Robust Estimators in High Dimensions without the Computational Intractability

Ilias Diakonikolas[*]
CS, USC
diakonik@usc.edu

Gautam Kamath[†]
EECS, MIT
g@csail.mit.edu

Daniel Kane[‡]
CSE & Math, UCSD
dakane@cs.ucsd.edu

Jerry Li [§]
EECS, MIT
jerryzli@mit.edu

Ankur Moitra[¶]
Math & CSAIL, MIT
moitra@mit.edu

Alistair Stewart[||]
CS, USC
alistais@usc.edu

April 22, 2016

Abstract

We study high-dimensional distribution learning in an agnostic setting where an adversary is allowed to arbitrarily corrupt an ε-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.

Contents

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 [FOS05, KMV10, MV10, BS10, HK13], phylogenetic trees [CGG02, MR05], HMMs [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 ε-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 ε-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 *et al.* [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 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 depth [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, (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* in 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) in the fraction ε 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 ε-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 θ and θ' are two sets of parameters that define distributions P_θ and $P_{\theta'}$ in a given class $\mathcal{D}$, a learning algorithm often relies on establishing the following type of relation[1] between $d_{\text{TV}}(P_\theta, P_{\theta'})$ and the parameter distance $d_p(\theta, \theta')$:

$$\text{poly}(d_p(\theta, \theta'), 1/d) \leq d_{\text{TV}}(P_\theta, P_{\theta'}) \leq \text{poly}(d_p(\theta, \theta'), d) . \tag{1}$$

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

$$\text{poly}(d_p(\theta, \theta')) \leq d_{\text{TV}}(P_\theta, P_{\theta'}) \leq \text{poly}(d_p(\theta, \theta')) \tag{2}$$

which allows us to reformulate our goal of designing robust estimators, with distribution-independent error guarantees, as the goal of robustly estimating θ 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 ℓ_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 much better sample complexity, in some cases matching the information-theoretic limit. Notably, either approach can be used to give all of our concrete learning applications with nearly identical error guarantees. In what follows, we specialize to the problem of robustly learning the mean μ 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

---

[1]For example, the work of Kalai *et al.* [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.

points that are obviously outliers) can be $\Omega(\varepsilon\sqrt{d})$-far from the true mean in ℓ_2-distance. The filter approach shows that either the sample mean is already a good estimate for μ 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, so that the weighted empirical average $\widehat{\mu}_w = \sum_{i=1}^N w_i X_i$ is close to μ. 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 μ, *which is unknown*. Nevertheless, it turns out that we can use the same types 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 μ.

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 ε-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 μ and identity covariance, that every $(1-\varepsilon)N$ set of samples produces a good estimate for μ. 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 μ and covariance Σ, 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 Σ is the unknown covariance matrix:

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

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

We can alternatively establish Theorem 1.2 with a slightly worse sample complexity via our filtering technique. See Section 8.

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 ℓ_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 χ^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:

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

For the sake of simplicity in the presentation, we did not make an effort to optimize the sample complexity of our robust estimators. 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 learning mixtures of the above families is rather non-trivial even in the noiseless case. First, we study 2-mixtures of c-balanced products, which stipulates that the coordinates of the mean vector of each component are in the range $(c, 1 - c)$. We prove:

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

This generalizes the algorithm of Freund and Mansour [FM99] to the agnostic setting. An interesting open question is to improve the ε-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. *Let* $\mathcal{M}$ *be a mixture of* k *Gaussians with spherical covariances, and let* $\varepsilon, \tau > 0$ *and* k *be a constant. There is a polynomial time (for constant* k) *algorithm which given* ε, τ, *and an* ε-*corrupted set of* N *samples from* $\mathcal{M}$ *with* $N \geq \mathrm{poly}(k, d, 1/\varepsilon, \log(1/\tau))$, *outputs an* $\mathcal{M}'$ *so that with probability* $1 - \tau$, *we have* $d_{\mathrm{TV}}(\mathcal{M}, \mathcal{M}') \leq \tilde{O}(\mathrm{poly}(k) \cdot \sqrt{\varepsilon})$.

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 ℓ_2-distance.

4

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$^+$94] initiated a systematic investigation of the computational complexity of distribution learning. Since then, efficient learning algorithms have been developed for a wide range of distributions in both low and high-dimensions [Das99, FM99, AK01, VW02, CGG02, MR05, BV08, KMV10, MV10, BS10, DDS12a, DDS12b, CDSS13, DDO$^+$13, CDSS14a, CDSS14b, HP15, ADLS15, DDS15b, DDKT16, DKS15, DKS16].

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 removing 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 [FOS05]. Given the vast body of work on high-dimensional distribution learning, there is 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 on 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 ε-fraction of both the samples and their labels. A sequence of works studied this problem for the class of halfspaces [Ser01, Ser03, KLS09] in the setting where the underlying distribution is a Gaussian, culminating in the work of Awasthi *et al.* [ABL14], who gave an efficient algorithm that finds a halfspace with agreement $O(\varepsilon)$. Our work and theirs are not directly comparable, since we work in an unsupervised setting. Moreover, their algorithms need to assume that the underlying Gaussian distribution is in isotropic position. In fact, our results are complementary to theirs: One could use our algorithms (on the unlabeled examples) to learn an affine transformation that puts the underlying Gaussian distribution in approximately isotropic position, even in the presence of malicious errors, so that one can then directly apply the [ABL14] algorithm.

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 than 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, LMTZ12]. 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 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 5 illustrate our convex programming framework, while Sections 6–8 illustrate our filter framework.

More specifically, in Section 4 we apply our convex programming framework to a single Gaussian with unknown mean and unknown covariance. In Section 5, we generalize these techniques to obtain an agnostic algorithm for mixtures of spherical Gaussians with unknown means. In Section 6, we apply our filter techniques to a binary product distribution. In Section 7 we generalize the techniques of Section 6 to obtain an agnostic learning algorithm for a mixture of two balanced binary product distributions. Finally, in Section 8 we apply our filter technique to obtain an alternate agnostic learning algorithm for a single Gaussian with unknown mean and unknown covariance.

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 a Gaussian with unknown mean, in Section 4.3. Similarly, for the filter framework, we suggest either the balanced product distribution in Section 6.1 or the Gaussian with unknown mean in Section 8.1.

2 Preliminaries

2.1 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 2.1. *Given $\varepsilon > 0$ and a class of distributions $\mathcal{D}$, the* oblivious adversary *chooses a distribution P so 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 2.2. *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 $m' \sim Bin(\varepsilon, m)$.*

The adversary is allowed to inspect the samples, and 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 $\mathrm{Bin}(\varepsilon, m)$).

We rely on the following well-known fact:

Fact 2.3. *Let P, D be two distributions so that $d_{\mathrm{TV}}(P, D) = \varepsilon$. Then there are distributions N_1 and N_2 so 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 2.4. *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 ε_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 each sample Z_i that came from N_1, we can delete and replace 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 $m' \sim \mathrm{Bin}(\varepsilon_1 + \varepsilon_2, m)$, although m' has and needs to have some dependence on the original samples from D. $\qquad\square$

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 2.5. *We say a a set of samples $X_1, X_2, \ldots, X_m$ is an ε-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.

2.2 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. 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.

2.3 Distributions of Interest

One object of study in this paper is the Gaussian (or Normal) distribution.

Definition 2.6. *A Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with mean μ and covariance Σ 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 2.7. *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 2.8. *A* mixture P *of distributions $P_1, \ldots, P_k$ with mixing weights $\alpha_1, \ldots, \alpha_k$ 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 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 non-agnostic 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 in 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, and as we will see this mitigates the problem.

Lemma 2.9. *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 2ε-fraction have been arbitrarily corrupted, returns a list of M distributions whose probability density functions are explicitly computable and which can be effectively sampled from so that with $1-\tau/2$ probability there exists a $\Pi' \in M$ 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 Π, an ε-fraction of which have been arbitrarily corrupted, returns a single distribution Π' so that with $1 - \tau$ probability $d_{\mathrm{TV}}(\Pi', \Pi) < O(\delta + \varepsilon)$.*

Proof. Firstly, 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 2ε-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 Π and another set, S_2, of

8

$C(\log(|\mathcal{M}|) + \log(1/\tau))/\varepsilon^2$ samples from Π and can arbitrary corrupt a 2ε-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 M 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, that 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 that for any $P, Q \in \mathcal{M}$ that

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

This follows by Chernoff bounds and a union bound over the $|\mathcal{M}|^2$ possibilities for P and Q. Since the total variational distance between the uniform distributions over S_2 and S'_2 is at most 2ε, 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

$$\Pr_{x \in_u S'_2}(x \in A_{PQ}) = \Pr_{x \sim \Pi}(x \in A_{PQ}) + O(\varepsilon) \geq \Pr_{x \sim P}(x \in A_{PQ}) + O(\varepsilon) - \delta \geq \Pr_{x \sim Q}(x \in A_{PQ}) + \delta + C\varepsilon/5.$$

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 so 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 Π'. $\square$

2.5 Bounds on TV Distance

The *Kullbeck-Liebler 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 so that $\det(\Sigma_1), \det(\Sigma_2) \neq 0$. Then*

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

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 lemmas, which allow us to relate parameter distance between two Gaussians to their total variation distance. The first 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 $d_{\mathrm{TV}}\left(\mathcal{N}(\mu_1, I), \mathcal{N}(\mu_2, I)\right) \leq \frac{1}{2}\|\mu_2 - \mu_1\|_2$.*

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

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

as desired. □

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

Corollary 2.14. *Let $\delta > 0$ be sufficiently small. Let Σ_1, Σ_2 so 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 $\sum_{i=1}^d \lambda_i - \log(1 + \lambda_i)$, where we are guaranteed that $(\sum_{i=1}^d \lambda_i^2)^{1/2} = \delta$. By the linear approximation to $\ln(1 + x)$, for x small, we have that for δ 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 δ 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 μ, and (2) We are given samples from $\mathcal{N}(0, \Sigma)$ and our goal is to learn Σ. 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 will also need the following lemma, bounding the total variation distance between two product distributions by the chi-squared distance between the corresponding mean vectors.

Lemma 2.15. *Let P, Q be binary product distributions with mean vectors $p, q \in (0,1)^d$. We have that*

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

The proof is deferred to Section A.

2.6 Miscellaneous Lemmata

We will frequently use the following version of the Chernoff bound:

Theorem 2.16. *Let $Z_1, \ldots, Z_n$ be independent random variables with Z_i supported on $[a_i, b_i]$. Let $Z = \sum_{i=1}^{d} Z_i$. Then for any $T > 0$,*

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

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 Fact 2.13 in order to achieve an estimate that is $O(\varepsilon)$-close in total variation distance (for a Gaussian when μ 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 agnostic learning in high dimensions is to agnostically learn along each coordainate 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 a ε-fraction of the samples are corrupted, our estimate can be off by ε 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 $\prod_{i=1}^{m} F(X_i)$. 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

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

$$= N \log \sqrt{2\pi} + \frac{1}{2} \sum_{i=1}^{N} \|X_i - \mu\|_2^2 \ ,$$

11

and so the μ which minimizes $\text{NLL}(\mathcal{N}(\mu, I), X_1, \ldots, X_N)$ 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 ε-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. Hence 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:

$$\text{med}(S) \triangleq \min_v \sum_{x \in S} \|x - v\|_2 \ .$$

Next, we show that the geometric median also does not give dimension independent error guarantees for learning the mean of a Gaussian with known covariance. Suppose we have a set S_1 of $(1 - \varepsilon)N$ samples from $\mathcal{N}(0, I)$ and a set S_2 of εN samples $\in \{0, 1\}^d$. Each element $x \in S_2$ is generated independently by choosing a random set $Q_x \subset [d]$ of cardinality $d/2$ and setting $x_i = 1$ if $i \in Q_x$, and 0 otherwise. Call this distribution D.

Let $f(v) = \sum_{x \in S_1 \cup S_2} \|x_j - v\|_2$, and let v^* be the value which minimizes this function, which is also the geometric median of this set of points. We know that the gradient is 0 at this point, and thus, for all i, $(\nabla f(v^*))_i = 0$. Assume that $\|v^*\|_2 \leq \varepsilon$, where we set $\varepsilon \geq \Omega(1/\sqrt{d})$. We will show to be a contradiction.

Writing the expression for the gradient, we have

$$(\nabla f(v))_i = \frac{1}{2} \sum_{x \in S_1} \frac{v_i - x_i}{\|x - v\|_2} + \frac{1}{2} \sum_{x \in S_2} \frac{v_i - x_i}{\|x - v\|_2} \ .$$

We will bound each term of this in turn at $v = v^*$, assuming $\|v^*\| \leq \varepsilon$. First, observe that

$$\left| \frac{1}{(1 - \varepsilon)N} \sum_{x \in S_1} \frac{v_i^*}{\|x - v^*\|_2} - v_i^* \mathbb{E}\left[\frac{1}{\|x - v^*\|_2} \right] \right| \to 0$$

as $N \to \infty$ by a Chernoff bound, and by basic concentration, we have that $\mathbb{E}\left[\frac{1}{\|x - v^*\|_2} \right] = \Theta(1/\sqrt{d})$, if

$\|v^*\| \leq O(1)$. We also have

$$
\left| \frac{1}{(1-\varepsilon)N} \sum_{x \in S_1} \frac{x_i}{\sqrt{\sum_{j=1}^d (x_j - v_j^*)^2}} - \frac{1}{(1-\varepsilon)N} \sum_{x \in S_1} \frac{x_i}{\sqrt{\sum_{j \neq i}(x_j - v_j^*)^2}} \right|
$$

$$
\leq \frac{1}{(1-\varepsilon)N} \sum_{x \in S_1} |x_i| \cdot \left| \frac{\sqrt{\sum_{j \neq i}(x_j - v_j^*)^2} - \sqrt{\sum_{j=1}^d (x_j - v_j^*)^2}}{\sqrt{\left(\sum_{j=1}^d (x_j - v_j^*)^2\right)\left(\sum_{j \neq i}(x_j - v_j^*)^2\right)}} \right|
$$

$$
\leq \frac{1}{(1-\varepsilon)N} \sum_{x \in S_1} |x_i| \frac{\widetilde{O}(1)}{\widetilde{\Omega}(d)} \leq \widetilde{O}\left(\frac{1}{d}\right) \,,
$$

from basic Gaussian concentration. Moreover, we have

$$
\left| \frac{1}{(1-\varepsilon)N} \sum_{x \in S_1} \frac{x_i}{\sqrt{\sum_{j \neq i}(x_j - v_j^*)^2}} \right| \to 0
$$

as $N \to \infty$, by another Chernoff bound. Putting these terms together, we have that

$$
\left| \frac{1}{(1-\varepsilon)N} \sum_{x \in S_1} \frac{v_i^* - x_i}{\|x - v\|_2} - v_i^* \mathop{\mathbb{E}}_{x \sim \mathcal{N}(0,I)}\left[\frac{1}{\|x - v^*\|_2}\right] \right| \leq \widetilde{O}\left(\frac{1}{d}\right) + o(1) \,.
$$

Similarly, we also have that since $\mathbb{E}_{x \sim D}[x_i] = 1/2$,

$$
\left| \frac{1}{\varepsilon N} \sum_{x \in S_2} \frac{v_i^* - x_i}{\|x - v\|_2} - (v_i^* + 1/2) \mathop{\mathbb{E}}_{x \sim D}\left[\frac{1}{\|x - v^*\|_2}\right] \right| \leq \widetilde{O}\left(\frac{1}{d}\right) + o(1) \,,
$$

and moreover, $\mathbb{E}_{x \sim D}\left[\frac{1}{\|x-v^*\|_2}\right] = \Theta\left(1/\sqrt{d}\right)$. Thus, altogether we have that

$$
0 = \frac{1}{(1-\varepsilon)N} \nabla f(v^*)_i = v_i^* \cdot \Theta\left(\frac{1}{\sqrt{d}}\right) + \varepsilon(v_i^* + 1/2) \cdot \Theta\left(\frac{1}{\sqrt{d}}\right) + O\left(\frac{1}{d}\right) + o(1) \,,
$$

which we see is impossible for our choice of $\varepsilon \geq \Omega(1/\sqrt{d})$ unless $|v_i^*| = \Omega(\varepsilon)$. In particular, this means that each coordinate of the geometric median must be at least $\Omega(\varepsilon)$, and so $\|v^*\|_2 \geq \Omega(\varepsilon\sqrt{d})$, which contradicts our original assumption. This finishes the analysis.

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 zero 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*

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

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

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

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 following notation

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

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 5 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

Here we will give rates of convergence for various statistics for Gaussians with covariance matrix I that we will make use of later. First, we will require the following well-known tail bounds:

Lemma 4.3 ([Ver10], Theorem 5.50)**.** *Let $Y_i \sim \mathcal{N}(0, I)$ be independent, for $i = 1, \ldots, N$, where we set $N = \Omega\left(\frac{d + \log(1/\tau)}{\delta^2}\right)$. We have that with probability $1 - \tau$,*

$$\left\| \sum_{i=1}^{N} \frac{1}{m} Y_i Y_i^T - I \right\|_2 \le \delta .$$

Lemma 4.4 (folklore)**.** *With the same setup as above, with probability $1 - \tau$,*

$$\left\| \frac{1}{N} \sum_{i=1}^{N} Y_i \right\|_2 \le \delta .$$

14

The above bounds have the property that as we take more samples, we can take δ to be arbitrarily small. We will also be interested in how well various statistics concentrate around their expectation, when we take the worst-case set of weights in $S_{N,\varepsilon}$. This is more subtle 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 proof to Section B.

Lemma 4.5. *Fix $\varepsilon \le 1/2$ and $\tau \le 1$. There is a $\delta_1 = O(\varepsilon\sqrt{\log 1/\varepsilon})$ so that if $Y_1, \ldots, Y_N$ are independent samples from $\mathcal{N}(0, I)$ and $N = \Omega\left(\frac{d + \log(1/\tau)}{\delta_1^2}\right)$, then*

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

A nearly identical argument (Use Hoeffding instead of Bernstein in the proof of Theorem 5.50 in [Ver10]) yields:

Lemma 4.6. *Fix ε and τ as above. There is a $\delta_2 = O(\varepsilon\sqrt{\log 1/\varepsilon})$ so 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*

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

Note that by Cauchy-Schwarz, this implies:

Corollary 4.7. *Fix ε and τ as above. There is a $\delta_2 = O(\varepsilon\sqrt{\log 1/\varepsilon})$ so that if $Y_1, \ldots, Y_N$ are independent samples from $\mathcal{N}(0, I)$ and $N \ge \Omega\left(\frac{d + \log(1/\tau)}{\delta_2^2}\right)$, then*

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

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

Fact 4.8. *Fix $\tau > 0$. Let $X_1, \ldots, X_N \sim \mathcal{N}(0, I)$. Then, with probability $1 - \tau$, we have that $\|X_i\|_2 \le O\left(\sqrt{d \log(N/\tau)}\right)$ 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 so 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

$$\|\widehat{\Sigma} - I\|_F = \max_{\|U\|_F = 1} \left|\left\langle \widehat{\Sigma} - I, U \right\rangle\right|$$

$$= \max_{\|U\|_F = 1} \left|\frac{1}{m} \sum_{i=1}^{m} \mathrm{tr}(X_i X_i^T U) - \mathrm{tr}(U)\right| .$$

We first state the following lemma, whose proof is deferred to Section B:

15

Lemma 4.9. *Let $U \succeq 0$ with $\|U\|_F = 1$. Then, for any $\tau' \geq 0$ and any $\delta \geq 0$, if $N = \Omega(\frac{d^2 + \log 1/\tau'}{\delta^2})$ then with probability $1 - \tau'$.*

$$\left| \frac{1}{N} \sum_{i=1}^{N} \mathrm{tr}(X_i X_i^T U) - \mathrm{tr}(U) \right| \leq \delta$$

Since there is a $1/4$-net over all PSD matrices with Frobenius norm 1 of size 9^{d^2}, the Vershynin-type union bound argument immediately gives us the following:

Corollary 4.10. *For any $\delta, \tau > 0$, if*

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

then $\|\widehat{\Sigma} - I\|_F \leq \delta$ with probability $1 - \tau$.

By the same union bound technique as used in the proof of Lemma 4.5, we obtain:

Corollary 4.11. *Fix $\varepsilon, \tau > 0$. There is a $\delta_1 = O(\varepsilon \sqrt{\log 1/\varepsilon})$ so 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

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

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 εm, and take the uniform weights over that set, then the empirical covariance is not too far away from the truth. More formally:

Corollary 4.12. *Fix $\varepsilon, \tau > 0$. There is a $\delta_2 = O(\varepsilon \sqrt{\log 1/\varepsilon})$ so 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

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

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 4.13. *For any matrix $M \in \mathbb{R}^{d \times d}$, let $M^\flat \in R^{d^2}$ 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 $M \in \mathbb{R}^{d \times d}$ so that $M^\flat = v$.*

16

We will also require the following definitions:

Definition 4.14. *Let $\mathcal{S}_{\text{sym}} = \{M^\flat \in \mathbb{R}^{d^2} : M \text{ is symmetric}\}$, and 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 Π_S and $\Pi_{\mathcal{S}^\perp}$ denote the projection operators onto $\mathcal{S}$ and $\mathcal{S}^\perp$ respectively. Finally let

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

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

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

In fact, the projection of $v = A^\flat$ onto S where A 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 $\|v\|_{\mathcal{S}^\perp} = \frac{|\text{tr}(M)|}{\sqrt{d}}$.

The key result in this section is the following:

Theorem 4.15. *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 $\mathcal{S}_{\text{sym}}$, 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's theorem, and is deferred to Section B.

Theorem 4.16 (Isserlis's 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 $\{1, \ldots, k\}$.

4.2.4 Concentration of the Fourth Moment Tensor

We also need to show that the fourth moment tensor concentrates:

Theorem 4.17. *Fix $\varepsilon, \tau > 0$. Let $Y_i \sim \mathcal{N}(0, I)$ be independent, for $i = 1, \ldots, m$, where we set*

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

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 $\delta_3 = O(\varepsilon \log^2 1/\varepsilon)$ so that if $Y_1, \ldots, Y_N$, and $Z_1, \ldots, Z_m$ be as above, then we have

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

17

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 so that removing points cannot change the mean by too much. The proof appears in Section B.1.

4.3 Finding the Mean, Using a Separation Oracle

In this section, we consider the problem of approximating μ 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 non-empty for reasonable values of δ (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 μ. The challenge is that since μ 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_δ. 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:

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

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

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

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

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 an 2ε-fraction of the remaining points.

Then we have the following fact:

Fact 4.18. *Suppose that (7) holds. Then* NAIVEPRUNE *removes no uncorrupted points, and moreover, if X_i is not removed by* NAIVEPRUNE, *we have* $\|X_i - \mu\|_2 \leq O\left(\sqrt{d \log(N/\tau)} \right)$.

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.

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

Henceforth, for simplicity we shall assume that no point was removed by NAIVEPRUNE, and that for all $i = 1, \ldots, m$, we have $\|X_i - \mu\|_2 < O(\sqrt{d \log(m/\tau)})$. 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 $\mathcal{C}_\delta$. Our first step is to show that any set of weights that does not yield a good estimate for μ cannot be in the set $\mathcal{C}_\delta$:

Lemma 4.19. *Suppose that (8)-(9) holds. Let $\delta = \max(\delta_1, \delta_2)$. Let $w \in S_{N,\varepsilon}$ and set $\widehat{\mu} = \sum_{i=1}^N w_i X_i$ and $\Delta = \mu - \widehat{\mu}$. Further, suppose that $\|\Delta\| \geq \Omega(\delta)$. 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 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

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

On the other hand,

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

19

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

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

as claimed. □

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

Corollary 4.20. *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 4.21. *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 so that:*

1. *(Completeness) If $w = w^*$, then it outputs "YES".*

2. *(Soundness) If $w \notin \mathcal{C}_{c\delta}$, the algorithm outputs a hyperplane $\ell : \mathbb{R}^m \to \mathbb{R}$ so that $\ell(w) \geq 0$ but $\ell(w^*) < 0$. Moreover, if the algorithm ever outputs a hyperplane ℓ, then $\ell(w^*) < 0$.*

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

Remark 4.22. *The conditions that the separation oracle given here satisfy 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 $\mathcal{C}_{c\delta}$, the separation oracle acts exactly as a separation oracle for w^*. Thus, as long as the algorithm continues to query points outside of $\mathcal{C}_{c\delta}$, the action of the algorithm is equivalent to one with a separation oracle for w^*. If*

The separation oracle is given in Algorithm 2.

Algorithm 2 Separation oracle sub-procedure for agnostically learning the mean.

1: **function** SEPARATIONORACLEUNKNOWNMEAN($w, X_1, \ldots, X_N$)
2:     Let $\widehat{\mu} = \sum_{i=1}^{N} w_i X_i$.
3:     For $i = 1, \ldots, m$, define $Y_i = X_i - \widehat{\mu}$.
4:     Let λ be the eigenvalue of largest magnitude of $M = \sum_{i=1}^{N} w_i Y_i Y_i^T - I$
5:     Let v be its associated eigenvector.
6:     **if** $|\lambda| < \frac{c}{2}\delta$ **then**
7:         **return** "YES".
8:     **else if** $\lambda > \frac{c}{2}\delta$ **then**
9:         **return** the hyperplane $\ell(w) = \left(\sum_{i=1}^{N} w_i \langle Y_i, v \rangle^2 - 1 \right) - \lambda$.
10:    **else**
11:        **return** the hyperplane $\ell(w) = \lambda - \left(\sum_{i=1}^{N} w_i \langle Y_i, v \rangle^2 - 1 \right)$.

Next, we prove correctness for our approximate separation oracle:

Proof of Theorem 4.21. Again, let $\Delta = \mu - \widehat{\mu}$. By expanding out the formula for M, we get:

$$\sum_{i=1}^{N} w_i Y_i Y_i^T - I = \sum_{i=1}^{N} w_i(X_i - \mu + \Delta)(X_i - \mu + \Delta)^T - I$$

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

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

Let us now prove completeness.

Claim 4.23. *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 4.20 we have that $\|\Delta\|_2 \leq O(\varepsilon\sqrt{\log 1/\varepsilon})$. Thus

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

$$\leq \delta_1 + O(\varepsilon^2 \log 1/\varepsilon) < \frac{c\delta}{2} .$$

$\square$

We now turn our attention to soundness.

Claim 4.24. *Suppose that $w \notin \mathcal{C}_{c\delta}$. Then $|\lambda| > \frac{c}{2}\delta$.*

Proof. By the triangle inequality, we have

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

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 4.19, we have that

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

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

$\square$

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

$$\left\| \frac{1}{|G|} \sum_{i \in G}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^T - I \right\|_2 = \left\| \frac{1}{|G|} \sum_{i \in G}(X_i - \mu + \Delta)(X_i - \mu + \Delta)^T - I \right\|_2$$

$$\leq \underbrace{\left\| \frac{1}{|G|} \sum_{i \in G}(X_i - \mu)(X_i - \mu)^T - I \right\|_2}_{\leq \delta_1} + \underbrace{2\|\Delta\|_2 \left\| \frac{1}{|G|} \sum_{i \in G}(X_i - \mu) \right\|_2}_{\leq 2\delta_2 \|\Delta\|_2} + \|\Delta\|_2^2 \quad (11)$$

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. $\square$

4.3.3 The Full Algorithm

This separation oracle, along with the classical theory of convex optimization, implies that we have shown the following:

Corollary 4.25. *Fix $\varepsilon, \tau > 0$, and let $\delta = O(\varepsilon\sqrt{\log 1/\varepsilon})$. Let $X_1, \ldots, X_N$ be a ε-corrupted set of points satisfying (8)-(9), for $\delta_1, \delta_2 \leq \delta$. Let c be a sufficiently large constant. Then, there is an algorithm* LEARNAPPROXMEAN$(\varepsilon, \tau, X_1, \ldots, X_N)$ *which runs in time* $\mathrm{poly}(N, d, 1/\varepsilon, \log 1/\tau)$, *and outputs a set of weights $w' \in S_{N,\varepsilon}$ so that there is a $w \in C_{c\delta}$ so 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, ε') is a convex optimization routine, which given the SEPARATIONORACLEUNKNOWNMEAN separation oracle and a target error ε', outputs a w' so that $\|w - w'\|_\infty \leq \varepsilon'$. From the classical theory of optimization, we know such a routine exists and runs in polynomial time.

Algorithm 3 Convex programming algorithm for agnostically learning the mean.

1: **function** LEARNMEAN$(\varepsilon, \tau, X_1, \ldots, X_N)$
2:     Run NAIVEPRUNE$(X_1, \ldots, X_N)$. Let $\{X_i\}_{i\in I}$ be the pruned set of samples.
3:     For simplicity assume $I = [N]$.
4:     Let $w' \leftarrow$ LEARNAPPROXMEAN$(\varepsilon, \tau, X_1, \ldots, X_N)$.
5:     **return** $\sum_{i=1}^N w_i' X_i$.

We have:

Theorem 4.26. *Fix $\varepsilon, \tau > 0$, and let $\delta = O(\varepsilon\sqrt{\log 1/\varepsilon})$. Let $X_1, \ldots, X_N$ be an ε-corrupted set of samples, where*

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

Let $\widehat{\mu}$ be the output of LEARNMEAN$(X_1, \ldots, X_N)$. *Then, we have $\|\widehat{\mu} - \mu\|_2 \leq \delta$.*

Proof. By Fact 4.8, Lemma 4.5, and Lemma 4.6, we know that (7)-(9) hold with probability $1 - \tau$, with $\delta_1, \delta_2 \leq \delta$. Condition on the event that this event holds. After NAIVEPRUNE, by Fact 4.18 we may assume that no uncorrupted points are removed, and all points satisfy $\|X_i - \mu\|_2 \leq O(\sqrt{d \log(N/\tau)})$. Let w' be the output of the algorithm, and let $w \in C_{c\delta}$ be so that $\|w - w'\|_\infty < \varepsilon/(N\sqrt{d \log(N/\tau)})$. By Corollary 4.20, we know that $\|\sum_{i=1}^N w_i X_i - \mu\|_2 \leq O(\delta)$. Hence, we have

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

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

4.4 Finding the Covariance, Using a Separation Oracle

In this section, we consider the problem of approximating μ given m samples from $\mathcal{N}(0, \Sigma)$ in the full adversary model. Let $U_i = \Sigma^{-1/2} X_i$ so 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

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

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

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

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

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

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

for all $w \in S_{N,\varepsilon}$, and all sets $T \subseteq G$ of size $|T| \leq 2\varepsilon N$. 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 4.8, Corollary 4.11, Corollary 4.12, and Theorem 4.17 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 $\delta \geq \Omega(\varepsilon \sqrt{\log 1/\varepsilon})$ we have that $w^* \in C_\delta$.

Theorem 4.27. *Let* $\delta = O(\varepsilon \log 1/\varepsilon)$. *Suppose that (13), (14), and 15 hold for* $\delta_1, \delta_2 \leq O(\delta)$ *and* $\delta_3 \leq O(\delta \log 1/\varepsilon)$. *Then, there is a constant* c *and an algorithm so that, given any input* $w \in S_{N,\varepsilon}$ *we have:*

1. *(Completeness) If* $w = w^*$, *the algorithm outputs "YES".*

2. *(Soundness) If* $w \notin C_{c\delta}$, *the algorithm outputs a hyperplane* $\ell : \mathbb{R}^m \to \mathbb{R}$ *so that* $\ell(w) \geq 0$ *but we have* $\ell(w^*) < 0$. *Moreover, if the algorithm ever outputs a hyperplane* ℓ, *then* $\ell(w^*) < 0$.

By the classical theory of convex optimization this implies that we will find a point w *so that* $\|w - w'\|_\infty \leq \frac{\varepsilon}{\text{poly}(N)}$ *for some* $w' \in C_{c\delta}$, *using polynomially many calls to this oracle.*

The pseudocode for the separation oracle is given in Algorithm 4. Observe briefly that this algorithm does indeed run in polynomial time. Lines 2-7 require only taking top eigenvalues and eigenvectors, and so can be done in polynomial time. For any $\xi \in \{-1, +1\}$, line 8 can be run by sorting the samples by $w_i \left(\frac{\|Y_i\|^2}{\sqrt{d}} - \sqrt{d} \right)$ and seeing if there is a subset of the top $2\varepsilon N$ samples satisfying the desired condition, and line 9 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 λ be its associated eigenvalue.
5:     **if** $|\lambda| > \Omega(\varepsilon \log^2 1/\varepsilon)$ **then**
6:         Let $\xi = \text{sgn}(\lambda)$.
7:         **return** the hyperplane

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

8:     **else if** there exists a sign $\xi \in \{-1, 1\}$ and a set T of samples of size at most $2\varepsilon N$ so 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
9:         **return** the hyperplane

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

10:         **return** "YES".

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

Lemma 4.28. *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 $M^{-1} = \text{diag}(1 + \nu_1, \ldots, 1 + \nu_d)$, where by assumption $\sum \nu_i^2 < \delta^2/4$. By our assumption that $\delta < 1$, we have $|\nu_i| \leq 1/2$. 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. $\square$

Lemma 4.29. *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. $\qquad\blacksquare$

Lemma 4.30. *Let w_i for $i = 1, \ldots, N$ be a set of non-negative weights so that $\sum_{i=1}^{N} w_i = 1$, and let $a_i \in \mathbb{R}$ be arbitrary. Then*

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

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 $\operatorname{Var}_{X \sim P}[X] = \mathbb{E}_{X \sim P}[X^2] - \mathbb{E}_{X \sim P}[X]^2$ is always a non-negative quantity, by rearranging the desired conclusion follows. $\qquad\blacksquare$

Lemma 4.31. *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. $\qquad\blacksquare$

Proof of Theorem 4.27. Let us first prove completeness. Observe that by Theorem 4.15, we know that restricted to $\mathcal{S}$, we have that $M_4 = 2I$. Therefore, by (15) we will not output a hyperplane in line 7. Moreover, by (14), we will not output a hyperplane in line 8. 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 4.32. *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. $\qquad\blacksquare$

This allows us to show:

Claim 4.33. *Assume (13)–(15) hold with $\delta_1, \delta_2 \leq O(\delta)$ and $\delta_3 \leq O(\delta \log 1/\varepsilon)$, and assume furthermore that $\|A - I\|_F \geq c\delta$. Then, if we let $\delta' = \frac{(1-\varepsilon)c}{2}\delta = \Theta(\delta)$, we have*

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

Proof. Combining Lemma 4.28 and Fact 4.32 we have

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

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 $\|R\|_F \leq \delta_1$. By the definition of $\widehat{\Sigma}$ we have that $\sum_{i=1}^N w_i Y_i Y_i^T = I$, and so

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

Furthermore we have

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

by applying Lemma 4.29. And putting it all together we have

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

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

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

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 Pythagorean theorem. $\square$

Claim 4.33 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 4.34. *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 7, 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 7. 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 $M^\flat \in S$ so that $\|M^\flat\|_2 = \|M\|_F = 1$ and so 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 Lemma 4.30 and so this implies that

$$\sum_{i \in E} w_i \langle M^\flat, Z_i \rangle^2 \geq O\left(\frac{\delta'^2}{w_b}\right)$$

$$\geq O\left(\frac{\delta'^2}{\varepsilon}\right) \tag{18}$$

Let $\widetilde{\Sigma} = \widehat{\Sigma}^{-1}\Sigma$. By Theorem 4.15 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,

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

$$\leq \|\widetilde{\Sigma} - I\|_F^2 + 2\|\widetilde{\Sigma} - I\|_2 + (1 - w_g) + \|R\|\|\widetilde{\Sigma}\|^2$$

$$\leq 3\|\widetilde{\Sigma} - I\|_F^2 + \delta\|\widetilde{\Sigma}\|^2 + O(\varepsilon) \ .$$

$$\leq O\left(\delta'^2 + \delta'\right) \ , \tag{19}$$

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 λ, and so the algorithm will output a hyperplane in line 7. Letting ℓ 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. $\qquad\square$

Now let us assume that the first term on the LHS is less than $\frac{1}{2}\delta'$, so that the algorithm does not necessarily output a hyperplane in line 7. Thus, the second term on the LHS of Equation 16 is at least $\frac{1}{2}\delta'$. We now show that this implies that this implies that the algorithm will output a separating hyperplane in line 9.

Claim 4.35. *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 9, and moreover, it is a separating hyperplane.

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

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

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

It now suffices to show that $\ell(w^*) < 0$. 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,

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

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

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

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

These two claims in conjunction directly imply the correctness of the theorem. $\qquad\square$

4.4.1 The Full Algorithm

As before, this separation oracle and the classical theory of convex optimization shows that we have demonstrated an algorithm FINDAPPROXCOVARIANCE with the following properties:

Theorem 4.36. *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 ε-corrupted set of points satisfying (13-(15), for $\delta_1, \delta_2 \leq O(\delta)$ and $\delta_3 \leq O(\delta \log 1/\varepsilon)$. Then FINDAPPROXCOVARIANCE$(\varepsilon, \tau, X_1, \ldots, X_N)$ runs in time $\mathrm{poly}(N, d, 1/\varepsilon, \log 1/\tau)$, and outputs a u so that there is some $w \in C_{c\delta}$ so that $\|w - u\|_\infty \leq \varepsilon/(Nd \log(N/\tau))$.*

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 so that $\|\Sigma^{-1/2} X_i\|_2$ is extremely large. We show here that we can postprocess the u so that we can weed out these points. First, observe that we have the following lemma:

Lemma 4.37. *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 $\widetilde{w}^- \in R^N$ 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:

Lemma 4.38. *Let u be so that there is $w \in C_{c\delta}$ so that $\|u-w\|_\infty \leq \varepsilon/(Nd \log(N/\tau))$. 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 $\widetilde{u}_i^- \leq w_i$ 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.

We now show that this algorithm is correct.

Theorem 4.39. *Let $1/2 \geq \varepsilon > 0$, and $\tau > 0$. Let $\delta = O(\varepsilon \log 1/\varepsilon)$. Let $X_1, \ldots, X_N$ be a ε-corrupted set of samples from $\mathcal{N}(0, \Sigma)$ where*

$$N = \widetilde{\Omega}\left(\frac{d^3 \log^2 1/\tau}{\varepsilon^2}\right).$$

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

Algorithm 5 Full algorithm for learning the covariance agnostically

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

Proof. We first condition on the event that we satisfy (12)-(15) with $\delta_1, \delta_2 \leq O(\delta)$ and $\delta_3 \leq O(\delta \log 1/\varepsilon)$. By our choice of N, Fact 4.8, Corollary 4.10, Corollary 4.12, and Theorem 4.17, and a union bound, we know that this event happens with probability $1 - \tau$.

By Theorem 4.36, Lemma 4.37, and Lemma 4.38, we have that since ε 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, 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 I be the set of pruned points. Because no uncorrupted point is thrown out, we have that $|I| \geq (1 - 2\varepsilon)N$, and moreover, this set of points still satisfies (13)-(15)[3] and moreover, for ever $i \in I$, we have $\|\Sigma^{-1/2} X_i\|_2^2 \leq O(d \log N/\tau)$. Therefore, by Theorem 4.36, we have that there is some $u'' \in C_{c|I|}$ so 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

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

which completes the proof. $\qquad\square$

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 zero 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 ε-corrupted set of samples from $\mathcal{N}(\mu, \Sigma)$, where both μ and Σ are unknown. We will set

$$N = \widetilde{\Omega}\left(\frac{d^3 \log^2 1/\tau}{\varepsilon^2}\right) .$$

---

[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.

4.5.1 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 ε-corrupted, then the $X_1', \ldots, X_{N/2}'$ is a $N/2$-sized set of samples which is 2ε-corrupted. Thus, by using the results from Section 4.4, with probability $1 - \tau$, we can recover a $\widehat{\Sigma}$ so 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

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

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 $X_i'' \sim \mathcal{N}(0, \Sigma_1)$, where $\Sigma_1 = \widehat{\Sigma}^{-1/2}\Sigma\widetilde{\Sigma}^{-1/2}$. By Corollary 2.14 and Lemma 4.28, 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 2.4, 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 a $O(\varepsilon \log 1/\varepsilon)$-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 $\widehat{\mu}$ so 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 the following:

Algorithm 6 Algorithm for learning an arbitrary Gaussian robustly

1: **function** RECOVERROBUSTGUASSIAN($\varepsilon, \tau, X_1, \ldots, X_N$)
2:     For $i = 1, \ldots, N/2$, let $X_i' = (X_i - X_{N/2+i})/\sqrt{2}$.
3:     Let $\widehat{\Sigma} \leftarrow$ LEARNCOVARIANCE($\varepsilon, \tau, X_1', \ldots, X_{N/2}'$).
4:     For $i = 1, \ldots, N$, let $X_i'' = \widehat{\Sigma}^{-1/2} X_i$.
5:     Let $\widehat{\mu} \leftarrow$ LEARNMEAN($\varepsilon, \tau, X_1'', \ldots, X_N''$).
6:     **return** the Gaussian with mean $\widehat{\Sigma}^{-1/2}\widehat{\mu}$, and covariance $\widehat{\Sigma}$.

Theorem 4.40. *Fix $\varepsilon, \tau > 0$. Let $X_1, \ldots, X_N$ be an ε-corrupted set of samples from $\mathcal{N}(\mu, \Sigma)$, where μ, Σ are both unknown, and*

$$N = \widetilde{\Omega} \left(\frac{d^3 \log^2 1/\tau}{\varepsilon^2} \right) .$$

There is a polynomial-time algorithm RECOVERROBUSTGAUSSIAN$(\varepsilon, \tau, X_1, \ldots, X_N)$ *which with probability $1 - \tau$, outputs a $\widehat{\Sigma}, \widehat{\mu}$ so 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 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-GMM each of whose components are spherical. For $X \sim \mathcal{M}$, we write $X \sim_j \mathcal{M}$ if X was drawn from the jth component of $\mathcal{M}$.

Our main result of this section is the following theorem:

Theorem 5.1. *Fix $\varepsilon, \tau > 0$, and $k \in \mathbb{N}$. Let $X_1, \ldots, X_N$ be an ε-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 α_j, μ_j, and σ_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, LS15]. 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 mean lie. However, as we have observed already, 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 as 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 5.1. Section 5.2 states the various concentration inequalites we require. In Section 5.3, we go over our overall algorithm in more detail. Section 5.4 describes a first naive clustering step, which deals with components which are very well separated. Section 5.5 contains details on our separation oracle approach, allowing us to approximate the true covariance. Section 5.6 describes our spectral clustering approach to cluster components with means separated more than $\Omega_k(\log 1/\varepsilon)$. In Section 5.7, we describe how to exhaustively search over a particular subspace to obtain a good estimate for the component means. In Section 5.8, we go over how to limit the set of hypotheses in order to satisfy the conditions of Lemma 2.9. For clarity of exposition, all of the above describe the algorithm assuming all σ_j^2 are equal. In Section 5.9, we discuss the changes to algorithm which are required to handle unequal variances.

For clarity of presentation, many of the proofs are deferred to Section C.

5.1 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

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

Finally, we use the notation

$$Q = \sum_{j\in[k]} \alpha_j(\mu_j - \mu)(\mu_j - \mu)^T. \tag{22}$$

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

Definition 5.2. *For any symmetric matrix $M \in \mathbb{R}^{d\times d}$ with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_d$, 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 easy to see that the Schatten top-k norm is indeed a norm, as its name suggests:

Fact 5.3. $\|M\|_{T_k}$ *is a norm on symmetric matrices.*

Proof. The only property that is not immediate, that we need to verify, is that it satisfies the triangle inequality. Indeed, we can lower bound $\|M\|_{T_k}$ in terms of its dual characterization

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

where the maxima is taken over all X with orthonormal columns. And now the fact that the triangle inequality holds is immediate. □

5.2 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 5.1. The following two concentration lemmas follow from the same proofs as for Lemmas 42 and 44 in [LS15].

Lemma 5.4. *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 equation (22).

Lemma 5.5. *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 5.6. *Fix $\varepsilon \leq 1/2$ and $\tau \leq 1$. There is a $\delta = O(\varepsilon\sqrt{\log 1/\varepsilon})$ so 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*

$$\Pr\left[\exists w \in S_{N,\varepsilon} : \left\|\sum_{i=1}^{N} w_i(Y_i - \mu)(Y_i - \mu)^T - I - Q\right\|_2 \geq f(k, \gamma, \delta_1)\right] \leq \tau, \tag{23}$$

where Q is defined as in equation (22).

Lemma 5.7. *Fix $\varepsilon \leq 1/2$ and $\tau \leq 1$. There is a $\delta = O(\varepsilon\sqrt{\log 1/\varepsilon})$ so 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*

$$\Pr\left[\exists w \in S_{N,\varepsilon} : \left\|\sum_{i=1}^{N} w_i Y_i - \mu\right\|_2 \geq k^{1/2}\delta_2\right] \leq \tau. \tag{24}$$

5.3 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 $|\mathcal{S}| = \text{poly}(d, 1/\varepsilon, \log(1/\tau))^{k^2}$ with the guarantee that there is some $\mathcal{M}^* \in \mathcal{S}$ such that $d_{\text{TV}}(\mathcal{M}, \mathcal{M}^*) \leq \tilde{O}(\text{poly}(k) \cdot \sqrt{\varepsilon})$. 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 σ^2 and performing naive clustering until later. For reasons we justify in Section 5.8, 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:

$$\frac{|\{X_i \in G, X_i \sim_j \mathcal{M} : \|X_i - \mu_j\|_2^2 \geq \Omega(d\log k/\varepsilon)\}|}{|\{X_i \in G, X_i \sim_j \mathcal{M}\}|} \leq \varepsilon/k, \forall j = 1, \ldots, N \tag{25}$$

$$\left\|\sum_{i\in G} w_i(X_i - \mu)(X_i - \mu)^T - w_g I - w_g Q\right\|_2 \leq f(k, \gamma, \delta_1) \; \forall w \in S_{N,4\varepsilon}, \text{ and} \tag{26}$$

$$\left\|\sum_{i\in G} w_i(X_i - \mu)\right\|_2 \leq k^{1/2}\delta_2 \; \forall w \in S_{N,4\varepsilon}. \tag{27}$$

(25) follows from basic Gaussian concentration, and (26) and (27) follow from the results in Section 5.2 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 right-hand side 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 η. 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 η 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 5.5.

After this procedure, Lemma 5.13 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 about the true covariance matrix. Thus, we will appeal to their analysis, highlighting the differences between our approach and theirs. We note that [LS15] also follows a similar approach and the interested reader may also adapt their arguments instead.

First, if γ 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 5.6.

Finally, if γ is sufficiently small, we may perform an exhaustive search. The span of the means is the top $k - 1$ eigenvectors of the true covariance matrix, which we can approximate with our weighted empirical covariance matrix. Since γ 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 5.7.

To conclude our algorithm, we can apply Lemma 2.9. We note that this hypothesis selection problem has been studied before (see, i.e. [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 2.9 requires the hypotheses to come from some fixed finite set, while there are an infinite number of Gaussian mixture models. In Section 5.8, we discuss how to limit the number of hypotheses based on the set of uncorrupted samples in order to satisfy the conditions of Lemma 2.9.

5.4 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 $C_1, \ldots, C_{k'}$ be formed by merging clusters in $\mathcal{C}$ if they overlap, and stopping if no clusters overlap. We give the formal pseudocode in Algorithm 7.

Algorithm 7 Naive clustering algorithm for spherical GMMs.

1: **function** NAIVECLUSTERGMM($X_1, \ldots, X_n$)
2:     **for** $i = 1, \ldots, N$ **do**
3:         Let $S_i = \{i' : \|X_i - X_{i'}\|_2^2 \leq \Theta(dk\log 1/\varepsilon)\}$.
        Let $\mathcal{C} = \{S_i : |S_i| \geq 4\varepsilon N\}$.
4:     **while** $\exists C, C' \in \mathcal{C}$ so that $C \neq C'$ and $C \cup C' \neq \varnothing$ **do**
5:         Remove C, C' from $\mathcal{C}$
6:         Add $C \cup C'$ to $\mathcal{C}$
7:     **return** the set of clusters $\mathcal{C}$

We prove here that this process (which may throw away points) throws away only at most a ε fraction of

good points, and moreover, the resulting clustering only misclassifies at most an $O(\varepsilon)$-fraction of the good points, assuming (25).

Theorem 5.8. *Let $X_1, \ldots, X_m$ be a set of samples satisfying (25). Let $C_1, \ldots, C_{k'}$ be the set of clusters returned. For each component j, let $\ell(j)$ be the ℓ so that C_ℓ contains the most points from j. Then:*

1. *Then, for each ℓ, there is some j so 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 ε-corrupted set of samples from $\mathcal{M}$, we may cluster them in a way which misclassifies at most an ε/k fraction of the samples from any component, and so that within each cluster, the means of the associated components differ by at most $dk \log k/\varepsilon$. Thus, each separate cluster is simply a ε-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), so that $\|X_j - X_{j'}\|_2^2 \geq \Omega(dk \log 1/\varepsilon)$.

5.5 Estimating the Covariance Using Convex Programming

In this section, we will apply our separation oracle framework to estimate the covariance matrix. While in the non-agnostic 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 5.5.1, we prove lemmas indicating important properties of this set. In Section 5.5.2, we give a separation oracle for this convex set. We conclude with Lemma 5.13, which shows that we have obtained an accurate estimate of the true covariance.

5.5.1 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 5.9. *Suppose that (26) 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 μ, then it is not in our convex set.

Lemma 5.10. *Suppose that (26) and (27) 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 μ:

Corollary 5.11. *Suppose that (26) and (27) hold. Let $w \in \mathcal{C}_{h(k,\gamma,\delta)}$ for $\delta = \Omega(\varepsilon \log 1/\varepsilon)$. Then*

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

5.5.2 Separation Oracle

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

Theorem 5.12. *Fix $\varepsilon > 0$, and let $\delta = \Omega(\varepsilon \log 1/\varepsilon)$. Suppose that (26) and (27) hold. Let w^* denote the weights which are uniform on the uncorrupted points. Then there is a constant c and an algorithm so that:*

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 $\ell : \mathbb{R}^m \to \mathbb{R}$ so 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' so 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 Section C.1.

Algorithm 8 Separation oracle sub-procedure for agnostically learning the span of the means of a GMM.

1: **function** SEPARATIONORACLEGMM(w)
2:     Let $\widehat{\mu} = \sum_{i=1}^{N} w_i X_i$.
3:     For $i = 1, \ldots, N$, define $Y_i = X_i - \widehat{\mu}$.
4:     Let $M = \sum_{i=1}^{N} w_i Y_i Y_i^T - I$.
5:     **if** $\|M\|_{T_k} < \sum_{j \in [k]} \gamma_j + \frac{ckh(k,\gamma,\delta)}{2}$ **then**
6:         **return** "YES".
7:     **else**
8:         Let $\Lambda = \|M\|_{T_k}$.
9:         Let U be a $d \times k$ matrix with orthonormal columns which span the top k eigenvectors of M.
10:        **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 5.13. *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 triangle inequality and Corollary 5.11,

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

$\square$

5.6 Spectral Clustering

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

Lemma 5.14. *Given a weight vector w as output by Algorithm 8, if $\gamma \ge \Omega(\mathrm{poly}(k) \cdot \log 1/\varepsilon)$, there exists an algorithm which produces a unit vector v with the following guarantees:*

- *There exists a non-trival 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 $O(\mathrm{poly}(\varepsilon/k))$, 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]. Their main concentration lemma is Lemma 30, which states that they obtain a good estimate of the true covariance matrix, akin to our Lemma 5.13.

Lemma 31 argues that the largest eigenvector of this estimate is highly correlated with the top eigenvector of the true covariance matrix. Since γ 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 $\gamma = \Omega\left(\text{poly}(k) \cdot \log(d/\varepsilon)\right)$. 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 γ at the start of the step, and as such, using the same approach as them 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 $\text{poly}(\varepsilon/k)$-fraction of points are misclustered. As such, we can continue spectral clustering until $\gamma = O\left(\text{poly}(k) \cdot \log(1/\varepsilon)\right)$.

5.7 Exhaustive Search

The final stage of the algorithm is when we know that all γ_i's are sufficiently small. We can directly apply Lemma 7 of [SOAJ14], which states the following:

Lemma 5.15. *Given a weight vector w as output by Algorithm 8, 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

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

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 2γ centered at $\widehat{\mu}$. By taking an η-mesh over the points in this cube (for $\eta = \text{poly}(\varepsilon/dk)$ 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

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

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

5.8 Applying the Tournament Lemma

In this section, we discuss details about how to apply our hypothesis selection algorithm. First, in Section 5.8.1, we describe how to guess the mixing weights and the variance of the components. Then in Section 5.8.2, 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.

5.8.1 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 ε. If our misspecified parameters are accurate enough, the agnostic learning guarantee remains unchanged.

Guessing the mixing weights is fairly straightforward. For some $\nu = \mathrm{poly}(\varepsilon/k)$ sufficiently small, our algorithm generates a set of at most $(1/\nu)^k = \mathrm{poly}(k/\varepsilon)^k$ possible mixing weights by guessing the values $\{0, \varepsilon, \varepsilon + \nu, \varepsilon + 2\nu, \ldots, 1 - \nu, 1\}$ for each α_j. Note that we may assume each weight is at least ε, 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 σ^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 $k << 1/\varepsilon$, 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 $\mathrm{poly}(\varepsilon/dk)$ to get an estimate for σ^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 $\sigma^2 = 1$.

5.8.2 Pruning Our Hypotheses

In this section, we describe how to prune our set of hypotheses in order to apply Lemma 2.9. 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 2.9 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 are 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 so 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 εN 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 εN points within distance ζ. We prune our set of samples by removing any point with fewer than εN other points at distance less than 2ζ. 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 3ζ from all component means, due to the fact that the adversary may only move an ε-fraction of the points, and reverse triangle inequality.

Now, we consider the union of the balls of radius 3ζ 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 6ζ 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 ε/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 ε/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 2.9 are satisfied and we may run this procedure to select a sufficiently accurate hypothesis.

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

---

[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.

Lemma 5.16 (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_ℓ be the set of all points with at least εN points at squared-distance at most $2\left(1 + \frac{1}{k}\right)^{\ell-1}\sigma_{min}^2\left(1 + 4\sqrt{\frac{\log \frac{N^2}{\tau}}{d}}\right)$, for $\ell \in [k]$. Note that $H_\ell \subseteq H_{\ell+1}$. Let ℓ^* be the minimum ℓ such that $H_\ell = H_{\ell+1}$, or k if no such ℓ exists, and partition the set of samples into H_{ℓ^*} and $\overline{H}_{\ell^*}$. This partition will contain all samples from components with variance at most some threshold t, where $t \leq e\sigma_{min}^2$ in H_{ℓ^*}. All samples from components with variance at least t will fall into $\overline{H}_{\ell^*}$. We continue running the algorithm with H_{ℓ^*}, 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 σ_{min}^2, by the upper bound in Lemma 5.16. However, it may also contain samples from other components – in particular, those with variance at most $\gamma\sigma_{min}^2$, for

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

where the second inequality follows for d sufficiently large. Therefore, we compute H_2, which contains all samples from such components. This is repeated up to at most H_k, since if a set $H_{\ell+1}$ is distinct from H_ℓ, 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_{ℓ^*}.

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

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

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:

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

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.

6 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 coordinate-wise mean, which we denote by the vector p, and our first

---

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

goal is to estimate p within ℓ_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 off 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 ℓ_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 6.2) in Section 6.1.

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 ℓ_2-distance as a starting point for handling binary product distributions that have imbalanced coordinates. Instead, we control the total variation distance via the χ^2-distance between the mean vectors. Let P and Q be two product distributions whose means are p and q respectively. From Lemma 2.15, 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 ℓ_2-distance to p, we can interpret the right hand side above as giving a renormalization of how we should measure the distance between p and q so that being close (in χ^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 Σ 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 6.2 for further details.

6.1 The Balanced Case

The main result of this section is the following theorem:

Theorem 6.1. *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 ε-fraction of the points in S. There exists a polynomial time algorithm that returns a product distribution P' so 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 6.1 applies to all binary product distributions, and its performance guarantee relates the ℓ_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 ℓ_2-guarantee implies a similar total variation guarantee. Formally, we have:

Definition 6.2. *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 2.15:

Fact 6.3. *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 ℓ_2-distance between their mean vectors is a good proxy for their total variation distance. Using Fact 6.3, we obtain the following corollary of Theorem 6.1:

Corollary 6.4. *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 ε-fraction of the points in S. There exists a polynomial time algorithm that returns a product distribution P' so 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 6.5 (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 ε-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 ε-good (with respect to P) with high probability.

Lemma 6.6. *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 ε-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 τ for $N = \Omega((d^4 + d^2 \log(1/\tau))/\varepsilon^2)$. $\qquad\square$

The following definition quantifies the extent to which a multiset has been corrupted.

Definition 6.7. *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.*

Note that the definition of $\Delta(S, S')$ is not symmetric in its two arguments.

Our agnostic learning algorithm establishing Theorem 6.1 is obtained by repeated application of the efficient procedure whose performance guarantee is given in the following proposition:

Proposition 6.8. *Let P be a binary product distribution with mean vector p and $\varepsilon > 0$ be sufficiently small. Let S be ε-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 6.1 follows easily from Proposition 6.8.

Proof of Theorem 6.1. By the definition of $\Delta(S, S')$, since S' has been obtained from S by corrupting an ε-fraction of the points in S, we have that $\Delta(S, S') \leq 2\varepsilon$. By Lemma 6.6, the set S of uncorrupted samples is ε-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 6.8 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 $S_0 = S', S'_1, \ldots$, so 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. $\square$

6.1.1 Algorithm FILTER-BALANCED-PRODUCT: Proof of Proposition 6.8

In this section, we describe the efficient procedure establishing Proposition 6.8 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 λ^* 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 9 Filter algorithm for a balanced binary product distribution

1: **procedure** FILTER-BALANCED-PRODUCT(ε, S')
input: A multiset S' such that there exists an ε-good S with $\Delta(S, S') \leq 2\varepsilon$
output: Multiset S'' or mean vector p' satisfying Proposition 6.8
2:     Compute the sample mean $\mu^{S'} = \mathbb{E}_{X \in_u S'}[X]$ and the sample covariance M with zeroed diagonal,
3:     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 $M_{i,i} = 0$.
4:     Compute approximations for the largest absolute eigenvalue of M, $\lambda^* := \|M\|_2$, and the associated unit eigenvector v^*.
5:     **if** $\|M\|_2 \leq O(\varepsilon \log(1/\varepsilon))$, **then return** $\mu^{S'}$.
6:     Let $\delta := 3\sqrt{\varepsilon\|M\|_2}$. Find $T > 0$ such that

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

7:     **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 ε-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 ε-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 ε-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 ℓ_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.

6.1.2 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 ε 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 6.9. *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 ε-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. $\square$

The following sequence of lemmas 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 6.10. *If S is ε-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:

$$\begin{aligned}
(M_P)_{i,j} &= \mathop{\mathbb{E}}_{X \sim P}[(X_i - \mu_i^{S'})(X_j - \mu_j^{S'})] \\
&= \mathop{\mathbb{E}}_{X \sim P}[X_i X_j] - \mu_i^{S'} \mathop{\mathbb{E}}_{X \sim P}[X_j] - \mu_j^{S'} \mathop{\mathbb{E}}_{X \sim P}[X_i] + \mu_j^{S'} \mu_i^{S'} \\
&= \mathop{\Pr}_{X \sim P}((e_i + e_j) \cdot X \geq 2) - \mu_i^{S'} \mathop{\Pr}_{X \sim P}(e_j \cdot X \geq 1) - \mu_j^{S'} \mathop{\Pr}_{X \sim P}(e_i \cdot X \geq 1) + \mu_j^{S'} \mu_i^{S'} .
\end{aligned}$$

A similar expression holds for M_S except with probabilities for $X \in_u S$. Since S is ε-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. $\square$

As a simple consequence of the above lemma, we obtain the following:

Claim 6.11. *If S is ε-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 6.10. $\square$

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 ℓ_2^2-norm between the corresponding mean vectors:

Lemma 6.12. *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. $\square$

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 6.13. *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

$$|S| \cdot \mathop{\Pr}_{X \in_u S}(X = x) \geq |L| \cdot \mathop{\Pr}_{X \in_u L}(X = x) . \tag{28}$$

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

$$|v^T M_L v| = \mathop{\mathbb{E}}_{X \in_u L}[|v \cdot (X - \mu^{S'})|^2].$$

47

The RHS is in turn bounded from above as follows:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{X \in_u L}[|v \cdot (X - \mu^{S'})|^2] &= 2 \int_0^{\sqrt{d}} \mathop{\Pr}_{X \in_u L} \left(|v \cdot (X - \mu^{S'})| > T \right) \cdot T dT \\
&\leq 2 \int_0^{\sqrt{d}} \min\left\{ 1, |S|/|L| \cdot \mathop{\Pr}_{X \in_u S} \left(|v \cdot (X - \mu^{S'})| > T \right) \right\} T dT \\
&\ll \int_0^{4\sqrt{\log(|S|/|L|)} + \|\mu^{S'} - p\|_2} T dT \\
&\quad + (|S|/|L|) \int_{4\sqrt{\log(|S|/|L|)} + \|\mu^{S'} - p\|_2}^{\sqrt{d}} \left(\exp(-(T - \|\mu^{S'} - p\|_2)^2/2)T + \varepsilon T/d \right) dT \\
&\ll \log(|S|/|L|) + \|\mu^{S'} - p\|_2^2 + \varepsilon \cdot |S|/|L| \,,
\end{aligned}
$$

where the second line follows from (28) and the third line follows from Claim 6.9. 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 $x \ln(1/x)$, and the fact that $|L|/|S| \leq 2\varepsilon$. The observation $|S|/|S'| \leq 1 + 2\varepsilon \leq 2$ completes the proof of the second part of the lemma. $\qquad \square$

Claim 6.11 combined with Lemmas 6.12 and 6.13 and the triangle inequality yield the following:

Corollary 6.14. *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.

6.1.3 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 $\|M\|_2 \leq O(\varepsilon \log(1/\varepsilon))$. We start with the following simple claim:

Claim 6.15. *Let* μ^E, μ^L *be the mean vectors of* E *and* L *respectively. Then,* $\|\mu^E - \mu^{S'}\|_2^2 \leq \|M_E\|_2$ *and* $\|\mu^L - \mu^{S'}\|_2^2 \leq \|M_L\|_2$.

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

$$
v^T M_E v = \mathop{\mathbb{E}}_{X \in_u E}[|v \cdot (X - \mu^{S'})|^2] \geq |v \cdot (\mu^E - \mu^{S'})|^2.
$$

Let $w = \mu^E - \mu^{S'}$ and take $v = w/\|w\|_2$. We conclude that $\|M_E\|_2 \geq \|w\|_2^2$, as desired. $\qquad \square$

The following crucial lemma, bounding from above the distance $\|\mu^{S'} - p\|_2$ as a function of ε and $\|M\|_2$, will be important for both this and the following subsections.

Lemma 6.16. *We have that* $\|\mu^{S'} - p\|_2 \leq 2\sqrt{\varepsilon \|M\|_2} + O(\varepsilon \sqrt{\log(1/\varepsilon)})$.

Proof. First we observe that the mean vector μ^S of the uncorrupted sample set S is close to p. Since S is ε-good, this follows from the fact that for any $i \in [d]$, we have

$$
|\mu_i^S - p_i| = |\mathop{\Pr}_{X \in_u S}[e_i \cdot X \geq 1] - \mathop{\Pr}_{X \sim P}[e_i \cdot X \geq 1]| \leq \varepsilon/d.
$$

Therefore, we get that $\|\mu^S - p\|_2 \le \varepsilon/\sqrt{d}$.

Consider μ^E and μ^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 6.15, while the third uses Lemma 6.13 and Corollary 6.14. Finally, the last couple of lines assume that ε is sufficiently small. The proof of Lemma 6.16 is now complete. $\qquad\square$

We can now deduce the correctness of Step 5 of the algorithm FILTER-BALANCED-PRODUCT, since for $\|M\|_2 \le O(\varepsilon \log(1/\varepsilon))$, Lemma 6.16 directly implies that $\|\mu^{S'} - p\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$.

6.1.4 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 6. This corresponds to the case that $\|M\|_2 \ge C\varepsilon \log(1/\varepsilon)$, for a sufficiently large universal constant $C > 0$. We will show that the multiset $S'' \subset S'$ computed in Step 6 satisfies $\Delta(S, S'') \le \Delta(S, S') - 2\varepsilon/d$.

We start by noting that, as a consequence of Lemma 6.16, we have the following:

Claim 6.17. *We have that $\|\mu^{S'} - p\|_2 \le \delta := 3\sqrt{\varepsilon\|M\|_2}$.*

Proof. By Lemma 6.16, we have that $\|\mu^{S'} - p\|_2 \le 2\delta/3 + O(\varepsilon\sqrt{\log(1/\varepsilon)})$. Recalling that $\|M\|_2 \ge C\varepsilon \log(1/\varepsilon)$, if $C > 0$ is sufficiently large, the term $O(\varepsilon\sqrt{\log(1/\varepsilon)})$ is at most $\delta/3$. $\qquad\square$

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

$$\mathop{\mathbb{E}}_{X \in_u E}[|v^* \cdot (X - \mu^{S'})|^2] = (v^*)^T M_E v^* \ge \frac{\delta^2 |S'|}{20\varepsilon|E|} , \tag{29}$$

where the equality holds by definition, and the inequality follows from Corollary 6.14 and Claim 6.17 using the fact that ε is sufficiently small and the constant C is sufficiently large (noting that the constant in the RHS of Corollary 6.14 does not depend on C).

We show that (29) implies the existence of a $T > 0$ with the properties specified in Step 6 of the algorithm FILTER-BALANCED-PRODUCT. More specifically, we have the following crucial lemma:

Lemma 6.18. *If $\|M\|_2 \ge C\varepsilon \log(1/\varepsilon)$, for a sufficiently large constant $C > 0$, there exists a $T > 0$ satisfying the property in Step 6 of the algorithm* 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

$$\Pr_{X \in_u S'} (|v^* \cdot (X - \mu^{S'})| \geq T + \delta) \leq 8 \exp(-T^2/2) + 8\varepsilon/d \ . \tag{30}$$

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 (30) implies that for all $T > 0$

$$\Pr_{Y \in_u E} (|v^* \cdot (Y - \mu^{S'})| \geq T + \delta) \ll (|S'|/|E|)(\exp(-T^2/2) + \varepsilon/d) \ . \tag{31}$$

Using (29) and (31), we have the following sequence of inequalities:

$$\delta^2 |S'|/(\varepsilon|E|) \ll \mathop{\mathbb{E}}_{Y \in_u E}[|v^* \cdot (Y - \mu^{S'})|^2]$$

$$= 2 \int_0^\infty \Pr_{Y \in_u E} \left(|v^* \cdot (Y - \mu^{S'})| \geq T\right) \cdot T dT$$

$$\ll (|S'|/|E|) \int_0^{O(\sqrt{d})} \min \left\{|E|/|S'|, \exp(-(T - \delta)^2/2) + \varepsilon/d\right\} T dT$$

$$\ll \int_0^{4\sqrt{\log(|S'|/|E|)}+\delta} T dt + \int_{4\sqrt{\log(|S'|/|E|)}+\delta}^{\infty} (|S'|/|E|) \exp(-(T - \delta)^2/2) T dT + \int_0^{O(\sqrt{d})} \frac{\varepsilon|S'|}{d|E|} T dT$$

$$\ll \log(|S'|/|E|) + \delta^2 + \frac{\varepsilon|S'|}{|E|} \ .$$

This yields the desired contradiction recalling that the assumption $\|M\|_2 \geq C\varepsilon \log(1/\varepsilon)$ and the definition of δ imply that $\delta \geq C'\varepsilon\sqrt{\log(1/\varepsilon)}$ for an appropriately large $C' > 0$. $\qquad\square$

The following simple claim completes the proof of Proposition 6.8:

Claim 6.19. *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 Claim 6.9 and Claim 6.17, 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 6 of the algorithm 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. $\qquad\square$

6.2 Agnostically Learning Arbitrary Binary Product Distributions

In this subsection, we build on the approach of the previous subsection to show the following:

Theorem 6.20. *Let P be a binary product distribution in d dimensions and $\varepsilon, \tau > 0$. There is a polynomial time algorithm that, given ε and a set of $\Theta(d^6 \log(1/\tau)/\varepsilon^3)$ independent samples from P, an ε fraction of which have been arbitrarily corrupted, outputs (the mean vector of) a binary product distribution $\widetilde{P}$ so that, with probability at least $1 - \tau$, $d_{\mathrm{TV}}(P, \widetilde{P}) \leq O(\sqrt{\varepsilon \log(1/\varepsilon)})$.*

By Lemma 2.15, the total variation distance between two binary product distributions can be bounded from above by the square root by the χ^2-distance between the corresponding means. For the case of balanced product distributions, the χ^2-distance and the ℓ_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 χ^2-distance and the ℓ_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 χ^2 in place of ℓ_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 somewhat weaker than in the balanced case.

Similarly 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 6.21 (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 (ε, η)-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*

$$|\Pr_{X \in_u S_T}(L(X) \geq 0) - \Pr_{X \sim P_T}(L(X) \geq 0)| \leq \varepsilon^{3/2}/d^2 .$$

We note that a sufficiently large set of samples from P will satisfy the above properties with high probability:

Lemma 6.22. *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 Section D.

We will also require a notion of the number of coordinates on which S non-trivially depends:

Definition 6.23. *For S a multiset of elements in $\{0, 1\}^d$, let $\mathrm{supp}(S)$ be the subset of $[d]$ consisting of indices i so that the ith coordinate of elements of S is not constant.*

Similarly to the balanced case, our algorithm is obtained by repeated application of an efficient filter procedure, whose precise guarantee is described below.

Proposition 6.24. *Let P be a binary product distribution in d dimensions and $\varepsilon > 0$. Suppose that S is an (ε, η)-good multiset with respect to P with $\eta > 10\varepsilon$ and S' be any multiset with $\Delta(S, S') \leq 2\varepsilon$. There exists a polynomial time algorithm which, given ε and S', returns one of the following:*

   (i) The mean vector of a product distribution P' with $d_{\mathrm{TV}}(P, P') = O(\sqrt{\varepsilon \ln(1/\varepsilon)})$.

(ii) A multiset S'' of elements of $\{0,1\}^d$ so 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

$$\Delta(\widetilde{S}, S'') + \varepsilon|\operatorname{supp}(S'')|/d + \|p - \widetilde{p}\|_1/6 < \Delta(S, S') + \varepsilon|\operatorname{supp}(S')|/d .$$

Our agnostic learning algorithm is then obtained by iterating this procedure. Indeed, the proof of Theorem 6.20 given Proposition 6.24 is quite simple.

Proof of Theorem 6.20. We draw $\Theta(d^6 \log(1/\tau)/\varepsilon^3))$ samples forming a set S, which is $(\varepsilon, 1/5)$ good with probability $1 - \tau$ by Lemma 6.22. We condition on this event. The adversary corrupts an ε-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 (ε, η_i)-good for some binary product distribution P_i and some sets S_i'. We note that $\Delta(S_i, S_i') + \varepsilon|\operatorname{supp}(S_i')|/d$ is monotonically decreasing and never more than 3ε. Since $|\mu^{P_i} - \mu^{P_{i+1}}| \leq d_{\mathrm{TV}}(P_i, P_{i+1})$, in the ith iteration, the aforementioned quantity decreases by at least $d_{\mathrm{TV}}(P_i, P_{i+1})$, and thus $\sum_i d_{\mathrm{TV}}(P_i, P_{i+1}) \leq 3\varepsilon$. Therefore, $d_{\mathrm{TV}}(P, P_k) \leq 3\varepsilon$. Our algorithm outputs a product distribution P' with $d_{\mathrm{TV}}(P_k, P') = O(\sqrt{\varepsilon \ln(1/\varepsilon)})$. The triangle inequality completes the proof. $\square$

6.2.1 Algorithm FILTER-PRODUCT: Proof of Proposition 6.24

In this section, we describe and analyze the efficient routine establishing Proposition 6.24. Our efficient filtering procedure is presented in detailed pseudocode below.

Algorithm 10 Filter algorithm for an arbitrary binary product distribution

1: **procedure** FILTER-PRODUCT(ε, S')

input: $\varepsilon > 0$ and multiset S' such that there exists an ε-good S with $\Delta(S, S') \leq 2\varepsilon$

output: Multiset S'' or mean vector p' satisfying Proposition 6.24

2:     Compute the sample mean $\mu^{S'} = \mathbb{E}_{X \in_u S'}[X]$ and the sample covariance matrix M

3:     i.e., $M = (M_{i,j})_{1 \leq i,j \leq d}$ with $M_{i,j} = \mathbb{E}_{X \in S'}[(X_i - \mu_i^{S'})(X_j - \mu_j^{S'})]$.

4:     **if** there exists $i \in [d]$ with $0 < \mu_i^{S'} < \varepsilon/d$ or $0 < 1 - \mu_i^{S'} < \varepsilon/d$, **then**

5:         let S'' be the subset of elements of S' in which those coordinates take their most common value.

6:         **return** S''.

      /* For the later steps, we ignore any coordinates not in $\operatorname{supp}(S')$. */

7:     Compute approximations for the largest magnitude eigenvalue λ' of DMD, $\lambda' := \|DMD\|_2$, where $D = \operatorname{Diag}(1/\sqrt{\mu_i^{S'}(1 - \mu_i^{S'})})$, and the associated unit eigenvector v'.

8:     **if** $\|DMD\|_2 < O(\log(1/\varepsilon))$, **then return** $\mu^{S'}$ (re-inserting all coordinates affected by Step 6).

9:     Let $\delta := 3\sqrt{\varepsilon\|DMD\|_2}$. Find $T > 0$ such that

$$\Pr_{X \in_u S'}(|v^* \cdot (X - \mu^{S'})| > T + \delta) > 20/T^2 + 4\varepsilon^{3/2}/d^2 ,$$

    where $v^* := Dv'$.

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

This completes the description of the algorithm. We now proceed to prove correctness.

6.2.2 Chi-Squared Distance and Basic Reductions

As previously mentioned, our algorithm will use the chi-squared 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-squared distance used in Lemma 2.15 We will instead require the following asymmetric version of the chi-squared distance:

Definition 6.25. *The χ^2-distance of $x, y \in \mathbb{R}^d$ is defined by $\chi^2(x, y) \stackrel{\text{def}}{=} \sum_{i=1}^{d} \frac{(x_i - y_i)^2}{x_i(1 - x_i)}$.*

The following fact follows directly from Lemma 2.15.

Fact 6.26. *Let P, Q be binary product distributions with mean vectors p, q respectively. Then, $d_{\text{TV}}(P, Q) = O(\sqrt{\chi^2(p, q)})$.*

There are two problems with using the chi-squared distance between the mean vectors as a proxy for the total variation distance. The first is that the chi-squared 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 pre-processing in Step 6. The second is that the above asymmetric notion of the χ^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 6 exists, then removing the uncommon values of that coordinate increases $\Delta(S, S')$ by at most ε/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 6, 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 $\widetilde{P}$ so 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 ε in 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 6.27. *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 ith 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 ith-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 $\sum_{j \in T} \widetilde{p}_j(1 - \widetilde{p}_j) \leq \eta - p_i$, we need to show that

$$| \Pr_{X \in_u \widetilde{S}_T} (L(X) > 0) - \Pr_{X \sim \widetilde{P}_T} (L(X) > 0)| \leq \varepsilon^{3/2}/d^2 .$$

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,

$$\sum_{j \in \widetilde{T}} p_j(1 - p_j) = p_i(1 - p_i) + \sum_{j \in T} \widetilde{p}_j(1 - \widetilde{p}_j) \leq \eta - p_i + p_i(1 - p_i) \leq \eta.$$

Since S is good for P, we have that

$$| \Pr_{X \in_u S_{\widetilde{T}}} (L(X) \geq 0) - \Pr_{X \sim P_{\widetilde{T}}} (L(X) \geq 0)| \leq \varepsilon^{3/2}/d^2 .$$

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 ith coordinate is 1. Since S is (ε, η)-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 ith-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 ith-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. $\square$

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 $1 - \mu_i^{S'} \geq (1 - p_i)/3$.

6.2.3 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. We start with the following simple claim:

Claim 6.28. *For any $v \in \mathbb{R}^d$ satisfying $\sum_{i=1}^d v_i^2 \mu_i^{S'}(1 - \mu_i^{S'}) \leq 1$, the following statements hold:*

(i) $\mathrm{Var}_{X \sim P}[v \cdot X] \leq 9$ *and* $|v \cdot (p - \mu^{S'})| \leq \sqrt{\chi^2(\mu^{S'}, p)}$, *and*

(ii) $\mathrm{Pr}_{X \sim P}\left(|v \cdot X - \mu^{S'}| \geq T + \sqrt{\chi^2(\mu^{S'}, p)}\right) \leq 9/T^2$.

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

$$\mathrm{Var}_{X \sim P}\left[\sum_{i=1}^d v_i X_i\right] = \sum_{i=1}^d v_i^2 \mathrm{Var}[X_i] = \sum_{i=1}^d v_i^2 p_i(1 - p_i) \leq 9 \sum_{i=1}^d v_i^2 \mu_i^{S'}(1 - \mu_i^{S'}) \leq 9 \ ,$$

where we used that $p_i \leq 3\mu_i^{S'}$, $(1 - p_i) \leq 3(1 - \mu_i^{S'})$ and the assumption in the claim statement. For the second part of (i), note that

$$|v \cdot (p - \mu^{S'})| = \left|\sum_{i=1}^d v_i \sqrt{\mu_i^{S'}(1 - \mu_i^{S'})} \cdot \frac{p_i - \mu_i^{S'}}{\sqrt{\mu_i^{S'}(1 - \mu_i^{S'})}}\right| \leq \sqrt{\sum_{i=1}^d v_i^2 \mu_i^{S'}(1 - \mu_i^{S'})} \cdot \sqrt{\chi^2(\mu^{S'}, p)} \leq \sqrt{\chi^2(\mu^{S'}, p)} \ ,$$

where the first inequality is Cauchy-Schwarz, and the second follows from the assumption in the claim statement that $\sum_{i=1}^d v_i^2 \mu_i^{S'}(1 - \mu_i^{S'}) \leq 1$. This proves (i).

To prove (ii), we note that Chebyshev's inequality gives

$$\mathrm{Pr}_{X \sim P}(|v \cdot (X - p)| \geq T) \leq \mathrm{Var}_{X \sim P}[v \cdot X]/T^2 \leq 9/T^2 \ ,$$

where the second inequality follows from (i). To complete the proof note the inequality

$$|v \cdot (X - \mu^{S'})| \geq T + \sqrt{\chi^2(\mu^{S'}, p)}$$

implies that

$$|v \cdot (X - p)| \geq |v \cdot (X - \mu^{S'})| - |v \cdot (p - \mu^{S'})| \geq T \ ,$$

where we used the triangle inequality and the second part of (i). $\qquad\square$

Let $\mathrm{Cov}[S]$ denote the sample covariance matrix with respect to S, and $\mathrm{Cov}[P]$ denote the covariance matrix of P. We will need the following lemma:

Lemma 6.29. *We have the following:*

(i) $\left|\sqrt{\chi^2(\mu^{S'}, \mu^S)} - \sqrt{\chi^2(\mu^{S'}, p)}\right| \leq O(\varepsilon/d)$, and

(ii) $\|D\left(\mathrm{Cov}[S] - \mathrm{Cov}[P]\right)D\|_2 \leq O(\sqrt{\varepsilon})$.

Proof. For (i): Since S is good, for any $i \in [d]$, we have

$$|\mu_i^S - p_i| = \left|\Pr_{X \in_u S}(e_i \cdot X \geq 1) - \Pr_{X \sim P}(e_i \cdot X \geq 1)\right| \leq \varepsilon^{3/2}/d^2 .$$

Therefore, by the triangle inequality we get

$$\left|\sqrt{\chi^2(\mu^{S'}, \mu^S)} - \sqrt{\chi^2(\mu^{S'}, p)}\right| \leq \sqrt{\sum_{i=1}^d \frac{(\mu_i^S - p_i)^2}{\mu_i^{S'}(1 - \mu_i^{S'})}} \leq \sqrt{\frac{d \cdot (\varepsilon^3/d^4)}{\varepsilon/(2d)}} \leq O(\varepsilon/d) ,$$

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

$$\left|\mathbb{E}_{X \in_u S}[X_i X_j - p_i p_j]\right| = \left|\Pr_{X \in_u S}[(e_i + e_j) \cdot X \geq 1] - \Pr_{X \sim P}[(e_i + e_j) \cdot X \geq 1]\right| \leq \varepsilon^{3/2}/d^2 .$$

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,

$$\|D\left(\mathrm{Cov}[S] - \mathrm{Cov}[P]\right)D\|_2 \leq O(\sqrt{\varepsilon}) .$$

$\square$

Combining Claim 6.28 and Lemma 6.29 we obtain:

Corollary 6.30. *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

$$
\begin{aligned}
|\mathrm{Var}_{X \in_u S}[v \cdot X] - \mathrm{Var}_{Y \sim P}[v \cdot Y]| &= v^T\left(\mathrm{Cov}[S] - \mathrm{Cov}[P]\right)v \\
&\leq \|D^{-1}v\|_2^2 \cdot \|D\left(\mathrm{Cov}[S] - \mathrm{Cov}[P]\right)D\|_2 \\
&\leq O(\sqrt{\varepsilon}) \\
&\leq 1 ,
\end{aligned}
$$

where the second line uses Lemma 6.29 (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 ε. Thus, using Claim 6.28 (i), we get that

$$\mathrm{Var}_{X \in_u S}[v \cdot X] \leq \mathrm{Var}_{Y \sim P}[v \cdot Y] + 1 \leq 10 .$$

56

By the Cauchy-Schwarz inequality and Lemma 6.29, 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 6.28 (ii) and the assumption that S is good for P. $\qquad \square$

Lemma 6.31. *For all $i, j \in [d]$, we have that $|(M_S)_{i,j} - (M_P)_{i,j}| \leq O(\varepsilon^{3/2}/d^2)$ and therefore*

$$\left\||S'|M - |S|M_P + |E|M_E - |L|M_L\right\|_2 = O(|S'|\varepsilon^{3/2}/d) \ .$$

Proof. The bound $|(M_S)_{i,j} - (M_P)_{i,j}| \leq O(\varepsilon^{3/2}/d^2)$ comes from expanding the LHS in terms of the differences of linear threshold functions on S and P in the same way as the proof of Lemma 6.29. The lemma then follows from the fact that that $|S'|M = |S|M_S + |E|M_E - |L|M_L$ and the observation $|S| \leq |S'|/(1 - 2\varepsilon) \leq 2|S'|$. $\qquad \square$

Combining the above we obtain:

Corollary 6.32. *We have that $\|D(|S'|M - |S|M_P - |E|M_E + |L|M_L)D\|_2 = O(|S'| \cdot \sqrt{\varepsilon})$.*

Proof. Note that $\|D\|_2 = \max_i \left(1/\sqrt{\mu_i^{S'}(1 - \mu_i^{S'})}\right) \leq \sqrt{2d/\varepsilon}$, and therefore

$$\begin{aligned}
\left\|D(|S'|M - |S|M_P - |E|M_E + |L|M_L)D\right\|_2 &\leq \|D\|_2^2 \cdot \left\||S'|M - |S|M_P - |E|M_E + |L|M_L\right\|_2 \\
&\leq (2d/\varepsilon) \cdot O(|S'|\varepsilon^{3/2}/d) \\
&= O(|S'| \cdot \sqrt{\varepsilon}) \ .
\end{aligned}$$

$\qquad \square$

We have the following lemma:

Lemma 6.33. *We have that $\|DM_PD\|_2 \leq 9 + \chi^2(\mu^{S'}, p)$.*

Proof. Note that $M_P = (\mu^{S'} - p)(\mu^{S'} - p)^T + \mathrm{Diag}(p_i(1 - p_i))$. For any v' with $\|v'_2\| \leq 1$, the vector $v = Dv'$ satisfies $\sum_{i=1}^d v_i^2 \mu_i^{S'}(1 - \mu_i^{S'}) \leq 1$. 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 6.28 (i), we get

$$(v.(\mu^{S'} - p))^2 \leq \chi^2(\mu^{S'}, p)$$

and

$$|v^T \mathrm{Diag}(p_i(1 - p_i))v| = |\operatorname*{Var}_{X \sim P}(v \cdot (X - p))| \leq 9 \ .$$

This completes the proof. $\qquad \square$

The following crucial lemma bounds from above the contribution of the subtractive error:

Lemma 6.34. *The spectral norm $\|DM_LD\|_2 = O(|S'|/|L| + \chi^2(\mu^{S'}, p))$.*

Proof. Similarly, we need to bound from above the quantity $|v'^T D M_L D v'|$ for all $v' \in \mathbb{R}^d$ with $\|v'\|_2 \leq 1$. Note that $|v'^T D M_L D v'| = |v^T M_L v| = \mathbb{E}_{X \in_u L}[|v \cdot (X - \mu^{S'})|^2]$, where the vector $v = Dv'$ satisfies $\sum_{i=1}^d v_i^2 \mu_i^{S'}(1 - \mu_i^{S'}) \leq 1$. The latter expectation is bounded from above as follows:

$$
\begin{aligned}
\mathbb{E}_{X \in_u L}[(v \cdot (X - \mu^{S'}))^2] &\leq 2 \mathbb{E}_{X \in_u L}[(v \cdot (X - p))^2] + 2(v \cdot (\mu^{S'} - p))^2 \\
&\leq 2 \mathbb{E}_{X \in_u L}[(v \cdot (X - p))^2] + 2\chi^2(\mu^{S'}, p) \\
&\leq (2|S|/|L|) \cdot \mathbb{E}_{X \in_u S}[(v \cdot (X - p))^2] + 2\chi^2(\mu^{S'}, p) \\
&\leq 20|S|/|L| + 2\chi^2(\mu^{S'}, p) \\
&\leq 21|S'|/|L| + 2\chi^2(\mu^{S'}, p) ,
\end{aligned}
$$

where the first line uses the triangle inequality, the second line uses Claim 6.28 (i), the third line follows from the fact that $L \subseteq S$, the fourth line uses Corollary 6.30 (i), and the last line uses the fact that ε is small enough. $\qquad \square$

The above lemmas and the triangle inequality yield the following:

Corollary 6.35. *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.

6.2.4 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 ε fraction of the points in S could have that any coordinate in N does not have its most common value. Therefore, at most a 2ε 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 the analysis below that $N = \varnothing$.

We begin by bounding various χ^2-distances by the spectral norm of appropriate matrices.

Lemma 6.36. *Let* μ^E, μ^L *be the mean vector of* E *and* L, *respectively. Then,* $\chi^2(\mu^{S'}, \mu^E) \leq \|D M_E D\|_2$ *and* $\chi^2(\mu^{S'}, \mu^L) \leq \|D M_L D\|_2$.

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,

$$
\|D M_E D\|_2 \geq v^T M_E v \geq |v \cdot (\mu^E - \mu^{S'})|^2 = \chi^2(\mu^{S'}, \mu^E) .
$$

$\qquad \square$

We can now prove that the output in Step 8 has the desired guarantee:

Lemma 6.37. *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

$$(|S|/|S'|)\sqrt{\chi^2(\mu^{S'}, \mu^S)} \leq (|E|/|S'|)\sqrt{\chi^2(\mu^{S'}, \mu^E)} + (|L|/|S'|)\sqrt{\chi^2(\mu^{S'}, \mu^L)} . \tag{32}$$

First note that, by Lemma 6.29, $|\sqrt{\chi^2(\mu^{S'}, \mu^S)} - \sqrt{\chi^2(\mu^{S'}, p)}| \leq O(\varepsilon/d)$. Lemma 6.36 and Corollary 6.35 give that

$$(|E|/|S'|)^2 \chi^2(\mu^{S'}, \mu^E) \leq (|E|/|S'|)^2 \|DM_E D\|_2 + O(\varepsilon) \leq (|E|/|S'|)\|DMD\|_2 + O(\varepsilon(1 + \chi^2(\mu^{S'}, p))) .$$

Thus,

$$(|E|/|S'|)\sqrt{\chi^2(\mu^{S'}, \mu^E)} \leq \sqrt{(|E|/|S'|)\|DMD\|_2} + \sqrt{\varepsilon} \cdot O(1 + \sqrt{\chi^2(\mu^{S'}, p)}) .$$

Lemmas 6.34 and 6.36 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,

$$(|L|/|S'|)\sqrt{\chi^2(\mu^{S'}, \mu^L)} \leq O((|L|/|S'|)\sqrt{\chi^2(\mu^{S'}, p)}) + O(\sqrt{\varepsilon}) .$$

Substituting these into (32), yields

$$(|S|/|S'|)\sqrt{\chi^2(\mu^{S'}, p)} \leq \sqrt{(|E|/|S'|)\|DMD\|_2} + O\left(\sqrt{\varepsilon}\left(1 + \sqrt{\chi^2(\mu^{S'}, p)}\right)\right) .$$

For ε sufficiently small, we have that the $\sqrt{\chi^2(\mu^{S'}, p)}$ terms satisfy

$$(|S|/|S'|) - O(\sqrt{\varepsilon}) \geq 1 - 2\varepsilon - O(\sqrt{\varepsilon}) \geq \frac{1}{\sqrt{2}}.$$

Recalling that $|E|/|S'| \leq \Delta(S, S')|S|/|S'| \leq (5/2)\varepsilon$, we now have:

$$\sqrt{\chi^2(\mu^{S'}, p)} \leq (5/2)\sqrt{\varepsilon \|DMD\|_2} + O(\sqrt{\varepsilon}) ,$$

as required. $\qquad \square$

Corollary 6.38. *For some universal constant C, if $\delta \leq C\sqrt{\varepsilon \ln(1/\varepsilon)}$, then $\sqrt{\chi^2(\mu^{S'}, p)} \leq O(\sqrt{\varepsilon \log(1/\varepsilon)})$. Otherwise, we have $\sqrt{\chi^2(\mu^{S'}, p)} \leq \delta$.*

Proof. From the algorithm we have $\delta := 3\sqrt{\varepsilon|\lambda|} = 3\sqrt{\varepsilon \|DMD\|_2}$. By Lemma 6.37, we have that

$$\sqrt{\chi^2(\mu^{S'}, p)} \leq \frac{5}{6}\delta + O(\sqrt{\varepsilon}) .$$

If C is sufficiently large, when $\delta > C\sqrt{\varepsilon \ln(1/\varepsilon)}$, this $O(\sqrt{\varepsilon})$ is at most $C\sqrt{\varepsilon \ln(1/\varepsilon)}/6$. $\qquad \square$

Claim 6.39. *If the algorithm terminates at Step 8, then we have $d_{\mathrm{TV}}(P, P') \leq O(\sqrt{\varepsilon \ln(1/\varepsilon)})$, where P' is the product distribution with mean vector $\mu^{S'}$.*

Proof. By Corollary 6.38, we have that $\sqrt{\chi^2(\mu^{S'}, p)} \leq O(\sqrt{\varepsilon \log(1/\varepsilon)})$. Thus, by Corollary 6.26, the total variation distance between the product distributions with means p and $\mu^{S'}$ is $O(\sqrt{\varepsilon \log(1/\varepsilon)})$. $\qquad \square$

6.2.5 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 6.38, it follows that $\sqrt{\chi^2(\mu^{S'}, p)} \leq \delta$.

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 6.30 to it.

We will show that $\Pr_{X \in_u E}(|v^* \cdot (X - \mu^{S'})| > T + \delta)$ is larger than this quantity for some $T > 0$.

Corollary 6.40. *If we do not terminate at Step 8, then $v^{*T}M_E v^* \geq \delta^2 |S'|/(2\varepsilon|E|)$.*

Proof. By Corollary 6.35, we have that

$$\|D(M - (|E|/|S'|)M_E)D\|_2 = O(1 + \chi^2(\mu^{S'}, p)) \ .$$

Since $\|D^{-1}v\|_2 \leq 1$, we have that

$$v^{*T}(M - (|E|/|S'|)M_E)v^* = O(1 + \chi^2(\mu^{S'}, p)) = O(1 + \delta) \ .$$

Recall that we can assume that $\delta^2/\varepsilon \ln(1/\varepsilon)$ is a sufficiently large constant. So,

$$(|E|/|S'|)v^{*T}M_E v^* \geq v'^T DMDv' - O(1 + \delta) = \delta^2/\varepsilon - O(1 + \delta) \geq \delta^2/2\varepsilon \ .$$

$\square$

Lemma 6.41. *If $\|DMD\|_2 = \Omega(1)$, there exists a $T > 0$ satisfying the property in Step 9 of the algorithm, i.e., $\Pr_{X \in_u S'}[v^* \cdot (X - \mu^{S'})| > T + \delta] \geq 20/T^2 + 4\varepsilon^{3/2}/d^2$.*

Proof. Assume for the sake of contradiction that this is not the case, i.e., that for all $T > 0$ we have that

$$\Pr_{X \in_u S'}(|v^* \cdot (X - \mu^{S'})| \geq T + \delta) < 20/T^2 + 4\varepsilon^{3/2}/d^2 \ .$$

This implies that for all $T > 0$

$$|E| \Pr_{Y \in_u E}(|v^* \cdot (Y - \mu^{S'})| > T + \delta) \ll |S'|(1/T^2 + \varepsilon^{3/2}/d^2).$$

Since

$$|v_i(Y_i - \mu_i^{S'})| \leq 1/\sqrt{p_i(1 - p_i)} \leq 2\sqrt{d/\varepsilon}$$

for all i, we have

$$\Pr_{Y \in_u E}[|v^* \cdot (Y - \mu^{S'})|) > 2d^{3/2}/\sqrt{\varepsilon}] = 0 \ .$$

Therefore, we have that

$$
\begin{aligned}
\delta^2/\varepsilon &\ll (|E|/|S'|)(v^*)^T M_E v^* \\
&= (|E|/|S'|) \mathop{\mathbb{E}}_{Y \in_u E}[|v^* \cdot (Y - \mu^{S'})|^2] \\
&= 2(|E|/|S'|) \int_0^\infty \mathop{\Pr}_{Y \in_u E}(|v^* \cdot (Y - \mu^{S'})| > T) T dT \\
&\ll \int_0^{2d^{3/2}/\sqrt{\varepsilon}} \min\{(|E|/|S'|), 1/(T-\delta)^2 + (\varepsilon/d)^2\} T dT \\
&\leq \int_0^{2d^{3/2}/\sqrt{\varepsilon}} \min\{(5/2)\varepsilon, 1/(T-\delta)^2 + (\varepsilon/d)^2\} T dT \\
&\ll \int_0^{\sqrt{1/\varepsilon}+\delta} \varepsilon T dt + \int_{\sqrt{1/\varepsilon}+\delta}^\infty T/(T-\delta)^2 dT + \int_0^{2d^{3/2}/\sqrt{\varepsilon}} \varepsilon^{3/2}/d^2 T dT \\
&\ll (1 + \delta^2 \varepsilon) + (\log(1/\varepsilon) + \delta\sqrt{\varepsilon}) + \varepsilon/\sqrt{d} \\
&\ll \log(1/\varepsilon) + \delta\sqrt{\varepsilon} + \delta^2 \varepsilon \ .
\end{aligned}
$$

This yields the desired contradiction recalling that $\delta \geq C\sqrt{\varepsilon \log(1/\varepsilon)}$ for an appropriately large $C > 0$. $\square$

The following simple claim completes the proof of Proposition 6.24, and thus proves Theorem 6.20.

Claim 6.42. *If we return an S'' it will satisfy $\Delta(S, S'') \leq \Delta(S, S') - \varepsilon^{3/2}/d^2$.*

Proof. We can write $S'' = (S \setminus L'') \cup E''$ for disjoint multisets L'' and E''. Note that S'' is produced from S' by removing points with $|v \cdot X - \mu^{S'}| \geq T + \delta$. Thus, $E'' \subset E$ and $L'' \supseteq L$.

By our choice of T, the fraction of $x \in S'$ with $|v \cdot X - \mu^{S'}| \geq T + \delta$ is at least $20/T^2 + 4\varepsilon^{3/2}/d^2$. Thus,

$$
|E \setminus E''| + |L'' \setminus L| \geq (20/T^2 + 4\varepsilon^{3/2}/d^2) \cdot |S'| \geq (18/T^2 + 3\varepsilon^{3/2}/d^2)|S| \ .
$$

On the other hand, the fraction of $x \in S$ with the desired property is

$$
\mathop{\Pr}_{X \in_u S}[|v^* \cdot X - \mu^{S'}| \geq T + \delta] \leq 9/T^2 + \varepsilon^{3/2}/d^2 \ ,
$$

by Corollary 6.30 (ii). Thus,

$$
|L'' \setminus L| \leq (9/T^2 + \varepsilon^{3/2}/d^2)|S| \ ,
$$

and therefore

$$
|E \setminus E''| \geq (9/T^2 + 2\varepsilon^{3/2}/d^2)|S|.
$$

In summary, we have that

$$
\Delta(S, S') - \Delta(S, S'') = \frac{|E| + |L| - |E''| - |L''|}{|S|} \geq \varepsilon^{3/2}/d^2 \ .
$$

$\square$

7 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 coordinate-wise 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. But even if only the top absolute eigenvalue is large, we know that both p and q are close to the line $\mu + cv$, where μ 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 λ 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. Moreover, there is a 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.

7.1 The Full Algorithm

This section is devoted to the proof of the following theorem:

Theorem 7.1. *Let Π 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 Π, an ε-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 Π' 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 7.2) 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 7.3) 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 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 $\mathbb{R}^d$ so that both mean vectors of the target product distributions are guaranteed to be close to this line in ℓ_2-distance. The assumption that our product distributions are balanced implies that Π 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 Π 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, so that at

least one is sufficiently close to Π. We then run the tournament described by Lemma 2.9 in order to find a particular candidate that is sufficiently close to the target. To ensure that all 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 ε/d, and similarly round the mixing weight to the nearest multiple of ε. This introduces at most $O(\varepsilon)$ additional error.

Algorithm 11 Filter algorithm for agnostically learning a mixture of two balanced products

---

1: **procedure** LEARNPRODUCTMIXTURE(ε, τ, S')

input: a set of $\mathrm{poly}(d, 1/\varepsilon)\log(1/\tau)$ samples of which an ε-fraction have been corrupted.

output: a mixture of two balanced binary products that is $O(\varepsilon^{1/6})$-close to the target

2:     Run the procedure FILTER-BALANCED-PRODUCT$(2\varepsilon^{1/6}, S'_1)$ for up to $d+1$ iterations on a set S'_1 of corrupted samples of size $\Theta(d^4 \log(1/\tau)/\varepsilon^{1/3})$.

3:     **for** each $1 \le i^* \le d$ **do**

4:         Run the procedure FILTER-PRODUCT-MIXTURE-ANCHOR$(i^*, \varepsilon, S'_{2,i^*})$ for up to $d+1$ iterations on a set S'_{2,i^*} of corrupted samples of size $\Theta(d^4 \log(1/\tau)/\varepsilon^{13/6})$.

5:     Run the procedure FILTER-PRODUCT-MIXTURE-CLOSE$(\varepsilon, S'_3, \delta := \varepsilon^{1/6})$ for up to $d+1$ iterations on a set S'_3 of corrupted samples of size $\Theta(d^4 \log(1/\tau)/\varepsilon^{13/6})$.

6:     Run a tournament among all mixtures output by any of the previous steps. Output the winner.

7.2 Mixtures of Products Whose Means Differ Significantly in One Coordinate

We will use the following notation. Let Π be a mixture of two c-balanced binary product distributions. We will write Π 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 7.2. *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 Π, an ε-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 the 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 7.3. *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 ε-good with respect to P if for every affine function $L : \mathbb{R}^d \to \mathbb{R}$ it holds*

$$\left| \Pr_{X \in_u S}(L(X) > 0) - \Pr_{X \sim P}(L(X) > 0) \right| \le \varepsilon/d.$$

We will also need this to hold after conditioning on the last coordinate.

Definition 7.4. *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 (ε, i)-good with respect to P if S is ε-good with respect to P, and $S^j \stackrel{\mathrm{def}}{=} \{x \in S : x_i = j\}$ is ε-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 7.5. *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 (ε, i)-good with respect to Π if we can write $S = S_P \cup S_Q$, where S_P is (ε, i)-good with respect to P, S_Q is (ε, i)-good with respect to Q, and $\left| \frac{|S_P|}{|S|} - \alpha \right| \leq \varepsilon/d^2$.*

We now show that taking random samples from Π produces such a set with high probability.

Lemma 7.6. *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 Π. Then, with probability at least $1 - \tau$, S is (ε, i)-good with respect to Π for all $i \in [d]$.*

The proof of this lemma is deferred to Section E.

We claim that given a good set with an ε-fraction of its entries corrupted, we can still determine Π from it. In particular, this is achieved by iterating the following proposition.

Proposition 7.7. *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 (ε, d)-good multiset for Π, 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 $\mathcal{S}$ so 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 7.2.

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 the similar notation 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 12 Filter algorithm for a mixture of two binary products whose means differ significantly in some coordinate

---

1: **procedure** FILTER-PRODUCT-MIXTURE-ANCHOR(i^*, ε, S')
2:     Let μ be the sample mean of S' without the i^* coordinate. Let Σ be the sample covariance matrix of S' without the i^* row and column.
3:     Let S'_0 and S'_1 be the subsets of S' with a 0 or 1 in their i^* coordinates, respectively.
4:     Let $\mu^{(j)}$ be the sample mean of S'_j without the i^* coordinate.
5:     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.*/
6:     **if** $\lambda \leq O(1)$ **then**
7:         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 $|i| \leq 1 + \sqrt{d}/\varepsilon^{1/6}$.
8:         **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 $\varepsilon^{1/6}$.
9:     Let $\delta = C(\varepsilon^{1/6}\sqrt{\lambda} + \varepsilon^{2/3} \log(1/\varepsilon))$ for a sufficiently large constant C.
10:    Find a real number $T > 0$ such that

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

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

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 lemmata relating the means of some of these distributions.

Lemma 7.8. *Let P be a binary product distribution with mean vector p. Let S be an ε-good multiset for P in the sense of Definition 7.3. 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, $\|p - \mu^{\tilde{S}}\|_2 \leq O(\varepsilon\sqrt{\log(1/\varepsilon)})$.*

Proof. Since S is ε-good, $\|\mu^S - p\|_2 \leq \varepsilon/\sqrt{d}$. Let $L = S \setminus \tilde{S}$. We can apply appropriate lemmata from Section 6.1. Note that Lemma 6.13 and Claim 6.15, 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 6.13 with $\mu^{S'} := p$, we have $\|\mathbb{E}_{X \in_u L}[(X - p)(X - p)^T]\|_2 \leq O(\log(|S|/|L|) + \varepsilon|S|/|L|)$. By Claim 6.15 again with $\mu^{S'} := p$, it follows that $(|L|/|S|)\|\mu^L - p\|_2 \leq O(\varepsilon\sqrt{\log(1/\varepsilon)})$. 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

$$\|\mu^S - \mu^{\tilde{S}}\|_2 \leq (|L|/|\tilde{S}|)\|\mu^L - \mu^S\|_2 \leq O(\varepsilon^2/\sqrt{d}) + O(1 + \varepsilon)(|L|/|S|)\|\mu^L - p\|_2 \leq O(\varepsilon\sqrt{\log(1/\varepsilon)}).$$

By the triangle inequality, $\|p - \mu^{\tilde{S}}\|_2 \leq \varepsilon/\sqrt{d} + O(\varepsilon\sqrt{\log(1/\varepsilon)}) = O(\varepsilon\sqrt{\log(1/\varepsilon)})$. □

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 ε-good, we can bound the error introduced by $\mu^{S_P} - p, \mu^{S_Q} - q_{-d}$ and Lemma 7.8 allow 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 7.9. *For some scalars $a = O(\varepsilon)$, $b^0 = O(|E^0|/|S'|)$, $b^1 = O(|E^1|/|S'|)$, we have*

$$\|(1 - \mu_d)\mu_d u - (\alpha(1 - \alpha)(p_d - q_d) + a)(p_{-d} - q_{-d}) - b^0(\mu^{E^0} - \mu) - b^1(\mu^{E^1} - \mu)\|_2 \leq O(\varepsilon\log(1/\varepsilon)),$$

where E^j is the subset of E with last entry j, μ^{E^j} is the mean of E^j with d^{th} coordinate removed.

Proof. Let $S'^j_P, S'^j_Q, E^j, S'^j$ 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'^j_P, S'^j_Q, 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

$$|S'^j|\mu^{(j)} = |S_{\tilde{P}^j}|\mu^{S'_P{}^j} + |S'^j_Q|\mu^{S'_Q{}^j} + |E^j|\mu^{E^j}.$$

Therefore, using this and Lemma 7.8, we have that

$$|S'^j|\mu^{(j)} = |S'^j_P|p_{-d} + |S'^j_Q|q_{-d} + |E^j|\mu^{E^j} + O(\varepsilon\log(1/\varepsilon)|S^j|),$$

65

where $O(\varepsilon)$ denotes a vector of ℓ_2-norm $O(\varepsilon)$.

Thus, we have

$$
\begin{aligned}
|S'^0||S'^1|(\mu^{(1)} - \mu^{(0)}) &= (|S'^0||S_P'^1| - |S'^1||S_P'^0|)p_{-d} \\
&\quad + (|S'^0||S_Q'^1| - |S'^1||S_Q'^0|)q_{-d} \\
&\quad + |E^1||S'^0|\mu^{E^1} - |E^0||S'^1|\mu^{E^0} + O(\varepsilon \log(1/\varepsilon)(|S^1||S'^0| + |S^0||S'^1|)) \,. \quad (33)
\end{aligned}
$$

Since $|S'^j| = |S_P'^j| + |S_Q'^j| + |E^j|$, we have:

$$
\begin{aligned}
0 = |S'^0||S'^1| - |S'^1||S'^0| &= (|S'^0||S_P'^1| - |S'^1||S_P'^0|) \\
&\quad + (|S'^0||S_Q'^1| - |S'^1||S_Q'^0|) + |E^1||S'^0| - |E^0||S'^1| \,.
\end{aligned}
$$

Thus, the sum of the coefficients of the p_{-d} and q_{-d} terms in Equation (33) is $|E^0||S'^1| - |E^1||S'^0|$, which is bounded in absolute value by $|E||S'| \le O(\varepsilon|S|^2)$. Meanwhile, the p_{-d} coefficient of Equation (33) has:

$$
\begin{aligned}
&|S'^0||S_P'^1| - |S'^1||S_P'^0| \\
&= |S'^0||S_P^1| - |S'^1||S_P^0| + O(\varepsilon|S'|^2) = |S'^0||S|\alpha p_d - |S'^1||S|\alpha(1 - p_d) + O(\varepsilon|S'|^2) \\
&= |S^0||S|\alpha p_d - |S^1||S|\alpha(1 - p_d) + O(\varepsilon|S'|^2) \\
&= ((\alpha(1 - p_d) + (1 - \alpha)(1 - q_d))\alpha p_d - (\alpha p_d + (1 - \alpha)q_d)\alpha(1 - p_d) + O(\varepsilon))|S'|^2 \\
&= (\alpha(1 - \alpha)(1 - q_d)p_d - \alpha(1 - \alpha)q_d(1 - p_d) + O(\varepsilon))|S'|^2 = (\alpha(1 - \alpha)(p_d - q_d) + O(\varepsilon))|S'|^2 \,.
\end{aligned}
$$

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 Equation (33) as:

$$
\begin{aligned}
|S'^0||S'^1|(\mu^{(1)} - \mu^{(0)}) &= (\alpha(1 - \alpha)(p_d - q_d) + O(\varepsilon))|S'|^2(p_{-d} - q_{-d}) \\
&\quad + (|E^1||S'^0| - |E^0||S'^1|)(\alpha p_{-d} + (1 - \alpha)q_{-d}) \\
&\quad + |E^1||S'^0|\mu^{E^1} - |E^0||S'^1|\mu^{E^0} + O(\varepsilon \log(1/\varepsilon)|S'|^2) \,.
\end{aligned}
$$

We write $\mu^\Pi = \alpha p_{-d} + (1 - \alpha)q_{-d}$ and so, dividing by $|S'|^2$ and recalling that $|E|/|S'| \le O(\varepsilon)$, we get

$$
\begin{aligned}
\mu_d(1 - \mu_d)(\mu^{(1)} - \mu^{(0)}) &= (\alpha(1 - \alpha)(p_d - q_d) + O(\varepsilon))(p_{-d} - q_{-d}) + O(|E^1|/|S'|)(\mu^{E^1} - \mu^\Pi) \\
&\quad + O(|E^0|/|S'|)(\mu^{E^0} - \mu^\Pi) + O(\varepsilon \log(1/\varepsilon)) \,. \quad (34)
\end{aligned}
$$

If $\mu^\Pi = \mu$, then we would be done. So, we must bound the error introduced by making this substitution. We can express μ 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 Equation (34), gives the lemma. $\qquad\square$

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 7.10. *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 7.9 that

$$
|v \cdot (p_{-d} - q_{-d})| \leq
$$

$$
\leq \frac{1}{\alpha(1-\alpha)(p_d - q_d)} \cdot \Big(O(|E^0|/|S'|)v \cdot (\mu^{E^0} - \mu) + O(|E^1|/|S'|)v \cdot (\mu^{E^1} - \mu) + O(\varepsilon \log(1/\varepsilon))\|v\|_2 \Big)
$$

$$
= O\left(\frac{1}{\alpha(1-\alpha)(p_d - q_d)}\right) \left(\sqrt{\varepsilon(v^T \Sigma v)} + \varepsilon\log(1/\varepsilon)\right) .
$$

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. $\qquad\square$

We can now show that if we return R, some distribution returned is close to Π. First, we show that there are points on L close to p_{-d} and q_{-d}.

Lemma 7.11. *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 7.10, 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, $\|\tilde{p} - p_{-d}\|_2 \leq \delta$. $\qquad\square$

It is clear that even discretizing c and d, we can still find such a pair that satisfies this condition.

Lemma 7.12. *There are $p', q' \in L$ such that $\|p_{-d} - p'\|_2, \|q_{-d} - q'\|_2 \leq \delta + O(\varepsilon^{1/6})$.*

Proof. By Lemma 7.11, 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 $\|\tilde{p} - p_{-d}\|_2, \|\tilde{q} - q_{-d}\|_2 \leq \delta$.

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 $\|p_{-d} - p'\|_2 \leq \|\tilde{p}_{-d} - p\|_2 + \|p' - \tilde{p}\|_2 \leq \delta + \varepsilon^{1/6}$.

Note that we have $\|p_{-d} - \tilde{p}\|_2 \leq \|p_{-d} - \mu\|_2 \leq \sqrt{d}\|p_{-d} - \mu\|_\infty \leq \sqrt{d}$, which implies that $a \leq \sqrt{d}$. Thus, $|i| \leq 1 + \sqrt{d}/\varepsilon^{1/6}$. 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 $\|p_{-d} - p'\|_2 \leq \delta + O(\varepsilon^{1/6})$.

Similarly, we show that there is a $q' \in L$ such that $\|q - q'\|_2 \leq \delta + O(\varepsilon^{1/6})$. $\qquad\square$

Corollary 7.13. *If the algorithm terminates at Step 8, then there is a $\Pi' \in R$ with with $d_{\mathrm{TV}}(\Pi', \Pi) = O(\varepsilon^{1/6}/\sqrt{c})$.*

Proof. By Lemma 7.12, there exists $\tilde{p}, \tilde{q} \in L$ such that $\|p_{-d} - \tilde{p}\|_2, \|q_{-d} - \tilde{q}\|_2 \leq \delta + O(\varepsilon^{1/6})$. 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 $|\alpha' - \alpha| \leq \varepsilon^{1/6}$, $\|p'_{-d} - p_{-d}\|_2, \|q'_{-d} - q_{-d}\|_2 \leq O(\varepsilon^{1/6})$ 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 $d_{\mathrm{TV}}(P, P') \leq O(\|p - p'\|_2/\sqrt{c}) \leq O(\varepsilon^{1/6}/\sqrt{c})$ and

$$d_{\mathrm{TV}}(Q, Q') \leq O(\|q - q'\|_2/\sqrt{c}) \leq O(\varepsilon^{1/6}/\sqrt{c}).$$

Thus, $d_{\mathrm{TV}}(\Pi', \Pi) \leq \delta + O(\varepsilon^{1/6}/\sqrt{c})$. Since we terminated in Step 8, $\lambda \leq O(1)$, and so $\delta = C(\varepsilon^{1/6}\sqrt{\lambda} + \varepsilon^{2/3}\log(1/\varepsilon)) = O(\varepsilon^{1/6})$. $\qquad\square$

Now, we are ready to analyze the second part of our algorithm. The basic idea will be to show that if λ is large, then a large fraction of the variance in the v-direction is due to points in E.

Lemma 7.14. *If $\lambda \geq \Omega(1)$, 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

$$|S|\operatorname*{Var}_{X \in_u S'}[v^* \cdot X] = |S'_P| \left(\operatorname*{Var}_{X \in_u S'_P}[v^* \cdot X] + |v^* \cdot (\mu^{S'_P} - \mu)|^2 \right)$$

$$+ |S'_Q| \left(\operatorname*{Var}_{X \in_u S'_Q}[v^* \cdot X] + |v^* \cdot (\mu^{S'_Q} - \mu)|^2 \right)$$

$$+ |E| \operatorname*{\mathbb{E}}_{X \in_u E}[|v^* \cdot (X - \mu)|^2]] .$$

Since S_P and S_Q are ε-good, we have that

$$\operatorname*{Var}_{X\in_u S'_P}[v^* \cdot X] = \operatorname*{\mathbb{E}}_{X\in_u S'_P}[(v \cdot X - v^* \cdot \mu^{S'_P})^2]$$

$$\leq (|S_P|/|S'_P|) \operatorname*{\mathbb{E}}_{X\in_u S_P}[(v^* \cdot X - v^* \cdot \mu^{S'_P})^2]$$

$$= (|S_P|/|S'_P|) \left(\operatorname*{Var}_{X\in_u S_P}[v^* \cdot X] + (v^* \cdot (\mu^{S_P} - \mu^{S'_P}))^2 \right)$$

$$\leq (|S_P|/|S'_P|) \left(\operatorname*{Var}_{X\sim P}[v^* \cdot X] + (v^* \cdot (p_{-d} - \mu^{S'_P}) + O(\varepsilon\sqrt{\log(1/\varepsilon)}))^2 \right)$$

$$\leq (1 + O(\varepsilon/\alpha)) \cdot (O(1) + O(\varepsilon\sqrt{\log(1/\varepsilon)})^2) \leq O(1) ,$$

and similarly,

$$\operatorname*{Var}_{X\in_u S'_Q}[v^* \cdot X] = O(1) .$$

Thus, we have:

$$|S'| \operatorname*{Var}_{X\in_u S'}[v^* \cdot X] \leq |E| \operatorname*{\mathbb{E}}_{X\in_u E}[|v^* \cdot (X - \mu)|^2] + O(1 + |v^* \cdot (p_{-d} - \mu)|^2 + |v^* \cdot (q_{-d} - \mu)|^2)|S'| .$$

By Lemma 7.9, we have

$$|v \cdot (p_{-d} - \mu)|, |v^* \cdot (q_{-d} - \mu)| \leq O(1/(\alpha(1 - \alpha)(p_d - q_d)))$$

$$\cdot (O(|E_0|/|S'|)|v^* \cdot (\mu^{E^0} - \mu)| + O(|E^1|/|S'|)|v^* \cdot (\mu^{E^1} - \mu)| + O(\varepsilon \log(1/\varepsilon)))$$

$$\leq \sqrt{(|E|/|S'|) \operatorname*{\mathbb{E}}_{Y\in_u E}[|v^* \cdot (Y - \mu)|^2]} + O(\varepsilon \log(1/\varepsilon)) .$$

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 λ is larger than a sufficiently large constant, this completes the proof. $\qquad\square$

We next show that the threshold $T > 0$ required by our algorithm exists.

Lemma 7.15. *If $\lambda \geq \Omega(1)$, 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

$$\operatorname*{Pr}_{X\in_u S'}(|v^* \cdot (X - \mu)| \geq T + \delta) \leq 8\exp(-T^2/2) + 8\varepsilon/d .$$

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

$$\mathop{\mathbb{E}}_{Y \in_u E}[|v^* \cdot (Y - \mu)|^2] \ll \delta^2 + \mathop{\mathbb{E}}_{Y \in_u E}[\min(0, |v^* \cdot (Y - \mu)| - \delta)^2]$$

$$\ll \delta^2 + \int_0^{\sqrt{d}} \mathop{\Pr}_{Y \in_u E}(|v^* \cdot (Y - \mu)| > T + \delta) T \, dT$$

$$\ll \delta^2 + \int_0^{\sqrt{d}} (\varepsilon/d) T \, dT + \int_0^{2\sqrt{\log(|S'|/|E|)}} T \, dT$$

$$+ \int_{2\sqrt{\log(|S'|/|E|)}}^{\infty} (|S'|/|E|) \exp(-T^2/2) T \, dT$$

$$\ll \delta^2 + \varepsilon + \log(|S'|/|E|) .$$

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. $\qquad \square$

Finally, we show that S'' is closer to S than S' was.

Claim 7.16. *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

$$|S' \setminus S''| = \mathop{\Pr}_{X \in_u S'}(|v \cdot (X - \mu)| \geq T + \delta)|S'|$$

$$\geq (8\exp(-T^2/2) + 8\varepsilon/d)|S'| \geq (4\exp(-T^2/2) + 4\varepsilon/d)|S| .$$

By Hoeffding's inequality, we have that

$$\mathop{\Pr}_{X \sim P}(|v^* \cdot (X - p_{-d})| \geq T) \leq 2\exp(-T^2/2) .$$

By Lemma 7.10, we have that $|v^* \cdot (\mu - p_{-d})| \leq \delta$ and so

$$\mathop{\Pr}_{X \sim P}(|v^* \cdot (X - \mu)| \geq T + \delta) \leq 2\exp(-T^2/2) .$$

Since S is (ε, d)-good, we have

$$\mathop{\Pr}_{X \in_u S_P}(|v^* \cdot (X - \mu)| \geq T + \delta) \leq 2\exp(-T^2/2) + \varepsilon/d .$$

70

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:

$$
\begin{aligned}
|E \setminus E'| &= |S' \setminus S''| - |S_P' \setminus S_P''| - |S_Q' \setminus S_Q''| \\
&\geq (4\exp(-T^2/2) + 4\varepsilon/d)|S| - (2\exp(-T^2/2) + \varepsilon/d)|S| \\
&\geq (2\exp(-T^2/2) + 3\varepsilon/d)|S| \\
&\geq |S_P' \setminus S_P''| + |S_Q' \setminus S_Q''| + 2\varepsilon|S|/d \ ,
\end{aligned}
$$

which completes the proof. $\qquad \square$

7.3 Mixtures of Products Whose Means Are Close in Every Coordinate

In this section, we prove the following theorem:

Theorem 7.17. *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 Π. Let S' be obtained by adversarially changing an ε-fraction of the points in S. There exists an algorithm that runs in polynomial time and, with probability at least $1 - \tau$, returns 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 7.18. *Let S be a multiset in $\{0, 1\}^d$. We say that S is ε-good for the mixture Π if there exists a partition $S = S_P \cup S_Q$ so 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 7.19. *If Π 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 Π is good for Π.*

The proof of this lemma is in Section E.

Our theorem will follow from the following proposition:

Proposition 7.20. *Let Π be as above and S be a good multiset for Π. Let S' be any multiset with $\Delta(S, S') \leq 2\varepsilon$. There exists a polynomial time algorithm that, given S', $\varepsilon > 0$ and δ, returns either a multiset S'' with $\Delta(S, S'') \leq \Delta(S, S') - 2\varepsilon/d$ or a set of parameters of binary product distributions of size $O(d/(\varepsilon\delta^2))$ which contains the parameters of a Π' with $d_{\mathrm{TV}}(\Pi, \Pi') \leq O(\delta/\sqrt{c})$.*

71

Algorithm 13 Filter algorithm for mixture of two binary products whose means are close in every coordinate

1: **procedure** FILTER-PRODUCT-MIXTURE-CLOSE(ε, S', δ)
2:     Compute μ, the sample mean of S', and $\text{Cov}_0(S')$. Let C be a sufficiently large constant.
3:     **if** $\text{Cov}_0(S')$ has at most one eigenvector with absolute eigenvalue more than $C\delta^2$ **then**
4:         Let v^* be the unit eigenvector of $\text{Cov}_0(S')$ with largest absolute eigenvalue.
5:         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 $|i| \le 1 + \sqrt{d}/\delta$.
6:         **return** the set of distributions of the form $\Pi' = \alpha' P' + (1 - \alpha')Q'$ with the means of P' and Q' in L and α' is a multiple of ε in $[10\varepsilon, 1/2]$.
7:     Let v^* and u^* be orthogonal eigenvectors with eigenvalues more than $C\delta^2$.
8:     Find a number $t \ge 1 + 2\sqrt{\ln(1/\varepsilon)}$ and θ a multiple of δ^2/d so that $r = (\cos\theta)u^* + (\sin\theta)v^*$ satisfies
$$\Pr_{X \in_u S'}\left(\Pr_{Y \in_u S'}(|r \cdot (X - Y)| < t) < 2\varepsilon\right) > 12\exp(-t^2/4) + 3\varepsilon/d .$$
9:     **return** the set $S'' = \{x \in S' \mid \Pr_{Y \in_u S'}(|r \cdot (x - Y)| < t) \ge 2\varepsilon\}$.

Before we present the algorithm, we give one final piece of notation. For S a set of points, we let $\text{Cov}(S)$ be the sample covariance matrix of S and $\text{Cov}_0(S)$ be the sample covariance matrix with zeroed out diagonal. Our algorithm is presented in detailed pseudocode below.

To analyze this algorithm, we begin with a few preliminaries. Firstly, 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 7.21. *We have that* $\alpha\|p - \mu^{S'_P}\|_2, (1 - \alpha)\|q - \mu^{S'_Q}\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$.

Proof. This follows from Lemma 7.8. $\qquad\square$

A key part in our analysis involves showing that the matrix $\text{Cov}_0(S')$ is close to being PSD:

Lemma 7.22. *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,* $\|\text{Cov}_0(S') - \text{Cov}(T)\|_2 = O(\delta^2)$.

Proof. Noting that the mean of T is μ and $|T| = |S'|$, we have:

$$
\begin{aligned}
|S'|\text{Cov}(S') &= |S'_P| \mathop{\mathbb{E}}_{X \in_u S'_P}[(X - \mu)(X - \mu)^T] + |S'_Q| \mathop{\mathbb{E}}_{X \in_u S'_Q}[(X - \mu)(X - \mu)^T] \\
&\quad + |E| \mathop{\mathbb{E}}_{X \in_u E}[(X - \mu)(X - \mu)^T] \\
&= |S'_P|\left(\text{Cov}(S'_P) + (\mu^{S'_P} - \mu)(\mu^{S'_P} - \mu)^T\right) + |S'_Q|\left(\text{Cov}(S'_P) + (\mu^{S'_Q} - \mu)(\mu^{S'_Q} - \mu)^T\right) \\
&\quad + |E| \mathop{\mathbb{E}}_{X \in_u E}[(X - \mu)(X - \mu)^T] \\
&= |S'_P|\text{Cov}(S'_P) + |S'_Q|\text{Cov}(S'_Q) + |S'|\text{Cov}(T) .
\end{aligned}
\tag{35}
$$

Since P and Q are product distributions, $\text{Cov}(S'_P)$ and $\text{Cov}(S'_Q)$ can have large diagonal elements but small off-diagonal ones. On the other hand, we bound the elements on the diagonal of $\text{Cov}(T)$, but $\|\text{Cov}(T)\|_2$ may still be large due to off-diagonal elements.

By the triangle inequality, and Equation (35) with zeroed diagonal, we have:

$$\|\mathrm{Cov}_0(S') - \mathrm{Cov}(T)\|_2 \leq \|\mathrm{Cov}_0(S') - \mathrm{Cov}_0(T)\|_2 + \|\mathrm{Cov}_0(T) - \mathrm{Cov}(T)\|_2$$

$$\leq \left(\frac{|S'_P|}{|S'|}\right)\|\mathrm{Cov}_0(S'_P)\|_2 + \left(\frac{|S'_Q|}{|S'|}\right)\|\mathrm{Cov}_0(S'_Q)\|_2 + \|\mathrm{Cov}_0(T) - \mathrm{Cov}(T)\|_2 \ .$$

$$(36)$$

We will bound each of these terms separately.

Note that $\mathrm{Cov}_0(T) - \mathrm{Cov}(T)$ is a diagonal matrix and its non-zero entries are

$$(\mathrm{Cov}_0(T) - \mathrm{Cov}(T))_{i,i} = \mathrm{Var}_{X \in_u T}[X_i].$$

Since the mean of T is μ, for all i, we have that $\mathrm{Var}_{X \in_u T}[X_i] \leq \mathbb{E}_{X \in_u T}[\|X - \mu\|_\infty^2]$. We seek to bound the RHS from above.

Note that μ 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 $|S'| - |E| = (1 + O(\varepsilon))|S|$, $|S'_Q| = (1 - \alpha)|S| - O(\varepsilon)$, $|E| \leq O(\varepsilon)|S|$, we have

$$\|\mu - \mu^{S'_P}\|_\infty \leq (1 - \alpha + O(\varepsilon))\|\mu^{S'_Q} - \mu^{S'_P}\|_\infty + O(\varepsilon) \ .$$

Similarly,

$$\|\mu - \mu^{S'_Q}\|_\infty \leq (\alpha + O(\varepsilon))\|\mu^{S'_Q} - \mu^{S'_P}\|_\infty + O(\varepsilon) \ .$$

Since S is ε-good for Π, it follows that $\|\mu^{S_P} - p\|_\infty \leq \varepsilon/d$ and $\|\mu^{S_Q} - q\|_\infty \leq \varepsilon/d$. Also,

$$\||S_P|\mu^{S_P} - |S'_P|\mu^{S'_P}\|_\infty \leq |S_P| - |S'_P| \leq O(\varepsilon)|S| \ .$$

Thus,

$$\|\mu^{S_P} - \mu^{S'_P}\|_\infty \leq \|\mu^{S_P} - (|S'_P|/|S_P|)\mu^{S'_P}\|_\infty + (|S_P| - |S'_P|)/|S_P| \leq O(\varepsilon)|S|/|S_P| \leq O(\alpha\varepsilon) \ .$$

Similarly, we show that

$$\|\mu^{S_Q} - \mu^{S'_Q}\|_\infty \leq O((1 - \alpha)\varepsilon).$$

Finally, $\|p - q\|_\infty \leq \delta$. Thus, by the triangle inequality, we get

$$\|\mu^{S'_Q} - \mu^{S'_P}\|_\infty \leq O(\alpha\varepsilon) + \varepsilon/d + \delta + \varepsilon/d + O((1 - \alpha)\varepsilon) \leq \delta + O(\varepsilon) \ .$$

We have the following sequence of inequalities:

$$|S'|\mathrm{Var}_{X \in_u T}[X_i] \leq |S'|\mathbb{E}_{X \in_u T}[\|X - \mu\|_\infty^2]$$

$$= |S'_P|\|\mu - \mu^{S'_P}\|_\infty^2 + |S'_Q|\|\mu - \mu^{S'_Q}\|_\infty^2$$

$$+ |E|\mathbb{E}_{X \in_u T}[\|X - \mu\|_\infty^2]$$

$$\leq (|S'_P| + |S'_Q|)(\|\mu^{S'_Q} - \mu^{S'_P}\|_\infty + O(\varepsilon))^2 + |E|$$

$$\leq (\delta^2 + O(\varepsilon))|S'| \ .$$

Thus,

$$\|\mathrm{Cov}_0(T) - \mathrm{Cov}(T)\|_2 = \max_i (\mathrm{Cov}_0(T) - \mathrm{Cov}(T))_{i,i} = \max_i \mathrm{Var}(T_i) \leq O(\delta^2 + \varepsilon).$$

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 (36). To analyze the first of these terms, note that $\mathrm{Cov}_0(P) = \mathbf{0}$. We have that

$$\|\mathrm{Cov}_0(S'_P)\|_2 = \|\mathrm{Cov}_0(S'_P) - \mathrm{Cov}(P) + \mathrm{Diag}(\underset{X\sim P}{\mathrm{Var}}(X_i))\|_2$$

$$\leq \|\mathrm{Cov}(S'_P) - \mathrm{Cov}(P)\|_2 + \max_i(|\underset{X\in_u S'_P}{\mathrm{Var}}(X_i) - \underset{X\sim P}{\mathrm{Var}}(X_i)|) \ .$$

Noting that

$$|\underset{X\in_u S'_P}{\mathrm{Var}}(X_i) - \underset{X\sim P}{\mathrm{Var}}(X_i)| = e_i^T(\mathrm{Cov}(S'_P) - \mathrm{Cov}(P))e_i,$$

we have that

$$\max_i(|\underset{X\in_u S'_P}{\mathrm{Var}}(X_i) - \underset{X\sim P}{\mathrm{Var}}(X_i)|) \leq \|\mathrm{Cov}(S'_P) - \mathrm{Cov}(P)\|_2 \ ,$$

and so

$$\|\mathrm{Cov}_0(S'_P)\|_2 \leq 2\|\mathrm{Cov}(S'_P) - \mathrm{Cov}(P)\|_2.$$

By the triangle inequality,

$$\|\mathrm{Cov}(S'_P) - \mathrm{Cov}(P)\|_2 \leq \|\mathrm{Cov}(S'_P) - \mathrm{Cov}(S_P)\|_2 + \|\mathrm{Cov}(S_P) - \mathrm{Cov}(P)\|_2 \ .$$

Note that since S is good, the (i,j)-th entry of $\mathrm{Cov}(S_P) - \mathrm{Cov}(P)$ has absolute value at most ε/d. Thus,

$$\|\mathrm{Cov}(S_P) - \mathrm{Cov}(P)\|_2 \leq \|\mathrm{Cov}(S_P) - \mathrm{Cov}(P)\|_F \leq \varepsilon \ ,$$

which gives

$$\|\mathrm{Cov}_0(S'_P)\|_2 \leq 2\|\mathrm{Cov}(S'_P) - \mathrm{Cov}(S_P)\|_2 + O(\varepsilon).$$

We have

$$\|\mathrm{Cov}(S'_P) - \mathrm{Cov}(S_P)\|_2 = \sup_{|v|_2=1}\left(|\underset{X\in_u S'_P}{\mathrm{Var}}(v\cdot X) - \underset{X\in_u S_P}{\mathrm{Var}}(v\cdot X)|\right) \ .$$

Since $S'_P \subseteq S_P$,

$$|S'_P|\underset{X\in_u S'_P}{\mathrm{Var}}(v\cdot X) \leq |S_P|\underset{X\in_u S_P}{\mathbb{E}}[(v\cdot X - \mu^{S'_P}]$$

$$\leq |S_P|(\underset{X\in_u S_P}{\mathrm{Var}}(v\cdot X) + \|\mu^{S'_P} - \mu^{S_P}\|_2^2)$$

$$\leq (1 + O(\varepsilon/\alpha))|S'_P| \cdot (\underset{X\in_u S_P}{\mathrm{Var}}(v\cdot X) + O(\varepsilon^2\log(1/\varepsilon)/\alpha^2)) \ .$$

Thus,

$$|\underset{X\in_u S'_P}{\mathrm{Var}}(v\cdot X) - \underset{X\in_u S_P}{\mathrm{Var}}(v\cdot X)| \leq O(\varepsilon/\alpha)\underset{X\in_u S_P}{\mathrm{Var}}(v\cdot X) + O(\varepsilon^2\log(1/\varepsilon)/\alpha^2))$$

$$\leq O(\varepsilon/\alpha)\underset{X\sim P}{\mathrm{Var}}(v\cdot X) + O(\varepsilon^2\log(1/\varepsilon)/\alpha^2))$$

$$\text{(since } \|\mathrm{Cov}(S_P) - \mathrm{Cov}(P)\|_2 \leq \varepsilon\text{)}$$

$$\leq O(\varepsilon/\alpha) + O(\varepsilon^2\log(1/\varepsilon)/\alpha^2)) \qquad \text{(since } \|\mathrm{Cov}(P)\|_2 \leq 1\text{)}$$

$$\leq O(\varepsilon\log(1/\varepsilon)/\alpha) \ . \qquad\qquad \text{(since } \alpha \geq \varepsilon\text{)}$$

Thus, we have that

$$\|\mathrm{Cov}_0(S'_P)\|_2 \leq 2 \cdot O(\varepsilon\log(1/\varepsilon)/\alpha) + O(\varepsilon) \leq O(\varepsilon\log(1/\varepsilon)/\alpha).$$

74

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 δ in Theorem 7.17, $\delta^2 = \Omega(\varepsilon \log(1/\varepsilon))$, and the proof is complete. $\qquad \square$

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 7.23. *For $u \in \mathbb{R}^d$, we have*

$$\alpha(u \cdot (\mu^{S'_P} - \mu))^2 + (1 - \alpha)(u \cdot (\mu^{S'_P} - \mu))^2 \le 2u^T \mathrm{Cov}_0(S')u + O(\delta^2)\|u\|_2^2.$$

Proof. Using Lemma 7.22, 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

$$\begin{aligned}
\mathrm{Var}_{X \in_u T}(u \cdot X) &\ge \left(\frac{|S'_P|}{|S'|}\right)(u \cdot (\mu^{S'_P} - \mu))^2 + \left(\frac{|S'_Q|}{|S'|}\right)(u \cdot (\mu^{S'_Q} - \mu))^2 + \frac{|E|}{|S'|}\mathrm{Var}_{X \in_u E}(u \cdot X) \\
&\ge (\alpha - 2\varepsilon)(u \cdot (\mu^{S'_P} - \mu))^2 + (1 - \alpha - 2\varepsilon)(u \cdot (\mu^{S'_Q} - \mu))^2 \\
&\ge \alpha/2 \cdot (u \cdot (\mu^{S'_P} - \mu))^2 + (1 - \alpha)/2 \cdot (u \cdot (\mu^{S'_Q} - \mu))^2 \ . \qquad \text{(since } \alpha, 1 - \alpha \ge 4\varepsilon\text{)}
\end{aligned}$$

$\qquad \square$

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 7.24. *There are $\tilde{p}, \tilde{q} \in L$ such that $\|p - \tilde{p}\|_2 \le O(\delta/\sqrt{\alpha})$ and $\|q - \tilde{q}\|_2 \le O(\delta/\sqrt{1 - \alpha})$.*

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

$$(p' - p)^T \mathrm{Cov}_0(S')(p' - p) \le C(\delta^2 + \varepsilon\sqrt{\log(1/\varepsilon)})\|p' - p\|_2^2 \ .$$

We thus obtain:

$$\begin{aligned}
\|p' - p\|_2^4 &= (p' - p) \cdot (p - \mu)^2 \\
&\le 2(p' - p) \cdot (p - \mu^{S_P})^2 + 2(p' - p) \cdot (p - \mu^{S_P})^2 \\
&\le O(\varepsilon^2 \log(1/\varepsilon)/\alpha^2)\|p' - p\|_2^2 + (4/\alpha) \cdot (p' - p)^T \mathrm{Cov}_0(S')(p' - p)^T + O(\delta^2/\alpha)\|p' - p\|_2^2 \\
&\le O((\delta^2 + \varepsilon \log(1/\varepsilon))/\alpha)\|p' - p\|_2^2 \qquad \text{(since } \alpha \ge \varepsilon\text{)} \\
&\le O(\delta^2/\alpha)\|p' - p\|_2^2 \ ,
\end{aligned}$$

where the second line uses Lemmas 7.21 and 7.23. We thus have that $\|p' - p\|_2 \le O(\delta/\sqrt{\alpha})$. Letting $i\delta$ be the nearest integer multiple to a, we have that $\tilde{p} := \mu + i\delta v^*$ has

$$\|p - \tilde{p}\|_2 \le \|p' - p\|_2 + \|p' - \tilde{p}\|_2 \le O(\delta/\sqrt{\alpha}).$$

Note that we have $\|p - p'\|_2 \le \|p - \mu\|_2 \le \sqrt{d}\|p - \mu\|_\infty \le \sqrt{d}$. So, $a \le \sqrt{d}/\delta$. Thus, $|i| \le 1 + \sqrt{d}/\delta$. 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 $\|q - \tilde{q}\|_2 \le O(\delta/\sqrt{1-\alpha})$, which completes the proof. $\qquad\square$

Corollary 7.25. *If the algorithm outputs a set of distributions in Step 6, then one of those distributions has* $d_{\mathrm{TV}}(\Pi', \Pi) \le O(\delta/\sqrt{c})$.

Proof. There is a distribution in the set $\Pi = \alpha' P' + (1 - \alpha')Q'$, where $|\alpha - \alpha'| \le \varepsilon$ and the means of P' and Q' are $\tilde{p}$ and $\tilde{q}$ as in Lemma 7.24. Then, we have $d_{\mathrm{TV}}(P, P') \le \|p - \tilde{p}\|/\sqrt{c} \le O(\delta/\sqrt{\alpha c})$ and $d_{\mathrm{TV}}(Q, Q') \le \|p - \tilde{p}\|/\sqrt{c} \le O(\delta/\sqrt{(1-\alpha)c})$. Thus, we have

$$d_{\mathrm{TV}}(\Pi', \Pi) \le O(\varepsilon) + \alpha d_{\mathrm{TV}}(P, P') + (1-\alpha)d_{\mathrm{TV}}(Q, Q') \le O(\varepsilon) + O((\sqrt{\alpha} + \sqrt{1-\alpha})\delta/\sqrt{c}) \le O(\delta/\sqrt{c}) \ .$$

$\qquad\square$

Next, we analyze the second case of the algorithm. We must show that Step 8 will find an r and t. First, we claim that there is a θ which makes r nearly perpendicular to $\mu^{S'_P} - \mu^{S'_Q}$.

Lemma 7.26. *There exists a* $r = (\cos\theta)u^* + (\sin\theta)v^*$, *with* θ *a multiple of* δ^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 θ be the nearest multiple of δ^2/d to θ'. 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

$$\begin{aligned}
|r \cdot z| &= |(\cos\theta)(u^* \cdot z) + (\sin\theta)(v^* \cdot z)| \\
&\le |(\cos\theta')(u^* \cdot z) + (\sin\theta')(v^* \cdot z)| + |\theta - \theta'|\sqrt{d} \\
&= |\sin\theta'||u^* \cdot z + (\cot\theta')(v^* \cdot z)| + |\theta - \theta'|\sqrt{d} \\
&\le 0 + \delta^2/\sqrt{d} \ .
\end{aligned}$$

$\qquad\square$

We now need to show that for this r, Step 8 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 7.27. *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*

$$\Pr_{X \in_u T}(r \cdot (X - \mu^{S'_P}) > 2t) > 12\exp(-(t-1)^2/4) + 3\varepsilon/d \ .$$

76

Proof. First, we wish to show that $\mathbb{E}_{X \in_u E}[(r \cdot (X - \mu^{S'_P}))^2]$ is large.

Since $r \in \text{span}(u^*, v^*)$, $|r^T\text{Cov}_0(S')r| \geq C\delta^2$. By Lemma 7.22, we have that

$$\underset{X \in_u T}{\text{Var}}(r \cdot X) = r^T\text{Cov}(T)r \geq r^T\text{Cov}_0(S')r - O(\delta^2) \geq (C - O(1))\delta^2 \geq (C/2)\delta^2 \ ,$$

for sufficiently large C and we also have that $r^T\text{Cov}_0(S')r$ is positive.

We note that

$$\begin{aligned}
r^T\text{Cov}(T)r &= \text{Var}(r \cdot T) \\
&= (|E|/|S'|) \underset{X \in_u E}{\text{Var}}(r \cdot X) + O(\alpha)(r \cdot (\mu - \mu^{S'_P}))^2 + O(1 - \alpha)(r \cdot (\mu - \mu^{S'_Q}))^2 \\
&\quad + (|E|/|S'|)(r \cdot (\mu - \mu^E))^2 \\
&= (|E|/|S'|)\left(\underset{X \in_u E}{\text{Var}}(r \cdot X) + (r \cdot (\mu - \mu^E))^2\right) + O(\delta^2) \ .
\end{aligned} \tag{37}$$

Now,

$$\underset{X \in_u E}{\mathbb{E}}[(r \cdot (X - \mu^{S'_P}))^2] = \underset{X \in_u E}{\text{Var}}(r \cdot X) + (r \cdot (\mu^{S'_P} - \mu^E))^2 \ .$$

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 (37), we have

$$(|E|/|S'|) \underset{X \in_u E}{\mathbb{E}}[(r \cdot (X - \mu^{S'_P}))^2] = (|E|/|S'|)[\underset{X \in_u E}{\text{Var}}[r \cdot X] + (r \cdot (\mu^{S'_P} - \mu^E))^2] - O(\delta^4) \gg C/2\delta^2.$$

Thus, for C sufficiently large,

$$\underset{X \in_u E}{\mathbb{E}}[(r \cdot (X - \mu^{S'_P}))^2] \gg \delta^2/\varepsilon.$$

Suppose for a contradiction that this lemma does not hold. Then, since $E \subset T$, we have

$$\underset{X \in_u E}{\text{Pr}}\left(r \cdot (X - \mu^{S'_P}) > 2t)\right) \leq (|S'|/|E|)12\exp(-t^2/2) + 3\varepsilon/d \ .$$

Thus, we have

$$\underset{X \in_u E}{\text{Pr}}(r \cdot (X - \mu^{S'_P}) > t)) \leq (|S'|/|E|)12\exp(-(t - 1)^2/4) + 3\varepsilon/d \ ,$$

and we can write

$$|S'|\delta^2 \ll |E| \underset{X\in_u E}{\mathbb{E}}[(r.X - r.\mu^{S_P})^2]$$

$$= |E| \int_0^{\sqrt{d}} \underset{X\in_u E}{\Pr}(r\cdot(X - \mu^{S_P}) > t))tdt$$

$$\ll |E| \int_0^{1+\sqrt{\ln(|S'|/|E|)/2}} tdt + |S'| \int_{1+\sqrt{\ln(|S'|/|E|)/2}}^{\infty} \exp(-(t-1)^2/4)tdt + \int_0^{\sqrt{n}} \varepsilon/dtdt$$

$$\ll |E|\ln(|S'|/|E|) + |E| + |S'|(|E|/|S|) + \varepsilon$$

$$\leq |S'|\cdot O(\varepsilon\log(1/\varepsilon)) .$$

Since we assumed that $\delta^2 \geq \Omega(\varepsilon\log(1/\varepsilon))$, this is a contradiction.

$\square$

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 7.28. *If $t \geq 1$,*

$$(1 - |E|/|S'|) \underset{X\in_u S'_P \cup S'_Q}{\Pr} \left(r\cdot(X - \mu^{S_P}) > t\right) \leq (5/4)\exp(-(t-1)^2/2) + \varepsilon/5d.$$

If $t \geq 1 + \sqrt{2\ln 6/\varepsilon}$, this is strictly less than $2\varepsilon/3$.

Proof.

$$\underset{X\in_u S'_P}{\Pr}(r\cdot(X - \mu^{S_P}) \leq t) \leq (|S_P|/|S'_P|)\underset{X\in_u S_P}{\Pr}(r\cdot(X - \mu^{S_P}) \leq t)$$

$$\leq (1 + O(\varepsilon)/\alpha)\cdot\left(\underset{X\sim P}{\Pr}(r\cdot(X - \mu^{S_P}) \leq t) + \varepsilon/(12d)\right)$$

$$= (1 + O(\varepsilon)/\alpha)\cdot\left(\underset{X\sim P}{\Pr}\left(r\cdot(X - p) \leq t - (r\cdot(\mu^{S_P} - p))\right) + \varepsilon/(6d)\right)$$

$$\leq (1 + O(\varepsilon)/\alpha)\cdot\left(2\exp(-(t-1/2)^2/2) + \varepsilon/6d\right) .$$
$$\text{(using Lemma 7.21 and Hoeffding's inequality)}$$

Similarly,

$$\underset{X\sim S'_Q}{\Pr}(r\cdot(X - \mu^{S_Q}) \leq t) \leq (1 + O(\varepsilon)/(1-\alpha))\cdot(2\exp(-(t-1/2)^2/2) + \varepsilon/6d).$$

Since $|r\cdot(\mu^{S_Q} - \mu^{S_P})| \leq \delta^2/\sqrt{d} \leq 1/2$, we have

$$\underset{X\sim S'_Q}{\Pr}\left(r\cdot(X - \mu^{S_Q}) \leq t\right) \leq (1 + O(\varepsilon)/(1-\alpha))\cdot\left(2\exp(-(t-1)^2/2) + \varepsilon/6d\right) .$$

78

Noting that $1 - (|S'_P| + |S'_Q|)/|S'| = |E|/|S'| \geq 4\varepsilon/3$, we have

$$(1 - |E|/|S'|) \Pr_{X \in_u S'_P \cup S'_Q} \left(r \cdot (X - \mu^{S'_P}) > t \right)$$

$$= (|S'_P|/|S'|) \Pr_{X \sim S'_P} \left(r \cdot (X - \mu^{S'_P}) > t \right) + (|S'_Q|/|S'|) \Pr_{X \sim S'_Q} \left(r \cdot (X - \mu^{S'_P}) > t \right)$$

$$= ((\alpha + O(\varepsilon))(1 + O(\varepsilon)/\alpha) + (1 - \alpha + O(\varepsilon))(1 + O(\varepsilon)/(1 - \alpha)) \cdot (2\exp(-(t-1)^2/2) + \varepsilon/6d)$$

$$\leq (1 + O(\varepsilon)) \cdot (2\exp(-(t-1)^2/2) + \varepsilon/6d)$$

$$\leq (5/2)\exp(-(t-1)^2/2) + \varepsilon/5d .$$

If $t \geq 1 + \sqrt{2\ln 6/\varepsilon}$, this expression is $(5/2)(\varepsilon/6) + \varepsilon/5d \leq 2\varepsilon/3$. $\qquad\square$

Now we can finally show that a t exists for this r, so Step 8 will succeed:

Lemma 7.29. *There is a $t \geq 1 + 2\sqrt{\ln(9/\varepsilon)}$ such that*

$$\Pr_{X \in_u S'} \left(\Pr_{Y \in_u S'} (r \cdot (X - Y) > t) < 2\varepsilon \right) > 12\exp(-(t-1)^2/4) + 3\varepsilon/d.$$

Proof. By Lemma 7.27, there exists a $t \geq 1$ such that

$$\Pr_{X \in_u T} \left(r \cdot (X - \mu^{S'_P}) > 2t) \right) > 12\exp(-(t-1)^2/4) + 3\varepsilon/d.$$

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

$$\Pr_{X \in_u S'} \left(r \cdot (X - \mu^{S'_P}) > 2t \right) \geq \Pr_{X \in_u T} \left(r \cdot (X - \mu^{S'_P}) > 2t \right) > 12\exp(-(t-1)^2/4) + 3\varepsilon/d. \quad (38)$$

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, $t \geq 1 + 2\sqrt{\ln(9/\varepsilon)}$.

Thus, by Lemma 7.28, 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 Equation (38) gives the lemma. $\qquad\square$

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 7.30. *For any unit vector r' and $t \geq 2\sqrt{\ln(1/\varepsilon)}$, we have*

$$(1 - |E|/|S'|) \Pr_{X \in_u S'_P \cup S'_Q} \left(\Pr_{Y \in_u S'} (r' \cdot (X - Y) \leq t) < 2\varepsilon \right) \leq 3\exp(-t^2/4) + \varepsilon/4d .$$

Proof. Using Hoeffding's inequality, we have

$$
|S_P'| \Pr_{X \in_u S_P'} (r \cdot (p - X) > t/2) \leq |S_P| \Pr_{X \in_u S_P} \left(|r' \cdot (X - p)| > t/2 \right)
$$

$$
\leq |S_P| \left(\Pr_{X \sim P} (|r' \cdot (X - p)| > t/2) + \varepsilon/6d \right)
$$

$$
\leq |S_P| \left(2 \exp(-t^2/4) + \varepsilon/6d \right) . \tag{39}
$$

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

$$
\Pr_{Y \in_u S'} (r' \cdot (x - Y) \leq t) \geq (|S_P|/|S'|) - (|S_P|/|S'|)(2 \exp(-t^2/4) + \varepsilon/6d) .
$$

When $t \geq 2\sqrt{\ln(1/\varepsilon)}$, $(|S_P/|S'|)(2 \exp(-t^2/4) + \varepsilon/6d) \leq (1+2\varepsilon) \cdot (2\varepsilon + \varepsilon/6d) \leq 3\varepsilon$. Also, $(|S_P|/|S'|) \leq ((\alpha - \varepsilon/6)|S|)/(|S|(1 - 2\varepsilon)) \leq \alpha - 3\varepsilon \leq 7\varepsilon$. Thus, we have $\Pr_{Y \in_u S'}(r \cdot (x - Y) \leq t) \geq 4\varepsilon > 2\varepsilon$.

But inequality (39) gives a bound on the number of x in S_P that do not satisfy this condition. That is,

$$
|S_P'| \Pr_{X \in_u S_P'} \left(\Pr_{Y \in_u S'} (r' \cdot (X - Y) \leq t) < 2\varepsilon \right) \leq |S_P| \left(2 \exp(-t^2/4) + \varepsilon/6d \right) .
$$

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 $|S_Q'| \Pr_{X \in_u S_Q'} (r' \cdot (X - p) > t/2) \leq (2 \exp(-t^2/4) + \varepsilon/6d)$. Thus,

$$
|S_Q'| \Pr_{X \in_u S_Q'} \left(\Pr_{Y \in_u S'} (r' \cdot (X - Y) \leq t) < 2\varepsilon \right) \leq |S_Q| \left(2 \exp(-t^2/4) + \varepsilon/6d \right) .
$$

Summing these gives

$$
(|S_P'| + |S_Q'|) \Pr_{X \in_u S_P' \cup S_Q'} \left(\Pr_{Y \in_u S'} (r' \cdot (X - Y) \leq t) < 2\varepsilon \right) \leq |S| \left(2 \exp(-t^2/4) + \varepsilon/6d \right) .
$$

Dividing by $|S'|$ and noting that $|S| \leq (1 + 2\varepsilon)|S'| \leq (3/2)|S'|$ completes the proof. $\square$

Now, we can show that the filter improves $\Delta(S, S'')$, and so that the algorithm is correct in the filter case.

Claim 7.31. *If we reach Step 9 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 so 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, we found a vector r and $t \geq 1 + 2\sqrt{\ln(1/\varepsilon)}$ such that

$$\Pr_{X \in_u S'} \left(\Pr_{Y \in_u S'} (|r \cdot (X - Y)| < t) < 2\varepsilon \right) > 12 \exp(-(t-1)^2/4) + 3\varepsilon/d \ .$$

Then in Step 9, we remove at least a $12 \exp(-t^2/4) + 3\varepsilon/d$ fraction of points. That is,

$$|S'' \setminus S'| \geq (12 \exp(-t^2/4) + 3\varepsilon/d)|S'| \ .$$

The fact that $t \geq 1 + 2\sqrt{\ln(1/\varepsilon)}$ allows us to use Lemma 7.30, with $r' = r$, yielding that:

$$(1 - |E|/|S'|) \Pr_{X \in_u S'_P \cup S'_Q} \left(\Pr_{Y \in_u S'} (r \cdot (X - Y) \leq t - 1) < 2\varepsilon \right) \leq 3 \exp(-(t-1)^2/4) + \varepsilon/(4d) \ .$$

This implies that

$$(1 - |E|/|S'|) \Pr_{X \in_u S'_P \cup S'_Q} \left(\Pr_{Y \in_u S'} (r \cdot (X - Y) < t) < 2\varepsilon \right) \leq 3 \exp(-(t-1)^2/4) + \varepsilon/(4d) \ .$$

Thus,

$$|S''_P \setminus S'_P| + |S''_Q \setminus S'_Q| \leq (3 \exp(-(t-1)^2/4) + \varepsilon/4d)|S'| \ ,$$

and we have

$$
\begin{aligned}
\Delta(S, S') - \Delta(S, S'') &\geq \left(12 \exp(-(t-1)^2/4) + 3\varepsilon/d - 2 \left(3 \exp(-(t-1)^2/4) + \varepsilon/4d \right) \right) |S'|/|S| \\
&\geq 2\varepsilon/d \ ,
\end{aligned}
$$

since $|S'| \geq |S|(1 - \Delta(S, S')) \geq (1 - 2\varepsilon)|S| \geq 5|S|/6$. $\qquad\square$

8 Agnostically Learning a Gaussian, via Filters

8.1 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:

Theorem 8.1. *Let G be a Gaussian with identity covariance matrix and $\varepsilon, \tau > 0$. Let S be a multiset of $\Omega(d^3 \log(d/\varepsilon\tau))^4/\varepsilon^2)$ independent samples from G and S' be a multiset obtained by arbitrarily changing an ε-fraction of the points in S. There exists a polynomial time algorithm that returns a Gaussian G' so that with probability at least $1 - \tau$ we have $d_{\mathrm{TV}}(G, G') = O(\varepsilon\sqrt{\log(1/\varepsilon)})$.*

Let $\mu^G, \mu^{G'}$ denote the mean vectors of G, G' respectively. Our output hypothesis will be a Gaussian G' with identity covariance matrix, such that $\|\mu^G - \mu^{G'}\|_2 = O(\varepsilon\sqrt{\log(1/\varepsilon)})$. The desired bound on the total variation distance follows from Corollary 2.13.

Notation. We will denote by $\mu^S = \frac{1}{|S|}\sum_{X \in S} X$ and $\mathrm{Cov}[S] = \frac{1}{|S|}\sum_{X \in S}(X - \mu^G)(X - \mu^G)^T$ the sample mean and sample covariance matrix of the set S.

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

Definition 8.2. *Let G be a Gaussian in d dimensions with identity covariance and $\varepsilon, \tau > 0$. We say that a multiset S of elements in $\mathbb{R}^d$ is (ε, τ)-good with respect to G if the following conditions are satisfied:*

(i) *For all $x \in S$ we have $\|x - \mu_G\|_2 \leq O(\sqrt{d \log(|S|/\tau)})$.*

(ii) *For every affine function $L : \mathbb{R}^d \to \mathbb{R}$ we have $|\Pr_{X \in_u S}(L(X) \geq 0) - \Pr_{X \sim G}(L(X) \geq 0)| \leq \varepsilon/(d \log(d/\varepsilon\tau))$.*

(iii) *We have that $\|\mu^S - \mu^G\|_2 \leq \varepsilon$.*

(iv) *We have that $\|\mathrm{Cov}[S] - I\|_2 \leq \varepsilon/d$.*

Next, we show that a sufficiently large set of independent samples from G is (ε, τ)-good (with respect to G) with high probability. The proof of the following lemma is deferred to Section F.

Lemma 8.3. *Let G be a Gaussian with identity covariance and $\varepsilon, \tau > 0$. If the multiset S is obtained by taking $\Omega(d^3 \log(d/\varepsilon\tau)/\varepsilon^2)$ independent samples from G, it is (ε, τ)-good with respect to G with probability at least $1 - \tau$.*

We recall our definition of $\Delta(S, S')$ that quantifies the extent to which a multiset has been corrupted.

Definition 8.4. *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 8.5. *Let G be a Gaussian with identity covariance, $\varepsilon > 0$ be sufficiently small and $\tau > 0$. Let S be an (ε, τ)-good set with respect to G. Let S' be any multiset with $\Delta(S, S') \leq 2\varepsilon$ and for any $x, y \in S'$, $\|x - y\|_2 \leq O(\sqrt{d \log(d\varepsilon/\tau)})$. 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 μ such that $d_{\mathrm{TV}}(\mathcal{N}(\mu, I), G) = O(\varepsilon\sqrt{\log(1/\varepsilon)})$.*

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

We start by showing how Theorem 8.1 follows easily from Proposition 8.5.

Proof of Theorem 8.1. By the definition of $\Delta(S, S')$, since S' has been obtained from S by corrupting an ε-fraction of the points in S, we have that $\Delta(S, S') \leq 2\varepsilon$. By Lemma 8.3, the set S of uncorrupted samples is (ε, τ)-good with respect to G with probability at least $1 - \tau$. We henceforth condition on this event.

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

Then we iteratively apply the FILTER-GAUSSIAN-UNKNOWN-MEAN procedure of Proposition 8.5 until it terminates returning a mean vector μ with $d_{\mathrm{TV}}(\mathcal{N}(\mu, I), G) = O(\varepsilon\sqrt{\log(1/\varepsilon)})$. We claim that we need at most $O(d \log(d/\varepsilon\tau))$ iterations for this to happen. Indeed, the sequence of iterations results in a sequence of sets S_i', so that $\Delta(S, S_i') \leq \Delta(S, S') - i \cdot \varepsilon/(d \log(d/\varepsilon\tau))$. Thus if we do not terminate in the first $2d \log(d/\varepsilon\tau)$ iterations, in the next iteration the algorithm terminates outputting the sample mean of S. $\qquad\square$

8.1.1 Algorithm FILTER-GAUSSIAN-UNKOWN-MEAN: Proof of Proposition 8.5

In this subsection, we describe the efficient algorithm establishing Proposition 8.5 and prove its correctness. Our filter algorithm 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 Σ has no large eigenvalues, we return $\mu^{S'}$. Otherwise, we use the eigenvector v^* corresponding to the maximum magnitude eigenvalue of Σ and the mean vector $\mu^{S'}$ to define a filter. Our efficient filtering procedure is presented in detailed pseudocode below.

Algorithm 14 Filter algorithm for a Gaussian with unknown mean and identity covariance

1: **procedure** FILTER-GAUSSIAN-UNKOWN-MEAN(S', ε, τ)
input: A multiset S' such that there exists an (ε, τ)-good S with $\Delta(S, S') \leq 2\varepsilon$
output: Multiset S'' or mean vector p' satisfying Proposition 8.5

2:    Compute the sample mean $\mu^{S'} = \mathbb{E}_{X \in_u S'}[X]$ and the sample covariance matrix Σ, i.e., $\Sigma = (\Sigma_{i,j})_{1 \leq i,j \leq d}$ 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** $\|\Sigma - I\|_2 \leq O(\varepsilon \log(1/\varepsilon))$, **then return** $\mu^{S'}$.

5:    Let $\delta := 3\sqrt{\varepsilon \|\Sigma - I\|_2}$. Find $T > 0$ such that

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

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

The rest of this section is devoted to the analysis of our algorithm FILTER-GAUSSIAN-UNKOWN-MEAN.

8.1.2 Proof of Correctness of FILTER-GAUSSIAN-UNKOWN-MEAN

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 $\Delta(S, S') \leq 2\varepsilon$ 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 ε is a sufficiently small constant.

We define $\mu^G, \mu^S, \mu^{S'}, \mu^L$, and μ^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 mean.

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

Fact 8.6. *Let $w \in \mathbb{R}^d$ be any unit vector, then for any $T > 0$, $\Pr_{X \in_u S}(|w \cdot (X - \mu^G)| > T) \leq 2 \exp(-T^2/2) + \varepsilon/(d \log(d/\varepsilon\tau))$, and $\Pr_{X \sim G}(|w \cdot (X - \mu^G)| > T) \leq 2 \exp(-T^2/2)$.*

Proof. The second line is a standard concentration result for Gaussians, and the former follows from it using the goodness of S. □

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

Claim 8.7. *Let $w \in \mathbb{R}^d$ be any unit vector, then for any $T > 0$,*

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

and

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

Proof. This follows from Fact 8.6 upon noting that $|w \cdot (X - \mu^{S'})| > T + \|\mu^{S'} - \mu^G\|_2$ only if $|w \cdot (X - \mu^G)| > T$. □

We can also use this to prove concentration bounds for L. In particular:

Corollary 8.8. $\|M_L\|_2 = O(\log(|L|/|S|))$.

Proof. We need to bound $v^T M_L v = \mathbb{E}_{X \in_u L}[(v \cdot (X - \mu^G))^2]$ for all unit vectors v. This is

$$\int_0^\infty \Pr_{X \in_u L}(|v \cdot (X - \mu^G)| > T) 2T \, dT = \int_0^{O(d \log(d/\varepsilon\tau))} \Pr_{X \in_u L}(|v \cdot (X - \mu^G)| > T) 2T \, dT.$$

Our bound then follows from noting that, since L is a subset of S, we have that

$$\Pr_{X \in_u L}(|v \cdot (X - \mu^G)| > T) \leq \min \left\{ 1, (|S|/|L|) \cdot \Pr_{X \in_u S}(|v \cdot (X - \mu^G)| > T) \right\},$$

and using Claim 8.7. □

Next, we prove a bound on $M_{S'}$.

Lemma 8.9. *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. Notice that

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

where the second line is by the goodness of S. Therefore, we have that

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

□

We get a similarly useful bound on the difference between the mean vectors:

Lemma 8.10. *We have that*

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

where the $O(\varepsilon\sqrt{\log(1/\varepsilon)})$ denotes a vector with ℓ_2 norm at most $O(\varepsilon\sqrt{\log(1/\varepsilon)})$.

Proof.

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

Since S is good $\|\mu^S - \mu^G\|_2 = O(\varepsilon)$. Since $\|M_L\| \gg (\mu^L - \mu^G)(\mu^L - \mu^G)^T$, we have that $\|\mu^L - \mu^G\|_2 \le O(\sqrt{\log(|L|/|S|)})$. Therefore, we have that

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

as desired. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

From this we conclude that $\Sigma - I$ is approximately proportional to M_E.

Corollary 8.11.

$$\Sigma - I = (|E|/|S'|)M_E + O(\varepsilon\log(1/\varepsilon)) + O(|E|/|S|)^2\|M_E\|_2.$$

Proof. We have that

$$\Sigma - I = M_{S'} - I - (\mu^{S'} - \mu^G)(\mu^{S'} - \mu^G)^T.$$

By Lemmas 8.9 and 8.10, this is

$$(|E|/|S'|)M_E + O(\varepsilon\log(1/\varepsilon)) + O((|E|/|S|)^2\|\mu^E - \mu^G\|_2^2) = (|E|/|S'|)M_E + O(\varepsilon\log(1/\varepsilon)) + O(|E|/|S|)^2\|M_E\|_2.$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$$

From this we can immediately see that if our algorithm returns on Step 4, then it returns a correct answer. In particular, if $\lambda^* = \|\Sigma - I\|_2 = O(\varepsilon\log(1/\varepsilon))$, we have by Corollary 8.11 that

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

Therefore,

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

On the other hand, we have $\|M_E\|_2 \ge \|\mu^E - \mu^G\|_2^2$, and therefore by Lemma 8.10,

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

Next, we need to show that if $\lambda^* > C\varepsilon\log(1/\varepsilon)$ that a T as in Step 5 exists. To show this, we note again by Corollary 8.11 that

$$\lambda^* = (v^*)^T(\Sigma - I)v^* = (|E|/|S'| + O(|E|/|S|)^2)(v^*)^T M_E v^* + O(\varepsilon\log(1/\varepsilon)).$$

Therefore,

$$\mathop{\mathbb{E}}_{X \in_u E}[|v^* \cdot (X - \mu^G)|^2] = (v^*)^T M_E v^* \gg (|S'|/|E|)\lambda^*.$$

On the other hand,

$$\mathbb{E}_{X \in_u E}[|v^* \cdot (X - \mu^G)|^2] = \int_0^{O(\sqrt{d \log(d/\varepsilon\tau)})} \Pr_{X \in_u E}(|v^* \cdot (X - \mu^G)| > T)2TdT$$

$$\leq |S'|/|E| \int_{\sqrt{\lambda^*}/10}^{O(\sqrt{d \log(d/\varepsilon\tau)})} \Pr_{X \in_u S'}(|v^* \cdot (X - \mu^G)| > T)2TdT + \lambda^*/100.$$

Therefore, it must be the case that for some value of $T' > \sqrt{\lambda^*}/10$ for which $\Pr_{X \in_u S'}(|v^* \cdot (X - \mu^G)| > T') > \exp(-T'^2/4) + 8\varepsilon/(d \log(d\varepsilon/\tau))$. Letting $T = T' - \delta$ yields the necessary T.

Finally, we need to show that if our algorithm returns an S'' in Step 6 that $\Delta(S, S'') < \Delta(S, S') - 2\varepsilon/(d \log(d/\varepsilon\tau))$.

Claim 8.12. *We have that* $\Delta(S, S'') \leq \Delta(S, S') - 2\varepsilon/(d \log(d/\varepsilon\tau))$.

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

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

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

8.2 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 8.13. *Let G be a Gaussian in d dimensions with mean 0 and unknown covariance, and let $\varepsilon, \tau > 0$. Let S be a set of N independent samples from G for $N = \Omega((d \log(d/\varepsilon\tau))^6/\varepsilon^2)$. Let S' be a set of N elements of $\mathbb{R}^d$ obtained from S by adversarially changing at most an ε-fraction of the samples of S. There exists an algorithm A that, given S' and ε, returns the parameters of a Gaussian distribution G' so that with probability at least $1 - \tau$ over the samples from S, it holds $d_{\mathrm{TV}}(G, G') \leq \delta = O(\varepsilon \log(1/\varepsilon))$.*

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

Definition 8.14. *Let G be a d-dimensional Gaussian with mean 0 and covariance matrix Σ. Let S be a set of N points of $\mathbb{R}^d$, and $\varepsilon, \tau > 0$. We say that S is (ε, τ)-good with respect to G if the following conditions hold:*

- *For all $x \in S$ we have that $x^T \Sigma^{-1} x = O(d \log(N/\tau))$.*

- *For any even degree-2 polynomial $p : \mathbb{R}^d \to \mathbb{R}$ we have that $|\mathbb{E}_{X \sim G}[p(X)] - \mathbb{E}_{X \in_u S}[p(X)]| \leq \varepsilon \sqrt{\mathbb{E}_{X \sim G}[p(X)^2]}$, $|\mathbb{E}_{X \sim G}[p^2(X)] - \mathbb{E}_{X \in_u S}[p^2(X)]| \leq \varepsilon \mathbb{E}_{X \sim G}[p(X)^2]$, and*

$$\Pr_{X \in_u S}(p(X) \geq 0) \leq 2 \Pr_{X \sim G}(p(X) \geq 0) + \varepsilon/(d \log(N/\tau))^2 \ .$$

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

Lemma 8.15. *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 Section F.

Lemma 8.16. *Let G be a d-dimensional Gaussian with mean 0 and let $\varepsilon, \tau > 0$. Then, for $\Omega((d \log(d/\varepsilon\tau))^6/\varepsilon^2) \leq N \leq \mathrm{poly}(d/\varepsilon\tau)$, we have that if S is obtained by taking N independent samples from G, then S is (ε, τ)-good with respect to G with probability at least $1 - \tau$.*

As in Definition 8.4, $\Delta(S, S')$ is the be the size of the symmetric difference of S and S' divided by $|S|$.

The basic thrust of our algorithm is as follows. By Lemma 8.16, with probability $1 - \tau$ we have that S is (ε, τ)-good with respect to G. The algorithm is then handed a new set S' so 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'' so that $\Delta(S, S'') < \Delta(S, S')$. In the latter case, the algorithm will recurse on S''. We formalize this idea below:

Proposition 8.17. *There is an algorithm that given a set finite $S' \subset \mathbb{R}^d$, so that there is a mean 0 Gaussian G and a set S that is (ε, τ)-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 8.17, the proof of Theorem 8.13 is straightforward. By Lemma 8.16 the original set S is (ε, τ)-good with respect to G with probability at least $1 - \tau$. Then, S' satisfies the hypotheses of Proposition 8.17. We then repeatedly iterate the algorithm from Proposition 8.17 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.

8.2.1 Analysis of Filter-based Algorithm: Proof of Proposition 8.17

We now turn our attention to the proof of Proposition 8.17. We first define the matrix Σ' to be $\mathbb{E}_{X \in S'}[XX^T]$, and let G' be the mean 0 Gaussian with covariance matrix Σ'. Our goal will be to either obtain proof 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 $\|M\|_F = 1$ so that

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

is far from 0. On the other hand, we know that the distribution of $p(X) = (\Sigma^{-1/2}X)^T M(\Sigma^{-1/2}X) - \text{tr}(M)$ 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 15 Filter algorithm for a Gaussian with unknown covariance matrix.

1: **procedure** FILTER-GAUSSIAN-UNKNOWN-COVARIANCE(S', ε, τ)

input: A multiset S' such that there exists an (ε, τ)-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 $d_{\text{TV}}(G, G') = O(\varepsilon \log(1/\varepsilon))$.

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

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

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

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

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

6:     Define two quadratic forms on L

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

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

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

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

9:         **return** G'

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

11:    Find a $T \geq C'$ so that

$$\mathop{\Pr}_{X \in_u S'}(|p^*(X) - \mu| \geq T) \geq 12 \exp(-(T - 4/3) + 3\varepsilon/(d\log(N/\tau))^2$$

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

    Thr function FIND-MAX-POLY uses similar notation to SEPARATIONORACLEUNKNOWNCOVARIANCE so that FILTER-GAUSSIAN-UNKNOWN-COVARIANCE and SEPARATIONORACLEUNKNOWNCOVARIANCE can be more easily compared.

Algorithm 16 Algorithm for maximizing $Q_{S'}(p)/Q_{G'}(p)$.

1: **function** FIND-MAX-POLY(S', Σ')

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 $\lambda^* = Q_{S'}(p^*)$.

2:       Compute an approximate eigendecomposition of Σ' and use it to compute $\Sigma'^{-1/2}$

3:       Let $x_{(1)}, \ldots, x_{(|S'|)}$ be the elements of S'.

4:       For $i = 1, \ldots, |S'|$, let $y_{(i)} = \Sigma'^{-1/2} x_{(i)}$ and $z_{(i)} = y_{(i)}^{\otimes 2}$.

5:       Let $T_{S'} = I^\flat I^{\flat T} + (1/|S'|) \sum_{i=1}^{|S'|} z_{(i)} z_{(i)}^T$.

6:       Approximate the top eigenvalue λ^* and corresponding unit eigenvector v^* of $T_{S'}$..

7:       Let $p^*(x) = \frac{1}{\sqrt{2}}((\Sigma'^{-1/2} x)^T v^{*\sharp}(\Sigma'^{-1/2} x) - \text{tr}(v^{*\sharp}))$

8:       **return** p^* and λ^*.

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

Claim 8.18. *Algorithm* FIND-MAX-POLY *is correct and* FILTER-GAUSSIAN-UNKNOWN-COVARIANCE *runs time* $\text{poly}(d \log \tau / \varepsilon)$.

Proof. First, assume that we can compute all eigenvalues and eigenvectors exactly. By Lemma 8.15 all even polynomials with degree-2 that have $\mathbb{E}_{X \sim G}[p(X)] = 0$ can be written as $p(x) = (\Sigma'^{-1/2} x)^T P_2 (\Sigma'^{-1/2} x) - \text{tr}(P_2)$ 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 $\text{Var}_{X \sim G'}[p(X)] = 2\|P_2\|_F = \|v_2\| = 1$.

Note that since the covariance matrix of S' is Σ', we have

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

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

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

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 $\text{poly}(\varepsilon/d)$ using standard algorithms to compute the eigen-decomposition 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 $\text{poly}(|S'|d\log(1\tau)/\varepsilon)$ time. $\qquad\square$

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 Σ' 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 Σ, and thus that either Σ' is very close to Σ or else Σ_E is very large in some direction.

Lemma 8.19. *We have that*

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

Proof. Note that, since the matrix inner product is an inner product,

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

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

Since $\text{tr}(M\Sigma^{-1/2}\Sigma_{S \setminus L}\Sigma^{-1/2}) = \text{tr}(M^T\Sigma^{-1/2}\Sigma_{S \setminus L}\Sigma^{-1/2}) = \text{tr}(\frac{1}{2}(M + M^T)\Sigma^{-1/2}\Sigma_{S \setminus L}\Sigma^{-1/2})$ and $\|\frac{1}{2}(M + M^T)\|_F \leq \frac{1}{2}(\|M\|_F + \|M^T\|_F) = 1$, we may assume WLOG that M is symmetric.

Consider such an M. We note that

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

Let $p(x)$ denote the quadratic polynomial

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

By Lemma 8.15, $\mathbb{E}_{X \sim G}[p(X)] = 0$ and $\text{Var}_{X \sim G}[p(X)] = 2\|M\|_F^2 = 2$.

Since S is (ε, τ)-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

$$\frac{\sum_{X \in L} p(X)}{|S|} = O(\varepsilon \log(1/\varepsilon)).$$

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 ε-fraction of the mass cannot change the value of the

mean by very much. In particular, we have that

$$\frac{\sum_{x \in L} p(x)}{|S|} \leq \int_0^\infty \frac{|\{x \in L : p(x) > t\}|}{|S|} dt$$

$$\leq \int_0^\infty \min(\varepsilon, \Pr_{X \in_u S}(p(X) > t)) dt$$

$$\leq \int_0^{O(d \log(N/\tau))} \min(\varepsilon, 2\Pr(p(G) > t)) + \varepsilon/(d \log(N/\tau))^2 dt$$

$$\leq \varepsilon + \int_0^\infty \min(\varepsilon, 2\exp(-\Omega(t-1))) dt$$

$$\leq \varepsilon + \int_0^{C \log(1/\varepsilon)} \varepsilon dt + \int_{C \log(1/\varepsilon)-1}^\infty 2\exp(-\Omega(t)) dt$$

$$\leq \varepsilon + O(\varepsilon \log(1/\varepsilon)) + O(\varepsilon)$$

$$= O(\varepsilon \log(1/\varepsilon)) \,,$$

where C above is a sufficiently large constant. The third line uses that

$$\max_{x \in S} |p(x)| \leq \max_{x \in S} \|\Sigma^{-1/2} x\|_2^2 \|M\|_2 + \mathrm{tr} M \leq O(d \log(N/\tau)) \cdot 1 + \sqrt{d} = O(d \log(N/\tau)) \,.$$

This completes the proof. $\qquad\square$

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

Corollary 8.20.
$$\Sigma' \geq (1 - O(\varepsilon \log(1/\varepsilon)))\Sigma.$$

Proof. Lemma 8.19 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, $\Sigma_{S\setminus L} \geq (1 + O(\varepsilon \log(1/\varepsilon)))\Sigma$. Our result now follows from noting that $\Sigma' = \Sigma_{S\setminus L}((|S|-|L|)/|S'|) + \Sigma_E(|E|/|S'|)$, and $\Sigma_E = \mathbb{E}_{X \in_u E}[XX^T] \geq 0$. $\qquad\square$

The first step in verifying correctness is to note that if our algorithm returns on Step 4 that it does so correctly.

Claim 8.21. *If our algorithm returns on Step 4, 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 (ε, τ)-goodness, all $x \in S$ satisfy

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

for C sufficiently large. $\qquad\square$

Next, we need to show that if our algorithm returns a G' in Step 9 that $d_{\mathrm{TV}}(G, G')$ is small.

Claim 8.22. *If our algorithm returns in Step 9, 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

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

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 $\|M\|_F = 1$ so that

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

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.

Letting $p(X)$ be the polynomial

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

Using Lemma 8.15, $\mathbb{E}_{X \sim G}[p(X)] = 0$ and $\text{Var}_{X \sim G}[p(X)] = 2$. 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 8.19 we can show that $\mathbb{E}_{X \in_u S\setminus L}[p^2(X)]$ is not much less than 2. In particular,

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

On the one hand, we have that

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

by assumption. On the other hand, we have that

$$\frac{\sum_{X \in L} p^2(X)}{|S|} \leq \int_0^\infty \frac{|\{X \in L : p^2(X) > t\}|}{|S|} dt$$
$$\leq \int_0^\infty \min(\varepsilon, \Pr_{X \in_u S}(p^2(X) > t)) dt$$
$$\leq \int_0^{O(d\log(N/\tau))^2} \min(\varepsilon, 2\Pr(p^2(G) > t)) + \varepsilon/(d\log(N/\tau)^2) dt$$
$$\leq \varepsilon + \int_0^\infty \min(\varepsilon, 2\exp(-\Omega(\sqrt{t-1}))) dt$$
$$\leq \varepsilon + \int_0^{C\log^2(1/\varepsilon)} \varepsilon dt + \int_{C\log^2(1/\varepsilon)-1}^\infty 2\exp(-\Omega(\sqrt{t})) dt$$
$$\leq \varepsilon + O(\varepsilon \log^2(1/\varepsilon)) + O(\varepsilon \log(1/\varepsilon))$$
$$= O(\varepsilon \log^2(1/\varepsilon)).$$

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 $Q_{S'}(p) \leq 2 + O(\varepsilon \log^2(1/\varepsilon))$, this implies that $(|E|/|S'|)\mathbb{E}_{X \in_u E}[p^2(X)] = O(\varepsilon \log^2(1/\varepsilon))$. By Cauchy-Schwartz, this implies that

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

Thus,

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

This shows that if the algorithm returns in this step, it does so correctly. $\qquad\square$

Next, we need to show that if the algorithm reaches Step 11 that such a T exists.

Claim 8.23. *If the algorithm reaches Step 11, then there exists a $T > 1$ so that*

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

Proof. Before we begin, we will need the following critical Lemma:

Lemma 8.24.

$$\operatorname*{Var}_{X \sim G}[p(X)] \leq 1 + O(\varepsilon \log(1/\varepsilon)).$$

Proof. We note that since $\operatorname{Var}_{X \sim G'}(p(G')) = Q_{G'}(p) = 1$, that we just need to show that the this 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 8.20.

Using Lemma 8.15, we can write

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

where $\|P_2\|_F = \frac{1}{2} \operatorname{Var}_{X \sim G'}(p(G')) = \frac{1}{2}$ and $p_0 = \mu + \operatorname{tr}(P_2)$. 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 $\operatorname{Var}_{X \sim G}[p(X)] = \|M\|_F$. Here, M is the matrix $\Sigma^{1/2}\Sigma'^{-1/2} P_2 \Sigma'^{-1/2}\Sigma^{1/2}$. By Corollary 8.20, it holds $\Sigma' \geq (1 - O(\varepsilon \log(1/\varepsilon)))\Sigma$. 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

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

$\square$

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

$$\left|\mu - \operatorname*{\mathbb{E}}_{X \sim G}[p(X)]\right| \leq 2\sqrt{\operatorname*{Var}_{X \sim G}[p(X)]} \leq 3 . \tag{40}$$

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.

Let $\hat{\mu}$ be $\mathbb{E}_{X \sim G}[p(X)]$. In particular, we have that $\operatorname{Var}_{X \in_u S'}(p(X)) = Q_{S'}(p) \geq 1 + C\varepsilon \log^2(1/\varepsilon)$, which means that

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

Now, because S is good, we know that

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

93

Therefore, we have that

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

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

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

and therefore,

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

On the other hand, we have that

$$\sum_{X \in E} |p(X) - \mu|^2 = \int_0^\infty \{X \in E : |p(X) - \mu| > t\} 2t dt$$

$$\le \int_0^{C^{1/4} \log(1/\varepsilon)} O(t\varepsilon|S'|) dt + \int_{C^{1/4} \log(1/\varepsilon)}^{(d \log(N/\tau))} \{X \in E : |p(X) - \mu| > t\} 2t dt$$

$$\le O(C^{1/2}\varepsilon \log^2(1/\varepsilon)|S'|) + |S'| \int_{C^{1/4} \log(1/\varepsilon)}^{(d \log(N/\tau))} \Pr_{X \in_u S'}(|p(X) - \mu| > t) 2t dt .$$

Therefore, we have that

$$\int_{C^{1/4} \log(1/\varepsilon)}^{(d \log(N/\tau))} \Pr_{X \in_u S'}(|p(X) - \mu| > t) 2t dt \ge (C/4)\varepsilon \log^2(1/\varepsilon) . \tag{41}$$

Assume for sake of contradiction that

$$\Pr_{X \in_u S'}(|p(X) - \mu| \ge T) < 12 \exp(-(T - 4)/3) + 3\varepsilon/(d \log(N/\tau))^2 ,$$

for all $T > 1$. Plugging this into Equation (41), we find that

$$\int_{C^{1/4} \log(1/\varepsilon)-4}^{(d \log(N/\tau))} 24 \exp(-t/3) dt + \int_{C^{1/4} \log(1/\varepsilon)}^{(d \log(N/\tau))} 6t\varepsilon/(d \log(N/\tau))^2 dt \ge (C/4)\varepsilon \log^2(1/\varepsilon) ,$$

which yields a contradiction. $\square$

Finally, we need to verify that if our algorithm returns output in Step 12, that it is correct.

Claim 8.25. *If the algorithm returns during Step 12, then $\Delta(S, S'') \le \Delta(S, S') - \varepsilon/(d \log(N/\tau))^2$.*

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

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

For this, it suffices to show that

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

or that

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

94

By assumption, we have that

$$|\{X \in S' : |p(X) - \mu| > T\}| > (12 \exp(-(T - 4)/3) + 3\varepsilon/(d \log(N/\tau))^2)|S'| .$$

On the other hand, we have that

$$|\{X \in S : |p(X) - \mu| > T\}| \leq |S|(2 \Pr(|p(G) - \mu| > T) + \varepsilon/(d \log(N/\tau))^2)$$
$$\leq 3|S'| \Pr_{X \sim G}(|p(X) - \mu| > T) + 3\varepsilon/(d \log(N/\tau))^2)|S'| .$$

On the other hand by the last claim, $p(X)$ for $X \sim G$ is an even degree-2 polynomial of a Gaussian with variance at most $3/2$. Therefore, by Lemma 8.15

$$\Pr_{X \sim G}(|p(X) - \mathbb{E}[p(X)]| \geq T) \leq 2e^{1/3 - 2T/3 \operatorname{Var}[p(X)]} \leq \exp((T - 1)/3)) .$$

However, from equation (40), $\mathbb{E}[p(X)] - \mu \leq 3$ and so

$$\Pr_{X \sim G}(|p(X) - \mu| > T) \leq 2 \exp(-(T - 4)/3).$$

This completes our proof. □

Acknowledgements

References

[ABL14] P. Awasthi, M. F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. In *STOC 2014*, pages 449–458, 2014.

[ADLS15] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. *CoRR*, abs/1506.00671, 2015.

[AGHK13] A. Anandkumar, R. Ge, D. Hsu, and S. Kakade. A tensor spectral approach to learning mixed membership community models. In *COLT 2013*, pages 867–881, 2013.

[AGM12] S. Arora, R. Ge, and A. Moitra. Learning topic models - going beyond SVD. In *FOCS 2012*, pages 1–10, 2012.

[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(1):215–236, 2015.

[AHK12] A. Anandkumar, D. Hsu, and S. Kakade. A method of moments for mixture models and Hidden Markov Models. *Journal of Machine Learning Research - Proceedings Track*, 23:33.1–33.34, 2012.

[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)*, pages 1682–1686, 2014.

[AK01] S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd Symposium on Theory of Computing*, pages 247–257, 2001.

[BBBB72] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions*. Wiley, New York, 1972.

[BD15]   S. Balmand and A. Dalalyan. Convex programming approach to robust estimation of a multivariate gaussian model, 2015.

[Ber06]   T. Bernholt. Robust estimators are hard to compute. Technical report, University of Dortmund, Germany, 2006.

[Bru09]   S. C. Brubaker. Robust PCA and clustering in noisy mixtures. In *SODA 2009*, pages 1078–1087, 2009.

[BS10]   M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010.

[BV08]   S. C. Brubaker and S. Vempala. Isotropic PCA and Affine-Invariant Clustering. In *Proc. 49th IEEE Symposium on Foundations of Computer Science*, pages 551–560, 2008.

[CDSS13] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, pages 1380–1394, 2013.

[CDSS14a] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014.

[CDSS14b] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *NIPS*, pages 1844–1852, 2014.

[CGG02]   M. Cryan, L. Goldberg, and P. Goldberg. Evolutionary trees can be learned in polynomial time in the two state general Markov model. *SIAM Journal on Computing*, 31(2):375–397, 2002.

[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 STOC'16*, 2016.

[CLMW11] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.

[CW01]   A. Carbery and J. Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in R^n. *Mathematical Research Letters*, 8(3):233–248, 2001.

[Das99]   S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 634–644, 1999.

[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*, STOC '16, New York, NY, USA, 2016. ACM.

[DDO$^+$13] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R.A. Servedio, and L. Tan. Learning Sums of Independent Integer Random Variables. In *FOCS*, pages 217–226, 2013.

[DDS12a] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In *SODA*, pages 1371–1385, 2012.

[DDS12b] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In *STOC*, pages 709–728, 2012.

[DDS15a] C. Daskalakis, I. Diakonikolas, and R. A. Servedio. Learning poisson binomial distributions. *Algorithmica*, 72(1):316–357, 2015.

[DDS15b] A. De, I. Diakonikolas, and R. Servedio. Learning from satisfying assignments. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 478–497, 2015.

[DG85] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons, 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*, COLT '14, pages 1183–1213, 2014.

[DKS15] I. Diakonikolas, D. M. Kane, and A. Stewart. Optimal learning via the fourier transform for sums of independent integer random variables. *CoRR*, abs/1505.00662, 2015.

[DKS16] I. Diakonikolas, D. M. Kane, and A. Stewart. The fourier transform of poisson multinomial distributions and its algorithmic applications. In *Proceedings of STOC'16*, 2016.

[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*, FOCS '15, Washington, DC, USA, 2015. IEEE Computer Society.

[DL96] L. Devroye and G. Lugosi. A universally acceptable smoothing factor for kernel density estimates. *Annals of Statistics*, 24:2499–2512, 1996.

[DL97] L. Devroye and G. Lugosi. Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *Annals of Statistics*, 25:2626–2637, 1997.

[DL01] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics, Springer, 2001.

[FM99] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proceedings of the 12th Annual COLT*, pages 183–192, 1999.

[FOS05] J. Feldman, R. O'Donnell, and R. Servedio. Learning mixtures of product distributions over discrete domains. In *Proc. 46th Symposium on Foundations of Computer Science (FOCS)*, pages 501–510, 2005.

[GLS88] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*, volume 2. Springer, 1988.

[Hau92] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[HK13] D. Hsu and S. M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Innovations in Theoretical Computer Science, ITCS '13*, pages 11–20, 2013.

[HM13] M. Hardt and A. Moitra. Algorithms and hardness for robust subspace recovery. In *COLT 2013*, pages 354–375, 2013.

[HP15]   M. Hardt and E. Price. Sharp bounds for learning a mixture of two Gaussians. In *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, STOC '15, pages 753–760, New York, NY, USA, 2015. ACM.

[HR09]   P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley 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*. Wiley 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*, pages 221–233. University of California Press, 1967.

[JP78]   D. S. Johnson and F. P. Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6:93–107, 1978.

[KL93]   M. J. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[KLS09]   A. Klivans, P. Long, and R. Servedio. Learning halfspaces with malicious noise. *Proc. 17th Internat. Colloq. on Algorithms, Languages and Programming (ICALP)*, 2009.

[KMR+94]   M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proc. 26th STOC*, pages 273–282, 1994.

[KMV10]   A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *STOC*, pages 553–562, 2010.

[KSS94]   M. Kearns, R. Schapire, and L. Sellie. Toward Efficient Agnostic Learning. *Machine Learning*, 17(2/3):115–141, 1994.

[LM00]   B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[LMTZ12]   G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang. Robust computation of linear models, or how to find a needle in a haystack. *CoRR*, abs/1202.4044, 2012.

[LS15]   J. Li and L. Schmidt. A nearly optimal and agnostic algorithm for properly learning a mixture of k gaussians, for any constant k. *CoRR*, abs/1506.01367, 2015.

[LT15]   P. L. Loh and X. L. Tan. High-dimensional robust precision matrix estimation: Cellwise corruption under -contamination. Available at arXiv:1509.07229, 2015.

[MR05]   E. Mossel and S. Roch. Learning nonsingular phylogenies and Hidden Markov Models. In *Proceedings of the 37th Annual Symposium on Theory of Computing (STOC)*, 2005.

[MV10]   A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

[Sco92]   D.W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York, 1992.

[Ser01]   R. Servedio. Smooth boosting and learning with malicious noise. In *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory*, pages 473–489, 2001.

[Ser03]   R. Servedio. Smooth boosting and learning with malicious noise. *JMLR*, 4:633–648, 2003.

[Sil86]   B. W. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

[SOAJ14]   A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *NIPS*, pages 1395–1403, 2014.

[Tuk75]   J.W. Tukey. Mathematics and picturing of data. In *Proceedings of ICM*, volume 6, pages 523–531, 1975.

[Val85]   L. Valiant. Learning disjunctions of conjunctions. In *Proc. 9th IJCAI*, pages 560–566, 1985.

[Ver10]   R. Vershynin. Introduction to the non-asymptotic analysis of random matrices, 2010.

[VW02]   S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *FOCS*, pages 113–122, 2002.

[Whi82]   H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

[Yat85]   Y. G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics*, 13:768–774, 1985.

[ZL14]   T. Zhang and G. Lerman. A novel m-estimator for robust pca. *J. Mach. Learn. Res.*, 15(1):749–808, January 2014.

A   Deferred Proofs from Section 2

Proof of Lemma 2.15: We include the simple proof for completeness. By Kraft's inequality, 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

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

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

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

Therefore,

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

where the last inequality follows from the union bound.   □

B Deferred Proofs from Section 4

This section contains deferred proofs of several concentration inequalities.

Proof of Lemma 4.5: Recall that for any $J \subseteq [m]$, we let $w^J \in \mathbb{R}^m$ 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

$$\Pr \left[\exists J : |J| = (1 - 2\varepsilon)N, \text{ and } \left\| \sum_{i=1}^N w_i^J Y_i Y_i^T - (1 - 2\varepsilon)I \right\|_2 \geq \delta_1 \right] \leq \tau .$$

For any fixed w^J we have

$$\sum_{i=1}^N w_i^J Y_i Y_i^T - I = \frac{1}{(1 - 2\varepsilon)N} \sum_{i \in J} Y_i Y_i^T - I$$

$$= \frac{1}{(1 - 2\varepsilon)N} \sum_{i=1}^N Y_i Y_i^T - \frac{1}{1 - 2\varepsilon}I - \left(\frac{1}{(1 - 2\varepsilon)N} \sum_{i \notin J} Y_i Y_i^T - \left(\frac{1}{1 - 2\varepsilon} - 1 \right) I \right) .$$

Therefore, by the triangle inequality, we have

$$\left\| \sum_{i=1}^N w_i^J Y_i Y_i^T - (1 - 2\varepsilon)I \right\|_2 \leq \left\| \frac{1}{(1 - 2\varepsilon)N} \sum_{i=1}^N Y_i Y_i^T - \frac{1}{1 - 2\varepsilon}I \right\|_2 + \left\| \frac{1}{(1 - 2\varepsilon)N} \sum_{i \notin J} Y_i Y_i^T - \left(\frac{1}{1 - 2\varepsilon} - 1 \right) I \right\|_2 .$$

By Lemma 4.3, the first term is at most δ_1 with probability $1 - \tau$ by setting

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

We now wish to show that with probability $1 - \tau$, there does not exist a J so that the second term is more than δ_1, for our choice of N. By multiplying both sides by $\gamma = (1 - 2\varepsilon)/\varepsilon$, this event is equivalent to the event that there is a J so that

$$\left\| \frac{1}{2\varepsilon N} \sum_{i \notin J} Y_i Y_i^T - I \right\|_2 > \gamma_1 \delta_1 .$$

By Lemma 4.3, for any fixed τ', this happens with probability at most τ' if we choose

$$N \geq c \frac{d + \log(1/\tau')}{\gamma \delta_1^2} , \tag{42}$$

for some universal constant c. We wish to union bound over all J of size $(1 - 2\varepsilon)N$. There are $\binom{N}{(1-2\varepsilon)N}$ such events. Thus, we take $\tau' = \binom{N}{(1-2\varepsilon)N}^{-1} \tau$. We now need to show that for our choice of N and τ', (42) is satisfied. By algebraic manipulations, and the fact that $\log \binom{n}{\alpha n} = O(nH(\alpha))$ for n sufficiently large, where $H(x)$ is the binary entropy function, the RHS of (42) simplifies to

$$c \frac{d + \log 1/\tau'}{\gamma \delta_1^2} = c \frac{d + \log 1/\tau + O(NH(2\varepsilon))}{\gamma \delta_1^2} .$$

100

Thus, (42) is satisfied if

$$N \left(1 - \frac{H(2\varepsilon)}{\gamma \delta_1^2}\right) \geq c \frac{d + \log 1/\tau}{\gamma \delta_1^2} .$$

By our choice of δ_1, $\left(1 - \frac{H(2\varepsilon)}{\gamma \delta_1^2}\right) \geq 1/2$, and hence, (42) is satisfied for our choice of m. □

Proof of Lemma 4.9: Let $U = V \Lambda V^T$ be the diagonalization of U. Then by the properties of trace, we have

$$\left| \frac{1}{N} \sum_{i=1}^{N} \mathrm{tr}(X_i X_i^T U) - \mathrm{tr}(U) \right| = \left| \frac{1}{N} \sum_{i=1}^{N} \mathrm{tr}(X_i X_i^T V \Lambda V^T) - \mathrm{tr}(\Lambda) \right|$$

$$= \left| \frac{1}{N} \sum_{i=1}^{N} \mathrm{tr}(X_i^T V \Lambda V^T X_i) - \mathrm{tr}(\Lambda) \right| .$$

Let $Y_i = (Y_{i1}, \ldots, Y_{id}) = V^T X_i$. Since V is an orthogonal matrix, and because a spherical Gaussian is rotationally invariant, we have $Y_i \sim \mathcal{N}(0, I)$. By computation, the lemma is equivalent to the statement that

$$\left| \sum_{i=1}^{N} \sum_{j=1}^{d} \frac{\lambda_j}{N} (Y_{ij}^2 - 1) \right| \leq \delta .$$

Since $Y_i \sim \mathcal{N}(0, I)$, we have that the $Y_{ij}^2 - 1$ are all independent random variables with $\|Y_{ij}^2 - 1\|_{\psi_1} \leq c$ for some universal constant c. Thus, we may invoke Bernstein's unequality (Prop 5.16), where our weight vector $a \in \mathbb{R}^{Nd}$ is given by $a_{ij} = \lambda_j/N$, and thus satisfies $\|a\|_\infty = \|U\|/N$ and

$$\|a\|_2^2 = \sum_{i=1}^{N} \sum_{j=1}^{d} \frac{\lambda_j^2}{N^2}$$

$$= \sum_{i=1}^{N} \frac{1}{N^2} = \frac{1}{N}$$

since by assumption $\sum_{i=1}^{d} \lambda_j^2 = \|U\|_F^2 = 1$. Thus, Bernstein's inequality gives us that

$$\Pr \left[\sum_{i=1}^{N} \sum_{j=1}^{d} \frac{\lambda_j}{N} (Y_{ij}^2 - 1) > \delta \right] \leq 2 \exp \left[-K \min \left(\frac{\delta^2}{c^2/N}, \frac{\delta}{c \|U\|/N} \right) \right]$$

$$\leq 2 \exp \left[-\Omega \left(\min \left(\delta^2 N, \delta N \right) \right) \right]$$

$$\leq \tau ,$$

for our choice of N, where the second line follows because the spectral norm of any matrix with Frobenius norm 1 is at most 1. □

Proof of Theorem 4.15: 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 a 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

$$
\begin{aligned}
\sum_{i,j=1}^{d} \lambda_i \lambda_j \langle B^T u_i, B^T u_j \rangle^2 &= \sum_{i,j} \lambda_i \lambda_j \langle (B^T u_i)^{\otimes 2}, (B^T u_j)^{\otimes 2} \rangle \\
&= \sum_{i,j=1}^{d} \lambda_i \lambda_j \langle (B^T \otimes B^T) u_i \otimes u_i, (B^T \otimes B^T) u_j \otimes u_j \rangle \\
&= \sum_{i,j=1}^{d} \lambda_i \lambda_j (u_i \otimes u_i) \Sigma^{\otimes 2} (u_j \otimes u_j) \\
&= v^T \Sigma^{\otimes 2} v \,.
\end{aligned}
$$

$\square$

B.1 Proof of Theorem 4.17

We will proceed in a couple of steps. First, we prove concentration of the empirical fourth moment, without worrying about removing points. We will use the following theorem:

Theorem B.1 (Theorem 5.44 of Vershynin). *Let $A_1, \ldots, A_N \in \mathbb{R}^d$ be independent random variables with identical second moment $M = \mathbb{E}[A_i A_i^T]$. Let R be so that $\|A_i\| \leq \sqrt{R}$ almost surely for all i. Then, for all $t \geq 0$, we have that with probability at least $1 - n \cdot \exp(-ct^2)$,*

$$\left\| \frac{1}{N} \sum_{i=1}^{N} A_i A_i^T - \Sigma \right\|_2 \leq \max\left(\|\Sigma\|_2^{1/2} \gamma, \gamma^2 \right) \quad \textit{where} \quad \gamma = t\sqrt{\frac{R}{d}} \ .$$

With some massaging, this allows us to show:

Lemma B.2. *Fix $\varepsilon, \tau > 0$. Let M_4 denote the canonical flattening of the fourth moment tensor of $\mathcal{N}(0, I)$. We have that with probability $1 - \tau$,*

$$\left\| \frac{1}{N} \sum_{i=1}^{N} Z_i Z_i^T - M_4 \right\|_2 \leq O\left(\sqrt{\frac{d^3 \log \frac{1}{\tau} \log \frac{N}{\tau \varepsilon}}{N}} \right) + \varepsilon \ .$$

In particular, if

$$N = \widetilde{\Omega}\left(\frac{d^3 \log^2 \frac{1}{\tau}}{\varepsilon^2} \right) \ ,$$

then $\left\| \frac{1}{N} \sum_{i=1}^{N} Z_i Z_i^T - M_4 \right\|_2 \leq O(\varepsilon)$.

We briefly remark that with somewhat more involved (but standard) methods, we believe we can strengthen this bound to

$$N = \widetilde{\Omega}\left(\frac{d^2 + \log^2 1/\tau}{\varepsilon^2} + \frac{d^3}{\varepsilon} \right) \ .$$

However, for simplicity of exposition we only present the weaker bound.

Proof. Let D be the distribution over $\mathbb{R}^{d^2}$ which is given by $X^{\otimes 2}$ where $X \sim \mathcal{N}(0, I)$ if $\|X\|_2^2 \leq O(d \log(N/(\tau \varepsilon)))$, and 0 otherwise. That is, a sample from D is drawn by sampling $X^{\otimes 2}$, and zeroing it out if its norm is too large. Let $I \subseteq [N]$ be the set of indices so that $\|X_i\|_2^2 \leq O(d \log \frac{N}{\tau \varepsilon})$. For each $i \in I$, we have that $X_i \sim D$. Let $M_4' = \mathbb{E}_{X \sim D}[(X^{\otimes 2})(X^{\otimes 2})^T]$. Since we are only removing large points, it is evident that $M_4' \preceq M_4$, and $\|M_4\| = O(d^2)$. By Theorem B.1, we have

$$\left\| \frac{1}{|I|} \sum_{i \in I} Z_i Z_i^T - M_4' \right\|_2 \leq \max\left(O(d) \gamma, \gamma^2 \right) \quad \text{where} \quad \gamma = O\left(\sqrt{\frac{d \log \frac{1}{\tau} \log \frac{m}{\tau \varepsilon}}{|I|}} \right) \ ,$$

except with probability $1 - O(\tau)$. By basic Gaussian concentration, we know that the probability that we throw away Z_i from this process is at most $O(\tau/m)$. Hence with probability $1 - O(\tau)$, we have that $I = [N]$. Union bounding over these two events and simplifying yields that

$$\left\| \frac{1}{N} \sum_{i=1}^{N} Z_i Z_i^T - M_4' \right\|_2 \leq O\left(\sqrt{\frac{d^3 \log \frac{1}{\tau} \log \frac{N}{\tau \varepsilon}}{N}} \right) \ .$$

It now suffices to show that $\|M_4 - M_4'\|_2 \leq O(\varepsilon)$. We have

$$
\begin{aligned}
\|M_4 - M_4'\|_2 &= \left\| \mathbb{E}[(X \otimes X)(X \otimes X)^T] - \mathbb{E}\left[(X \otimes X)(X \otimes X)^T \,\big|\, \|X\|_2^2 \leq O\left(d \log \frac{N}{\tau \varepsilon}\right)\right] \right\|_2 \\
&\leq \mathbb{E}\left[\|(X \otimes X)(X \otimes X)^T\|_2 \mathbb{1}_{\{\|X\|_2^2 \geq \Omega(d \log \frac{N}{\tau \varepsilon})\}} \right] \\
&= \mathbb{E}\left[\|X\|_2^4 \mathbb{1}_{\{\|X\|_2^2 \geq \Omega(d \log \frac{N}{\tau \varepsilon})\}} \right] \\
&\leq \int_{\Omega\left(d \log \frac{N}{\tau \varepsilon}\right)}^{\infty} t^2 \Pr[\|X\|_2^2 \geq t] dt \\
&\leq O(\varepsilon) \,,
\end{aligned}
$$

by basic Gaussian concentration. $\qquad\square$

We have now shown that spectral concentration holds when we don't allow any subselection of the points. To complete the proof of the theorem, we now need to use concentration of low-degree polynomials to show that this still holds when we allow ourselves to remove points.

Proof of Theorem 4.17. We will first prove a couple of key claims. First, we shall show that there cannot be too many i so that $p^2(X_i)$ is large:

Claim B.3. *Fix $\tau, \varepsilon > 0$. Let $N \geq \Omega\left(\log(1/\tau)/\varepsilon^2\right)$. There is a universal constant (independent of ε, τ) c' so that for all $c > c'$, we have that with probability $1 - O(\tau)$,*

$$
\frac{1}{N} \left| \{i : p^2(X_i) \geq c \log^2(1/\varepsilon)\} \right| \leq 2\varepsilon \,.
$$

Proof. Let $U_i = \mathbb{1}_{\{p^2(X_i) \geq c \log^2(1/\varepsilon)\}}$. Then, $\mathbb{E}[U_i] = \Pr_{X \sim \mathcal{N}(0,I)}[p^2(X) \geq c \log^2(1/\varepsilon)] \leq O(\varepsilon)$. The result then follows from a Hoeffding bound on $\frac{1}{N} \sum_{i=1}^{N} U_i$. $\qquad\square$

We now show that if we remove all points so that $p^2(X_i)$ is too large, then this cannot change the expectation by too much:

Claim B.4. *Let $c > 0$ be sufficiently large. Fix $\tau, \varepsilon > 0$, and let $N \geq \Omega\left(\log(1/\tau)/\varepsilon^2\right)$. Let $A_i = p^2(X_i)$ if $p^2(X_i) \leq c \log^2(1/\varepsilon)$, and 0 otherwise. Let $\alpha = \mathbb{E}_{X \sim \mathcal{N}(0,I)}[p^2(X)]$. Then*

$$
\left| \frac{1}{N} \sum_{i=1}^{N} A_i - \alpha \right| \leq O(c \cdot \varepsilon \log^2 1/\varepsilon) \,.
$$

Proof. Let $X \sim \mathcal{N}(0, I)$, and let $A = p^2(X)$ if $p^2(X) \leq c \log^2 1/\varepsilon$, and 0 otherwise. Let A be distributed as D. Then $A_i \sim D$ for all i. In particular, each A_i is a bounded random variable. Hence, by Hoeffding's inequality, by our choice of N, we have

$$
\left| \frac{1}{N} \sum_{i=1}^{N} A_i - \alpha' \right| \leq O(\varepsilon) \,,
$$

where $\alpha' = \mathbb{E}_{A \sim D}[p^2(A)]$. It thus suffices to show that $|\alpha' - \alpha| < O(c \cdot \varepsilon \log^2 1/\varepsilon)$. We have

$$
\begin{aligned}
|\alpha' - \alpha| &= \mathbb{E}_{X \sim \mathcal{N}(0,I)} \left[p^2(X) 1_{\{p^2(X) \geq c \log^2 1/\varepsilon\}} \right] \\
&= \int_{\sqrt{c} \log 1/\varepsilon}^{\infty} t^2 \Pr[p(X) \geq t] dt \\
&\leq \int_{\sqrt{c} \log 1/\varepsilon}^{\infty} t^2 \exp(-\Omega(t)) dt \\
&\leq O(c \cdot \varepsilon \log^2 1/\varepsilon) .
\end{aligned}
$$

$\square$

Let us now prove the theorem. We first condition on the event that for all homogenous degree 2 polynomials p with $\|p\|_2 = 1$, we have

$$
\left| \frac{1}{N} \sum_{i=1}^{N} p^2(X_i) - 1 \right| \leq O(\varepsilon) . \tag{43}
$$

We observe that every $v = M^\flat \in \mathcal{S}$ corresponds to a homogenous degree two polynomial $p(X) = X^T M X$, and moreover, $\|p\|_2 = 2\|M\|_F$. Then Lemma 7 asserts that by our choice of N, (43) holds with probability $1 - O(\tau)$. Moreover, this implies that to show the theorem, it suffices to show that for all homogenous degree two polynomials p with $\|p\|_2 = 1$, we have

$$
\left| \sum_{i=1}^{N} w_i p^2(X_i) - 1 \right| \leq O(\varepsilon \log^2(1/\varepsilon)) ,
$$

for all $w \in S_{m,\varepsilon}$.

Moreover, let $\mathcal{C}$ be a $(1/3)$-net for the set of all homogenous degree 2 polynomials with $\|p\|_2 = 1$. It is well-known that we can take $|\mathcal{C}| = 2^{O(d^2)}$. Let us also condition on the event that for all $p \in \mathcal{C}$, we have

$$
\frac{1}{N} \left| \{ i : p^2(X_i) \geq c \log^2(1/\varepsilon) \} \right| \leq 2\varepsilon, \text{ and} \tag{44}
$$

$$
\left| \frac{1}{N} \sum_{i=1}^{m} A_i - 1 \right| \leq O(\varepsilon \log^2 1/\varepsilon) , \tag{45}
$$

where A_i is defined as in Claim B.4 depends on p. By claims B.3 and B.4, by our choice of N, for any fixed p, these hold with probability $1 - O(\tau')$, where $\tau' = O(\tau/|\mathcal{C}|)$, so that by a union bound this event occurs with probability $1 - O(\tau)$.

Fix any $p \in \mathcal{C}$. Let J^* be the set of $2\varepsilon N$ indices with largest $(p^2(X_i) - 1)$, and let $J_1^* = \{ i : p^2(X_i) \geq c \log^2(1/\varepsilon) \}$. Observe that $J_1^* \subseteq J^*$, by (44). Moreover, by (45), we have

$$
\left| \frac{1}{N} \sum_{i \notin J_1^*} (p^2(X_i) - 1) \right| \leq O(\varepsilon \log^2(1/\varepsilon)) . \tag{46}
$$

Fix any $I \subseteq [m]$ so that $|I| = (1 - 2\varepsilon)N$. Partition $[N] \setminus I$ into two sets $J^+ \cup J^-$ where $J^+ = \{ i \notin I :$

$p^2(X_i) \geq 1\}$, and $J^- = \{i \notin I : p^2(X_i) < 1\}$. We then have

$$\frac{1}{|I|}\sum_{i\in I}(p^2(X_i) - 1) \leq \frac{1}{|I|}\sum_{i\in I\cup J^+}(p^2(X_i) - 1) - \frac{1}{|I|}\sum_{i\in J^-}(p^2(X_i) - 1)$$

$$\leq \left|\frac{1}{|I|}\sum_{i=1}^{N}p^2(X_i) - 1\right| + 2\frac{1}{|I|}\left|\sum_{i\in J^-}(p^2(X_i) - 1)\right|$$

$$\leq O(\varepsilon) + \frac{|J^-|}{|I|}$$

$$\leq O(\varepsilon) . \tag{47}$$

where the penultimate line follows from (43) and since $p^2(X_i) \geq 0$ and hence $|p^2(X_i) - 1| \leq 1$ for all $i \in J^-$. Moreover, we have that

$$\frac{1}{|I|}\sum_{i\in I}(p^2(X_i) - 1) \geq \frac{1}{(1 - 2\varepsilon)N}\sum_{i\notin J_1^*}(p^2(X_i) - 1)$$

$$\geq (1 - O(\varepsilon))\frac{1}{N}\sum_{i\notin J_1^*}(p^2(X_i) - 1)$$

$$\geq -O(\varepsilon \log^2 1/\varepsilon) ,$$

by (46) and hence this along with (47) implies that $\left|\frac{1}{|I|}\sum_{i\in I}(p^2(X_i) - 1)\right| \leq O(\varepsilon \log^2 1/\varepsilon)$. Since this holds for all I with $|I| = (1 - 2\varepsilon)N$, by convexity this implies that for all $w \in S_{N,\varepsilon}$, we have

$$\left|\sum_{i\in I}w_i(p^2(X_i) - 1)\right| \leq O(\varepsilon \log^2 1/\varepsilon) .$$

Finally, by the usual net argument, this implies that for all degree 2 p with $\|p\|_2 = 1$, we have

$$\left|\sum_{i\in I}w_i(p^2(X_i) - 1)\right| \leq O(\varepsilon \log^2 1/\varepsilon) ,$$

which by the logic above, suffices to prove the theorem. $\qquad\square$

C   Deferred Proofs from Section 5

Proof of Theorem 5.8: The first two properties follow directly from (25). We now show the third property. Suppose this does not happen, that is, there are j, j' so that $\ell = \ell(j) = \ell(j')$ so that $\|\mu_j - \mu_{j'}\|_2^2 \geq \Omega(dk \log k/\varepsilon)$. That means that by (25) there is some sequence of clusters $S_1, \ldots, S_t$ so 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 a $X_i \in S_1$ so that $\|X_i - \mu_1\|_2^2 \leq O(d \log k/\varepsilon)$ and an $X_{i'} \in S_t$ so that $\|X_{i'} - \mu_2\|_2^2 \leq O(d \log k/\varepsilon)$. But by (25), we know that each S_i contains an point $X_{i''}$ so 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 $i = 1, \ldots, t - 1$, and 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. $\quad\square$

Proof of Lemma 5.9: 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 (26), we have

$$\sum_{i=1}^{N} w_i (X_i - \mu)(X_i - \mu)^T \succeq \sum_{i \in G} w_i (X_i - \mu)(X_i - \mu)^T$$
$$\succeq w_g (I + Q) - f(k, \gamma, \delta_1) I$$
$$\succeq (1 - \varepsilon)(I + Q) - f(k, \gamma, \delta_1) I \ ,$$

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 an $d \times k$ 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 μ.

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 $\text{span}(Q) \subseteq \text{span}(U)$, we have

$$
\begin{aligned}
\|C\|_{T_k} \geq \text{Tr}\left(U^T C U\right) &= \text{Tr}\left(U^T \left(\sum_{i=1}^{N} w_i(X_i - \mu)(X_i - \mu)^T - (I + Q)\right) U\right) + \|Q\|_{T_k} \\
&= \text{Tr}\left(U^T(C - Q)U\right) + \sum_{j \in [k]} \gamma_j \\
&= v^T(C - Q)v + \sum_{i=1}^{k-1} u_i^T(C - Q)u_i + \sum_{j \in [k]} \gamma_j \\
&\geq ckh(k, \gamma, \delta) - (k - 1)\frac{c}{2}h(k, \gamma, \delta) + \sum_{j \in [k]} \gamma_j \\
&\geq \frac{c}{2}kh(k, \gamma, \delta) + \sum_{j \in [k]} \gamma_j ,
\end{aligned}
$$

as claimed.

$\square$

Proof of Lemma 5.10: By Fact 4.2 and (27) we have $\|\sum_{i=G} \frac{w_i}{w_g}X_i - \mu\| \leq k^{1/2}\delta_2$. 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

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

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,

$$
\left\|\sum_{i \in G} w_i(X_i - \mu)(X_i - \mu)^T - I\right\|_2 \leq \left\|\sum_{i \in G} w_i(X_i - \mu)(X_i - \mu)^T - w_gI\right\|_2 + w_b \leq f(k, \gamma, \delta_1) + w_b.
$$

where in the last inequality we have used Fact 4.2 and (26). 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.

$\square$

C.1 Proof of Theorem 5.12

Once more, let $\Delta = \mu - \hat{\mu}$ and expand the formula for M:

$$\sum_{i=1}^{N} w_i Y_i Y_i^T - I = \sum_{i=1}^{N} w_i (X_i - \mu + \Delta)(X_i - \mu + \Delta)^T - I$$

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

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

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 (26), this implies that $w^* \in \mathcal{C}_{f(k,\gamma,\delta_1)}$. By Corollary 5.11, $\|\Delta\|_2 \leq O(\varepsilon\sqrt{\log 1/\varepsilon})$.

$$\left\| \sum_{i=1}^{m} w_i^* (X_i - \mu)(X_i - \mu)^T - I - \Delta\Delta^T \right\|_{T_k}$$

$$\leq \left\| \sum_{i=1}^{N} w_i^* (X_i - \mu)(X_i - \mu)^T - I - \sum_{j \in [k]} \alpha_j (\mu_j - \mu)(\mu_j - \mu)^T \right\|_{T_k} + \left\| \sum_{j \in [k]} \alpha_i (\mu_j - \mu)(\mu_j - \mu)^T \right\|_{T_k} + \|\Delta\Delta^T\|_2$$

$$\leq kf(k,\gamma,\delta_1) + \sum_{j \in [k]} \gamma_j + O(\varepsilon^2 \log 1/\varepsilon)$$

$$< \sum_{j \in [k]} \gamma_j + \frac{ckh(k,\gamma,\delta)}{2} .$$

$\square$

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 into two cases. In the first case, $\|\Delta\|_2^2 \leq \frac{ckh(k,\gamma,\delta)}{10}$. By Lemma 5.9, we have that

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

By the triangle inequality,

$$\|M\|_{T_k} \geq \sum_{j \in [k]} \gamma_j + \frac{3ckh(k,\gamma,\delta)}{4} - \|\Delta\|_2^2 \geq \sum_{i \in [k]} \gamma_j + \frac{ckh(k,\gamma,\delta)}{2},$$

as desired.

In the other case, $\|\Delta\|_2^2 \geq \frac{ckh(k,\gamma,\delta)}{10}$. Recall that $Q = \sum_{j\in[k]} \alpha_j(\mu_j - \mu)(\mu_j - \mu)^T$ from (22). Write M as follows:

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

$$= \left(\sum_{i\in G} w_i(X_i - \mu)(X_i - \mu)^T - w_g I - w_g Q\right) + w_g Q + \sum_{i\in E} w_i(X_i - \mu)(X_i - \mu)^T - w_b I - \Delta\Delta^T$$

Now taking the Schatten top-k norm of M, we have

$$\left\|\left(\sum_{i\in G} w_i(X_i - \mu)(X_i - \mu)^T - w_g I - w_g Q\right) + w_g Q + \sum_{i\in E} w_i(X_i - \mu)(X_i - \mu)^T - w_b I - \Delta\Delta^T\right\|_{T_k}$$

$$\geq \left\|w_g Q + \sum_{i\in E} w_i(X_i - \mu)(X_i - \mu)^T\right\|_{T_k} - \left\|\sum_{i\in G} w_i(X_i - \mu)(X_i - \mu)^T - w_g I - w_g Q\right\|_{T_k} - \|w_b I\|_2 - \|\Delta\Delta^T\|_2$$

$$\geq \left\|w_g Q + \sum_{i\in E} w_i(X_i - \mu)(X_i - \mu)^T\right\|_{T_k} - kf(k,\gamma,\delta_1) - 4\varepsilon - \|\Delta\|_2^2$$

$$\geq \left(\sum_{j\in[k]} \gamma_j - 4\varepsilon k\gamma\right) + \left\|\sum_{i\in E} w_i(X_i - \mu)(X_i - \mu)^T\right\|_{T_k} - kf(k,\gamma,\delta_1) - 4\varepsilon - \|\Delta\|_2^2$$

$$\geq \sum_{j\in[k]} \gamma_j + \Omega\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right) \tag{48}$$

$$\geq \sum_{j\in[k]} \gamma_j + \frac{ckh(k,\gamma,\delta)}{2}.$$

The first inequality is the triangle inequality, the second is by (26) and Fact 4.2, the third is because the summed matrices are positive semidefinite, the fourth follows from Lemma 5.10, and the fifth holds for all c sufficiently large. $\qquad\square$

By construction, we have that $\ell(w) \geq 0$. It remains to show that $\ell(w^*) < 0$.

$$\left\|\frac{1}{|G|}\sum_{i\in G}(X_i - \hat\mu)(X_i - \hat\mu)^T - I\right\|_{T_k}$$

$$= \left\|\frac{1}{|G|}\sum_{i\in G}(X_i - \mu + \Delta)(X_i - \mu + \Delta)^T - I\right\|_{T_k}$$

$$\leq \left\|\frac{1}{|G|}\sum_{i\in G}(X_i - \mu)(X_i - \mu)^T - I - \sum_{j\in[k]}\alpha_j(\mu_j - \mu)(\mu_j - \mu)^T\right\|_{T_k}$$

$$+ \left\|\sum_{j\in[k]}\alpha_j(\mu_j - \mu)(\mu_j - m)^T\right\|_{T_k} + 2\|\Delta\|_2\left\|\frac{1}{|G|}\sum_{i\in G}(X_i - \mu)\right\|_{T_k} + \|\Delta\|_2^2$$

$$\leq kf(k,\gamma,\delta_1) + \sum_{j\in[k]}\gamma_j + 2k^{1/2}\delta_2\|\Delta\|_2 + \|\Delta\|_2^2$$

110

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 $\|\Delta\|_2^2 \leq \frac{ckh(k,\gamma,\delta)}{10}$, then

$$\ell(w^*) \leq \sum_{j \in [k]} \gamma_j + kh(k, \gamma, \delta) + \frac{2k\delta\sqrt{ch(k, \gamma, \delta)}}{\sqrt{10}} + \frac{ckh(k, \gamma, \delta)}{10} - \Lambda.$$

We wish to show that

$$\frac{2k\delta\sqrt{ch(k, \gamma, \delta)}}{\sqrt{10}} \leq \frac{ckh(k, \gamma, \delta)}{10},$$

or equivalently, that

$$\delta \leq \frac{\sqrt{ch(k, \gamma, \delta)}}{2\sqrt{10}}.$$

But this is true for c sufficiently large, as $\sqrt{h(k, \gamma, \delta)} \geq \sqrt{\delta}$. Therefore,

$$\ell(w^*) \leq \sum_{j \in [k]} \gamma_j + \frac{(c+5)kh(k, \gamma, \delta)}{5} - \Lambda \leq 0,$$

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 $\|\Delta\|_2^2 \geq \frac{ckh(k,\gamma,\delta)}{10}$. By (48), we know that

$$\Lambda \geq \sum_{j \in [k]} \gamma_j + \Omega\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right).$$

Then we know

$$\ell(w^*) \leq kf(k, \gamma, \delta) + 2k^{1/2}\delta\|\Delta\|_2 + \|\Delta\|_2^2 - \Omega\left(\frac{\|\Delta\|_2^2}{\varepsilon}\right).$$

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

$$\|\Delta\|_2 \geq O(\sqrt{h(k, \gamma, \delta)}) \geq O(\sqrt{k\delta^2}) = O(k^{1/2}\delta\varepsilon)$$

Therefore in this case as well, $\ell(w^*) < 0$, as desired.

D   Deferred Proofs from Section 6

Proof of Lemma 6.22: By Lemma 6.6 applied with $\varepsilon' := \varepsilon^{3/2}/10d$ in place of ε, since we have $\Omega(d^4 \log(1/\tau)/\varepsilon'^2)$ samples from P, with probability at least $1 - \tau$, the set S is such that for all affine functions L, it holds $|\Pr_{X \in_u S}(L(X) \geq 0) - \Pr_{X \sim P}(L(X) \geq 0)| \leq \varepsilon'/d$. 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 $\Pr_P(C_T) \ge 3/5$. 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 $|\Pr_S(C_T) - \Pr_P[C_T]| \le \varepsilon'/d$. Hence, we have that $\Pr_S(C_T) \ge 1/2$.

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

$$\left| \Pr_{X \in_u S}(L(X) > 0) - \Pr_{X \sim P}(L(X) > 0) \right| = \left| \frac{\Pr_{X \in_u S}(L_T(X) > 0)}{\Pr_{X \in_u S}(C_T)} - \frac{\Pr_{X \sim P}(L_T(X) > 0)}{\Pr_{X \sim P}(C_T)} \right|$$

$$= \frac{|\Pr_{X \in_u S}(L_T(X) > 0) \Pr_{X \sim P}(C_T) - \Pr_{X \sim P}(L_T(X) > 0) \Pr_{X \in_u S}(C_T)|}{\Pr_{X \in_u S}(C_T) \Pr_{X \sim P}(C_T)}$$

$$\le (10/3) \cdot \left(\Pr_{X \in_u S}(L_T(X) > 0) \cdot \left(\Pr_{X \sim P}(C_T) - \Pr_{X \in_u S}(C_T) \right) \right.$$

$$\left. - \Pr_{X \in_u S}(C_T) \left(\Pr_{X \sim P}(L_T(X) > 0) - \Pr_{X \in_u S}(L_T(X) > 0) \right) \right)$$

$$\le (10/3) \cdot 2\varepsilon'/d \le \varepsilon^{3/2}/d^2 .$$

This completes the proof of Lemma 6.22. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

E Deferred Proofs from Section 7

Proof of Lemma 7.6: Let $S_P \subseteq S$ be the set of samples drawn from P and $S_Q \subseteq S$ be the set of samples drawn from Q. Firstly, we note that by a Chernoff bound, $||S_P|/|S| - \alpha| \le O(\varepsilon/d^2)$ with probability $1 - \tau/3$. Assuming this holds, it follows that $|S_P| \ge (\alpha/2)|S| \ge (\varepsilon^{1/6}/2)|S| = \Omega(d^4 \log(1/\tau)/\varepsilon^2)$. Similarly, $|S_Q| \ge (1 - \alpha)|S|/2 \ge \Omega(d^4 \log(1/\tau)/\varepsilon^2)$. By Lemma 6.6 applied with $\varepsilon' := (c^2/4) \cdot \varepsilon$ in place of ε, since we have $\Omega((d^4 + d^2 \log(\tau))/\varepsilon'^2)$ samples, with probability $1 - \tau/3$, the set S_P is ε'-good for P, i.e., it satisfies that for all affine functions L, $|\Pr_{X \in_u S_P}(L(X) > 0) - \Pr_{X \sim P}(L(X) > 0)| \le \varepsilon'/d$. We show that assuming S is ε'-good, it is (ε, i) good for each $1 \le i \le d$.

Note that $\Pr_{X \sim P}[X_i = 1] = p_i \ge c$ and $\Pr_{X \sim P}[X_i = 0] = 1 - p_i \ge c$. Since $|\Pr_{X \sim P}[X_i = 1] - \Pr_{X \in_u S_P}[X_i = 1]| \le c^2 \varepsilon/(4d)$, it follows that $\Pr_{X \in_u S}[X_i = 1] \ge c/2$. 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:

$$\left| \Pr_{X \in_u S_P^1} (L(X) > 0) - \Pr_{X \sim P} (L(X) > 0 \mid X_i = 1) \right| =$$

$$= \left| \frac{\Pr_{X \in_u S_P^1} (L^1(X) > 0)}{\Pr_{X \in_u S_P^1} (X_i > 0)} - \frac{\Pr_{X \sim P} (L^1(X) > 0)}{p_i} \right|$$

$$\leq (2/c^2) \left(\Pr_{X \in_u S_P^1} (L^1(X) > 0) \, p_i - \Pr_{X \sim P} (L^1(X) > 0) \Pr_{X \in_u S_1} (X_i > 0) \right)$$

$$\leq (2/c^2) \left(p_i \left(\Pr_{X \in_u S_P^1} (L^1(X) > 0) - \Pr_{X \sim P} (L^1(X) > 0) \right) - \Pr_{X \sim P} (L^1(X) > 0) \left(\Pr_{X \in_u S_P^1} (X_i > 0) - p_i \right) \right)$$

$$\leq 2/c^2 \cdot 2\varepsilon'/d \leq \varepsilon/d \, .$$

Similarly, we obtain that

$$\left| \Pr_{X \in_u S_1} (L(X) > 0) - \Pr_{X \sim \Pi} (L(X) > 0) \right| \leq \varepsilon/d \, .$$

So, we have that S_P is (ε, i) good for P for all $1 \leq i \leq d$ with probability $1 - \tau/3$. Similarly, S_Q is (ε, i) good for Q for all $1 \leq i \leq d$ with probability $1 - \tau/3$. Thus, we have that $\left| |S_P|/|S| - \alpha \right| \leq \varepsilon/d^2$, S_P is (ε, i) good for P and S_Q is (ε, i) good for Q for all $1 \leq i \leq d$ with probability $1 - \tau$. That is, S is (ε, i)-good for Π for all $1 \leq i \leq d$ with probability at least $1 - \tau$. $\quad\square$

Proof of Lemma 7.19: Let $S_P \subseteq S$ be the set of samples drawn from P and $S_Q \subseteq S$ be the set of samples drawn from Q. Firstly, we note that by a Chernoff bound, $\left| |S_P|/|S| - \alpha \right| \leq O(\varepsilon/d^2)$ with probability at least $1 - \tau/3$. Assuming this holds, $|S_P| \geq (\alpha/2)|S| \geq \delta|S| = \Omega(d^4 \log(1/\tau)/\varepsilon^2)$. Similarly, $|S_Q| \geq (1-\alpha)|S|/2 \geq \Omega(d^4 \log(1/\tau)/\varepsilon^2)$.

By Lemma 6.6 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 ε-good for P. Similarly, with probability at least $1 - \tau/3$, the set S_Q is ε-good for Q. Thus, with probability $1 - \tau$, we have that $\left| \frac{|S_P|}{|S|} - \alpha \right| \leq \varepsilon$ and that S_P and S_Q are ε-good for P and Q respectively. $\quad\square$

F Deferred Proofs from Section 8

Proof of Lemma 8.3: By Fact 4.8, the first property holds with probability at least $1 - \tau/3$. After translating by μ^G, we note that (iii) and (iv) follow immediately from Lemmas 4.3 and 4.4 with probability $1 - \tau/3$. It remans to show that, conditioned on (i), (ii) holds with probability at least $1 - \tau/3$.

To simplify some expressions, let $\delta := \varepsilon/(d \log(d/\varepsilon\tau))$ and $R = C\sqrt{d \log(|S|/\tau)}$.

We need to show that, if f is the indicator function of any halfspace, with probability at least $1 - \tau/4$ it holds

$$\left| \mathop{\mathbb{E}}_{X \sim G}[f(X)] - \mathop{\mathbb{E}}_{x \in_u S}[f(X)] \right| \leq \delta.$$

In fact, we will show that with probability at least $1 - \tau/8$ we have

$$\mathop{\mathbb{E}}_{x \in_u S}[f(X)] \leq \mathop{\mathbb{E}}_{X \sim G}[f(X)] + \delta \, ,$$

as the other direction will hold with a similar probability.

In order to do this, we will demonstrate a finite collection $\mathcal{C}$ of indicators of halfspaces so that for any halfspace indicator f, there is a $g \in \mathcal{C}$ so that

- $g \geq f$ on $B(\mu^G, R)$, and

- $\mathbb{E}_{X \sim G}[g(X)] \leq \mathbb{E}_{X \sim G}[f(X)] + \delta/2$.

We note that given such a $\mathcal{C}$, by an application of the Chernoff and union bounds, we have that so long as $|S| \gg \log(|\mathcal{C}|)/\delta^2$ it holds

$$\left| \mathbb{E}_{X \sim G}[g(X)] - \mathbb{E}_{x \in_u S}[g(X)] \right| < \delta/2$$

for all $g \in \mathcal{C}$. This would imply that for all f in question, if $g \in \mathcal{C}$ is as guaranteed above, we have

$$\begin{aligned} \mathbb{E}_{x \in_u S}[f(X)] &\leq \mathbb{E}_{x \in_u S}[g(X)] \\ &\leq \mathbb{E}_{X \sim G}[g(X)] + \delta/2 \\ &\leq \mathbb{E}_{X \sim G}[f(X)] + \delta , \end{aligned}$$

where the first line above holds since S is supported on $B(\mu^G, R)$.

It remains to construct $\mathcal{C}$. We claim that it suffices to take $\mathcal{C}$ be the the set of indicators of halfspaces of the form $v \cdot (x - \mu^G) \geq T$, where v is taken from a $\delta/(CR)$-cover of the unit sphere and T is a multiple of δ/C with $|T| \leq 2R$. By standard results, v can be taken from a set of size $O(\delta/R)^d$ and T from a set of size $O(R/\delta)$, thus $|\mathcal{C}| = O(\delta/R)^{d+1}$. Hence, condition (ii) will be satisfied with appropriately high probability so long as $|S| \gg d \log(\delta/R)/\varepsilon^2$, which is sufficient for our purposes.

What remains to show is that the $\mathcal{C}$ described above is sufficient. Consider f, the indicator of the halfspace $w \cdot (x - \mu^G) \geq T$ for some unit vector w. If $|T| \geq R$, f is constant on $B(\mu^G, R)$, and replacing T by $\mathrm{sgn}(T)R$ does not change f on $B(\mu^G, R)$ and changes the expectation of f by a negligible amount. Therefore, we can assume that $|T| \leq R$. Let v be a unit vector with $\|v - w\|_2 \leq (\delta/(CR))$ and $T' \in [T - 2\delta/C, T - \delta/C]$ be a multiple of δ/C. We claim that the $g \in \mathcal{C}$ given by the halfspace $v \cdot (x - \mu^G) \geq T'$ suffices.

To verify the first property, note that if $w \cdot (x - \mu^G) \geq T$, then

$$v \cdot (x - \mu^G) = w \cdot (x - \mu^G) + (v - w) \cdot (x - \mu^G) \geq T - R(\delta/(CR)) \geq T'.$$

For the second property, we note that the expectation of f is an error function of T, while the expectation of g is an error function of T'. Since the error function is Lipschitz, the difference is $O(T - T') = O(\delta/C) < \delta/2$ for C sufficiently large. This completes the proof. $\square$

Proof of Lemma 8.15: Note that an even polynomial has no degree-1 terms. Thus, we may write $p(x) = \sum_i p_{i,i} x_i^2 + \sum_{i > j} p_{i,j} x_i x_j + p_o$. 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 that $p(x) = x^T P_2' x + p_0$. 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 Λ is diagonal be an eigen-decomposition 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 χ^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

$$\mathrm{Var}[p(X)] = \mathrm{Var}\left[\sum_i \lambda_i Y_i^2 + p_0 \right] = \sum_i \lambda_i^2 = \|P_F\|_2 .$$

114

Lemma F.1 (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:

$$\Pr\left(\sum_i \lambda_i(Y_i^2 - 1) > 2\sqrt{(\sum_i \lambda_i^2)t} + 2(\max_i \lambda_i)t\right) \leq e^{-t} .$$

Noting that $\operatorname{tr}(P_2) = \sum_i \lambda_i, \sum_i \lambda_i^2 = \|P_2\|_F$ 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 for $-p(x)$ instead of $p(x)$ and putting these together, 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 $2\|P_2\|_F^2 = \operatorname{Var}_{X \sim G}(p(X))$ gives:

$$\Pr(|p(X) - \mathbb{E}_{X \sim G}[p(X)]| \geq T) \leq 2e^{1/3 - 2T/3 \operatorname{Var}_{X \sim G}[p(X)]} .$$

The final property is a consequence of the following anti-concentration inequality:

Theorem F.2 ([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. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Proof of Lemma 8.16: Note that we can reduce this to the case where $\Sigma = I$. We can consider the multiset T of y with $y = \Sigma^{-1/2}x$ for the corresponding $x \in S$. Then, T is a set of independent samples from $\mathcal{N}(0, I)$ and note that if it is (ε, τ)good for $\mathcal{N}(0, I)$, then S is (ε, τ)-good for G. Thus, we assume that $\Sigma = I$.

By Fact 4.8, the first property holds with probability at least $1 - \tau/4$.

Let M_S be the empirical second moment matrix of S, i.e., the $d \times d$ matrix with $(M_S)_{i,j} = \mathbb{E}_{X \in_u S}[X_i X_j]$. Then, by Lemma 4.3, with probability at least $1 - \tau/4$, $\|M_S - I\|_2 \leq \varepsilon/4d^{3/2}$.

Let T_S be $d^2 \times d^2$ matrix that is given by the canonical flattening of the empirical fourth moment tensor of S. Then, by B.2, we have that with probability at least $1 - \tau/4$, $\|T_S - M_4\|_2 \leq \varepsilon/2$, where M_4 is the flattened fourth moment tensor of $G = \mathcal{N}0, I$.

We claim that these bounds on M_S and T_S imply the the conditions about the expectation of even degree-2 polynomials. Using Lemma 8.15, we can write $p(x) = x^T P_2 x + p_0$ for a symmetric matrix $P_2 \in \mathbb{R}^{d \times d}$ and $p_0 \in R$. Recall that we write $P_2^\flat \in \mathbb{R}^{d^2}$ for the flattened out version of P_2. Then, in terms of these, we can write

$$\mathbb{E}_{X \in_u S}[p(X)] = p_0 + \operatorname{tr}(P_2^T M_S) .$$

Lemma 8.15, gives that $\mathbb{E}_{X \sim G}[p(X)] = p_0 + \operatorname{tr}(P_2)$. Thus, we have

$$| \underset{X \in_u S}{\mathbb{E}}[p(X)] - \underset{X \sim G}{\mathbb{E}}[p(X)]| = |\operatorname{tr}(P_2^T(M_S - I))| \leq \|P_2\|_F \|M_S - I\|_F$$

$$\leq \sqrt{\underset{X \sim G}{\operatorname{Var}}[p(X)]/2} \cdot d\|M_S - I\|_2 \qquad \text{(using Lemma 8.15)}$$

$$\leq \varepsilon \sqrt{\underset{X \sim G}{\mathbb{E}}[p(X)^2]} \; .$$

Again, in terms of P_2 and p_0, we can write:

$$\underset{X \in_u S}{\mathbb{E}}[p(X)^2] = p_0^2 + 2p_0 \operatorname{tr}(P_2^T M_S) + P_2^{\flat T} T_S P_2^\flat \; ,$$

and

$$\underset{X \sim G}{\mathbb{E}}[p(X)^2] = p_0^2 + 2p_0 \operatorname{tr}(P_2) + P_2^{\flat T} M_4 P_2^\flat \; .$$

We can take norms of the difference between these equations, giving

$$| \underset{X \in_u S}{\mathbb{E}}[p(X)^2] - \underset{X \sim G}{\mathbb{E}}[p(X)^2]| \leq 2p_0 \|P_2\|_F \|M_S - I\|_F + \|P_2\|_F^2 \|T_S - M_4\|_2 \; .$$

We need to express the RHS in terms of $\mathbb{E}_{X \sim G}[p(X)^2] = \operatorname{Var}_{X \sim G}[p(X)] + \mathbb{E}_{X \sim G}[p(X)]^2$. Using Lemma 8.15, $\|P_2\|_F^2 = \operatorname{Var}_{X \sim G}[p(X)]/2 \leq \mathbb{E}_{X \sim G}[p(X)^2]/2$ and $|p_0| = |\mathbb{E}_{X \sim G}[p(X)] + \operatorname{tr}(P_2)| \leq \sqrt{\mathbb{E}_{X \sim G}[p(X)^2]} + \sqrt{d}\|P_2\|_F \leq \sqrt{d}\sqrt{\mathbb{E}_{X \sim G}[p(X)^2]}$. Thus, we have

$$| \underset{X \in_u S}{\mathbb{E}}[p(X)^2] - \underset{X \sim G}{\mathbb{E}}[p(X)^2]| \leq 2p_0 \|P_2\|_F \|M_S - I\|_F + \|P_2\|_F^2 \|T_S - M_4\|_2$$

$$\leq (2\sqrt{d}\|M_S - I\|_F + \|T_S - M_4\|_2) \underset{X \sim G}{\mathbb{E}}[p(X)^2]$$

$$\leq (2d^{3/2}\|M_S - I\|_2 + \|T_S - M_4\|_2) \underset{X \sim G}{\mathbb{E}}[p(X)^2]$$

$$\leq \varepsilon \underset{X \sim G}{\mathbb{E}}[p(X)^2] \; .$$

For the last property, we need a finite set of degree-2 polynomials so that we can take a union bound.

Claim F.3. *For any $\gamma > 0$, there exists a set $\mathcal{C}$ of degree-2 polynomials on $\mathbb{R}^d$ of size $(\gamma/d \log(N/\tau)(d^2 + 1))^{d^2+2}$, such that for every degree-2 polynomial $p(x)$ on $\mathbb{R}^d$ such no coefficient of $p(x)$ has absolute value bigger than 1, except possibly for the constant term which has absolute value at most $O(d \log(N/\tau))$, there is a $q(x) \in \mathcal{C}$ such that for all x with $\|x\|_2^2 = O(d \log(N/\tau))$, $|p(x) - q(x)| \leq \gamma$.*

Proof. Note that any degree 2 polynomial on $\mathbb{R}^d$ has at most $d^2 + 1$ coefficients.

We take $\mathcal{C}$ to be the set of polynomials with coefficients that are multiples of $\gamma/\log(N/\tau)d(d^2 + 1)$ with absolute value no bigger than 1, and constant term a multiple of $\gamma/\log(N/\tau)d(d^2 + 1)$ with absolute value no bigger than $O(d \log(N/\tau))$. Then, $|\mathcal{C}| \leq (\gamma/d \log(N/\tau)(d^2 + 1))^{d^2+2}$. Now, given $p(x)$, we take $q(x) \in \mathcal{C}$ to be the polynomial whose coefficients are the corresponding coefficients of $p(x)$ rounded towards zero to a multiple of $\gamma/\log(N/\tau)d(d^2 + 1)$. When $x^T \Sigma^{-1} x = O(d \log(N/\tau))$, the absolute value of any monomial of degree at most 2 is bounded from above by $\|x\|_\infty^2 \leq O(d \log(N/\tau))$. Thus, the absolute value of the error between corresponding terms of $p(x)$ and $q(x)$ is at most $\gamma/(d^2 + 1)$ and so $|p(x) - q(x)| \leq \gamma$. $\square$

Let $\delta = \varepsilon/C(d \log(N/\tau))^2$ for some sufficiently high constant C. We apply the above with $\gamma = \delta^2$. Let $\mathcal{C}'$ be the set containing $q(x)$, $q(x) - \delta^2$ and $q(x) + \delta^2$ for all $q(x) \in \mathcal{C}$. Then $|\mathcal{C}'| \leq \operatorname{poly}((d \log(N/\tau))/\varepsilon)^{d^2}$.

By Chernoff and union bound for $N \geq \Omega((d \log(d/\varepsilon\tau))^6/\varepsilon^2)$, then with probability at least $1 - \tau/4$, for all $q \in \mathcal{C}'$ that

$$\left| \underset{X \in_u S}{\Pr}(q(X) \geq 0) - \underset{X \sim G}{\Pr}(q(X) \geq 0) \right| < \varepsilon/2(d \log(N/\tau))^2 \; .$$

We now ahow that if for all $x \in S$ we have that $x^T \Sigma^{-1} x = O(d \log(N/\tau))$, this implies the last property.

Note that if $p(x)$ satifies the three claims then so does $\alpha p(x)$ for any $\alpha > 0$. Thus, we may assume without loss of generality that $\text{Var}_{X \sim G} \, p(x) = 1$. Note that if $p_0 = 0$ and $x \in S$, then $|p(x)| \leq x^T P_2 x \leq \|x\|_2^2 \|P_2\|_2 \leq O(d \log(N/\tau)) \cdot \|P_2\|_F = O(d \log(N/\tau))$, since by Lemma 8.15, $2\|P_2\|_F = \text{Var}_{X \sim G} \, p(x) = 1$. Thus, if $|p_0| \geq \Omega(d \log(N/\tau))$ then $\Pr_{X \in_u S}(p(X) \geq 0)$ is 0 or 1. In this case, the claim for the minimum value $|p_0|$ can take and have this probabiltiy 0 or 1 respectively is stronger. We can therefore assume that $|p_0| = O(d \log(N/\tau))$. Since $\|P_2\|_F = 1/2$, the absolute value of each other coefficient is less than 1.

Thus, there is a $q(x) \in \mathcal{C}$ with $|p(x) - q(x)| \leq \delta^2$ for all $x \in S$. Note that this $q(x)$ has $\text{Var}_{X \sim G}[q(x)] = \Theta(1)$ and thus, by anti-concentration bounds, $\Pr_{X \sim G}(|q(X)|^2 \leq \delta^2) \leq O(\delta)$. Also, $\Pr_{X \in G}[\|X\|_2 \geq \Omega(\sqrt{d \log(N/\tau)}] \leq \tau/N \leq \delta$. Hence, we have:

$$
\begin{aligned}
\Pr_{X \in_u S}(p(x) \geq 0) &\geq \Pr_{X \in_u S}(q(x) - \delta^2 \geq 0) \\
&\geq \Pr_{X \sim G}(q(X) - \delta^2 \geq 0) - C\delta/2 && \text{(since } q(x) - \delta^2 \in \mathcal{C}') \\
&\geq \Pr_{X \sim G}(q(X) + \delta^2 \geq 0) - (C/2 + O(1))\delta \\
&\geq \Pr_{X \sim G}(p(x) \geq 0) - (C/2 + O(1))\delta - \Pr_{X \in G}[\|X\|_2 \geq \Omega(\sqrt{d \log(N/\tau)}] \\
&\geq \Pr_{X \sim G}(p(x) \geq 0) - C\delta \ .
\end{aligned}
$$

Similarly $\Pr_{X \in_u S}(p(x) \geq 0) \leq \Pr_{X \sim G}(p(x) \geq 0) + C\delta$. Thus, we have

$$
\Pr_{X \in_u S}(p(X) \geq 0) \leq \Pr_{X \sim G}(p(X) \geq 0) + \varepsilon/(d \log(N/\tau))^2 \ ,
$$

which is stronger than the required bound.

By a union bound, all four of the assumptions we made that hold with probability at least $1 - \tau/4$, hold with probability at least $1 - \tau$.

$\square$