Testing Convex Truncation

Anindya De* Shivam Nadimpalli† Rocco A. Servedio‡

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 lower bounds on the sample complexity of distinguishing algorithms (computationally efficient or otherwise) for various classes of convex truncations; in some cases these lower bounds match our algorithms up to logarithmic or even constant factors.

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, KTZ19] 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 $\mathrm{poly}(n)$. 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

*University of Pennsylvania. Supported by NSF grants CCF-1910534, CCF-1926872, and CCF-2045128.

†Columbia University. Supported by NSF grants IIS-1838154, CCF-2106429, CCF-2211238, CCF-1763970, and CCF-2107187.

‡Columbia University. Supported by NSF grants IIS-1838154, CCF-2106429, and CCF-2211238.

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 1.1. (SYMMETRIC CONVEX TRUNCATIONS, INFORMAL STATEMENT) *There is an algorithm* SYMM-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 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 1.2. (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 1.3. (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 1.2 and Theorem 1.3, 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.

Finally, we also give a different and more efficient algorithm (as a function of n) algorithm for the special case in which the truncation set is the simplest possible convex set, namely a halfspace:

THEOREM 1.4. (HALFSPACE TRUNCATIONS) *There is an algorithm* LTF-DISTINGUISHER *which uses* $O(\sqrt{n}/\varepsilon^2 + (\log(1/\varepsilon))^2/\varepsilon^4)$ *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 halfspace with Gaussian volume at most* $1 - \varepsilon$.

[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.

1.1.2 Information-Theoretic Lower Bounds We show that the performance of LTF-Distinguisher is essentially best possible, by giving an $\tilde{\Omega}(\sqrt{n})$-sample lower bound for any algorithm that successfully distinguishes $N(0, I_n)$ from $N(0, I_n)|_K$ when K is an origin-centered halfspace:

Theorem 1.5. (Lower bound for halfspace truncations, informal theorem 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 halfspace, must use* $\Omega(\sqrt{n}/\log n)$ *samples.*

We also show that the same lower bound holds even for arguably the simplest class of symmetric convex sets, namely "slabs" (intersections of two parallel halfspaces):

Theorem 1.6. (Lower bound for symmetric slab truncations, informal theorem 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 symmetric slab with* $\mathrm{Vol}(K) = 1/2$*, must use* $\Omega(\sqrt{n}/\log n)$ *samples.*

Finally, we also show that the algorithm Symm-Convex-Distinguisher is essentially best possible for mixtures of symmetric convex sets:

Theorem 1.7. (Lower bound for mixtures of symmetric convex truncations) *Any algorithm which distinguishes (with probability at least* $9/10$*) between the standard* $N(0, I_n)$ *distribution and a distribution* $\mathcal{D}$ *which is a normal distribution conditioned on a mixture of symmetric convex sets must use* $\Omega(n)$ *samples, even if* $\mathcal{D}$ *is guaranteed to satisfy* $\mathrm{d_{TV}}(N(0, I_n), \mathcal{D}) = 1$.

1.2 Techniques Upper Bounds. To build intuition, let us first consider the case of a single symmetric convex body K. It is not difficult to see, using symmetry and convexity, that draws from $\mathcal{N}(0, I_n)|_K$ will on average lie closer to the origin than draws from $\mathcal{N}(0, I_n)$, so it is natural to use this as the basis for a distinguisher. We thus are led to consider our first estimator,

$$\mathbf{M} := \frac{1}{T}\sum_{i=1}^{T} \|\boldsymbol{x}^{(i)}\|^2, \tag{1.1}$$

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]

Finally, we use a more efficient tester for the special case of halfspaces to get an improved $O(\sqrt{n})$ bound for this case. The estimator we use is

$$\mathbf{N} := \left\| \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{x}^{(i)} \right\|^2, \tag{1.2}$$

[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].

namely the squared Euclidean length of the center of mass of the received samples. The intuition behind this estimator is that when K is a halfspace, the center of mass of $N(0, I_n)|_K$ is noticeably far from the origin, whereas when there is no conditioning and the samples are from the standard normal distribution, the empirical center of mass approaches the origin as the sample size grows large. Exploiting the rotational symmetry of the Gaussian, the analysis of this estimator reduces to the case when the halfspace is in the direction of e_1. With this, the rest of the analysis, is a somewhat long but ultimately straightforward calculation. We note that our estimator and its analysis bears some resemblance to the so-called Dicker's estimator [Dic14] used in estimating *goodness of fit* in noisy linear regression. In particular, [Dic14] gives a $O(\sqrt{n})$ sample complexity algorithm to estimate the variance of the noise given noisy labeled samples from a linear model. (We emphasize that while there is some resemblance in the calculations between our halfspace estimator and Dicker's estimator, the problems are substantially different and there is no reduction between the two problems.)

Lower Bounds. For both single halfspaces and "slabs" (symmetric convex sets that are the intersection of two parallel halfspaces), we use coupling arguments to reduce to the problem of distinguishing between two multivariate normal distributions with slightly different covariance matrices. A recent bound due to Devroye et al. [DMR20] on the total variation distance between multivariate normal distributions completes the proofs of those results.

Our most technically involved, and quantitatively strongest, lower bound is for normal distributions conditioned on a mixture of symmetric convex sets. We first show that $N(0, I_n)$ is indistinguishable, given cn samples, from $N(0, (1-\delta)I_n)$ for a suitable $\delta = \Theta(1/n)$. Next, we show that $N(0, (1-\delta)I_n)$ can be very accurately approximated (to variation distance $1/n^{\omega(1)}$) by a mixture P of $N(0, I_n)|_K$ distributions where each K is a ball intersected with an $n-1$-dimensional subspace. (The subspaces are Haar-uniform, and the radii of the balls are distributed according to a carefully designed distribution.) Finally, we adapt an idea from [RS09] and argue that $\sqrt{T}$ samples from P are indistinguishable from $\sqrt{T}$ samples from $\mathcal{D}$, where $\mathcal{D}$ is a subsampled version of the mixture $\mathcal{M}$ (a uniform mixture of T distributions sampled from the mixture). Given this a simple argument shows that $\mathcal{D}$ is both indistinguishable from $N(0, I_n)$ and statistically far from $N(0, I_n)$ as desired.

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 $\mathrm{poly}(n)$ 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 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 i^{th} 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 $\mathrm{Ball}(r)$ to denote the origin-centered ball of radius r in $\mathbb{R}^n$, i.e.,

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

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 and Chi-Squared 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

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

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

$$\mathrm{d_{TV}}(N(0, I_n)|_K, N(0, I_n)) = 1 - \mathrm{Vol}(K), \tag{2.3}$$

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 $\mathrm{Vol}(K) \leq 1 - \varepsilon$.

The squared norm $\|\boldsymbol{x}\|^2$ of $\boldsymbol{x} \sim N(0, I_n)$ is distributed according to the chi-squared distribution $\chi(n)^2$ with n degrees of freedom. The following tail bound for $\chi(n)^2$ (see [Joh01]) will be useful:

Lemma 2.1. (Tail bound for the chi-squared distribution) *Let $\boldsymbol{X} \sim \chi(n)^2$. Then we have*

$$\mathbf{Pr}\left[|\boldsymbol{X} - n| \geq tn\right] \leq e^{-(3/16)nt^2}, \quad \textit{for all } t \in [0, 1/2).$$

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 2.1. (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 $O(nm) + \mathrm{poly}(n \log(1/\delta))$ 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$\left(\{\boldsymbol{x}^{(j)}\}, \delta\right)$ *outputs a vector $\boldsymbol{L}$ such that*

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

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 2.1. (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 $\Omega(1/\mathrm{d_{TV}}(P, Q))$ many samples.*

The squared Hellinger distance provides a more refined lower bound on the sample complexity of this task (in fact it characterizes the sample complexity, though we will only need the lower bound). Recall that the squared Hellinger distance between two distributions P, Q over $\mathbb{R}^n$ that are absolutely continuous with respect to Lebesgue measure λ is

$$H^2(P, Q) = \int_{\mathbb{R}^n} \left(\sqrt{\frac{dP}{d\lambda}} - \sqrt{\frac{dQ}{d\lambda}}\right)^2 d\lambda.$$

Fact 2.2. (Squared Hellinger distance distinguishing lower bound, [BY02], Theorem 4.7) *Under the same conditions on P, Q and A as Fact 2.1, A must use at least $\Omega(1/H^2(P, Q))$ 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 2.1. (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*

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

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

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

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 2.1.

FACT 2.3. *For Lebesgue measurable $K \subseteq \mathbb{R}^n$, we have*

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,I_n)_K} \left[\boldsymbol{x}_i^2 \right] = 1 - \frac{\sqrt{2} \cdot \mathbf{Inf}_{e_i}[K]}{\mathrm{Vol}(K)}. \tag{2.4}$$

We also have that

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,I_n)_K} \left[\|\boldsymbol{x}\|^2 \right] = n - \frac{\sqrt{2} \cdot \mathbf{I}[K]}{\mathrm{Vol}(K)}. \tag{2.5}$$

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

PROPOSITION 2.2. (POINCARÉ FOR CONVEX INFLUENCES FOR SYMMETRIC CONVEX SETS) *For symmetric convex $K \subseteq \mathbb{R}^n$, we have*

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

The following variant of Proposition 2.2 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.

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

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

PROPOSITION 2.3. (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*

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

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 2.4. (BRASCAMP-LIEB INEQUALITY) *Let $\mathcal{D}$ be a probability distribution on $\mathbb{R}^n$ with density $e^{-V(x)}$. $\varphi_n(x)$ 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*

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

Vempala [Vem10] obtained a quantitative version of Proposition 2.4 in one dimension, which we state next. Note in particular that the following holds for non-centered Gaussians.

PROPOSITION 2.5. (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

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

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

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

$$e^{-\tau(x)}\varphi_n(x) = \mathbf{E}[A_{\boldsymbol{t}}(x)] \cdot \varphi_n(x) \tag{2.6}$$

where $A_t(x) = \mathbf{1}[e^{-\tau(x)} \geq t]$ and the expectation in (2.6) 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. ∎

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. Figure 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 3.1. *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

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

[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.

where the equality is Equation (2.5), the first inequality is Proposition 2.2 (Poincaré for convex influences for symmetric convex sets), and the second inequality holds because $\mathrm{Vol}(K) \le 1 - \varepsilon$. ∎

LEMMA 3.2. *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*

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

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 $f : \mathbb{R}^n \to \mathbb{R}$,

$$\mathop{\mathbf{Var}}_{\boldsymbol{x}\sim\mathcal{D}}[f(\boldsymbol{x})] \le \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{D}}[\|\nabla f(\boldsymbol{x})\|^2]. \tag{3.7}$$

Now, we may assume without loss of generality that $v = e_1$. Taking $f(x) = x_1$ in Equation (3.7), we get that

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

which finishes the proof. ∎

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

Proof. Taking $f(x) := \|x\|^2$ in Equation (3.7), we have that

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

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 3.2. ∎

3.2 An $O(n/\varepsilon^2)$-Sample Algorithm for Symmetric Convex Sets We recall Theorem 1.1:

THEOREM 3.1. (RESTATEMENT OF THEOREM 1.1) *For a sufficiently large constant $C > 0$, the algorithm* SYMM-CONVEX-DISTINGUISHER *(Figure 1) has the following performance guarantee: given any $\varepsilon > 0$ and access to independent samples from any unknown distribution $\mathcal{D} \in \mathcal{P}_{\mathrm{symm}}$, 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 $\mathrm{d_{TV}}(\mathcal{D}, N(0, I_n)) \ge \varepsilon$, 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 (1.1). We now turn to the proof of Theorem 3.1.

Proof. 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

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{D}_G}[\|\boldsymbol{x}\|^2] = n; \quad \mathop{\mathbf{Var}}_{\boldsymbol{x}\sim\mathcal{D}_G}[\|\boldsymbol{x}\|^2] = 3n. \tag{3.8}$$

On the other hand, if $\mathrm{d_{TV}}(\mathcal{D}, N(0, I_n)) \ge \varepsilon$ (equivalently, $\mathrm{Vol}(K) \le 1 - \varepsilon$), then using Lemma 3.1 and Lemma 3.3, it follows that

$$\mathbf{E}_{\boldsymbol{x}\sim\mathcal{D}_T}[\|\boldsymbol{x}\|^2] \le n - c\varepsilon; \quad \mathbf{Var}_{\boldsymbol{x}\sim\mathcal{D}_T}[\|\boldsymbol{x}\|^2] \le 4n. \tag{3.9}$$

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 $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(T)} \sim \mathcal{D}$.
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".

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

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

$$\begin{aligned} \mathbf{E}[\mathbf{M}] = n \quad \text{and} \quad \mathbf{Var}[\mathbf{M}] = \frac{3n}{T} \qquad &\text{when } \mathcal{D} = \mathcal{D}_G, \\ \mathbf{E}[\mathbf{M}] = n - c\varepsilon \quad \text{and} \quad \mathbf{Var}[\mathbf{M}] \leq \frac{4n}{T} \qquad &\text{when } \mathcal{D} = \mathcal{D}_T. \end{aligned}$$

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. ∎

3.3 An $O(n/\varepsilon^2)$-Sample Algorithm for Mixtures of Symmetric Convex Sets By extending the above analysis, we can show that Figure 1 succeeds for mixtures of (an arbitrary number of) symmetric convex sets as well. In particular, we have the following:

THEOREM 3.2. *For a sufficiently large constant $C > 0$,* SYMM-CONVEX-DISTINGUISHER *(Figure 1) has the following performance guarantee: given any $\varepsilon > 0$ and access to independent samples from any unknown distribution $\mathcal{D} \in \text{Mix}(\mathcal{P}_{\text{symm}})$, 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 $\mathrm{d}_{\text{TV}}(\mathcal{D}, N(0, I_n)) \geq \varepsilon$, then with probability at least $9/10$ the algorithm outputs "truncated."*

The following lemma, which characterizes the mean and variance of a distribution in $\text{Mix}(\mathcal{P}_{\text{symm}})$ in terms of the components of the mixture, will crucial to the proof of Theorem 3.2:

LEMMA 3.4. *Let $\mathcal{X}$ denote a distribution over Gaussians truncated by symmetric convex sets. Suppose $\mathcal{D}_{\mathcal{X}} \in \text{Mix}(\mathcal{P}_{\text{symm}})$ is the mixture of $N(0, I_n)|_{\mathbf{K}}$ for $\mathbf{K} \sim \mathcal{X}$. Let $\boldsymbol{a}_{\mathbf{K}}$ denote the random variable*

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

Then

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}} \left[\|\boldsymbol{x}\|^2\right] = \mathop{\mathbf{E}}_{\mathbf{K} \sim \mathcal{X}} [\boldsymbol{a}_{\mathbf{K}}], \tag{3.10}$$

$$\mathop{\mathbf{Var}}_{\boldsymbol{x} \sim \mathcal{D}_{\mathcal{X}}} \left[\|\boldsymbol{x}\|^2\right] \leq 4n + \mathop{\mathbf{Var}}_{\mathbf{K} \sim \mathcal{X}} [\boldsymbol{a}_{\mathbf{K}}]. \tag{3.11}$$

Proof. Note that Equation (3.10) follows from linearity of expectation and the definition of $\boldsymbol{a}_{\mathbf{K}}$. For Equation (3.11), note that for any symmetric convex set K, by definition of variance we have

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

where the inequality is by Lemma 3.3. By linearity of expectation, it now follows that

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

Combining with Equation (3.10), we get Equation (3.11). ∎

We are now ready to prove Theorem 3.2.

Proof. 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 3.1, we have that

$$\mathbf{E}[\mathbf{M}] = n, \quad \mathbf{Var}[\mathbf{M}] = \frac{3n}{T} \qquad \text{when } \mathcal{D} = \mathcal{D}_G. \tag{3.12}$$

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 $\mathbf{M} \ge n - \varepsilon/2$.

Now we analyze the case that $\mathcal{D} = \mathcal{D}_{\mathcal{X}}$ has $\mathrm{d_{TV}}(\mathcal{D}, N(0, I_n)) \ge \varepsilon$. From Lemma 3.4, it follows that in this case

$$\mathbf{E}[\mathbf{M}] = \mathop{\mathbf{E}}_{\mathbf{K}\sim\mathcal{X}}[\boldsymbol{a}_{\mathbf{K}}], \tag{3.13}$$

$$\mathbf{Var}[\mathbf{M}] = \frac{\mathbf{Var}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{X}}}\left[\|\boldsymbol{x}\|^2\right]}{T} \le \frac{4n}{T} + \frac{\mathbf{Var}_{\mathbf{K}\sim\mathcal{X}}[\boldsymbol{a}_{\mathbf{K}}]}{T}. \tag{3.14}$$

Next, observe that

$$\mathop{\mathbf{E}}_{\mathbf{K}\sim\mathcal{X}}[(n - \boldsymbol{a}_{\mathbf{K}})] \ge c \cdot \mathop{\mathbf{E}}_{\mathbf{K}\sim\mathcal{X}}[1 - \mathrm{Vol}(\mathbf{K})] \ge c \cdot \mathrm{d_{TV}}(\mathcal{D}, N(0, I_n)) \ge c\varepsilon, \tag{3.15}$$

where the first inequality uses Lemma 3.1 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 (3.14) that

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

By Equation (2.5) and Proposition 2.2, we have that $0 \le a_K \le n$ for any symmetric convex K. Thus, we can further upper bound the right hand side to obtain

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

Recalling from Equation (3.15) that $\mathbf{E}_{\mathbf{K}\sim\mathcal{X}}[n - \boldsymbol{a}_{\mathbf{K}}] \ge c\varepsilon$, a routine computation shows that for a sufficiently large constant C, we have

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

Equation (3.13) and Chebyshev's inequality now give that when $\mathcal{D} = \mathcal{D}_{\mathcal{X}}$, with probability at least $9/10$ we have $\mathbf{M} \le n - c\varepsilon/2$, completing the proof. ∎

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:

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

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 $\boldsymbol{L} :=$ MEAN-ESTIMATOR$\big(\{\boldsymbol{x}^{(j)}\}, 0.01\big)$.
3. Output "truncated" if either
   (a) $\mathbf{M} \le n - c\varepsilon/2$, or
   (b) $\|\boldsymbol{L}\|^2 \ge 0.05$;

   and output "un-truncated" otherwise.

Figure 2: Distinguisher for General Convex Sets

THEOREM 4.1. *There is an algorithm,* CONVEX-DISTINGUISHER *(Figure 2), with the following performance guarantee: Given any $\varepsilon > 0$ and access to independent samples from any unknown distribution $\mathcal{D} \in \mathcal{P}_{\text{conv}}$, the algorithm uses $O(n/\varepsilon^2)$ samples, runs in* $\text{poly}(n, 1/\varepsilon)$ *time, and*

1. *If $\mathcal{D} = N(0, I_n)$, then with probability at least* $9/10$ *the algorithm outputs "un-truncated;"*
2. *If $\mathrm{d}_{\text{TV}}(\mathcal{D}, N(0, I_n)) \ge \varepsilon$, then with probability at least* $9/10$ *the algorithm outputs "truncated."*

Note that the estimator $\mathbf{M}$ in Figure 2 is identical to the estimator $\mathbf{M}$ in Figure 1 to distinguish Gaussians restricted to (mixtures of) symmetric convex sets. As we will see, the analysis of Figure 1 via the Poincaré inequality for convex influences (cf. Proposition 2.2) 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 3):

- **Case 1:** When $\|\mu\| \gg 0$, we detect truncation via estimating the mean $\boldsymbol{L}$ using Proposition 2.1.
- **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 2.5).

4.1 Useful Preliminaries Below are two useful consequences of Vempala's quantitative one-dimensional Brascamp-Lieb inequality (Proposition 2.5) which will be useful in our analysis of Figure 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 4.1. *Given a convex set $K \subseteq \mathbb{R}^n$ and $v \in S^{n-1}$, if*

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

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

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

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

is a log-concave function (this is immediate from the Prékopa-Leindler inequality [Pré73, Lei72]). Furthermore, note that by Fact 2.3,

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

and so the result follows by Proposition 2.5. ∎

We also require a version of Proposition 2.5 for log-concave functions whose center of mass with respect to the standard normal distribution is not at the origin. Looking ahead, Proposition 4.2 will come in handy when analyzing Figure 2 for Gaussians restricted to convex sets with small inradius and with center of mass close to the origin.

PROPOSITION 4.2. *Let $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a one-dimensional log-concave function with*

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

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

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

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

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)} \left[\boldsymbol{x}^2 f(\boldsymbol{x})\right] \leq \left(1 + \mu^2 - \frac{1}{2\pi} e^{-(\varepsilon - \mu)^2}\right) \cdot \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)} [f(\boldsymbol{x})]. \tag{4.16}$$

We prove Proposition 4.2 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 2.5.

Proof. Note that it suffices to prove Equation (4.16). 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

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(-\mu,1)} \left[\widetilde{f}(\boldsymbol{x})\right] = \int_{\mathbb{R}} f(x + \mu)\varphi(x + \mu)\, dx = \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)} [f(\boldsymbol{x})]. \tag{4.17}$$

We also have that

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

where we made the substitution $y = x - \mu$. Therefore, by Proposition 2.5, we have that

$$\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(-\mu,1)} \left[\boldsymbol{x}^2 \widetilde{f}(\boldsymbol{x})\right] \leq \left(1 - \frac{1}{2\pi} e^{-(\varepsilon - \mu)^2}\right) \cdot \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(-\mu,1)} \left[\widetilde{f}(\boldsymbol{x})\right]. \tag{4.18}$$

However, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(-\mu,1)} \left[\boldsymbol{x}^2 \widetilde{f}(\boldsymbol{x})\right] &= \int_{\mathbb{R}} x^2 f(x + \mu)\varphi(x + \mu)\, dx \\
&= \int_{\mathbb{R}} (y - \mu)^2 f(y)\varphi(y)\, dy \\
&= \mathop{\mathbf{E}}_{\boldsymbol{y} \sim N(0,1)} \left[\boldsymbol{y}^2 f(\boldsymbol{y})\right] - \mathop{\mathbf{E}}_{\boldsymbol{y} \sim N(0,1)} \left[\mu^2 \cdot f(\boldsymbol{y})\right]. \qquad (4.19)
\end{aligned}$$

Equation (4.16) now follows from Equations (4.17) to (4.19). ∎

4.2 Proof of Theorem 4.1 We can now turn to the proof of Theorem 4.1.

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

$$\mathbf{E}[\mathbf{M}] = \frac{1}{T}\sum_{j=1}^{T} \mathbf{E}\left[\|\boldsymbol{x}^{(j)}\|^2\right] = \frac{1}{T}\sum_{j=1}^{T} n = n. \tag{4.20}$$

We also have that

$$\mathbf{Var}[\mathbf{M}] = \frac{1}{T^2}\sum_{j=1}^{T} \mathbf{Var}\left[\|\boldsymbol{x}^{(j)}\|^2\right] = \frac{1}{T}\left(\mathop{\mathbf{Var}}_{\boldsymbol{x}\sim N(0,I_n)}\left[\|\boldsymbol{x}\|^2\right]\right) = \frac{1}{T}\sum_{i=1}^{n} \mathop{\mathbf{Var}}_{\boldsymbol{x}_i\sim N(0,1)}\left[\boldsymbol{x}_i^2\right] = \frac{2n}{T}, \tag{4.21}$$

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 2.1 we have that

$$\|\boldsymbol{L}\|^2 \le 0.01 \tag{4.22}$$

with probability at least 0.99.

Next, suppose that $\mathcal{D} = N(0, I_n)_K$ for convex $K \subseteq \mathbb{R}^n$ with $\mathrm{d_{TV}}(\mathcal{D}, N(0, I_n)) \ge \varepsilon$. 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

$$\mathbf{E}[\mathbf{M}] = \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\boldsymbol{x}\|^2\right] \le n - \Omega(\varepsilon). \tag{4.23}$$

by Equation (2.3), Fact 2.3, and Proposition 2.3. By independence of the $\boldsymbol{x}^{(j)}$'s, we also have that

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

Note, however, that by Proposition 2.4 we have

$$\mathop{\mathbf{Var}}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\boldsymbol{x}\|^2\right] \le 4 \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\boldsymbol{x}\|^2\right] \qquad \text{and so} \qquad \mathbf{Var}[\mathbf{M}] \le \frac{4n}{T}, \tag{4.24}$$

where the second inequality follows from Equation (4.23). From Equations (4.20) and (4.23), 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 (4.21) and (4.24) 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

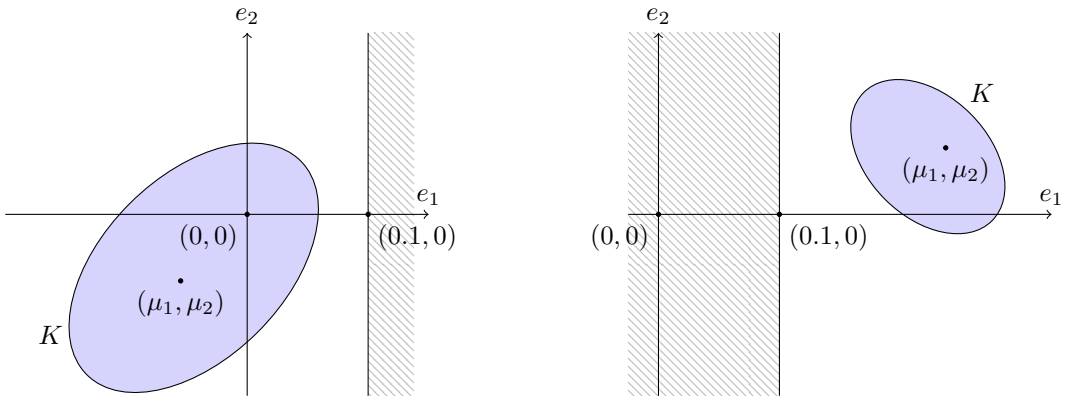
$$K \subseteq \{x \in \mathbb{R}^n : x_1 < 0.1\} \qquad \text{or} \qquad K \subseteq \{x \in \mathbb{R}^n : x_1 \ge 0.1\},$$

corresponding to (a) and (b) respectively in Figure 3. Writing μ for the center of mass of $\mathcal{D}$, i.e.

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

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 $\|\mu\|^2 \ge 0.06$, then we claim that Step 3(b) of Figure 2 will correctly output "truncated" with probability at least 99/100. Indeed, by the Brascamp-Lieb inequality, we have that $\mathrm{tr}(\Sigma) \le n$ where Σ is the



(a) (b)

Figure 3: The "small inradius" ($r_{\text{in}} \le 0.1$) setting in the analysis of Figure 2, with μ denoting the center of mass of K. Our estimator for (a) is $\text{Avg}(\|\boldsymbol{x}^{(j)}\|^2)$, whereas for (b) we simply estimate μ.

covariance matrix of $\mathcal{D}$, and so Proposition 2.1 implies that for a suitable choice of C, we will have $\|\mu - \boldsymbol{L}\| \le 0.001$ with probability at least 0.99, and hence $\|\boldsymbol{L}\|^2 \ge 0.05$.

Case 2. If $\|\mu\|^2 < 0.06$, then we will show that Figure 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" ($r_{\text{in}} \ge 0.1$) setting considered earlier. Recall that

$$\mathbf{E}[\mathbf{M}] = \sum_{i=1}^{n} \mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left[\boldsymbol{x}_i^2\right]. \tag{4.25}$$

For $i \in \{3,\dots,n\}$, as $\mu_i = 0$, we have by Proposition 4.1 that $\mathbf{Inf}_i[K] \ge 0$, and so

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left[\boldsymbol{x}_i^2\right] \le 1 \qquad \text{for } i \in \{3,\dots,n\} \tag{4.26}$$

by Fact 2.3.

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

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

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

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left[\boldsymbol{x}_i^2\right] = \frac{\mathbf{E}_{\boldsymbol{x}\sim N(0,1)}\left[\boldsymbol{x}^2 f_i(\boldsymbol{x})\right]}{\text{Vol}(K)}. \tag{4.27}$$

Since we have

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

it follows from Proposition 4.2 that

$$\frac{\mathbf{E}_{\boldsymbol{x}\sim N(0,1)}\left[\boldsymbol{x}^2 f_1(\boldsymbol{x})\right]}{\text{Vol}(K)} \le 1 + \mu_1^2 - \frac{1}{2\pi}e^{-(0.1-\mu_1)^2} \qquad \text{and} \qquad \frac{\mathbf{E}_{\boldsymbol{x}\sim N(0,1)}\left[\boldsymbol{x}^2 f_2(\boldsymbol{x})\right]}{\text{Vol}(K)} \le 1 + \mu_2^2 \tag{4.28}$$

(note that we used the fact that $\text{supp}(f_1) \subseteq (-\infty, 0.1]$ in the first inequality above). Combining Equations (4.27) and (4.28) and recalling that $\|\mu\|^2 < 0.06$, we get that

$$\mathop{\mathbf{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left[\boldsymbol{x}_1^2 + \boldsymbol{x}_2^2\right] \le 2 + \|\mu\|^2 - \frac{1}{2\pi}e^{-(0.1-\mu_1)^2} < 2.06 - \frac{1}{2\pi}e^{-(0.1+\sqrt{0.06})^2} < 1.95. \tag{4.29}$$

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

Output: "Un-truncated" or "truncated"

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

1. Let $T := \frac{C\sqrt{n}}{\varepsilon^2} + \frac{C(\log(1/\varepsilon))^2}{\varepsilon^4}$ and sample $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)} \sim \mathcal{D}$.
2. Set $\mathbf{N} := \left\| \frac{1}{T} \sum_{i=1}^{T} \boldsymbol{x}^{(i)} \right\|^2$.
3. Output "truncated" if $\mathbf{N} \geq \frac{n}{T} + c\varepsilon^2$, and "un-truncated" otherwise.

Figure 4: Distinguisher for LTFs

Combining Equations (4.25), (4.26) and (4.29), we get that

$$\mathbf{E}[\mathbf{M}] = \mathbf{E}\left[\|\boldsymbol{x}\|^2\right] \leq n - 0.05. \tag{4.30}$$

As in Equation (4.24), by the Brascamp-Lieb inequality (Proposition 2.4) we have that

$$\mathbf{Var}[\mathbf{M}] \leq \frac{4n}{T}, \tag{4.31}$$

and so by Equation (4.30), Equation (4.31) 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. □

5 An $O(\sqrt{n}/\varepsilon^2 + (\log(1/\varepsilon))^2/\varepsilon^4)$-Sample Algorithm for LTFs

In this section, we obtain a distinguisher for LTFs with better sample complexity than the distinguisher for convex sets ($(O(\sqrt{n})$ versus $O(n)$). More precisely, we have the following:

THEOREM 5.1. *There is an algorithm* LTF-DISTINGUISHER *(Figure 4) with the following performance guarantee: Given any $\varepsilon > 0$ and access to independent samples from any unknown distribution $\mathcal{D} \in \mathcal{P}_{\mathrm{LTF}}$, the algorithm uses $O(\sqrt{n}/\varepsilon^2 + (\log(1/\varepsilon))^2/\varepsilon^4)$ samples, and*

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

Looking ahead, we will show in Theorem 6.1 that this is essentially the best possible sample complexity. The intuition behind LTF-DISTINGUISHER starts with the easy observation that the center of mass of a LTF-truncated Gaussian is not at the origin. Employing Proposition 2.1 to estimate the center of mass, however, results in a $O(n)$-sample complexity, matching that of Figure 2. We therefore employ a different statistic, namely the *norm* of the empirical center of mass (cf. Equation (1.2)), which allows us to successfully distinguish LTF-truncated Gaussians using fewer samples.

5.1 Moments of Univariate Truncated Gaussians Our proof of Theorem 5.1 will rely on expressions for the first four moments of the truncated univariate standard Gaussian distribution. We start by introducing some convenient notation.

DEFINITION 5.1. *For $b \in \mathbb{R}$, the* Mills ratio $\mathrm{Mills}(b)$ *is*

$$\mathrm{Mills}(b) := \frac{1 - \Phi(b)}{\varphi(b)}.$$

NOTATION 5.1. *Let $K := [b, \infty) \subseteq \mathbb{R}$, and let $\boldsymbol{x} \sim N(0,1)|_K$. We write $\mathcal{M}_b(p)$ for the p^{th} raw moment of $\boldsymbol{x}$, i.e.*

$$\mathcal{M}_p(b) := \mathop{\mathbf{E}}_{\boldsymbol{x} \sim N(0,1)|_K}[\boldsymbol{x}^p].$$

When b is clear from context, we will simply write $\mathcal{M}_p$ instead.

The following expressions can be obtained via repeated integration by parts; alternatively, see [Orj14].

FACT 5.1. (MOMENTS OF TRUNCATED GAUSSIAN) *Let $K := [b, \infty) \subseteq \mathbb{R}$, and let $\boldsymbol{x} \sim N(0,1)|_K$. Then*

$$\mathcal{M}_1 = \frac{1}{\text{Mills}(b)}, \quad \mathcal{M}_2 = 1 + \frac{b}{\text{Mills}(b)}, \quad \mathcal{M}_3 = \frac{2 + b^2}{\text{Mills}(b)}, \quad \textit{and} \quad \mathcal{M}_4 = 3 + \frac{b^3 + 3b}{\text{Mills}(b)}.$$

Standard tail bounds on the Gaussian distribution (see e.g. [Wai19]) imply that

$$\frac{1}{x} \geq \text{Mills}(x) \geq \frac{x}{1 + x^2} \qquad \text{for } x \geq 0. \tag{5.32}$$

It is also easy to check that

$$\text{Mills}(x) \geq \sqrt{\frac{\pi}{2}} \qquad \text{for } x < 0. \tag{5.33}$$

We will also require the following estimate on the *Gaussian isoperimetric function* $\varphi \circ \Phi^{-1}(\cdot)$, a proof of which can be found in Proposition 27 of [DNS21a].

PROPOSITION 5.1. *For all $\varepsilon \in (0,1)$, we have*

$$\varphi \circ \Phi^{-1}(\varepsilon) \geq \sqrt{\frac{2}{\pi}} \min\{\varepsilon, 1 - \varepsilon\}.$$

5.2 Proof of Theorem 5.1 We can now proceed to the proof of Theorem 5.1.

Proof. Suppose first that $\mathcal{D} = N(0, I_n)$. Writing $\overline{\boldsymbol{x}}$ for the empirical mean $\frac{1}{T}\sum_{i=1}^T \boldsymbol{x}^{(i)}$, it is clear that $\mathbf{E}[\overline{\boldsymbol{x}}] = 0^n$ (the expected location of the empirical mean vector is the origin). Let $\mathbf{N}$ be the random variable which is the square of the Euclidean distance from $\overline{\boldsymbol{x}}$ to the origin, i.e. $\mathbf{N} = \|\overline{\boldsymbol{x}}\|^2$. Then

$$\mathbf{E}[\mathbf{N}] = \mathbf{E}\left[\sum_{i \in [n]} \overline{\boldsymbol{x}}_i^2\right] = \mathbf{E}\left[\sum_i \left(\sum_{j=1}^T \frac{\boldsymbol{x}_i^{(j)}}{T}\right)^2\right] = \mathbf{E}\left[\frac{1}{T^2}\sum_i \sum_{j_1, j_2} \boldsymbol{x}_i^{(j_1)} \boldsymbol{x}_i^{(j_2)}\right] = \frac{n}{T}, \tag{5.34}$$

and the variance of $\mathbf{N}$ is

$$\mathbf{Var}[\mathbf{N}] = \mathbf{Var}\left[\sum_{i=1}^n \overline{\boldsymbol{x}}_i^2\right] = n \cdot \mathbf{Var}\left[\overline{\boldsymbol{x}}_1^2\right] = n \cdot \left(\mathbf{E}\left[\overline{\boldsymbol{x}}_1^4\right] - \mathbf{E}\left[\overline{\boldsymbol{x}}_1^2\right]^2\right).$$

Note that $\mathbf{E}\left[\overline{\boldsymbol{x}}_1^2\right] = \frac{1}{T}$, and that

$$\mathbf{E}\left[\overline{\boldsymbol{x}}_1^4\right] = \mathbf{E}\left[\left(\frac{1}{T}\sum_{j=1}^T \boldsymbol{x}_1^{(j)}\right)^4\right] = \frac{1}{T^4}\sum_{i,j,k,l=1}^T \mathbf{E}\left[\boldsymbol{x}_1^{(i)} \boldsymbol{x}_1^{(j)} \boldsymbol{x}_1^{(k)} \boldsymbol{x}_1^{(l)}\right].$$

Note that we have a non-zero contribution to the sum in the final expression above when

- $i = j = k = l$ with $\mathbf{E}\left[\boldsymbol{x}_1^{(i)} \boldsymbol{x}_1^{(j)} \boldsymbol{x}_1^{(k)} \boldsymbol{x}_1^{(l)}\right] = 3$, contributing $3T$ to the sum; and

- i equals exactly one of the other three indices (there are three ways to choose which one) and the other two are equal; in this case $\mathbf{E}\left[\boldsymbol{x}_1^{(i)}\boldsymbol{x}_1^{(j)}\boldsymbol{x}_1^{(k)}\boldsymbol{x}_1^{(l)}\right] = 1$, so these contribute $3(T^2 - T)$ to the sum.

It follows that

$$\mathbf{Var}[\mathbf{N}] = n\left(\frac{3}{T^2} - \frac{1}{T^2}\right) = \frac{2n}{T^2}. \tag{5.35}$$

Now, suppose that $\mathcal{D} \in \mathcal{P}_{\mathrm{LTF}}$ with $\mathrm{d_{TV}}(\mathcal{D}, N(0, I_n)) \geq \varepsilon$. Thanks to spherical symmetry of the Gaussian distribution, we can assume that the halfspace is given by

$$K := \{x \in \mathbb{R}^n : x_1 \geq b\},$$

with $\Phi(b) \geq \varepsilon$. For this distribution, we have that

$$\mathbf{E}[\overline{\boldsymbol{x}}] = (\mathbf{E}[\boldsymbol{x}_1], 0, \ldots, 0) \qquad \text{where} \qquad \boldsymbol{x}_1 \sim N(0,1)_{[b,\infty)}.$$

As above let $\mathbf{N} = \|\overline{\boldsymbol{x}}\|^2$, and now we have that

$$\begin{aligned}
\mathbf{E}[\mathbf{N}] &= \mathbf{E}\left[\sum_{i\in[n]} \overline{\boldsymbol{x}}_i^2\right] \\
&= \mathbf{E}\left[\left(\sum_{j\in[T]} \frac{\boldsymbol{x}_1^{(j)}}{T}\right)^2\right] + \mathbf{E}\left[\sum_{i=2}^{n}\left(\sum_{j\in[T]} \frac{\boldsymbol{x}_i^{(j)}}{T}\right)^2\right] \\
&= \frac{1}{T^2}\sum_{j_1,j_2=1}^{T} \mathbf{E}\left[\boldsymbol{x}_1^{(j_1)}\boldsymbol{x}_1^{(j_2)}\right] + \frac{(n-1)}{T} \\
&= \left(\frac{\mathcal{M}_2(b)}{T} + \frac{(T-1)\mathcal{M}_1(b)^2}{T}\right) + \frac{(n-1)}{T} \qquad (5.36) \\
&= \frac{n}{T} + \mathcal{M}_1(b)^2 + \frac{\mathcal{M}_1(b)}{T}(b - \mathcal{M}_1(b)), \qquad (5.37)
\end{aligned}$$

where we used the expression for $\mathcal{M}_2(b)$ from Fact 5.1. When $\varepsilon \geq 0.5$, we have $b \geq 0$ and that

$$\mathbf{E}[\mathbf{N}] \geq \frac{n}{T} + \Omega(1) = \frac{n}{T} + \Omega\big(\varepsilon^2\big)$$

using Equation (5.32). On the other hand, when $\varepsilon < 0.5$, we have $b < 0$; in this case, we have from Fact 5.1 and Proposition 5.1 that

$$\mathcal{M}_1(b) \geq \frac{\varphi \circ \Phi^{-1}(\varepsilon)}{1-\varepsilon} = \Omega(\varepsilon).$$

(Note that we used the fact that $\varphi(x)$ is monotone increasing for $x \leq 0$.) Combining this with Equation (5.33) we also get that

$$\frac{\mathcal{M}_1(b)}{T}(b - \mathcal{M}_1(b)) \geq \frac{\Omega(\varepsilon)}{T}\left(b - \sqrt{\frac{2}{\pi}}\right).$$

Recalling that since $\varepsilon < 0.5$ we have $\Theta(\sqrt{\ln(1/\varepsilon)}) < b < 0$, it follows from Equation (5.37) and our choice of $T = Cn/\varepsilon^2$ that

$$\mathbf{E}[\mathbf{N}] \geq \frac{n}{T} + \Omega\big(\varepsilon^2\big). \tag{5.38}$$

Turning to the variance of $\mathbf{N}$, we have

$$\mathbf{Var}\left[\sum_{i\in[n]} \overline{\boldsymbol{x}}_i^2\right] = \mathbf{Var}\left[\overline{\boldsymbol{x}}_1^2\right] + \sum_{i=2}^{n} \mathbf{Var}\left[\overline{\boldsymbol{x}}_i^2\right] = \mathbf{Var}\left[\left(\sum_{j\in[T]} \frac{\boldsymbol{x}_1^{(j)}}{T}\right)^2\right] + \frac{2(n-1)}{T^2}. \tag{5.39}$$

Writing

$$\mathbf{Var}\left[\left(\sum_{j\in[T]}\frac{\boldsymbol{x}_1^{(j)}}{T}\right)^2\right] = \mathbf{E}\left[\left(\sum_{j\in[T]}\frac{\boldsymbol{x}_1^{(j)}}{T}\right)^4\right] - \mathbf{E}\left[\left(\sum_{j\in[T]}\frac{\boldsymbol{x}_1^{(j)}}{T}\right)^2\right]^2, \tag{5.40}$$

the contributions to the sum in the first term above, $\frac{1}{T^4}\sum_{i,j,k,\ell=1}^{T}\mathbf{E}\left[\boldsymbol{x}_1^{(i)}\boldsymbol{x}_1^{(j)}\boldsymbol{x}_1^{(k)}\boldsymbol{x}_1^{(\ell)}\right]$, break down as follows:

- When $i=j=k=\ell$ the expectation is $\mathcal{M}_4$; there are T ways for this to happen so this contributes $T\mathcal{M}_4$ to the sum;
- When three of the indices are equal and the fourth is distinct the expectation is $\mathcal{M}_3\mathcal{M}_1$; there are $4T(T-1)$ ways for this to happen so it contributes $4T(T-1)\mathcal{M}_3\mathcal{M}_1$ to the sum;
- When two indices equal each other and so do the other two the expectation is $\mathcal{M}_2^2$; there are $3T(T-1)$ ways for this to happen so it contributes $3T(T-1)\mathcal{M}_2^2$ to the sum;
- When two indices equal each other and the other two are two distinct values, the expectation is $\mathcal{M}_2\mathcal{M}_1^2$; there are $6T(T-1)(T-2)$ ways for this to happen so it contributes $6T(T-1)(T-2)\mathcal{M}_2\mathcal{M}_1^2$ to the sum;
- When all four indices are distinct the expectation is $\mathcal{M}_1^4$; there are $T(T-1)(T-2)(T-3)$ ways for this to happen so it contributes $T(T-1)(T-2)(T-3)\mathcal{M}_1^4$ to the sum.

In particular, we have

$$\mathbf{E}\left[\left(\sum_{j\in[T]}\frac{\boldsymbol{x}_1^{(j)}}{T}\right)^4\right] \le O\Big(\frac{\mathcal{M}_4}{T^3} + \frac{\mathcal{M}_3\mathcal{M}_1+\mathcal{M}_2^2}{T^2} + \frac{\mathcal{M}_2\mathcal{M}_1^2}{T}\Big) + \mathcal{M}_1^4. \tag{5.41}$$

We also have from Equation (5.36) and Fact 5.1 that

$$\begin{aligned}\mathbf{E}\left[\left(\sum_{j\in[T]}\frac{\boldsymbol{x}_1^{(j)}}{T}\right)^2\right]^2 &= \Big(\frac{\mathcal{M}_2}{T} + \frac{(T-1)\mathcal{M}_1^2}{T}\Big)^2 \\ &= \Big(\mathcal{M}_1^2 + \frac{(\mathcal{M}_2-\mathcal{M}_1^2)}{T}\Big)^2. \end{aligned} \tag{5.42}$$

It follows from Equations (5.40) to (5.42) that

$$\mathbf{Var}\left[\left(\sum_{j\in[T]}\frac{\boldsymbol{x}_1^{(j)}}{T}\right)^2\right] \le O\Big(\frac{\mathcal{M}_4}{T^3} + \frac{\mathcal{M}_3\mathcal{M}_1+\mathcal{M}_2^2}{T^2} + \frac{\mathcal{M}_2\mathcal{M}_1^2+\mathcal{M}_1^4}{T}\Big). \tag{5.43}$$

Recalling Equation (5.39), when $\varepsilon \ge 0.5$ (i.e. $b \ge 0$) we have that the RHS of Equation (5.43) is

$$(5.43) \le O\Big(\frac{\mathcal{M}_1^4}{T}\Big) = O\Big(\frac{b^4}{T}\Big) \quad \text{and so} \quad \mathbf{Var}\left[\sum_{i\in[n]}\overline{\boldsymbol{x}}_i^2\right] \le O\Big(\frac{n}{T^2}+\frac{b^4}{T}\Big) \le O\Big(\frac{n}{T^2}+\frac{\log(1/\varepsilon)^2}{T}\Big), \tag{5.44}$$

where we used $|b| \le \Theta(\sqrt{\log(1/\varepsilon)})$ for the last inequality. On the other hand, when $\varepsilon < 0.5$ (i.e. $b < 0$) we have that the RHS of Equation (5.43) is

$$O\Big(\frac{1}{T}+\frac{1}{T^2}\Big) \qquad \text{and so} \qquad \mathbf{Var}\left[\sum_{i\in[n]}\overline{\boldsymbol{x}}_i^2\right] \le O\Big(\frac{n}{T^2}+\frac{1}{T}\Big).$$

Summarizing Equation (5.34), Equation (5.35), Equation (5.38) and Equation (5.44), we have that

$$\mathbf{E}[\mathbf{N}] = \frac{n}{T} \quad \text{and} \qquad \mathbf{Var}[\mathbf{N}] = \frac{2n}{T^2} \qquad \text{when } \mathcal{D} = N(0, I_n),$$

$$\mathbf{E}[\mathbf{N}] \geq \frac{n}{T} + \Omega(\varepsilon^2) \quad \text{and} \qquad \mathbf{Var}[\mathbf{N}] \leq O\left(\frac{n}{T^2} + \frac{\log(1/\varepsilon)^2}{T}\right) \qquad \text{when } \mathrm{d_{TV}}(\mathcal{D}, N(0, I_n)) \geq \varepsilon.$$

The correctness of Figure 4 follows by our choice of $T = C\sqrt{n}/\varepsilon^2$ (for a sufficiently large constant C) and Chebyshev's inequality. This finishes the proof. ∎

6 Lower Bounds

In Section 6.1, we present a $\widetilde{\Omega}(\sqrt{n})$ lower-bound for distinguishing Gaussians truncated to LTFs, followed by a $\widetilde{\Omega}(n)$ lower bound for distinguishing Gaussians truncated to symmetric convex sets in Section 6.2. (The specific symmetric convex set we use for the latter is the slab.) Finally, in Section 6.3, we present a $\Omega(n)$ lower bound for distinguishing a mixture of Gaussians truncated to symmetric convex sets.

6.1 A $\tilde{\Omega}(\sqrt{n})$-Sample Lower Bound for Halfspaces Our first lower bound, Theorem 6.1, shows that $\Omega(\sqrt{n}/\log n)$ samples are needed to distinguish $N(0, I_n)$ from $N(0, I_n)|_K$ where K is an unknown halfspace whose separating hyperplane passes through the origin. Some of the ideas in this lower bound will recur in our lower bound for symmetric "slabs" (intersections of two parallel halfspaces) given in Section 6.2, but by taking advantage of the fact that origin-centered halfspaces are odd functions, we can sidestep some aspects of the argument that arise for slabs; hence we give a self-contained proof for halfspaces below.

THEOREM 6.1. *Let A be any algorithm which is given access to samples from an unknown distribution $\mathcal{D}$ over $\mathbb{R}^n$ and has the following performance guarantee:*

1. *If $\mathcal{D} = N(0, I_n)$, then with probability at least $2/3$ the algorithm outputs "un-truncated";*
2. *If $\mathcal{D} = N(0, I_n)|_K$ where K is an unknown zero-threshold LTF, then with probability at least $2/3$ the algorithm outputs "truncated."*

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

Proof. We consider two different distributions, $\mathcal{D}_1$ and $\mathcal{D}_2$, each of which is a distribution over sequences of m points in $\mathbb{R}^n$.

1. A draw of $\overline{\boldsymbol{g}} = (\boldsymbol{g}^1, \ldots, \boldsymbol{g}^m)$ from $\mathcal{D}_1$ is obtained by having each $\boldsymbol{g}^i \in \mathbb{R}^n$ be distributed independently according to $N(0, I_n)$.
2. A draw from $\mathcal{D}_2$ is obtained by independently drawing $\boldsymbol{g}, \boldsymbol{g}^1, \ldots, \boldsymbol{g}^m$ from $N(0, I_n)$ and outputting $(\mathrm{sign}(\boldsymbol{g} \cdot \boldsymbol{g}^1)\boldsymbol{g}^1, \ldots, \mathrm{sign}(\boldsymbol{g} \cdot \boldsymbol{g}^m)\boldsymbol{g}^m)$.

Observe that a draw from $\mathcal{D}_2$ is distributed as a sample of m independent draws from $N(0, I_n)|_{\mathbf{K}}$ where $\mathbf{K} = \{x \in \mathbb{R}^n : \boldsymbol{g} \cdot x \geq 0\}$ is a random zero-threshold LTF defined by a normal vector $\boldsymbol{g}$ that is drawn from $N(0, I_n)$ (here is where we are using that any zero-threshold LTF is an odd function). Thus to prove Theorem 6.1 it suffices to show that any algorithm that determines (with correctness probability at least $2/3$) whether an m-element sample came from $\mathcal{D}_1$ or $\mathcal{D}_2$ must have $m = \Omega(\sqrt{n}/\log n)$. This is an immediate consequence of the following claim:

CLAIM 6.1. *There is a universal constant $c > 0$ such that for $m = c\sqrt{n}/\log n$, we have $\mathrm{d_{TV}}(\mathcal{D}_1, \mathcal{D}_2) \leq 0.1$.*

Proof of Claim 6.1. We first observe that a draw $(\boldsymbol{g}^1, \ldots, \boldsymbol{g}^m) \sim \mathcal{D}_1$ is distributed identically to $(\boldsymbol{u}_1\boldsymbol{g}^1, \ldots, \boldsymbol{u}_m\boldsymbol{g}^m)$ where each $\boldsymbol{u}_i$ is an independent uniform $\{-1, +1\}$ random variable. Next, we observe that

$$\begin{aligned} \mathrm{d_{TV}}(\mathcal{D}_1, \mathcal{D}_2) &= \mathrm{d_{TV}}((\boldsymbol{u}_1\boldsymbol{g}^1, \ldots, \boldsymbol{u}_m\boldsymbol{g}^m), (\mathrm{sign}(\boldsymbol{g} \cdot \boldsymbol{g}^1)\boldsymbol{g}^1, \ldots, \mathrm{sign}(\boldsymbol{g} \cdot \boldsymbol{g}^m)\boldsymbol{g}^m)) \\ &\leq \mathop{\mathbf{E}}_{\overline{\boldsymbol{g}}}[\mathrm{d_{TV}}((\boldsymbol{u}_1\boldsymbol{g}^1, \ldots, \boldsymbol{u}_m\boldsymbol{g}^m), (\mathrm{sign}(\boldsymbol{g} \cdot \boldsymbol{g}^1)\boldsymbol{g}^1, \ldots, \mathrm{sign}(\boldsymbol{g} \cdot \boldsymbol{g}^m)\boldsymbol{g}^m))] \quad (6.45) \\ &= \mathop{\mathbf{E}}_{\overline{\boldsymbol{g}}}[\mathrm{d_{TV}}((\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m), (\mathrm{sign}(\boldsymbol{g} \cdot \boldsymbol{g}^1), \ldots, \mathrm{sign}(\boldsymbol{g} \cdot \boldsymbol{g}^m)))] \quad (6.46) \end{aligned}$$

where $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$ is uniform over $\{-1,+1\}^m$; it should be noted that in Equation (6.45) the randomness in the two random variables whose variation distance is being considered is only over $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ and $\boldsymbol{g}$; and the equality between (6.45) and (6.46) holds because the $N(0, I_n)$ distribution has no atoms of mass. So it remains to bound the expected variation distance between the two m-bit strings $\boldsymbol{u}$ and $(\mathrm{sign}(\boldsymbol{g}\cdot\boldsymbol{g}^1), \ldots, \mathrm{sign}(\boldsymbol{g}\cdot\boldsymbol{g}^m))$, where the expectation is over $\overline{\boldsymbol{g}} = (\boldsymbol{g}^1, \ldots, \boldsymbol{g}^m)$. We may view a uniform $\boldsymbol{u} \sim \{-1,+1\}^m$ as being distributed according to $(\mathrm{sign}(\boldsymbol{g}\cdot e^1), \ldots, \mathrm{sign}(\boldsymbol{g}\cdot e^m))$, so it suffices to show that with probability at least 0.95 over $\overline{\boldsymbol{g}}$, we have

$$\mathrm{d_{TV}}((\mathrm{sign}(\boldsymbol{g}\cdot e^1), \ldots, \mathrm{sign}(\boldsymbol{g}\cdot e^m)), (\mathrm{sign}(\boldsymbol{g}\cdot\boldsymbol{g}^1), \ldots, \mathrm{sign}(\boldsymbol{g}\cdot\boldsymbol{g}^m))) \leq 0.05. \tag{6.47}$$

We say that a specific outcome $(g^1, \ldots, g^m) \in (\mathbb{R}^n)^m$ of $(\boldsymbol{g}^1, \ldots, \boldsymbol{g}^m) \sim (N(0, I_n))^m$ is *bad* if there is a pair $i \neq j$ such that the normalized inner product $\frac{|g^i \cdot g^j|}{\|g^i\| \cdot \|g^j\|}$ is greater than $C\sqrt{\frac{\log n}{n}}$, and otherwise we say that $(g^1, \ldots, g^m)$ is *good*. Intuitively, the tuple $(g^1, \ldots, g^m)$ is bad if the m vectors $g^1, \ldots, g^m$ are not "pairwise nearly orthogonal"; as we now show, this is very unlikely to happen. Fix any $1 \leq i \neq j \leq m$; we have that $\frac{\boldsymbol{g}^i}{\|\boldsymbol{g}^i\|}$ and $\frac{\boldsymbol{g}^j}{\|\boldsymbol{g}^j\|}$ are independent uniform vectors on the n-dimensional unit sphere $\mathbb{S}^{n-1}$, and so the distribution of $\frac{|\boldsymbol{g}^i \cdot \boldsymbol{g}^j|}{\|\boldsymbol{g}^i\| \cdot \|\boldsymbol{g}^j\|} = \left| \frac{\boldsymbol{g}^i}{\|\boldsymbol{g}^i\|} \cdot \frac{\boldsymbol{g}^j}{\|\boldsymbol{g}^j\|} \right|$ is the same as the distribution of $|\boldsymbol{v}_1|$ for a uniform random unit vector $\boldsymbol{v} \sim \mathbb{S}^{n-1}$. It is well known (see e.g. Lemma 2.2 of [Bal97]) that $\mathbf{Pr}_{\boldsymbol{v}\sim\mathbb{S}^{n-1}}[|\boldsymbol{v}_1| > C\sqrt{\log(n)/n}]$ is at most $1/n^2$ for a suitable choice of the absolute constant C. Thus by a union bound over all $m^2 = o(n)$ many pairs $i \neq j$, we have that

$$\mathbf{Pr}[(\boldsymbol{g}^1, \ldots, \boldsymbol{g}^m) \text{ is bad}] \leq \frac{m^2}{n^2} < 0.05.$$

Hence to establish (6.47) (and hence Claim 6.1 and Theorem 6.1) it suffices to prove the following:

Claim 6.2. *Fix a good $(g^1, \ldots, g^m) \in (\mathbb{R}^n)^m$, where $m = c\sqrt{n}/\log n$. Then for $\boldsymbol{g} \sim N(0, I_n)$ we have that*

$$\mathrm{d_{TV}}((\mathrm{sign}(e^1 \cdot \boldsymbol{g}), \ldots, \mathrm{sign}(e^m \cdot \boldsymbol{g})), (\mathrm{sign}(g^1 \cdot \boldsymbol{g}), \ldots, \mathrm{sign}(g^m \cdot \boldsymbol{g}))) \leq 0.05. \tag{6.48}$$

To prove Claim 6.2 we first observe that (6.48) is equivalent to

$$\mathrm{d_{TV}}((\mathrm{sign}(e^1 \cdot \boldsymbol{g}), \ldots, \mathrm{sign}(e^m \cdot \boldsymbol{g})), (\mathrm{sign}(g'^1 \cdot \boldsymbol{g}), \ldots, \mathrm{sign}(g'^m \cdot \boldsymbol{g}))) \leq 0.05 \tag{6.49}$$

where we define $g'^i := \frac{g^i}{\|g^i\|}$ to be g^i normalized to unit length. Next we recall a recent upper bound on the total variation distance between n-dimensional Gaussians due to Devroye, Mehrabian and Reddad:

Theorem 6.2. (Theorem 1.1 of [DMR20]) *Let Σ_1, Σ_2 be two $m \times m$ positive definite covariance matrices and let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $\Sigma_1^{-1}\Sigma_2 - I_m$. Then*

$$\mathrm{d_{TV}}(N(0^m, \Sigma_1), N(0^m, \Sigma_2)) \leq \frac{3}{2} \cdot \min\left\{1, \sqrt{\lambda_1^2 + \cdots + \lambda_m^2}\right\}$$

To apply Theorem 6.2 we take Σ_1 to be the identity matrix I_m and Σ_2 to be the $m \times m$ matrix whose (i,j) entry is $g'^i \cdot g'^j$, so a draw from $N(0, \Sigma_1)$ is distributed as $(e_1 \cdot \boldsymbol{g}, \ldots, e_m \cdot \boldsymbol{g})$ and a draw from $N(0, \Sigma_2)$ is distributed as $(g'^1 \cdot \boldsymbol{g}, \ldots, g'^m \cdot \boldsymbol{g})$. Applying the data processing inequality for total variation distance (see e.g. Proposition B.1 of [DDO+13], it follows that the LHS of (6.49) is at most $\mathrm{d_{TV}}(N(0^m, \Sigma_1), N(0^m, \Sigma_2))$; we proceed to upper bound the RHS of Theorem 6.2.

Let A denote the matrix $\Sigma_1^{-1}\Sigma_2 - I_m$, so $\lambda_1^2, \ldots, \lambda_m^2$ are the eigenvalues of A^2 and we have $\sqrt{\lambda_1^2 + \cdots + \lambda_m^2} = \sqrt{\mathrm{tr}(A^2)}$. Since each g'^i is a unit vector the matrix A has zero entries on the diagonal, and A has off-diagonal entries that are each at most $C\sqrt{\frac{\log n}{n}}$ in magnitude (because $(g^1, \ldots, g^m)$ is good). Hence any diagonal element of A^2 is at most $\frac{C^2 m \log n}{n}$ in magnitude, and consequently $\mathrm{tr}(A^2) \leq \frac{C^2 m^2 \log n}{n}$, which is at most $\frac{1}{900}$ for a suitable choice of $m = \Theta(\sqrt{n/\log n})$. Having $\mathrm{tr}(A^2) \leq \frac{1}{900}$ gives that $\frac{3}{2} \cdot \min\left\{1, \sqrt{\lambda_1^2 + \cdots + \lambda_m^2}\right\} \leq 0.05$ as required, and the proofs of Claim 6.2, Claim 6.1 and Theorem 6.1 are complete. ∎

6.2 A $\tilde{\Omega}(\sqrt{n})$-Sample Lower Bound for Symmetric Convex Bodies (Slabs) Given a unit vector $v \in \mathbb{S}^{n-1}$ and a positive value $r > 0$, we say that the *symmetric slab of width* $2r$ *in direction* v is the symmetric convex set $\mathrm{Slab}_{v,r} := \{x \in \mathbb{R}^n : |v \cdot x| \leq r\}$. Our second lower bound shows that $\Omega(\sqrt{n}/\log n)$ samples are needed to distinguish $N(0, I_n)$ from $N(0, I_n)|_K$ where K is an unknown symmetric slab that is promised to have volume roughly 1/2: œ

THEOREM 6.3. *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 $\mathcal{D} = N(0, I_n)|_K$ where K is an unknown symmetric slab with $\mathrm{Vol}(K) \in [0.49, 0.51]$, then with probability at least 9/10 the algorithm outputs "truncated."*

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

Proof. Let $\psi : \mathbb{R} \to \{0, 1\}$ be the function $\psi(t) = \mathbf{1}[|t| \leq c]$ where $c \approx 0.68$ is the real number satisfying $\mathbf{Pr}_{\boldsymbol{g} \sim N(0,1)}[|\boldsymbol{g}| \leq c] = 1/2$. We define a distribution $\mathcal{K}$ over symmetric slabs as follows: to draw a symmetric slab $\mathbf{K} \sim \mathcal{K}$, first draw a standard Normal vector $\boldsymbol{g} \sim N(0, I_n)$, and let $\mathbf{K} = \mathbf{K}_{\boldsymbol{g}}$ be the symmetric slab $\mathbf{K} = \{x \in \mathbb{R}^n : \psi(\frac{\boldsymbol{g}}{\sqrt{n}} \cdot x) = 1\}$.

We observe that given a particular outcome g of $\boldsymbol{g}$, the width of the slab $K = \mathbf{K}_g$ is $\frac{2c\sqrt{n}}{\|g\|}$, and hence the volume $\mathrm{Vol}(K)$ is $\mathbf{Pr}_{\boldsymbol{x} \sim N(0,1)}[|\boldsymbol{x}| < \frac{c\sqrt{n}}{\|g\|}]$. Standard tail bounds on the chi-distribution (which is the distribution of $\|\boldsymbol{g}\|$ for $\boldsymbol{g} \sim N(0, I_n)$) imply that $\mathbf{Pr}_{\mathbf{K} \sim \mathcal{K}}[\mathrm{Vol}(\mathbf{K}) \notin [0.49, 0.51]]$ is extremely small, in particular at most $1/n^{\omega(1)}$. Hence to prove Theorem 6.3, it suffices to prove an $\Omega(\sqrt{n}/\log n)$ lower bound on the sample complexity of any algorithm A that outputs "un-truncated" with probability at least 9/10 if it is given samples from $N(0, I_n)$ and outputs "truncated" with probability at least 91/100 if it is given samples from $N(0, I_n)|_{\mathbf{K}}$ where $\mathbf{K} \sim \mathcal{K}$. We do this in the rest of the proof.

We consider two different distributions, $\mathcal{D}_1$ and $\mathcal{D}_2$, each of which is a distribution over sequences of $m = c\sqrt{n}/\log n$ points in $\mathbb{R}^n$:

1. A draw of $\overline{\boldsymbol{x}} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m)$ from $\mathcal{D}_1$ is obtained by having each $\boldsymbol{x}^i \in \mathbb{R}^n$ be distributed independently according to $N(0, I_n)$.
2. A draw of $\overline{\boldsymbol{x}} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m)$ from $\mathcal{D}_2$ is obtained by drawing (once and for all) a set $\mathbf{K} = \mathbf{K}_{\boldsymbol{g}} \sim \mathcal{K}$ (so $\boldsymbol{g} \sim N(0, I_n)$) and having each $\boldsymbol{x}^i$ be distributed independently according to $N(0, I_n)|_{\mathbf{K}}$, i.e. each $\boldsymbol{x}^i$ is drawn from $N(0, I_n)$ conditioned on satisfying $\psi(\frac{\boldsymbol{g}}{\sqrt{n}} \cdot x) = 1$.

We further define two more distributions $\mathcal{D}_1'$ and $\mathcal{D}_2'$, each of which is a distribution over sequences of $3m$ points in $\mathbb{R}^n$:

1. A draw of $\overline{\boldsymbol{x}} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{3m})$ from $\mathcal{D}_1'$ is obtained as follows: $\boldsymbol{x}^i$ is taken to be $\boldsymbol{b}_i \boldsymbol{g}^i$ where each $\boldsymbol{b}_i$ is an independent uniform draw from $\{0, 1\}$ and each $\boldsymbol{g}^i$ is an independent draw from $N(0, I_n)$.
2. A draw of $\overline{\boldsymbol{z}} = (\boldsymbol{z}^1, \ldots, \boldsymbol{z}^{3m})$ from $\mathcal{D}_2'$ is obtained as follows: once and for all draw $\boldsymbol{g} \sim N(0, I_n)$, and take $\boldsymbol{z}^i$ to be $\boldsymbol{g}^i \cdot \psi(\frac{\boldsymbol{g}}{\sqrt{n}} \cdot \boldsymbol{g}^i)$ where each $\boldsymbol{g}^i$ is an independent draw from $N(0, I_n)$.

Suppose that algorithm A successfully distinguishes between $\mathcal{D}_1$ and $\mathcal{D}_2$, i.e. it outputs "un-truncated" with probability at least 9/10 when run on a draw from $\mathcal{D}_1$ and outputs "truncated" with probability at least 91/100 when run on a draw from $\mathcal{D}_2$. We claim that then there is an algorithm A' which successfully distinguishes between $\mathcal{D}_1'$ and $\mathcal{D}_2'$, i.e. it outputs "un-truncated" with probability at least 89/100 when run on a draw from $\mathcal{D}_1'$ (recall that this is a sequence of $3m$ points in $\mathbb{R}^n$) and outputs "truncated" with probability at least 90/100 when run on a draw from $\mathcal{D}_2'$. This algorithm A' simply takes the first m nonzero points in its $3m$-point input sequence and uses them as input to A. (We may suppose that A' fails if there are fewer than m nonzero points in the $3m$-point sample; it is easy to see that whether the input to A' is drawn from $\mathcal{D}_1'$ or $\mathcal{D}_2'$, the probability of failure is at most $1/n^{\omega(1)}$ and hence is negligible.) If the input sequence is a draw from $\mathcal{D}_1'$ then the input that A' gives to

A is distributed according to $\mathcal{D}_1$, and if the input is a draw from $\mathcal{D}_2'$ then the input sequence that A' gives to A is distributed according to $\mathcal{D}_2$ (note that we may safely ignore the probability-0 event that some $\boldsymbol{g}^i$ drawn from $N(0, I_n)$ is the zero vector).

Given the above claim, to prove Theorem 6.3 it suffices to show that any algorithm A' with the performance guarantee described above must have $3m = \Omega(\sqrt{n}/\log n)$. This follows from the statement

$$\text{If } m = c\sqrt{n}/\log n, \text{ then } \quad \mathrm{d_{TV}}(\mathcal{D}_1', \mathcal{D}_2') \leq 0.01. \tag{6.50}$$

The rest of the proof establishes (6.50).

Consider a coupling of the two distributions $\mathcal{D}_1'$ and $\mathcal{D}_2'$ as follows. A draw from the joint coupled distribution of $(\mathcal{D}_1', \mathcal{D}_2')$ is generated in the following way:

- Let $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{3m})$ be a uniform random string from $\{0,1\}^{3m}$.
- Let $\boldsymbol{g}, \boldsymbol{g}^1, \ldots, \boldsymbol{g}^m$ be $3m+1$ independent draws from $N(0, I_n)$.
- The draw from the joint coupled distribution is $(\overline{\boldsymbol{x}}, \overline{\boldsymbol{z}})$ where

$$\text{for each } i = 1, \ldots, 3m, \quad \boldsymbol{x}^i = \boldsymbol{b}_i \boldsymbol{g}^i \quad \text{and} \quad \boldsymbol{z}^i = \psi\left(\frac{\boldsymbol{g}}{\sqrt{n}} \cdot \boldsymbol{g}^i\right) \boldsymbol{g}^i.$$

It is easily verified that this is indeed a valid coupling of $\mathcal{D}_1'$ and $\mathcal{D}_2'$. Writing $\overline{\boldsymbol{g}}$ to denote $(\boldsymbol{g}^1, \ldots, \boldsymbol{g}^m)$, by the same reasoning as in Equation (6.45) and Equation (6.46), we have that

$$\begin{aligned}
\mathrm{d_{TV}}(\mathcal{D}_1', \mathcal{D}_2') &= \mathrm{d_{TV}}((\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{3m}), (\boldsymbol{z}^1, \ldots, \boldsymbol{z}^{3m})) \\
&= \mathrm{d_{TV}}\left((\boldsymbol{b}_1 \boldsymbol{g}^1, \ldots, \boldsymbol{b}_{3m} \boldsymbol{g}^{3m}), \left(\psi\left(\frac{\boldsymbol{g}}{\sqrt{n}} \cdot \boldsymbol{g}^1\right) \boldsymbol{g}^1, \psi\left(\frac{\boldsymbol{g}}{\sqrt{n}} \cdot \boldsymbol{g}^{3m}\right) \boldsymbol{g}^{3m}\right)\right) \\
&\leq \mathop{\mathbf{E}}_{\overline{\boldsymbol{g}}}\left[\mathrm{d_{TV}}\left((\boldsymbol{b}_1 \boldsymbol{g}^1, \ldots, \boldsymbol{b}_{3m} \boldsymbol{g}^{3m}), \left(\psi\left(\frac{\boldsymbol{g}}{\sqrt{n}} \cdot \boldsymbol{g}^1\right) \boldsymbol{g}^1, \psi\left(\frac{\boldsymbol{g}}{\sqrt{n}} \cdot \boldsymbol{g}^{3m}\right) \boldsymbol{g}^{3m}\right)\right)\right] \\
&\leq \mathop{\mathbf{E}}_{\overline{\boldsymbol{g}}}\left[\mathrm{d_{TV}}\left((\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{3m}), \left(\psi\left(\frac{\boldsymbol{g}}{\sqrt{n}} \cdot \boldsymbol{g}^1\right), \ldots, \psi\left(\frac{\boldsymbol{g}}{\sqrt{n}} \cdot \boldsymbol{g}^{3m}\right)\right)\right)\right] \qquad (6.51)
\end{aligned}$$

It remains to show that if $m = c\sqrt{n}/\log n$ then (6.51) ≤ 0.01.

We observe that for $\boldsymbol{g} \sim N(0, I_n)$, we have that $\mathbf{Pr}[\psi(e_i \cdot \boldsymbol{g}) = 1] = \mathbf{Pr}[\psi(\boldsymbol{g}_i) = 1] = 1/2$. Since the random variables $\{\boldsymbol{g}_i\}_{i=1,\ldots,3m}$ are independent, we have that the joint distribution of $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{3m})$ is identical to the joint distribution of $(\psi(e_1 \cdot \boldsymbol{g}), \ldots, \psi(e_{3m} \cdot \boldsymbol{g}))$, namely, both distributions are uniform over $\{0,1\}^{3m}$. Thus we have that

$$(6.51) = \mathop{\mathbf{E}}_{\overline{\boldsymbol{g}}}\left[\mathrm{d_{TV}}\left((\psi(e_1 \cdot \boldsymbol{g}), \ldots, \psi(e_{3m} \cdot \boldsymbol{g})), \left(\psi\left(\frac{\boldsymbol{g}^1}{\sqrt{n}} \cdot \boldsymbol{g}\right), \ldots, \psi\left(\frac{\boldsymbol{g}^{3m}}{\sqrt{n}} \cdot \boldsymbol{g}\right)\right)\right)\right]. \tag{6.52}$$

The rest of the argument proceeds similarly to the proof of Claim 6.1 following (6.47): briefly, for most outcomes $(g^1, \ldots, g^{3m})$ of $\boldsymbol{g}^1, \ldots, \boldsymbol{g}^{3m}$, the covariance structure of the two $3m$-dimensional Gaussians $(e_1 \cdot \boldsymbol{g}, \ldots, e_{3m} \cdot \boldsymbol{g})$ and $(\frac{g^1}{\sqrt{n}} \cdot \boldsymbol{g}, \ldots, \frac{g^{3m}}{\sqrt{n}} \cdot \boldsymbol{g})$ are similar enough that we can bound the variation distance between those two Gaussians, which implies a bound on the variation distance between $(\psi(e_1 \cdot \boldsymbol{g}), \ldots, \psi(e_{3m} \cdot \boldsymbol{g}))$ and $\left(\psi\left(\frac{g^1}{\sqrt{n}} \cdot \boldsymbol{g}\right), \ldots, \psi\left(\frac{g^m}{\sqrt{n}} \cdot \boldsymbol{g}\right)\right)$. Details are given below.

We require a slight refinement of the notion of a "good" tuple of vectors from Section 6.1 which now takes into account the lengths of the vectors as well as the angles between them. We say that a specific outcome $(g^1, \ldots, g^{3m}) \in (\mathbb{R}^n)^{3m}$ of $(\boldsymbol{g}^1, \ldots, \boldsymbol{g}^{3m}) \sim (N(0, I_n))^{3m}$ is *atypical* if either

(a) for some $i \in [3m]$ we have $\frac{\|g^i\|}{\sqrt{n}} \notin \left[1 - C_1\sqrt{\frac{\log n}{n}}, 1 + C_1\sqrt{\frac{\log n}{n}}\right]$, or

(b) for some pair $i \neq j$ we have $|\frac{g^i \cdot g^j}{n}| > C_1\sqrt{\frac{\log n}{n}}$.

Otherwise we say that $(g^1, \ldots, g^m)$ is *typical*. We defer the proof of the following claim until later:

CLAIM 6.3. *For a suitable choice of the absolute constant C_1, we have $\mathbf{Pr}[(\boldsymbol{g}^1, \ldots, \boldsymbol{g}^{3m})$ is atypical$] \le 0.005$.*

Hence to show that (6.52) ≤ 0.01 (and finish the proof of Theorem 6.3 modulo Claim 6.3), it suffices to prove the following:

CLAIM 6.4. *Fix a typical $(g^1, \ldots, g^{3m}) \in (\mathbb{R}^n)^{3m}$, where $m = c\sqrt{n}\log n$. Then for $\boldsymbol{g} \sim N(0, I_n)$ we have that*

$$\mathrm{d_{TV}}\left((\psi(e_1 \cdot \boldsymbol{g}), \ldots, \psi(e_{3m} \cdot \boldsymbol{g})), \left(\psi\left(\frac{g^1}{\sqrt{n}} \cdot \boldsymbol{g}\right), \ldots, \psi\left(\frac{g^{3m}}{\sqrt{n}} \cdot \boldsymbol{g}\right)\right)\right) \le 0.005. \tag{6.53}$$

Similar to Claim 6.2, we prove Claim 6.4 using Theorem 6.2. To apply Theorem 6.2 we take Σ_1 to be the identity matrix I_{3m} and Σ_2 to be the $3m \times 3m$ matrix whose (i,j) entry is $\frac{g^i \cdot g^j}{n}$, so a draw from $N(0, \Sigma_1)$ is distributed as $(e_1 \cdot \boldsymbol{g}, \ldots, e_{3m} \cdot \boldsymbol{g})$ and a draw from $N(0, \Sigma_2)$ is distributed as $(\frac{g^1}{\sqrt{n}} \cdot \boldsymbol{g}, \ldots, \frac{g^{3m}}{\sqrt{n}} \cdot \boldsymbol{g})$. Applying the data processing inequality for total variation distance, it follows that the LHS of (6.53) is at most $\mathrm{d_{TV}}(N(0^{3m}, \Sigma_1), N(0^{3m}, \Sigma_2))$.

Let A denote the matrix $\Sigma_1^{-1}\Sigma_2 - I_{3m}$, so $\lambda_1^2, \ldots, \lambda_{3m}^2$ are the eigenvalues of A^2 and we have $\sqrt{\lambda_1^2 + \cdots + \lambda_{3m}^2} = \sqrt{\mathrm{tr}(A^2)}$. By part (a) of the definition of "typical" we know that each diagonal entry of A has magnitude at most $3C_1\sqrt{\frac{\log n}{n}}$, and by part (b) we know that each off-diagonal entry has magnitude at most $C_1\sqrt{\frac{\log n}{n}}$. The rest of the proof of Claim 6.4 follows the proof of Claim 6.2 with obvious minor modifications. □

We will finally prove Claim 6.3.

Proof. Recalling that $m = c\sqrt{n}/\log n$, a union bound together with Lemma 2.1 shows that for a suitable choice of C_1, part (a) of the "atypical" definition holds with probability at most 0.0025.

For part (b), by a union bound over all $\binom{m}{2}$ choices of $i \neq j$ and $m = c\sqrt{n}/\log n$, it is enough to show that for $\boldsymbol{g}, \boldsymbol{g}'$ independent $N(0, I_n)$ random variables we have

$$\mathbf{Pr}[\boldsymbol{g} \cdot \boldsymbol{g}' > C_1\sqrt{n \log n}] < 1/n \tag{6.54}$$

(note that we have used the symmetry of the random variable $\boldsymbol{g} \cdot \boldsymbol{g}'$). By the radial symmetry of the $N(0, I_n)$ distribution we may assume that $\boldsymbol{g} = (\boldsymbol{r}, 0, \ldots, 0)$ where $\boldsymbol{r}^2 \sim \chi_n^2$, and hence (since $\boldsymbol{g}_1'$ is distributed as $N(0,1)$) we have that $\boldsymbol{g} \cdot \boldsymbol{g}'$ is distributed as $\boldsymbol{r} \cdot N(0,1)$. By Lemma 2.1 we have that $\boldsymbol{r}^2 < 2n$ except with failure probability at most $\exp(-\Theta(n))$, so we may safely assume that $|r| \le \sqrt{2}n$. By the standard Gaussian tail bound $\mathbf{Pr}[N(0,1) \ge t] \le e^{-t^2/2}$ for $t > 0$, we see that Equation (6.54) holds for a suitable choice of the absolute constant C_1, and Claim 6.3 is proved. □

6.3 An $\Omega(n)$-Sample Lower Bound for Mixtures of Symmetric Convex Sets In this section we show that $\Omega(n)$ samples are needed to distinguish the distribution $N(0, I_n)$ from an unknown distribution in $\mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$, even if the distribution in $\mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$ is guaranteed to have variation distance 1 (the largest possible value) from $\mathcal{N}(0, I_n)$:

THEOREM 6.4. *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 $\mathcal{D} \in \mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$ and has $\mathrm{d_{TV}}(\mathcal{D}, N(0,1)|_K) = 1$ then with probability at least $9/10$ the algorithm outputs "truncated."*

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

Overview. Theorem 6.4 is the most involved of our lower bounds so we give an overview of the steps of the argument before entering into the actual proof.

1. We first use Fact 2.2 to show that $\Omega(1/\delta)$ samples are required to distinguish $N(0, I_n)$ from $N(0, (1-\delta)I_n)$. (We will take $\delta := C/n$ for an absolute constant C.)

2. We define a distribution P that is an infinite mixture over distributions $N(0, I_n)|_K$ where each K in the mixture is a symmetric convex set. The mixture P is carefully designed so that $\mathrm{d_{TV}}(N(0, (1-\delta)I_n), P)$ is small, in particular $o(1/n)$ (in fact we will show that it is at most some $\kappa = 1/n^{\omega(1)}$). Consequently by Fact 2.1, at least $\Omega(1/\kappa)$ samples are required to distinguish P from $N(0, (1-\delta)I_n)$.

   In more detail, each symmetric convex set K in the mixture is the intersection of an $(n-1)$-dimensional hyperplane through the origin with an n-dimensional ball. The radii of these balls are chosen according to a carefully designed distribution, and the directions of the orthogonal vectors to the hyperplanes are (Haar)-uniform random over all possible directions. Consequently P is a radially symmetric distribution; this greatly simplifies the analysis and the proof that indeed P closely approximates the radially symmetric distribution $N(0, (1-\delta)I_n)$.

3. Finally, we define a distribution $\mathcal{F}_P$ over probability distributions Q that are derived from the mixture distribution P. Each distribution Q in the support of $\mathcal{F}_P$ is a finite mixture of n^2 many $N(0, I_n)|_K$ components that are independently chosen from the mixture defining P. Since the support of each Q has n-dimensional Gaussian volume 0 (since it is contained in the union of n^2 many $(n-1)$-dimensional hyperplanes) we have that each Q satisfies $\mathrm{d_{TV}}(Q, N(0, I_n)) = 1$ as claimed.

   We argue that for a random distribution $\mathbf{Q} \sim \mathcal{F}_P$, the variation distance between (the distribution of an i.i.d sample of n points drawn from $\mathbf{Q}$) and (the distribution of an i.i.d sample of n points drawn from P) is very small. This implies that $\Omega(n)$ samples are required to distinguish between a random distribution $\mathbf{Q}$ drawn from $\mathcal{F}_P$ and P itself. We note that this high-level approach, of constructing a distribution over distributions from a mixture distribution and using that distribution over distributions for a distinguishing lower bound, is similar in broad outline to the lower bound approach used in [RS09], see Section 4 of that paper. However, the technical arguments required here are entirely different from [RS09] (and significantly more involved).

4. Combining items 1 through 3 above, we get that

$$\min\{\Omega(1/\delta), \Omega(1/\kappa), \Omega(n)\} = \Omega(n)$$

   samples are required to distinguish a random distribution $\mathbf{Q}$ from $N(0, I_n)$, which gives Theorem 6.4.

We now enter into the formal proof of Theorem 6.4. Suppose that A is an algorithm which receives cn draws from a distribution $\mathcal{D}$ (where $c > 0$ is a suitably small absolute constant) and outputs "un-truncated" with probability at least 9/10 if the distribution $\mathcal{D}$ is $N(0, I_n)$. We will show that any such algorithm must also output "un-truncated" with probability at least 0.8 when the distribution $\mathcal{D}$ is some element of $\mathrm{Mix}(\mathcal{P}_{\mathrm{symm}})$.

Claim 6.5. (Indistinguishability of $N(0, I_n)$ and $N(0, (1-\delta)I_n)$.) *Let $\delta = o(1/\sqrt{n})$. Any algorithm that distinguishes (with correctness probability at least 9/10) whether it is being run on samples from $N(0, I_n)$ or $N(0, (1-\delta)I_n)$ must use $\Omega(\frac{1}{\delta^2 n})$ samples.*

Proof. The squared Hellinger distance between two n-dimensional normal distributions $P = N(0, \Sigma)$ and $Q = N(0, \Sigma')$ is ([Par06], p. 51)

$$H^2(P, Q) = 1 - \frac{\det(\Sigma)^{1/4} \det(\Sigma')^{1/4}}{\det\left(\frac{\Sigma+\Sigma'}{2}\right)}. \tag{6.55}$$

In our setting of $\Sigma_1 = I_n$ and $\Sigma_2 = (1-\delta)I_n$, this simplifies to

$$\begin{aligned}
1 - \frac{(1-\delta)^{n/4}}{(1-\frac{\delta}{2})^{n/2}} &= 1 - \frac{\exp(\frac{n}{4}\ln(1-\delta))}{\exp(\frac{n}{2}\ln(1-\frac{\delta}{2}))} \\
&= 1 - \frac{\exp(-\frac{n}{4}(\delta + \frac{\delta^2}{2} + \frac{\delta^3}{3} + \cdots)}{\exp(-\frac{n}{2}(-\frac{\delta}{2} + \frac{\delta^2}{8} + \frac{\delta^3}{2}4 + \cdots)} \\
\text{(since } \delta = o(1/\sqrt{n})\text{)} \qquad &= 1 - \exp\left(-\frac{\delta^2 n}{16} + O(\delta^3 n)\right) \\
&= \frac{\delta^2 n}{16} \pm O(\delta^3 n),
\end{aligned}$$

from which the claim follows by Fact 2.2. □

We mention that if δ is chosen too small then it would not be possible to closely approximate $N(0,(1-\delta)I_n)$ with a mixture distribution as we do in Lemma 6.1; on the other hand, $1/\delta$ is a bottleneck on the quantitative lower bound that the overall proof will yield, so we would like δ to be as small as possible. We fix $\delta := C/n$ for the rest of the argument for a (large) absolute constant C; this choice gives us the following:

COROLLARY 6.1. *The algorithm A must output "un-truncated" with probability at least* $89/100$ *if it is run on cn draws from $N(0,(1-\delta)I_n)$.*

6.3.1 The Mixture Distribution P We now describe the mixture distribution P; its construction and the analysis establishing Lemma 6.1 are the main part of the proof of Theorem 6.4. In particular, we will approximate $N(0,(1-\delta)I_n)$ with a mixture P over distributions $N(0,I_n)|_K$ where each K is a symmetric convex body.

REMARK 6.1. *We define a parameter δ' to satisfy $1-\delta = (1+\delta')^{-1}$, so $N(0,(1-\delta)I_n)$ is the same as $N(0,(1+\delta')^{-1}I_n)$; this simplifies notation in our analysis below. We observe that like δ, the value δ' is C/n (for a different large absolute constant C).*

We now describe the mixture distribution P. The distribution over the convex bodies $\mathbf{K}$ defining the mixture is as follows: a random convex body $\mathbf{K}$ from the distribution is the intersection of an origin-centered $(n-1)$-dimensional hyperplane with a ball of radius $\sqrt{\boldsymbol{R}}$, where we choose the direction of the hyperplane according to the Haar measure on S^{n-1}, and we draw $\boldsymbol{R} \in [0,\infty)$ with probability $\lambda(\boldsymbol{R})$ (defined below in Definition 6.1).

REMARK 6.2. *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 $K = H \cap \mathrm{Ball}(\sqrt{R})$ (where $H = H_0$ is the hyperplane $\{x \in \mathbb{R}^n : v \cdot x = 0\}$ is taking a sequence of slabs $H_\varepsilon = \{x : |v \cdot x| \le \varepsilon\}$ and letting $\varepsilon \to 0$). In the limit the distribution induced for $N(0,I_n)|_H$ is a symmetric distribution restricted to H, where the probability assigned to points of squared Euclidean distance x from the origin is $\chi^2(n-1,1)(x)$. This will be used in Proposition 6.1 below.*

Recall that for $\boldsymbol{x} \sim N(0,I_n)$, the random variable $\|\boldsymbol{x}\|^2$ is distributed according to a $\chi^2(n)$ distribution. In particular, a draw $\boldsymbol{x} \sim N(0,I_n)$ can be viewed as

1. First drawing $\boldsymbol{v} \sim S^{n-1}$ according to the Haar measure, then
2. Drawing $\boldsymbol{X} \sim \chi^2(n)$;

and then outputting $\boldsymbol{v} \cdot \sqrt{\boldsymbol{X}}$. A draw from $N(0,(1+\delta')^{-1}I_n)$ can be similarly viewed, except with $\boldsymbol{X}$ drawn from a *scaled* $\chi^2(n)$ distribution. For convenience, we introduce the following notation.

NOTATION 6.1. *We will write $\chi^2(n,\sigma^2)$ to denote the distribution of $\|\boldsymbol{x}\|^2$ for $\boldsymbol{x} \sim N(0,\sigma^2 I_n)$. We will also write $\chi^2(n,\sigma^2)|_R$ for the truncated distribution $\chi^2(n,\sigma^2)|_{[0,R]}$.*

It is easy to check that if $\boldsymbol{X} \sim \chi^2(n, \sigma^2)$, then $\boldsymbol{X}/\sigma^2 \sim \chi^2(n, 1) = \chi^2(n)$. In particular, we have the following expression for the density of $\boldsymbol{X}$:

FACT 6.1. *The density of the $\chi^2(n, \sigma^2)$ distribution is*

$$\chi^2(n, \sigma^2)(x) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} \left(\frac{x}{\sigma^2}\right)^{n/2-1} e^{-x/2\sigma^2}.$$

NOTATION 6.2. *We will write $p(\cdot)$ to denote the density of $\chi^2\left(n, (1+\delta')^{-1}\right)$, $q_R(\cdot)$ for the density of $\chi^2(n-1, 1)|_R$, and $\psi(R)$ for the c.d.f. of $\chi^2(n-1, 1)$, i.e. we have*

$$\psi(R) := \mathbf{Pr}\left[\chi^2(n-1, 1) \le R\right].$$

In particular, the following is immediate from Fact 6.1:

$$p(x) = \frac{(1+\delta')^{n/2-1} x^{n/2-1} e^{-(1+\delta')x/2}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}, \tag{6.56}$$

$$q_R(x) = \frac{x^{(n-1)/2-1} e^{-x/2}}{2^{(n-1)/2}\Gamma\left(\frac{n-1}{2}\right)} \cdot \frac{\mathbf{1}(x \le R)}{\psi(R)}. \tag{6.57}$$

Note that a draw from the mixture P can thus be viewed as

1. First drawing $\boldsymbol{v} \sim S^{n-1}$ according to the Haar measure, then
2. Drawing $\boldsymbol{Y} \sim S$;

and outputting $\boldsymbol{v} \cdot \sqrt{\boldsymbol{Y}}$. Defining the univariate distribution S over $[0, \infty)$ whose density is

$$S(x) = \int_R \lambda(R) \cdot q_R(x)\, dR, \tag{6.58}$$

by rotational symmetry, we have the following:

PROPOSITION 6.1. *We have*

$$\mathrm{d_{TV}}\left(P, N(0, (1+\delta')^{-1} I_n)\right) = \mathrm{d_{TV}}\left(S, \chi^2\left(n, (1+\delta')^{-1}\right)\right).$$

We thus want to come up with mixing weights $\lambda(R)$ such that the total variation distance between the univariate distributions p and S is small. It will turn out by our choice of $\lambda(R)$ below that in fact $S(x)$ will not just approximate $p(x)$, but will in fact be *exactly* equal to $p(x)$ for all $x \ge a^*$ for a carefully chosen a^*. In particular, our choice of $\lambda(R)$ (see Definition 6.1) will ensure that:

1. $S(a^*) = p(a^*)$; and
2. $S'(x) = p'(x)$ for all $x \ge a^*$.

We start by simplifying the expressions for $S(\cdot)$ and $p(\cdot)$ (cf. Equations (6.56) and (6.57)). Indeed, having $S(x) = p(x)$ is equivalent to having

$$\frac{x^{(n-1)/2-1} e^{-x/2}}{2^{(n-1)/2}\Gamma\left(\frac{n-1}{2}\right)} \int_R \frac{\lambda(R)}{\psi(R)} \cdot \mathbf{1}(x \le R)\, dR = \frac{(1+\delta')^{n/2-1} x^{n/2-1} e^{-(1+\delta')x/2}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)},$$

which can be rearranged to get

$$\int_x^\infty \frac{\lambda(R)}{\psi(R)}\, dR = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot (1+\delta')^{n/2-1} \cdot \left(\sqrt{x} e^{-\delta' x/2}\right). \tag{6.59}$$

Next, we ensure that that $S'(x) = p'(x)$, i.e. the second item above, before choosing a^*. Differentiating both sides of Equation (6.59), via the fundamental theorem of calculus we get that

$$\frac{\lambda(x)}{\psi(x)} = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\cdot\Gamma\left(\frac{n}{2}\right)} \cdot (1+\delta')^{n/2-1} \cdot e^{-\delta' x/2}\left(\frac{\delta' x - 1}{2\sqrt{x}}\right) \tag{6.60}$$

This suggests taking

$$\lambda(x) := \psi(x) \cdot (\text{RHS of Equation (6.60)}),$$

but this clearly does not result in a valid distribution over R as $\lambda(x)$ will be negative for $x < 1/\delta'$; see Figure 5. However, as the following claim shows, for the above choice of λ we do have that $\int_0^\infty \lambda(R)\,dR = 1$:

Claim 6.6. *For $\lambda(R)$ defined as in Equation (6.60), we have*

$$\int_R \lambda(R)\,dR = 1.$$

Proof. We have

$$\int_R \lambda(R)\,dR = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\cdot\Gamma\left(\frac{n}{2}\right)} \cdot (1+\delta')^{n/2-1} \int_R \underbrace{e^{-\delta' R/2}\left(\frac{\delta' R - 1}{2\sqrt{R}}\right)}_{=:v'(R)} \psi(R)\,dR$$

for $v(x) := -\sqrt{x}e^{-\delta' x/2}$; integrating by parts then gives us

$$= \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\cdot\Gamma\left(\frac{n}{2}\right)} \cdot (1+\delta')^{n/2-1}\left([v(R)\psi(R)]_0^\infty - \int_R v(R)\psi'(R)\,dR\right).$$

Note, however, that $\lim_{x\to\infty} v(x) = 0$ and $\psi(0) = 0$. We also have that $\psi'(R) = \chi^2(n-1,1)(R)$, and so using Fact 6.1 the above simplifies to

$$\begin{aligned}
&= \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\cdot\Gamma\left(\frac{n}{2}\right)} \cdot (1+\delta')^{n/2-1} \int_R \sqrt{R}e^{-\delta' R/2} \cdot \frac{R^{(n-1)/2-1}e^{-R/2}}{2^{(n-1)/2}\Gamma\left(\frac{n-1}{2}\right)}\,dR \\
&= \frac{1}{2^{n/2}\cdot\Gamma\left(\frac{n}{2}\right)} \cdot (1+\delta')^{n/2-1} \int_R R^{n/2-1}e^{-(1+\delta')R/2}\,dR \\
&= \int_R p(R)\,dR \\
&= 1,
\end{aligned}$$

completing the proof. ∎

Claim 6.6 and the fact that $\lambda(R) < 0$ iff $R < 1/\delta$ together suggest a natural way to obtain a valid distribution over $\mathbb{R}$ from $\{\lambda(R)\}$. This is by truncating the support of $\lambda(R)$ to $[a^*, \infty)$, where we take

$$a^* > \frac{1}{\delta'} \qquad \text{such that} \qquad -\int_0^{1/\delta'} \lambda(R)\,dR = \int_{1/\delta'}^{a^*} \lambda(R)\,dR. \tag{6.61}$$

(See also Figure 5.) We thus have the following:

Definition 6.1. (Mixing weights) *For $R \geq 0$, we define the mixing weight $\lambda(R)$ to be*

$$\lambda(R) := \begin{cases} 0 & R < a^* \\ \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\cdot\Gamma\left(\frac{n}{2}\right)} \cdot (1+\delta')^{n/2-1} \cdot e^{-\delta' R/2}\left(\frac{\delta' R - 1}{2\sqrt{R}}\right) \cdot \psi(R) & R \geq a^* \end{cases}.$$

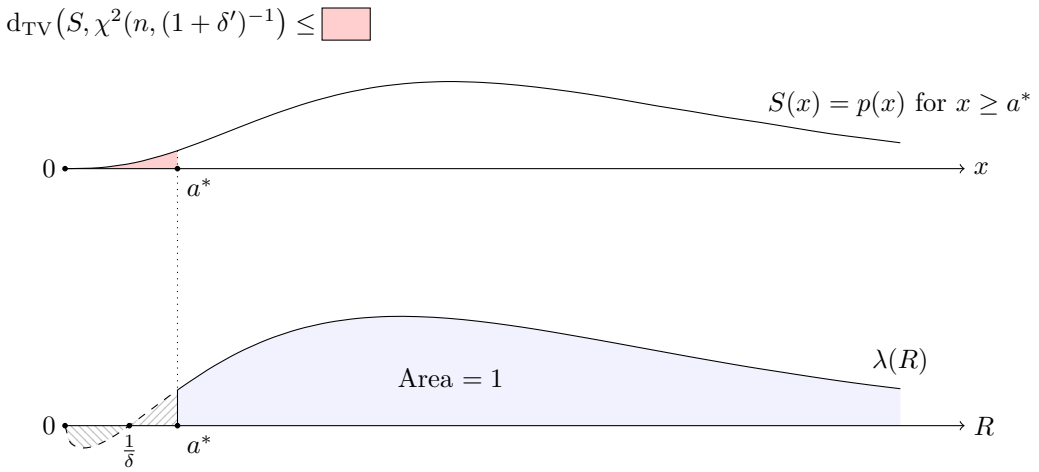


Figure 5: Constructing the mixture distribution S. The bottom plot, of $\lambda(R)$, is the function defined by Equation (6.60); a^* is the value for which the two cross-hatched areas are the same. The revised definition of $\lambda(R)$ given in Definition 6.1, together with Equation (6.58), determines S, depicted in the top plot.

By construction, we have that $\lambda_R \geq 0$ for all R, and that $\int \lambda_R \, dR = 1$. Furthermore, for this choice of $\lambda(\cdot)$, we have from Equations (6.59) and (6.60) that $S'(x) = p'(x)$ for all $x \geq a^*$. We will now show that we also have $S(a^*) = p(a^*)$, establishing the first item.

CLAIM 6.7. *For a^* as in Equation* (6.61)*, we have $S(a^*) = p(a^*)$.*

Proof. The mass at a^* under S is given by

$$\begin{aligned} S(a^*) &= \int_R \lambda(R) q_R(a^*) \, dR \\ &= \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} \cdot (1+\delta')^{n/2-1} \int_R e^{-\delta' R/2} \left(\frac{\delta' R - 1}{2\sqrt{R}}\right) e^{-a^*/2} \cdot (a^*)^{(n-1)/2-1} \mathbf{1}(a^* \leq R) \, dR \\ &= \frac{(1+\delta')^{n/2-1} e^{-a^*/2} \cdot (a^*)^{(n-1)/2-1}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} \int_{a^*}^{\infty} e^{-\delta' R/2} \left(\frac{\delta' R - 1}{2\sqrt{R}}\right) dR. \end{aligned}$$

As before, integration by parts gives

$$\int_{a^*}^{\infty} e^{-\delta' R/2} \left(\frac{\delta' R - 1}{2\sqrt{R}}\right) dR = e^{-\delta' a^*/2} \sqrt{a^*},$$

and so we have

$$S(a^*) = \frac{(1+\delta')^{n/2-1} e^{-(1+\delta')a^*/2} \cdot (a^*)^{n/2-1}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} = p(a^*),$$

completing the proof. ∎

We finally turn to establishing the closeness of P and $N(0, (1+\delta')^{-1} I_n)$ (recall that by our choice of $\delta' = C/n$ we have that $N(0, (1+\delta')^{-1} I_n)$ is the same as $N(0, (1-\delta) I_n)$).

LEMMA 6.1. (CLOSENESS OF P AND $N(0, (1+\delta')^{-1} I_n)$) $\mathrm{d_{TV}}(P, N(0, (1+\delta')^{-1} I_n)) \leq \kappa := 1/n^{\omega(1)}$.

Proof. Note that by Proposition 6.1, it suffices to show that

$$\mathrm{d_{TV}}\big(S, \chi^2\big(n, (1+\delta')^{-1}\big)\big) \leq \kappa.$$

However, by our construction of S above (see also Figure 5), we have that $S(x) = p(x)$ for all $x \geq a^*$, and so

$$\mathrm{d_{TV}}\big(S, \chi^2\big(n, (1+\delta')^{-1}\big)\big) \leq \mathop{\mathbf{Pr}}_{\boldsymbol{X} \sim \chi^2(n,(1+\delta')^{-1})}[\boldsymbol{X} \leq a^*]; \tag{6.62}$$

hence to prove Lemma 6.1 it suffices to show that the RHS of Equation (6.62) is $1/n^{\omega(1)}$.

Recall from Equation (6.61) that $a^* > 1/\delta' = n/C$ for some absolute constant C, and was chosen such that

$$-\int_0^{1/\delta'} \lambda(R)\, dR = \int_{1/\delta'}^{a^*} \lambda(R)\, dR \tag{6.63}$$

where $\lambda(R)$ is given by Equation (6.60). In particular, for our choice of δ' we have

$$\lambda(R) = \underbrace{\frac{\Gamma\big(\frac{n-1}{2}\big)}{\sqrt{2}\cdot\Gamma\big(\frac{n}{2}\big)} \cdot (1+\delta')^{n/2-1}}_{=\Theta\left(\frac{1}{\sqrt{n}}\right)} e^{-\delta' R/2}\left(\frac{\delta' R - 1}{2\sqrt{R}}\right) \cdot \psi(R).$$

For $R \leq n/C$ for large enough C, we thus have (using standard tail bounds on the χ^2 distribution, cf. Lemma 2.1) that $|\lambda(R)| = O(e^{-cn})$. Hence by Equation (6.63) we get that

$$\int_{1/\delta'}^{a^*} \lambda(R)\, dR \leq \frac{1}{\delta'} \cdot O(e^{-cn}) = \frac{n}{C' e^{c' n}} \tag{6.64}$$

where c' and C' are absolute constants. We will establish the following claim:

Claim 6.8. $a^* \leq n - n^{3/4}$.

Given Claim 6.8, we have that the RHS of Equation (6.62) is $\mathbf{Pr}_{\boldsymbol{X}\sim\chi^2(n,1)}[\boldsymbol{X} \leq (n-n^{3/4})(1+\delta')]$, which is at most $2^{-\Theta(n^{1/2})}$ by Lemma 2.1, which gives Lemma 6.1.

To establish Claim 6.8, suppose (for the sake of contradiction) that $a^* \geq n - n^{3/4}$. For each $x \in [n-2n^{3/4}, n-n^{3/4}]$ we have that

$$\begin{aligned}
\text{(by Fact 6.1)} \qquad \chi^2(n,1)(x) &= \frac{1}{2^{n/2}\Gamma\big(\frac{n}{2}\big)} x^{n/2-1} e^{-x/2} \\
&\geq \frac{1}{2^{n/2}\lfloor \frac{n}{2} - 1\rfloor!}\Big(n - 2n^{3/4}\Big)^{n/2-1} e^{-(n-2n^{3/4})/2} \qquad (6.65) \\
&\geq \frac{1}{\mathrm{poly}(n)\cdot 2^{n/2}\cdot\big(\frac{n}{2e}\big)^{n/2-1}} n^{n/2-1}\left(1 - \frac{2}{n^{1/4}}\right)^{n/2-1} e^{-(n-2n^{3/4})/2} \\
&= \frac{1}{\mathrm{poly}(n)}\left(1 - \frac{2}{n^{1/4}}\right)^{n/2-1} e^{n^{3/4}} \\
\text{(using } 1 - x \geq e^{-2x} \text{ for } 0 < x < 0.1) \qquad &\geq \frac{1}{\mathrm{poly}(n)} e^{-2n^{3/4}} e^{n^{3/4}} \\
&\geq e^{-1.5n^{3/4}}.
\end{aligned}$$

where (6.65) is by unimodality of the $\chi^2(n,1)$ distribution and the fact that its mode is $n - 2 > n - n^{3/4}$. Hence

$$\mathop{\mathbf{Pr}}_{\boldsymbol{X}\sim\chi^2(n,1)}[\boldsymbol{X} \leq n - 1.5n^{3/4}] \geq \mathop{\mathbf{Pr}}_{\boldsymbol{X}\sim\chi^2(n,1)}[\boldsymbol{X} \in [n-2n^{3/4}, n-1.5n^{3/4}] \geq (1/2)n^{3/4}\cdot e^{-1.5n^{3/4}},$$

which implies that $\mathbf{Pr}_{\boldsymbol{X}\sim\chi^2(n,(1+\delta')^{-1})}[\boldsymbol{X} \leq n - n^{3/4}] \geq (1/2)n^{3/4}\cdot e^{-1.5n^{3/4}}$. Comparing with Equation (6.64) this gives Claim 6.8, which in turn completes the proof of Lemma 6.1. ∎

Combining Lemma 6.1 and Corollary 6.1 we obtain the following:

Corollary 6.2. *The algorithm A must output "un-truncated" with probability at least* $88/100$ *if it is run on cn draws from P.*

6.3.2 The Distribution $\mathcal{F}_P$ We now describe the distribution $\mathcal{F}_P$, which is a distribution over distributions Q over $\mathbb{R}^n$. A draw of a distribution $\mathbf{Q}$ from $\mathcal{F}_P$ is obtained as follows: $\mathbf{Q}$ is a uniform mixture of n^2 many distributions, where the i-th distribution in the mixture is $N(0, I_n)|_{\mathbf{K}_i}$, where each $\mathbf{K}_i$ is i.i.d. drawn from the distribution $\mathcal{K}$.

It is clear that each distribution Q in the support of $\mathcal{F}_P$ has $\mathrm{Vol}(\mathrm{supp}(Q)) = 0$ and hence $\mathrm{d_{TV}}(N(0, I_n), Q) = 1$. Let $\boldsymbol{S}$ be a random variable which takes values in $(\mathbb{R}^n)^{cn}$, where a draw from $\boldsymbol{S}$ is obtained by making cn independent draws from P. Let $\boldsymbol{S}'$ be a random variable which also takes values in $(\mathbb{R}^n)^{cn}$, where a draw from $\boldsymbol{S}'$ is obtained by (i) first randomly drawing a $\mathbf{Q} \sim \mathcal{F}_P$, and then (ii) making cn independent draws from $\mathbf{Q}$.

CLAIM 6.9. $\mathrm{d_{TV}}(\boldsymbol{S}, \boldsymbol{S}') \leq \frac{1}{100}$.

Proof. The idea is similar to the proof of Claim 7 of [RS09]; for completeness we recall the simple argument here. By the Birthday Paradox, with probability at least 99/100 the cn independent draws from $\mathbf{Q}$ come from $N(0, I_n)|_{\mathbf{K}_{i_1}}, \ldots, N(0, I_n)|_{\mathbf{K}_{i_{cn}}}$ such that $i_1, \ldots, i_{cn}$ are cn distinct values from $[n^2]$. If this happens then the distribution of $\boldsymbol{S}'$ is identical to the distribution of $\boldsymbol{S}$, since in both cases each of the cn vectors in $\mathbb{R}^n$ is independently drawn by first selecting a component $\mathbf{K} \sim \mathcal{K}$ and then making a draw from $N(0, I_n)|_{\mathbf{K}}$. This gives the claim. □

From Claim 6.9 and Corollary 6.2 we obtain the following, which completes the proof of Theorem 6.4:

COROLLARY 6.3. *The algorithm A must output "un-truncated" with probability at least* 87/100 *if it is run on cn draws from P.*

References

[Bal97] Keith Ball. An elementary introduction to modern convex geometry. In *Flavors of geometry*, volume 31 of *Math. Sci. Res. Inst. Publ.*, pages 1–58. Cambridge Univ. Press, Cambridge, 1997.

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

[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.

[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.

[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.

[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.

[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.

[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.

[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.

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

[DDO+13] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R. A. Servedio, and L.-Y. Tan. Learning sums of independent integer random variables. In *54th Annual IEEE Symposium on Foundations of Computer Science*, pages 217–226, 2013.

[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.

[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.

[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.

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

[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.

[DMR20] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. arXiv:1810.08693v5, 22 May 2020, 2020.

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

[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.

[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.

[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.

[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.

[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.

[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.

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

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

[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.

[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.

[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.

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

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

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

[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.

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

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

[Orj14] Eric Orjebin. A recursive formula for the moments of a truncated univariate normal distribution. Available at `https://people.smp.uq.edu.au/YoniNazarathy/teaching_projects/studentWork/EricOrjebin_TruncatedNormalMoments.pdf`, September 2014.

[Par06] L. Pardo. *Statistical Inference Based on Divergence Measures*. Chapman and Hall/CRC, 2006.

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

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

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

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

[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.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[Wik22] Wikipedia contributors. Borel-Kolmogorov paradox. Wikipedia, The Free Encyclopedia, accessed July 11, 2022. `https://en.wikipedia.org/wiki/Borel%E2%80%93Kolmogorov_paradox`.

A Hardness for Mixtures of General Convex Sets

Theorem 1.2 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 1.3 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 A.1. *Let* $\mathrm{Mix}(\mathcal{P}_{\mathrm{conv}})$ *denote the class of all convex combinations (mixtures) of distributions 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* $\mathrm{d_{TV}}(N(0, I_n), \mathcal{D}) \geq \varepsilon$*.*

We sketch a proof of Theorem A.1 below.

Proof. 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 $\mathrm{d_{TV}}(N(0, I_n), N(0, I_n)|_{\boldsymbol{S}}) = \varepsilon$, 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 $N(0, I_n)$. ∎