Testing Convex Truncation[*]

Anindya De
University of Pennsylvania

Shivam Nadimpalli
Columbia University

Rocco A. Servedio
Columbia University

May 8, 2023

Abstract

We study the basic statistical problem of testing whether normally distributed n-dimensional data has been *truncated*, i.e. altered by only retaining points that lie in some unknown truncation set $S \subseteq \mathbb{R}^n$. As our main algorithmic results,

1. We give a computationally efficient $O(n)$-sample algorithm that can distinguish the standard normal distribution $N(0, I_n)$ from $N(0, I_n)$ conditioned on an unknown and arbitrary *convex set S*.

2. We give a different computationally efficient $O(n)$-sample algorithm that can distinguish $N(0, I_n)$ from $N(0, I_n)$ conditioned on an unknown and arbitrary *mixture of symmetric convex sets*.

These results stand in sharp contrast with known results for learning or testing convex bodies with respect to the normal distribution or learning convex-truncated normal distributions, where state-of-the-art algorithms require essentially $n^{\sqrt{n}}$ samples. An easy argument shows that no finite number of samples suffices to distinguish $N(0, I_n)$ from an unknown and arbitrary mixture of general (not necessarily symmetric) convex sets, so no common generalization of results (1) and (2) above is possible.

We also prove that any algorithm (computationally efficient or otherwise) that can distinguish $N(0, I_n)$ from $N(0, I_n)$ conditioned on an unknown symmetric convex set must use $\Omega(n)$ samples. This shows that the sample complexity of each of our algorithms is optimal up to a constant factor.

---

1   Introduction

Understanding distributions which have been *truncated*, i.e. subjected to some type of conditioning, is one of the oldest and most intensively studied questions in probability and statistics. Research on truncated distributions goes back the work of Bernoulli [Ber60], Galton [Gal97], Pearson [Pea02], and other pioneers; we refer the reader to the introductions of [DGTZ18, KTZ19] for historical context, and to [Sch86, BC14, Coh16] for contemporary book-length studies of statistical truncation.

In recent years a nascent line of work [DKTZ21, FKT20, DGTZ19, DGTZ18] has considered various different learning and inference problems for truncated distributions from a modern theoretical computer science perspective (see Section 1.3 for a more detailed discussion of these works and how they relate to the results of this paper). The current paper studies an arguably more basic statistical problem than learning or inference, namely *distinguishing* between a null hypothesis (that there has been no truncation) and an alternative hypothesis (that some unknown truncation has taken place).

In more detail, we consider a high-dimensional version of the fundamental problem of determining whether given input data was drawn from a known underlying probability distribution $\mathcal{P}$, versus from $\mathcal{P}$ conditioned on some unknown *truncation set* S (we write $\mathcal{P}|_S$ to denote such a truncated distribution). In our work the known high-dimensional distribution $\mathcal{P}$ is the n-dimensional standard normal distribution $N(0, I_n)$, and we consider a very broad and natural class of possible truncations, corresponding to conditioning on an unknown *convex set* (and variations of this class).

As we discuss in detail in Section 1.3, the sample complexity and running time of known algorithms for a number of related problems, such as learning convex-truncated normal distributions [KTZ19], learning convex sets under the normal distribution [KOS08], and testing whether an unknown set is convex under the normal distribution [CFSS17], all scale exponentially in $\sqrt{n}$. In sharp contrast, all of our distinguishing algorithms have sample complexity *linear* in n and running time at most poly(n). Thus, our results can be seen as an exploration of one of the most fundamental questions in testing—namely, *can we test faster than we can learn?* What makes our work different is that we allow the algorithm only to have access to random samples, which is weaker than the more powerful query access that is standardly studied in the complexity theoretic literature on property testing. However, from the vantage point of statistics and machine learning, having only sample access is arguably more natural than allowing queries. Indeed, motivated by the work of Dicker [Dic14] in statistics, a number of recent results in computer science [KV18, CDS20, KBV20] have explored the distinction between *testing versus learning* from random samples, and our work is another instantiation of this broad theme. To complement our algorithmic upper bounds, we also give a number of information theoretic lower bounds on sample complexity, which in some cases nearly match our algorithmic results. We turn to a detailed discussion of our results below.

1.1   Our Results

We give algorithms and lower bounds for a range of problems on distinguishing the normal distribution from various types of convex truncations.

1.1.1   Efficient Algorithms

Our most basic algorithmic result is an algorithm for symmetric convex sets:

Theorem 1 (Symmetric convex truncations, informal statement)**.** There is an algorithm SYMM-CONVEX-DISTINGUISHER which uses $O(n/\varepsilon^2)$ samples, runs in poly($n, 1/\varepsilon$) time, and distinguishes

between the standard $N(0, I_n)$ distribution and any distribution $\mathcal{D} = N(0, I_n)|_S$ where $S \subset \mathbb{R}^n$ is any symmetric convex set with Gaussian volume at most[1] $1 - \varepsilon$.

The algorithm SYMM-CONVEX-DISTINGUISHER is quite simple: it estimates the expected squared length of a random draw from the distribution and checks whether this value is significantly smaller than it should be for the $N(0, I_n)$ distribution. (See Section 1.2 for a more thorough discussion of SYMM-CONVEX-DISTINGUISHER and the techniques underlying its analysis.) By extending the analysis of SYMM-CONVEX-DISTINGUISHER we are able to show that the same algorithm in fact succeeds for a broader class of truncations, namely truncation by any mixture of symmetric convex distributions:

Theorem 2 (Mixtures of symmetric convex truncations, informal statement)**.** The algorithm SYMM-CONVEX-DISTINGUISHER uses $O(n/\varepsilon^2)$ samples, runs in $\mathrm{poly}(n, 1/\varepsilon)$ time, and distinguishes between the standard $N(0, I_n)$ distribution and any distribution $\mathcal{D}$ which is a normal distribution conditioned on a mixture of symmetric convex sets such that $\mathrm{d_{TV}}(N(0, I_n), \mathcal{D}) \geq \varepsilon$ (where $\mathrm{d_{TV}}(\cdot, \cdot)$ denotes total variation distance).

It is not difficult to see that the algorithm SYMM-CONVEX-DISTINGUISHER, which only uses the empirical mean of the squared length of samples from the distribution, cannot succeed in distinguishing $N(0, I_n)$ from a truncation of $N(0, I_n)$ by a general (non-symmetric) convex set. To handle truncation by general convex sets, we develop a different algorithm which uses both the estimator of SYMM-CONVEX-DISTINGUISHER and also a second estimator corresponding to the squared length of the empirical mean of its input data points. We show that this algorithm succeeds for general convex sets:

Theorem 3 (General convex truncations, informal statement)**.** There is an algorithm CONVEX-DISTINGUISHER which uses $O(n/\varepsilon^2)$ samples, runs in $\mathrm{poly}(n, 1/\varepsilon)$ time, and distinguishes between the standard $N(0, I_n)$ distribution and any distribution $\mathcal{D} = N(0, I_n)|_S$ where $S \subset \mathbb{R}^n$ is any convex set such that $\mathrm{d_{TV}}(N(0, I_n), N(0, I_n)|_S) \geq \varepsilon$.

Given Theorem 2 and Theorem 3, it is natural to wonder about a common generalization to mixtures of general convex sets. However, an easy argument (which we sketch in Appendix A) shows that no finite sample complexity is sufficient for this distinguishing problem, so no such common generalization is possible.

1.1.2 An Information-Theoretic Lower Bound

We show that the sample complexity of both our algorithms CONVEX-DISTINGUISHER and SYMM-CONVEX-DISTINGUISHER are essentially the best possible, by giving an $\Omega(n)$-sample lower bound for any algorithm that successfully distinguishes $N(0, I_n)$ from $N(0, I_n)|_K$ where K is a (randomly-oriented) origin-centered hyperplane:

Theorem 4 (Lower bound, informal statement)**.** Any algorithm which distinguishes (with probability at least $9/10$) between the standard $N(0, I_n)$ distribution and $N(0, I_n)|_K$, where K is an unknown origin-centered hyperplane, must use $\Omega(n)$ samples.

Since an origin-centered hyperplane is a symmetric convex set of Gaussian volume zero, this immediately implies the optimality (up to constants) of the sample complexity of each of our algorithms.

---

[1]Note that a Gaussian volume upper bound on S is a necessary assumption, since the limiting case where the Gaussian volume of S equals 1 is the same as having no truncation.

1.2 Techniques

In this section, we give a technical overview of our upper and lower bounds, starting with the former.

Upper Bounds. To build intuition, let us first consider the case of a single symmetric convex body K. It can be shown, using symmetry and convexity of K, that draws from $N(0, I_n)|_K$ will on average lie closer to the origin than draws from $N(0, I_n)$, so it is natural to use this as the basis for a distinguisher. The proof of this relies on the background distribution being $N(0, I_n)$ in a crucial manner. We thus are led to consider our first estimator,

$$\mathbf{M} := \frac{1}{T} \sum_{i=1}^{T} \|\boldsymbol{x}^{(i)}\|^2, \tag{1}$$

where $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}$ are independent draws from the unknown distribution (which is either $N(0, I_n)$ or $N(0, I_n)_K$). We analyze this estimator using the notion of *convex influence* from the recent work [DNS22]. In particular, we use a version of Poincaré's inequality for convex influence to relate the mean of $\mathbf{M}$ to the Gaussian volume $\mathrm{Vol}(K)$ of the truncation set K, and combine this with the fact that the statistical distance between $N(0, I_n)$ and $N(0, I_n)|_K$ is precisely $1 - \mathrm{Vol}(K)$. With some additional technical work in the analysis, this same tester turns out to works even for conditioning on a mixture of symmetric convex sets rather than a single symmetric convex set.

The estimator described above will not succeed for general (non-symmetric) convex sets; for example, if K is a convex set that is "far from the origin," then $\mathbf{E}_{\boldsymbol{x} \sim N(0, I_n)|_K}[\|\boldsymbol{x}\|]$ can be larger than $\mathbf{E}_{\boldsymbol{x} \sim N(0, I_n)}[\|\boldsymbol{x}\|]$. However, if K is "far from the origin," then the center of mass of a sample of draws from $N(0, I_n)|_K$ should be "far from the origin," whereas the center of mass of a sample of draws from the standard normal distribution should be "close to the origin;" this suggests that a distinguisher based on estimating the center of mass should work for convex sets K that are far from the origin. The intuition behind our distinguisher for general convex sets is to trade off between the two cases that K is "far from the origin" versus "close to the origin." This is made precise via a case analysis based on whether or not the set K contains a "reasonably large" origin-centered ball.[2]

Lower Bound. As stated earlier, our lower bound is achieved for symmetric convex truncations in which the unknown truncation set $\mathbf{K}$ is an $(n-1)$-dimensional hyperplane which passes through the origin and is randomly oriented (i.e. the normal vector to $\mathbf{K}$ is a Haar random unit vector in $\mathbb{R}^n$).

Our first observation is that for a suitable choice of a limiting operation (so that conditioning on $\mathbf{K}$, which is a set of Gaussian volume 0, is well-defined), the distribution $N(0, I_n)|_{\mathbf{K}}$ corresponds to an $(n-1)$-dimensional standard Normal distribution supported on $\mathbf{K}$ (see Remark 28). The next observation is that since the direction of $\mathbf{K}$ is Haar random, "all of the information" in a sample of p many draws from $N(0, I_n)|_{\mathbf{K}}$ is contained in the $p \times p$ matrix of inner products between the draws, i.e. the Gram matrix of the sample (see Claim 29).

Since the Gram matrix of a collection of $p < m$ draws from $N(0, I_m)$ is a draw from the Wishart ensemble $\mathrm{Wis}(p, m)$, given the above observations it suffices to analyze the two Wishart ensembles

---

[2]Splitting into these two cases is reminiscent of the case split in the analysis of a weak learning algorithm for convex sets in [DS21], though the technical details of the analysis are quite different in our work versus [DS21]. In particular, [DS21] relies on a "density increment" result for sets with large inradius, whereas we do not use a density increment argument but instead make crucial use of an extension of the Brascamp-Lieb inequality due to Vempala [Vem10].

Wis(p, n) and Wis$(p, n-1)$. In more detail, in order to prove a cn lower bound on the number of samples that are necessary for distinguishing $N(0, I_n)$ from $N(0, I_n)|_{\mathbf{K}}$ (for some absolute constant $0 < c < 1$), it suffices to show that for $p = cn$, the total variation distance between the two distributions Wis(p, n) and Wis$(p, n-1)$ is bounded away from 1. We do this by directly analyzing the probability density function of the Wishart ensemble and using a central limit theorem, due to Jonsson [Jon82], for the log-determinant of a matrix drawn from the Wishart ensemble Wis(p, n) when $p = \Theta(n)$.

1.3 Related Work

As noted earlier in the introduction, this paper can be viewed in the context of a recent body of work [DKTZ21, FKT20, DGTZ19, DGTZ18, KTZ19] studying a range of statistical problems for truncated distributions from a theoretical computer science perspective. In particular, [DKTZ21] gives algorithms for non-parametric density estimation of sufficiently smooth multi-dimensional distributions in low dimension, while [FKT20] gives algorithms for parameter estimation of truncated product distributions over discrete domains, and [DGTZ19] gives algorithms for truncated linear regression.

The results in this line of research that are closest to our paper are those of [DGTZ18] and [KTZ19], both of which deal with truncated normal distributions (as does our work). [DGTZ18] considers the problem of inferring the parameters of an *unknown* high-dimensional normal distribution given access to samples from a *known* truncation set S, which is provided via access to an oracle for membership in S. Note that in contrast, in our work the high-dimensional normal distribution is known to be $N(0, I_n)$ but the truncation set is unknown, and we are interested only in detecting whether or not truncation has occurred rather than performing any kind of estimation or learning. Like [DGTZ18], the subsequent work of [KTZ19] considered the problem of estimating the parameters of an unknown high-dimensional normal distribution, but allowed for the truncation set S to also be unknown. They gave an estimation algorithm whose performance depends on the Gaussian surface area $\Gamma(S)$ of the truncation set S; when the set S is an unknown convex set in n dimensions, the sample complexity and running time of their algorithm is $n^{O(\sqrt{n})}$. In contrast, our algorithm for the distinguishing problem requires only $O(n)$ samples and poly(n) running time when S is an unknown n-dimensional convex set.

Other prior works which are related to ours are [KOS08] and [CFSS17], which dealt with Boolean function learning and property testing, respectively, of convex sets under the normal distribution. [KOS08] gave an $n^{O(\sqrt{n})}$-time and sample algorithm for (agnostically) learning an unknown convex set in $\mathbb{R}^n$ given access to labeled examples drawn from the standard normal distribution, and proved an essentially matching lower bound on sample complexity. [CFSS17] studied algorithms for testing whether an unknown set $S \subset \mathbb{R}^n$ is convex versus far from every convex set with respect to the normal distribution, given access to random labeled samples drawn from the standard normal distribution. [CFSS17] gave an $n^{O(\sqrt{n})}$-sample algorithm and proved a near-matching $2^{\Omega(\sqrt{n})}$ lower bound on sample-based testing algorithms.

We mention that our techniques are very different from those of [DGTZ18, KTZ19] and [KOS08, CFSS17]. [KOS08] is based on analyzing the Gaussian surface area and noise sensitivity of convex sets using Hermite analysis, while [CFSS17] uses a well-known connection between testing and learning [GGR98] to leverage the [KOS08] learning algorithm result for its testing algorithm, and analyzes a construction due to Nazarov [Naz03] for its lower bound. [DGTZ18] uses a projected stochastic gradient descent algorithm on the negative log-likelihood function of the samples together with other tools from convex optimization, while (roughly speaking) [KTZ19] combines elements from both [KOS08] and [DGTZ18] together with moment-based methods. In contrast, our approach

4

mainly uses ingredients from the geometry of Gaussian space, such as the Brascamp-Lieb inequality and its extensions due to Vempala [Vem10], and the already-mentioned "convex influence" notion of [DNS22].

Finally, we note that the basic distinguishing problem we consider is similar in spirit to a number of questions that have been studied in the field of property testing of probability distributions [Can20]. These are questions of the general form "given access to samples drawn from a distribution that is promised to satisfy thus-and-such property, is it the uniform distribution or far in variation distance from uniform?" Examples of works of this flavor include the work of Batu et al. [BKR04] on testing whether an unknown monotone or unimodal univariate distribution is uniform; the work of Daskalakis et al. [DDS+13] on testing whether an unknown k-modal distribution is uniform; the work of Rubinfeld and Servedio [RS09] on testing whether an unknown monotone high-dimensional distribution is uniform; and others. The problems we consider are roughly analogous to these, but where the unknown distribution is now promised to be normal conditioned on (say) a convex set, and the testing problem is whether it is actually the normal distribution (analogous to being actually the uniform distribution, in the works mentioned above) versus far from normal.

2 Preliminaries

In Section 2.1, we set up basic notation and background. We recall preliminaries from convex and log-concave geometry in Sections 2.2 and 2.3, and formally describe the classes of distributions we consider in Section 2.4.

2.1 Basic Notation and Background

Notation. We use boldfaced letters such as $\boldsymbol{x}, \boldsymbol{f}, \boldsymbol{A}$, etc. to denote random variables (which may be real-valued, vector-valued, function-valued, set-valued, etc.; the intended type will be clear from the context). We write "$\boldsymbol{x} \sim \mathcal{D}$" to indicate that the random variable $\boldsymbol{x}$ is distributed according to probability distribution $\mathcal{D}$. For $i \in [n]$, we will write $e_i \in \mathbb{R}^n$ to denote the ith standard basis vector.

Geometry. For $r > 0$, we write $S^{n-1}(r)$ to denote the origin-centered sphere of radius r in $\mathbb{R}^n$ and $\text{Ball}(r)$ to denote the origin-centered ball of radius r in $\mathbb{R}^n$, i.e.,

$$S^{n-1}(r) = \left\{x \in \mathbb{R}^n : \|x\| = r\right\} \quad \text{and} \quad \text{Ball}(r) = \left\{x \in \mathbb{R}^n : \|x\| \leq r\right\},$$

where $\|x\|$ denotes the ℓ_2-norm $\|\cdot\|_2$ of $x \in \mathbb{R}^n$. We also write S^{n-1} for the unit sphere $S^{n-1}(1)$.

Recall that a set $C \subseteq \mathbb{R}^n$ is convex if $x, y \in C$ implies $\alpha x + (1-\alpha)y \in C$ for all $\alpha \in [0,1]$. Recall that convex sets are Lebesgue measurable.

For sets $A, B \subseteq \mathbb{R}^n$, we write $A + B$ to denote the Minkowski sum $\{a + b : a \in A \text{ and } b \in B\}$. For a set $A \subseteq \mathbb{R}^n$ and $r > 0$ we write rA to denote the set $\{ra : a \in A\}$. Given a point $a \in \mathbb{R}^n$ and a set $B \subseteq \mathbb{R}^n$, we use $a + B$ and $B - a$ to denote $\{a\} + B$ and $B + \{-a\}$ for convenience.

Gaussians Distributions. We write $N(0, I_n)$ to denote the n-dimensional standard Gaussian distribution, and denote its density function by φ_n, i.e.

$$\varphi_n(x) = (2\pi)^{-n/2} e^{-\|x\|^2/2}.$$

When the dimension is clear from context, we may simply write φ instead of φ_n. We write $\Phi : \mathbb{R} \to [0,1]$ to denote the cumulative density function of the one-dimensional standard Gaussian

distribution, i.e.

$$\Phi(x) := \int_{-\infty}^{x} \varphi(y)\,dy.$$

We write $\mathrm{Vol}(K)$ to denote the Gaussian volume of a (Lebesgue measurable) set $K \subseteq \mathbb{R}^n$, that is

$$\mathrm{Vol}(K) := \Pr_{\boldsymbol{x} \sim N(0,I_n)}[\boldsymbol{x} \in K].$$

For a Lebesgue measurable set $K \subseteq \mathbb{R}^n$, we write $N(0,I_n)|_K$ to denote the standard Normal distribution conditioned on K, so the density function of $N(0,I_n)|_K$ is

$$\frac{1}{\mathrm{Vol}(K)} \cdot \varphi_n(x) \cdot K(x)$$

where we identify K with its 0/1-valued indicator function. Note that the total variation distance between $N(0,I_n)$ and $N(0,I_n)|_K$ is

$$\mathrm{d}_{\mathrm{TV}}\big(N(0,I_n)|_K, N(0,I_n)\big) = 1 - \mathrm{Vol}(K), \tag{2}$$

and so the total variation distance between $N(0,I_n)$ and $N(0,I_n)|_K$ is at least ε if and only if $\mathrm{Vol}(K) \leq 1 - \varepsilon$.

Mean Estimation in High Dimensions. We will also require the following celebrated result of Hopkins [Hop20] for computationally-efficient mean estimation in high-dimensions (extending an earlier result, due to [LM18], that had the same sample complexity but was not computationally efficient).

Proposition 5 (Theorem 1.2 of [Hop20]). For every $n, m \in \mathbb{N}$ and $\delta > 2^{-O(n)}$, there is an algorithm MEAN-ESTIMATOR which runs in time $O(nm) + \mathrm{poly}\big(n\log(1/\delta)\big)$ such that for every random variable $\boldsymbol{x}$ on $\mathbb{R}^n$, given i.i.d. copies $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ of $\boldsymbol{x}$, MEAN-ESTIMATOR$\big(\{\boldsymbol{x}^{(j)}\}, \delta\big)$ outputs a vector $\boldsymbol{L}$ such that

$$\Pr\left[\|\mu - \boldsymbol{L}\| > O\left(\sqrt{\frac{\mathrm{tr}(\Sigma)}{m}} + \sqrt{\frac{\|\Sigma\|\log(1/\delta)}{m}}\right)\right] \leq \delta$$

where $\mu := \mathbf{E}[\boldsymbol{x}]$ and $\Sigma := \mathbf{E}\left[(\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)^{\mathrm{T}}\right]$.

Distinguishing Distributions. We recall the basic fact that variation distance provides a lower bound on the sample complexity needed to distinguish two distributions from each other.

Fact 6 (Variation distance distinguishing lower bound). Let P, Q be two distributions over $\mathbb{R}^n$ and let A be any algorithm which is given access to independent samples that are either from P or from Q. If A determines correctly (with probability at least $9/10$) whether its samples are from P or from Q, then A must use at least $\Omega(1/\mathrm{d}_{\mathrm{TV}}(P,Q))$ many samples.

2.2 Convex Influences

In what follows, we will identify a set $K \subseteq \mathbb{R}^n$ with its 0/1-valued indicator function. The following notion of *convex influence* was introduced in [DNS21b, DNS22] as an analog of the well-studied notion of *influence of a variable on a Boolean function* (cf. Chapter 2 of [O'D14]). [DNS21b, DNS22] defined this notion only for symmetric convex sets; we define it below more generally for arbitrary (Lebesgue measurable) subsets of $\mathbb{R}^n$.

Definition 7 (Convex influence)**.** Given a Lebesgue measurable set $K \subseteq \mathbb{R}^n$ and a unit vector $v \in S^{n-1}$, we define the *convex influence of v on K*, written $\mathbf{Inf}_v[K]$, as

$$\mathbf{Inf}_v[K] := \operatorname*{\mathbf{E}}_{\boldsymbol{x} \sim N(0,I_n)} \left[K(\boldsymbol{x}) \left(\frac{1 - \langle v, \boldsymbol{x} \rangle^2}{\sqrt{2}} \right) \right].$$

Furthermore, we define the *total convex influence of K*, written $\mathbf{I}[K]$, as

$$\mathbf{I}[K] := \sum_{i=1}^{n} \mathbf{Inf}_{e_i}[K] = \operatorname*{\mathbf{E}}_{\boldsymbol{x} \sim N(0,I_n)} \left[K(\boldsymbol{x}) \left(\frac{n - \|\boldsymbol{x}\|^2}{\sqrt{2}} \right) \right].$$

In Proposition 20 of [DNS22] it is shown that the influence of a direction v captures the rate of change of the Gaussian measure of the set K under a dilation along v. Also note that that total convex influence of a set is invariant under rotations. The following is immediate from Definition 7.

Fact 8. For Lebesgue measurable $K \subseteq \mathbb{R}^n$, we have

$$\operatorname*{\mathbf{E}}_{\boldsymbol{x} \sim N(0,I_n)_K} \left[\boldsymbol{x}_i^2 \right] = 1 - \frac{\sqrt{2} \cdot \mathbf{Inf}_{e_i}[K]}{\mathrm{Vol}(K)}. \tag{3}$$

We also have that

$$\operatorname*{\mathbf{E}}_{\boldsymbol{x} \sim N(0,I_n)_K} \left[\|\boldsymbol{x}\|^2 \right] = n - \frac{\sqrt{2} \cdot \mathbf{I}[K]}{\mathrm{Vol}(K)}. \tag{4}$$

The following Poincaré-type inequality for convex influences was obtained as Proposition 23 in the full version of [DNS22] (available at [DNS21a]).

Proposition 9 (Poincaré for convex influences for symmetric convex sets)**.** For symmetric convex $K \subseteq \mathbb{R}^n$, we have

$$\frac{\mathbf{I}[K]}{\mathrm{Vol}(K)} \geq \Omega\big(1 - \mathrm{Vol}(K)\big).$$

The following variant of Proposition 9 for arbitrary convex sets (not necessarily symmetric) is implicit in the proof of Theorem 22 of [DNS22] (see Equation 16 of [DNS22]). Given a convex set $K \subseteq \mathbb{R}^n$, we denote its inradius by $r_{\mathrm{in}}(K)$, i.e.

$$r_{\mathrm{in}}(K) := \max \big\{ r : \mathrm{Ball}(r) \subseteq K \big\}.$$

When K is clear from context, we will simply write r_{in} instead.

Proposition 10 (Poincaré for convex influences for general convex sets)**.** For convex $K \subseteq \mathbb{R}^n$ with $r_{\mathrm{in}} > 0$ (and hence $\mathrm{Vol}(K) > 0$), we have

$$\frac{\mathbf{I}[K]}{\mathrm{Vol}(K)} \geq r_{\mathrm{in}} \cdot \Omega\big(1 - \mathrm{Vol}(K)\big).$$

2.3 The Brascamp-Lieb Inequality

The following result of Brascamp and Lieb [BL76] generalizes the Gaussian Poincaré inequality to measures which are more log-concave than the Gaussian distribution.

Proposition 11 (Brascamp-Lieb inequality)**.** Let $\mathcal{D}$ be a probability distribution on $\mathbb{R}^n$ with density $e^{-V(x)} \cdot \varphi_n(x)$ for a convex function $V : \mathbb{R}^n \to \mathbb{R}$. Then for any differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\mathbf{Var}_{\boldsymbol{x} \sim \mathcal{D}}[f(\boldsymbol{x})] \leq \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}}\left[\|\nabla f(\boldsymbol{x})\|^2\right].$$

Vempala [Vem10] obtained a quantitative version of Proposition 11 in one dimension, which we state next. Note in particular that the following holds for non-centered Gaussians.

Proposition 12 (Lemma 4.7 of [Vem10])**.** Fix $\theta \in \mathbb{R}$ and let $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a log-concave function such that

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(\theta,1)}[\boldsymbol{x} f(\boldsymbol{x})] = 0.$$

Then $\mathbf{E}[\boldsymbol{x}^2 f(\boldsymbol{x})] \leq \mathbf{E}[f(\boldsymbol{x})]$ for $\boldsymbol{x} \sim N(\theta,1)$, with equality if and only if f is a constant function. Furthermore, if $\mathrm{supp}(f) \subseteq (-\infty, \varepsilon]$, then

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(\theta,1)}\left[\boldsymbol{x}^2 f(\boldsymbol{x})\right] \leq \left(1 - \frac{1}{2\pi} e^{-\varepsilon^2}\right) \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(\theta,1)}\left[f(\boldsymbol{x})\right].$$

2.4 The Classes of Distributions We Consider

We say that a distribution over $\mathbb{R}^n$ with density φ is *symmetric* if $\varphi(x) = \varphi(-x)$ for all x, and that a set $K \subseteq \mathbb{R}^n$ is symmetric if $-x \in K$ whenever $x \in K$.

We let $\mathcal{P}_{\mathrm{symm}}$ denote the class of all distributions $N(0, I_n)|_K$ where $K \subseteq \mathbb{R}^n$ may be any symmetric convex set, $\mathcal{P}_{\mathrm{conv}}$ denote the class of all such distributions where K may be any convex set (not necessarily symmetric), and $\mathcal{P}_{\mathrm{LTF}}$ denote the class of all such distributions where K may be any linear threshold function $\mathrm{sign}(v \cdot x \geq \theta)$. We let $\mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$ denote the class of all convex combinations (mixtures) of distributions from $\mathcal{P}_{\mathrm{symm}}$, and we remark that a distribution in $\mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$ can be viewed as $N(0, I_n)$ conditioned on a *mixture* of symmetric convex sets.

The following alternate characterization of $\mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$ may be of interest. Let $\mathcal{P}_{\mathrm{slcg}}$ denote the class of all symmetric distributions that are log-concave relative to the standard normal distribution, i.e. all distributions that have a density of the form $e^{-\tau(x)}\varphi_n(x)$ where $\tau(\cdot)$ is a symmetric convex function. Let $\mathrm{Mix}(\mathcal{P}_{\mathrm{slcg}})$ denote the class of all mixtures of distributions in $\mathcal{P}_{\mathrm{slcg}}$.

Claim 13. $\mathrm{Mix}(\mathcal{P}_{\mathrm{slcg}}) = \mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$.

Proof. We will argue below that $\mathcal{P}_{\mathrm{slcg}} \subseteq \mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$. Given this, it follows that any mixture of distributions in $\mathcal{P}_{\mathrm{slcg}}$ is a mixture of distributions in $\mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$, but since a mixture of distributions in $\mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$ is itself a distribution in $\mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$, this means that $\mathrm{Mix}(\mathcal{P}_{\mathrm{slcg}}) \subseteq \mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$. For the other direction, we observe that any distribution in $\mathcal{P}_{\mathrm{symm}}$ belongs to $\mathcal{P}_{\mathrm{slcg}}$,[3] and hence $\mathrm{Mix}(\mathcal{P}_{\mathrm{symm}}) \subseteq \mathrm{Mix}(\mathcal{P}_{\mathrm{slcg}})$.

Fix any distribution $\mathcal{D}$ in $\mathcal{P}_{\mathrm{slcg}}$ and let $e^{-\tau(x)}\varphi_n(x)$ be its density. We have that

$$e^{-\tau(x)}\varphi_n(x) = \mathbf{E}[A_{\boldsymbol{t}}(x)] \cdot \varphi_n(x) \tag{5}$$

where $A_t(x) = \mathbf{1}[e^{-\tau(x)} \geq t]$ and the expectation in (5) is over a uniform $\boldsymbol{t} \sim [0,1]$. Since τ is a symmetric convex function we have that the level set $\{x \in \mathbb{R}^n : e^{-\tau(x)} \geq t\}$ is a symmetric convex set, so $\mathcal{D}$ is a mixture of distributions in $\mathcal{P}_{\mathrm{symm}}$ as claimed above. $\qquad\square$

---

[3]Recall that a distribution in $\mathcal{P}_{\mathrm{symm}}$ has a density which is $\mathrm{Vol}(K)^{-1} \cdot K(x) \cdot \varphi_n(x)$ for some symmetric convex K.

3 An $O(n/\varepsilon^2)$-Sample Algorithm for Symmetric Convex Sets and Mixtures of Symmetric Convex Sets

In this section, we give an algorithm (cf. Algorithm 1) to distinguish Gaussians from (mixtures of) Gaussians truncated to a symmetric convex set.

3.1 Useful Structural Results

We record a few important lemmas which are going to be useful for the analysis in this section.

Lemma 14. Let $K \subseteq \mathbb{R}^n$ be a centrally symmetric convex set. If $\mathrm{Vol}(K) \leq 1 - \varepsilon$, then,

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,I_n)|_K}[\|\boldsymbol{x}\|^2] \leq n - c\varepsilon$$

for some absolute constant $c > 0$.

Proof. We have

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,I_n)|_K}\left[\|\boldsymbol{x}\|^2\right] = n - \frac{\sqrt{2} \cdot \mathbf{I}[K]}{\mathrm{Vol}(K)} \leq n - \sqrt{2} \cdot c'(1 - \mathrm{Vol}(K)) \leq n - \sqrt{2} \cdot c'\varepsilon,$$

where the equality is Equation (4), the first inequality is Proposition 9 (Poincaré for convex influences for symmetric convex sets), and the second inequality holds because $\mathrm{Vol}(K) \leq 1 - \varepsilon$. $\square$

Lemma 15. Let $K \subseteq \mathbb{R}^n$ be a convex set (not necessarily symmetric) and let $\mathcal{D} = N(0, I_n)|_K$. Then for any unit vector v, we have

$$\mathop{\mathbf{Var}}_{\boldsymbol{x} \sim \mathcal{D}}[v \cdot \boldsymbol{x}] \leq 1.$$

Proof. Given $c > 0$, we define $V_c : \mathbb{R}^n \to \{c, +\infty\}$ to be

$$V_c(x) = \begin{cases} c & \text{if } x \in K \\ +\infty & \text{if } x \notin K. \end{cases}$$

We note that $V_c(\cdot)$ is a convex function for any choice of $c > 0$, and that for a suitable choice of c, the density function of $\mathcal{D}$ is $e^{-V_c(x)} \cdot \gamma_n(x)$. Thus, we can apply the Brascamp-Lieb inequality to get that for any differentiable $f : \mathbb{R}^n \to \mathbb{R}$,

$$\mathop{\mathbf{Var}}_{\boldsymbol{x} \sim \mathcal{D}}[f(\boldsymbol{x})] \leq \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}}[\|\nabla f(\boldsymbol{x})\|^2]. \tag{6}$$

Now, we may assume without loss of generality that $v = e_1$. Taking $f(x) = x_1$ in Equation (6), we get that

$$\mathbf{Var}_{\boldsymbol{x} \sim \mathcal{D}}[\boldsymbol{x}_1] \leq 1,$$

which finishes the proof. $\square$

Now we can bound the variance of $\|\boldsymbol{x}\|^2$ when $\boldsymbol{x} \sim N(0, I_n)|_K$ for a symmetric convex set K.

Lemma 16. Let $\mathcal{D} = N(0, I_n)|_K$ for a symmetric convex set K. Then, $\mathbf{Var}_{\boldsymbol{x} \sim \mathcal{D}}[\|\boldsymbol{x}\|^2] \leq 4n$.

Proof. Taking $f(x) := \|x\|^2$ in Equation (6), we have that

$$\mathop{\mathbf{Var}}_{\boldsymbol{x} \sim \mathcal{D}}[\|\boldsymbol{x}\|^2] \leq 4 \cdot \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}}[\boldsymbol{x}_1^2 + \ldots + \boldsymbol{x}_n^2].$$

Since K is symmetric, for each $i \in [n]$ we have $\mathbf{E}_{\boldsymbol{x} \sim \mathcal{D}}[\boldsymbol{x}_i] = 0$ and hence $\mathbf{E}_{\boldsymbol{x} \sim \mathcal{D}}[\boldsymbol{x}_i^2] = \mathbf{Var}[e_i \cdot \boldsymbol{x}]$, which is at most 1 by Lemma 15. $\square$

3.2 An $O(n/\varepsilon^2)$-Sample Algorithm for Symmetric Convex Sets

We recall Theorem 1:

Theorem 17 (Restatement of Theorem 1). For a sufficiently large constant $C > 0$, the algorithm SYMM-CONVEX-DISTINGUISHER (Algorithm 1) has the following performance guarantee: given any $\varepsilon > 0$ and access to independent samples from any unknown distribution $\mathcal{D} \in \mathcal{P}_{\text{symm}}$, the algorithm uses Cn/ε^2 samples, and

1. If $\mathcal{D} = N(0, I_n)$, then with probability at least 9/10 the algorithm outputs "un-truncated";

2. If $d_{\text{TV}}(\mathcal{D}, N(0, I_n)) \geq \varepsilon$, then with probability at least 9/10 the algorithm outputs "truncated."

As alluded to in Section 1.2, SYMM-CONVEX-DISTINGUISHER uses the estimator from Equation (1).

Input: $\mathcal{D} \in \mathcal{P}_{\text{conv}}$, $\varepsilon > 0$

Output: "Un-truncated" or "truncated"

SYMM-CONVEX-DISTINGUISHER$(\mathcal{D}, \varepsilon)$:

1. For $T = C \cdot n/\varepsilon^2$, sample points $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)} \sim \mathcal{D}$.

2. Let $\mathbf{M} := \frac{1}{T} \sum_{i=1}^{T} \|\boldsymbol{x}^{(i)}\|^2$.

3. If $\mathbf{M} \geq n - c\varepsilon/2$, output "un-truncated," else output "truncated".

Algorithm 1: Distinguisher for (Mixtures of) Symmetric Convex Sets

Proof of Theorem 17. Let $\mathcal{D}_G := N(0, I_n)$ and $\mathcal{D}_T := N(0, I_n)|_K$. Then, for $\boldsymbol{x} \sim \mathcal{D}_G$, the random variable $\|\boldsymbol{x}\|^2$ follows the χ^2 distribution with n degrees of freedom, and thus we have

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}_G}[\|\boldsymbol{x}\|^2] = n; \quad \mathop{\mathbf{Var}}_{\boldsymbol{x} \sim \mathcal{D}_G}[\|\boldsymbol{x}\|^2] = 3n. \tag{7}$$

On the other hand, if $d_{\text{TV}}(\mathcal{D}, N(0, I_n)) \geq \varepsilon$ (equivalently, $\text{Vol}(K) \leq 1 - \varepsilon$), then using Lemma 14 and Lemma 16, it follows that

$$\mathbf{E}_{\boldsymbol{x} \sim \mathcal{D}_T}[\|\boldsymbol{x}\|^2] \leq n - c\varepsilon; \quad \mathbf{Var}_{\boldsymbol{x} \sim \mathcal{D}_T}[\|\boldsymbol{x}\|^2] \leq 4n. \tag{8}$$

Since in Algorithm 1 the samples $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}$ are independent, we have the following:

$$\mathbf{E}[\mathbf{M}] = n \quad \text{and} \quad \mathbf{Var}[\mathbf{M}] = \frac{3n}{T} \qquad \text{when } \mathcal{D} = \mathcal{D}_G,$$

$$\mathbf{E}[\mathbf{M}] = n - c\varepsilon \quad \text{and} \quad \mathbf{Var}[\mathbf{M}] \leq \frac{4n}{T} \qquad \text{when } \mathcal{D} = \mathcal{D}_T.$$

By choosing $T = Cn/\varepsilon^2$ (for a sufficiently large constant C), it follows that when $\mathcal{D} = \mathcal{D}_G$ (resp. $\mathcal{D} = \mathcal{D}_T$), with probability at least 9/10 we have $\mathbf{M} \geq n - c\varepsilon/2$ (resp. $\mathbf{M} < n - c\varepsilon/2$). This finishes the proof. $\qquad\square$

3.3 An $O(n/\varepsilon^2)$-Sample Algorithm for Mixtures of Symmetric Convex Sets

By extending the above analysis, we can show that Algorithm 1 succeeds for mixtures of (an arbitrary number of) symmetric convex sets. In particular, we have the following:

Theorem 18. For a sufficiently large constant $C > 0$, SYMM-CONVEX-DISTINGUISHER (Algorithm 1) has the following performance guarantee: given any $\varepsilon > 0$ and access to independent samples from any unknown distribution $\mathcal{D} \in \mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$, the algorithm uses Cn/ε^2 samples, and

1. If $\mathcal{D} = N(0, I_n)$, then with probability at least $9/10$ the algorithm outputs "un-truncated";

2. If $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}, N(0, I_n)) \geq \varepsilon$, then with probability at least $9/10$ the algorithm outputs "truncated."

The following lemma, which characterizes the mean and variance of a distribution in $\mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$ in terms of the components of the mixture, will crucial to the proof of Theorem 18:

Lemma 19. Let $\mathcal{X}$ denote a distribution over Gaussians truncated by symmetric convex sets. Suppose $\mathcal{D}_{\mathcal{X}} \in \mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$ is the mixture of $N(0, I_n)|_{\mathbf{K}}$ for $\mathbf{K} \sim \mathcal{X}$. Let $\boldsymbol{a_K}$ denote the random variable

$$\boldsymbol{a_K} = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0, I_n)|_{\mathbf{K}}}\left[\|\boldsymbol{x}\|^2\right] \qquad \text{where } \mathbf{K} \sim \mathcal{X}.$$

Then

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}}\left[\|\boldsymbol{x}\|^2\right] = \mathop{\mathbf{E}}_{\mathbf{K} \sim \mathcal{X}}[\boldsymbol{a_K}], \tag{9}$$

$$\mathop{\mathbf{Var}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}}\left[\|\boldsymbol{x}\|^2\right] \leq 4n + \mathop{\mathbf{Var}}_{\mathbf{K} \sim \mathcal{X}}[\boldsymbol{a_K}]. \tag{10}$$

Proof. Note that Equation (9) follows from linearity of expectation and the definition of $\boldsymbol{a_K}$. For Equation (10), note that for any symmetric convex set K, by definition of variance we have

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0, I_n)|_K}\left[\|\boldsymbol{x}\|^4\right] = \left(\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0, I_n)|_K}\left[\|\boldsymbol{x}\|^2\right]\right)^2 + \mathop{\mathbf{Var}}_{\boldsymbol{x} \sim N(0, I_n)|_K}\left[\|\boldsymbol{x}\|^2\right]$$
$$\leq \boldsymbol{a}_K^2 + 4n,$$

where the inequality is by Lemma 16. By linearity of expectation, it now follows that

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}}\left[\|\boldsymbol{x}\|^4\right] \leq 4n + \mathop{\mathbf{E}}_{\mathbf{K} \sim \mathcal{X}}\left[\boldsymbol{a}_{\mathbf{K}}^2\right].$$

Combining with Equation (9), we get Equation (10). $\qquad \square$

We are now ready to prove Theorem 18.

Proof of Theorem 18. Let $\mathcal{X}$ denote a distribution over symmetric convex sets. Define $\mathcal{D}_{\mathcal{X}} \in \mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$ to be the mixture of $N(0, I_n)_{\mathbf{K}}$ for $\mathbf{K} \sim \mathcal{X}$ and define $\mathcal{D}_G := N(0, I_n)$. Using the fact that the samples $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}$ are independent, as in the proof of Theorem 17, we have that

$$\mathbf{E}[\mathbf{M}] = n, \quad \mathbf{Var}[\mathbf{M}] = \frac{3n}{T} \qquad \text{when } \mathcal{D} = \mathcal{D}_G. \tag{11}$$

As $T = Cn/\varepsilon^2$ (for a sufficiently large constant C), it follows that when $\mathcal{D} = \mathcal{D}_G$, with probability at least $9/10$ we have that $\mathbf{M} \geq n - \varepsilon/2$.

11

Now we analyze the case that $\mathcal{D} = \mathcal{D}_\mathcal{X}$ has $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}, N(0, I_n)) \geq \varepsilon$. From Lemma 19, it follows that in this case

$$\mathbf{E}[\mathbf{M}] = \mathop{\mathbf{E}}_{\mathbf{K} \sim \mathcal{X}}[\boldsymbol{a}_\mathbf{K}], \tag{12}$$

$$\mathbf{Var}[\mathbf{M}] = \frac{\mathbf{Var}_{\boldsymbol{x} \sim \mathcal{D}_\mathcal{X}}\left[\|\boldsymbol{x}\|^2\right]}{T} \leq \frac{4n}{T} + \frac{\mathbf{Var}_{\mathbf{K} \sim \mathcal{X}}[\boldsymbol{a}_\mathbf{K}]}{T}. \tag{13}$$

Next, observe that

$$\mathop{\mathbf{E}}_{\mathbf{K} \sim \mathcal{X}}\left[(n - \boldsymbol{a}_\mathbf{K})\right] \geq c \cdot \mathop{\mathbf{E}}_{\mathbf{K} \sim \mathcal{X}}\left[1 - \mathrm{Vol}(\mathbf{K})\right] \geq c \cdot \mathrm{d}_{\mathrm{TV}}(\mathcal{D}, N(0, I_n)) \geq c\varepsilon, \tag{14}$$

where the first inequality uses Lemma 14 and the second inequality follows from the definition of TV distance. Now, observing that variance of a random variable is invariant under negation and translation and that $T = Cn/\varepsilon^2$, it follows from Equation (13) that

$$\mathbf{Var}[\mathbf{M}] \leq \frac{4n}{T} + \frac{\mathbf{Var}_{\mathbf{K} \sim \mathcal{X}}[\boldsymbol{a}_\mathbf{K}]}{T} \leq \frac{4\varepsilon^2}{C} + \frac{\varepsilon^2 \cdot \mathbf{Var}_{\mathbf{K} \sim \mathcal{X}}[n - \boldsymbol{a}_\mathbf{K}]}{Cn} \leq \frac{4\varepsilon^2}{C} + \frac{\varepsilon^2 \cdot \mathbf{E}_{\mathbf{K} \sim \mathcal{X}}\left[(n - \boldsymbol{a}_\mathbf{K})^2\right]}{Cn}.$$

By Equation (4) and Proposition 9, we have that $0 \leq a_K \leq n$ for any symmetric convex K. Thus, we can further upper bound the right hand side to obtain

$$\mathbf{Var}[\mathbf{M}] \leq \frac{4\varepsilon^2}{C} + \frac{\varepsilon^2 \cdot \mathbf{E}_{\mathbf{K} \sim \mathcal{X}}[n - \boldsymbol{a}_\mathbf{K}]}{C}.$$

Recalling from Equation (14) that $\mathbf{E}_{\mathbf{K} \sim \mathcal{X}}[n - \boldsymbol{a}_\mathbf{K}] \geq c\varepsilon$, a routine computation shows that for a sufficiently large constant C, we have

$$\mathbf{Var}[\mathbf{M}] \leq \frac{4\varepsilon^2}{C} + \frac{\varepsilon^2 \cdot \mathbf{E}_{\mathbf{K} \sim \mathcal{X}}[n - \boldsymbol{a}_\mathbf{K}]}{C} \leq \frac{\mathbf{E}_{\mathbf{K} \sim \mathcal{X}}\left[n - c\varepsilon/2 - \boldsymbol{a}_\mathbf{K}\right]^2}{100}.$$

Equation (12) and Chebyshev's inequality now give that when $\mathcal{D} = \mathcal{D}_\mathcal{X}$, with probability at least $9/10$ we have $\mathbf{M} \leq n - c\varepsilon/2$, completing the proof. $\qquad\square$

4 An $O(n/\varepsilon^2)$-Sample Algorithm for General Convex Sets

In this section we present a $O(n/\varepsilon^2)$-sample algorithm for distinguishing the standard normal distribution from the standard normal distribution restricted to an arbitrary convex set. More precisely, we prove the following:

Theorem 20. There is an algorithm, CONVEX-DISTINGUISHER (Algorithm 2), with the following performance guarantee: Given any $\varepsilon > 0$ and access to independent samples from any unknown distribution $\mathcal{D} \in \mathcal{P}_{\mathrm{conv}}$, the algorithm uses $O(n/\varepsilon^2)$ samples, runs in $\mathrm{poly}(n, 1/\varepsilon)$ time, and

1. If $\mathcal{D} = N(0, I_n)$, then with probability at least $9/10$ the algorithm outputs "un-truncated;"

2. If $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}, N(0, I_n)) \geq \varepsilon$, then with probability at least $9/10$ the algorithm outputs "truncated."

Note that the estimator $\mathbf{M}$ in Algorithm 2 is identical to the estimator $\mathbf{M}$ in Algorithm 1 to distinguish Gaussians restricted to (mixtures of) symmetric convex sets. As we will see, the analysis of Algorithm 1 via the Poincaré inequality for convex influences (cf. Proposition 9) extends to arbitrary convex sets with "large inradius." For the "small inradius" case, we further consider sub-cases depending on how close the center of mass of $\mathcal{D}$, denoted μ, is to the origin (see Figure 1):

Input: $\mathcal{D} \in \mathcal{P}_{\mathrm{conv}}$, $\varepsilon > 0$

Output: "un-truncated" or "truncated"

CONVEX-DISTINGUISHER($\mathcal{D}, \varepsilon$):

1. For $T = C \cdot n/\varepsilon^2$, sample points $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)} \sim \mathcal{D}$.

2. Set $\mathbf{M} := \frac{1}{T} \sum_{j=1}^{T} \|\boldsymbol{x}^{(j)}\|^2$ and $\boldsymbol{L} := \text{MEAN-ESTIMATOR}\Big(\{\boldsymbol{x}^{(j)}\}, 0.01\Big)$.

3. Output "truncated" if either

   (a) $\mathbf{M} \leq n - c\varepsilon/2$, or
   (b) $\|\boldsymbol{L}\|^2 \geq 0.05$;

   and output "un-truncated" otherwise.

Algorithm 2: Distinguisher for General Convex Sets

- **Case 1:** When $\|\mu\| \gg 0$, we detect truncation via estimating the mean $\boldsymbol{L}$ using Proposition 5.

- **Case 2:** When $\|\mu\| \approx 0$, we show that we can detect truncation via $\mathbf{M}$. This is our most technically-involved case and relies crucially on (small extensions of) Vempala's quantitative Brascamp-Lieb inequality (Proposition 12).

4.1 Useful Preliminaries

Below are two useful consequences of Vempala's quantitative one-dimensional Brascamp-Lieb inequality (Proposition 12) which will be useful in our analysis of Algorithm 2.

The following proposition says that if the center of mass of a convex body (with respect to the standard normal distribution) along a direction $v \in S^{n-1}$ is the origin, then the convex influence of v on the body is non-negative.

Proposition 21. Given a convex set $K \subseteq \mathbb{R}^n$ and $v \in S^{n-1}$, if

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0, I_n)} \big[K(\boldsymbol{x})\langle v, \boldsymbol{x}\rangle\big] = 0,$$

then $\mathbf{Inf}_v[K] \geq 0$.

Proof. We may assume without loss of generality that $v = e_1$. Note that the function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ defined by

$$f(x) := \mathop{\mathbf{E}}_{\boldsymbol{y} \sim N(0, I_{n-1})} \big[K(x, \boldsymbol{y})\big],$$

is a log-concave function (this is immediate from the Prékopa-Leindler inequality [Pré73, Lei72]). Furthermore, note that by Fact 8,

$$\sqrt{2} \cdot \mathbf{Inf}_v[K] = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)} \Big[f(\boldsymbol{x})(1 - \boldsymbol{x}^2)\Big],$$

and so the result follows by Proposition 12. $\qquad\square$

We also require a version of Proposition 12 for log-concave functions whose center of mass with respect to the standard normal distribution is not at the origin. Looking ahead, Proposition 22 will come in handy when analyzing Algorithm 2 for Gaussians restricted to convex sets with small inradius and with center of mass close to the origin.

Proposition 22. Let $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a one-dimensional log-concave function with

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)} \left[\boldsymbol{x} f(\boldsymbol{x}) \right] = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)} \left[\mu \cdot f(\boldsymbol{x}) \right]$$

for some $\mu \in \mathbb{R}$. Then

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)} \left[\boldsymbol{x}^2 f(\boldsymbol{x}) \right] \leq \left(1 + \mu^2 \right) \cdot \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)} \left[f(\boldsymbol{x}) \right].$$

Furthermore, if $\operatorname{supp}(f) \subseteq (-\infty, \varepsilon]$, then

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)} \left[\boldsymbol{x}^2 f(\boldsymbol{x}) \right] \leq \left(1 + \mu^2 - \frac{1}{2\pi} e^{-(\varepsilon-\mu)^2} \right) \cdot \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)} \left[f(\boldsymbol{x}) \right]. \tag{15}$$

We prove Proposition 22 by translating the log-concave function f so that its center of mass (with respect to a shifted Gaussian) is the origin, and then appealing to Proposition 12.

Proof. Note that it suffices to prove Equation (15). Consider the one-dimensional log-concave function $\widetilde{f} : \mathbb{R} \to \mathbb{R}_{\geq 0}$ given by

$$\widetilde{f}(x) := f(x + \mu).$$

It is clear that $\operatorname{supp}(\widetilde{f}) \subseteq (-\infty, \varepsilon - \mu]$ if $\operatorname{supp}(f) \subseteq (-\infty, \varepsilon]$. Note that

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(-\mu,1)} \left[\widetilde{f}(\boldsymbol{x}) \right] = \int_{\mathbb{R}} f(x + \mu) \varphi(x + \mu) \, dx = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)} \left[f(\boldsymbol{x}) \right]. \tag{16}$$

We also have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(-\mu,1)} \left[\boldsymbol{x} \widetilde{f}(\boldsymbol{x}) \right] &= \int_{\mathbb{R}} x f(x + \mu) \varphi(x + \mu) \, dx \\
&= \int_{\mathbb{R}} (y - \mu) f(y) \varphi(y) \, dy \\
&= \mathop{\mathbf{E}}_{\boldsymbol{y} \sim N(0,1)} \left[\boldsymbol{y} f(\boldsymbol{y}) \right] - \mathop{\mathbf{E}}_{\boldsymbol{y} \sim N(0,1)} \left[\mu \cdot f(\boldsymbol{y}) \right] \\
&= 0,
\end{aligned}$$

where we made the substitution $y = x - \mu$. Therefore, by Proposition 12, we have that

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(-\mu,1)} \left[\boldsymbol{x}^2 \widetilde{f}(\boldsymbol{x}) \right] \leq \left(1 - \frac{1}{2\pi} e^{-(\varepsilon-\mu)^2} \right) \cdot \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(-\mu,1)} \left[\widetilde{f}(\boldsymbol{x}) \right]. \tag{17}$$

However, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(-\mu,1)} \left[\boldsymbol{x}^2 \widetilde{f}(\boldsymbol{x}) \right] &= \int_{\mathbb{R}} x^2 f(x + \mu) \varphi(x + \mu) \, dx \\
&= \int_{\mathbb{R}} (y - \mu)^2 f(y) \varphi(y) \, dy \\
&= \mathop{\mathbf{E}}_{\boldsymbol{y} \sim N(0,1)} \left[\boldsymbol{y}^2 f(\boldsymbol{y}) \right] - \mathop{\mathbf{E}}_{\boldsymbol{y} \sim N(0,1)} \left[\mu^2 \cdot f(\boldsymbol{y}) \right]. \tag{18}
\end{aligned}$$

Equation (15) now follows from Equations (16) to (18). $\qquad \square$

4.2 Proof of Theorem 20

We can now turn to the proof of Theorem 20.

Proof of Theorem 20. Suppose first that $\mathcal{D} = N(0, I_n)$. In this case,

$$\mathbf{E}\,[\mathbf{M}] = \frac{1}{T} \sum_{j=1}^{T} \mathbf{E}\left[\|\boldsymbol{x}^{(j)}\|^2\right] = \frac{1}{T} \sum_{j=1}^{T} n = n. \tag{19}$$

We also have that

$$\mathbf{Var}\,[\mathbf{M}] = \frac{1}{T^2} \sum_{j=1}^{T} \mathbf{Var}\left[\|\boldsymbol{x}^{(j)}\|^2\right] = \frac{1}{T}\left(\mathbf{Var}_{\boldsymbol{x} \sim N(0,I_n)}\left[\|\boldsymbol{x}\|^2\right]\right) = \frac{1}{T} \sum_{i=1}^{n} \mathbf{Var}_{\boldsymbol{x}_i \sim N(0,1)}\left[\boldsymbol{x}_i^2\right] = \frac{2n}{T}, \tag{20}$$

where we used the fact that $\mathbf{Var}_{\boldsymbol{x} \sim N(0,1)}[\boldsymbol{x}^2] = 2$. Looking ahead, we also note that in this case, by Proposition 5 we have that

$$\|\boldsymbol{L}\|^2 \le 0.01 \tag{21}$$

with probability at least 0.99.

Next, suppose that $\mathcal{D} = N(0, I_n)_K$ for convex $K \subseteq \mathbb{R}^n$ with $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}, N(0, I_n)) \ge \varepsilon$. Let us write r_{in} for the in-radius of K. Suppose first that $r_{\mathrm{in}} \ge 0.1$. In this case, we have that

$$\mathbf{E}\,[\mathbf{M}] = \mathbf{E}_{\boldsymbol{x} \sim \mathcal{D}}\left[\|\boldsymbol{x}\|^2\right] \le n - \Omega(\varepsilon). \tag{22}$$

by Equation (2), Fact 8, and Proposition 10. By independence of the $\boldsymbol{x}^{(j)}$'s, we also have that

$$\mathbf{Var}[\mathbf{M}] = \frac{1}{T^2} \sum_{j=1}^{T} \mathbf{Var}_{\boldsymbol{x}^{(j)} \sim \mathcal{D}}\left[\|\boldsymbol{x}^{(j)}\|^2\right].$$

Note, however, that by Proposition 11 we have

$$\mathbf{Var}_{\boldsymbol{x} \sim \mathcal{D}}\left[\|\boldsymbol{x}\|^2\right] \le 4 \, \mathbf{E}_{\boldsymbol{x} \sim \mathcal{D}}\left[\|\boldsymbol{x}\|^2\right] \qquad \text{and so} \qquad \mathbf{Var}[\mathbf{M}] \le \frac{4n}{T}, \tag{23}$$

where the second inequality follows from Equation (22). From Equations (19) and (22), we have that the means of $\mathbf{M}$ under $N(0, I_n)$ versus $N(0, I_n)|_K$ differ by $\Omega(\varepsilon)$, and from Equations (20) and (23) we have that the standard deviations in both settings are on the order of $O(\sqrt{n/T})$. This shows that CONVEX-DISTINGUISHER indeed succeeds in distinguishing $\mathcal{D} = N(0, I_n)$ from $\mathcal{D} = N(0, I_n)_K$ with $O(n/\varepsilon^2)$ samples in the case that $r_{\mathrm{in}} \ge 0.1$.

For the rest of the proof we can therefore assume that $r_{\mathrm{in}} < 0.1$. It follows from the hyperplane separation theorem that there exists $x^* \in S^{n-1}(0.1)$ such that K lies entirely on one side of the hyperplane that is tangent to $S^{n-1}(0.1)$ at x^*. Recalling that the standard normal distribution is invariant under rotation, we can suppose without generality that x^* is the point $(0.1, 0^{n-1})$, so we have that either

$$K \subseteq \{x \in \mathbb{R}^n : x_1 < 0.1\} \qquad \text{or} \qquad K \subseteq \{x \in \mathbb{R}^n : x_1 \ge 0.1\},$$

corresponding to (a) and (b) respectively in Figure 1. Writing μ for the center of mass of $\mathcal{D}$, i.e.

$$\mu := \mathbf{E}_{\boldsymbol{x} \sim \mathcal{D}}\,[\boldsymbol{x}],$$

15

Figure 1: The "small inradius" ($r_{\text{in}} \leq 0.1$) setting in the analysis of Algorithm 2, with μ denoting the center of mass of K. Our estimator for (a) is $\text{Avg}\left(\|\boldsymbol{x}^{(j)}\|^2\right)$, whereas for (b) we simply estimate μ.

we can apply another rotation to obtain $\mu = (\mu_1, \mu_2, 0^{n-2})$ while maintaining that $x^* = (0.1, 0^{n-1})$. Now we consider two cases based on the norm of μ:

Case 1. If $\|\mu\|^2 \geq 0.06$, then we claim that Step 3(b) of Algorithm 2 will correctly output "truncated" with probability at least $99/100$. Indeed, by the Brascamp-Lieb inequality, we have that $\text{tr}(\Sigma) \leq n$ where Σ is the covariance matrix of $\mathcal{D}$, and so Proposition 5 implies that for a suitable choice of C, we will have $\|\mu - \boldsymbol{L}\| \leq 0.001$ with probability at least 0.99, and hence $\|\boldsymbol{L}\|^2 \geq 0.05$.

Case 2. If $\|\mu\|^2 < 0.06$, then we will show that Algorithm 2 will output "untruncated" with probability at least $9/10$ in Step 3(a). We will do this by proceeding analogously to the "large inradius" ($r_{\text{in}} \geq 0.1$) setting considered earlier. Recall that

$$\mathbf{E}\left[\mathbf{M}\right] = \sum_{i=1}^{n} \mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}}\left[\boldsymbol{x}_i^2\right]. \tag{24}$$

For $i \in \{3, \ldots, n\}$, as $\mu_i = 0$, we have by Proposition 21 that $\mathbf{Inf}_i[K] \geq 0$, and so

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}}\left[\boldsymbol{x}_i^2\right] \leq 1 \qquad \text{for } i \in \{3, \ldots, n\} \tag{25}$$

by Fact 8.

We now consider coordinates 1 and 2. Consider the one-dimensional log-concave functions $f_1, f_2 : \mathbb{R} \to \mathbb{R}_{\geq 0}$ defined by

$$f_1(x) := \mathop{\mathbf{E}}_{\boldsymbol{y} \sim N(0, I_{n-1})}\left[K(x, \boldsymbol{y})\right] \qquad \text{and} \qquad f_2(x) := \mathop{\mathbf{E}}_{\boldsymbol{y} \sim N(0, I_{n-1})}\left[K(\boldsymbol{y}_1, x, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n-1})\right].$$

Note that $\mathbf{E}[f_1] = \mathbf{E}[f_2] = \text{Vol}(K)$. It is also immediate that

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}}\left[\boldsymbol{x}_i^2\right] = \frac{\mathbf{E}_{\boldsymbol{x} \sim N(0,1)}\left[\boldsymbol{x}^2 f_i(\boldsymbol{x})\right]}{\text{Vol}(K)}. \tag{26}$$

Since we have

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim N(0,1)}\left[\boldsymbol{x}f_1(\boldsymbol{x})\right] = \mu_1 \cdot \mathrm{Vol}(K) \qquad \text{and} \qquad \mathop{\mathbf{E}}_{\boldsymbol{x}\sim N(0,1)}\left[\boldsymbol{x}f_2(\boldsymbol{x})\right] = \mu_2 \cdot \mathrm{Vol}(K),$$

it follows from Proposition 22 that

$$\frac{\mathbf{E}_{\boldsymbol{x}\sim N(0,1)}\left[\boldsymbol{x}^2 f_1(\boldsymbol{x})\right]}{\mathrm{Vol}(K)} \leq 1 + \mu_1^2 - \frac{1}{2\pi}e^{-(0.1-\mu_1)^2} \qquad \text{and} \qquad \frac{\mathbf{E}_{\boldsymbol{x}\sim N(0,1)}\left[\boldsymbol{x}^2 f_2(\boldsymbol{x})\right]}{\mathrm{Vol}(K)} \leq 1 + \mu_2^2 \quad (27)$$

(note that we used the fact that $\mathrm{supp}(f_1) \subseteq (-\infty, 0.1]$ in the first inequality above). Combining Equations (26) and (27) and recalling that $\|\mu\|^2 < 0.06$, we get that

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left[\boldsymbol{x}_1^2 + \boldsymbol{x}_2^2\right] \leq 2 + \|\mu\|^2 - \frac{1}{2\pi}e^{-(0.1-\mu_1)^2} < 2.06 - \frac{1}{2\pi}e^{-(0.1+\sqrt{0.06})^2} < 1.95. \qquad (28)$$

Combining Equations (24), (25) and (28), we get that

$$\mathbf{E}[\mathbf{M}] = \mathbf{E}\left[\|\boldsymbol{x}\|^2\right] \leq n - 0.05. \qquad (29)$$

As in Equation (23), by the Brascamp-Lieb inequality (Proposition 11) we have that

$$\mathbf{Var}[\mathbf{M}] \leq \frac{4n}{T}, \qquad (30)$$

and so by Equation (29), Equation (30) and Chebyshev's inequality, for a suitable choice of C algorithm CONVEX-DISTINGUISHER will output "truncated" in Step 3(a) with probability at least 0.9. $\qquad \square$

5 An $\Omega(n)$-Sample Lower Bound for Testing Truncation

In this section, we present an $\Omega(n)$-sample lower bound for testing convex truncation. Our lower bound is information-theoretic and applies to all algorithms, computationally efficient or otherwise. More formally, we prove the following:

Theorem 23. Let A be any algorithm which is given access to samples from an unknown distribution $\mathcal{D}$ and has the following performance guarantee:

1. If $\mathcal{D} = N(0, I_n)$, then with probability at least $9/10$ the algorithm outputs "un-truncated";

2. If $\mathcal{D} \in \mathcal{P}_{\mathrm{symm}}$ and has $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}, N(0, I_n)) = 1$ then with probability at least $9/10$ the algorithm outputs "truncated."[4]

Then A must use at least $\Omega(n)$ samples from $\mathcal{D}$.

As $\mathcal{P}_{\mathrm{symm}} \subseteq \mathcal{P}_{\mathrm{conv}}$ and $\mathcal{P}_{\mathrm{symm}} \subseteq \mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$, Theorem 23 immediately that the algorithms CONVEX-DISTINGUISHER and SYMM-CONVEX-DISTINGUISHER are optimal in terms of sample complexity for testing convex truncation and truncation by a mixture of symmetric convex sets respectively.

---

[4]We note that if $\mathcal{D} \in \mathcal{P}_{\mathrm{symm}}$ has $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}, N(0, I_n)) = 1$, then $\mathcal{D}$ is obtained by conditioning the Gaussian distribution $N(0, I_n)$ on a symmetric convex set of zero measure. In order for this to be a well-defined operation (see for e.g. [Wik22]), we need to specify a suitable limiting sequence of sets of positive measure; we defer this discussion to Remark 28.

5.1 Useful Preliminaries

Before proceeding to the proof of Theorem 23, we introduce some useful notation and recall preliminaries on the *Wishart distribution*. Let $\mathfrak{C}_p$ denote the cone of $p \times p$ positive semi-definite matrices. We will write $(x)_+ := \max(0, x)$.

Definition 24 (Wishart distribution). For $p \leq n$, let $\mathrm{Wis}(p, n)$ denote the *Wishart distribution on $p \times p$ matrices with n degrees of freedom*, i.e. the distribution on symmetric positive-semidefinite matrices obtained by

1. Drawing $\boldsymbol{G}_i \sim N(0, I_p)$ for $i \in [n]$.

2. Outputting $\boldsymbol{S} := \sum_{i=1}^n \boldsymbol{G}_i \boldsymbol{G}_i^T$.

We will make use of the following well-known expression for the density of $\mathrm{Wis}(p, n)$.

Fact 25 (see e.g. [Eat07]). Let $\Psi_{p,n}$ denote the density of $\mathrm{Wis}(p, n)$. We have

$$\Psi_{p,n}(A) = \frac{\det(A)^{(n-p-1)/2} \cdot \exp\left(-\mathrm{Tr}(A)/2\right)}{2^{np/2} \pi^{p(p-1)/4} \prod_{i=1}^p \Gamma\left(\frac{1}{2}(n+1-i)\right)} \cdot \mathbf{1}\{A \succeq 0\}. \tag{31}$$

We will also require the following central limit theorem for the log-determinant of a matrix drawn according to the Wishart distribution $\mathrm{Wis}(p, n)$ when $p = \Theta(n)$. The following statement was obtained by Jonsson [Jon82].

Theorem 26 (Theorem 5.1 of [Jon82]). Suppose $n, p = p(n) \in \mathbb{N}$ are such that

$$\lim_{n \to \infty} \frac{p}{n} = y \in (0, 1).$$

Then as $n \to \infty$, the following holds:

$$\frac{1}{\sqrt{-2\log(1-y)}} \log\left(\frac{\det(\mathbf{W})}{(n-1)\cdots(n-p)}\right) \longrightarrow_L N(0, 1) \qquad \text{for } \mathbf{W} \sim \mathrm{Wis}(p, n),$$

where the convergence is in distribution.

5.2 Proof of Theorem 23

We start by formally defining our distribution over truncations of $N(0, I_n)$ by symmetric convex sets.

Notation 27. Let $\mathcal{D}_{\mathrm{plane}}$ be the distribution on origin-centered hyperplanes induced by the Haar measure on S^{n-1}, i.e. it is the distribution on hyperplanes obtained by

1. First drawing a Haar-random vector $\boldsymbol{v} \sim S^{n-1}$; and then

2. Outputting the origin-centered hyperplane $\boldsymbol{v}^\perp := \left\{x \in \mathbb{R}^n : \langle x, \boldsymbol{v} \rangle = 0\right\}$.

In what follows, we will show that no algorithm can distinguish between $N(0, I_n)$ and $N(0, I_n)|_{\mathbf{K}}$ for $\mathbf{K} \sim \mathcal{D}_{\mathrm{plane}}$ with cn samples for some sufficiently small absolute constant c; note that this immediately implies Theorem 23. We first show that distinguishing between $N(0, I_n)$ and $N(0, I_n)|_{\mathbf{K}}$ is equivalent to distinguishing between Wishart distributions with n and $(n-1)$ degrees of freedom.

18

Remark 28. The attentive reader may notice that the sets $\mathbf{K}$ defined above have $\mathrm{Vol}(\mathbf{K}) = 0$, and thus $N(0, I_n)|_{\mathbf{K}}$ is the standard normal distribution conditioned on an event of measure zero. For this to be a well defined operation (see e.g. [Wik22]), we need to specify how our measure-zero sets are obtained as the limit of a sequence of sets of positive measure. The limiting process we use for a set $K = \boldsymbol{v}^{\perp}$ is taking a sequence of slabs

$$\boldsymbol{v}_\varepsilon^{\perp} := \{x : |v \cdot x| \le \varepsilon\}$$

and letting $\varepsilon \to 0$. In the limit, the distribution induced for $N(0, I_n)|_{\mathbf{K}}$ is a symmetric distribution restricted to $\mathbf{K}$ corresponding to $N(0, I_{n-1})$.

Claim 29. For $p \le n - 1$, there exists an algorithm $\mathcal{A}$ which can distinguish between $N(0, I_n)$ and $N(0, I_n)|_{\mathbf{K}}$ where $\mathbf{K} \sim \mathcal{D}_{\mathrm{plane}}$ (i.e. outputs "un-truncated" with probability $9/10$ given samples from the former and outputs "truncated" with probability $9/10$ given samples from the latter) with p samples if and only if there exists an algorithm $\mathcal{A}'$ that can distinguish between $\mathrm{Wis}(p, n)$ and $\mathrm{Wis}(p, n-1)$ with 1 sample.

Proof. We start with the "if" direction – namely, the existence of $\mathcal{A}'$ immediately allows us to distinguish between $N(0, I_n)$ and $N(0, I_n)|_{\mathbf{K}}$ given p samples $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)}$ by computing the $p \times p$ matrix $\mathbf{W}$ given by

$$\mathbf{W}_{i,j} := \left\langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)} \right\rangle. \tag{32}$$

Note that if $\boldsymbol{x}^{(i)} \sim N(0, I_n)$ for $i \in [p]$, then $\mathbf{W} \sim \mathrm{Wis}(p, n)$. On the other hand, if $\boldsymbol{x}^{(i)} \sim N(0, I_n)|_{\mathbf{K}}$ for $i \in [p]$, then since the distribution of the inner product of two draws from $N(0, I_n)|_K$ is the same for all $(n-1)$-dimensional origin-centered hyperplanes K, it is easy to see that $\mathbf{W} \sim \mathrm{Wis}(p, n-1)$.

Before proceeding with the proof for the "only if" direction, i.e., how the existence of $\mathcal{A}$ implies the existence of $\mathcal{A}'$, we make a few important observations.

First, observe that any $p \times p$ matrix W where

$$W_{i,j} = \left\langle x^{(i)}, x^{(j)} \right\rangle \qquad \text{where } x^{(i)}, x^{(j)} \in \mathbb{R}^n$$

uniquely determines the collection of points $\{x^{(1)}, \ldots, x^{(p)}\}$ up to an orthogonal transformation.

Given any symmetric $p \times p$ psd matrix W, we define $\Sigma(W)$ as

$$\Sigma(W) := \left\{ \left(x^{(1)}, \ldots, x^{(p)} \right) \in \mathbb{R}^{np} : W_{i,j} = \left\langle x^{(i)}, x^{(j)} \right\rangle \right\}.$$

Note that $\Sigma(W)$ is a compact Hausdorff set (under the usual topology on $\mathbb{R}^{np}$). As the orthogonal group $\mathrm{O}(n)$ acts transitively on $\Sigma(W)$ for all symmetric $p \times p$ matrices W—where the group action is given by the map

$$\left(Q, (x^{(1)}, \ldots, x^{(p)}) \right) \mapsto \left(Qx^{(1)}, \ldots, Qx^{(p)} \right)$$

for $Q \in \mathrm{O}(n)$—it follows by Weil's theorem (cf. Theorem 6.2 of [DS14]) that there is a unique $\mathrm{O}(n)$-invariant probability measure on $\Sigma(W)$. Call this probability measure $\mathcal{D}_{\Sigma(W)}$.

Given matrix $W \in \mathfrak{C}_p$, consider the process that (a) computes a canonical choice of p points $(x^{(1)}, \ldots, x^{(p)}) \in \Sigma(W)$; and then (b) applies a Haar-random orthogonal transformation $\mathbf{Q} \sim \mathrm{O}(n)$ to output the points

$$\mathbf{Q}(W) := (\mathbf{Q}x^{(1)}, \ldots, \mathbf{Q}x^{(p)}).$$

Given any $W \in \mathfrak{C}_p$, the distribution of $\mathbf{Q}(W)$ on $\Sigma(W)$ is $\mathrm{O}(n)$-invariant under the above group action. Furthermore, by the preceding discussion, this is the unique such measure on $\Sigma(W)$, $\mathcal{D}_{\Sigma(W)}$. We record this fact below.

Fact 30. For every matrix $W \in \mathfrak{C}_p$, there is a unique $\mathrm{O}(n)$-invariant probability measure $\mathcal{D}_{\Sigma(W)}$ on the set $\Sigma(W)$, and further, the algorithm given by (a) and (b) above samples from $\mathcal{D}_{\Sigma(W)}$.

The next claim states that samples in the untruncated case can be simulated by sampling $\mathbf{W} \sim \mathrm{Wis}(p, n)$ and then sampling from $\mathcal{D}_{\Sigma(\mathbf{W})}$.

Claim 31. For $p < n$, the distribution of $(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)})$ where each $\boldsymbol{x}^{(i)} \sim N(0, I_n)$ is the same as sampling $\mathbf{W} \sim \mathrm{Wis}(p, n)$ and then sampling $(\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(p)}) \sim \mathcal{D}_{\Sigma(\mathbf{W})}$.

Proof. Consider a sample $(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)})$ where each $\boldsymbol{x}^{(i)} \sim N(0, I_n)$. Corresponding to this draw, we define $\mathbf{W}$ given by Equation (32). Note that the distribution of $\mathbf{W}$ is given by $\mathrm{Wis}(p, n)$. Thus, the distribution of $(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)})$ can be equivalently defined as sampling $\mathbf{W} \sim \mathrm{Wis}(p, n)$ and then sampling from the distribution $(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)}) \sim N(0, I_n)^p$ conditioned on the Gram matrix of $(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)})$ being $\Sigma(\mathbf{W})$ (we denote this conditional distribution by $\mathcal{X}_{\mathbf{W}}$).

The crucial observation is that the distribution of $(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)})$ is $\mathrm{O}(n)$-invariant. Since the image of $\Sigma(W)$ (for any W) under $\mathrm{O}(n)$ is $\Sigma(W)$ itself, it follows that for every W, the conditional distributions $\mathcal{X}_W$ are $\mathrm{O}(n)$-invariant. By applying Fact 30, it follows that for every W, $\mathcal{X}_W$ is the same as $\mathcal{D}_{\Sigma(W)}$, thus finishing the proof. $\square$

We now have the analogous claim for the truncated case as well.

Claim 32. For $p < n$, the distribution of $(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)})$ where each $\boldsymbol{x}^{(i)} \sim N(0, I_n)|_{\mathbf{K}}$ where $\mathbf{K} \sim \mathcal{D}_{\mathrm{plane}}$ is the same as sampling $\mathbf{W} \sim \mathrm{Wis}(p, n-1)$ and then sampling $(\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(p)}) \sim \mathcal{D}_{\Sigma(\mathbf{W})}$.

Proof. Note that in this case as well, the distribution $(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(p)})$ is invariant under the action of $\mathrm{O}(n)$. With this observation, the proof in this case is identical to that of Claim 31 except that the matrix $\mathbf{W}$ given by Equation (32) is now distributed as $\mathrm{Wis}(p, n-1)$. $\square$

Thus, now consider the process where given $\mathbf{W}$, we output $(\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(p)}) \sim \mathcal{D}_{\Sigma(W)}$. We observe that

1. If $\mathbf{W} \sim \mathrm{Wis}(p, n)$, then $(\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(p)})$ are distributed as p i.i.d. samples from $N(0, I_n)$.

2. If $\mathbf{W} \sim \mathrm{Wis}(p, n-1)$, then $(\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(p)})$ are distributed as p i.i.d. samples from $N(0, I_n)|_{\mathbf{K}}$ where $\mathbf{K} \sim \mathcal{D}_{\mathrm{plane}}$.

Thus, an algorithm $\mathcal{A}$ which can distinguish $N(0, I_n)$ versus $N(0, I_n)|_{\mathbf{K}}$ (where $\mathbf{K} \sim \mathcal{D}_{\mathrm{plane}}$) with p samples can distinguish between $\mathrm{Wis}(p, n)$ and $\mathrm{Wis}(p, n-1)$ with one sample (with the same distinguishing probability). $\square$

With this in hand, Theorem 23 is immediately implied by the following:

$$\mathrm{d}_{\mathrm{TV}}\big(\mathrm{Wis}(p, n), \mathrm{Wis}(p, n-1)\big) \leq 0.1 \qquad \text{for } p = 0.00001(n-1). \tag{33}$$

The rest of the proof establishes Equation (33) for the above choice of p. Writing $\mu_{p,n-1}$ for the measure on $\mathfrak{C}_p$ corresponding to the density $\Psi_{p,n-1}$, we have

$$\mathrm{d}_{\mathrm{TV}}\big(\mathrm{Wis}(p, n), \mathrm{Wis}(p, n-1)\big) = \int_{\mathfrak{C}_p} \left(1 - \frac{\Psi_{p,n}(A)}{\Psi_{p,n-1}(A)}\right)_+ d\mu_{p,n-1}(A). \tag{34}$$

For notational convenience, we define

$$\alpha_{p,n}(A) := \log\left(\frac{\Psi_{p,n}(A)}{\Psi_{p,n-1}(A)}\right).$$

20

Note that in order to establish Equation (33), it suffices to show that

$$\big|\alpha_{p,n}(A)\big| \le 0.02 \text{ with probability at least } 0.99 \text{ under } \mu_{n-1}. \tag{35}$$

To see this, note that $0 \le \left(1 - \frac{\Psi_n(A)}{\Psi_{n-1}(A)}\right)_+ \le 1$ for all A, and so using Equation (34) we have

- A contribution of at most 0.01 to the variation distance from outcomes of A such that $|\alpha_{p,n}(A)| > 0.02$; and

- A contribution of at most $1 - \exp(-0.02) \le 0.04$ to the variation distance from outcomes of A such that $|\alpha_{p,n}(A)| \le 0.02$.

Using Fact 25, for $A \in \mathfrak{C}_p$ we can write

$$
\begin{aligned}
\alpha_{p,n}(A) &= \log\left(\frac{\det(A)^{1/2}}{2^{p/2}} \cdot \frac{\Gamma\big((n-p)/2\big)}{\Gamma\big(n/2\big)}\right) \\
&= \frac{\log \det(A) - p}{2} - \sum_{i=\frac{n-p}{2}}^{\frac{n-2}{2}} \log i \\
&= \frac{\log \det(A) - p}{2} - \sum_{i=1}^{\frac{p}{2}} \log\left(\frac{n-p}{2} + i - 1\right) \\
&= \frac{\log \det(A) - p}{2} - \left(\sum_{i=1}^{\frac{p}{2}} \log\big(n - p + 2(i-1)\big)\right) + \frac{p}{2} \\
&= \frac{\log \det(A)}{2} - \sum_{i=1}^{\frac{p}{2}} \log\big(n - p + 2(i-1)\big). \tag{36}
\end{aligned}
$$

We pause to recall Theorem 26, from which we know that the $\log \det(\mathbf{W})$ where $\mathbf{W} \sim \text{Wis}(n - 1, p)$ converges in distribution to a $N(\mu, \sigma^2)$ random variable where

$$\mu := \sum_{j=1}^{p} \log\big(n - p + (j-2)\big) \quad \text{and} \quad \sigma^2 := 2\log\left(\frac{1}{1 - p/(n-1)}\right) = 2\log\left(\frac{1}{0.99999}\right).$$

This in turn implies that for every $t \in \mathbb{R}$, there exists n large enough such that

$$\Pr_{\mathbf{W} \sim \text{Wis}(p, n-1)}\big[\big|\log \det(\mathbf{W}) - \mu\big| > t\big] \le 0.005 + \Pr_{\boldsymbol{g} \sim N(0, \sigma^2)}\big[|\boldsymbol{g}| > t\big]. \tag{37}$$

21

Returning to Equation (36), we can write

$$\alpha_{p,n}(A) = \frac{\log\det(A) - \mu}{2} + \frac{\mu - 2\sum_{i=1}^{\frac{p}{2}}\log\left(n - p + 2(i-1)\right)}{2}$$

$$= \frac{\log\det(A) - \mu}{2} + \frac{1}{2}\left(\sum_{j=1}^{p}\log(n - p + (j-2)) - 2\sum_{i=1}^{\frac{p}{2}}\log(n - p + 2(i-1))\right)$$

$$= \frac{\log\det(A) - \mu}{2} + \frac{1}{2}\sum_{i=1}^{p/2}\log\left(\frac{n - p + 2j - 1}{n - p + 2j - 2}\right)$$

$$\leq \frac{\log\det(A) - \mu}{2} + \frac{p}{4}\log\left(1 + \frac{1}{n-p}\right)$$

$$\leq \frac{\log\det(A) - \mu}{2} + 0.001$$

where the final inequality uses the fact that $1 + x \leq e^x$ and our choice of $p = 0.00001n$. Combining this with Equation (37), we have that for sufficiently large n (where we take $t = 0.038$),

$$\Pr_{\mathbf{W}\sim\text{Wis}(p,n-1)}\left[|\alpha_{p,n}(\mathbf{W})| > 0.02\right] \leq \Pr_{\mathbf{W}\sim\text{Wis}(p,n-1)}\left[\left|\log\det(\mathbf{W}) - \mu\right| > 0.038\right]$$

$$\leq 0.005 + \Pr_{\mathbf{g}\sim N(0,\sigma^2)}\left[|\mathbf{g}| > 0.038\right]$$

$$\leq 0.01,$$

establishing Equation (35). This in turn establishes Equation (34), which together with Claim 29 completes the proof of Theorem 23. □

Acknowledgements

References

[BC14]   N. Balakrishnan and Erhard Cramer. *The art of progressive censoring*. Springer, 2014. 1

[Ber60]   Daniel Bernoulli. Essai d'une nouvelle analyse de la mortalité causeé par la petite vérole, et des avantages de l'inoculation pour la preévenir. *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem*, pages 1–45, 1760. 1

[BFR⁺13]   T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing Closeness of Discrete Distributions. *J. ACM*, 60(1):4, 2013. 26

[BKR04]   Tugkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *ACM Symposium on Theory of Computing*, pages 381–390, 2004. 5

[BL76]   H. Brascamp and E. Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log-concave functions and with an application to the diffusion equation. *Journal of Functional Analysis*, 22:366–389, 1976. 8

[BY02]   Ziv Bar-Yossef. *The Complexity of Massive Data Set Computations*. PhD thesis, UC Berkeley, 2002. Adviser: Christos Papadimitriou. Available at http://webee.technion.ac.il/people/zivby/index_files/Page1489.html.

[Can20]   Clément L. Canonne. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Number 9 in Graduate Surveys. Theory of Computing Library, 2020. 5

[CDS20]   Xue Chen, Anindya De, and Rocco A. Servedio. Testing noisy linear functions for sparsity. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 610–623. ACM, 2020. 1

[CFSS17]   X. Chen, A. Freilich, R. Servedio, and T. Sun. Sample-based high-dimensional convexity testing. In *Proceedings of the 17th Int. Workshop on Randomization and Computation (RANDOM)*, pages 37:1–37:20, 2017. 1, 4

[Coh16]   A. Clifford Cohen. *Truncated and censored samples: theory and applications*. CRC Press, 2016. 1

[DDS+13]   C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA 2013*, pages 729–746, 2013. 5

[DGTZ18]   C. Daskalakis, T. Gouleakis, C. Tzamos, and M. Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018*, pages 639–649. IEEE Computer Society, 2018. 1, 4

[DGTZ19]   Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Computationally and statistically efficient truncated regression. In *Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pages 955–960, 2019. 1, 4

[Dic14]   L. Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284, 2014. 1

[DKTZ21]   Constantinos Daskalakis, Vasilis Kontonis, Christos Tzamos, and Emmanouil Zampetakis. A Statistical Taylor Theorem and Extrapolation of Truncated Densities. In *Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pages 1395–1398, 2021. 1, 4

[DNS21a]   Anindya De, Shivam Nadimpalli, and Rocco A. Servedio. Convex Influences. Manuscript, available at https://arxiv.org/abs/2109.03107, 2021. 7

[DNS21b] Anindya De, Shivam Nadimpalli, and Rocco A. Servedio. Quantitative correlation inequalities via semigroup interpolation. In James R. Lee, editor, *12th Innovations in Theoretical Computer Science Conference, ITCS 2021, January 6-8, 2021, Virtual Conference*, volume 185 of *LIPIcs*, pages 69:1–69:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. 7

[DNS22] Anindya De, Shivam Nadimpalli, and Rocco A. Servedio. Convex influences. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference, ITCS*, volume 215 of *LIPIcs*, pages 53:1–53:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022. 3, 5, 7

[DS14] Joe Diestel and Angela Spalsbury. *The joys of Haar measure.* American Mathematical Soc., 2014. 19

[DS21] Anindya De and Rocco A. Servedio. Weak learning convex sets under normal distributions. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 1399–1428. PMLR, 2021. 3

[Eat07] Morris L Eaton. The wishart distribution. In *Multivariate Statistics*, volume 53, pages 302–334. Institute of Mathematical Statistics, 2007. 18

[FKT20] Dimitris Fotakis, Alkis Kalavasis, and Christos Tzamos. Efficient parameter estimation of truncated boolean product distributions. In *Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 1586–1600, 2020. 1, 4

[Gal97] Francis Galton. An examination into the registered speeds of American trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London*, 62(379-387):310–315, 1897. 1

[GGR98] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45:653–750, 1998. 4

[GR00] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *Electronic Colloqium on Computational Complexity*, 7(20), 2000. 26

[Hop20] Samuel B. Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193 – 1213, 2020. 6

[Joh01] Iain M. Johnstone. Chi-square oracle inequalities. In *State of the art in probability and statistics*, pages 399–418. Institute of Mathematical Statistics, 2001.

[Jon82] Dag Jonsson. Some limit theorems for the eigenvalues of a sample covariance matrix. *Journal of Multivariate Analysis*, 12(1):1–38, 1982. 4, 18

[KBV20] Weihao Kong, Emma Brunskill, and Gregory Valiant. Sublinear Optimal Policy Value Estimation in Contextual Bandits. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4377–4387. PMLR, 2020. 1

[KOS08] A. Klivans, R. O'Donnell, and R. Servedio. Learning geometric concepts via Gaussian surface area. In *Proc. 49th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 541–550, 2008. 1, 4

[KTZ19]   Vasilis Kontonis, Christos Tzamos, and Manolis Zampetakis. Efficient truncated statis-
          tics with unknown truncation. In David Zuckerman, editor, *60th IEEE Annual Sym-
          posium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA,
          November 9-12, 2019*, pages 1578–1595. IEEE Computer Society, 2019. 1, 4

[KV18]   W. Kong and G. Valiant. Estimating learnability in the sublinear data regime. *Advances
          in Neural Information Processing Systems*, 31, 2018. 1

[Lei72]   L. Leindler. On a certain converse of Hölder's inequality. II. *Acta Universitatis Szege-
          diensis. Acta Scientiarum Mathematicarum*, 33(3-4):217–223, 1972. 13

[LM18]   Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random
          vector. *The Annals of Statistics*, 47(2):783–794, 2018. 6

[Naz03]   F. Nazarov. On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaus-
          sian measure. In *Geometric aspects of functional analysis (2001-2002)*, pages 169–187.
          Lecture Notes in Math., Vol. 1807, Springer, 2003. 4

[O'D14]   Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. 7

[Pea02]   Karl Pearson. On the systematic fitting of frequency curves. *Biometrika*, 2:2–7, 1902. 1

[Pré73]   András Prékopa. On logarithmic concave measures and functions. *Acta Universitatis
          Szegediensis. Acta Scientiarum Mathematicarum*, 34:335–343, 1973. 13

[RS09]   Ronitt Rubinfeld and Rocco A. Servedio. Testing monotone high-dimensional distribu-
          tions. *Random Struct. Algorithms*, 34(1):24–44, 2009. 5

[Sch86]   Helmut Schneider. *Truncated and censored samples from normal populations*. Marcel
          Dekker, Inc., 1986. 1

[Vem10]   Santosh S. Vempala. Learning convex concepts from gaussian distributions with PCA.
          In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010,
          October 23-26, 2010, Las Vegas, Nevada, USA*, pages 124–130. IEEE Computer Society,
          2010. 3, 5, 8

[Wik22]   Wikipedia contributors. Borel-Kolmogorov paradox. *Wikipedia, The Free Encyclopedia*,
          2022. 17, 19

A   Hardness for Mixtures of General Convex Sets

Theorem 2 gives an efficient ($O(n)$-sample) algorithm that distinguishes $N(0, I_n)$ from $N(0, I_n)$ conditioned on a mixture of (any number of) symmetric convex sets, and Theorem 3 gives an efficient ($O(n)$-sample) algorithm that distinguishes $N(0, I_n)$ from $N(0, I_n)$ conditioned on any single convex set (which may not be symmetric). We observe here that no common generalization of these results, to mixtures of arbitrary convex sets, is possible with any finite sample complexity, no matter how large:

Theorem 33. Let $\mathrm{Mix}(\mathcal{P}_{\mathrm{conv}})$ denote the class of all convex combinations (mixtures) of distribu-
tions from $\mathcal{P}_{\mathrm{conv}}$, and let N be an arbitrarily large integer (N may depend on n, e.g. we may have
$N = 2^{2^n}$). For any $0 < \varepsilon < 1$, no N-sample algorithm can successfully distinguish between the
standard $N(0, I_n)$ distribution and an unknown distribution $\mathcal{D} \in \mathrm{Mix}(\mathcal{P}_{\mathrm{conv}})$ which is such that
$\mathrm{d}_{\mathrm{TV}}(N(0, I_n), \mathcal{D}) \geq \varepsilon$.

Proof sketch: The argument is essentially that of the the well-known $\Omega(\sqrt{L})$-sample lower bound for testing whether an unknown distribution over the discrete set $\{1, \ldots, L\}$ is uniform or $\Omega(1)$-far from uniform [GR00, BFR$^+$13]. Let $M = \omega(\frac{N^2}{1-\varepsilon})$, and consider a(n extremely fine) gridding of $\mathbb{R}^n$ into disjoint hyper-rectangles R each of which has $\mathrm{Vol}(R) = 1/M$. (For convenience we may think of M as being an n-th power of some integer, and of ε as being of the form $1/k$ for k an integer that divides M.) We note that for any set S that is a union of such hyper-rectangles, the distribution $N(0, I_n)|_S$ is an element of $\mathrm{Mix}(\mathcal{P}_{\mathrm{conv}})$.

Let $\boldsymbol{S}$ be the union of a random collection of exactly $(1-\varepsilon)M$ many of the hyper-rectangles R. We have $\mathrm{Vol}(\boldsymbol{S}) = (1-\varepsilon)M$, so $\mathrm{d}_{\mathrm{TV}}(N(0, I_n), N(0, I_n)|_{\boldsymbol{S}}) = \varepsilon$, and consequently a successful N-sample distinguishing algorithm as described in the theorem must be able to distinguish $N(0, I_n)$ from the distribution $\mathcal{D} = N(0, I_n)|_{\boldsymbol{S}}$. But it is easy to see that any $o(\sqrt{(1-\varepsilon)M})$-sample algorithm will, with $1-o(1)$ probability, receive a sample of points that all come from distinct hyper-rectangles; if this occurs, then the sample will be distributed precisely as a sample of the same size drawn from $N(0, I_n)$. $\qquad\square$