\|D\,(M - (|E|/|S'|)M_E)\,D\|_2 = O(1+\chi^2(\mu^{S'}, p))$ .

We are now ready to analyze the two cases of the algorithm FILTER-PRODUCT.

**7.2.4. The case of small spectral norm.** We start by considering the case where the vector $\mu^{S'}$ is returned. It suffices to show that in this case $d_{\mathrm{TV}}(P, P') = O(\sqrt{\varepsilon \log(1/\varepsilon)})$.

Let $N$ be the set of coordinates not in $\mathrm{supp}(S')$. We note that only an $\varepsilon$-fraction of the points in $S$ could have that any coordinate in $N$ does not have its most common value. Therefore, at most a $2\varepsilon$-fraction of samples from $P$ have this property. Hence, the contribution to the variation distance coming from these coordinates is $O(\varepsilon)$. So, it suffices to consider only the coordinates not in $N$ and show that $d_{\mathrm{TV}}(P_{\overline{N}}, P'_{\overline{N}}) = O(\sqrt{\varepsilon\log(1/\varepsilon)})$. Thus, we may assume for the sake of the analysis below that $N = \varnothing$.

We begin by bounding various $\chi^2$-distances by the spectral norm of appropriate matrices.

Lemma 7.35. *Let $\mu^E, \mu^L$ be the mean vector of $E$ and $L$, respectively. Then,* $\chi^2(\mu^{S'}, \mu^E) \leq \|DM_E D\|_2$ *and* $\chi^2(\mu^{S'}, \mu^L) \leq \|DM_L D\|_2$.

*Proof.* We prove the first inequality, the proof of the second being very similar.

Note that for any vector $v$, $v^T M_E v = \mathbb{E}_{X\in_u E}[|v\cdot(X-\mu^{S'})|^2] \geq |v\cdot(\mu^E-\mu^{S'})|^2$. Let $v\in\mathbb{R}^d$ be the vector defined by

$$v_i = \frac{\mu_i^E - \mu_i^{S'}}{\mu_i^{S'}(1-\mu_i^{S'})\sqrt{\chi^2(\mu^{S'},\mu^E)}} \,.$$

We have that

$$\|D^{-1}v\|_2^2 = \sum_{i=1}^d v_i^2 \mu_i^{S'}(1-\mu_i^{S'}) = \frac{1}{\chi^2(\mu^{S'},\mu^E)}\sum_{i=1}^d \frac{(\mu_i^E-\mu_i^{S'})^2}{\mu_i^{S'}(1-\mu_i^{S'})} = 1.$$

Therefore,

$$\|DM_E D\|_2 \geq v^T M_E v \geq |v\cdot(\mu^E-\mu^{S'})|^2 = \chi^2(\mu^{S'},\mu^E)\,.$$

We can now prove that the output in step 7 of Algorithm 13 has the desired guarantee.

Lemma 7.36. *We have that* $\sqrt{\chi^2(\mu^{S'},p)} \leq 2\sqrt{\varepsilon\|DMD\|_2} + O(\sqrt{\varepsilon})$.

*Proof.* Since $S' = (S\setminus L)\cup E$, we have that $|S'|\mu^{S'} = |S|\mu^S + |E|\mu^E - |L|\mu^L$. Recalling that $L, E$ are disjoint, the latter implies that

$$(|S|/|S'|)\sqrt{\chi^2(\mu^{S'},\mu^S)} \leq (|E|/|S'|)\sqrt{\chi^2(\mu^{S'},\mu^E)} + (|L|/|S'|)\sqrt{\chi^2(\mu^{S'},\mu^L)}\,. \tag{39}$$

First note that, by Lemma 7.28, $|\sqrt{\chi^2(\mu^{S'},\mu^S)} - \sqrt{\chi^2(\mu^{S'},p)}| \leq O(\varepsilon/d)$. Lemma 7.35 and Corollary 7.34 give that

$$\begin{aligned}(|E|/|S'|)^2\chi^2(\mu^{S'},\mu^E) &\leq (|E|/|S'|)^2\|DM_E D\|_2 + O(\varepsilon)\\ &\leq (|E|/|S'|)\|DMD\|_2 + O(\varepsilon(1+\chi^2(\mu^{S'},p)))\,.\end{aligned}$$

Thus,

$$(|E|/|S'|)\sqrt{\chi^2(\mu^{S'},\mu^E)} \leq \sqrt{(|E|/|S'|)\|DMD\|_2} + \sqrt{\varepsilon}\cdot O\left(1+\sqrt{\chi^2(\mu^{S'},p)}\right)\,.$$

Lemmas 7.33 and 7.35 give that

$$(|L|/|S'|)^2\chi^2(\mu^{S'},\mu^L) \leq (|L|/|S'|)^2\|DM_L D\|_2 \leq O((|L|/|S'|)^2\chi^2(\mu^{S'},p)+\varepsilon)\,.$$

Thus,

$$(|L|/|S'|)\sqrt{\chi^2(\mu^{S'},\mu^L)} \le O((|L|/|S'|)\sqrt{\chi^2(\mu^{S'},p)}) + O(\sqrt{\varepsilon})\ .$$

Substituting these into (39) yields

$$(|S|/|S'|)\sqrt{\chi^2(\mu^{S'},p)} \le \sqrt{(|E|/|S'|)\|DMD\|_2} + O\left(\sqrt{\varepsilon}\left(1+\sqrt{\chi^2(\mu^{S'},p)}\right)\right)\ .$$

For $\varepsilon$ sufficiently small, we have that the $\sqrt{\chi^2(\mu^{S'},p)}$ terms satisfy

$$(|S|/|S'|) - O(\sqrt{\varepsilon}) \ge 1 - 2\varepsilon - O(\sqrt{\varepsilon}) \ge \frac{1}{\sqrt{2}}.$$

Recalling that $|E|/|S'| \le \Delta(S,S')|S|/|S'| \le (5/2)\varepsilon$, we now have

$$\sqrt{\chi^2(\mu^{S'},p)} \le (5/2)\sqrt{\varepsilon\|DMD\|_2} + O(\sqrt{\varepsilon})\ ,$$

as required. ∎

COROLLARY 7.37. *Let* $\delta := 3\sqrt{\varepsilon|\lambda|}$. *For some universal constant* $C$, *if* $\delta \le C\sqrt{\varepsilon\log(1/\varepsilon)}$, *then* $\sqrt{\chi^2(\mu^{S'},p)} \le O(\sqrt{\varepsilon\log(1/\varepsilon)})$. *Otherwise, we have* $\sqrt{\chi^2(\mu^{S'},p)} \le \delta$.

*Proof.* By Lemma 7.36, we have that

$$\sqrt{\chi^2(\mu^{S'},p)} \le \frac{5}{6}\delta + O(\sqrt{\varepsilon})\ .$$

If $C$ is sufficiently large, when $\delta > C\sqrt{\varepsilon\log(1/\varepsilon)}$, this $O(\sqrt{\varepsilon})$ is at most $C\sqrt{\varepsilon\log(1/\varepsilon)}$ /6. ∎

CLAIM 7.38. *If the algorithm terminates at step* 7 *of Algorithm* 13, *then we have* $d_{\mathrm{TV}}(P,P') \le O(\sqrt{\varepsilon\log(1/\varepsilon)})$, *where* $P'$ *is the product distribution with mean vector* $\mu^{S'}$.

*Proof.* By Corollary 7.37, we have that $\sqrt{\chi^2(\mu^{S'},p)} \le O(\sqrt{\varepsilon\log(1/\varepsilon)})$. Thus, by Fact 7.25, the total variation distance between the product distributions with means $p$ and $\mu^{S'}$ is $O(\sqrt{\varepsilon\log(1/\varepsilon)})$. ∎

**7.2.5. The case of large spectral norm.** We next need to show the correctness of the algorithm if it returns a filter. If we reach this step, then we have $\|DMD\|_2 = \Omega(1)$, indeed $|v'DMDv'^T| = \Omega(1)$, and by Corollary 7.37, it follows that $\sqrt{\chi^2(\mu^{S'},p)} \le \delta$, where $\delta := 3\sqrt{\varepsilon\|DMD\|_2}$.

Since $\|v'\|_2 = 1$, $Dv'$ satisfies $\sum_{i=1}^d (Dv')_i^2 \mu_i^{S'}(1-\mu_i^{S'}) = \sum_{i=1}^m v_i'^2 = 1$. Thus, we can apply Corollary 7.29 to it.

LEMMA 7.39. *We have* $\mathbb{E}_Z[\Delta(S,S'')] \le \Delta(S,S')$.

*Proof.* Let $a = \max_{x\in S'} |v^*\cdot x - \mu^{S'}|$. First, we look at the expected number of samples we reject:

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z}[|S''|] - |S'| &= \mathop{\mathbb{E}}_{Z}\left[|S'| \Pr_{X\in_u S'}[|X-\mu^{S'}| \geq aZ]\right] \\
&= |S'| \int_0^1 \Pr_{X\in_u S'}\left[|v^*\cdot(X-\mu^{S'})| \geq ax\right] 2x dx \\
&= |S'| \int_0^a \Pr_{X\in_u S'}\left[|v^*\cdot(X-\mu^{S'})| \geq T\right] (2T/a) dT \\
&= |S'| \mathop{\mathbb{E}}_{X\in_u S'}\left[(v^*\cdot(X-\mu^{S'}))^2\right]/a \\
&= (|S'|/a)\cdot v^{*T} M v^* = (|S'|/a)\lambda' \ .
\end{aligned}$$

Next, we look at the expected number of false positive samples we reject. If we write $S'' = S \cup L' \setminus E'$ for disjoint multisets $L'$ and $E'$, then these are the elements of $L' \setminus L$. We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z}[|L'|] - |L| &= \mathop{\mathbb{E}}_{Z}\left[(|S|-|L|) \Pr_{X\in_u S\setminus L}\left[|X-\mu^{S'}| \geq T\right]\right] \\
&\leq \mathop{\mathbb{E}}_{Z}\left[|S| \Pr_{X\in_u S}[|v^*\cdot(X-\mu^{S'})| \geq aZ]\right] \\
&= |S| \int_0^1 \Pr_{X\in_u S}[|v^*\cdot(X-\mu^{S'})| \geq ax] 2x\, dx \\
&= |S| \int_0^a \Pr_{X\in_u S}[|v^*\cdot(X-\mu^{S'})| \geq T](2T/a)\, dT \\
&\leq |S| \int_0^\infty \Pr_{X\in_u S}[|v^*\cdot(X-\mu^{S'})| \geq T](2T/a)\, dT \\
&= |S| \mathop{\mathbb{E}}_{X\in_u S}\left[(v^*\cdot(X-\mu^{S'}))^2\right]/a \\
&= (|S'|/a)\cdot v^{*T} M_S v^* = (|S'|/a)\cdot v'^T D M_S D v' \\
&\leq (|S'|/a)\cdot \|D M_S D\|_2 \\
&\leq (|S'|/a)\cdot \|D M_P D\|_2 + (|S'|/a)\cdot \|D(M_P - M_S)D\|_2 \\
&\leq (|S'|/a)\cdot (\sqrt{\varepsilon} + 9 + \chi^2(\mu^{S'}, p)) \\
&\leq (|S'|/a)\cdot O(1+\delta^2) \leq (|S'|/a)\cdot O(1+\varepsilon\lambda')\ ,
\end{aligned}$$

where the penultimate line uses Lemmas 7.30 and 7.32. When $\lambda'$ is at least a sufficiently large constant, $\lambda'$ is bigger than $2\cdot O(1+\varepsilon\lambda')$, and so $\mathbb{E}_Z[S''] - S' \geq 2(\mathbb{E}_Z[L'] - L)$. Now consider that $|S''| = |S| + |E'| - |L'| = |S'| - |E| + |E'| + |L| - |L'|$, and thus $|S''| - |S'| = |E| - |E'| + |L'| - |L|$. This yields that $|E| - \mathbb{E}_Z[|E'|] \geq (\mathbb{E}_Z[L'] - L)$, which can be rearranged to $\mathbb{E}_Z[|E'| + |L'|] \leq |E| + |L|$ or in other terms $\mathbb{E}_Z[\Delta(S, S'')] \leq \Delta(S, S')$. ∎

**8. Agnostically learning mixtures of two balanced binary products, via filters.** In this section, we study the problem of agnostically learning a mixture of two balanced binary product distributions. Let $p$ and $q$ be the coordinatewise means of the two product distributions. Let $u = \frac{p}{2} - \frac{q}{2}$. Then, when there is no noise, the empirical covariance matrix is $\Sigma = uu^T + D$, where $D$ is a diagonal matrix whose entries are $\frac{p_i+q_i}{2} - \frac{(p_i-q_i)^2}{4}$. Thus, it can already have a large eigenvalue. Now in the presence of corruptions it turns out that we can construct a filter when the *second* absolute eigenvalue is also large. When it is the case that only the top absolute eigenvalue is large, we know that both $p$ and $q$ are close to one-dimensional affine subspace (a.k.a. line) $\{\mu + cv : c \in \mathbb{R}\}$, where $\mu$ is the empirical mean and $v$ is the top eigenvector. And by performing a grid search over $c$, we will find a good candidate hypothesis.

Unfortunately, bounds on the top absolute eigenvalue do not translate as well into bounds on the total variation distance of our estimate to the true distribution, as they did in all previous cases (e.g., if the top absolute eigenvalue is small in the case of learning the mean of a Gaussian with identity covariance, we can just use the empirical mean, etc.). In fact, an eigenvalue $\lambda$ could just mean that $p$ and $q$ differ by $\sqrt{\lambda}$ along the direction $v$. However, we can proceed by zeroing out the diagonals. If $uu^T$ has any large value along the diagonal, this operation can itself produce large eigenvalues. So, this strategy only works when $\|u\|_\infty$ is appropriately bounded. When $\|u\|_\infty$ is large, there is a separate strategy to deal with large entries in $u$ by guessing a coordinate whose value is large and conditioning on it, and once again setting up a modified eigenvalue problem. Our overall algorithm then follows from balancing all of these different cases, and we describe the technical components in more detail in the next subsection.

**8.1. The full algorithm.** This section is devoted to the proof of the following theorem.

THEOREM 8.1. *Let* $\Pi$ *be a mixture of two c-balanced binary product distributions in d dimensions. Given* $\varepsilon > 0$ *and* $\mathrm{poly}(d, 1/\varepsilon)\log(1/\tau)$ *independent samples from* $\Pi$*, an* $\varepsilon$*-fraction of which have been arbitrarily corrupted, there is a polynomial-time algorithm that, with probability at least* $1-\tau$*, outputs a mixture of two binary product distributions* $\Pi'$ *such that* $d_{\mathrm{TV}}(\Pi, \Pi') = O(\varepsilon^{1/6}/\sqrt{c})$.

Recall that our overall approach is based on two strategies that succeed under different assumptions. Our first algorithm (section 8.2) assumes that there exists a coordinate in which the means of the two component product distributions differ by a substantial amount. Under this assumption, we can use the empirical mean vectors conditioned on this coordinate being 0 and 1. We show that the difference between these conditional mean vectors is almost parallel to the difference between the mean vectors of the product distributions. Considering eigenvectors perpendicular to this difference will prove a critical part of the analysis of this case. Our second algorithm (section 8.3) succeeds under the assumption that the mean vectors of the two product distributions are close in all coordinates. This assumption allows us to zero out the diagonal of the covariance matrix without introducing too much error.

Both of these algorithms give an iterative procedure that produces filters which improve the sample set until they produce an output. We note that these algorithms essentially only produce a line in $\mathbb{R}^d$ such that both mean vectors of the target product distributions are guaranteed to be close to this line in $\ell_2$-distance. The assumption that our product distributions are balanced implies that $\Pi$ is close in variation distance to some mixture of two products whose mean vectors lie exactly on the given line. Given this line, we can exhaustively compare $\Pi$ to a polynomial number of such mixtures and run a tournament to find one that is sufficiently close.

We note that together these algorithms will cover all possible cases. Our final algorithm runs all of these procedures in parallel, obtaining a polynomial number of candidate hypothesis distributions, such that at least one is sufficiently close to $\Pi$.

We then run the tournament described by Lemma 2.23 in order to find a particular candidate that is sufficiently close to the target. To ensure that all of the distributions returned are in some finite set $\mathcal{M}$, we round each of the probabilities of each of the products to the nearest multiple of $\varepsilon/d$, and similarly round the mixing weight to the nearest multiple of $\varepsilon$. This introduces at most $O(\varepsilon)$ additional error.

**Algorithm 14** Filter algorithm for agnostically learning a mixture of two balanced products.

```
1: procedure LearnProductMixture(ε, τ, S')
input: a set of poly(d, 1/ε) log(1/τ) samples of which an ε-fraction have been corrupted
output: a mixture of two balanced binary products that is O(ε^{1/6})-close to the target
2:    Run the procedure Filter-Balanced-Product(2ε^{1/6}, S'_1) for up to d + 1
      iterations on a set S'_1 of corrupted samples of size Θ(d^4 log(1/τ)/ε^{1/3}).
3:    for each 1 ≤ i* ≤ d do
4:        Run the procedure Filter-Product-Mixture-Anchor(i*, ε, S'_{2,i*}) for
      up to d + 1 iterations on a set S'_{2,i*} of corrupted samples of size
      Θ(d^4 log(1/τ)/ε^{13/6}).
5:    Run the procedure Filter-Product-Mixture-Close(ε, S'_3, δ := ε^{1/6}) for
      up to d+1 iterations on a set S'_3 of corrupted samples of size Θ(d^4 log(1/τ)/ε^{13/6}).
6:    Run a tournament among all mixtures output by any of the previous steps.
      Output the winner.
```

**8.2. Mixtures of products whose means differ significantly in one coordinate.** We will use the following notation. Let $\Pi$ be a mixture of two $c$-balanced binary product distributions. We will write $\Pi$ as $\alpha P+(1-\alpha)Q$, where $P, Q$ are binary product distributions with mean vectors $p, q$, and $\alpha \in [0,1]$. In this subsection, we prove the following theorem.

THEOREM 8.2. *Let $\Pi = \alpha P + (1-\alpha)Q$ be a mixture of two $c$-balanced binary product distributions in $d$ dimensions, with $\varepsilon^{1/6} \le \alpha \le 1-\varepsilon^{1/6}$, such that there exists $1 \le i^* \le d$ with $p_{i^*} \ge q_{i^*} + \varepsilon^{1/6}$. There is an algorithm that, given $i^*$, $\varepsilon > 0$, and $\Theta(d^4 \log(1/\tau)/\varepsilon^3)$ independent samples from $\Pi$, an $\varepsilon$-fraction of which have been arbitrarily corrupted, runs in polynomial time and, with probability at least $1-\tau$, outputs a set $R$ of candidate hypotheses such that there exists $\Pi' \in R$ satisfying $d_{\mathrm{TV}}(\Pi, \Pi') = O(\varepsilon^{1/6}/\sqrt{c})$.*

For simplicity of analysis, we will assume without loss of generality that $i^* = d$, unless otherwise specified. First, we determine some conditions under which our sample set will be sufficient. We start by recalling our condition of a good set for a balanced binary product distribution.

DEFINITION 8.3. *Let $P$ be a binary product distribution in $d$ dimensions, and let $\varepsilon > 0$. We say that a multiset $S$ of elements of $\{0,1\}^d$ is* $\varepsilon$-good *with respect to $P$ if* *for every affine function* $L : \mathbb{R}^d \to \mathbb{R}$ *it holds that*

$$\left| \Pr_{X \in_u S}(L(X) > 0) - \Pr_{X \sim P}(L(X) > 0) \right| \leq \varepsilon/d.$$

We will also need this to hold after conditioning on the last coordinate.

DEFINITION 8.4. *Let* $P$ *be a binary product distribution in* $d$ *dimensions, and let* $\varepsilon > 0$. *We say that a multiset* $S$ *of elements of* $\{0,1\}^d$ *is* $(\varepsilon, i)$-good *with respect to* $P$ *if* $S$ *is* $\varepsilon$*-good with respect to* $P$, *and* $S^j \stackrel{\text{def}}{=} \{x \in S : x_i = j\}$ *is* $\varepsilon$*-good for the restriction of* $P$ *to* $x_i = j$, *for* $j \in \{0, 1\}$.

Finally, we define the notion of a good set for a mixture of two balanced products.

DEFINITION 8.5. *Let* $\Pi = \alpha P + (1 - \alpha)Q$ *be a mixture of two binary product distributions. We say that a multiset* $S$ *of elements of* $\{0,1\}^d$ *is* $(\varepsilon, i)$-good *with respect to* $\Pi$ *if we can write* $S = S_P \cup S_Q$, *where* $S_P$ *is* $(\varepsilon, i)$*-good with respect to* $P$, $S_Q$ *is* $(\varepsilon, i)$*-good with respect to* $Q$, *and* $|\frac{|S_P|}{|S|} - \alpha| \leq \varepsilon/d^2$.

We now show that taking random samples from $\Pi$ produces such a set with high probability.

LEMMA 8.6. *Let* $\Pi = \alpha P + (1 - \alpha)Q$ *be a mixture of binary product distributions, where* $P, Q$ *are binary product distributions with mean vectors* $p, q$. *Let* $S$ *be a set obtained by taking* $\Omega(d^4 \log(1/\tau)/\varepsilon^{13/6})$ *independent samples from* $\Pi$. *Then, with probability at least* $1 - \tau$, $S$ *is* $(\varepsilon, i)$*-good with respect to* $\Pi$ *for all* $i \in [d]$.

The proof of this lemma is deferred to section E.

We claim that given a good set with an $\varepsilon$-fraction of its entries corrupted, we can still determine $\Pi$ from it. In particular, this is achieved by iterating the following proposition.

PROPOSITION 8.7. *Let* $\Pi = \alpha P + (1 - \alpha)Q$ *be a mixture of two* $c$*-balanced binary products, with* $p_d \geq q_d + \varepsilon^{1/6}$ *and* $\varepsilon^{1/6} < \alpha < 1 - \varepsilon^{1/6}$. *Let* $S$ *be an* $(\varepsilon, d)$*-good multiset for* $\Pi$, *and let* $S'$ *be any multiset with* $\Delta(S, S') \leq 2\varepsilon$. *There exists an algorithm which, given* $S'$ *and* $\varepsilon > 0$, *runs in polynomial time and returns either a multiset* $S''$ *with* $\Delta(S, S'') \leq \Delta(S, S') - 2\varepsilon/d$, *or returns a list of mixtures of two binary products* $\mathcal{S}$ *such that there exists a* $\Pi' \in \mathcal{S}$ *with* $d_{\mathrm{TV}}(\Pi, \Pi') = O(\varepsilon^{1/6}/\sqrt{c})$.

We note that iteratively applying this algorithm until it outputs a set $R$ of mixtures gives Theorem 8.2.

**Notation.** All vectors in this section should be assumed to be over the first $d-1$ coordinates only. We will write $p_{-d}$ and $q_{-d}$ for the first $d - 1$ coordinates of $p$ and $q$, but for other vectors we will use notation similar to that used elsewhere to denote $(d - 1)$-dimensional vectors.

The algorithm, written in terms of $i^*$ instead of $d$ for generality, is as follows.

**Algorithm 15** Filter algorithm for a mixture of two binary products whose means differ significantly in some coordinate.

**procedure** FILTER-PRODUCT-MIXTURE-ANCHOR($i^*, \varepsilon, S'$)
Let $\mu$ be the sample mean of $S'$ without the $i^*$ coordinate. Let $\Sigma$ be the sample covariance matrix of $S'$ without the $i^*$ row and column.
Let $S'_0$ and $S'_1$ be the subsets of $S'$ with a 0 or 1 in their $i^*$ coordinates, respectively.
Let $\mu^{(j)}$ be the sample mean of $S'_j$ without the $i^*$ coordinate.
Let $u = \mu^{(1)} - \mu^{(0)}$. Compute the unit vector $v^* \in \mathbb{R}^{d-1}$ with $v^* \cdot u = 0$ that maximizes $v^T \Sigma v$ and let $\lambda = v^{*T} \Sigma v^*$.
*/* Note that $v^*$ is the unit vector maximizing the quadratic form $v^T\Sigma v$ over the subspace $u \cdot v = 0$, and thus can be approximated using standard eigenvalue computations.*/*
**if** $\lambda \le \gamma$ **then**
*/* $\gamma$ is some absolute constant to be determined in the course of the analysis*/*
Let $L$ be the set of points $\mu + i(\varepsilon^{1/6}/\|u\|_2)u$ truncated to be in $[c, 1-c]^d$ for $i \in \mathbb{Z}$ with $|i| \le 1 + \sqrt{d}/\varepsilon^{1/6}$.
**return** the set of distributions $\Pi' = \alpha' P' + (1-\alpha')Q'$ with the means of $P'$ and $Q'$, $p', q'$ with $p'_{-i^*}, q'_{-i^*} \in L$ and $p'_{i^*}, q'_{i^*} \in [c, 1-c], \alpha' \in [0,1]$, multiples of $\varepsilon^{1/6}$.
Let $\delta = C(\varepsilon^{1/6}\sqrt{\lambda} + \varepsilon^{2/3}\log(1/\varepsilon))$ for a sufficiently large constant $C$.
Find a real number $T > 0$ such that

$$\Pr_{X \in_u S'} \left(|v^* \cdot (X_{-i^*} - \mu)| > T + \delta\right) > 8\exp(-T^2/2) + 8\varepsilon/d\,.$$

**return** the set $S'' = \{x \in S' : |v \cdot (x_{-i^*} - \mu)| \le T + \delta\}$.

We now proceed to prove correctness. We note that given $S = S_P \cup S_Q$, we can write

$$S' = S'_P \cup S'_Q \cup E,$$

where $S'_P \subseteq S_P$, $S'_Q \subseteq S_Q$, and $E$ is disjoint from $S_P \setminus S'_P$ and $S_Q \setminus S'_Q$. Thus, we have

$$\Delta(S, S') = \frac{|S_P \setminus S'_P| + |S_Q \setminus S'_Q| + |E|}{|S|}\,.$$

We use the notation $\mu^{S_P}, \mu^{S'_P}, \mu^E \in \mathbb{R}^{d-1}$ etc., for the means of $S_P$, $S'_P$, $E$, etc., excluding the last coordinate.

We next need some basic lemmas relating the means of some of these distributions.

LEMMA 8.8. *Let $P$ be a binary product distribution with mean vector $p$. Let $S$ be an $\varepsilon$-good multiset for $P$ in the sense of Definition* 8.3. *Let $\tilde{S}$ be a subset of $S$ with $|S| - |\tilde{S}| = O(\varepsilon|S|)$. Let $\mu^{\tilde{S}}$ be the mean of $\tilde{S}$. Then, $\|p - \mu^{\tilde{S}}\|_2 \le O(\varepsilon\sqrt{\log(1/\varepsilon)})$.*

*Proof.* Since $S$ is $\varepsilon$-good, $\|\mu^S - p\|_2 \le \varepsilon/\sqrt{d}$. Let $L = S \setminus \tilde{S}$. We can apply appropriate lemmas from section 7.1. Note that Lemma 7.12 and Claim 7.14 only depend on $\mu^{S'}$ as far as it appears in the definition of $M_L$, and we may treat it as a parameter that we will set to $p$. By Lemma 7.12 with $\mu^{S'} := p$, we have $\|\mathbb{E}_{X \in_u L}[(X - p)(X-p)^T]\|_2 \le O\left(\log(|S|/|L|) + \varepsilon|S|/|L|\right)$. By Claim 7.14 again with $\mu^{S'} := p$, it follows that $(|L|/|S|)\|\mu^L - p\|_2 \le O(\varepsilon\sqrt{\log(1/\varepsilon)})$. Since $|S|\mu^S = |\tilde{S}|\mu^{\tilde{S}} + |L|\mu^L$, we have $\mu^S - \mu^{\tilde{S}} = -(|L|/|\tilde{S}|)(\mu^L - \mu^S)$ and so

$$\begin{aligned}\|\mu^S - \mu^{\tilde{S}}\|_2 &\le (|L|/|\tilde{S}|)\|\mu^L - \mu^S\|_2 \\ &\le O(\varepsilon^2/\sqrt{d}) + O(1+\varepsilon)(|L|/|S|)\|\mu^L - p\|_2 \le O(\varepsilon\sqrt{\log(1/\varepsilon)}).\end{aligned}$$

By the triangle inequality, $\|p - \mu^{\tilde{S}}\|_2 \le \varepsilon/\sqrt{d} + O(\varepsilon\sqrt{\log(1/\varepsilon)}) = O(\varepsilon\sqrt{\log(1/\varepsilon)})$. ☐

We next show that $\mu^{(1)} - \mu^{(0)}$ is approximately parallel to $p_{-d} - q_{-d}$. Note that if we had $S = S'$ and $\mu^{S_P} = p_{-d}, \mu^{S_Q} = q_{-d}$, then $\mu^{(1)} - \mu^{(0)}$ would be a multiple of $p_{-d} - q_{-d}$. Since $S$ is $\varepsilon$-good, we can bound the error introduced by $\mu^{S_P} - p$, $\mu^{S_Q} - q_{-d}$, and Lemma 8.8 allows us to bound the error in taking $\mu^{S'_P}, \mu^{S'_Q}$ instead of $p_{-d}, q_{-d}$. However, we still have terms in the conditional means of $E$.

LEMMA 8.9. *For some scalars $a = O(\varepsilon)$, $b^0 = O(|E^0|/|S'|)$, $b^1 = O(|E^1|/|S'|)$, we have*

$$\begin{aligned}\|(1-\mu_d)\mu_d u - (\alpha(1-\alpha)(p_d - q_d) + a)(p_{-d} - q_{-d}) - b^0(\mu^{E^0} - \mu) - b^1(\mu^{E^1} - \mu)\|_2 \\ \le O(\varepsilon\log(1/\varepsilon))\ ,\end{aligned}$$

*where $E^j$ is the subset of $E$ with last entry $j$, and $\mu^{E^j}$ is the mean of $E^j$ with dth coordinate removed.*

*Proof.* Let $S_P'^j, S_Q'^j, E^j, S'^j$ denote the subset of the appropriate set in which the last coordinate is $j$. Let $\mu^{S_P'^j}, \mu^{S_Q'^j}, \mu^{E^j}$ denote the means of $S_P'^j, S_Q'^j$, and $E^j$ with the last entry truncated, respectively.

We note that

$$S'^j = S'_P \cup S'_Q \cup E^j.$$

Taking the means of the subsets of $S'^j$, we find that

$$|S'^j|\mu^{(j)} = |S_{\tilde{P}^j}|\mu^{S_P'^j} + |S_Q'^j|\mu^{S_Q'^j} + |E^j|\mu^{E^j}.$$

Therefore, using this and Lemma 8.8, we have that

$$|S'^j|\mu^{(j)} = |S_P'^j|p_{-d} + |S_Q'^j|q_{-d} + |E^j|\mu^{E^j} + O(\varepsilon\log(1/\varepsilon)|S^j|),$$

where $O(\varepsilon)$ denotes a vector of $\ell_2$-norm $O(\varepsilon)$.

Thus, we have

$$\begin{aligned}|S'^0||S'^1|(\mu^{(1)} - \mu^{(0)}) &= (|S'^0||S_P'^1| - |S'^1||S_P'^0|)p_{-d} \\ &+ (|S'^0||S_Q'^1| - |S'^1||S_Q'^0|)q_{-d} \\ &+ |E^1||S'^0|\mu^{E^1} - |E^0||S'^1|\mu^{E^0} \\ &+ O(\varepsilon\log(1/\varepsilon)(|S^1||S'^0| + |S^0||S'^1|))\ . \end{aligned} \tag{40}$$

Since $|S'^j| = |S_P'^j| + |S_Q'^j| + |E^j|$, we have

$$\begin{aligned}0 = |S'^0||S'^1| - |S'^1||S'^0| &= (|S'^0||S_P'^1| - |S'^1||S_P'^0|) \\ &+ (|S'^0||S_Q'^1| - |S'^1||S_Q'^0|) + |E^1||S'^0| - |E^0||S'^1|\ .\end{aligned}$$

Thus, the sum of the coefficients of the $p_{-d}$ and $q_{-d}$ terms in (40) is $|E^0||S'^1| - |E^1||S'^0|$, which is bounded in absolute value by $|E||S'| \leq O(\varepsilon|S|^2)$. Meanwhile, the $p_{-d}$ coefficient of (40) has

$$\begin{aligned}
& |S'^0||S'^1_P| - |S'^1||S'^0_P| \\
= \ & |S'^0||S^1_P| - |S'^1||S^0_P| + O(\varepsilon|S'|^2) = |S'^0||S|\alpha p_d - |S'^1||S|\alpha(1-p_d) + O(\varepsilon|S'|^2) \\
= \ & |S^0||S|\alpha p_d - |S^1||S|\alpha(1-p_d) + O(\varepsilon|S'|^2) \\
= \ & ((\alpha(1-p_d) + (1-\alpha)(1-q_d))\alpha p_d - (\alpha p_d + (1-\alpha)q_d)\alpha(1-p_d) + O(\varepsilon))|S'|^2 \\
= \ & (\alpha(1-\alpha)(1-q_d)p_d - \alpha(1-\alpha)q_d(1-p_d) + O(\varepsilon))|S'|^2 \\
= \ & (\alpha(1-\alpha)(p_d - q_d) + O(\varepsilon))|S'|^2 \ .
\end{aligned}$$

Noting that $(|E^1||S'^0| - |E^0||S'^1|)\alpha = O(\varepsilon|S'|^2)$ and $(|E^1||S'^0| - |E^0||S'^1|)(1-\alpha) = O(\varepsilon|S'|^2)$, we can write (40) as

$$\begin{aligned}
|S'^0||S'^1|(\mu^{(1)} - \mu^{(0)}) = \ & (\alpha(1-\alpha)(p_d - q_d) + O(\varepsilon))|S'|^2(p_{-d} - q_{-d}) \\
& + (|E^1||S'^0| - |E^0||S'^1|)(\alpha p_{-d} + (1-\alpha)q_{-d}) \\
& + |E^1||S'^0|\mu^{E^1} - |E^0||S'^1|\mu^{E^0} + O(\varepsilon\log(1/\varepsilon)|S'|^2) \ .
\end{aligned}$$

We write $\mu^{\Pi} = \alpha p_{-d} + (1-\alpha)q_{-d}$ and so, dividing by $|S'|^2$ and recalling that $|E|/|S'| \leq O(\varepsilon)$, we get

$$\begin{aligned}
\mu_d(1-\mu_d)(\mu^{(1)} - \mu^{(0)}) = \ & (\alpha(1-\alpha)(p_d - q_d) + O(\varepsilon))(p_{-d} - q_{-d}) \\
& + O(|E^1|/|S'|)(\mu^{E^1} - \mu^{\Pi}) \\
& + O(|E^0|/|S'|)(\mu^{E^0} - \mu^{\Pi}) + O(\varepsilon\log(1/\varepsilon)) \ .
\end{aligned} \tag{41}$$

If $\mu^{\Pi} = \mu$, then we would be done. So, we must bound the error introduced by making this substitution. We can express $\mu$ as

$$\begin{aligned}
|S'|\mu &= |S'_P|\mu^{S'_P} + |S'_Q|\mu^{S'_Q} + |E|\mu^E \\
&= |S|\mu^{\Pi} + O(\varepsilon|S|)(p_{-d} - q_{-d}) + O(\varepsilon\log(1/\varepsilon)|S'|) + |E^1|\mu^{E^1} + |E^0|\mu^{E^0} \ ,
\end{aligned}$$

and so

$$|S|(\mu^{\Pi} - \mu) = O(\varepsilon|S|)(p_{-d} - q_{-d}) + O(\varepsilon\log(1/\varepsilon)|S|) + |E^1|(\mu^{E^1} - \mu) + |E^0|(\mu^{E^0} - \mu) \ .$$

Thus, we have

$$\mu^{\Pi} = \mu + O(\varepsilon)(p_{-d} - q_{-d}) + O(\varepsilon\log(1/\varepsilon)) + O(|E^1|/|S'|)(\mu^{E^1} - \mu) + O(|E^0|/|S'|)(\mu^{E^0} - \mu) \ .$$

Substituting this into (41) gives the lemma. ∎

We now show that, for any vector $v$ perpendicular to $u$, if the variance of $S'$ in the $v$-direction is small, then $v \cdot p_{-d}$ and $v \cdot q_{-d}$ are both approximately $v \cdot \mu$.

LEMMA 8.10. *For any $v$ with $\|v\|_2 = 1$, $v \cdot u = 0$, we have that $|v \cdot (p_{-d} - \mu)| \leq \delta$ and $|v \cdot (q_{-d} - \mu)| \leq \delta$ for $\delta := C(\varepsilon^{1/6}\|\Sigma\|_2 + \varepsilon^{2/3}\log(1/\varepsilon))$ for a sufficiently large constant $C$ as defined in the algorithm.*

*Proof.* We begin by noting that

$$\begin{aligned} v^T \Sigma v &= \operatorname*{Var}_{X \in_u S'}(v \cdot X) = \mathop{\mathbb{E}}_{X \in_u S'}[|v \cdot (X - \mu)|^2] \\ &\geq (|E^j|/|S'|) \mathop{\mathbb{E}}_{X \in_u E^j}[|v \cdot (X - \mu)|^2] \\ &\geq (|E^j|/|S'|)|v \cdot (\mu^{E^j} - \mu)|^2 \ . \end{aligned}$$

Next, since $v \cdot u = 0$, we have by Lemma 8.9 that

$$\begin{aligned} & |v \cdot (p_{-d} - q_{-d})| \\ \leq \ & \frac{1}{\alpha(1-\alpha)(p_d - q_d)} \cdot \Big( O(|E^0|/|S'|)v \\ & \cdot (\mu^{E^0} - \mu) + O(|E^1|/|S'|)v \cdot (\mu^{E^1} - \mu) + O(\varepsilon \log(1/\varepsilon))\|v\|_2 \Big) \\ = \ & O\left( \frac{1}{\alpha(1-\alpha)(p_d - q_d)} \right) \left( \sqrt{\varepsilon(v^T \Sigma v)} + \varepsilon \log(1/\varepsilon) \right). \end{aligned}$$

However, we have that $|S'|\mu = |S'_P|\mu^{S'_p} + |S'_Q|\mu^{S'_q} + |E|\mu^E + |S'|O(\varepsilon \log(1/\varepsilon))$, and so

$$(|S'| - |E|)(\mu - \mu^{S'_p}) = |S'_Q|(\mu^{S'_Q} - \mu^{S'_P}) + |E|(\mu^E - \mu) + |S'|O(\varepsilon \log(1/\varepsilon)) \ .$$

Now, we have

$$\mu - p_{-d} = (1 - \alpha + O(\varepsilon))(q_{-d} - p_{-d}) + O(|E|/|S'|)(\mu^E - \mu) + O(\varepsilon \log(1/\varepsilon)) \ .$$

Thus,

$$|v \cdot (p_{-d} - \mu)| = O(v \cdot (p_{-d} - q_{-d})) + O(|E|/|S'|)(v \cdot (\mu^E - \mu) - v \cdot (\mu - p_{-d})) + O(\varepsilon \log(1/\varepsilon)).$$

Therefore,

$$|v \cdot (p_{-d} - \mu)| = O\left( \frac{1}{\alpha(1-\alpha)(p_d - q_d)} \right) \left( \sqrt{\varepsilon(v^T \Sigma v)} + \varepsilon \log(1/\varepsilon) \right) \ .$$

Inserting our assumptions that $\alpha(1-\alpha) \geq \varepsilon^{1/6}/2$ and $p_d - q_d \geq \varepsilon^{1/6}$ gives

$$|v \cdot (p_{-d} - \mu)| = O(\varepsilon^{1/6} \sqrt{\|\Sigma\|_2} + \varepsilon^{2/3} \log(1/\varepsilon)) \leq \delta \ ,$$

when $C$ is sufficiently large.

The other claim follows symmetrically. ∎

We can now show that if we return $R$, some distribution returned is close to $\Pi$. First, we show that there are points on $L$ close to $p_{-d}$ and $q_{-d}$.

LEMMA 8.11. *There are $c, d \in \mathbb{R}$ such that $\tilde{p} = \mu + cu$ and $\tilde{q} = \mu + du$ have $\|\tilde{p} - p_{-d}\|_2, \|\tilde{q} - q_{-d}\|_2 \leq \delta$.*

*Proof.* If we take the $c$ that minimizes $\|\tilde{p} - p_{-d}\|_2$, then $u \cdot (\tilde{p} - p_{-d}) = 0$. Thus, we can apply Lemma 8.10, giving that $|(\tilde{p} - p_{-d}) \cdot (p_{-d} - \mu)| \leq \|\tilde{p} - p_{-d}\|_2 \delta$.

However, $\tilde{p} - \mu = cu$ so we have $(\tilde{p} - p_{-d}) \cdot (\tilde{p} - \mu) = 0$ and thus

$$\|\tilde{p} - p\|_2^2 = |(\tilde{p} - p_{-d}) \cdot (p_{-d} - \mu)| \leq \|\tilde{p} - p_{-d}\|_2 \delta.$$

Therefore, $\|\tilde{p} - p_{-d}\|_2 \leq \delta$. ∎

It is clear that even discretizing $c$ and $d$, we can still find such a pair that satisfies this condition.

LEMMA 8.12. *There are* $p', q' \in L$ *such that* $\|p_{-d} - p'\|_2, \|q_{-d} - q'\|_2 \le \delta + O(\varepsilon^{1/6})$ .

*Proof.* By Lemma 8.11, there exist points $\tilde{p} = \mu + (a/\|u\|_2)u$ and $\tilde{q} = \mu + (b/\|u\|_2)u$ with $a, b \in \mathbb{R}$ that have $\|\tilde{p} - p_{-d}\|_2, \|\tilde{q} - q_{-d}\|_2 \le \delta$.

Letting $i\varepsilon^{1/6}$ be the nearest integer multiple of $\varepsilon^{1/6}$ to $a$, we have that $p' := \mu + i(\varepsilon^{1/6}/\|u\|_2)u$ has $\|p_{-d} - p'\|_2 \le \|\tilde{p}_{-d} - p\|_2 + \|p' - \tilde{p}\|_2 \le \delta + \varepsilon^{1/6}$.

Note that we have $\|p_{-d} - \tilde{p}\|_2 \le \|p_{-d} - \mu\|_2 \le \sqrt{d}\|p_{-d} - \mu\|_\infty \le \sqrt{d}$, which implies that $a \le \sqrt{d}$. Thus, $|i| \le 1 + \sqrt{d}/\varepsilon^{1/6}$. If $p' \notin [c, 1-c]$, then replacing any coordinates less than $c$ with $c$ and more than $1-c$ with $1-c$ can only decrease the distance to $p$ since $p \in [c, 1-c]^d$. Thus, there is a point $p' \in L$ with $\|p_{-d} - p'\|_2 \le \delta + O(\varepsilon^{1/6})$.

Similarly, we show that there is a $q' \in L$ such that $\|q - q'\|_2 \le \delta + O(\varepsilon^{1/6})$. ∎

COROLLARY 8.13. *If the algorithm terminates at step* 8 *of Algorithm* 15, *then there is a* $\Pi' \in R$ *with* $d_{\mathrm{TV}}(\Pi', \Pi) = O(\varepsilon^{1/6}/\sqrt{c})$.

*Proof.* By Lemma 8.12, there exists $\tilde{p}$, $\tilde{q} \in L$ such that $\|p_{-d} - \tilde{p}\|_2, \|q_{-d} - \tilde{q}\|_2 \le \delta + O(\varepsilon^{1/6})$. But now there is a distribution $\Pi' \in R$, where $\Pi' = \alpha' P' + (1-\alpha')Q'$ for binary products $P'$ and $Q'$, whose mean vectors are $p', q'$ and with $|\alpha' - \alpha| \le \varepsilon^{1/6}$, $\|p'_{-d} - p_{-d}\|_2, \|q'_{-d} - q_{-d}\|_2 \le O(\varepsilon^{1/6})$ and $|p'_d - p_d|, |q'_d - q_d| = O(\varepsilon^{1/6})$. Note that this implies that $\|p' - p\|_2, \|q' - q\|_2 = O(\varepsilon^{1/6})$.

Since $P$ and $Q$ are $c$-balanced, we have $d_{\mathrm{TV}}(P, P') \le O(\|p - p'\|_2/\sqrt{c}) \le O(\varepsilon^{1/6}/\sqrt{c})$ and

$$d_{\mathrm{TV}}(Q, Q') \le O(\|q - q'\|_2/\sqrt{c}) \le O(\varepsilon^{1/6}/\sqrt{c}).$$

Thus, $d_{\mathrm{TV}}(\Pi', \Pi) \le \delta + O(\varepsilon^{1/6}/\sqrt{c})$. Since we terminated in step 8 of Algorithm 15, $\lambda \le O(1)$, and so $\delta = C(\varepsilon^{1/6}\sqrt{\lambda} + \varepsilon^{2/3}\log(1/\varepsilon)) = O(\varepsilon^{1/6})$. ∎

Now, we are ready to analyze the second part of our algorithm. The basic idea will be to show that if $\lambda$ is large, then a large fraction of the variance in the $v$-direction is due to points in $E$.

LEMMA 8.14. *If* $\lambda \ge \Omega(1)$, *then*

$$\operatorname*{Var}_{X \in_u S'}[v^* \cdot X] \ll \frac{|E| \, \mathbb{E}_{Y \in_u E}[|v^* \cdot (Y - \mu)|^2]}{|S'|(\alpha(1-\alpha)(p_d - q_d))^2}.$$

*Proof.* We have that

$$\begin{aligned} |S| \operatorname*{Var}_{X \in_u S'}[v^* \cdot X] &= |S'_P| \left( \operatorname*{Var}_{X \in_u S'_P}[v^* \cdot X] + |v^* \cdot (\mu^{S'_P} - \mu)|^2 \right) \\ &+ |S'_Q| \left( \operatorname*{Var}_{X \in_u S'_Q}[v^* \cdot X] + |v^* \cdot (\mu^{S'_Q} - \mu)|^2 \right) \\ &+ |E| \operatorname*{\mathbb{E}}_{X \in_u E}[|v^* \cdot (X - \mu)|^2] \, . \end{aligned}$$

Since $S_P$ and $S_Q$ are $\varepsilon$-good, we have that

$$\begin{aligned}\operatorname*{Var}_{X\in_u S'_P}[v^*\cdot X] &= \operatorname*{\mathbb{E}}_{X\in_u S'_P}[(v\cdot X - v^*\cdot\mu^{S'_P})^2]\\ &\le (|S_P|/|S'_P|)\operatorname*{\mathbb{E}}_{X\in_u S_P}[(v^*\cdot X - v^*\cdot\mu^{S'_P})^2]\\ &= (|S_P|/|S'_P|)\left(\operatorname*{Var}_{X\in_u S_P}[v^*\cdot X] + (v^*\cdot(\mu^{S_P}-\mu^{S'_P}))^2\right)\\ &\le (|S_P|/|S'_P|)\left(\operatorname*{Var}_{X\sim P}[v^*\cdot X] + (v^*\cdot(p_{-d}-\mu^{S'_P}) + O(\varepsilon\sqrt{\log(1/\varepsilon)}))^2\right)\\ &\le (1+O(\varepsilon/\alpha))\cdot(O(1)+O(\varepsilon\sqrt{\log(1/\varepsilon)})^2) \le O(1)\ ,\end{aligned}$$

and similarly,

$$\operatorname*{Var}_{X\in_u S'_Q}[v^*\cdot X] = O(1)\ .$$

Thus, we have

$$|S'|\operatorname*{Var}_{X\in_u S'}[v^*\cdot X] \le |E|\operatorname*{\mathbb{E}}_{X\in_u E}[|v^*\cdot(X-\mu)|^2] + O(1+|v^*\cdot(p_{-d}-\mu)|^2+|v^*\cdot(q_{-d}-\mu)|^2)|S'|\ .$$

By Lemma 8.9, we have

$$\begin{aligned}|v\cdot(p_{-d}-\mu)|, |v^*\cdot(q_{-d}-\mu)| &\le O(1/(\alpha(1-\alpha)(p_d-q_d)))\\ &\quad\cdot(O(|E_0|/|S'|)|v^*\cdot(\mu^{E^0}-\mu)|\\ &\quad+O(|E^1|/|S'|)|v^*\cdot(\mu^{E^1}-\mu)| + O(\varepsilon\log(1/\varepsilon)))\\ &\le \sqrt{(|E|/|S'|)\operatorname*{\mathbb{E}}_{Y\in_u E}[|v^*\cdot(Y-\mu)|^2]} + O(\varepsilon\log(1/\varepsilon))\ .\end{aligned}$$

However,

$$\lambda = \operatorname*{Var}_{X\in_u S'}[v^*\cdot X] \ll \frac{|E|\,\mathbb{E}_{Y\in_u E}[|v^*\cdot(Y-\mu)|^2]}{|S|(\alpha(1-\alpha)(p_d-q_d))^2} + O(1)\ .$$

Since $\lambda$ is larger than a sufficiently large constant, this completes the proof. ∎

We next show that the threshold $T>0$ required by our algorithm exists.

LEMMA 8.15. *If $\lambda \ge \Omega(1)$, there exists a $T>0$ such that*

$$\operatorname*{Pr}_{X\in_u S'}(|v^*\cdot(X-\mu)| > T+\delta) > 8\exp(-T^2/2) + 8\varepsilon/d\ .$$

*Proof.* Assume for the sake of contradiction that this is not the case, i.e., that for all $T>0$ we have that

$$\operatorname*{Pr}_{X\in_u S'}(|v^*\cdot(X-\mu)| \ge T+\delta) \le 8\exp(-T^2/2) + 8\varepsilon/d\ .$$

Using the fact that $E\subset S'$, this implies that for all $T>0$

$$|E|\operatorname*{Pr}_{Y\in_u E}(|v^*\cdot(Y-\mu)| > T+\delta) \ll |S'|(\exp(-T^2/2)+\varepsilon/d)\ .$$

Therefore, we have that

$$\begin{aligned}\mathop{\mathbb{E}}_{Y\in_u E}[|v^*\cdot(Y-\mu)|^2] &\ll \delta^2+\mathop{\mathbb{E}}_{Y\in_u E}[\min(0,|v^*\cdot(Y-\mu)|-\delta)^2]\\ &\ll \delta^2+\int_0^{\sqrt{d}}\Pr_{Y\in_u E}(|v^*\cdot(Y-\mu)|>T+\delta)TdT\\ &\ll \delta^2+\int_0^{\sqrt{d}}(\varepsilon/d)TdT+\int_0^{2\sqrt{\log(|S'|/|E|)}}TdT\\ &\quad+\int_{2\sqrt{\log(|S'|/|E|)}}^{\infty}(|S'|/|E|)\exp(-T^2/2)TdT\\ &\ll \delta^2+\varepsilon+\log(|S'|/|E|)\ .\end{aligned}$$

On the other hand, we know that

$$\mathop{\mathbb{E}}_{Y\in_u E}[|v^*\cdot(Y-\mu)|^2]\gg(\alpha(1-\alpha)(p_d-q_d))^2\lambda|S'|/|E|\gg\log(|S'|/|E|)\ .$$

Combining with the above we find that

$$\delta^2=O(\varepsilon^{1/3}\lambda)\gg(\alpha(1-\alpha)(p_d-q_d))^2\lambda|S'|/|E|\ .$$

Or in other words,

$$\varepsilon^{4/3}\geq\varepsilon^{1/3}|E|/|S'|\gg(\alpha(1-\alpha)(p_d-q_d))^2\geq\varepsilon^{2/3}\ ,$$

which provides a contradiction. ∎

Finally, we show that $S''$ is closer to $S$ than $S'$ was.

CLAIM 8.16. *If the algorithm returns $S''$, then $\Delta(S,S'')\leq\Delta(S,S')-2\varepsilon/d$.*

*Proof.* Since $S''\subset S$, we can write $S''=S''_P\cup S''_Q\cup E''$ for $S''_P\subseteq S'_P$, $S''_Q\subseteq S_Q$ and $E''\subset E$, where $E''$ has disjoint support from $S''_P\setminus S_P$ and $S''_Q\setminus S_Q$. Thus, we need to show that

$$|E''\setminus E|\geq 2\varepsilon|S|/d+|S'_P\setminus S''_P|+|S'_Q\setminus S''_Q|\ .$$

We have that

$$\begin{aligned}|S'\setminus S''|&=\Pr_{X\in_u S'}(|v\cdot(X-\mu)|\geq T+\delta)|S'|\\ &\geq(8\exp(-T^2/2)+8\varepsilon/d)|S'|\geq(4\exp(-T^2/2)+4\varepsilon/d)|S|\ .\end{aligned}$$

By Hoeffding's inequality, we have that

$$\Pr_{X\sim P}(|v^*\cdot(X-p_{-d})|\geq T)\leq 2\exp(-T^2/2)\ .$$

By Lemma 8.10, we have that $|v^*\cdot(\mu-p_{-d})|\leq\delta$ and so

$$\Pr_{X\sim P}(|v^*\cdot(X-\mu)|\geq T+\delta)\leq 2\exp(-T^2/2)\ .$$

Since $S$ is $(\varepsilon,d)$-good, we have

$$\Pr_{X\in_u S_P}(|v^*\cdot(X-\mu)|\geq T+\delta)\leq 2\exp(-T^2/2)+\varepsilon/d\ .$$

We get the same bound for $X \in_u S_Q$, and so

$$\begin{aligned} &\Pr_{X \in_u S}(|v^* \cdot (X - \mu)| \geq T + \delta) \\ =\ &(|S_P|/|S|) \Pr_{X \in_u S_P}(|v^* \cdot (X - \mu)| \geq T + \delta) + (|S_Q|/|S|) \Pr_{X \in_u S_Q}(|v^* \cdot (X - \mu)| \geq T + \delta) \\ \leq\ &2\exp(-T^2/2) + \varepsilon/d \ . \end{aligned}$$

Since $L''_P \cup L''_Q \subseteq S$ but any $x \in (S'_P \setminus S''_P) \cup (S'_Q \setminus S''_Q)$ has $v^* \cdot (x - \mu) \geq T + \delta$, we have that

$$\begin{aligned} |S'_P \setminus S''_P| + |S'_Q \setminus S''_Q| &\leq \Pr_{X \in_u S}(|v^* \cdot (X - \mu)| \geq T + \delta)|S| \\ &\leq (2\exp(-T^2/2) + \varepsilon/d)|S| \ . \end{aligned}$$

Finally, we have that

$$\begin{aligned} |E \setminus E'| &= |S' \setminus S''| - |S'_P \setminus S''_P| - |S'_Q \setminus S''_Q| \\ &\geq \left(4\exp(-T^2/2) + \frac{4\varepsilon}{d}\right)|S| - \left(2\exp(-T^2/2) + \frac{\varepsilon}{d}\right)|S| \\ &\geq \left(2\exp(-T^2/2) + \frac{3\varepsilon}{d}\right)|S| \\ &\geq |S'_P \setminus S''_P| + |S'_Q \setminus S''_Q| + \frac{2\varepsilon}{d} \ , \end{aligned}$$

which completes the proof. ∎

**8.3. Mixtures of products whose means are close in every coordinate.** In this section, we prove the following theorem.

THEOREM 8.17. *Let $\varepsilon, \tau > 0$ and let $\Pi = \alpha P + (1 - \alpha)Q$ be a $d$-dimensional mixture of two $c$-balanced product distributions $P$ and $Q$ whose means $p$ and $q$ satisfy $\|p - q\|_\infty \leq \delta$, for $\delta \geq \sqrt{\varepsilon \log(1/\varepsilon)}$, and $c \leq p_i, q_i \leq 1 - c$ for $i \in [d]$. Let $S$ be a multiset of $\Omega(d^4 \log(1/\tau)/(\varepsilon^2\delta))$ independent samples from $\Pi$. Let $S'$ be obtained by adversarially changing an $\varepsilon$-fraction of the points in $S$. There exists an algorithm that runs in polynomial time and, with probability at least $1 - \tau$, returns a set of distributions $R$ such that some $\Pi' \in R$ has $d_{\mathrm{TV}}(\Pi, \Pi') \leq O(\delta/\sqrt{c})$.*

We will assume without loss of generality that $\alpha \leq 1/2$. We may also assume that $\alpha > 10\delta \geq 10\varepsilon$ since, otherwise, we can make use of our algorithm for learning a single product distribution.

In this context, we require the following slightly different definition of a good set.

DEFINITION 8.18. *Let $S$ be a multiset in $\{0, 1\}^d$. We say that $S$ is $\varepsilon$-*good *for the mixture $\Pi$ if there exists a partition $S = S_P \cup S_Q$ such that $\left|\frac{|S_P|}{|S|} - \alpha\right| \leq \varepsilon$ and that $S_P$ and $S_Q$ are $\varepsilon/6$-good for the component product distributions $P$ and $Q$, respectively.*

LEMMA 8.19. *If $\Pi$ has mixing weights $\delta \leq \alpha \leq 1 - \delta$, with probability at least $1 - \tau$, a set $S$ of $\Omega(d^4 \log 1/\tau/(\varepsilon^2\delta))$ samples drawn from $\Pi$ is good for $\Pi$.*

The proof of this lemma is in Appendix E. Our theorem will follow from the following proposition.

PROPOSITION 8.20. *Let $\Pi$ be as above and $S$ be a good multiset for $\Pi$. Let $S'$ be any multiset with $\Delta(S, S') \leq 2\varepsilon$. There exists a polynomial-time algorithm that, given*

*$S'$, $\varepsilon > 0$, and $\delta$, returns either a multiset $S''$ with $\Delta(S, S'') \le \Delta(S, S') - 2\varepsilon/d$ or a set of parameters of binary product distributions of size $O(d/(\varepsilon\delta^2))$ which contains the parameters of a $\Pi'$ with $d_{\mathrm{TV}}(\Pi, \Pi') \le O(\delta/\sqrt{c})$.*

Before we present the algorithm, we give one final piece of notation. For $S$ a set of points, we let $\mathrm{Cov}(S)$ be the sample covariance matrix of $S$ and $\mathrm{Cov}_0(S)$ be the sample covariance matrix with zeroed out diagonal. Our algorithm is presented in detailed pseudocode in Algorithm 16.

**Algorithm 16** Filter algorithm for mixture of two binary products whose means are close in every coordinate.

**procedure** FILTER-PRODUCT-MIXTURE-CLOSE$(\varepsilon, S', \delta)$
    Compute $\mu$, the sample mean of $S'$, and $\mathrm{Cov}_0(S')$. Let $C$ be a sufficiently large constant.
    **if** $\mathrm{Cov}_0(S')$ has at most one eigenvector with an absolute eigenvalue more than $C\delta^2$ **then**
        Let $v^*$ be the unit eigenvector of $\mathrm{Cov}_0(S')$ with largest absolute eigenvalue.
        Let $L$ be the set of points $\mu + i\delta v^*$ truncated to be in $[c, 1-c]^d$, for $i \in \mathbb{Z}$ with $|i| \le 1 + \sqrt{d}/\delta$.
        **return** the set of distributions of the form $\Pi' = \alpha' P' + (1-\alpha')Q'$ with the means of $P'$ and $Q'$ in $L$ and $\alpha'$ a multiple of $\varepsilon$ in $[10\varepsilon, 1/2]$.
    Let $v^*$ and $u^*$ be orthogonal eigenvectors with eigenvalues more than $C\delta^2$.
    Find a number $t \ge 1 + 2\sqrt{\log(1/\varepsilon)}$ and $\theta$ a multiple of $\delta^2/d$ such that $r = (\cos\theta)u^* + (\sin\theta)v^*$ satisfies

$$\Pr_{X \in_u S'}\left(\Pr_{Y \in_u S'}(|r \cdot (X - Y)| < t) < 2\varepsilon\right) > 12\exp(-t^2/4) + 3\varepsilon/d \, .$$

    **return** the set $S'' = \{x \in S' \mid \Pr_{Y \in_u S'}(|r \cdot (x - Y)| < t) \ge 2\varepsilon\}$ .

To analyze this algorithm, we begin with a few preliminaries. First, we recall that $S = S_P \cup S_Q$. We can write $S' = S'_P \cup S'_Q \cup E$, where $S'_P \subset S_P$, $S'_Q \subset S_Q$, and

$$|S|\Delta(S, S') = |S_P \setminus S'_P| + |S_Q \setminus S'_Q| + |E| \, .$$

Let $\mu^{S'_P}$ and $\mu^{S'_Q}$ be the sample means of $S'_P$ and $S'_Q$, respectively.

LEMMA 8.21. *We have that $\alpha\|p - \mu^{S'_P}\|_2, (1-\alpha)\|q - \mu^{S'_Q}\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$ .*

*Proof.* The proof follows from Lemma 8.8. ∎

We will require that the matrix $\mathrm{Cov}_0(S')$ is close to being PSD. The proof of this fact is rather technical and we defer it to Appendix E.

LEMMA 8.22. *Let $T$ be the multiset obtained from $S'$ by replacing all points of $S'_P$ with copies of $\mu^{S'_P}$ and all points of $S'_Q$ with copies of $\mu^{S'_Q}$. Then, $\|Cov_0(S') - Cov(T)\|_2 = O(\delta^2)$ .*

We are now prepared to show that the first return condition outputs a correct answer. We begin by showing that vectors $u$ with large inner products with $\mu^{S'_P} - \mu$ or $\mu^{S'_Q} - \mu$ correspond to large eigenvectors of $\mathrm{Cov}_0(S')$.

LEMMA 8.23. *For $u \in \mathbb{R}^d$, we have*

$$\alpha(u \cdot (\mu^{S'_P} - \mu))^2 + (1-\alpha)(u \cdot (\mu^{S'_P} - \mu))^2 \le 2u^T Cov_0(S')u + O(\delta^2)\|u\|_2^2.$$

*Proof.* Using Lemma 8.22, we have $u^T \mathrm{Cov}_0(S')u = \mathrm{Var}_{X \in_u T}(u \cdot X) + O(\delta^2)\|u\|_2^2$. From the definition of $T$ it follows that

$$\begin{aligned}\operatorname*{Var}_{X\in_u T}(u\cdot X) &\geq \left(\frac{|S'_P|}{|S'|}\right)(u\cdot(\mu^{S'_P}-\mu))^2 + \left(\frac{|S'_Q|}{|S'|}\right)(u\cdot(\mu^{S'_Q}-\mu))^2 + \frac{|E|}{|S'|}\operatorname*{Var}_{X\in_u E}(u\cdot X)\\ &\geq (\alpha-2\varepsilon)(u\cdot(\mu^{S'_P}-\mu))^2 + (1-\alpha-2\varepsilon)(u\cdot(\mu^{S'_Q}-\mu))^2\\ &\geq \alpha/2\cdot(u\cdot(\mu^{S'_P}-\mu))^2 + (1-\alpha)/2\cdot(u\cdot(\mu^{S'_Q}-\mu))^2 . \qquad \square\end{aligned}$$

Next, we show that if there is only one large eigenvalue of $\mathrm{Cov}_0(S')$, the means in question are both close to a given line.

LEMMA 8.24. *There are $\tilde{p}, \tilde{q} \in L$ such that $\|p - \tilde{p}\|_2 \leq O(\delta/\sqrt{\alpha})$ and $\|q - \tilde{q}\|_2 \leq O(\delta/\sqrt{1-\alpha})$.*

*Proof.* Let $p' = \mu + av^*$, $q' = \mu + bv^*$ with $a, b \in \mathbb{R}$ be the closest points to $p$ and $q$ on the line $\mu + cv^*$, for $c \in \mathbb{R}$. Then, $v^* \cdot (p' - p) = 0$ and since $v^*$ is the only eigenvector of the symmetric matrix $\mathrm{Cov}_0(S')$ with eigenvalue more than $C(\delta^2 + \varepsilon \log(1/\varepsilon))$, we have that

$$(p'-p)^T \mathrm{Cov}_0(S')(p'-p) \leq C(\delta^2 + \varepsilon\sqrt{\log(1/\varepsilon)})\|p'-p\|_2^2 .$$

We thus obtain

$$\begin{aligned}\|p'-p\|_2^4 &= (p'-p)\cdot(p-\mu)^2\\ &\leq 2(p'-p)\cdot(p-\mu^{S_P})^2 + 2(p'-p)\cdot(p-\mu^{S_P})^2\\ &\leq O(\varepsilon^2\log(1/\varepsilon)/\alpha^2)\|p'-p\|_2^2 + (4/\alpha)\cdot(p'-p)^T\mathrm{Cov}_0(S')(p'-p)^T\\ &\quad + O(\delta^2/\alpha)\|p'-p\|_2^2\\ \text{(since } \alpha\geq\varepsilon) \quad &\leq O((\delta^2+\varepsilon\log(1/\varepsilon))/\alpha)\|p'-p\|_2^2\\ &\leq O(\delta^2/\alpha)\|p'-p\|_2^2 ,\end{aligned}$$

where the second line uses Lemmas 8.21 and 8.23. We thus have that $\|p' - p\|_2 \leq O(\delta/\sqrt{\alpha})$. Letting $i\delta$ be the nearest integer multiple to $a$, we have that $\tilde{p} := \mu + i\delta v^*$ has

$$\|p-\tilde{p}\|_2 \leq \|p'-p\|_2 + \|p'-\tilde{p}\|_2 \leq O(\delta/\sqrt{\alpha}).$$

Note that we have $\|p - p'\|_2 \leq \|p - \mu\|_2 \leq \sqrt{d}\|p - \mu\|_\infty \leq \sqrt{d}$. So, $a \leq \sqrt{d}/\delta$. Thus, $|i| \leq 1 + \sqrt{d}/\delta$. If $\tilde{p} \notin [c, 1 - c]$, then replacing any coordinates less than $c$ with $c$ and more than $1 - c$ with $1 - c$ can only decrease the distance to $p$, since $p \in [c, 1 - c]^d$.

Similarly, we show that there is a $\tilde{q} \in L$ such that $\|q - \tilde{q}\|_2 \leq O(\delta/\sqrt{1-\alpha})$, which completes the proof. $\square$

COROLLARY 8.25. *If the algorithm outputs a set of distributions in step 6 of Algorithm 16, then one of those distributions has $d_{\mathrm{TV}}(\Pi', \Pi) \leq O(\delta/\sqrt{c})$.*

*Proof.* There is a distribution in the set $\Pi = \alpha' P' + (1 - \alpha')Q'$, where $|\alpha - \alpha'| \leq \varepsilon$ and the means of $P'$ and $Q'$ are $\tilde{p}$ and $\tilde{q}$ as in Lemma 8.24. Then, we have $d_{\mathrm{TV}}(P, P') \leq \|p - \tilde{p}\|/\sqrt{c} \leq O(\delta/\sqrt{\alpha c})$ and $d_{\mathrm{TV}}(Q, Q') \leq \|p - \tilde{p}\|/\sqrt{c} \leq O(\delta/\sqrt{(1-\alpha)c})$. Thus, we have

$$\begin{aligned}d_{\mathrm{TV}}(\Pi',\Pi) \leq O(\varepsilon) + \alpha d_{\mathrm{TV}}(P,P') + (1-\alpha)d_{\mathrm{TV}}(Q,Q') &\leq O(\varepsilon)\\ + O((\sqrt{\alpha}+\sqrt{1-\alpha})\delta/\sqrt{c}) &\leq O(\delta/\sqrt{c}) . \qquad \square\end{aligned}$$

Next, we analyze the second case of the algorithm. We must show that step 8 of Algorithm 16 will find an $r$ and $t$. First, we claim that there is a $\theta$ which makes $r$ nearly perpendicular to $\mu^{S'_P} - \mu^{S'_Q}$.

Lemma 8.26. *There exists an $r = (\cos\theta)u^* + (\sin\theta)v^*$, with $\theta$ a multiple of $\delta^2/d$, that has*

$$|r \cdot (\mu^{S'_P} - \mu^{S'_Q})| \le \delta^2/\sqrt{d}.$$

*Proof.* Let $z = (\mu^{S'_P} - \mu^{S'_Q})$. If $u^* \cdot z = 0$, then $\theta = 0$ suffices. Otherwise, we take $\theta' = \cot^{-1}(\frac{v^*\cdot z}{u^* \cdot z})$. Then, let $\theta$ be the nearest multiple of $\delta^2/d$ to $\theta'$. Note that $|\cos\theta - \cos\theta'|, |\sin\theta - \sin\theta'| \le |\theta - \theta'|$ and $|u^* \cdot z|, |v^* \cdot z| \le \sqrt{\|z\|_2} \le \sqrt{d}$. Then, we have

$$\begin{aligned}
|r \cdot z| &= |(\cos\theta)(u^* \cdot z) + (\sin\theta)(v^* \cdot z)| \\
&\le |(\cos\theta')(u^* \cdot z) + (\sin\theta')(v^* \cdot z)| + |\theta - \theta'|\sqrt{d} \\
&= |\sin\theta'||u^* \cdot z + (\cot\theta')(v^* \cdot z)| + |\theta - \theta'|\sqrt{d} \\
&\le 0 + \delta^2/\sqrt{d}\,.
\end{aligned}$$

∎

We now need to show that for this $r$, step 8 of Algorithm 16 will find a $t$. For this $r$, $r \cdot \mu^{S'_P}$ and $r \cdot \mu^{S'_Q}$ are close. We need to show that $E$ contains many elements $x$ whose $r \cdot x$ is far from these. We can express this in terms of $T$.

Lemma 8.27. *Let $r$ be a unit vector in $r \in \langle u^*, v^* \rangle$ with $|r \cdot (\mu^{S'_P} - \mu^{S'_Q})| \le \delta^2/\sqrt{d}$. Then, there is a $t > 1$ such that*

$$\Pr_{X \in_u T}(r \cdot (X - \mu^{S'_P}) > 2t) > 12\exp(-(t-1)^2/4) + \frac{3\varepsilon}{d}\,.$$

*Proof.* First, we wish to show that $\mathbb{E}_{X \in_u E}[(r \cdot (X - \mu^{S'_P}))^2]$ is large.

Since $r \in \mathrm{span}(u^*, v^*)$, $|r^T \mathrm{Cov}_0(S')r| \ge C\delta^2$. By Lemma 8.22, we have that

$$\mathop{\mathrm{Var}}_{X \in_u T}(r \cdot X) = r^T \mathrm{Cov}(T) r \ge r^T \mathrm{Cov}_0(S')r - O(\delta^2) \ge (C - O(1))\delta^2 \ge (C/2)\delta^2$$

for sufficiently large $C$, and we also have that $r^T \mathrm{Cov}_0(S')r$ is positive.

We note that

$$\begin{aligned}
r^T \mathrm{Cov}(T) r &= \mathrm{Var}(r \cdot T) \\
&= (|E|/|S'|) \mathop{\mathrm{Var}}_{X \in_u E}(r \cdot X) + O(\alpha)(r \cdot (\mu - \mu^{S'_P}))^2 + O(1-\alpha)(r \cdot (\mu - \mu^{S'_Q}))^2 \\
&\quad + (|E|/|S'|)(r \cdot (\mu - \mu^E))^2 \\
&= (|E|/|S'|)\left(\mathop{\mathrm{Var}}_{X \in_u E}(r \cdot X) + (r \cdot (\mu - \mu^E))^2\right) + O(\delta^2)\,. \qquad (42)
\end{aligned}$$

Now,

$$\mathop{\mathbb{E}}_{X \in_u E}[(r \cdot (X - \mu^{S'_P}))^2] = \mathop{\mathrm{Var}}_{X \in_u E}(r \cdot X) + (r \cdot (\mu^{S'_P} - \mu^E))^2\,.$$

We also have that

$$\begin{aligned}
|S'|(r \cdot \mu) &= (|S'| - |E|)(r \cdot \mu^{S'_P}) + |S'_Q|(r \cdot (\mu^{S'_P} - \mu^{S'_Q})) + |E|(r \cdot \mu^E) \\
&= (|S'| - |E|)(r \cdot \mu^{S'_P}) + |E|(r \cdot \mu^E) + |S'|O(\delta^2)\,.
\end{aligned}$$

Thus,

$$(|S'| - |E|)(r \cdot (\mu - \mu^E)) = (|S'| - |E|)(r \cdot (\mu^{S'_P} - \mu^E)) + |S'|O(\delta^2)$$

or

$$(r \cdot (\mu - \mu^E)) = (r \cdot (\mu^{S'_P} - \mu^E)) + O(\delta^2) \ .$$

This implies that

$$(r \cdot (\mu^{S'_P} - \mu^E))^2 \geq (r \cdot (\mu - \mu^E))^2/2 - O(\delta^4) \ .$$

Substituting into (42), we have

$$(|E|/|S'|) \mathop{\mathbb{E}}_{X \in_u E}[(r \cdot (X - \mu^{S'_P}))^2] = (|E|/|S'|) \left[ \mathop{\mathrm{Var}}_{X \in_u E}[r \cdot X] + (r \cdot (\mu^{S'_P} - \mu^E))^2 \right] - O(\delta^4) \gg C/2\delta^2.$$

Thus, for $C$ sufficiently large,

$$\mathop{\mathbb{E}}_{X \in_u E}[(r \cdot (X - \mu^{S'_P}))^2] \gg \delta^2/\varepsilon.$$

Suppose for a contradiction that this lemma does not hold. Then, since $E \subset T$, we have

$$\Pr_{X \in_u E}\left(r \cdot (X - \mu^{S'_P}) > 2t\right) \leq (|S'|/|E|) 12 \exp(-t^2/2) + \frac{3\varepsilon}{d} \ .$$

Thus, we have

$$\Pr_{X \in_u E}(r \cdot (X - \mu^{S'_P}) > t) \leq (|S'|/|E|) 12 \exp(-(t-1)^2/4) + \frac{3\varepsilon}{d} \ ,$$

and we can write

$$\begin{aligned}
|S'|\delta^2 &\ll |E| \mathop{\mathbb{E}}_{X \in_u E}[(r.X - r.\mu^{S_P})^2] \\
&= |E| \int_0^{\sqrt{d}} \Pr_{X \in_u E}(r \cdot (X - \mu^{S'_P}) > t) t dt \\
&\ll |E| \int_0^{1+\sqrt{\log(|S'|/|E|)/2}} t dt + |S'| \int_{1+\sqrt{\log(|S'|/|E|)/2}}^{\infty} \exp(-(t-1)^2/4) t dt \\
&\quad + \int_0^{\sqrt{n}} \varepsilon/d t dt \\
&\ll |E| \log(|S'|/|E|) + |E| + |S'|(|E|/|S|) + \varepsilon \\
&\leq |S'| \cdot O(\varepsilon \log(1/\varepsilon)) \ .
\end{aligned}$$

Since we assumed that $\delta^2 \geq \Omega(\varepsilon \log(1/\varepsilon))$, this is a contradiction. ◻

To get a similar result for $S'$, we first need to show that $S'_P$ and $S'_Q$ are suitably concentrated about their means.

LEMMA 8.28. *If $t \geq 1$,*

$$(1 - |E|/|S'|) \Pr_{X \in_u S'_P \cup S'_Q}\left(r \cdot (X - \mu^{S'_P}) > t\right) \leq \frac{5}{4} \exp(-(t-1)^2/2) + \frac{\varepsilon}{5d}.$$

*If $t \geq 1 + \sqrt{2 \log 6/\varepsilon}$, this is strictly less than $2\varepsilon/3$.*

*Proof.* We present

$$\begin{aligned}\Pr_{X\in_u S'_P}(r\cdot(X-\mu^{S'_P})\le t) &\le (|S_P|/|S'_P|)\Pr_{X\in_u S_P}(r\cdot(X-\mu^{S'_P})\le t)\\ &\le \left(1+\frac{O(\varepsilon)}{1-\alpha}\right)\cdot\left(\Pr_{X\sim P}(r\cdot(X-\mu^{S'_P})\le t)+\frac{\varepsilon}{12d}\right)\\ &=\left(1+\frac{O(\varepsilon)}{1-\alpha}\right)\\ &\quad\cdot\left(\Pr_{X\sim P}\left(r\cdot(X-p)\le t-(r\cdot(\mu^{S'_P}-p))\right)+\frac{\varepsilon}{6d}\right)\end{aligned}$$

(using Lemma 8.21 and Hoeffding's inequality)

$$\le\left(1+\frac{O(\varepsilon)}{1-\alpha}\right)\cdot\left(2\exp(-(t-1/2)^2/2)+\frac{\varepsilon}{6d}\right)\ .$$

Similarly,

$$\Pr_{X\sim S'_Q}(r\cdot(X-\mu^{S'_Q})\le t)\le\left(1+\frac{O(\varepsilon)}{1-\alpha}\right)\cdot\left(2\exp(-(t-1/2)^2/2)+\frac{\varepsilon}{6d}\right)\ .$$

Since $|r\cdot(\mu^{S_Q}-\mu^{S_P})|\le\delta^2/\sqrt{d}\le 1/2$, we have

$$\Pr_{X\sim S'_Q}\left(r\cdot(X-\mu^{S'_Q})\le t\right)\le\left(1+\frac{O(\varepsilon)}{1-\alpha}\right)\cdot\left(2\exp(-(t-1)^2/2)+\frac{\varepsilon}{6d}\right)\ .$$

Noting that $1-(|S'_P|+|S'_Q|)/|S'|=|E|/|S'|\ge 4\varepsilon/3$, we have

$$\begin{aligned}&(1-|E|/|S'|)\Pr_{X\in_u S'_P\cup S'_Q}\left(r\cdot(X-\mu^{S'_P})>t\right)\\ &=(|S'_P|/|S'|)\Pr_{X\sim S'_P}\left(r\cdot(X-\mu^{S'_P})>t\right)+(|S'_Q|/|S'|)\Pr_{X\sim S'_Q}\left(r\cdot(X-\mu^{S'_P})>t\right)\\ &=(\alpha+O(\varepsilon))\left(1+\left(1+\frac{O(\varepsilon)}{\alpha}\right)\right)+(1-\alpha+O(\varepsilon))\left(1+\left(1+\frac{O(\varepsilon)}{1-\alpha}\right)\right)\\ &\quad\cdot\left(2\exp(-(t-1)^2/2)+\frac{\varepsilon}{6d}\right)\\ &\le(1+O(\varepsilon))\cdot\left(2\exp(-(t-1)^2/2)+\frac{\varepsilon}{6d}\right)\\ &\le\frac{5}{2}\exp(-(t-1)^2/2)+\frac{\varepsilon}{5d}\ ,\end{aligned}$$

for $\varepsilon$ sufficiently small. If $t\ge 1+\sqrt{2\log 6/\varepsilon}$, this expression is $(5/2)(\varepsilon/6)+\varepsilon/5d\le 2\varepsilon/3$. ∎

Now we can finally show that a $t$ exists for this $r$, so step 8 of Algorithm 16 will succeed.

LEMMA 8.29. *There is a* $t\ge 1+2\sqrt{\log(9/\varepsilon)}$ *such that*

$$\Pr_{X\in_u S'}\left(\Pr_{Y\in_u S'}(r\cdot(X-Y)>t)<2\varepsilon\right)>12\exp(-(t-1)^2/4)+\frac{3\varepsilon}{d}\ .$$

*Proof.* By Lemma 8.27, there exists a $t\ge 1$ such that

$$\Pr_{X\in_u T}\left(r\cdot(X-\mu^{S'_P})>2t\right)>12\exp(-(t-1)^2/4)+\frac{3\varepsilon}{d}\ .$$

Using the definition of $T$, the points when $x = \mu^{S'_P}$ or $x = \mu^{S'_Q}$ do not contribute to this probability so all points in $T$ that satisfy $r \cdot (x - \mu^{S'_P}) > 2t$ come from $E$. Since $E \subset S'$ and $|S'| = |T|$, we have

(43)
$$\Pr_{X \in_u S'}\left(r \cdot (X - \mu^{S'_P}) > 2t\right) \geq \Pr_{X \in_u T}\left(r \cdot (X - \mu^{S'_P}) > 2t\right) > 12\exp(-(t-1)^2/4) + \frac{3\varepsilon}{d} .$$

Noting that $|E|/|S'| \leq 4\varepsilon/3$, all except a $4\varepsilon/3$-fraction of points $x \in T$ have $r \cdot (x - \mu^{S'_P}) = O(\delta^2)$. So, $4\varepsilon/3 \geq 12\exp(-(t-1)^2/4)$. Therefore, $t \geq 1 + 2\sqrt{\log(9/\varepsilon)}$.

Thus, by Lemma 8.28, we have $(1 - |E|/|S'|)\Pr_{X \in_u S'_P \cup S'_Q}\left(r \cdot (X - \mu^{S'_P}) > t\right) < 2\varepsilon/3$. Again, using that $|E|/|S'| \leq 4\varepsilon/3$, we have that

$$\Pr_{X \in_u S'}\left(r \cdot (X - \mu^{S'_P}) > t\right) < 2\varepsilon .$$

Consequently, if $x$ satisfies $r \cdot (x - \mu^{S'_P}) > 2t$, then it satisfies $\Pr_{Y \in_u S'}(r \cdot (x - Y) \leq t) < 2\varepsilon$. Substituting this condition into (43) gives the lemma. ∎

Again we need to show that any filter does not remove too many points of $S$. We need to show this for an arbitrary $r$, not just one nearly parallel to $\mu^{S'_P} - \mu^{S'_Q}$.

LEMMA 8.30. *For any unit vector $r'$ and $t \geq 2\sqrt{\log(1/\varepsilon)}$, we have*

$$(1 - |E|/|S'|)\Pr_{X \in_u S'_P \cup S'_Q}\left(\Pr_{Y \in_u S'}(r' \cdot (X - Y) \leq t) < 2\varepsilon\right) \leq 3\exp(-t^2/4) + \frac{\varepsilon}{4d} .$$

*Proof.* Using Hoeffding's inequality, we have

$$\begin{aligned} |S'_P| \Pr_{X \in_u S'_P}(r \cdot (p - X) > t/2) &\leq |S_P| \Pr_{X \in_u S_P}(|r' \cdot (X - p)| > t/2) \\ &\leq |S_P|\left(\Pr_{X \sim P}(|r' \cdot (X - p)| > t/2) + \frac{\varepsilon}{6d}\right) \\ &\leq |S_P|\left(2\exp(-t^2/4) + \frac{\varepsilon}{6d}\right) . \end{aligned} \tag{44}$$

Every point $x$ with $|r' \cdot (x - p)| \leq t/2$ has $|r' \cdot (x - y) \leq t|$ for all $y$ with $|r' \cdot (y - p)| \leq t/2$. Thus, for $x$ with $|r' \cdot (x - p)| \leq t/2$, we have

$$\Pr_{Y \in_u S'}(r' \cdot (x - Y) \leq t) \geq \frac{|S_P|}{|S'|} - \frac{|S_P|}{|S'|}\left(2\exp(-t^2/4) + \frac{\varepsilon}{6d}\right) .$$

When $t \geq 2\sqrt{\log(1/\varepsilon)}$, we have

$$\frac{|S_P|}{|S'|}\left(2\exp(-t^2/4) + \frac{3\varepsilon}{d}\right) \leq (1 + 2\varepsilon) \cdot \left(2\varepsilon + \frac{\varepsilon}{6d}\right) \leq 3\varepsilon .$$

Also, we have

$$\frac{|S_P|}{|S'|} \leq \frac{(\alpha - \varepsilon/6)|S|}{|S|(1 - 2\varepsilon)} \leq \alpha - 3\varepsilon \leq 7\varepsilon .$$

Thus, we have $\Pr_{Y \in_u S'}(r \cdot (x - Y) \leq t) \geq 4\varepsilon > 2\varepsilon$.

But inequality (44) gives a bound on the number of $x$ in $S_P$ that do not satisfy this condition. That is,

$$|S'_P| \Pr_{X \in_u S'_P}\left(\Pr_{Y \in_u S'}(r' \cdot (X - Y) \leq t) < 2\varepsilon\right) \leq |S_P|\left(2\exp(-t^2/4) + \frac{\varepsilon}{6d}\right) .$$

Similarly, every point $x$ with $|r' \cdot (x-q)| \leq t/2$ has

$$\Pr_{y \in_u S'}(r' \cdot (x-y) \leq t) > 2\varepsilon$$

and

$$|S'_Q| \Pr_{X \in_u S'_Q}(r' \cdot (X-p) > t/2) \leq \left(2\exp(-t^2/4) + \frac{\varepsilon}{6d}\right) .$$

Thus,

$$|S'_Q| \Pr_{X \in_u S'_Q}\left(\Pr_{Y \in_u S'}(r' \cdot (X-Y) \leq t) < 2\varepsilon\right) \leq |S_Q| \left(2\exp(-t^2/4) + \frac{\varepsilon}{6d}\right) .$$

Summing these gives

$$(|S'_P| + |S'_Q|) \Pr_{X \in_u S'_P \cup S'_Q}\left(\Pr_{Y \in_u S'}(r' \cdot (X-Y) \leq t) < 2\varepsilon\right) \leq |S| \left(2\exp(-t^2/4) + \frac{\varepsilon}{6d}\right) .$$

Dividing by $|S'|$ and noting that $|S| \leq (1+2\varepsilon)|S'| \leq (3/2)|S'|$ completes the proof. ∎

Now, we can show that the filter improves $\Delta(S, S'')$, such that the algorithm is correct in the filter case.

CLAIM 8.31. *If we reach step* 9 *of Algorithm* 16 *and return* $S''$, *then* $\Delta(S, S'') \leq \Delta(S, S') - 2\varepsilon/d$.

*Proof.* We can write $S'' = S''_P \cup S''_Q \cup E''$, where $E''$ has disjoint support from $S_P \setminus S''_P$ and $S_Q \setminus S''_Q$. Note that, since we have $S'' \subset S'$, we can define these sets such that $S''_P \subseteq S'_P$, $S''_Q \subseteq S'_Q$, and $E'' \subseteq E$. We assume that we do. Now we have that

$$\Delta(S, S') - \Delta(S, S'') = \frac{|E'' \setminus E'| - |S''_P \setminus S'_P| - |S''_Q \setminus S'_Q|}{|S|} .$$

Therefore,

$$\Delta(S, S') - \Delta(S, S'') = \frac{|S'' \setminus S'| - 2(|S''_P \setminus S'_P| + |S''_Q \setminus S'_Q|)}{|S|} .$$

In step 8 of Algorithm 16, we found a vector $r$ and $t \geq 1 + 2\sqrt{\log(1/\varepsilon)}$ such that

$$\Pr_{X \in_u S'}\left(\Pr_{Y \in_u S'}(|r \cdot (X-Y)| < t) < 2\varepsilon\right) > 12\exp(-(t-1)^2/4) + \frac{3\varepsilon}{d} .$$

Then in step 9 of Algorithm 16, we remove at least a $12\exp(-t^2/4) + 3\varepsilon/d$-fraction of points. That is,

$$|S'' \setminus S'| \geq \left(12\exp(-t^2/4) + \frac{3\varepsilon}{d}\right)|S'| .$$

The fact that $t \geq 1 + 2\sqrt{\log(1/\varepsilon)}$ allows us to use Lemma 8.30, with $r' = r$, yielding

$$(1 - |E|/|S'|) \Pr_{X \in_u S'_P \cup S'_Q}\left(\Pr_{Y \in_u S'}(r \cdot (X-Y) \leq t-1) < 2\varepsilon\right) \leq 3\exp(-(t-1)^2/4) + \frac{\varepsilon}{4d} .$$

This implies that

$$(1 - |E|/|S'|) \Pr_{X \in_u S'_P \cup S'_Q} \left( \Pr_{Y \in_u S'} (r \cdot (X - Y) < t) < 2\varepsilon \right) \leq 3\exp(-(t-1)^2/4) + \frac{\varepsilon}{4d} .$$

Thus,

$$|S''_P \setminus S'_P| + |S''_Q \setminus S'_Q| \leq \left( 3\exp(-(t-1)^2/4) + \frac{\varepsilon}{4d} \right) |S'| ,$$

and we have

$$\begin{aligned} \Delta(S, S') - \Delta(S, S'') &\geq \Big( 12\exp(-(t-1)^2/4) + 3\varepsilon/d \\ &\qquad - 2\left( 3\exp(-(t-1)^2/4) + \frac{\varepsilon}{4d} \right) \Big) |S'|/|S| \\ &\geq \frac{2\varepsilon}{d} , \end{aligned}$$

since $|S'| \geq |S|(1 - \Delta(S, S')) \geq (1 - 2\varepsilon)|S| \geq 5|S|/6$. ∎

**Appendix A. Deferred proofs from section 4.** This section contains deferred proofs of several concentration inequalities.

*Proof of Lemma* 4.3. Recall that for any $J \subseteq [N]$, we let $w^J \in \mathbb{R}^N$ be the vector which is given by $w_i^J = \frac{1}{|J|}$ for $i \in J$ and $w_i^J = 0$ otherwise. By convexity, it suffices to show that

$$\Pr\left[ \exists J : |J| = (1-\varepsilon)N \text{ and } \left\| \sum_{i=1}^N w_i^J Y_i Y_i^\top - (1-\varepsilon) I \right\|_2 \geq \delta_1 \right] \leq \tau .$$

For any fixed $w^J$ we have

$$\begin{aligned} \sum_{i=1}^n w_i^J Y_i Y_i^\top - I &= \frac{1}{(1-\varepsilon)N} \sum_{i \in J} Y_i Y_i^\top - I \\ &= \frac{1}{(1-\varepsilon)N} \sum_{i=1}^N Y_i Y_i^\top - \frac{1}{1-2\varepsilon} I \\ &\quad - \left( \frac{1}{(1-\varepsilon)N} \sum_{i \notin J} Y_i Y_i^\top - \left( \frac{1}{1-\varepsilon} - 1 \right) I \right) . \end{aligned}$$

Therefore, by the triangle inequality, we have

$$\begin{aligned} \left\| \sum_{i=1}^N w_i^I Y_i Y_i^\top - (1-\varepsilon) I \right\|_2 &\leq \left\| \frac{1}{(1-\varepsilon)N} \sum_{i=1}^N Y_i Y_i^\top - \frac{1}{1-\varepsilon} I \right\|_2 \\ &\quad + \left\| \frac{1}{(1-\varepsilon)N} \sum_{i \notin J} Y_i Y_i^\top - \left( \frac{1}{1-\varepsilon} - 1 \right) I \right\|_2 . \end{aligned}$$

Observe that the first term on the RHS does not depend on the choice of $J$. Let $E_1$ denote the event that

$$\left\| \frac{1}{(1-\varepsilon)N} \sum_{i=1}^N Y_i Y_i^\top - \frac{1}{1-\varepsilon} I \right\|_2 \leq \delta_1 . \tag{45}$$

By Lemma 2.22, this happens with probability $1-\tau$ so long as

$$N = \Omega\left(\frac{d+\log(1/\tau)}{\delta_1^2}\right) .$$

For any $J \subset [n]$ so that $|J| = (1-\varepsilon)n$, let $E_2(J)$ denote the event that

$$\left\| \frac{1}{(1-\varepsilon)N} \sum_{i \notin J} Y_i Y_i^\top - \left(\frac{1}{1-\varepsilon} - 1\right) I \right\|_2 \leq \delta_1 .$$

Fix any such $J$. By multiplying both sides by $\rho = (1-\varepsilon)/\varepsilon$, the event $E_2(J)$ is equivalent to the event that

$$\left\| \frac{1}{\varepsilon N} \sum_{i \notin J} Y_i Y_i^\top - I \right\|_2 > \rho\delta_1 .$$

Let $A, B$ be as in Lemma 2.22. Observe that $\rho\delta_1 = \Omega(\log 1/\varepsilon) \geq 1$ for $\varepsilon$ sufficiently small. Then, by Lemma 2.22, we have that for any fixed $J$,

$$\Pr\left[\left\| \frac{1}{\varepsilon N} \sum_{i \notin J} Y_i Y_i^\top - I \right\|_2 > \rho\delta_1\right] \leq 4\exp\left(Ad - B\varepsilon N\rho\delta_1\right) .$$

Let $H(\varepsilon)$ denote the binary entropy function. We now have

$$\begin{aligned}
&\Pr\left[\left(\bigcap_{J:|J|=(1-\varepsilon)N} E_2(J)\right)^c\right] \\
&\overset{(a)}{\leq} 4\exp\left(\log\binom{N}{\varepsilon N} + Ad - B\varepsilon N\rho\delta_1\right) \\
&\overset{(b)}{\leq} 4\exp\left(NH(\varepsilon) + Ad - B\varepsilon N\rho\delta_1\right) \\
&\overset{(c)}{\leq} 4\exp\left(\varepsilon N(O(\log 1/\varepsilon) - N\rho) + Ad\right) \\
&\overset{(d)}{\leq} 4\exp\left(-\varepsilon N/2 + Ad\right) \overset{(e)}{\leq} O(\tau) ,
\end{aligned}$$

as claimed, where (a) follows by a union bound over all sets $J$ of size $(1-\varepsilon)N$, (b) follows from the bound $\log\binom{n}{\varepsilon n} \leq \varepsilon H(\varepsilon)$, (c) follows since $H(\varepsilon) = O(\varepsilon\log 1/\varepsilon)$ as $\varepsilon \to 0$, (d) follows from our choice of $\delta_1$, and (e) follows from our choice of $n$. This completes the proof. ∎

*Proof of Theorem* 4.12. We first recall Isserlis' theorem, which we will require in this proof.

THEOREM A.1 (Isserlis' theorem). *Let $a_1, \ldots, a_k \in \mathbb{R}^d$ be fixed vectors. Then if $X \sim \mathcal{N}(0, I)$, we have*

$$\mathbb{E}\left[\prod_{i=1}^k \langle a_i, X\rangle\right] = \sum\prod\langle a_i, a_j\rangle ,$$

*where the $\sum\prod$ is over all matchings of $\{1, \ldots, k\}$.*

Let $v = A^\flat \in \mathcal{S}_{\mathrm{sym}}$. We will show that

$$\langle v, Mv\rangle = 2v^T \left(\Sigma^{\otimes 2}\right) v + v^T \left(\Sigma^\flat\right) \left(\Sigma^\flat\right)^T v \ .$$

Since $M$ is a symmetric operator on $\mathbb{R}^{d^2}$, its quadratic form uniquely identifies it and this suffices to prove the claim.

Since $A$ is symmetric, it has an eigenvalue expansion $A = \sum_{i=1}^d \lambda_i u_i u_i^T$, which immediately implies that $v = \sum_{i=1}^d \lambda_i u_i \otimes u_i$. Let $X \sim \mathcal{N}(0, \Sigma)$. We compute the quadratic form

$$\begin{aligned}
\langle v, Mv\rangle &= \sum_{i,j=1}^d \lambda_i \lambda_j \langle u_i \otimes u_i, \mathbb{E}[(X\otimes X)(X \otimes X)^T] u_j \otimes u_j\rangle \\
&= \sum_{i,j=1}^d \lambda_i \lambda_j \, \mathbb{E}\left[\langle u_i \otimes u_i, (X\otimes X)(X \otimes X)^T u_j \otimes u_j\rangle\right] \\
&= \sum_{i,j=1}^d \lambda_i \lambda_j \, \mathbb{E}\left[\langle u_i, X\rangle^2 \langle u_j, X\rangle^2\right] \\
&= \sum_{i,j=1}^d \lambda_i \lambda_j \, \mathbb{E}\left[\langle B^T u_i, Y\rangle^2 \langle B^T u_j, Y\rangle^2\right] \\
&= \sum_{i,j=1}^d \lambda_i \lambda_j \left(\langle B^T u_i, B^T u_i\rangle \langle B^T u_j, B^T u_j\rangle + 2\langle B^T u_i, B^T u_j\rangle^2\right) \ ,
\end{aligned}$$

where the last line follows by invoking Isserlis's theorem. We now manage both sums individually. We have

$$\begin{aligned}
\sum_{i,j=1}^d \lambda_i \lambda_j \langle B^T u_i, B^T u_i\rangle \langle B^T u_j, B^T u_j\rangle &= \left(\sum_{i=1}^d \lambda_i u_i^T \Sigma u_i\right)^2 \\
&= \left(\sum_{i=1}^d \lambda_i \left(u_i \otimes u_i\right)^T \left(\Sigma^\flat\right)\right)^2 \\
&= v^T \left(\Sigma^\flat\right) \left(\Sigma^\flat\right)^T v
\end{aligned}$$

and

$$\begin{aligned}
\sum_{i,j=1}^d \lambda_i \lambda_j \langle B^T u_i, B^T u_j\rangle^2 &= \sum_{i,j} \lambda_i \lambda_j \langle (B^T u_i)^{\otimes 2}, (B^T u_j)^{\otimes 2}\rangle \\
&= \sum_{i,j=1}^d \lambda_i \lambda_j \langle (B^T \otimes B^T) u_i \otimes u_i, (B^T \otimes B^T) u_j \otimes u_j\rangle \\
&= \sum_{i,j=1}^d \lambda_i \lambda_j (u_i \otimes u_i) \Sigma^{\otimes 2} (u_j \otimes u_j) \\
&= v^T \Sigma^{\otimes 2} v \ .
\end{aligned}$$

This completes the proof. ∎

*Proof of Corollary* 4.9. Let $\mathfrak{S}_m = \{S \subseteq [N] : |S| = m\}$ denote the set of subsets of $[N]$ of size $m$. The same Bernstein-style analysis as in the proof of Lemma 4.3 yields that there exist universal constants $A, B$ so that

$$\Pr\left[\exists T \in \mathfrak{S}_m : \left\| \frac{1}{m} \sum_{i \in T} X_i X_I^\top - I \right\|_F \geq O\left(\delta_2 \frac{N}{m}\right)\right] \leq 4 \exp\left(\log \binom{N}{m} + Ad^2 - B\delta_2 N\right) .$$

Thus, union bounding over all $m \in \{1, \ldots, \varepsilon N\}$ yields that

$$\Pr\left[\exists T \text{ s.t. } |T| \leq \varepsilon N : \left\| \frac{1}{|T|} \sum_{i \in T} X_i X_I^\top - I \right\|_F \geq O\left(\delta_2 \frac{N}{|T|}\right)\right] \leq 4 \exp\left(\log(\varepsilon N) + \log \binom{N}{\varepsilon N} + Ad^2 - B\delta_2 n\right) \leq \tau ,$$

by the same manipulations as in the proof of Lemma 4.3. ∎

**A.1. Proof of Theorem 4.13.** This follows immediately from Lemmas 5.17 and 5.20.

## Appendix B. Deferred proofs from section 5.

### B.1. Proof of Lemma 5.3.

*Proof of Lemma* 5.3. Let $N = \Omega((d/\varepsilon^2) \operatorname{poly}\log(d/\varepsilon\tau))$ be the number of samples drawn from $G$. For (i), the probability that a coordinate of a sample is at least $\sqrt{2\nu \log(Nd/3\tau)}$ is at most $\tau/3dN$ by Fact 5.6. By a union bound, the probability that all coordinates of all samples are smaller than $\sqrt{2\nu \log(Nd/3\tau)}$ is at least $1-\tau/3$. In this case, $\|x\|_2 \leq \sqrt{2\nu d \log(Nd/3\tau)} = O(\sqrt{d\nu \log(N\nu/\tau)})$.

After translating by $\mu^G$, we note that (iii) follows immediately from Lemma 2.21 and (iv) follows from Theorem 5.50 of [Ver10], as long as $N = \Omega(\nu^4 d \log(1/\tau)/\varepsilon^2)$, with probability at least $1 - \tau/3$. It remains to show that, conditioned on (i), (ii) holds with probability at least $1 - \tau/3$.

To simplify some expressions, let $\delta := \varepsilon/(\log(d \log d/\varepsilon\tau))$ and $R = C\sqrt{d \log(|S|/\tau)}$. We need to show that for all unit vectors $v$ and all $0 \leq T \leq R$ that

$$\left| \Pr_{X \in_u S}[|v \cdot (X - \mu^G)| > T] - \Pr_{X \sim G}[|v \cdot (X - \mu^G) > T \geq 0] \right| \leq \frac{\delta}{T^2} . \tag{46}$$

First, we show that for all unit vectors $v$ and $T > 0$,

$$\left| \Pr_{X \in_u S}[|v \cdot (X - \mu^G)| > T] - \Pr_{X \sim G}[|v \cdot (X - \mu^G)| > T \geq 0] \right| \leq \frac{\delta}{10\nu \ln(1/\delta)}$$

with probability at least $1 - \tau/6$. Since the Vapnik–Chervonenkis dimension (VC dimension) of the set of all halfspaces is $d + 1$, this follows from the VC inequality [DL01], since we have more than $\Omega(d/(\delta/(10\nu \log(1/\delta))^2))$ samples. We thus need only consider the case when $T \geq \sqrt{10\nu \ln(1/\delta)}$.

LEMMA B.1. *For any fixed unit vector $v$ and $T > \sqrt{10\nu \ln(1/\delta)}$, except with probability* $\exp(-N\delta/(6C\nu))$, *we have that*

$$\Pr_{X \in_u S}[|v \cdot (X - \mu^G)| > T] \leq \frac{\delta}{CT^2} ,$$

*where* $C = 8$.

*Proof.* Let $E$ be the event that $|v \cdot (X - \mu^G)| > T$. Since $G$ is sub-Gaussian, Fact 5.6 yields that $\Pr_G[E] = \Pr_{Y\sim G}[|v \cdot (X - \mu^G)| > T] \le \exp(-T^2/(2\nu))$. Note that, thanks to our assumption on $T$, we have that $T \le \exp(T^2/(4\nu))/2C$, and therefore, $T^2 \Pr_G[E] \le \exp(-T^2/(4\nu))/2C \le \delta/2C$.

Consider $\mathbb{E}_S[\exp(t^2/(3\nu) \cdot N \Pr_S[E])]$. Each individual sample $X_i$ for $1 \le i \le N$ is an independent copy of $Y \sim G$, and hence,

$$\begin{aligned}\mathop{\mathbb{E}}_S\left[\exp\left(\frac{T^2}{3\nu} \cdot N \Pr_S[E]\right)\right] &= \mathop{\mathbb{E}}_S\left[\exp\left(\frac{T^2}{3\nu}\right) \cdot \sum_{i=1}^n 1_{X_i \in E}\right] \\ &= \prod_{i=1}^N \mathop{\mathbb{E}}_{X_i}\left[\exp\left(\frac{T^2}{3\nu}\right) \cdot \sum_{i=1}^n 1_{X_i \in E}\right] \\ &= \left(\exp\left(\frac{T^2}{3\nu}\right) \Pr_G[G] + 1\right)^N \\ &\overset{(a)}{\le} \left(\exp\left(\frac{T^2}{6\nu}\right) + 1\right)^N \\ &\overset{(b)}{\le} (1 + \delta^{5/3})^N \\ &\overset{(c)}{\le} \exp(N\delta^{5/3}) ,\end{aligned}$$

where (a) follows from sub-Gaussianity, (b) follows from our choice of $T$, and (c) comes from the fact that $1 + x \le e^x$ for all $x$.

Hence, by Markov's inequality, we have

$$\begin{aligned}\Pr\left[\Pr_S[E] \ge \frac{\delta}{CT^2}\right] &\le \exp\left(N\delta^{5/3} - \frac{\delta N}{3C}\right) \\ &= \exp(N\delta(\delta^{2/3} - 1/(3C))) .\end{aligned}$$

Thus, if $\delta$ is a sufficiently small constant and $C$ is sufficiently large, this yields the desired bound. ∎

Now let $\mathcal{C}$ be a 1/2-cover in Euclidean distance for the set of unit vectors of size $2^{O(d)}$. By a union bound, for all $v' \in \mathcal{C}$ and $T'$ a power of 2 between $\sqrt{4\nu \ln(1/\delta)}$ and $R$, we have that

$$\Pr_{X\in_u S}[|v' \cdot (X - \mu^G)| > T'] \le \frac{\delta}{8T^2}$$

except with probability

$$2^{O(d)} \log(R) \exp(-N\delta/6C\nu) = \exp\left(O(d) + \log\log R - N\delta/6C\nu\right) \le \tau/6 .$$

However, for any unit vector $v$ and $\sqrt{4\nu \ln(1/\delta)} \le T \le R$, there are a $v' \in \mathcal{C}$ and a $T'$ such that for all $x \in \mathbb{R}^d$, we have $|v \cdot (X - \mu^G)| \ge |v' \cdot (X - \mu^G)|/2$, and so $|v' \cdot (X - \mu^G)| > 2T'$ implies $|v' \cdot (X - \mu^G)| > T$.

Then, by a union bound, (46) holds simultaneously for all unit vectors $v$ and all $0 \le T \le R$, with probability of at least $1 - \tau/3$. This completes the proof. ∎

## B.2. Proof of Lemma 5.16.

*Proof of Lemma* 5.16. Note that an even polynomial has no degree-1 terms. Thus, we may write $p(x) = \sum_i p_{i,i}x_i^2 + \sum_{i>j} p_{i,j}x_ix_j + p_o$. Taking $(P_2)_{i,i} = p_{i,i}$ and $(P_2')_{i,j} = (P_2')_{j,i} = \frac{1}{2}p_{i,j}$, for $i > j$, gives $p(x) = x^T P_2' x + p_0$. Taking $P_2 = \Sigma^{1/2} P_2' \Sigma^{1/2}$, we have $p(x) = (\Sigma^{-1/2}x)^T P_2 (\Sigma^{-1/2}x) + p_0$, for a $d \times d$ symmetric matrix $P_2$ and $p_0 \in \mathbb{R}$.

Let $P_2 = U^T \Lambda U$, where $U$ is orthogonal and $\Lambda$ is diagonal, be an eigendecomposition of the symmetric matrix $P_2$. Then, $p(x) = (U\Sigma^{-1/2}x)^T P_2 (U\Sigma^{-1/2}x)$. Let $X \sim G$ and $Y = U\Sigma^{-1/2}X$. Then, $Y \sim \mathcal{N}(0, I)$ and $p(X) = \sum_i \lambda_i Y_i^2 + p_0$ for independent Gaussians $Y_i$. Thus, $p(X)$ follows a generalized $\chi^2$-distribution.

Thus, we have

$$\mathbb{E}[p(X)] = \mathbb{E}\left[\sum_i \lambda_i Y_i^2 + p_0\right] = p_0 + \sum_i \lambda_i = p_0 + \mathrm{tr}(P_2)$$

and

$$\mathrm{Var}[p(X)] = \mathrm{Var}\left[\sum_i \lambda_i Y_i^2 + p_0\right] = 2\sum_i \lambda_i^2 = 2\|P_2\|_F^2 \ .$$

Lemma B.2 (cf. Lemma 1 from [LM00]). *Let $Z = \sum_i a_i Y_i^2$, where $Y_i$ are independent random variables distributed as $\mathcal{N}(0,1)$. Let $a$ be the vector with coordinates $a_i$. Then,*

$$\Pr(Z \geq 2\|a\|_2\sqrt{x} + 2\|a\|_\infty x) \leq \exp(-x) \ .$$

We thus have

$$\Pr\left(\sum_i \lambda_i(Y_i^2 - 1) > 2\sqrt{\left(\sum_i \lambda_i^2\right)t} + 2\left(\max_i \lambda_i\right)t\right) \leq e^{-t} \ .$$

Noting that $\mathrm{tr}(P_2) = \sum_i \lambda_i, \sum_i \lambda_i^2 = \|P_2\|_F$, and $\max_i \lambda_i = \|P_2\|_2 \leq \|P_2\|$, for $\mu_p = \mathbb{E}[p(X)]$ we have

$$\Pr(p(X) - \mu_p > 2\|P_2\|_F(\sqrt{t} + t)) \leq e^{-t} \ .$$

Noting that $2\sqrt{a} = 1 + a - (1 - \sqrt{a})^2 \leq 1 + a$ for $a > 0$, we have

$$\Pr(p(X) - \mu_p > \|P_2\|_F(3t + 1)) \leq e^{-t} \ .$$

Applying this inequality on both $p(x)$ and $-p(x)$ simultaneously, we get

$$\Pr(|p(X) - \mu_p| > \|P_2\|_F(3t + 1)) \leq 2e^{-t} \ .$$

Substituting $t = T/3\|P_2\|_F - 1/3$ and $2\|P_2\|_F^2 = \mathrm{Var}_{X\sim G}(p(X))$ gives

$$\Pr(|p(X) - \mathop{\mathbb{E}}_{X\sim G}[p(X)]| \geq T) \leq 2e^{1/3 - 2T/3\,\mathrm{Var}_{X\sim G}[p(X)]} \ .$$

The final property is a consequence of the following anticoncentration inequality.

Theorem B.3 (see [CW01]). *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a degree-$d$ polynomial. Then, for $X \sim \mathcal{N}(0, I)$, we have*

$$\Pr|p(X)| \leq \varepsilon\sqrt{\mathbb{E}[p(X)^2]} \leq O(d\varepsilon^{1/d}) \ .$$

This completes the proof. ∎

### B.3. Proof of Lemma 5.17.

*Proof of Lemma* 5.17. First, we note that it suffices to prove this for the case $\Sigma = I$, since for $X \sim \mathcal{N}(0, \Sigma)$, $Y = \Sigma^{-1/2} X$ is distributed as $\mathcal{N}(0, I)$, and all the conditions transform to those for $G = \mathcal{N}(0, I)$ under this transformation.

Condition 1 of Definition 5.15 follows by standard concentration bounds on $\|x\|_2^2$. Condition 2 of Definition 5.15 follows by estimating the entrywise error between $\mathrm{Cov}(S)$ and $I$. These two conditions hold by Lemma 5.3, since they follow from (i), (iii), and (iv) of $(\varepsilon, \tau)$-goodness in the sense of Definition 5.2.

Condition 3 of Definition 5.15 is slightly more involved. Let $\{p_i\}$ be an orthonormal basis for the set of even, degree-2, mean-0 polynomials with respect to $G$. Define the matrix $M_{i,j} = \mathbb{E}_{x \in_u S}[p_i(x) p_j(x)] - \delta_{i,j}$. This condition is equivalent to $\|M\|_2 = O(\varepsilon)$. Thus, it suffices to show that for every $v$ with $\|v\|_2 = 1$ we have $v^T M v = O(\varepsilon)$. It actually suffices to consider a cover of such $v$'s. Note that this cover will be of size $2^{O(d^2)}$. For each $v$, let $p_v = \sum_i v_i p_i$. We need to show that $\mathrm{Var}(p_v(S)) = 1 + O(\varepsilon)$. We can show this happens with probability $1 - \tau 2^{-\Omega(d^2)}$, and thus it holds for all $v$ in our cover by a union bound.

Condition 4 of Definition 5.15 is substantially the most difficult of these conditions to prove. Naively, we would want to find a cover of all possible $p$ and all possible $T$, and bound the probability that the desired condition fails. Unfortunately, the best a priori bounds on $\Pr(|p(G)| > T)$ are on the order of $\exp(-T)$. As our cover would need to be of size $2^{d^2}$ or so, to make this work with $T = d$, we would require $\Omega(d^3)$ samples in order to make this argument work.

However, we will note that this argument is sufficient to cover the case of $T < 10 \log(1/\varepsilon) \log^2(d/\varepsilon)$.

Fortunately, most such polynomials $p$ satisfy much better tail bounds. Note that any even, mean-0 polynomial $p$ can be written in the form $p(x) = x^T A x - \mathrm{tr}(A)$ for some matrix $A$. We call $A$ the associated matrix to $p$. We note by the Hanson–Wright inequality that $\Pr(|p(G)| > T) = \exp(-\Omega(\min((T/\|A\|_F)^2, T/\|A\|_2)))$. Therefore, the tail bounds above are only as bad as described when $A$ has a single large eigenvalue. To take advantage of this, we will need to break $p$ into parts based on the size of its eigenvalues. We begin with a definition.

Definition B.4. *Let $\mathcal{P}_k$ be the set of even, mean-0, degree-2 polynomials, such that the associated matrix $A$ satisfies*

1. $\mathrm{rank}(A) \le k$,
2. $\|A\|_2 \le 1/\sqrt{k}$.

Note for $p \in \mathcal{P}_k$ that $|p(x)| \le |x|^2/\sqrt{k} + \sqrt{k}$.

Importantly, any polynomial can be written in terms of these sets.

Lemma B.5. *Let $p$ be an even, degree-2 polynomial with $\mathbb{E}[p(G)] = 0, \mathrm{Var}(p(G)) = 1$. Then if $t = \lfloor \log_2(d) \rfloor$, it is possible to write $p = 2(p_1 + p_2 + \cdots + p_{2^t} + p_d)$, where $p_k \in \mathcal{P}_k$.*

*Proof.* Let $A$ be the associated matrix to $p$. Note that $\|A\|_F = \mathrm{Var}\, p = 1$. Let $A_k$ be the matrix corresponding to the top $k$ eigenvalues of $A$. We now let $p_1$ be the polynomial associated to $A_1/2$, $p_2$ be associated to $(A_2 - A_1)/2$, $p_4$ be associated to $(A_4 - A_2)/2$, and so on. It is clear that $p = 2(p_1 + p_2 + \cdots + p_{2^t} + p_d)$. It is also clear that the matrix associated to $p_k$ has rank at most $k$. If the matrix associated to $p_k$ had an eigenvalue more than $1/\sqrt{k}$, it would need to be the case that the $k/2^{nd}$ largest eigenvalue of $A$ had size at least $2/\sqrt{k}$. This is impossible since the sum of the squares of the eigenvalues of $A$ is at most 1.

This completes our proof. ∎

We will also need covers of each of these sets $\mathcal{P}_k$. We will assume that Condition 1 of Definition 5.15 holds, i.e., that $\|x\|_2 \leq \sqrt{R}$, where $R = O(d \log(d/\varepsilon\tau))$. Under this condition, $p(x)$ cannot be too large and this affects how small a variance polynomial we can ignore.

LEMMA B.6. *For each $k$, there exists a set $\mathcal{C}_k \subset \mathcal{P}_k$ such that*

1. *For each $p \in \mathcal{P}_k$ there exists a $q \in \mathcal{C}_k$ such that* $\mathrm{Var}(p(G) - q(G)) \leq 1/R^2 d^2$.
2. $|\mathcal{C}_k| = 2^{O(dk \log R)}$.

*Proof.* We note that any such $p$ is associated to a matrix $A$ of the form $A = \sum_{i=1}^k \lambda_i v_i v_i^T$, for $\lambda_i \in [0, 1/\sqrt{k}]$ and $v_i$ orthonormal. It suffices to let $q$ correspond to the matrix $A' = \sum_{i=1}^k \mu_i w_i w_i^T$ with $|\lambda_i - \mu_i| < 1/R^2 d^3$ and $|v_i - w_i| < 1/R^2 d^3$ for all $i$. It is easy to let $\mu_i$ and $w_i$ range over covers of the interval and the sphere with appropriate errors. This gives a set of possible $q$'s of size $2^{O(dk \log R)}$ as desired. Unfortunately, some of these $q$ will not be in $\mathcal{P}_k$ as they will have eigenvalues that are too large. However, this is easily fixed by replacing each such $q$ by the closest element of $\mathcal{P}_k$. This completes our proof. ∎

Next, we will show that these covers are sufficient to express any polynomial.

LEMMA B.7. *Let $p$ be an even degree-2 polynomial with $\mathbb{E}[p(G)] = 0$ and $\mathrm{Var}(p(G)) = 1$. It is possible to write $p$ as a sum of $O(\log(d))$ elements of some $\mathcal{C}_k$ plus another polynomial of variance at most $O(1/R^2)$.*

*Proof.* Combining the above two lemmas we have that any such $p$ can be written as

$$p = (q_1 + p_1) + (q_2 + p_2) + \cdots + (q_{2^t} + p_{2^t}) + (q_d + p_d) = q_1 + q_2 + \cdots + q^{2^t} + q^d + p' ,$$

where $q_k$ above is in $\mathcal{C}_k$ and $\mathrm{Var}[p_k(G)] < 1/R^2 d^2$. Thus, $p' = p_1 + p_2 + \cdots + p_{2^t} + p_d$ has $\mathrm{Var}[p'(G)] \leq O(1/R^2)$. This completes the proof. ∎

The key observation now is that if $|p(x)| \geq T$ for $\|x\|_2 \leq \sqrt{d/\varepsilon}$, then writing $p = q_1 + q_2 + q_4 + \cdots + q_d + p'$ as above, it must be the case that $|q_k(x)| > (T-1)/(2\log(d))$ for some $k$. Therefore, to prove our main result, it suffices to show that, with high probability over the choice of $S$, for any $T \geq 10 \log(1/\varepsilon) \log^2(d/\varepsilon)$ and any $q \in \mathcal{C}_k$ for some $k$, that $\Pr_{x \in_u S}(|q(x)| > T/(2\log(d))) < \varepsilon/(2T^2 \log^2(T) \log(d))$. Equivalently, it suffices to show that for $T \geq 10 \log(1/\varepsilon) \log(d/\varepsilon)$ it holds that $\Pr_{x \in_u S}(|q(x)| > T/(2\log(d))) < \varepsilon/(2T^2 \log^2(T) \log^2(d))$. Note that this holds automatically for $T > R$, as $p(x)$ cannot possibly be that large for $\|x\|_2 \leq \sqrt{R}$. Furthermore, note that losing a constant factor in the probability, it suffices to show this only for $T$ a power of 2.

Therefore, it suffices to show for every $k \leq d$, every $q \in \mathcal{C}_k$, and every $R/\sqrt{k} \gg T \gg \log(1/\varepsilon) \log R$ that with probability at least $1 - \tau 2^{-\Omega(dk \log R)}$ over the choice of $S$ we have that $\Pr_{x \in_u S}(|q(x)| > T) \ll \varepsilon/(T^2 \log^4(R))$. However, by the Hanson–Wright inequality, we have that

$$\Pr(|q(G)| > T) = \exp(-\Omega(\min(T^2, T\sqrt{k}))) < (\varepsilon/(T^2 \log^4 R))^2 .$$

Therefore, by Chernoff bounds, the probability that more than an $\varepsilon/(T^2 \log^4 R)$-

fraction of the elements of $S$ satisfy this property is at most

$$\begin{aligned}&\exp(-\Omega(\min(T^2, T\sqrt{k}))|S|\varepsilon/(T^2\log^4 R))\\&= \exp(-\Omega(|S|\varepsilon/(\log^4 R)\min(1,\sqrt{k}/T)))\\&\leq \exp(-\Omega(|S|k\varepsilon^2/R(\log^4 R)))\\&\leq \exp(-\Omega(|S|k\varepsilon/d(\log(d/\varepsilon\tau))(\log^4(d/\log(1/\varepsilon\tau)))))\\&\leq \tau\exp(-\Omega(dk\log(d/\varepsilon)))\ ,\end{aligned}$$

as desired.

This completes our proof. ∎

**Appendix C. Deferred proofs from section 6.**

*Proof of Theorem* 6.8. The first two properties follow directly from (32). We now show the third property. Suppose this does not happen, that is, there are $j, j'$ such that $\ell = \ell(j) = \ell(j')$ such that $\|\mu_j - \mu_{j'}\|_2^2 \geq \Omega(dk\log k/\varepsilon)$. That means that by (32) there is some sequence of clusters $S_1, \ldots, S_t$ such that $S_i \cap S_{i+1} \neq \varnothing$ for each $i$, $|S_i| \geq 4\varepsilon N$ for each $i$, and moreover, there is an $X_i \in S_1$ such that $\|X_i - \mu_1\|_2^2 \leq O(d\log k/\varepsilon)$ and an $X_{i'} \in S_t$ such that $\|X_{i'} - \mu_2\|_2^2 \leq O(d\log k/\varepsilon)$. But by (32), we know that each $S_i$ contains a point $X_{i''}$ such that $\|X_{i''} - \mu_{r_i}\|_2^2 \leq O(d\log k/\varepsilon)$. In particular, by the triangle inequality, this means that if $\|\mu_{r_i} - \mu_{r_{i+1}}\|_2^2 \leq O(d\log k/\varepsilon)$ for all $i = 1, \ldots, t-1$, then we can set $\mu_{r_1} = \mu_j$ and $\mu_{r_t} = \mu_{j'}$.

Construct an auxiliary graph on $k$ vertices, where we put an edge between nodes $r_i$ and $r_{i+1}$. By the above, there must be a path from $j$ to $j'$ in this graph. Since this graph has $k$ nodes, there must be a path of length at most $k$ from $j$ to $j'$; moreover, by the above, we know that this implies that $\|\mu_j - \mu_{j'}\|_2^2 \leq O(kd\log k/\varepsilon)$.

Finally, the fourth property follows from the same argument as the proof of the third. ∎

*Proof of Lemma* 6.9. Let $C = \sum_{i=1}^N w_i(X_i - \mu)(X_i - \mu)^T - I$. Let $v$ be the top eigenvector of

$$\sum_{i=1}^N w_i(X_i - \mu)(X_i - \mu)^T - I - \sum_{j\in[k]} \alpha_j(\mu_j - \mu)(\mu_j - \mu)^T\ .$$

Observe that by (33), we have

$$\begin{aligned}\sum_{i=1}^N w_i(X_i - \mu)(X_i - \mu)^T &\succeq \sum_{i\in G} w_i(X_i - \mu)(X_i - \mu)^T\\&\succeq w_g(I+Q) - f(k,\gamma,\delta_1)I\\&\succeq (1-\varepsilon)(I+Q) - f(k,\gamma,\delta_1)I\ ,\end{aligned}$$

and so, in particular,

$$\sum_{i=1}^N w_i(X_i - \mu)(X_i - \mu)^T - (I+Q) \succeq -\varepsilon(I+Q) - f(k,\gamma,\delta_1)I\ .$$

Therefore, for any unit vector $u \in \mathbb{R}^d$, we must have

$$u^T \left( \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - (I + Q) \right) u \geq -\varepsilon u^T (I+Q) u^T - f(k, \gamma, \delta_1)$$
$$\geq -\frac{c}{2} h(k, \gamma, \delta) \; .$$

In particular, since $\left| v^T \left( \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - (I+Q) \right) v \right| \geq ckh(k, \gamma, \delta)$, we must have

$$v^T \left( \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - (I + Q) \right) v > 0,$$

and hence

$$v^T \left( \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - (I + Q) \right) v \geq ckh(k, \gamma, \delta) \; .$$

Let $U = [v, u_1, \ldots, u_{d-1}]$ be a $d \times k$ matrix with orthonormal columns, where the columns span the set of vectors $\{(\mu_j - \mu) \; : \; j \in [k]\} \cup \{v\}$. We note the rank of this set is at most $k$ due to the definition of $\mu$.

Using the dual characterization of the Schatten top-$k$ norm, we have that

$$\|C\|_{T_k} = \max_{X \in \mathbb{R}^{d \times k}} \mathrm{Tr}(X^T C X) \geq \mathrm{Tr}(U^T C U).$$

Observe that since $\mathrm{span}(Q) \subseteq \mathrm{span}(U)$, we have

$$\begin{aligned}
\|C\|_{T_k} \geq \mathrm{Tr}\left(U^T C U\right) &= \mathrm{Tr}\left( U^T \left( \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - (I+Q) \right) U \right) + \|Q\|_{T_k} \\
&= \mathrm{Tr}\left(U^T (C - Q) U\right) + \sum_{j \in [k]} \gamma_j \\
&= v^T (C - Q) v + \sum_{i=1}^{k-1} u_i^T (C - Q) u_i + \sum_{j \in [k]} \gamma_j \\
&\geq ckh(k, \gamma, \delta) - (k-1) \frac{c}{2} h(k, \gamma, \delta) + \sum_{j \in [k]} \gamma_j \\
&\geq \frac{c}{2} kh(k, \gamma, \delta) + \sum_{j \in [k]} \gamma_j \; ,
\end{aligned}$$

as claimed. ◻

*Proof of Lemma* 6.10. By Fact 4.2 and (34) we have $\| \sum_{i=G} \frac{w_i}{w_g} X_i - \mu \|_2 \leq k^{1/2} \delta_2$. Now, by the triangle inequality, we have

$$\left\| \sum_{i \in E} w_i (X_i - \mu) \right\|_2 \geq \|\Delta\|_2 - \left\| \sum_{i \in G} w_i (X_i - \mu) - w_g \mu \right\|_2 \geq \Omega(\|\Delta\|_2).$$

Using the fact that variance is nonnegative, we have

$$\sum_{i \in E} \frac{w_i}{w_b} (X_i - \mu)(X_i - \mu)^T \succeq \left( \sum_{i \in E} \frac{w_i}{w_b} \left(X_i - \mu\right) \right) \left( \sum_{i \in E} \frac{w_i}{w_b} \left(X_i - \mu\right) \right)^T \; ,$$

and therefore,

$$\left\| \sum_{i \in E} w_i (X_i - \mu)(X_i - \mu)^T \right\|_2 \geq \Omega\left( \frac{\|\Delta\|_2^2}{w_b} \right) \geq \Omega\left( \frac{\|\Delta\|_2^2}{\varepsilon} \right).$$

On the other hand,

$$\left\| \sum_{i \in G} w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 \leq \left\| \sum_{i \in G} w_i (X_i - \mu)(X_i - \mu)^T - w_g I \right\|_2 + w_b \leq f(k, \gamma, \delta_1) + w_b,$$

where in the last inequality we have used Fact 4.2 and (33). Hence, altogether this implies that

$$\left\| \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I \right\|_2 \geq \Omega\left( \frac{\|\Delta\|_2^2}{\varepsilon} \right) - w_b - f(k, \gamma, \delta_1) \geq \Omega\left( \frac{\|\Delta\|_2^2}{\varepsilon} \right) ,$$

as claimed. ∎

**C.1. Proof of Theorem 6.12.** Once more, let $\Delta = \mu - \hat{\mu}$ and expand the formula for $M$:

$$\begin{aligned} \sum_{i=1}^N w_i Y_i Y_i^T - I &= \sum_{i=1}^N w_i (X_i - \mu + \Delta)(X_i - \mu + \Delta)^T - I \\ &= \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I + \sum_{i=1}^N w_i (X_i - \mu) \Delta^T \\ &\quad + \Delta \sum_{i=1}^N w_i (X_i - \mu)^T + \Delta \Delta^T \\ &= \sum_{i=1}^N w_i (X_i - \mu)(X_i - \mu)^T - I - \Delta \Delta^T . \end{aligned}$$

We start by proving completeness.

CLAIM C.1. *Suppose that* $w = w^*$. *Then* $\|M\|_{T_k} \leq \sum_{i \in [k]} \gamma_j + \frac{ckh(k,\gamma,\delta_1)}{2}$.

*Proof.* $w^*$ are the weights that are uniform on the uncorrupted points. Because $E \leq 2\varepsilon N$, we have that $w^* \in S_{N,\varepsilon}$. Using (33), this implies that $w^* \in \mathcal{C}_{f(k,\gamma,\delta_1)}$. By Corollary 6.11, $\|\Delta\|_2 \leq O(\varepsilon \sqrt{\log 1/\varepsilon})$. With this in hand, we proceed by taking the

Schatten top-$k$ norm of the above expansion of $M$.

$$\begin{aligned}
&\left\| \sum_{i=1}^{m} w_i^*(X_i-\mu)(X_i-\mu)^T - I - \Delta\Delta^T \right\|_{T_k} \\
&\le \left\| \sum_{i=1}^{N} w_i^*(X_i-\mu)(X_i-\mu)^T - I - \sum_{j\in[k]} \alpha_j(\mu_j-\mu)(\mu_j-\mu)^T \right\|_{T_k} \\
&\quad + \left\| \sum_{j\in[k]} \alpha_i(\mu_j-\mu)(\mu_j-\mu)^T \right\|_{T_k} + \|\Delta\Delta^T\|_2 \\
&\le kf(k,\gamma,\delta_1) + \sum_{j\in[k]} \gamma_j + O(\varepsilon^2 \log 1/\varepsilon) \\
&< \sum_{j\in[k]} \gamma_j + \frac{ckh(k,\gamma,\delta)}{2} . \qquad \square
\end{aligned}$$

CLAIM C.2. *Suppose that* $w \notin \mathcal{C}_{ckh(k,\gamma,\delta)}$. *Then* $\|M\|_{T_k} > \sum_{i\in[k]} \gamma_j + \frac{ckh(k,\gamma,\delta_1)}{2}$.

*Proof.* We split the proof into two cases. In the first case, $\|\Delta\|_2^2 \le \frac{ckh(k,\gamma,\delta)}{10}$. By Lemma 6.9, we have that

$$\left\| \sum_{i=1}^{N} w_i(X_i-\mu)(X_i-\mu)^T - I \right\|_{T_k} \ge \sum_{j\in[k]} \gamma_j + \frac{3ckh(k,\gamma,\delta)}{4}.$$

By the triangle inequality,

$$\|M\|_{T_k} \ge \sum_{j\in[k]} \gamma_j + \frac{3ckh(k,\gamma,\delta)}{4} - \|\Delta\|_2^2 \ge \sum_{i\in[k]} \gamma_j + \frac{ckh(k,\gamma,\delta)}{2},$$

as desired.

In the other case, $\|\Delta\|_2^2 \ge \frac{ckh(k,\gamma,\delta)}{10}$. Recall that $Q = \sum_{j\in[k]} \alpha_j(\mu_j-\mu)(\mu_j-\mu)^T$ from (29). Write $M$ as follows:

$$\begin{aligned}
M &= \sum_{i=1}^{N} w_i(X_i-\mu)(X_i-\mu)^T - I - \Delta\Delta^T \\
&= \left( \sum_{i\in G} w_i(X_i-\mu)(X_i-\mu)^T - w_g I - w_g Q \right) \\
&\quad + w_g Q + \sum_{i\in E} w_i(X_i-\mu)(X_i-\mu)^T - w_b I - \Delta\Delta^T.
\end{aligned}$$

Now taking the Schatten top-$k$ norm of $M$, we have

$$
\begin{aligned}
&\left\| \left( \sum_{i \in G} w_i (X_i - \mu)(X_i - \mu)^T - w_g I - w_g Q \right) + w_g Q + \sum_{i \in E} w_i (X_i - \mu)(X_i - \mu)^T \right. \\
&\left. \quad - w_b I - \Delta\Delta^T \right\|_{T_k} \\
&\geq \left\| w_g Q + \sum_{i \in E} w_i (X_i - \mu)(X_i - \mu)^T \right\|_{T_k} - \left\| \sum_{i \in G} w_i (X_i - \mu)(X_i - \mu)^T - w_g I - w_g Q \right\|_{T_k} \\
&\quad - \|w_b I\|_2 - \left\|\Delta\Delta^T\right\|_2 \\
&\geq \left\| w_g Q + \sum_{i \in E} w_i (X_i - \mu)(X_i - \mu)^T \right\|_{T_k} - kf(k, \gamma, \delta_1) - 4\varepsilon - \|\Delta\|_2^2 \\
&\geq \left( \sum_{j \in [k]} \gamma_j - 4\varepsilon k \gamma \right) + \left\| \sum_{i \in E} w_i (X_i - \mu)(X_i - \mu)^T \right\|_{T_k} - kf(k, \gamma, \delta_1) - 4\varepsilon - \|\Delta\|_2^2 \\
&\text{(47)} \\
&\geq \sum_{j \in [k]} \gamma_j + \Omega\left( \frac{\|\Delta\|_2^2}{\varepsilon} \right) \\
&\geq \sum_{j \in [k]} \gamma_j + \frac{ckh(k, \gamma, \delta)}{2}.
\end{aligned}
$$

The first inequality is the triangle inequality, the second is by (33) and Fact 4.2, the third is because the summed matrices are positive semidefinite, the fourth follows from Lemma 6.10, and the fifth holds for all $c$ sufficiently large. ∎

By construction, we have that $\ell(w) \geq 0$. It remains to show that $\ell(w^*) < 0$:

$$
\begin{aligned}
&\left\| \frac{1}{|G|} \sum_{i \in G} (X_i - \hat{\mu})(X_i - \hat{\mu})^T - I \right\|_{T_k} \\
&= \left\| \frac{1}{|G|} \sum_{i \in G} (X_i - \mu + \Delta)(X_i - \mu + \Delta)^T - I \right\|_{T_k} \\
&\leq \left\| \frac{1}{|G|} \sum_{i \in G} (X_i - \mu)(X_i - \mu)^T - I - \sum_{j \in [k]} \alpha_j (\mu_j - \mu)(\mu_j - \mu)^T \right\|_{T_k} \\
&+ \left\| \sum_{j \in [k]} \alpha_j (\mu_j - \mu)(\mu_j - m)^T \right\|_{T_k} + 2\|\Delta\|_2 \left\| \frac{1}{|G|} \sum_{i \in G} (X_i - \mu) \right\|_{T_k} + \|\Delta\|_2^2 \\
&\leq kf(k, \gamma, \delta_1) + \sum_{j \in [k]} \gamma_j + 2k^{1/2}\delta_2 \|\Delta\|_2 + \|\Delta\|_2^2.
\end{aligned}
$$

Therefore,

$$
\ell(w^*) \leq kf(k, \gamma, \delta) + \sum_{j \in [k]} \gamma_j + 2k^{1/2}\delta \|\Delta\|_2 + \|\Delta\|_2^2 - \Lambda.
$$

If $\|\Delta\|_2^2 \le \frac{ckh(k,\gamma,\delta)}{10}$, then

$$\ell(w^*) \le \sum_{j\in[k]} \gamma_j + kh(k,\gamma,\delta) + \frac{2k\delta\sqrt{ch(k,\gamma,\delta)}}{\sqrt{10}} + \frac{ckh(k,\gamma,\delta)}{10} - \Lambda.$$

We wish to show that

$$\frac{2k\delta\sqrt{ch(k,\gamma,\delta)}}{\sqrt{10}} \le \frac{ckh(k,\gamma,\delta)}{10},$$

or, equivalently, that

$$\delta \le \frac{\sqrt{ch(k,\gamma,\delta)}}{2\sqrt{10}}.$$

But this is true for $c$ sufficiently large, as $\sqrt{h(k,\gamma,\delta)} \ge \sqrt{\delta}$. Therefore,

$$\ell(w^*) \le \sum_{j\in[k]} \gamma_j + \frac{(c+5)kh(k,\gamma,\delta)}{5} - \Lambda \le 0,$$

where the second inequality holds since $\Lambda > \sum_{j\in[k]} \gamma_j + \frac{ckh(k,\gamma,\delta)}{2}$.

On the other hand, consider when $\|\Delta\|_2^2 \ge \frac{ckh(k,\gamma,\delta)}{10}$. By (47), we know that

$$\Lambda \ge \sum_{j\in[k]} \gamma_j + \Omega\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right).$$

Then we know

$$\ell(w^*) \le kf(k,\gamma,\delta) + 2k^{1/2}\delta\|\Delta\|_2 + \|\Delta\|_2^2 - \Omega\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right).$$

The first and third terms are immediately dominated by $\Omega\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right)$; it remains to show that

$$k^{1/2}\delta\|\Delta\|_2 = O\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right).$$

Or, equivalently, $k^{1/2}\delta\varepsilon = O\left(\|\Delta\|_2\right)$. This follows since

$$\|\Delta\|_2 \ge O(\sqrt{h(k,\gamma,\delta)}) \ge O(\sqrt{k\delta^2}) = O(k^{1/2}\delta\varepsilon).$$

Therefore, in this case as well, $\ell(w^*) < 0$, as desired.

**Appendix D. Deferred proofs from section 7.**

*Proof of Lemma* 7.21. By Lemma 7.6, applied with $\varepsilon' := \varepsilon^{3/2}/10d$ in place of $\varepsilon$, since we have $\Omega(d^4\log(1/\tau)/\varepsilon'^2)$ samples from $P$, with probability of at least $1-\tau$, the set $S$ is such that for all affine functions $L$, it holds that $|\Pr_{X\in_u S}(L(X) \ge 0) - \Pr_{X\sim P}(L(X)\ge 0)| \le \varepsilon'/d$. We henceforth condition on this event.

Let $C_T$ be the event that all coordinates in $T$ take their most common value. For a single coordinate $i$, the probability that it does not take its most common value, $\min\{p_i, 1-p_i\}$, satisfies

$$\min\{p_i,1-p_i\} = p_i(1-p_i)/\max\{p_i,1-p_i\} \le 2p_i(1-p_i).$$

Thus, by a union bound, we have that $\Pr_P(C_T) \geq 3/5$. Let $\#_T(x)$ be the number of coordinates of $x$ in $T$ which do not have their most common value, and observe that $\#_T(x)$ is an affine function of $x$. Noting that for $x \in \{0,1\}^d$, we have that $1-\#_T(x) > 0$ if and only if $C_T$ holds for $x$, it follows that $|\Pr_S(C_T)-\Pr_P[C_T]| \leq \varepsilon'/d$. Hence, we have that $\Pr_S(C_T) \geq 1/2$.

For any affine function $L(x)$, let

$$L_T(x) = L(x) - \#_T(x) \cdot \max_{y\in\{0,1\}^d} L(y).$$

Note that for $x \in \{0,1\}^d$, we have that $L_T(x) > 0$ if and only if $L(x) > 0$ and $C_T$ holds for $x$. Therefore, we can write

$$\begin{aligned}
&\left|\Pr_{X\in_u S}(L(X) > 0) - \Pr_{X\sim P}(L(X) > 0)\right| \\
&= \left|\frac{\Pr_{X\in_u S}(L_T(X) > 0)}{\Pr_{X\in_u S}(C_T)} - \frac{\Pr_{X\sim P}(L_T(X) > 0)}{\Pr_{X\sim P}(C_T)}\right| \\
&= \frac{|\Pr_{X\in_u S}(L_T(X) > 0)\Pr_{X\sim P}(C_T) - \Pr_{X\sim P}(L_T(X) > 0)\Pr_{X\in_u S}(C_T)|}{\Pr_{X\in_u S}(C_T)\Pr_{X\sim P}(C_T)} \\
&\leq (10/3)\cdot\Big(\Pr_{X\in_u S}(L_T(X) > 0)\cdot\Big(\Pr_{X\sim P}(C_T) - \Pr_{X\in_u S}(C_T)\Big) \\
&- \Pr_{X\in_u S}(C_T)\Big(\Pr_{X\sim P}(L_T(X) > 0) - \Pr_{X\in_u S}(L_T(X) > 0)\Big)\Big) \\
&\leq (10/3)\cdot 2\varepsilon'/d \leq \varepsilon^{3/2}/d^2\ .
\end{aligned}$$

This completes the proof of Lemma 7.21. ∎

## Appendix E. Deferred proofs from section 8.

*Proof of Lemma* 8.6. Let $S_P \subseteq S$ be the set of samples drawn from $P$, and let $S_Q \subseteq S$ be the set of samples drawn from $Q$. First, we note that by a Chernoff bound, $||S_P|/|S| - \alpha| \leq O(\varepsilon/d^2)$ with probability $1-\tau/3$. Assuming this holds, it follows that $|S_P| \geq (\alpha/2)|S| \geq (\varepsilon^{1/6}/2)|S| = \Omega(d^4\log(1/\tau)/\varepsilon^2)$. Similarly, $|S_Q| \geq (1-\alpha)|S|/2 \geq \Omega(d^4\log(1/\tau)/\varepsilon^2)$. By Lemma 7.6 applied with $\varepsilon' := (c^2/4)\cdot\varepsilon$ in place of $\varepsilon$, since we have $\Omega((d^4 + d^2\log(\tau))/\varepsilon'^2)$ samples, with probability $1-\tau/3$, the set $S_P$ is $\varepsilon'$-good for $P$, i.e., it satisfies that for all affine functions $L$, $|\Pr_{X\in_u S_P}(L(X) > 0) - \Pr_{X\sim P}(L(X) > 0)| \leq \varepsilon'/d$. We show that assuming $S$ is $\varepsilon'$-good, it is $(\varepsilon, i)$-good for each $1 \leq i \leq d$.

Note that $\Pr_{X\sim P}[X_i = 1] = p_i \geq c$ and $\Pr_{X\sim P}[X_i = 0] = 1-p_i \geq c$. Since $|\Pr_{X\sim P}[X_i = 1] - \Pr_{X\in_u S_P}[X_i = 1]| \leq c^2\varepsilon/(4d)$, it follows that $\Pr_{X\in_u S}[X_i = 1] \geq c/2$. For any affine function $L$, define $L^{(0)}(x) := L(x) - x_i(\max_y |L(y)|)$ and $L^{(1)}(x) :=$

$L(x) - (1 - x_i)(\max_y |L(y)|)$. Then, we have the following:

$$
\begin{aligned}
&\left| \Pr_{X \in_u S_P^1} (L(X) > 0) - \Pr_{X \sim P} (L(X) > 0 \mid X_i = 1) \right| \\
&= \left| \frac{\Pr_{X \in_u S_P^1} \left(L^1(X) > 0\right)}{\Pr_{X \in_u S_P^1} (X_i > 0)} - \frac{\Pr_{X \sim P} \left(L^1(X) > 0\right)}{p_i} \right| \\
&\le (2/c^2) \left( \Pr_{X \in_u S_P^1} \left(L^1(X) > 0\right) p_i - \Pr_{X \sim P} \left(L^1(X) > 0\right) \Pr_{X \in_u S_1} (X_i > 0) \right) \\
&\le (2/c^2) \left( p_i \left( \Pr_{X \in_u S_P^1} \left(L^1(X) > 0\right) - \Pr_{X \sim P} \left(L^1(X) > 0\right) \right) \right. \\
&\qquad \left. - \Pr_{X \sim P} \left(L^1(X) > 0\right) \left( \Pr_{X \in_u S_P^1} (X_i > 0) - p_i \right) \right) \\
&\le 2/c^2 \cdot 2\varepsilon'/d \le \varepsilon/d \,.
\end{aligned}
$$

Similarly, we obtain that

$$
\left| \Pr_{X \in_u S_1} (L(X) > 0) - \Pr_{X \sim \Pi} (L(X) > 0) \right| \le \varepsilon/d \,.
$$

So, we have that $S_P$ is $(\varepsilon, i)$-good for $P$ for all $1 \le i \le d$ with probability $1 - \tau/3$ . Similarly, $S_Q$ is $(\varepsilon, i)$-good for $Q$ for all $1 \le i \le d$ with probability $1 - \tau/3$. Thus, we have that $||S_P|/|S| - \alpha| \le \varepsilon/d^2$, $S_P$ is $(\varepsilon, i)$-good for $P$, and $S_Q$ is $(\varepsilon, i)$-good for $Q$ for all $1 \le i \le d$ with probability $1 - \tau$. That is, $S$ is $(\varepsilon, i)$-good for $\Pi$ for all $1 \le i \le d$ with probability at least $1 - \tau$. ∎

*Proof of Lemma* 8.19. Let $S_P \subseteq S$ be the set of samples drawn from $P$, and let $S_Q \subseteq S$ be the set of samples drawn from $Q$. First, we note that by a Chernoff bound, $||S_P|/|S| - \alpha| \le O(\varepsilon/d^2)$ with probability at least $1 - \tau/3$. Assuming this holds, $|S_P| \ge (\alpha/2)|S| \ge \delta|S| = \Omega(d^4 \log(1/\tau)/\varepsilon^2)$. Similarly, $|S_Q| \ge (1 - \alpha)|S|/2 \ge \Omega(d^4 \log(1/\tau)/\varepsilon^2)$.

By Lemma 7.6 applied with $\varepsilon' := \varepsilon/6$, since we have $\Omega(d^4 \log(1/\tau)/\varepsilon'^2)$ samples, with probability at least $1-\tau/3$, the set $S_P$ is $\varepsilon$-good for $P$. Similarly, with probability at least $1 - \tau/3$, the set $S_Q$ is $\varepsilon$-good for $Q$. Thus, with probability $1 - \tau$, we have that $\left|\frac{|S_P|}{|S|} - \alpha\right| \le \varepsilon$ and that $S_P$ and $S_Q$ are $\varepsilon$-good for $P$ and $Q$, respectively. ∎

*Proof of Lemma* 8.22. Noting that the mean of $T$ is $\mu$ and $|T| = |S'|$, we have

$$
\begin{aligned}
|S'|\mathrm{Cov}(S') &= |S'_P| \mathop{\mathbb{E}}_{X \in_u S'_P}[(X - \mu)(X - \mu)^T] + |S'_Q| \mathop{\mathbb{E}}_{X \in_u S'_Q}[(X - \mu)(X - \mu)^T] \\
&\quad + |E| \mathop{\mathbb{E}}_{X \in_u E}[(X - \mu)(X - \mu)^T] \\
&= |S'_P| \left( \mathrm{Cov}(S'_P) + (\mu^{S'_P} - \mu)(\mu^{S'_P} - \mu)^T \right) \\
&\quad + |S'_Q| \left( \mathrm{Cov}(S'_P) + (\mu^{S'_Q} - \mu)(\mu^{S'_Q} - \mu)^T \right) \\
&\quad + |E| \mathop{\mathbb{E}}_{X \in_u E}[(X - \mu)(X - \mu)^T] \\
&= |S'_P|\mathrm{Cov}(S'_P) + |S'_Q|\mathrm{Cov}(S'_Q) + |S'|\mathrm{Cov}(T) \,. \qquad (48)
\end{aligned}
$$

Since $P$ and $Q$ are product distributions, $\mathrm{Cov}(S'_P)$ and $\mathrm{Cov}(S'_Q)$ can have large diagonal elements but small off-diagonal ones. On the other hand, we bound the elements on the diagonal of $\mathrm{Cov}(T)$, but $\|\mathrm{Cov}(T)\|_2$ may still be large due to off-diagonal elements.

By the triangle inequality, and (48) with zeroed diagonal, we have

$$\begin{aligned}\|\mathrm{Cov}_0(S') - \mathrm{Cov}(T)\|_2 &\le \|\mathrm{Cov}_0(S') - \mathrm{Cov}_0(T)\|_2 + \|\mathrm{Cov}_0(T) - \mathrm{Cov}(T)\|_2 \\ &\le \left(\frac{|S'_P|}{|S'|}\right)\|\mathrm{Cov}_0(S'_P)\|_2 + \left(\frac{|S'_Q|}{|S'|}\right)\|\mathrm{Cov}_0(S'_Q)\|_2 \\ &\quad + \|\mathrm{Cov}_0(T) - \mathrm{Cov}(T)\|_2 \,. \end{aligned} \tag{49}$$

We will bound each of these terms separately.

Note that $\mathrm{Cov}_0(T) - \mathrm{Cov}(T)$ is a diagonal matrix and its nonzero entries are

$$(\mathrm{Cov}_0(T) - \mathrm{Cov}(T))_{i,i} = \operatorname*{Var}_{X\in_u T}[X_i].$$

Since the mean of $T$ is $\mu$, for all $i$, we have that $\mathrm{Var}_{X\in_u T}[X_i] \le \mathbb{E}_{X\in_u T}[\|X-\mu\|_\infty^2]$. We seek to bound the RHS from above.

Note that $\mu$ satisfies $|S'|\mu = |S'_P|\mu^{S'_P} + |S'_Q|\mu^{S'_Q} + |E|\mu^E$. Since $|S'| - |E| = |S'_P| + |S'_Q|$, we have $(|S'| - |E|)(\mu - \mu^{S'_P}) = |S'_Q|(\mu^{S'_Q} - \mu^{S'_P}) + |E|(\mu^E - \mu)$. Using that $|S'| - |E| = (1 + O(\varepsilon))|S|$, $|S'_Q| = (1-\alpha)|S| - O(\varepsilon)$, $|E| \le O(\varepsilon)|S|$, we have

$$\|\mu - \mu^{S'_P}\|_\infty \le (1 - \alpha + O(\varepsilon))\|\mu^{S'_Q} - \mu^{S'_P}\|_\infty + O(\varepsilon)\,.$$

Similarly,

$$\|\mu - \mu^{S'_Q}\|_\infty \le (\alpha + O(\varepsilon))\|\mu^{S'_Q} - \mu^{S'_P}\|_\infty + O(\varepsilon)\,.$$

Since $S$ is $\varepsilon$-good for $\Pi$, it follows that $\|\mu^{S_P} - p\|_\infty \le \varepsilon/d$ and $\|\mu^{S_Q} - q\|_\infty \le \varepsilon/d$. Also,

$$\||S_P|\mu^{S_P} - |S'_P|\mu^{S'_P}\|_\infty \le |S_P| - |S'_P| \le O(\varepsilon)|S|\,.$$

Thus,

$$\begin{aligned}\|\mu^{S_P} - \mu^{S'_P}\|_\infty &\le \|\mu^{S_P} - (|S'_P|/|S_P|)\mu^{S'_P}\|_\infty \\ &\quad + (|S_P| - |S'_P|)/|S_P| \le O(\varepsilon)|S|/|S_P| \le O(\alpha\varepsilon)\,.\end{aligned}$$

Similarly, we show that

$$\|\mu^{S_Q} - \mu^{S'_Q}\|_\infty \le O((1-\alpha)\varepsilon).$$

Finally, $\|p - q\|_\infty \le \delta$. Thus, by the triangle inequality, we get

$$\|\mu^{S'_Q} - \mu^{S'_P}\|_\infty \le O(\alpha\varepsilon) + \varepsilon/d + \delta + \varepsilon/d + O((1-\alpha)\varepsilon) \le \delta + O(\varepsilon)\,.$$

We have the following sequence of inequalities:

$$\begin{aligned}|S'|\operatorname*{Var}_{X\in_u T}[X_i] &\le |S'|\operatorname*{\mathbb{E}}_{X\in_u T}[\|X-\mu\|_\infty^2] \\ &= |S'_P|\|\mu - \mu^{S'_P}\|_\infty^2 + |S'_Q|\|\mu - \mu^{S'_Q}\|_\infty^2 \\ &\quad + |E|\operatorname*{\mathbb{E}}_{X\in_u T}[\|X-\mu\|_\infty^2] \\ &\le (|S'_P| + |S'_Q|)(\|\mu^{S'_Q} - \mu^{S'_P}\|_\infty + O(\varepsilon))^2 + |E| \\ &\le (\delta^2 + O(\varepsilon))|S'|\,.\end{aligned}$$

Thus,

$$\|\mathrm{Cov}_0(T) - \mathrm{Cov}(T)\|_2 = \max_i (\mathrm{Cov}_0(T) - \mathrm{Cov}(T))_{i,i} = \max_i \mathrm{Var}(T_i) \le O(\delta^2 + \varepsilon).$$

It remains to bound the $\left(\frac{|S'_P|}{|S'|}\right) \|\mathrm{Cov}_0(S'_P)\|_2 + \left(\frac{|S'_Q|}{|S'|}\right) \|\mathrm{Cov}_0(S'_Q)\|_2$ terms in (49). To analyze the first of these terms, note that $\mathrm{Cov}_0(P) = \mathbf{0}$. We have that

$$\begin{aligned}\|\mathrm{Cov}_0(S'_P)\|_2 &= \|\mathrm{Cov}_0(S'_P) - \mathrm{Cov}(P) + \mathrm{Diag}(\mathop{\mathrm{Var}}_{X\sim P}(X_i))\|_2 \\ &\le \|\mathrm{Cov}(S'_P) - \mathrm{Cov}(P)\|_2 + \max_i (|\mathop{\mathrm{Var}}_{X\in_u S'_P}(X_i) - \mathop{\mathrm{Var}}_{X\sim P}(X_i)|) .\end{aligned}$$

Noting that

$$|\mathop{\mathrm{Var}}_{X\in_u S'_P}(X_i) - \mathop{\mathrm{Var}}_{X\sim P}(X_i)| = e_i^T(\mathrm{Cov}(S'_P) - \mathrm{Cov}(P))e_i,$$

we have that

$$\max_i \left(|\mathop{\mathrm{Var}}_{X\in_u S'_P}(X_i) - \mathop{\mathrm{Var}}_{X\sim P}(X_i)|\right) \le \|\mathrm{Cov}(S'_P) - \mathrm{Cov}(P)\|_2 ,$$

and so

$$\|\mathrm{Cov}_0(S'_P)\|_2 \le 2\|\mathrm{Cov}(S'_P) - \mathrm{Cov}(P)\|_2.$$

By the triangle inequality,

$$\|\mathrm{Cov}(S'_P) - \mathrm{Cov}(P)\|_2 \le \|\mathrm{Cov}(S'_P) - \mathrm{Cov}(S_P)\|_2 + \|\mathrm{Cov}(S_P) - \mathrm{Cov}(P)\|_2 .$$

Note that since $S$ is good, the $(i, j)$th entry of $\mathrm{Cov}(S_P) - \mathrm{Cov}(P)$ has absolute value at most $\varepsilon/d$. Thus,

$$\|\mathrm{Cov}(S_P) - \mathrm{Cov}(P)\|_2 \le \|\mathrm{Cov}(S_P) - \mathrm{Cov}(P)\|_F \le \varepsilon ,$$

which gives

$$\|\mathrm{Cov}_0(S'_P)\|_2 \le 2\|\mathrm{Cov}(S'_P) - \mathrm{Cov}(S_P)\|_2 + O(\varepsilon).$$

We have

$$\|\mathrm{Cov}(S'_P) - \mathrm{Cov}(S_P)\|_2 = \sup_{\|v\|_2=1} \left(\left|\mathop{\mathrm{Var}}_{X\in_u S'_P}(v\cdot X) - \mathop{\mathrm{Var}}_{X\in_u S_P}(v\cdot X)\right|\right) .$$

Since $S'_P \subseteq S_P$,

$$\begin{aligned}|S'_P| \mathop{\mathrm{Var}}_{X\in_u S'_P}(v\cdot X) &\le |S_P| \mathop{\mathbb{E}}_{X\in_u S_P}[v\cdot X - \mu^{S'_P}] \\ &\le |S_P| \left(\mathop{\mathrm{Var}}_{X\in_u S_P}(v\cdot X) + \|\mu^{S'_P} - \mu^{S_P}\|_2^2\right) \\ &\le (1 + O(\varepsilon/\alpha))|S'_P| \cdot \left(\mathop{\mathrm{Var}}_{X\in_u S_P}(v\cdot X) + O(\varepsilon^2\log(1/\varepsilon)/\alpha^2)\right) .\end{aligned}$$

Thus,

$$\begin{aligned}\left|\mathop{\mathrm{Var}}_{X\in_u S'_P}(v\cdot X) - \mathop{\mathrm{Var}}_{X\in_u S_P}(v\cdot X)\right| &\le O(\varepsilon/\alpha) \mathop{\mathrm{Var}}_{X\in_u S_P}(v\cdot X) + O(\varepsilon^2\log(1/\varepsilon)/\alpha^2) \\ &\le O(\varepsilon/\alpha) \mathop{\mathrm{Var}}_{X\sim P}(v\cdot X) + O(\varepsilon^2\log(1/\varepsilon)/\alpha^2) \\ &\le O(\varepsilon/\alpha) + O(\varepsilon^2\log(1/\varepsilon)/\alpha^2) \\ &\le O(\varepsilon\log(1/\varepsilon)/\alpha) .\end{aligned}$$

Thus, we have that

$$\|\mathrm{Cov}_0(S'_P)\|_2 \le 2 \cdot O(\varepsilon \log(1/\varepsilon)/\alpha) + O(\varepsilon) \le O(\varepsilon \log(1/\varepsilon)/\alpha).$$

Therefore, combining the above we have that

$$\left(\frac{|S'_P|}{|S'|}\right) \|\mathrm{Cov}_0(S'_P)\|_2 = (\alpha + O(\varepsilon))\|\mathrm{Cov}_0(S'_P)\|_2 = O(\varepsilon \log(1/\varepsilon)) \,.$$

A similar argument shows

$$\left(\frac{|S'_Q|}{|S'|}\right) \|\mathrm{Cov}_0(S'_Q)\|_2 = O(\varepsilon \log(1/\varepsilon)).$$

Combining this with the above gives that

$$\|\mathrm{Cov}_0(S') - \mathrm{Cov}(T)\|_2 = O(\delta^2 + \varepsilon \log(1/\varepsilon)).$$

By the assumption on $\delta$ in Theorem 8.17, $\delta^2 = \Omega(\varepsilon \log(1/\varepsilon))$, and the proof is complete. ∎

**Acknowledgements.** The fourth author would like to thank Michael B. Cohen and Samuel B. Hopkins for some very helpful discussions. The authors thank the reviewers for their detailed feedback, and Lili Su for pointing out an error in the proof of Lemma 5.3.

## REFERENCES

[ABL17] P. Awasthi, M. F. Balcan, and P. M. Long, *The power of localization for efficiently learning linear separators with noise*, J. ACM, 63 (2017), 50.

[ADLS17] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt, *Sample-optimal density estimation in nearly-linear time*, in Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 2017, pp. 1278–1289.

[AGHK13] A. Anandkumar, R. Ge, D. Hsu, and S. Kakade, *A tensor spectral approach to learning mixed membership community models*, in Proceedings of the 26th Annual Conference on Learning Theory, 2013, pp. 867–881.

[AGM12] S. Arora, R. Ge, and A. Moitra, *Learning topic models—going beyond SVD*, in Proceedings of the 53rd Annual Symposium on Foundations of Computer Science, IEEE Computer Soc., Los Alamitos, CA, 2012, pp. 1–10.

[AGMS15] S. Arora, R. Ge, A. Moitra, and S. Sachdeva, *Provable ICA with unknown Gaussian noise, and implications for Gaussian mixtures and autoencoders*, Algorithmica, 72 (2015), pp. 215–236.

[AHK12] A. Anandkumar, D. Hsu, and S. Kakade, *A method of moments for mixture models and hidden Markov models*, J. Mach. Learn. Res. - Proceedings Track, 23 (2012), pp. 33.1–33.34.

[AJOS14] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, *Sorting with adversarial comparators and application to density estimation*, in Proceedings of the 2014 IEEE International Symposium on Information Theory (ISIT), 2014, pp. 1682–1686.

[AK01] S. Arora and R. Kannan, *Learning mixtures of arbitrary Gaussians*, in Proceedings of the 33rd Annual ACM Symposium on Theory of Computing, ACM, New York, 2001, pp. 247–257.

[BBBB72] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk, *Statistical Inference under Order Restrictions*, John Wiley & Sons, London, New York, Sydney, 1972.

[BD15] S. Balmand and A. Dalalyan, *Convex Programming Approach to Robust Estimation of a Multivariate Gaussian Model*, preprint, https://arxiv.org/abs/1512.04734, 2015.

[BDLS17] S. Balakrishnan, S. S. Du, J. Li, and A. Singh, *Computationally efficient robust sparse estimation in high dimensions*, in Proceedings of the 30th Annual Conference on Learning Theory, 2017, pp. 169–212.

[Ber06] T. Bernholt, *Robust Estimators Are Hard to Compute*, Technical Report, University of Dortmund, Dortmund, Germany, 2006.

[Bru09] S. C. Brubaker, *Extensions of Principle Components Analysis*, PhD. thesis, Georgia Institute of Technology, Atlanta, GA, 2009.

[BS10] M. Belkin and K. Sinha, *Polynomial learning of distribution families*, in Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Soc., Los Alamitos, CA, 2010, pp. 103–112.

[BV08] S. C. Brubaker and S. Vempala, *Isotropic PCA and affine-invariant clustering*, in Proceedings of the 49th IEEE Symposium on Foundations of Computer Science, IEEE Computer Soc., Los Alamitos, CA, 2008, pp. 551–560.

[CDSS13] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun, *Learning mixtures of structured distributions over discrete domains*, in Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 2013, pp. 1380–1394.

[CDSS14a] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun, *Efficient density estimation via piecewise polynomial approximation*, in Proceedings of the 46th Annual ACM Symposium on Theory of Computing, ACM, New York, 2014, pp. 604–613.

[CDSS14b] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun, *Near-optimal density estimation in near-linear time using variable-width histograms*, in Proceedings of the 27th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, 2014, pp. 1844–1852.

[CGG02] M. Cryan, L. A. Goldberg, and P. W. Goldberg, *Evolutionary trees can be learned in polynomial time in the two-state general Markov model*, SIAM J. Comput., 31 (2002), pp. 375–397, https://doi.org/10.1137/S0097539798342496.

[CLM+16] M. B. Cohen, Y. T. Lee, G. L. Miller, J. W. Pachocki, and A. Sidford, *Geometric median in nearly linear time*, in Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, ACM, New York, 2016, pp. 9–21.

[CLMW11] E. J. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?*, J. ACM, pp. 58 (2011), 11.

[CSV17] M. Charikar, J. Steinhardt, and G. Valiant, *Learning from untrusted data*, in Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, ACM, New York, 2017, pp. 47–60.

[CT06] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed., John Wiley & Sons, Hoboken, NJ, 2006.

[CW01] A. Carbery and J. Wright, *Distributional and $L^q$ norm inequalities for polynomials over convex bodies in $R^n$*, Math. Res. Lett., 8 (2001), pp, 233–248.

[Das99] S. Dasgupta, *Learning mixtures of Gaussians*, in Proceedings of the 40th Annual Symposium on Foundations of Computer Science, IEEE Computer Soc., Los Alamitos, CA, 1999, pp. 634–644.

[DBS17] S. S. Du, S. Balakrishnan, and A. Singh, *Computationally Efficient Robust Estimation of Sparse Functionals*, preprint, https://arxiv.org/abs/1702.07709, 2017.

[DDKT16] C. Daskalakis, A. De, G. Kamath, and C. Tzamos, *A size-free CLT for Poisson multinomials and its applications*, in Proceedings of the 48th Annual ACM Symposium on the Theory of Computing, ACM, New York, 2016, pp. 1074–1086.

[DDO+13] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R.A. Servedio, and L. Tan, *Learning sums of independent integer random variables*, in Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Soc., Los Alamitos, CA, 2013 pp. 217–226.

[DDS12] C. Daskalakis, I. Diakonikolas, and R. A. Servedio, *Learning k-modal distributions via testing*, in Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, 2012, pp. 1371–1385.

[DDS15a] C. Daskalakis, I. Diakonikolas, and R. A. Servedio, *Learning Poisson binomial distributions*, Algorithmica, 72 (2015), pp. 316–357.

[DDS15b] A. De, I. Diakonikolas, and R. Servedio, *Learning from satisfying assignments*, in Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 2015, pp. 478–497.

[DG85] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The $L_1$ View*, John Wiley & Sons, New York, 1985.

[DK14] C. Daskalakis and G. Kamath, *Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians*, in Proceedings of the 27th Annual Conference on Learning Theory, 2014, pp. 1183–1213.

[DKK+16] 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 the 57th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, Washington, DC, 2016, pp. 655–664.

[DKK+17] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, *Being robust (in high dimensions) can be practical*, in Proceedings of the 34th International Conference on Machine Learning, JMLR, Inc., 2017, pp. 999–1008.

[DKK+18] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, J. Steinhardt, and A. Stewart, *Sever: A Robust Meta-Algorithm for Stochastic Optimization*, preprint, https://arxiv.org/abs/1803.02815, 2018.

[DKS16a] I. Diakonikolas, D. M. Kane, and A. Stewart, *The Fourier transform of Poisson multinomial distributions and its algorithmic applications*, in Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, ACM, New York, 2016, pp. 1060–1073.

[DKS16b] I. Diakonikolas, D. M. Kane, and A. Stewart, *Optimal learning via the Fourier transform for sums of independent integer random variables*, in Proceedings of the 29th Annual Conference on Learning Theory, 2016, pp. 831–849.

[DKS16c] I. Diakonikolas, D. M. Kane, and A. Stewart, *Robust Learning of Fixed-Structure Bayesian Networks*, preprint, https://arxiv.org/abs/1606.07384, 2016.

[DKS17] I. Diakonikolas, D. M. Kane, and A. Stewart, *Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures*, in Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, Washington, DC, 2017, pp. 73–84.

[DKS18a] I. Diakonikolas, D. M. Kane, and A. Stewart, *Learning geometric concepts with nasty noise*, in Proceedings of the 50th Annual ACM Symposium on Theory of Computing, ACM, New York, 2018, pp. 1061–1073.

[DKS18b] 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 Symposium on Theory of Computing, ACM, New York, 2018, pp. 1047–1060.

[DKT15] C. Daskalakis, G. Kamath, and C. Tzamos, *On the structure, covering, and learning of Poisson multinomial distributions*, in Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, Washington, DC, 2015, pp. 1203–1217.

[DL96] L. Devroye and G. Lugosi, *A universally acceptable smoothing factor for kernel density estimates*, Ann. Statist., 24 (1996), pp. 2499–2512.

[DL97] L. Devroye and G. Lugosi, *Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes*, Ann. Statist., 25 (1997), pp. 2626–2637.

[DL01] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*, Springer Ser. Statist., Springer-Verlag, New York, 2001.

[FM99] Y. Freund and Y. Mansour, *Estimating a mixture of two product distributions*, in Proceedings of the Twelfth Annual Conference on Computational Learning Theory, ACM, New York, 1999, pp. 53–62.

[FOS08] J. Feldman, R. O'Donnell, and R. A. Servedio, *Learning mixtures of product distributions over discrete domains*, SIAM J. Comput., 37 (2008), pp. 1536–1564, https://doi.org/10.1137/060670705.

[GLS88] M. Grötschel, L. Lovász, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Vol. 2, Springer-Verlag, Berlin, 1988.

[Hau92] D. Haussler, *Decision theoretic generalizations of the PAC model for neural net and other learning applications*, Inform. and Comput., 100 (1992), pp. 78–150.

[HK13] D. Hsu and S. M. Kakade, *Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions*, in Innovations in Theoretical Computer Science, ACM, New York, 2013, pp. 11–19.

[HL18] S. B. Hopkins and J. Li, *Mixture models, robustness, and sum of squares proofs*, in Proceedings of the 50th Annual ACM Symposium on Theory of Computing, ACM, New York, 2018, pp. 1021–1034.

[HM13] M. Hardt and A. Moitra, *Algorithms and hardness for robust subspace recovery*, in Proceedings of the 26th Annual Conference on Learning Theory, 2013, pp. 354–375.

[HP15] M. Hardt and E. Price, *Sharp bounds for learning a mixture of two Gaussians*, in Proceedings of the 47th Annual ACM Symposium on Theory of Computing, ACM, New York, 2015, pp. 753–760.

[HR09] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, John Wiley & Sons, New York, 2009.

[HRRS86] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics. The Approach Based on Influence Functions*, John Wiley & Sons, New York, 1986.

[Hub67] P. J. Huber, *The behavior of maximum likelihood estimates under nonstandard conditions*, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, University of California Press, Berkeley, CA, 1967, pp. 221–233.

[JP78] D. S. Johnson and F. P. Preparata, *The densest hemisphere problem*, Theoret. Comput. Sci., 6 (1978), pp. 93–107.

[KKM18] A. Klivans, P. K. Kothari, and R. Meka, *Efficient Algorithms for Outlier-Robust Regression*, preprint, https://arxiv.org/abs/1803.03241, 2018.

[KL93] M. J. Kearns and M. Li, *Learning in the presence of malicious errors*, SIAM J. Comput., 22 (1993), pp. 807–837, https://doi.org/10.1137/0222052.

[KLS09] A. Klivans, P. Long, and R. Servedio, *Learning halfspaces with malicious noise*, in Automata, Languages and Programming. Part I, Lecture Notes in Comput. Sci. 5555, Springer, Berlin, 2009, pp. 609–621.

[KMR+94] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie, *On the learnability of discrete distributions*, in Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing, ACM, New York, 1994, pp. 273–282.

[KMV10] A. T. Kalai, A. Moitra, and G. Valiant, *Efficiently learning mixtures of two Gaussians*, in Proceedings of the 2010 ACM International Symposium on Theory of Computing, ACM, New York, pp. 553–562.

[KSS94] M. Kearns, R. Schapire, and L. Sellie, *Toward efficient agnostic learning*, Mach. Learn., 17 (1994), pp. 115–141.

[KSS18] P. K. Kothari, J. Steinhardt, and D. Steurer, *Robust moment estimation and improved clustering via sum of squares*, in Proceedings of the 50th Annual ACM Symposium on Theory of Computing, ACM, New York, 2018, pp. 1035–1046.

[Li17] J. Li, *Robust Sparse Estimation Tasks in High Dimensions*, preprint, https://arxiv.org/abs/1702.05860, 2017.

[LM00] B. Laurent and P. Massart, *Adaptive estimation of a quadratic functional by model selection*, Ann. Statist., 28 (2000), pp. 1302–1338.

[LMTZ15] G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang, *Robust computation of linear models by convex relaxation*, Found. Comput. Math., 15 (2015), pp. 363–410.

[LRV16] K. A. Lai, A. B. Rao, and S. Vempala, *Agnostic estimation of mean and covariance*, in Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, Los Alamitos, CA, 2016, pp. 665–674.

[LS17] J. Li and L. Schmidt, *Robust proper learning for mixtures of Gaussians via systems of polynomial inequalities*, in Proceedings of the 30th Annual Conference on Learning Theory, 2017, pp. 1302–1382.

[LT15] P. L. Loh and X. L. Tan, *High-Dimensional Robust Precision Matrix Estimation: Cellwise Corruption under $\varepsilon$-Contamination*, preprint, https://arxiv.org/abs/1509.07229, 2015.

[MR05] E. Mossel and S. Roch, *Learning nonsingular phylogenies and hidden Markov models*, in Proceedings of the 37th Annual ACM Symposium on Theory of Computing, ACM, New York, 2005, pp. 366–375.

[MV10] A. Moitra and G. Valiant, *Settling the polynomial learnability of mixtures of Gaussians*, in Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Soc., Los Alamitos, CA, 2010, pp. 93–102.

[MV17] M. Meister and G. Valiant, *A Data Prism: Semi-Verified Learning in the Small-Alpha Regime*, preprint, https://arxiv.org/abs/1708.02740, 2017.

[PSBR18] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar, *Robust Estimation via Robust Gradient Estimation*, preprint, https://arxiv.org/abs/1802.06485, 2018.

[QV18] M. Qiao and G. Valiant, *Learning discrete distributions from untrusted batches*, in Proceedings of the 9th Conference on Innovations in Theoretical Computer Science, ACM, New York, 2018, 47.

[RH17] P. Rigollet and J.-C. Hütter, *High-Dimensional Statistics*, available at http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf, 2017.

[Sco92] D. W. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization*, John Wiley & Sons, Hoboken, NJ, 1992.

[SCV18] J. Steinhardt, M. Charikar, and G. Valiant, *Resilience: A criterion for learning in the presence of arbitrary outliers*, in Proceedings of the 9th Conference on Innovations in Theoretical Computer Science, ACM, New York, 2018, 45.

[Ser01] R. Servedio, *Smooth boosting and learning with malicious noise*, in Proceedings of the Fourteenth Annual Conference on Computational Learning Theory, 2001, pp. 473–489.

[Ser03] R. Servedio, *Smooth boosting and learning with malicious noise*, J. Mach. Learn. Res., 4 (2003), pp. 633–648.

[Sil86] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, 1986.

[SOAJ14] A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour, *Near-optimal-sample estimators for spherical Gaussian mixtures*, in Proceedings of the 27th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, 2014, pp. 1395–1403.

[Tuk75] J. W. Tukey, *Mathematics and picturing of data*, in Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2, Canad. Math. Congress, Montreal, Que., Canada, 1975, pp. 523–531.

[Val85] L. Valiant, *Learning disjunctions of conjunctions*, in Proceedings of the 9th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, 1985, pp. 560–566.

[Ver10] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, in Compressed Sensing, Cambridge University Press, Cambridge, UK, 2012, pp. 210–268.

[VW02] S. Vempala and G. Wang, *A spectral algorithm for learning mixtures of distributions*, in Proceeding of the 43rd Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Soc., Los Alamitos, CA, 2002, pp. 113–122.

[Whi82] H. White, *Maximum likelihood estimation of misspecified models*, Econometrica, 50 (1982), pp. 1–25.

[Yat85] Y. G. Yatracos, *Rates of convergence of minimum distance estimators and Kolmogorov's entropy*, Ann. Statist., 13 (1985), pp. 768–774.

[ZL14] T. Zhang and G. Lerman, *A novel m-estimator for robust PCA*, J. Mach. Learn. Res., 15 (2014), pp. 749–808.