Efficient Truncated Statistics with Unknown Truncation

Vasilis Kontonis
University of Wisconsin-Madison
Madison, WI, USA
kontonis@wisc.edu

Christos Tzamos
University of Wisconsin-Madison
Madison, WI, USA
tzamos@wisc.edu

Manolis Zampetakis
MIT
Cambridge, MA, USA
mzampet@mit.edu

***Abstract*—We study the problem of estimating the parameters of a Gaussian distribution when samples are only shown if they fall in some (unknown) set. This core problem in truncated statistics has long history going back to Galton, Lee, Pearson and Fisher. Recent work by Daskalakis et al. (FOCS'18), provides the first efficient algorithm that works for arbitrary sets in high dimension when the set is known, but leaves as an open problem the more challenging and relevant case of unknown truncation set.**

Our main result is a computationally and sample efficient algorithm for estimating the parameters of the Gaussian under arbitrary unknown truncation sets whose performance decays with a natural measure of complexity of the set, namely its Gaussian surface area. Notably, this algorithm works for large families of sets including intersections of halfspaces, polynomial threshold functions and general convex sets. We show that our algorithm closely captures the tradeoff between the complexity of the set and the number of samples needed to learn the parameters by exhibiting a set with small Gaussian surface area for which it is information theoretically impossible to learn the true Gaussian with few samples.

***Keywords*-Truncated Statistics; Gaussian Distribution; Learning Theory; Unknown Truncation Set**

I. Introduction

A classical challenge in Statistics is estimation from truncated samples. Truncation occurs when samples falling outside of some subset S of the support of the distribution are not observed. Truncation of samples has myriad manifestations in business, economics, engineering, social sciences, and all areas of the physical sciences.

Statistical estimation under truncated samples has had a long history in Statistics, going back to at least the work of Galton [Gal97] who analyzed truncated samples corresponding to speeds of American trotting horses. Following Galton's work, Pearson and Lee [Pea02], [PL08], [Lee14] used the method of moments in order to estimate the mean and standard deviation of a truncated univariate normal distribution and later Fisher [Fis31] used the maximum likelihood method for the same estimation problem. Since then, there has been a large volume of research devoted to estimating the truncated normal distribution; see e.g. [Sch86], [Coh16], [BC14]. Nevertheless, the first algorithm that is provably computationally and statistically efficient was only recently developed by Daskalakis et al. [DGTZ18], under the assumption that the truncation set S is known.

In virtually all these works the question of estimation under unknown truncation set is raised. Our work resolves this question by providing tight sample complexity guarantees and an efficient algorithm for recovering the underlying Gaussian distribution. Although this estimation problem has clear and important practical and theoretical motivation too little was known prior to our work even in the asymptotic regime. In the early work of Shah and Jaiswal [SJ66] it was proven that the method of moments can be used to estimate a single dimensional Gaussian distribution when the truncation set is unknown but it is assumed to be an interval. In the other extreme where the set is allowed to be arbitrarily complex, Daskalakis et al. [DGTZ18] showed that it is information theoretically impossible to recover the parameters. We provide the first complete analysis of the number of samples needed for recovery taking into account the complexity of the underlying set.

Our Contributions.: Our work studies the estimation task when the truncation set belongs in a family $\mathcal{C}$ of "low complexity". We use two different notions for quantifying the complexity of sets: the VC-dimension and the Gaussian Surface Area.

Our first result is that for any set family with VC-dimension $\mathrm{VC}(\mathcal{C})$, the mean and covariance of the true d-dimensional Gaussian Distribution can be recovered up to accuracy ε using only $\tilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{\varepsilon}+\frac{d^2}{\varepsilon^2}\right)$ truncated samples.

Informal Theorem 1. *Let $\mathcal{C}$ be a class of sets with VC-dimension $\mathrm{VC}(\mathcal{C})$ and let $N=\tilde{O}\left(\frac{\mathrm{VC}(\mathcal{C})}{\varepsilon}+\frac{d^2}{\varepsilon^2}\right)$. Given N samples from a d-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ with unknown mean μ and covariance $\boldsymbol{\Sigma}$, truncated on a set $S\in\mathcal{C}$ with mass at least α, it is possible to find an estimate $(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ such that $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}),\mathcal{N}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}))\leq\varepsilon$.*

The estimation method computes the set of smallest mass that maximizes the likelihood of the data observed and learns the truncated distribution within error $O(\varepsilon)$ in total variation distance. To translate this error in total variation to parameter distance, we prove a general result showing that it is impossible to create a set (no matter the complexity) so that two Gaussians whose parameters are far have similar truncated distributions (see Lemma 8).

A simple but not successful approach would be to first

2575-8454/19/$31.00 ©2019 IEEE
DOI 10.1109/FOCS.2019.00095

try to learn an approximation of the truncation set with symmetric difference roughly ε^2/d^2 with the true set and then run the algorithm of [DGTZ18] using the approximate oracle. This approach would lead to a $\mathrm{VC}(\mathcal{S})d^2/\varepsilon^2$ sample complexity that is worse than what we get. More importantly, doing empirical risk minimization[1] using truncated samples does not guarantee that we will find a set of small *symmetric* difference with the true and it is not clear how one could achieve that.

Our result bounds the sample complexity of identifying the underlying Gaussian distribution in terms of the VC-dimension of the set but does not yield a computationally efficient method for recovery. Obtaining a computationally efficient algorithm seems unlikely, unless one restricts attention to simple specific set families, such as axis aligned rectangles. One would hope that exploiting the fact that samples are drawn from a "tame" distribution, such as a Gaussian, can lead to general computationally efficient algorithms and even improved sample complexity.

Indeed, our main result is an algorithm that is both computationally and statistically efficient for estimating the parameters of a spherical Gaussian and uses only $d^{O(\Gamma^2(\mathcal{C}))}$ samples, where $\Gamma(\mathcal{C})$ is the *Gaussian Surface Area* of the class $\mathcal{C}$, an alternative complexity measure introduced by Klivans et al. [KOS08]:

Informal Theorem 2. *Let $\mathcal{C}$ be a class of sets with Gaussian surface area at most $\Gamma(\mathcal{C})$ and let $k = \mathrm{poly}(1/\alpha, 1/\varepsilon)\Gamma(\mathcal{C})^2$. Given $N = d^k$ samples from a spherical d-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$, truncated on a set $S \in \mathcal{C}$ with mass at least α,in time $\mathrm{poly}(m)$, we can find an estimate $\hat{\mu}, \hat{\sigma}^2$ such that*

$$d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I}), \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2\boldsymbol{I})) \leq \varepsilon.$$

The notion of Gaussian surface area can lead to better sample complexity bounds even when the VC dimension is infinite. An example of such a case is when $\mathcal{C}$ is the class of all convex sets. Table I summarizes the known bounds for the Gaussian surface area of different concept classes and the implied sample complexity in our setting when combined with our main theorem.

Beyond spherical Gaussians, our main result extends to Gaussians with arbitrary diagonal covariance matrices. In addition, we provide an information theoretic result showing that the case with general covariance matrices can also be estimated using the same sample complexity bound by finding a Gaussian and a set that matches the moments of the true distribution. We remark our main algorithmic result Informal Theorem 3 uses Gaussian Surface Area whereas our sample complexity result Informal Theorem 2 uses VC-dimension. We discuss the differences of the two approaches in Section VII.

[1]That is finding a set of the family that contains all the observed samples.

Informal Theorem 3. *Let $\mathcal{C}$ be a class of sets with Gaussian surface area at most $\Gamma(\mathcal{C})$ and let $k = \mathrm{poly}(1/\alpha, 1/\varepsilon)\Gamma(\mathcal{C})^2$. Any truncated Gaussian with $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{S})$ with $\hat{S} \in \mathcal{C}$ that approximately matches the moments up to degree k of a truncated d-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ with $S \in \mathcal{C}$, satisfies $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})) \leq \varepsilon$. The number of samples to estimate the moments within the required accuracy is at most $d^{O(k)}$.*

This shows that the first few moments are sufficient to identify the parameters. Analyzing the guarantees of moment matching methods is notoriously challenging as it involves bounding the error of a system of many polynomial equations. Even for a single-dimensional Gaussian with truncation in an interval, where closed form solutions of the moments exist, it is highly non-trivial to bound these errors [SJ66]. In contrast, our analysis using Hermite polynomials allows us to easily obtain bounds for arbitrary truncation sets in high dimensions, even though no closed form expression for the moments exists.

We conclude by showing that the dependence of our sample complexity bounds both on the VC-dimension and the Gaussian Surface Area is tight *up to polynomial factors*. In particular, we construct a family in d dimensions with VC dimension 2^d and Gaussian surface area $O(d)$ for which it is not possible to learn the mean of the underlying Gaussian within 1 standard deviation using $o(2^{d/2})$ samples.

Informal Theorem 4. *There exists a family of sets $\mathcal{S}$ with $\Gamma(\mathcal{S}) = O(d)$ and VC-dimension 2^d such that any algorithm that draws N samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}, S)$ and computes an estimate $\widetilde{\boldsymbol{\mu}}$ with $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq 1$ must have $N = \Omega(2^{d/2})$.*

Our techniques and relation to prior work.: The work of Klivans et al. [KOS08] provides a computationally and sample efficient algorithm for learning geometric concepts from labeled examples drawn from a Gaussian distribution. On the other hand, the recent work of Daskalakis et al. [DGTZ18] provides efficient estimators for truncated statistics with *known* sets. One could hope to combine these two approaches for our setting, by first learning the set and then using the algorithm of [DGTZ18] to learn the parameters of the Gaussian. This approach, however, fails for two reasons. First, the results of Klivans et al. [KOS08] apply in the supervised learning setting where one has access to both positive and negative samples, while our problem can be thought of as observing only positive examples (those falling inside the set). In addition, any direct approach that extends their result to work with positive only examples requires that the underlying Gaussian distribution is known in advance.

One of our key technical contributions is to extend the techniques of Klivans et al. [KOS08] to work with *positive only examples* from an *unknown* Gaussian distribution, which is the major case of interest in truncated statistics. To

Concept Class	Gaussian Surface Area	Sample Complexity
Polynomial threshold functions of degree k	$O(k)$ [Kan11]	$d^{O(k^2)}$
Intersections of k halfspaces	$O(\sqrt{\log k})$ [KOS08]	$d^{O(\log k)}$
General convex sets	$O(d^{1/4})$ [Bal93]	$d^{O(\sqrt{d})}$

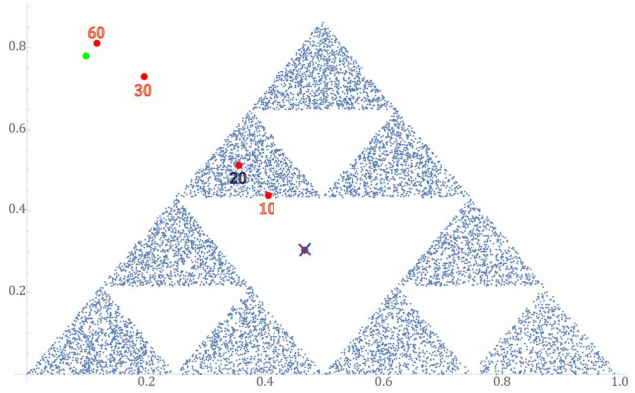
Table I
SUMMARY OF KNOWN RESULTS FOR GAUSSIAN SURFACE AREA. THE LAST COLUMN GIVES THE SAMPLE COMPLEXITY WE OBTAIN FOR OUR SETTING.

perform the set estimation Klivans et al. [KOS08], rely on a family of orthogonal polynomials with respect to the Gaussian distribution, namely the Hermite polynomials and show that the indicator function of the set is well approximated by its low degree Hermite expansion. While we cannot learn this function directly in our setting, we are able to recover an alternative function, that contains "entangled" information of both the true Gaussian parameters and the underlying set. After learning the function, we formulate an optimization problem whose solution enables us to decouple these two quantities and retrieve both the Gaussian parameters and the underlying set. We describe our estimation method in more detail in Section IV. As a corollary of our approach, we obtain the first efficient algorithm for learning geometric concepts from positive examples drawn from an unknown spherical Gaussian distribution.

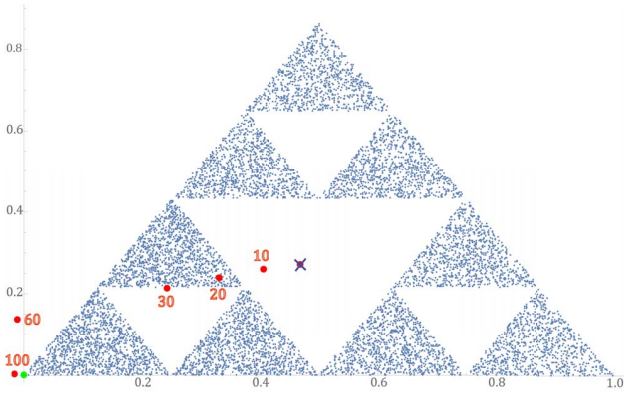
Simulations.: In addition to the theoretical guarantees of our algorithm, we empirically evaluate its performance using simulated data. We present the results that we get in Figure 1, where one can see that even when the truncation set is complex, our algorithm finds an accurate estimation of the mean of the untruncated distribution. Observe that our algorithm succeeds in estimating the true mean of the input distribution despite the fact that the set is unknown and the samples look similar in both cases.

A. Further Related Work

Our work is related to the field of robust statistics as it can robustly learn a Gaussian even in the presence of an adversary erasing samples outside a certain set. Recently, there has been a lot of theoretical work doing robust estimation of the parameters of multi-variate Gaussian distributions in the presence of arbitrary corruptions to a small ε fraction of the samples, allowing for both deletions of samples and additions of samples that can also be chosen adaptively [DKK+16], [CSV17], [LRV16], [DKK+17], [DKK+18]. When the corruption of the data is so powerful it is easy to see that the estimation error of the parameter depends on ε and cannot shrink to 0 as the number of samples grows to infinity. In our model the corruption is more restrictive but in return our results show how to estimate the parameters of a multi-variate Gaussian distribution *to arbitrary* accuracy even when the fraction of corruption is any constant less than 1.



(a) Execution of our algorithm for isotropic Gaussian distribution with $\boldsymbol{\mu}^* = (0.1, 0.78)$ and $\boldsymbol{\mu}_S = (0.48, 0.32)$.



(b) Execution of our algorithm for isotropic Gaussian distribution with $\boldsymbol{\mu}^* = (0, 0)$ and $\boldsymbol{\mu}_S = (0.47, 0.27)$.

Figure 1. Illustration of the results of our algorithm for an unknown truncation set. The $\times$ sign corresponds to the conditional mean of the truncated distribution, while the green point corresponds to the true mean and the red points correspond to the estimated true mean depending on the degree of the Hermite polynomials that are being used by the algorithm.

Our work also has connections with the literature of learning from positive examples. At the heart of virtually all of the results in this literature is the use of the exact knowledge of the original non-truncated distribution to be able to generate fake negative examples, e.g. [Den98], [LDG00]. When the original distribution is uniform, better algorithms are known. Diakonikolas et al. [DDS14] gave

efficient learning algorithms for DNFs and linear threshold functions, Frieze et al. [FJK96] and Anderson et al. [AGR13] gave efficient learning algorithms for learning d-dimensional simplices. Another line of work proves lower bounds on the sample complexity of recovering an unknown set from positive examples. Goyal et al. [GR09] showed that learning a convex set in d-dimensions to accuracy ε from positive samples, uniformly distributed inside the set, requires at least $2^{\Omega(\sqrt{d/\varepsilon})}$ samples, while the work of [Eld11] showed that $2^{\Omega(\sqrt{d})}$ samples are necessary even to estimate the mass of the set. To the best of our knowledge, no matching upper bounds are known for those results. Our estimation result implies that $d^{\mathrm{poly}(\frac{1}{\varepsilon})\sqrt{d}}$ are sufficient to learn the set and its mass when given positive samples from a Gaussian truncated on the convex set.

II. Preliminaries

Notation. We use small bold letters $\boldsymbol{x}$ to refer to real vectors in finite dimension $\mathbb{R}^d$ and capital bold letters $\boldsymbol{A}$ to refer to matrices in $\mathbb{R}^{d\times\ell}$. Similarly, a function with image in $\mathbb{R}^d$ is represented by a small and bold letter $\boldsymbol{f}$. Given a subset S of $\mathbb{R}^d$ we define $\mathbf{1}_S(\boldsymbol{x})$ to be its $0-1$ indicator. Let $\boldsymbol{A}\in\mathbb{R}^{d\times d}$, we define $\boldsymbol{A}^\flat\in\mathbb{R}^{d^2}$ to be the standard vectorization of $\boldsymbol{A}$. Let also $\mathcal{Q}_d$ be the set of all the symmetric $d\times d$ matrices. The *Frobenius norm* of a matrix $\boldsymbol{A}$ is defined as $\|\boldsymbol{A}\|_F=\left\|\boldsymbol{A}^\flat\right\|_2$.

Gaussian Distribution. Let $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ be the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with the following probability density function

$$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};\boldsymbol{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \quad (1)$$

Also, let $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};S)$ denote the *probability mass of a measurable set* S under this Gaussian measure. We shall also denote by $\mathcal{N}_0$ the standard Gaussian, whether it is single or multidimensional will be clear from the context.

Truncated Gaussian Distribution. Let $S\subseteq\mathbb{R}^d$ be a subset of the d-dimensional Euclidean space, we define the *S-truncated normal distribution* $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)$ the normal distribution $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ conditioned on taking values in the subset S. The probability density function of $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)$ is the following

$$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S;\boldsymbol{x}) = \frac{\mathbf{1}_S(\boldsymbol{x})}{\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};S)}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};\boldsymbol{x}). \quad (2)$$

We will assume that the covariance matrix Σ is full rank. The case where Σ is not full rank we can easily detect and solve the estimation problem in the linear subspace of samples.

The core complexity measure of Borel sets in $\mathbb{R}^d$ that we use is the notion of Gaussian Surface Area defined below.

Definition 1 (Gaussian Surface Area)**.** *For a Borel set $A\subseteq\mathbb{R}^d$, $\delta\geq 0$ let $A_\delta=\{x:\mathrm{dist}(x,A)\leq\delta\}$. The Gaussian surface area of A is*

$$\Gamma(A)=\liminf_{\delta\to 0}\frac{\mathcal{N}_0(A_\delta\setminus A)}{\delta}.$$

We define the Gaussian surface area of a family of sets $\mathcal{C}$ to be $\Gamma(\mathcal{C})=\sup_{C\in\mathcal{C}}\Gamma(C)$.

Problem Formulation. Given samples from a truncated Gaussian $\mathcal{N}_S^*\coloneqq\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)$, our goal is to learn the parameters $(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)$ and recover the set S. We denote by $\alpha^*=\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;S)$, the total mass contained in set S by the untruncated Gaussian $\mathcal{N}^*\coloneqq\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)$. Throughout this paper, we assume that we know an absolute constant $\alpha>0$ such that

$$\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;S)=\alpha^*\geq\alpha. \quad (3)$$

We next state the following simple lemma that connects the total variation distance of two Normal distributions with their parameter distance. For a proof see e.g. Corollaries 2.13 and 2.14 of [DKK+16].

Lemma 1. *Let $N_1=\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$, $N_2=\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ be two Normal distributions. Then*

$$d_{\mathrm{TV}}(N_1,N_2)\leq\frac{1}{2}\left\|\boldsymbol{\Sigma}_1^{-1/2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right\|_2 + \sqrt{2}\left\|\boldsymbol{I}-\boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2}\right\|_F$$

We readily use the following two lemmas from [DGTZ18]. The first suggests that we can accurately estimate the parameters (μ_S,Σ_S).

Lemma 2. *Let $(\boldsymbol{\mu}_S,\boldsymbol{\Sigma}_S)$ be the mean and covariance of the truncated Gaussian $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)$ for a set S such that $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};S)\geq\alpha$. Using $\tilde{O}(\frac{d}{\varepsilon^2}\log(1/\alpha)\log^2(1/\delta))$ samples, we can compute estimates $\widetilde{\boldsymbol{\mu}}_S$ and $\widetilde{\boldsymbol{\Sigma}}_S$ such that ,with probability at least $1-\delta$,*

$$\|\boldsymbol{\Sigma}^{-1/2}(\widetilde{\boldsymbol{\mu}}_S-\boldsymbol{\mu}_S)\|_2\leq\varepsilon$$
$$(1-\varepsilon)\boldsymbol{\Sigma}_S\preceq\widetilde{\boldsymbol{\Sigma}}_S\preceq(1+\varepsilon)\boldsymbol{\Sigma}_S$$

The second lemma suggests that the empirical estimates are close to the true parameters of underlying truncated Gaussian.

Lemma 3. *The empirical mean and covariance $\widetilde{\boldsymbol{\mu}}_S$ and $\widetilde{\boldsymbol{\Sigma}}_S$ computed using $\tilde{O}(d^2\log^2(1/\alpha\delta))$ samples from a truncated Normal $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)$ with $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};S)\geq\alpha$ satisfies with probability $1-\delta$ that:*

$$\|\boldsymbol{\Sigma}^{-1/2}(\widetilde{\boldsymbol{\mu}}_S-\boldsymbol{\mu})\|_2^2\leq O(\log\frac{1}{\alpha}),\quad\widetilde{\boldsymbol{\Sigma}}_S\succeq\Omega(\alpha^2)\boldsymbol{\Sigma},$$
$$\left\|\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}^{-1/2}-\boldsymbol{I}\right\|_F^2\leq O(\log\frac{1}{\alpha}).$$

Moreover, $\Omega(\alpha^2)\leq\left\|\widetilde{\boldsymbol{\Sigma}}_S^{-1/2}\boldsymbol{\Sigma}\widetilde{\boldsymbol{\Sigma}}_S^{-1/2}\right\|_2\leq O(1/\alpha^2)$.

In particular, the mean and covariance $\widetilde{\boldsymbol{\mu}}_S$ and $\widetilde{\boldsymbol{\Sigma}}_S$ that satisfy the conditions of Lemma 3, are in $(O(\log(1/\alpha)), 1 - O(\alpha^2))$-near isotropic position.

We will use the following very useful anti-concentration result about the Gaussian mass of sets defined by polynomials.

Theorem 1 (Theorem 8 of [CW01]). *Let $q, \gamma \in \mathbb{R}_+$, $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ such that $\boldsymbol{\Sigma}$ is symmetric positive semidefinite and $p : \mathbb{R}^d \to \mathbb{R}$ be a multivariate polynomial of degree at most ℓ, we define*

$$\bar{Q} = \left\{\boldsymbol{x} \in \mathbb{R}^d \mid |p(\boldsymbol{x})| \le \gamma\right\},$$

then there exists an absolute constant C such that

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \bar{Q}) \le \frac{Cq\gamma^{1/\ell}}{\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[|p(\boldsymbol{z})|^{q/\ell}\right]\right)^{1/q}}.$$

A. Hermite Polynomials, Ornstein-Uhlenbeck Operator, and Gaussian Surface Area.

We denote by $L^2(\mathbb{R}^d, \mathcal{N}_0)$ the vector space of all functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_0}[f^2(x)] < \infty$. The usual inner product for this space is $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_0}[f(\boldsymbol{x})g(\boldsymbol{x})]$. While, usually one considers the probabilists's or physicists' Hermite polynomials, in this work we define the *normalized* Hermite polynomial of degree i to be $H_0(x) = 1, H_1(x) = x, H_2(x) = \frac{x^2-1}{\sqrt{2}}, \ldots, H_i(x) = \frac{He_i(x)}{\sqrt{i!}}, \ldots$ where by $He_i(x)$ we denote the probabilists' Hermite polynomial of degree i. These normalized Hermite polynomials form a complete orthonormal basis for the single dimensional version of the inner product space defined above. To get an orthonormal basis for $L^2(\mathbb{R}^d, \mathcal{N}_0)$, we use a multi-index $V \in \mathbb{N}^d$ to define the d-variate normalized Hermite polynomial as $H_V(\boldsymbol{x}) = \prod_{i=1}^d H_{v_i}(x_i)$. The total degree of H_V is $|V| = \sum v_i \in V v_i$. Given a function $f \in L^2$ we compute its Hermite coefficients as $\hat{f}(V) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_0}[f(\boldsymbol{x})H_V(\boldsymbol{x})]$ and express it uniquely as $\sum_{V\in\mathbb{N}^d} \hat{f}(V)H_V(\boldsymbol{x})$. We denote by $S_k f(x)$ the degree k partial sum of the Hermite expansion of f, $S_k f(\boldsymbol{x}) = \sum_{|V|\le k} \hat{f}(V)H_V(\boldsymbol{x})$. Then, since the basis of Hermite polynomials is complete, we have $\lim_{k\to\infty} \mathbb{E}_{x\sim\mathcal{N}_0}[(f(\boldsymbol{x}) - S_k f(\boldsymbol{x}))^2] = 0$. We would like to quantify the convergence rate of $S_k f$ to f. Parseval's identity states that

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}_0}[(f(\boldsymbol{x}) - S_k f(\boldsymbol{x}))^2] = \sum_{|V|=k}^{\infty} \hat{f}(V)^2.$$

Definition 2 (HERMITE CONCENTRATION). *Let $\gamma(\varepsilon, d)$ be a function $\gamma : (0, 1/2) \times \mathbb{N} \mapsto \mathbb{N}$. We say that a class of functions $\mathcal{F}$ over $\mathbb{R}^d$ has a Hermite concentration bound of $\gamma(\varepsilon, d)$, if for all $d \ge 1$, all $\varepsilon \in (0, 1/2)$, and $f \in \mathcal{F}$ it holds $\sum_{|V|\ge\gamma(\varepsilon,d)} \hat{f}(V)^2 \le \varepsilon$.*

We now define the Gaussian Noise Operator as in [O'D14]. Using a different parametrization, which is not convenient for our purposes, these operators are also known as the Ornstein-Uhlenbeck semigroup, or the Mehler transform.

Definition 3. *The Gaussian Noise operator T_ρ is the linear operator defined on the space of functions $L^1(\mathbb{R}^d, \mathcal{N}_0)$ by*

$$T_\rho f(\boldsymbol{x}) = \mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mathcal{N}_0}\left[f(\rho\boldsymbol{x} + \sqrt{1-\rho^2}\boldsymbol{y})\right].$$

A nice property of operator $T_{1-\rho}$ that we will use is that it has a simple Hermite expansion

$$S_k(T_\rho f)(\boldsymbol{x}) = \sum_{V:|V|\le k} \rho^{|V|}\widehat{f}(V)H_V(\boldsymbol{x}) \tag{4}$$

We also define the noise sensitivity of a function f.

Definition 4 (NOISE SENSITIVITY). *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a function in $L^2(\mathbb{R}^d, \mathcal{N}_0)$. The noise sensitivity of f at $\rho \in [0, 1]$ is defined to be*

$$\mathbf{NS}_\rho[f] = 2 \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}[f(\boldsymbol{x})^2 - f(\boldsymbol{x})T_{1-\rho}f(\boldsymbol{x})]$$

Since, the vectors $\boldsymbol{x}$ and $\boldsymbol{z} = (1-\rho)\boldsymbol{x} + \sqrt{1-\rho^2}\boldsymbol{y}$ are jointly distributed according to

$$D_\rho = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{I} & (1-\rho)\boldsymbol{I}\\ (1-\rho)\boldsymbol{I} & \boldsymbol{I}\end{pmatrix}\right). \tag{5}$$

we can write

$$\begin{aligned}\mathbf{NS}_\rho[f] &\tag{6}\\ &= \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}\left[f(\boldsymbol{x})^2\right] + \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}\left[f(\boldsymbol{z})^2 - 2f(\boldsymbol{x})f(\boldsymbol{z})\right]\\ &= \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{z})\sim D_\rho}[(f(\boldsymbol{x}) - f(\boldsymbol{z}))^2]. &\tag{7}\end{aligned}$$

When f is an indicator of a set, the noise sensitivity is

$$\begin{aligned}\mathbf{NS}_\rho[\mathbf{1}_S] &= 2 \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{z})}[\mathbf{1}_S(\boldsymbol{x})(1 - \mathbf{1}_S(\boldsymbol{z}))]\\ &= 2 \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{z})}[\mathbf{1}_S(\boldsymbol{x})\mathbf{1}_{S^c}(\boldsymbol{z})], &\tag{8}\end{aligned}$$

which is equal to the probability of the correlated points $\boldsymbol{x}, \boldsymbol{z}$ landing at "opposite" sides of S.

Ledoux [Led94] and Pisier [Pis86] showed that the noise sensitivity of a set can be bounded by its Gaussian surface area.

Definition 5 (Gaussian Surface Area). *For a Borel set $A \subseteq \mathbb{R}^d$, its Gaussian surface area is $\Gamma(A) = \liminf_{\delta\to 0} \frac{\mathcal{N}_0(A_\delta \setminus A)}{\delta}$, where $A_\delta = \{x : \mathrm{dist}(x, A) \le \delta\}$.*

We will use the following lemma given in [KOS08].

Lemma 4 (Corollary 14 of [KOS08]). *For a Borel set $S \subseteq \mathbb{R}^d$ and $\rho \ge 0$, $\mathbf{NS}_\rho[\mathbf{1}_S(\boldsymbol{x})] \le \sqrt{\pi}\sqrt{\rho}\,\Gamma(S)$.*

For more details on the Gaussian space and Hermite Analysis (especially from the theoretical computer science perspective), we refer the reader to [O'D14]. Most of the facts about Hermite polynomials that we shall use in this work are well known properties and can be found, for example, in [Sze67].

III. Identifiability with bounded VC dimension

In this section we analyze the sample compexity of learning the true Gaussian parameters when the truncation set has bounded VC-dimension. In particular we show that the overhead over the d^2/ε^2 samples (which is the sample compexity of learning the parameters of the Gaussian without truncation) is proportional to the VC dimension of the class.

Theorem 2. *Let $\mathcal{S}$ be a family of sets of finite VC dimension, and let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ be a truncated Gaussian distribution such that $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S) \geq \alpha$. Given N samples with*

$$N = \mathrm{poly}(1/\alpha)\ \widetilde{O}\left(\frac{d^2}{\varepsilon^2} + \frac{\mathrm{VC}(\mathcal{S})}{\varepsilon}\right)$$

Then, with probability at least 99%, it is possible to identify $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$ that satisfy $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})) \leq \varepsilon$ and $\left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})\right\|_2 \leq \varepsilon$ and $\left\|\boldsymbol{I} - \boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\right\|_F \leq \varepsilon$.

Our algorithm works by first learning the truncated distribution within total variation distance ε. To do this, we first assume that we know the mean and covariance of the underlying Gaussian by guessing the parameters and accurately learn the underlying set. After drawing $N = \Theta(\frac{\mathrm{VC}(\mathcal{S})\log(1/\varepsilon)}{\varepsilon})$ samples from the distribution, any set in the class that contains the samples will only exclude at most an ε fraction of the total mass. Picking the set $\widetilde{S}$ that maximizes the likelihood of those samples, i.e. the set with minimum mass according to the guessed Gaussian distribution, guarantees that the total variation distance between the learned truncated distribution and the true is at most ε, if the guess of the parameters was accurate (Lemma 5). For the proof of Lemma 5 we refer the reader to the full version of the paper.

Lemma 5. *Let $\mathcal{S}$ be a family of subsets in $\mathbb{R}^d$ and Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S^*) = \mathcal{N}_S^*$ be a Normal distribution truncated on the set $S^* \in \mathcal{S}$. Fix $\varepsilon \in (0,1), \delta \in (0, 1/4)$ and let*

$$N = O\left(\frac{\mathrm{VC}(\mathcal{S})\log(1/\varepsilon)}{\varepsilon} + \log\left(\frac{1}{\delta}\right)\right)$$

Moreover, let $\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}$ be such that $d_{\mathrm{TV}}(\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \leq \varepsilon$. Assume that we draw N samples $\boldsymbol{x}_i$ from $\mathcal{N}_{S^*}$, Let S be the solution of the problem

$$\min_S \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}; S) \quad \textit{subject to } x_i \in S \textit{ for all } i \in [n]$$

Then with probability at least $1 - \delta$ we have $d_{\mathrm{TV}}(\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S}), \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)) \leq 3\varepsilon/(2\alpha)$.

This is because the total variation distance between two densities f and g can be written as $\int (f(x) - g(x))\mathbf{1}_{f(x)>g(x)}dx$. Note that by choosing the set of the smallest mass consistent with the samples, we guarantee that the guess will have higher density at every point apart from those outside the support $\widetilde{S}$. However, as we argued the outside mass is at most ε with respect to the true distribution which gives the bound in the total variation distance.

To remove the assumption that the true parameters are known, we build a cover of all possible mean and covariance matrices that the underlying Gaussian might have and run the tournament from [DK14] to identify the best one (Lemma 6). We will use the following statement of the tournament from [DK14]. See also [DL12].

Lemma 6 (Tournament [DK14])**.** *There is an algorithm, which is given sample access to some distribution X and a collection of distributions $\mathcal{H} = \{H_1, \ldots, H_N\}$ over some set, access to a PDF comparator for every pair of distributions H_i, $H_j \in \mathcal{H}$, an accuracy parameter $\varepsilon > 0$, and a confidence parameter $\delta > 0$. The algorithm makes $O(\log(1/\delta)\varepsilon^2)\log N)$ draws from each of $X, H_1, \ldots, H_N$ and returns some $H \in \mathcal{H}$ or declares "failure" If there is some $H \in \mathcal{H}$ such that $d_{\mathrm{TV}}(H, X) \leq \varepsilon$ then with probability at least $1-\delta$ the returned distribution H satisfies $d_{\mathrm{TV}}(H, X) \leq 512\varepsilon$. The total number of operations of the algorithm is $O(\log(1/\delta)(1/\varepsilon^2)(N \log N + \log 1/\delta))$.*

While there are $(d/\varepsilon)^{O(d^2)}$ such parameters, the number of samples needed for running the tournament is only logarithmic which shows that an additional $\widetilde{O}(d^2/\varepsilon^2)$ are sufficient to find a hypothesis in total variation distance ε (Lemma 7). The proof of Lemma 7 can be found the full version of the paper.

Lemma 7. *Let $S \in \mathcal{S}$ be a subset of $\mathbb{R}^d$ and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ be the corresponding truncated normal distribution. Then $\widetilde{O}\left(\mathrm{VC}(\mathcal{S})/\varepsilon + d^2/\varepsilon^2\right)$ samples are sufficient to find parameters $\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S}$ such that $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S), \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})) \leq \varepsilon$ with probability at least 99%.*

We finally argue that the ε error in total variation of the truncated distributions translates to an $O(\varepsilon)$ bound in total variation distance of the untruncated distributions (Lemma 8). We show that this is true in general and does not depend on the complexity of the set. To prove this statement, we consider two Gaussians with parameters that are far from each other and construct the worst possible set to make their truncated distributions as close as possible. We show that under the requirement that the set contains at least α mass, the total variation distance of the truncated distributions will be large.

Lemma 8 (Total Variation of Truncated Normals)**.** *Let $D_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, S_1)$ and $D_2 = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, S_2)$ be two truncated Normal distributions such that $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1; S_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2; S_2) \geq \alpha$. Then*

$$d_{\mathrm{TV}}(D_1, D_2) \geq C_\alpha\ d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2))$$

where $C_\alpha < \alpha/8$ is a positive constant that only depends on α, $C_\alpha = \Omega(\alpha^3)$.

Proof: Without loss of generality we assume that $D_1 = \mathcal{N}(\mathbf{0}, \boldsymbol{I}, S_1)$ and $D_2 = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, S_2)$, where $\boldsymbol{\Lambda}$ is a diagonal matrix. We want to find the worst sets S_1, S_2 so that $d_{\mathrm{TV}}(D_1, D_2)$ is small. If $D_1(S_1 \setminus S_2) \geq \alpha/2$ then the statement holds. Therefore, we consider the set $S = S_1 \cap S_2$ and relax the constraint that the truncated Gaussian D_2 integrates to 1. Taking into account the fact that the set $S = S_1 \cap S_2$ must have at least some mass $\alpha/2$ with respect to $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, the following optimization problem provides a lower bound on the total variation distance of D_1 and D_2.

$$\begin{aligned} \min_{S \in \mathcal{S}, \beta > 0} \quad & \frac{1}{\alpha} \int |\mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x}) - \frac{\alpha}{\beta} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}; \boldsymbol{x})| \ \mathbf{1}_S(\boldsymbol{x}) \mathrm{d}x \\ \text{subj. to} \quad & \int \mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x}) \ \mathbf{1}_S(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \geq \alpha/2, \end{aligned}$$

For any fixed $\beta > 0$ this is a fractional knapsack problem and therefore we should include in the set the points x in order of increasing ratio of weight that is contribution to the L_1 error $|\mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x}) - \frac{\alpha}{\beta} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}; \boldsymbol{x})|$, over value, that is density $\mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x})$ until we reach some threshold T. Therefore, the set is defined to be

$$\begin{aligned} S &= \left\{ \boldsymbol{x} \in \mathbb{R}^d : \frac{|\mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x}) - \frac{\alpha}{\beta} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}; \boldsymbol{x})|}{\mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x})} \leq T \right\} \\ &= \left\{ \boldsymbol{x} \in \mathbb{R}^d : |1 - \exp(p(\boldsymbol{x}))| \leq T \right\}, \end{aligned}$$

where $p(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{x})^T \boldsymbol{\Lambda}^{-1} (\boldsymbol{\mu} - \boldsymbol{x}) + \frac{1}{2} \boldsymbol{x}^T \boldsymbol{x} + \log(\alpha / (\sqrt{|\boldsymbol{\Lambda}|} \beta))$. Using Theorem 1 for the degree 2 polynomial $p(\boldsymbol{x})$ and setting $q = 4$, $\gamma = \alpha^2 (\mathbb{E}_{x \sim \mathcal{N}_0} p^2(x))^{1/2} / (256 C^2)$, where C is the absolute constant of Theorem 1, we get that

$$\mathcal{N}_0(\{\boldsymbol{z} : |p(\boldsymbol{z})| \leq \gamma\}) \leq \frac{\alpha}{4}.$$

To simplify notation set $Q = \{\boldsymbol{z} : |p(\boldsymbol{z})| \leq \gamma\}$. Therefore, for any $\boldsymbol{x}$ in the remaining $\alpha/4$ mass of the set S we know that $|p(\boldsymbol{x})| \geq \gamma$. Next, we lower bound γ in terms of the distance of the parameters of the two Gaussians. We have

$$\begin{aligned} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}_0} [p^2(\boldsymbol{x})] &\geq \mathop{\mathrm{Var}}_{\boldsymbol{x} \sim \mathcal{N}_0} [p(\boldsymbol{x})] \\ &= \mathop{\mathrm{Var}}_{\boldsymbol{x} \sim \mathcal{N}_0} \left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{x})^T \boldsymbol{\Lambda}^{-1} (\boldsymbol{\mu} - \boldsymbol{x}) + \frac{1}{2} \boldsymbol{x}^T \boldsymbol{x} \right] \\ &= \mathop{\mathrm{Var}}_{\boldsymbol{x} \sim \mathcal{N}_0} \left[\sum_{i=1}^d \left(\frac{\mu_i}{\lambda_i} x + x^2 \frac{(1 - 1/\lambda_i)}{2} \right) \right] \\ &= \sum_{i=1}^d \mathop{\mathrm{Var}}_{x \sim \mathcal{N}(0,1)} \left[\frac{\mu_i}{\lambda_i} x + x^2 \frac{(1 - 1/\lambda_i)}{2} \right] \\ &= \sum_{i=1}^d \frac{1}{2} \left(\frac{1}{\lambda_i} - 1 \right)^2 + \frac{\mu_i^2}{\lambda_i^2} \\ &= \frac{1}{2} \left\| \boldsymbol{\Lambda}^{-1} - \boldsymbol{I} \right\|_F^2 + \left\| \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\mu} \right\|_2^2 \end{aligned}$$

Therefore, using the inequality $\sqrt{2}\sqrt{x+y} \geq \sqrt{x} + \sqrt{y}$ we obtain

$$\begin{aligned} \gamma &\geq \frac{\alpha^2}{256\sqrt{2} C^2} \left(\frac{1}{\sqrt{2}} \left\| \boldsymbol{\Lambda}^{-1} - \boldsymbol{I} \right\|_F + \left\| \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\mu} \right\|_2 \right) \\ &\geq \frac{\alpha^2}{256 C^2} d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)), \end{aligned}$$

where we used Lemma 1. Assume first that $\gamma \leq 1$. We have that the L_1 distance between the functions $f(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x}) \mathbf{1}_S(\boldsymbol{x})$ and $g(\boldsymbol{x}) = \frac{\alpha}{\beta} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}; \boldsymbol{x}) \mathbf{1}_S(\boldsymbol{x})$ is

$$\begin{aligned} &\int |f(\boldsymbol{x}) - g(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} \\ &\quad = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}_0} [|1 - \exp(p(\boldsymbol{x}))| \mathbf{1}_S(\boldsymbol{x})] \\ &\quad \geq \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}_0} \left[\frac{|p(x)|}{2} \mathbf{1}_{S \setminus Q}(\boldsymbol{x}) \right] \\ &\quad \geq \gamma \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}_0} \left[\mathbf{1}_{S \setminus Q}(\boldsymbol{x}) \right] \geq \frac{\alpha \gamma}{4} \\ &\quad \geq C_\alpha d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)), \end{aligned}$$

where for the first inequality we used the inequality $|1 - \mathrm{e}^x| \geq |x|/2$ for $|x| \leq 1$. Note that $C_a = \Omega(\alpha^3)$. If $\gamma > 1$ we have

$$\begin{aligned} \int |f(\boldsymbol{x}) - g(\boldsymbol{x})| \mathrm{d}\boldsymbol{x} &= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}_0} [|1 - \exp(p(\boldsymbol{x}))| \mathbf{1}_S(\boldsymbol{x})] \\ &\geq \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}_0} \left[\frac{1}{2} \mathbf{1}_{S \setminus Q}(\boldsymbol{x}) \right] \geq \alpha/8, \end{aligned}$$

where we used the inequality $|1 - \mathrm{e}^x| \geq 1/2$ for $|x| > 1$. ■

A. Learning a Weighted Characteristic Function

Our goal in this section is to recover using conditional samples from $\mathcal{N}_S^*$ a weighted characteristic function of the set S. In particular, we will show that it is possible to learn a good approximation to the function

$$\psi(\boldsymbol{x}) = \frac{\mathbf{1}_S(\boldsymbol{x})}{\alpha^*} \frac{\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; \boldsymbol{x})}{\mathcal{N}(\mathbf{0}, \boldsymbol{I}; \boldsymbol{x})} = \frac{\mathbf{1}_S(\boldsymbol{x})}{\alpha^*} \frac{\mathcal{N}^*(\boldsymbol{x})}{\mathcal{N}_0(\boldsymbol{x})}. \tag{9}$$

We will later use the knowledge of this function to extract the unknown parameters and learn the set S.

1) Hermite Concentration: We start by showing that the function $\psi(\boldsymbol{x})$ admits strong Hermite concentration. This means that we can well-approximate $\psi(\boldsymbol{x})$ if we ignore the higher order terms in the Hermite expansion of $\psi(\boldsymbol{x})$.

Theorem 3. (LOW DEGREE APPROXIMATION) *Let $S_k \psi$ denote the degree k Hermite expansion of function ψ defined in* (9)*. We have that*

$$\begin{aligned} &\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}_0} \left[(S_k \psi(\boldsymbol{x}) - \psi(\boldsymbol{x}))^2 \right] \\ &= \sum_{|V| \geq k} \hat{\psi}(V)^2 \leq \mathrm{poly}(1/\alpha) \left(\frac{\sqrt{\Gamma(S)}}{k^{1/4}} + \frac{1}{k} \right). \end{aligned}$$

where $\Gamma(S)$ is the Gaussian surface area of S, and $a < \alpha^*$ is the absolute constant of (3)*.*

We note that the Hermite expansion of ψ is well-defined as $\psi(\boldsymbol{x}) \in L_2(\mathbb{R}^d, \mathcal{N}_0)$. This can be seen from the following lemma which will be useful in many calculations throughout the paper.

Lemma 9. *Let $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ be two $(B, \frac{1-\delta}{2k})$-isotropic Gaussians for some parameters $B, \delta > 0$ and $k \in \mathbb{N}$. Let $C = \frac{13k^2}{\delta} B$. It holds*

$$\exp(-C) \leq \mathop{\mathbb{E}}_{x \sim \mathcal{N}_0} \left[\left(\frac{\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1; \boldsymbol{x})}{\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2; \boldsymbol{x})} \right)^k \right] \leq \exp(C).$$

Lemma 9 applied for $\mathcal{N}_0$ and $\mathcal{N}^*$ for $k = 2$ implies that $\psi(\boldsymbol{x}) \in L_2(\mathbb{R}^d, \mathcal{N}_0)$.

To get the desired bound for Theorem 3 we use the following lemma, which allows us to bound the Hermite concentration of a function f through its noise stability.

Lemma 10. *For any function $f : \mathbb{R}^d \mapsto \mathbb{R}$ and parameter $\rho \in (0, 1)$, it holds*

$$\sum_{|V| \geq 1/\rho} \hat{f}(V)^2 \leq 2 \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})} \left[f(\boldsymbol{x})^2 - f(\boldsymbol{x}) T_{1-\rho} f(\boldsymbol{x}) \right]$$

Lemma 10 was originally shown in [KKMS05] for indicator functions of sets, but their proof extends to arbitrary real functions. For a proof see the full of the paper.

Using Lemma 10, we can obtain Theorem 3 by bounding the noise sensitivity of the function ψ. The following lemma directly gives the desired result.

Lemma 11. *For any $\rho \in (0, 1)$,*

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}_0} \left[\psi(\boldsymbol{x})^2 - \psi(\boldsymbol{x}) T_{1-\rho} \psi(\boldsymbol{x}) \right] \leq \mathrm{poly}(1/\alpha) \left(\sqrt{\Gamma(S)} \rho^{1/4} + \rho \right)$$

To prove Lemma 11, we will require the following lemma whose proof is provided in the full version of the paper.

Lemma 12. *Let $r(\boldsymbol{x}) \in L_2(\mathbb{R}^d, \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}))$ be differentiable at every $\boldsymbol{x} \in \mathbb{R}^d$. Then*

$$\frac{1}{2} \mathop{\mathbb{E}}_{(x,z) \sim D_\rho} [(r(\boldsymbol{x}) - r(\boldsymbol{z}))^2] \leq \rho \mathop{\mathbb{E}}_{x \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})} \left[\|\nabla r(\boldsymbol{x})\|_2^2 \right]$$

We now move on to the proof of Lemma 11.

Proof of Lemma 11: For ease of notation we define the following distribution

$$D_\rho = \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} \boldsymbol{I} & (1-\rho)\boldsymbol{I} \\ (1-\rho)\boldsymbol{I} & \boldsymbol{I} \end{pmatrix} \right).$$

We also denote by $r(x) = \mathcal{N}^*(\boldsymbol{x}) / \mathcal{N}_0(\boldsymbol{x})$ We can now write

$$\begin{aligned}
&2 \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}_0} \left[\psi(\boldsymbol{x})^2 - \psi(\boldsymbol{x}) T_{1-\rho} \psi(\boldsymbol{x}) \right] \\
&\quad = \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{z}) \sim D_\rho} \left[\psi(\boldsymbol{x})^2 - \psi(\boldsymbol{x}) \psi(\boldsymbol{z}) \right] \\
&\quad = \frac{1}{\alpha^{*2}} \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{z}) \sim D_\rho} [\mathbf{1}_S(\boldsymbol{x}) r^2(\boldsymbol{x}) - \mathbf{1}_S(\boldsymbol{x}) \mathbf{1}_S(\boldsymbol{z}) r^2(\boldsymbol{x})] + \\
&\quad \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{z}) \sim D_\rho} [\mathbf{1}_S(\boldsymbol{x}) \mathbf{1}_S(\boldsymbol{z}) r^2(\boldsymbol{x}) - \mathbf{1}_S(\boldsymbol{x}) \mathbf{1}_S(\boldsymbol{z}) r(\boldsymbol{x}) r(\boldsymbol{z})]
\end{aligned}$$

We bound each of the two terms separately. For the first term, using Schwarz's inequality we get

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{z}) \sim D_\rho} [\mathbf{1}_S(\boldsymbol{x}) r^2(\boldsymbol{x}) - \mathbf{1}_S(\boldsymbol{x}) \mathbf{1}_S(\boldsymbol{z}) r^2(\boldsymbol{x})] \\
&\leq \left(\mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{z}) \sim D_\rho} [\mathbf{1}_S(\boldsymbol{x}) \mathbf{1}_{\bar{S}}(\boldsymbol{z})] \right)^{1/2} \left(\mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{z}) \sim D_\rho} [r^4(\boldsymbol{x})] \right)^{1/2} \\
&\leq (\mathbf{NS}[S])^{1/2} \mathrm{poly}(1/\alpha) \leq \sqrt{\Gamma(S)} \rho^{1/4} \mathrm{poly}(1/\alpha)
\end{aligned}$$

where the bound on the expectation of $r^4(\boldsymbol{x})$ follows from Lemma 9 and the last inequality follows from Lemma 4.

For the second term, we have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{z}) \sim D_\rho} [\mathbf{1}_S(\boldsymbol{x}) \mathbf{1}_S(\boldsymbol{z}) (r^2(\boldsymbol{x}) - r(\boldsymbol{x}) r(\boldsymbol{z}))] \\
&= \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{z}) \sim D_\rho} \left[\mathbf{1}_S(\boldsymbol{x}) \mathbf{1}_S(\boldsymbol{z}) \left(\frac{r^2(\boldsymbol{x})}{2} + \frac{r^2(\boldsymbol{z})}{2} - r(\boldsymbol{x}) r(\boldsymbol{z}) \right) \right] \\
&= \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{z}) \sim D_\rho} \left[\mathbf{1}_S(\boldsymbol{x}) \mathbf{1}_S(\boldsymbol{z}) \frac{1}{2} (r(\boldsymbol{x}) - r(\boldsymbol{z}))^2 \right] \\
&\leq \frac{1}{2} \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{z}) \sim D_\rho} \left[(r(\boldsymbol{x}) - r(\boldsymbol{z}))^2 \right] \leq \rho \mathop{\mathbb{E}}_{\boldsymbol{x} \sim N_0} [\|\nabla r(\boldsymbol{x})\|_2^2],
\end{aligned}$$

where the last inequality follows from Lemma 12. It thus suffices to bound the expectation of the gradient of r. We have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x} \sim N_0} [\, \|\nabla r(\boldsymbol{x})\|_2^2] \\
&\quad = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim N_0} \left[\left\| -\boldsymbol{\Sigma}^{*-1} (\boldsymbol{x} - \boldsymbol{\mu}^*) + \boldsymbol{x} \right\|_2^2 r^2(\boldsymbol{x}) \right] \\
&\quad \leq 2 \mathop{\mathbb{E}}_{\boldsymbol{x} \sim N_0} [\left\| (\boldsymbol{I} - \boldsymbol{\Sigma}^*)^{-1} \boldsymbol{x} \right\|_2^2 r^2(\boldsymbol{x})] \\
&\quad + 2 \left\| \Sigma^{*-1} \boldsymbol{\mu}^* \right\|_2^2 \mathop{\mathbb{E}}_{\boldsymbol{x} \sim N_0} [r^2(\boldsymbol{x})] \\
&\quad \leq 2 \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{x} \sim N_0} [\|(\boldsymbol{I} - \boldsymbol{\Sigma}^{*-1}) \boldsymbol{x}\|_2^4] \mathop{\mathbb{E}}_{\boldsymbol{x} \sim N_0} [r^4(\boldsymbol{x})]} \\
&\quad + 2 \left\| \Sigma^{*-1} \boldsymbol{\mu}^* \right\|_2^2 \mathop{\mathbb{E}}_{\boldsymbol{x} \sim N_0} [r^2(\boldsymbol{x})] \leq \mathrm{poly}(1/\alpha)
\end{aligned}$$

where the bound on the expectation of $r^4(\boldsymbol{x})$ and $r^2(\boldsymbol{x})$ follows from Lemma 9 and the expectation

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x} \sim N_0} \left[\left\| (\boldsymbol{I} - \boldsymbol{\Sigma}^{*-1}) \boldsymbol{x} \right\|_2^4 \right] = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim N_0} \left[\left(\sum_i (1 - \lambda_i)^2 x_i^2 \right)^2 \right] \\
&\leq 3 \left(\sum_i (1 - \lambda_i)^2 \right)^2 \leq 3 \log^2(1/\alpha) \leq \mathrm{poly}(1/\alpha)
\end{aligned}$$

□

2) Learning the Hermite Expansion: In this section we deal with the sample complexity of estimating the coefficients of the Hermite expansion. We have

$$c_V = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)} [H_V(\boldsymbol{x})]$$

Using samples $\boldsymbol{x}_i$ from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$, we can estimate this expectation empirically with the unbiased estimate

$$\widetilde{c}_V = \frac{\sum_{i=1}^N H_V(\boldsymbol{x}_i)}{N}.$$

We now show an upper bound for the variance of the above estimate. The proof of this lemma can be found in the full version of the paper.

Lemma 13. *Let $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$ be the unknown truncated Gaussian. The variance of the following unbiased estimator of the Hermite coefficients $\widetilde{c}_V = \frac{\sum_{i=1}^N H_V(\boldsymbol{x}_i)}{N}$, is upper bounded*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}[(\widetilde{c}_V - c_V)^2] \le \mathrm{poly}(1/\alpha)\frac{5^{|V|}}{N}.$$

Theorem 4. *Let S be an arbitrary (Borel) subset of $\mathbb{R}^d$. Let α be the constant of* (3)*. Let $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$ be the corresponding truncated Gaussian in $(O\log(1/\alpha), 1/16)$-isotropic position (see Definition 6), Then, for the estimate*

$$\psi_k(\boldsymbol{x}) = \max\left(0, \sum_{V:0\le|V|\le k} \widetilde{c}_V H_V(\boldsymbol{x})\right),$$

where $\widetilde{c}_V = \frac{\sum_{i=1}^N H_V(\boldsymbol{x}_i)}{N}$ it holds for $k \ll d$, $\Gamma(S) > 1$,

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[(\psi_k(\boldsymbol{x}) - \psi(\boldsymbol{x}))^2\right]\right]$$
$$\le \mathrm{poly}(1/\alpha)\left(\frac{\sqrt{\Gamma(S)}}{k^{1/4}} + \frac{(5d)^k}{N}\right).$$

Alternatively, for $k = \mathrm{poly}(1/\alpha)\Gamma(S)^2/\varepsilon^4$ we obtain that with $N = d^{\mathrm{poly}(1/\alpha)\Gamma(S)^2/\varepsilon^4}$ samples, with probability at least $9/10$, it holds $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_0}[(\psi_{N,k}(\boldsymbol{x}) - \psi(\boldsymbol{x}))^2] \le \varepsilon$.

Proof: Instead of considering the positive part of the Hermite expansion, we will prove the claim for the empirical Hermite expansion of degree k and N samples

$$p_{N,k} = \sum_{V:0\le|V|\le k} \widetilde{c}_V H_V(\boldsymbol{x}).$$

As usual we denote by $S_k\psi(\boldsymbol{x})$ the true (exact) Hermite expansion of degree k of $\psi(\boldsymbol{x})$. Using the inequality $(a-b)^2 \le 2(a-c)^2 + 2(c-b)^2$ we obtain

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[(p_{N,k}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right]$$
$$\le 2\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[(p_{N,k}(\boldsymbol{x}) - S_k\psi(\boldsymbol{x}))^2\right]$$
$$+ 2\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[(S_k\psi(\boldsymbol{x}) - \psi(\boldsymbol{x}))^2\right]$$

Since Hermite polynomials form an orthonormal system with respect to $\mathcal{N}_0$, we obtain

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[(p_{N,k}(\boldsymbol{x}) - S_k\psi(\boldsymbol{x}))^2\right] = \sum_{V:0\le|V|\le k}(\widetilde{c}_V - c_V)^2.$$

Using Lemma 13 we obtain

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\sim\mathcal{N}^*}\left[\sum_{V:0\le|V|\le k}(\widetilde{c}_V - c_V)^2\right]$$
$$\le \frac{\mathrm{poly}(1/\alpha)}{N}\sum_{V:0\le|V|\le k} 5^{|V|} \le \frac{\mathrm{poly}(1/\alpha)}{N}\binom{d+k}{k}5^k,$$

where we used the fact that the number of all multi-indices V of d elements such that $0 \le |V| \le k$ is $\binom{d+k}{k}$. Moreover, from Theorem 3 we obtain that

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}\left[(S_k\psi(\boldsymbol{x}) - \psi(\boldsymbol{x}))^2\right] \le \mathrm{poly}(1/\alpha)\left(\frac{\sqrt{\Gamma(S)}}{k^{1/4}} + \frac{1}{k}\right).$$

The theorem follows. ∎

IV. Estimation Algorithm for bounded Gaussian Surface Area

In this section, we present the main steps of our estimation algorithm. In later sections, we provide details of the individual components. The algorithm can be thought of in 3 stages.

First Stage: In the first stage, our goal is to learn a weighted characteristic function of the underlying set. Even though we cannot access the underlying set directly, for any given function f we can evaluate the expectation $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)}[f(\boldsymbol{x})]$ using truncated samples.

This expectation can be equivalently written as $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}[f(\boldsymbol{x})\psi(\boldsymbol{x})]$ for the function

$$\psi(\boldsymbol{x}) \coloneqq \frac{\boldsymbol{1}_S(\boldsymbol{x})}{\alpha^*}\frac{\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; \boldsymbol{x})}{\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x})} = \frac{\boldsymbol{1}_S(\boldsymbol{x})}{\alpha^*}\frac{\mathcal{N}^*(\boldsymbol{x})}{\mathcal{N}_0(\boldsymbol{x})}.$$

By evaluating the above expectation for different functions f corresponding to the Hermite polynomials $H_V(\boldsymbol{x})$, we can recover $\psi(\boldsymbol{x})$, through its Hermite expansion:

$$\psi(\boldsymbol{x}) = \sum_{V\in\mathbb{N}^d}\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}[H_V(\boldsymbol{x})\psi(\boldsymbol{x})]H_V(\boldsymbol{x})$$
$$= \sum_{V\in\mathbb{N}^d}\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}[H_V(\boldsymbol{x})]H_V(\boldsymbol{x}).$$

Of course, it is infeasible to calculate the Hermite expansion for any $V \in \mathbb{N}^d$. In Section III-A, we show that by estimating only terms of degree at most k, we can achieve a good approximation to ψ where the error depends on the Gaussian surface area of the underlying set S. To do this, we show that most of the mass of the coefficients $c_V = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_0}[H_V(\boldsymbol{x})\psi(\boldsymbol{x})]$ is concentrated on low degree terms, i.e. $\sum_{|V|>k} c_V^2$ is significantly small. Moreover, we show that even though we can only estimate the coefficients c_V through sampling, the sampling error is significantly small.

Overall, after the first stage, we obtain a non-negative function ψ_k that is close to ψ. The approximation error guarantees are given in Theorem 4.

Second Stage: Given the function ψ_k that was recovered in the first stage, our goal is to decouple the influence of the set $\frac{\mathbf{1}_S(\boldsymbol{x})}{\alpha^*}$ and the influence of the underlying Gaussian distribution which corresponds to the multiplicative term $\frac{\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\boldsymbol{x})}{\mathcal{N}(\mathbf{0},\boldsymbol{I};\boldsymbol{x})}$. This would be easy if we had the exact function ψ in hand. In contrast, for the polynomial function ψ_k the problem is significantly challenging as it is only close to ψ on average but not pointwise.

To perform the decoupling and identify the underlying Gaussian we explicitly multiply the function ψ_k with a corrective term of the form $\frac{\mathcal{N}(\mathbf{0},\boldsymbol{I};\boldsymbol{x})}{\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};\boldsymbol{x})}$. We set up an optimization problem seeking to minimize the function $C(\boldsymbol{\mu},\boldsymbol{\Sigma})\,\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_S^*}[\frac{\mathcal{N}(\mathbf{0},\boldsymbol{I};\boldsymbol{x})}{\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};\boldsymbol{x})}\psi_k(\boldsymbol{x})]$ with an appropriate choice of $C(\boldsymbol{\mu},\boldsymbol{\Sigma})$ so that the unique solution corresponds to $(\boldsymbol{\mu},\boldsymbol{\Sigma}) = (\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)$. Under a reparameterization of $(\boldsymbol{u},\boldsymbol{B}) = (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},\boldsymbol{\Sigma}^{-1})$, we show that the corresponding problem is *strongly* convex. Still, optimizing it directly is non-trivial as it involves taking the expectation with respect to the unknown truncated Gaussian. Instead, we perform stochastic gradient descent (SGD) and show that it quickly converges in few steps to point close to the true minimizer (Algorithm 2).

This allows us to recover parameters $(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ so that the total variation distance between the recovered and the true (untruncated) Gaussian is very small, i.e. $d_{TV}\left(\mathcal{N}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}),\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)\right) \le \varepsilon$. Theorem 5 describes the guarantees of the second stage. Further details are provided in Section IV-A.

The guarantees of the algorithm: We first show our algorithmic results under the assumption that the untruncated Gaussian $\mathcal{N}^*$ is known to be in near-isotropic position.

Definition 6 (Near-Isotropic Position)**.** *Let $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ be a positive semidefinite symmetric matrix and $a, b > 0$. We say that $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is in (a,b)-isotropic position if the following hold.*

$$\|\boldsymbol{\mu}\|_2^2 \le a, \quad \|\boldsymbol{\Sigma}-\boldsymbol{I}\|_F^2 \le a, \quad (1-b)\boldsymbol{I} \preceq \boldsymbol{\Sigma} \preceq \frac{1}{1-b}\boldsymbol{I}$$

We later transform the more interesting case with an unknown mean and an unknown diagonal covariance matrix to the isotropic case.

Theorem 5. *Let $\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)$ be a d-dimensional Gaussian distribution that is in $(O(\log(1/\alpha^*),1/16)$-isotropic position and consider a set S such that $\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;S) \ge \alpha$. There exists an algorithm such that for all $\epsilon > 0$, the algorithm uses $n > d^{\mathrm{poly}(1/\alpha)\frac{\Gamma^2(S)}{\varepsilon^8}}$ samples and produces, in $\mathrm{poly}(n)$ time, estimates that, with probability at least 99%, satisfy $d_{TV}(\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*),\mathcal{N}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})) \le \varepsilon$.*

We can apply this theorem to estimate the parameters of any Gaussian distribution with an unknown mean and an unknown diagonal covariance matrix by bringing the Gaussian to an $(O(\log(1/\alpha^*),1/16)$-isotropic position. Lemma 3 shows that with high probability, we can obtain initial estimates $\widetilde{\boldsymbol{\mu}}_S$ and $\widetilde{\boldsymbol{\Sigma}}_S$ so that $\|\boldsymbol{\Sigma}^{-1/2}(\widetilde{\boldsymbol{\mu}}_S-\boldsymbol{\mu}^*)\|_2^2 \le O(\log\frac{1}{\alpha})$ and

$$\widetilde{\boldsymbol{\Sigma}}_S \succeq \Omega(\alpha^2)\boldsymbol{\Sigma}^*,$$
$$\text{and} \quad \left\|\boldsymbol{\Sigma}^{*-1/2}\widetilde{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}^{*-1/2}-\boldsymbol{I}\right\|_F^2 \le O(\log\frac{1}{\alpha}).$$

Given these estimates, we can transform the space so that $\widetilde{\boldsymbol{\mu}}_S = 0$, and $\widetilde{\boldsymbol{\Sigma}}_S = \boldsymbol{I}$. We note that after this transformation, the mean will be at the right distance from 0, while the eigenvalues λ_i of $\boldsymbol{\Sigma}^*$ will all be within the desired range $\frac{15}{16} \le \lambda_i \le \frac{16}{15}$ apart from at most $O(\log(1/\alpha))$. This is because the condition $\left\|\boldsymbol{\Sigma}^{*-1/2}\widetilde{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}^{*-1/2}-\boldsymbol{I}\right\|_F^2 \le O(\log\frac{1}{\alpha})$ implies that $\sum_i(1-\frac{1}{\lambda_i})^2 \le O(\log(1/\alpha))$. With this observation, since we know of the eigenvectors of Σ^*, we would be able to search over all possible corrections to the eigenvalues to bring the Gaussian in $(O(\log(1/\alpha)),\frac{1}{16})$-isotropic position as required by Theorem 5. We only need to correct $O(\log(1/\alpha))$ of them.

We can form a space of candidate hypotheses for the underlying distribution, for each choice of $O(\log(1/\alpha))$ out of the d vectors along with the all possible scalings. These hypotheses are at most $d^{O(\log(1/\alpha))}$ times $(\log(1/\alpha))^{O(\log(1/\alpha))}$ for all possible scalings. Thus, there are at most $d^{O(\log(1/\alpha))}$ hypotheses. Running the algorithm for each one of them, we would learn at least one distribution and one set that is accurate according to the guarantees of Theorems 5. Running the generic hypothesis testing algorithm of Lemma 6, we can identify one that is closest in total variation distance to the true distribution $\mathcal{N}_S^*$. The sample complexity and runtime would thus only increase by at most $d^{O(\log(1/\alpha))}$. As we showed in Lemma 8, knowing the truncated Gaussian in total variation distance suffices to learn in accuracy ε the parameters of the untruncated distribution. We thus obtain as corollary, that we can estimate the parameters when the covariance is spherical or diagonal. The same results hold when one wants to recover the underlying set in these cases.

A. Optimization of Gaussian Parameters

In this section we show that we can formulate a convex objective function that can be optimized to yield the unknown parameters $\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*$ of the truncated Gaussian. Let S be the unknown (Borel) subset of $\mathbb{R}^d$ such that $\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;S) = \alpha^*$ and let $\mathcal{N}_S^* = \mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)$ be the corresponding truncated Gaussian.

To find the parameters $\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*$, we define the function

$$M_f(\boldsymbol{u},\boldsymbol{B}) :- \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[\mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\mathbf{0},\boldsymbol{I};\boldsymbol{x})f(\boldsymbol{x})\right] \tag{10}$$

where $h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x}) = \frac{\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}}{2} - \frac{\mathrm{tr}((\boldsymbol{B}-\boldsymbol{I})(\widetilde{\boldsymbol{\Sigma}}_S+\widetilde{\boldsymbol{\mu}}_S\widetilde{\boldsymbol{\mu}}_S^T))}{2} - \boldsymbol{u}^T(\boldsymbol{x}-\widetilde{\boldsymbol{\mu}}_S) + \frac{d}{2}\log 2\pi$.

We will show that the minimizer of $M_f(\boldsymbol{u},\boldsymbol{B})$ for the polynomial function $f = \psi_k$, will satisfy $(\boldsymbol{B}^{-1}\boldsymbol{u},\boldsymbol{B}^{-1}) \approx$

$(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. Note that $M_f(\boldsymbol{u}, \boldsymbol{B})$ can be estimated through samples. Our goal will be to optimize it through stochastic gradient descent.

In order to make sure that SGD algorithm for M_{ψ_k} converges fast in the parameter space we need to project after every iteration to some subset of the space as we will see in more details later in this Section. Assuming that the pair $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ is in $(\sqrt{\log(1/\alpha^*)}, 1/16)$-isotropic position we define the following set

$$\mathcal{D} = \{(\boldsymbol{u}, \boldsymbol{B}) \mid (\boldsymbol{B}^{-1}\boldsymbol{u}, \boldsymbol{B}^{-1}) \text{ is in } (c \cdot \log(1/\alpha^*), 1/16)\text{-isotropic position}\} \quad (11)$$

Where c is the universal constant guaranteed to exist from Section II so that

$$\max\left\{\|\boldsymbol{\mu}^* - \tilde{\boldsymbol{\mu}}\|_{\boldsymbol{\Sigma}^*}, \left\|\boldsymbol{\Sigma}^* - \tilde{\boldsymbol{\Sigma}}\right\|_F\right\} \leq c \cdot \log(1/\alpha^*).$$

It is not hard to see that $\mathcal{D}$ is a convex set and that for any $(\boldsymbol{u}, \boldsymbol{B})$ the projection to $\mathcal{D}$ can be done efficiently. For more details we refer to Lemma 8 of [DGTZ18]. Since after every iteration of our algorithm we project to $\mathcal{D}$ we will assume for the rest of this Section that $(\boldsymbol{u}, \boldsymbol{B}) \in \mathcal{D}$.

An equivalent formulation of $M_f(\boldsymbol{u}, \boldsymbol{B})$ that will be useful for the analysis of the SGD algorithm is

$$\begin{aligned} &M_f(\boldsymbol{u}, \boldsymbol{B}) \\ &= e^{-\frac{1}{2}\left(\mathrm{tr}((\boldsymbol{B}-\boldsymbol{I})(\widetilde{\boldsymbol{\Sigma}}_S+\widetilde{\boldsymbol{\mu}}_S\widetilde{\boldsymbol{\mu}}_S^T))\right)+\boldsymbol{u}^T\boldsymbol{B}^{-1}\boldsymbol{u}-\boldsymbol{u}^T\widetilde{\boldsymbol{\mu}}_S}\sqrt{|\boldsymbol{B}|} \\ &\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[\frac{\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x})}{\mathcal{N}(\boldsymbol{B}^{-1}\boldsymbol{u}, \boldsymbol{B}^{-1}; \boldsymbol{x})} f(\boldsymbol{x})\right] \\ &:= C_{\boldsymbol{u},\boldsymbol{B}} \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[\frac{\mathcal{N}_0(\boldsymbol{x})}{\mathcal{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})} f(\boldsymbol{x})\right] \end{aligned} \quad (12)$$

Lemma 14. *For $(\boldsymbol{u}, \boldsymbol{B}) \in \mathcal{D}$, we have that $\mathrm{poly}(\alpha) \leq C_{\boldsymbol{u},\boldsymbol{B}} \leq \mathrm{poly}(1/\alpha)$.*

Proof: We have that

$$\begin{aligned} &|2\log C_{\boldsymbol{u},\boldsymbol{B}}| \\ &= \left|\mathrm{tr}((\boldsymbol{B}-\boldsymbol{I})(\widetilde{\boldsymbol{\Sigma}}_S+\widetilde{\boldsymbol{\mu}}_S\widetilde{\boldsymbol{\mu}}_S^T))\right) \\ &+ \boldsymbol{u}^T\boldsymbol{B}^{-1}\boldsymbol{u} - \boldsymbol{u}^T\widetilde{\boldsymbol{\mu}}_S - \log|\boldsymbol{B}|\Big| \\ &= \left|\mathrm{tr}(\boldsymbol{B}-\boldsymbol{I}) + \mathrm{tr}((\boldsymbol{B}-\boldsymbol{I})(\widetilde{\boldsymbol{\Sigma}}_S-\boldsymbol{I}))\right. \\ &+ \boldsymbol{u}^T\boldsymbol{B}^{-1}\boldsymbol{u} - \log|\boldsymbol{B}|\Big| \\ &\leq |\mathrm{tr}(\boldsymbol{B}-\boldsymbol{I}) - \log|\boldsymbol{B}|| \\ &+ \left|\mathrm{tr}((\boldsymbol{B}-\boldsymbol{I})(\widetilde{\boldsymbol{\Sigma}}_S-\boldsymbol{I}))\right| + \left|\boldsymbol{u}^T\boldsymbol{B}^{-1}\boldsymbol{u}\right| \end{aligned}$$

We now bound each of the terms separately. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of $\boldsymbol{B}$.

For the first term, we have that

$$\begin{aligned} &|\mathrm{tr}(\boldsymbol{B}-\boldsymbol{I}) - \log|\boldsymbol{B}|| \\ &= |\sum_{i=1}^d(\lambda_i - 1 - \log\lambda_i)| \leq \sum_{i=1}^d \frac{(\lambda_i-1)^2}{\lambda_i} \leq \frac{\|\boldsymbol{B}-\boldsymbol{I}\|_F^2}{\lambda_{min}}, \end{aligned}$$

where we used the fact that $0 \leq x - 1 - \log x \leq \frac{(x-1)^2}{x}$ for all $x > 0$. For the second term, we have that $\left|\mathrm{tr}((\boldsymbol{B}-\boldsymbol{I})(\widetilde{\boldsymbol{\Sigma}}_S-\boldsymbol{I}))\right| \leq \|\boldsymbol{B}-\boldsymbol{I}\|_F\|\widetilde{\boldsymbol{\Sigma}}_S-\boldsymbol{I}\|_F$ For the third term, we have that $\left|\boldsymbol{u}^T\boldsymbol{B}^{-1}\boldsymbol{u}\right| = \boldsymbol{u}^T\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{B}^{-1}\boldsymbol{u} \leq \lambda_{max}\|\boldsymbol{B}^{-1}\boldsymbol{u}\|_2^2$

Now from the assumption $(\boldsymbol{u}, \boldsymbol{B}) \in \mathcal{D}$ we have that $\|\boldsymbol{B}-\boldsymbol{I}\|_F \leq O(\sqrt{\log(1/\alpha^*)})$, $\left\|\boldsymbol{B}^{-1}\boldsymbol{u}\right\|_2 \leq O(\sqrt{\log(1/\alpha^*)})$, $\lambda_{min} \geq 15/16$ and $\lambda_{max} \leq 17/16$. Also from Lemma 3 we get that $\left\|\tilde{\boldsymbol{\Sigma}}_S - \boldsymbol{I}\right\|_F \leq O(\sqrt{\log(1/\alpha^*)})$ and hence $|2\log C_{\boldsymbol{u},\boldsymbol{B}}| \leq O(\log(1/\alpha^*))$. This means that $C_{\boldsymbol{u},\boldsymbol{B}} = \mathrm{poly}(1/\alpha)$ and the lemma follows. ∎

1) The Objective Function and its Approximation: To show that the minimizer of the function M_{ψ_k} is a good estimator for the unknown parameters $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$, we consider the function M_f', defined as $M_f(\boldsymbol{u}, \boldsymbol{B}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[\mathrm{e}^{h'(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x})f(\boldsymbol{x})\right]$ for $h'(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x}) = \frac{\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}}{2} - \frac{\mathrm{tr}((\boldsymbol{B}-\boldsymbol{I})(\boldsymbol{\Sigma}_S+\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T))}{2} - \boldsymbol{u}^T(\boldsymbol{x}-\boldsymbol{\mu}_S) + \frac{d}{2}\log 2\pi$. This function corresponds to an ideal situation where we know the parameters $\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S$ exactly. Similarly to (12), we can write M_f' as $C'_{\boldsymbol{u},\boldsymbol{B}}\,\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[\frac{\mathcal{N}_0(\boldsymbol{x})}{\mathcal{N}_{\boldsymbol{u},\boldsymbol{B}}(\boldsymbol{x})} f(\boldsymbol{x})\right]$. We argue that both M_f and M_f' are convex.

Claim 1. *For any function $f : \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$, $M_f(\boldsymbol{u}, \boldsymbol{B})$ and $M_f'(\boldsymbol{u}, \boldsymbol{B})$ are convex functions of the parameters $(\boldsymbol{u}, \boldsymbol{B})$.*

Proof: We show the statement for M_f. The proof for M_f' is identical. The proof follows by computing the Hessian of M_f and arguing that it is positive semidefinite.

The gradient with respect to $(\boldsymbol{u}, \boldsymbol{B})$ is

$$\begin{aligned} &\nabla M_f(\boldsymbol{u}, \boldsymbol{B}) \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)}\left[\nabla h(\boldsymbol{u}, \boldsymbol{B}; \boldsymbol{x})\mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x})f(\boldsymbol{x})\right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)}\left[\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S\tilde{\boldsymbol{\mu}}_S^T\right)^\flat \\ \tilde{\boldsymbol{\mu}}_S - \boldsymbol{x}\end{pmatrix}\right. \\ &\left.\mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x})f(\boldsymbol{x})\right] \end{aligned} \quad (13)$$

Moreover, the Hessian is

$$\begin{aligned} &\mathcal{H}_{M_f}(\boldsymbol{u}, \boldsymbol{B}) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)}\left[\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S\tilde{\boldsymbol{\mu}}_S^T\right)^\flat \\ \tilde{\boldsymbol{\mu}}_S - \boldsymbol{x}\end{pmatrix}\right. \\ &\left.\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S\tilde{\boldsymbol{\mu}}_S^T\right)^\flat \\ \tilde{\boldsymbol{\mu}}_S - \boldsymbol{x}\end{pmatrix}^T \mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x})f(\boldsymbol{x})\right] \end{aligned}$$

which is clearly positive semidefinite since for any $\boldsymbol{z} \in \mathbb{R}^{d\times d+d}$ we have

$$\begin{aligned} &\boldsymbol{z}^T\mathcal{H}_{M_f}(\boldsymbol{u}, \boldsymbol{B})\boldsymbol{z} \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)}\left[\left(\boldsymbol{z}^T\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S\tilde{\boldsymbol{\mu}}_S^T\right)^\flat \\ \tilde{\boldsymbol{\mu}}_S - \boldsymbol{x}\end{pmatrix}\right)^2\right. \end{aligned}$$

$$\mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})f(\boldsymbol{x})\Bigg] \geq 0.$$

■

We now argue that the minimizer of the convex function M'_ψ for the weighted characteristic function $\psi(\boldsymbol{x}) = \frac{\mathbf{1}_S(\boldsymbol{x})}{\alpha^*}\frac{\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\boldsymbol{x})}{\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})}$ is $(\boldsymbol{u},\boldsymbol{B}) = (\boldsymbol{\Sigma}^{*-1},\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*)$.

Claim 2. *The minimizer of $M'_\psi(\boldsymbol{u},\boldsymbol{B})$ is $(\boldsymbol{u},\boldsymbol{B}) = (\boldsymbol{\Sigma}^{*-1},\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*)$.*

Proof: The gradient of M'_ψ with respect to $(\boldsymbol{u},\boldsymbol{B})$ is

$$\begin{aligned}\nabla M'_\psi(\boldsymbol{u},\boldsymbol{B}) = &\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\Bigg[\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T-\boldsymbol{\Sigma}_S-\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T\right)^\flat\\ \boldsymbol{\mu}_S-\boldsymbol{x}\end{pmatrix}\\ &\mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})\frac{\mathbf{1}_S(\boldsymbol{x})}{\alpha^*}\frac{\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\boldsymbol{x})}{\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})}\Bigg]\\ = &\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\Bigg[\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T-\boldsymbol{\Sigma}_S-\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T\right)^\flat\\ \boldsymbol{\mu}_S-\boldsymbol{x}\end{pmatrix}\\ &\mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\frac{\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\boldsymbol{x})}{\alpha^*}\Bigg]\end{aligned}$$

For $(\boldsymbol{u},\boldsymbol{B}) = (\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^{*-1})$, this is equal to

$$\begin{aligned}&\nabla M'_\psi(\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^{*-1})\\ &\quad= C_{\boldsymbol{u},\boldsymbol{B}}\cdot\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\Bigg[\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T-\boldsymbol{\Sigma}_S-\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T\right)^\flat\\ \boldsymbol{\mu}_S-\boldsymbol{x}\end{pmatrix}\\ &\qquad\frac{1}{\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\boldsymbol{x})}\frac{\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*;\boldsymbol{x})}{\alpha^*}\Bigg]\\ &\quad= \frac{C_{\boldsymbol{u},\boldsymbol{B}}}{\alpha^*}\cdot\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\Bigg[\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T-\boldsymbol{\Sigma}_S-\boldsymbol{\mu}_S\boldsymbol{\mu}_S^T\right)^\flat\\ \boldsymbol{\mu}_S-\boldsymbol{x}\end{pmatrix}\Bigg]\end{aligned}$$

where $C_{\boldsymbol{u},\boldsymbol{B}}$ that does not depend on x. This is equal to 0 by definition of $\boldsymbol{\mu}_S$ and $\boldsymbol{\Sigma}_S$. ■

We want to show that the minimizer of M_{ψ_k} is close to that of M'_ψ. To do this, we bound the difference of the two functions pointwise. The proof of the following lemma is technical and can be found in the full version of the paper.

Lemma 15 (POINTWISE APPROXIMATION OF THE OBJECTIVE FUNCTION)**.** *Assume that we use Lemma 2 to estimate $\tilde{\boldsymbol{\mu}}_S,\tilde{\boldsymbol{\Sigma}}_S$ with $\varepsilon = \frac{1}{\mathrm{poly}(1/\alpha^*)}\varepsilon'$ and Theorem 4 with $\varepsilon = \frac{1}{p(1/\alpha^*)}\varepsilon'^2$ then*

$$\left|M_{\psi_k}(\boldsymbol{u},\boldsymbol{B}) - M'_\psi(\boldsymbol{u},\boldsymbol{B})\right| \leq \varepsilon'.$$

Now that we have established that M_{ψ_k} is a good approximation of M'_ψ we will prove that we can optimize M_{ψ_k} and get a solution that is very close to the optimal solution of M'_ψ.

2) Optimization of the Approximate Objective Function: Our goal in this section is to prove that using sample access to $\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)$ we can find the minimum of the function M_{ψ_k} defined in the previous section. First of all recall that M_{ψ_k} can be written as an expectation over $\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)$ in the following way

$$M_{\psi_k}(\boldsymbol{u},\boldsymbol{B}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[\mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})\psi_k(\boldsymbol{x})\right].$$

In Section III-A we prove that we can learn the function ψ_k and hence M_{ψ_k} can be written as

$$M_{\psi_k}(\boldsymbol{u},\boldsymbol{B}) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[m_{\psi_k}(\boldsymbol{u},\boldsymbol{B};x)\right]$$

where $m_{\psi_k}(\boldsymbol{u},\boldsymbol{B};x) = \mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})\psi_k(\boldsymbol{x})$, and for any $\boldsymbol{u},\boldsymbol{B}$ and $\boldsymbol{x}$ we can compute $m_{\psi_k}(\boldsymbol{u},\boldsymbol{B};x)$. Since M_{ψ_k} is convex we are going to use stochastic gradient descent to find its minimum. To prove the convergence of SGD and bound the number of steps that SGD needs to converge we will use the the formulation developed in Chapter 14 of [SSBD14]. To be able to use their results we have to define for any $(\boldsymbol{u},\boldsymbol{B})$ a random vector $\boldsymbol{v}(\boldsymbol{u},\boldsymbol{B})$ and prove the following

UNBIASED GRADIENT ESTIMATION
$$\mathbb{E}\left[\boldsymbol{v}(\boldsymbol{u},\boldsymbol{B})\right] = \nabla M_{\psi_k},$$

BOUNDED STEP VARIANCE
$$\mathbb{E}\left[\left\|\boldsymbol{v}(\boldsymbol{u},\boldsymbol{B})\right\|_2^2\right] \leq \rho,$$

STRONG CONVEXITY for any $\boldsymbol{z}\in\mathcal{D}$ it holds
$$\boldsymbol{z}^T\mathcal{H}_{M_f}(\boldsymbol{u},\boldsymbol{B})\boldsymbol{z} \geq \lambda.$$

We start with the definition of the random vector $\boldsymbol{v}$. Given a sample $\boldsymbol{x}$ from $\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)$, for any $(\boldsymbol{u},\boldsymbol{B})$ we define

$$\begin{aligned}\boldsymbol{v}(\boldsymbol{u},\boldsymbol{B}) =&\nabla_{\boldsymbol{u},\boldsymbol{B}}\ m_{\psi_k}(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x}) &(14)\\ =&\begin{pmatrix}\frac{1}{2}\left(\boldsymbol{x}\boldsymbol{x}^T-\tilde{\boldsymbol{\Sigma}}_S-\tilde{\boldsymbol{\mu}}_S\tilde{\boldsymbol{\mu}}_S^T\right)^\flat\\ \tilde{\boldsymbol{\mu}}_S-\boldsymbol{x}\end{pmatrix}\\ &\cdot\mathrm{e}^{h(\boldsymbol{u},\boldsymbol{B};\boldsymbol{x})}\mathcal{N}(\boldsymbol{0},\boldsymbol{I};\boldsymbol{x})\psi_k(\boldsymbol{x}) &(15)\end{aligned}$$

observe that the randomness of $\boldsymbol{v}$ only comes from the random sample $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)$. The fact that $\boldsymbol{v}(\boldsymbol{u},\boldsymbol{B})$ is an unbiased estimator of $\nabla M_f(\boldsymbol{u},\boldsymbol{B})$ follows directly from the fact calculation of $\nabla M_f(\boldsymbol{u},\boldsymbol{B})$ in Section IV-A1. For the other two properties that we need we have the following lemmas. The following lemma bounds the variance of the step of the SGD algorithm. It's rather technical proof can be found in the full version of the paper.

Lemma 16 (BOUNDED STEP VARIANCE)**.** *Let α be the constant of* (3)*. For every $(\boldsymbol{u},\boldsymbol{B})\in\mathcal{D}$ it holds*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_S^*}\left[\left\|\boldsymbol{v}(\boldsymbol{u},\boldsymbol{B})\right\|_2^2\right] \leq \mathrm{poly}(1/\alpha)\cdot d^{2k},$$

We are now going to prove the strong convexity of the objective function M_{ψ_k}. For this we are going to use the anti-concentration result, Theorem 1) for polynomial functions over the Gaussian measure.

The following lemma shows that our objective is strongly convex as long as the guess $\boldsymbol{u}, \boldsymbol{B}$ remains in the set $\mathcal{D}$. For the proof, we refer the reader to the full version of the paper.

Lemma 17 (STRONG CONVEXITY)**.** *Let α be the absolute constant of* (3)*. For every $(\boldsymbol{u}, \boldsymbol{B}) \in \mathcal{D}$, any $\boldsymbol{z} \in \mathbb{R}^d$ such that $\|\boldsymbol{z}\|_2 = 1$ and the first d^2 coordinated of $\boldsymbol{z}$ correspond to a symmetric matrix, then*

$$\boldsymbol{z}^T \mathcal{H}_{M_f}(\boldsymbol{u}, \boldsymbol{B}) \boldsymbol{z} \geq \mathrm{poly}(\alpha),$$

3) Recovering the Unconditional Mean and Covariance: The framework that we use for proving the fast convergence of our SGD algorithm is summarized in the following theorem and the following lemma.

Theorem 6 (Theorem 14.11 of [SSBD14].)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$. Assume that f is λ-strongly convex, that $\mathbb{E}\left[\boldsymbol{v}^{(i)} \mid \boldsymbol{w}^{(i-1)}\right] \in \partial f(\boldsymbol{w}^{(i-1)})$ and that $\mathbb{E}\left[\left\|\boldsymbol{v}^{(i)}\right\|_2^2\right] \leq \rho^2$. Let $\boldsymbol{w}^* \in \arg\min_{\boldsymbol{w} \in \mathcal{D}} f(\boldsymbol{w})$ be an optimal solution. Then,*

$$\mathbb{E}\left[f(\bar{\boldsymbol{w}})\right] - f(\boldsymbol{w}^*) \leq \frac{\rho^2}{2\lambda T}\left(1 + \log T\right),$$

where $\bar{\boldsymbol{w}}$ is the output projected stochastic gradient descent with steps $\boldsymbol{v}^{(i)}$ and projection set $\mathcal{D}$ after T iterations.

Lemma 18 (Lemma 13.5 of [SSBD14].)**.** *If f is λ-strongly convex and $\boldsymbol{w}^*$ is a minimizer of f, then, for any $\boldsymbol{w}$ it holds that*

$$f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \geq \frac{\lambda}{2}\left\|\boldsymbol{w} - \boldsymbol{w}^*\right\|_2^2.$$

Now we have all the ingredients to present the proof of Theorem 5.

Proof of Theorem 5: The estimation procedure starts by computing the polynomial function ψ_k using $d^{\mathrm{poly}(1/\alpha^*)\frac{\mathrm{r}^2(S)}{\varepsilon'^8}}$ samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$ as explained in Theorem 4 to get error $\mathrm{poly}(\alpha^*)\varepsilon'^2$. Then we compute $\tilde{\boldsymbol{\mu}}_S$ and $\tilde{\boldsymbol{\Sigma}}_S$ as explained in Section II with $\varepsilon = \frac{q(\alpha^*)}{8p(1/\alpha^*)}(\varepsilon')^2$ where p comes from Lemma 15 and q comes from Lemma 17. Our estimators for $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ are the outputs of Algorithm 2.

We analyze the accuracy of our estimation by proving that the minimum of M_{ψ_k} is close in the parameter space to the minimum of M'_{ψ}. Let $\boldsymbol{u}', \boldsymbol{B}'$ be the minimum of the convex function M'_{ψ} and $\boldsymbol{u}_k, \boldsymbol{B}_k$ be the minimum of the convex function M_{ψ_k}. Using Lemma 15 we have the following relations

$$\left|M'_{\psi}(\boldsymbol{u}', \boldsymbol{B}') - M_{\psi_k}(\boldsymbol{u}', \boldsymbol{B}')\right| \leq \varepsilon',$$
$$\left|M'_{\psi}(\boldsymbol{u}_k, \boldsymbol{B}_k) - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k)\right| \leq \varepsilon'$$

and also

$$M'_{\psi}(\boldsymbol{u}', \boldsymbol{B}') \leq M'_{\psi}(\boldsymbol{u}_k, \boldsymbol{B}_k),$$
$$M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k) \leq M_{\psi_k}(\boldsymbol{u}', \boldsymbol{B}').$$

These relations imply that

$$\begin{aligned}
&\left|M_{\psi_k}(\boldsymbol{u}', \boldsymbol{B}') - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k)\right| \\
&= M_{\psi_k}(\boldsymbol{u}', \boldsymbol{B}') - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k) \\
&\leq M_{\psi_k}(\boldsymbol{u}', \boldsymbol{B}') - M'_{\psi}(\boldsymbol{u}', \boldsymbol{B}') \\
&+ M'_{\psi}(\boldsymbol{u}_k, \boldsymbol{B}_k) - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k) \\
&\leq \left|M'_{\psi}(\boldsymbol{u}', \boldsymbol{B}') - M_{\psi_k}(\boldsymbol{u}', \boldsymbol{B}')\right| \\
&+ \left|M'_{\psi}(\boldsymbol{u}_k, \boldsymbol{B}_k) - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k)\right| \leq 2\varepsilon'.
\end{aligned}$$

But from Lemma 17 and Lemma 18 we get that $\left\|\begin{pmatrix} \boldsymbol{B}'^{\flat} \\ \boldsymbol{u}' \end{pmatrix} - \begin{pmatrix} \boldsymbol{B}_k^{\flat} \\ \boldsymbol{u}_k \end{pmatrix}\right\|_2 \leq \frac{\varepsilon'}{2}$. Now we can apply the Claim 2 which implies that

$$\left\|\begin{pmatrix} (\boldsymbol{\Sigma}^{*-1})^{\flat} \\ \boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^* \end{pmatrix} - \begin{pmatrix} \boldsymbol{B}_k^{\flat} \\ \boldsymbol{u}_k \end{pmatrix}\right\|_2 \leq \frac{\varepsilon'}{2}. \tag{16}$$

Therefore it suffices to find $(\boldsymbol{u}_k, \boldsymbol{B}_k)$ with accuracy $\varepsilon'/2$ to get our theorem.

Let $\boldsymbol{w}^* = \begin{pmatrix} \boldsymbol{B}_k^{\flat} \\ \boldsymbol{u}_k \end{pmatrix}$ To prove that Algorithm 2 converges to $\boldsymbol{w}^*$ we use Theorem 6 which together with Markov's inequality, Lemma 16 and Lemma 17 gives us

$$\begin{aligned}
\mathbb{P}\Bigg(& M_{\psi_k}(\hat{\boldsymbol{u}}, \hat{\boldsymbol{B}}) - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k) \\
& \geq \mathrm{poly}(1/\alpha^*) \cdot \frac{d^{2k}}{T}\left(1 + \log(T)\right)\Bigg) \leq \frac{1}{3}.
\end{aligned} \tag{17}$$

To get our estimation we first repeat the SGD procedure $K = \log(1/\delta)$ times independently, with parameters T, λ each time. We then get the set of estimates $\mathcal{E} = \{\bar{\boldsymbol{w}}_1, \bar{\boldsymbol{w}}_2, \ldots, \bar{\boldsymbol{w}}_K\}$. Because of (17) we know that, with high probability $1 - \delta$, for at least the 2/3 of the points $\bar{\boldsymbol{w}}$ in $\mathcal{E}$ it is true that $M_{\psi_k}(\boldsymbol{w}) - M_{\psi_k}(\boldsymbol{w}^*) \leq \eta$ where $\eta = \mathrm{poly}(1/\alpha^*) \cdot \frac{d^{2k}}{T}\left(1 + \log(T)\right)$. Moreover we will prove later that $M_{\psi_k}(\boldsymbol{w}) - M_{\psi_k}(\boldsymbol{w}^*) \leq \eta$ and this implies $\|\boldsymbol{w} - \boldsymbol{w}^*\| \leq c \cdot \eta$, where c is a universal constant. Therefore with high probability $1 - \delta$ for at least the 2/3 of the points $\bar{\boldsymbol{w}}, \bar{\boldsymbol{w}'}$ in $\mathcal{E}$ it is true that $\|\boldsymbol{w} - \boldsymbol{w}'\| \leq 2c \cdot \eta$. Hence if we set $\hat{\boldsymbol{w}}$ to be a point that is at least $2c \cdot \eta$ close to more that the half of the points in $\mathcal{E}$ then with high probability $1 - \delta$ we have that $f(\bar{\boldsymbol{w}}) - f(\boldsymbol{w}^*) \leq \eta$. Hence we can we lose probability at most δ if we condition on the event

$$\begin{aligned}
& M_{\psi_k}(\hat{\boldsymbol{u}}, \hat{\boldsymbol{B}}) - M_{\psi_k}(\boldsymbol{u}_k, \boldsymbol{B}_k) \\
& \quad \leq \mathrm{poly}(1/\alpha^*) \cdot \frac{d^{2k}}{T}\left(1 + \log(T)\right).
\end{aligned}$$

Using once again Lemma 18 we get that

$$\left\|\begin{pmatrix} \hat{\boldsymbol{B}}^{\flat} \\ \hat{\boldsymbol{u}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{B}_k^{\flat} \\ \boldsymbol{u}_k \end{pmatrix}\right\|_2 \leq \frac{\varepsilon'}{2}.$$

Figure 2. Projected Stochastic Gradient Descent given access to samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$.

$\boldsymbol{w}^{(0)} = \begin{pmatrix} (\boldsymbol{B}^{(0)})^\flat \\ \boldsymbol{u}^{(0)} \end{pmatrix} \leftarrow \begin{pmatrix} (\tilde{\boldsymbol{\Sigma}}_S^{-1})^\flat \\ \tilde{\boldsymbol{\Sigma}}_S^{-1} \tilde{\boldsymbol{\mu}}_S \end{pmatrix}$

for $i = 1, \ldots, T$ **do**

  Sample $\boldsymbol{x}^{(i)}$ from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$

  $\eta_i \leftarrow \frac{1}{\lambda \cdot i}$

  $\begin{pmatrix} (\boldsymbol{B}^{(i-1)})^\flat \\ \boldsymbol{u}^{(i-1)} \end{pmatrix} \leftarrow \boldsymbol{w}^{(i-1)}$

  $$\boldsymbol{v}^{(i)} \leftarrow \begin{pmatrix} \frac{1}{2}\left(\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T} - \tilde{\boldsymbol{\Sigma}}_S - \tilde{\boldsymbol{\mu}}_S \tilde{\boldsymbol{\mu}}_S^T\right)^\flat \\ \tilde{\boldsymbol{\mu}}_S - \boldsymbol{x}^{(i)} \end{pmatrix} \mathrm{e}^{h(\boldsymbol{u}^{(i-1)}, \boldsymbol{B}^{(i-1)}; \boldsymbol{x}^{(i)})} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}; \boldsymbol{x}^{(i)}) \psi_k\left(\boldsymbol{x}^{(i)}\right)$$

  {From Equation (13).}

  $\boldsymbol{r}^{(i)} \leftarrow \boldsymbol{w}^{(i-1)} - \eta_i \boldsymbol{v}^{(i)}$

  $\boldsymbol{w}^{(i)} \leftarrow \arg\min_{\boldsymbol{w} \in \mathcal{D}} \left\|\boldsymbol{w} - \boldsymbol{r}^{(i)}\right\|_2^2$
  {From Lemma 8 of [DGTZ18].}

end for

$\begin{pmatrix} \hat{\boldsymbol{B}}^\flat \\ \hat{\boldsymbol{u}} \end{pmatrix} \leftarrow \frac{1}{T} \sum_{i=1}^T \boldsymbol{w}^{(i)}$

$\hat{\boldsymbol{\Sigma}} \leftarrow \hat{\boldsymbol{B}}^{-1}$

$\hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{B}}^{-1} \hat{\boldsymbol{u}}$

return $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$

which together with (16) implies

$$\left\| \begin{pmatrix} \hat{\boldsymbol{B}}^\flat \\ \hat{\boldsymbol{u}} \end{pmatrix} - \begin{pmatrix} (\boldsymbol{\Sigma}^{*-1})^\flat \\ \boldsymbol{\Sigma}^{*-1} \boldsymbol{\mu}^* \end{pmatrix} \right\|_2 \leq \frac{\varepsilon'}{2}.$$

and the theorem follows as closeness in parameter distance implies closeness in total variation distance for the corresponding untruncated Gaussian distributions. □

V. Lower Bound for Learning the Mean of a Truncated Normal

Theorem 7. *There exists a family of sets $\mathcal{S}$ with $\Gamma(\mathcal{S}) = O(d)$ such that any algorithm that draws m samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}, S)$ and computes an estimate $\widetilde{\boldsymbol{\mu}}$ with $\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq 1$ must have $m = \Omega(2^{d/2})$.*

Proof: Let $H = [-1, 1]^{d+1}$ be the $d+1$-dimensional cube. We will also use the left and right subcubes $H_+ = [-1, 0] \times [-1, 1]^d$, $H_- = [0, 1] \times [-1, 1]^d$ respectively. Let $\mathcal{N}_+ = \mathcal{N}(\boldsymbol{e}_1, \boldsymbol{I})$ and $\mathcal{N}_- = \mathcal{N}(-e_1, \boldsymbol{I})$. We denote by r the (scaled) pointwise minimum of the two densities truncated at the cube H, that is

$$\begin{aligned} r(x) &= \frac{\min(\mathcal{N}_+(H; \boldsymbol{x}), \mathcal{N}_-(H; \boldsymbol{x}))}{c} \\ &= \frac{\boldsymbol{1}_H(\boldsymbol{x})}{c} \min(\mathcal{N}_+(\boldsymbol{x}), \mathcal{N}_-(\boldsymbol{x})), \end{aligned}$$

where $c = 1 - d_{\mathrm{TV}}(\mathcal{N}_+, \mathcal{N}_-)$.

To simplify notation we assume that we work in $\mathbb{R}^{d+1}$ instead of $\mathbb{R}^d$. Let $V = (v_1, \ldots, v_d) \in \{+1, -1\}^d$. For every V we define the set $G_V = H \cap \{\boldsymbol{y} \in \mathbb{R}^d : y_i v_i \geq 0\}$. We also define the subcubes $H_V = [0, 1] \times G_V$. We consider the following subset of H parameterized by the 2^d parameters $t_V \in [0, 1]$ and $\delta \in [-1, 1]$.

$$S_+ = [-1 + \delta, 0] \times [-1, 1]^d \cup \bigcup_{V \in \{-1,+1\}^d} [0, t_V] \times G_V$$

We will argue that there exists a distribution D^+ on the values t_V such that on expectation $d_{\mathrm{TV}}(\mathcal{N}_+^{S_+}, \mathcal{N}_-^{S_-})$ is $O(2^{-d})$. We show how to construct the distribution D_+ since the construction for D_- is the same. In fact we will show that both distributions are very close to $r(x)$. Notice that for some $(t, \boldsymbol{y}) \in \mathbb{R}^{d+1}$ we have We draw each t_V independently from the distribution with cdf

$$F(t) = \boldsymbol{1}_{[0,1)}(t)(1 - \mathrm{e}^{-2t}) + \boldsymbol{1}_{[1,+\infty)}(t)$$

Notice that for $t \in (0, 1)$ and any $\boldsymbol{y} \in \mathbb{R}^d$ we have that $1 - F(t) = \mathcal{N}_-(t, \boldsymbol{y}) / \mathcal{N}_+(t, \boldsymbol{y})$.

After we draw all t_V from F we choose δ so that $\mathcal{N}_+(S_+; x) = c$. We will show that on expectation over the t_V we have $\delta = 0$, which means that no correction is needed. In fact we show something stronger, namely that for all $x \in H_+$ we have that $\mathbb{E}_{S_+ \sim D_+}[N_+(S_+; \boldsymbol{x})] = r(\boldsymbol{x})$. Assume that $x \in H_V$. Indeed,

$$\begin{aligned} &\mathop{\mathbb{E}}_{S_+ \sim D_+}[\mathcal{N}_+(S_+; \boldsymbol{x})] \\ &= \frac{\mathcal{N}_+(\boldsymbol{x})}{c} \mathop{\mathbb{E}}_{S_+ \sim D_+}[\boldsymbol{1}_{S_+}(\boldsymbol{x})] = \frac{\mathcal{N}_+(\boldsymbol{x})}{c} \mathop{\mathbb{E}}_{S_+ \sim D_+}[\boldsymbol{1}_{\{x_1 \leq t_V\}}] \\ &= \frac{\mathcal{N}_+(\boldsymbol{x})}{c}(1 - F(t_V)) = \frac{\mathcal{N}_-(\boldsymbol{x})}{c} = r(\boldsymbol{x}) \end{aligned}$$

Moreover, observe that for all $\boldsymbol{x} \in H_- \cap S_+$ we have that $N_+(S_+; \boldsymbol{x}) = r(\boldsymbol{x})$ always (with probability 1). We now argue that in order to have constant probability to distinguish $N_+(S_+)$ from $r(x)$ one needs to draw $\Omega(2^d)$ samples. Since the expected density of $N_+(S_+)$ matches $r(x)$ for all $x \in H_+$, to be able to distinguish the two distributions one needs to observe at least two samples in the same cube H_V. Since we have 2^d disjoint cubes H_V the probability of a sample landing in H_V is at most $1/2^d$. Therefore, using the birthday problem, to have constant probability to observe a collision one needs to draw $\Omega(\sqrt{2^d}) = \Omega(2^{d/2})$ samples. Since for all $x \in H_- \cap S_+$, $N_+(S_+)$ exactly matches $r(x)$, to distinguish between the two distributions one needs to observe a sample $\boldsymbol{x}$ with $-1 + \delta < x_1 < -1$. Due to symmetry, N_+ assigns

to all cubes H_V equal probability, call that p. Moreover, we have that $c = 2^{d+1}p$. Now let p_V be the random variable corresponding to the probability that N_+ assigns to $[0, t_V] \times G_V$. We have that $\mathbb{E}_{t_V \sim F}[p_V] = p$ for all V. Since the independent random variables p_V are bounded in $[0, 1/2^d]$, Hoeffding's inequality implies that $|\sum_{V \in \{-1,1\}^d}(p_V - p)| < 1/2^{d/2}$ with probability at least $1 - 2/\mathrm{e}^2$. This means that with probability at least $3/4$ one will need to draw $\Omega(2^{d/2})$ samples in order to observe one with $x_1 < -1 + \delta$.

Since any set S in our family $\mathcal{S}$ has almost everywhere (that is except the set of its vertices which a finite set and thus of measure zero) smooth boundary we may use the following equivalent (see e.g. [Naz03]) definition of its surface area

$$\Gamma(S) = \int_{\partial S} \mathcal{N}_0(\boldsymbol{x}) \mathrm{d}\sigma(\boldsymbol{x}),$$

where $\mathrm{d}\sigma(x)$ is the standard surface measure on $\mathbb{R}^d$. Without loss of generality we assume that S corresponds to the set S_+ defined above (the proof is the same if we consider a set S_-). We have

$$\partial S \subseteq \bigcup_{V \in \{+1,-1\}^d} (\{t_V\} \times G_V) \cup \partial([-1,+1]^{d+1}) \cup \bigcup_{i=1}^{d+1} \{\boldsymbol{x} : x_i = 0\}.$$

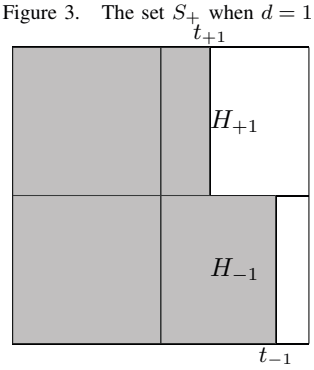
By the definition of Gaussian surface area it is clear that $\Gamma(A \cup B) \leq \Gamma(A) + \Gamma(B)$. From Table I we know that $\Gamma([-1,+1]^{d+1}) = O(\sqrt{\log d})$. Moreover, we know that a single halfspace has surface area at most $\sqrt{2/\pi}$ (see e.g. [KOS08]). Therefore $\Gamma\left(\bigcup_{i=1}^{d+1}\{x : x_i = 0\}\right) \leq \sum_{i=1}^{d+1}\sqrt{2/\pi} = O(d)$. Finally, we notice that for any point x on the hyperplane $\{\boldsymbol{x} : x_1 = 0\}$ and any $\boldsymbol{y}$ on $\{\boldsymbol{x} : x_1 = c\}$ (for any $c \geq 0$), we have $\mathcal{N}_0(\boldsymbol{x}) \geq \mathcal{N}_0(\boldsymbol{y})$. Therefore, the surface area of each set $t_V \times G_V$ is maximized for $t_V = 0$. In this case $\bigcup_{V \in \{+1,-1\}^d}(\{t_V\} \times G_V) \subseteq \{\boldsymbol{x} : x_1 = 0\}$, which implies that the set $\bigcup_{V \in \{+1,-1\}^d}(\{t_V\} \times G_V)$ contributes at most $\sqrt{2/\pi}$ to the total surface area. Putting everything together, we have that $\Gamma(S) = O(d)$.

■

VI. Identifiability with bounded Gaussian Surface Area

In this section we investigate the sample complexity of the problem of estimating the parameters of a truncated Gaussian using a different approach that does not depend on the VC dimension of the family $\mathcal{S}$ of the truncation sets to be finite. For example, we settle the sample complexity of learning the parameters of a Gaussian distribution truncated at an unknown convex set (recall that the class of convex sets has infinite VC dimension). Our method relies on finding a tuple $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})$ of parameters so that the moments of the corresponding truncated Gaussian $\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})$ are all close to the moments of the unknown truncated Gaussian distribution, for which we have unbiased estimates using samples. The main question that we need to answer to determine the sample complexity of this problem is how many moments are needed to be matched in order to be sure that our guessed parameters are close to the parameters of the unknown truncated Gaussian. We state now the main result. Its proof is based on Lemma 20 and can be found in the full version of the paper.

Figure 3. The set S_+ when $d = 1$.



Theorem 8 (Moment Matching)**.** *Let $\mathcal{S}$ be a family of subsets of $\mathbb{R}^d$ of bounded Gaussian surface area $\Gamma(\mathcal{S})$. Moreover, assume that if T is an affine map and $T(\mathcal{S}) = \{T(S) : S \in \mathcal{S}\}$ is the family of the images of the sets of $\mathcal{S}$, then it holds $\Gamma(T(\mathcal{S})) = O(\Gamma(\mathcal{S}))$. For some $S \in \mathcal{S}$, let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ be an unknown truncated Gaussian. $d^{O(\Gamma(\mathcal{S})/\varepsilon^4)}$ samples are sufficient to find parameters $\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S}$ such that $d_{\mathrm{TV}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S), \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{S})) \leq \varepsilon$.*

The key lemma of this section is Lemma 20. It shows that if two truncated normals are in total variation distance ε then there exists a moment where they differ. The main idea is to prove that there exists a polynomial that approximates well the indicator of the set $\{f_1 > f_2\}$. Notice that the total variation distance between two densities can be written as $\int \mathbf{1}_{\{f_1 > f_2\}}(\boldsymbol{x}) f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) \mathrm{d}x$. In our proof we use the chi squared divergence, which for two distributions with densities f_1, f_2 is defined as

$$D_{\chi^2}(f_1 \| f_2) = \int \frac{(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))^2}{f_2(\boldsymbol{x})} \mathrm{d}x$$

To prove it we need the following nice fact about chi squared divergence between Gaussian distributions. In general chi squared divergence may be infinite for some pairs of Gaussians. In the following lemma we prove that for any pair of Gaussians, there exists another Gaussian N such that $D_{\chi^2}(N_1 \| N)$ $D_{\chi^2}(N_2 \| N)$ are finite even if $D_{\chi^2}(N_1 \| N_2) = \infty$.

Lemma 19. *Let $N_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, and $N_2 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_2)$ be two Normal distributions that satisfy the conditions of*

Lemma 3. Then there exists a Normal distribution N such that

$$D_{\chi^2}(N_1\|N), D_{\chi^2}(N_2\|N) \leq \exp\left(2\left\|\boldsymbol{\Sigma}_1^{-1/2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right\|_2 + \frac{1}{2}\max(1,\|\boldsymbol{\Sigma}_1\|_2)\left\|\boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2}-\boldsymbol{I}\right\|_F^2\right)$$

Now we state the main lemma of this section. We give here a sketch of its proof. For a formal proof we refer the reader to the full version of the paper.

Lemma 20. *Let $\mathcal{S}$ be a family of subsets of $\mathbb{R}^d$ of bounded Gaussian surface area $\Gamma(\mathcal{S})$. Moreover, assume that if T is an affine map and $T(\mathcal{S}) = \{T(S) : S \in \mathcal{S}\}$ is the family of the images of the sets of $\mathcal{S}$, then it holds $\Gamma(T(\mathcal{S})) = O(\Gamma(\mathcal{S}))$. Let $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, S_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, S_2)$ be two truncated Gaussians with densities f_1, f_2 respectively. Let $k = O(\Gamma(\mathcal{S})/\varepsilon^4)$. If $d_{\mathrm{TV}}(f_1, f_2) \geq \varepsilon$, then there exists a $V \in \mathbb{N}^d$ with $|V| \leq k$ such that*

$$\left|\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1,S_1)}[\boldsymbol{x}^V] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2,S_2)}[\boldsymbol{x}^V]\right| \geq \varepsilon/d^{O(k)}.$$

Proof sketch.: Let $W = S_1 \cap S_2 \cap \{f_1 > f_2\} \cup S_1 \setminus S_2$, that is the set of points where the first density is larger than the second. We now write the L_1 distance between f_1, f_2 as

$$\int |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})|\mathrm{d}\boldsymbol{x} = \int \mathbf{1}_W(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$$

Denote $p(\boldsymbol{x})$ the polynomial that will do the approximation of the L_1 distance. From Lemma 19 we know that there exists a Normal distribution within small chi-squared divergence of both $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Call the density function of this distribution $g(\boldsymbol{x})$. We have

$$\begin{aligned}
&\left|\int |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})|\mathrm{d}\boldsymbol{x} - \int p(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))\right| \quad (18)\\
&\leq \int |\mathbf{1}_W(\boldsymbol{x}) - p(\boldsymbol{x})|\ |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})|\mathrm{d}\boldsymbol{x}\\
&\leq \int |\mathbf{1}_W(\boldsymbol{x}) - p(\boldsymbol{x})|\sqrt{g(\boldsymbol{x})}\ \frac{|f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})|}{\sqrt{g(\boldsymbol{x})}}\mathrm{d}x\\
&\leq \sqrt{\int (\mathbf{1}_W(\boldsymbol{x}) - p(\boldsymbol{x}))^2 g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}\\
&\cdot \sqrt{\int \frac{(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))^2}{g(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}}, \quad (19)
\end{aligned}$$

where we use Schwarzs' inequality. From Lemma 19 we know that

$$\int \frac{f_1(\boldsymbol{x})^2}{g(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} \leq \int \frac{\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1; \boldsymbol{x})^2}{g(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \exp(\mathrm{poly}(1/\alpha)).$$

Similarly, $\int \frac{f_2(\boldsymbol{x})^2}{g(\boldsymbol{x})}\mathrm{d}x = \exp(\mathrm{poly}(1/\alpha))$. Therefore, we have,

$$\begin{aligned}
&\left|\int |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})|\mathrm{d}\boldsymbol{x} - \int p(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))\right|\\
&\leq \exp(\mathrm{poly}(1/\alpha))\sqrt{\int (\mathbf{1}_W(\boldsymbol{x}) - p(\boldsymbol{x}))^2 g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}
\end{aligned}$$

Recall that $g(\boldsymbol{x})$ is the density function of a Gaussian distribution, and let $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ be the parameters of this Gaussian. Notice that it remains to show that there exists a good approximating polynomial $p(\boldsymbol{x})$ to the indicator function $\mathbf{1}_W$. We can now transform the space so that $g(\boldsymbol{x})$ becomes the standard normal. Notice that this is an affine transformation that also transforms the set W; Since the Gaussian surface area is "invariant" under linear transformations

Since $\mathbf{1}_W \in L^2(\mathbb{R}^d, \mathcal{N}_0)$ we can approximate it using Hermite polynomials. For some $k \in \mathcal{N}$ we set $p(\boldsymbol{x}) = S_k\mathbf{1}_W(x)$, that is

$$p_k(\boldsymbol{x}) = \sum_{V:|V|\leq k} \widehat{\mathbf{1}_W} H_V(\boldsymbol{x}).$$

Combining Lemma 10 and Lemma 4 we obtain

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}_0}[(\mathbf{1}_W(\boldsymbol{x}) - p_k(x))^2] = O\left(\frac{\Gamma(\mathcal{S})}{k^{1/2}}\right).$$

Therefore, $\left|\int |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})|\mathrm{d}\boldsymbol{x} - \int p_k(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))\right| = \exp(\mathrm{poly}(1/\alpha))\frac{\Gamma(\mathcal{S})^{1/2}}{k^{1/4}}$. Ignoring the dependence on the absolute constant α, to achieve error $O(\varepsilon)$ we need degree $k = O(\Gamma(\mathcal{S})^2/\varepsilon^4)$.

To complete the proof, it remains to obtain a bound for the coefficients of the polynomial $q(\boldsymbol{x}) = p_k(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}))$. Using known facts about the coefficients of Hermite polynomials we obtain that $\|q(\boldsymbol{x})\|_\infty \leq \binom{d+k}{k}^2(4d)^{k/2}(O(1/\alpha^2))^k$. To conclude the proof we notice that we can pick the degree k so that

$$\begin{aligned}
&\left|\int q(\boldsymbol{x})(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))\right|\\
&\quad = \left|\sum_{V:|V|\leq k} \boldsymbol{x}^V(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))\right| \geq \varepsilon/2.
\end{aligned}$$

Since the maximum coefficient of $q(\boldsymbol{x})$ is bounded by $d^{O(k)}$ we obtain the result. ∎

VII. VC-dimension vs Gaussian Surface Area

We use two different complexity measures of the truncation set to get sample complexity bounds, the VC-dimension and the Gaussian Surface Area (GSA) of the class of the sets. As we already mentioned in the introduction, there are classes, for example convex sets, that have bounded Gaussian surface area but infinite VC-dimension. However, this is not the main difference between the two complexity

measures in our setting. Having a class with bounded VC-dimension means that the empirical risk minimization needs finite samples. To get an efficient algorithm we still need to *implement the ERM for this specific class*. Therefore, it is not clear whether it is possible to get an algorithm that works for all sets of bounded VC-dimension. On the other hand, bounded GSA means that we can approximate the weighted indicator of the set using its low order Hermite coeffients. This approximation works for all sets of bounded GSA and does not depend on the specific class of sets. Therefore, using GSA we manage to get a unified approach that learns the parameters of the underlying Gaussian distribution using only the assumption that the truncation set has bounded GSA. In other words, our approach uses the information of the class that the truncation set belongs only to decide how large the degree of the approximating polynomial should be. Having said that, it is an interesting open problem to design algorithms that learn the parameters of the Gaussian and use the information that the truncation set belongs to some class (e.g. intersection of k-halfspaces) to beat the runtime of our generic approach that only depends on the GSA of the class.

REFERENCES

[AGR13] Joseph Anderson, Navin Goyal, and Luis Rademacher. Efficient learning of simplices. In *Conference on Learning Theory*, pages 1020–1045, 2013.

[Bal93] Keith Ball. The reverse isoperimetric problem for gaussian measure. *Discrete & Computational Geometry*, 10(1):411–420, 1993.

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

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

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 47–60, 2017.

[CW01] Anthony Carbery and James Wright. Distributional and lq̂ norm inequalities for polynomials over convex bodies in rn̂. *Mathematical research letters*, 8(3):233–248, 2001.

[DDS14] Anindya De, Ilias Diakonikolas, and Rocco A Servedio. Learning from satisfying assignments. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 478–497. SIAM, 2014.

[Den98] François Denis. Pac learning from positive statistical queries. In *International Conference on Algorithmic Learning Theory*, pages 112–126. Springer, 1998.

[DGTZ18] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *the 59th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2018.

[DK14] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 1183–1213, 2014.

[DKK+16] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 655–664, 2016.

[DKK+17] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 999–1008, 2017.

[DKK+18] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 2683–2702, 2018.

[DL12] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

[Eld11] Ronen Eldan. A polynomial number of random points does not determine the volume of a convex body. *Discrete & Computational Geometry*, 46(1):29–47, 2011.

[Fis31] RA Fisher. Properties and applications of Hh functions. *Mathematical tables*, 1:815–852, 1931.

[FJK96] Alan Frieze, Mark Jerrum, and Ravi Kannan. Learning linear transformations. In *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, pages 359–368. IEEE, 1996.

[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.

[GR09] Navin Goyal and Luis Rademacher. Learning convex bodies is hard. *arXiv preprint arXiv:0904.1227*, 2009.

[Kan11] Daniel M Kane. The gaussian surface area and noise sensitivity of degree-d polynomial threshold functions. *computational complexity*, 20(2):389–412, 2011.

[KKMS05] Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2005), 23-25 October 2005, Pittsburgh, PA, USA, Proceedings*, pages 11–20, 2005.

[KOS08] Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning geometric concepts via gaussian surface area. In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, pages 541–550, 2008.

[LDG00] Fabien Letouzey, François Denis, and Rémi Gilleron. Learning from positive and unlabeled examples. In *International Conference on Algorithmic Learning Theory*, pages 71–85. Springer, 2000.

[Led94] Michel Ledoux. Semigroup proofs of the isoperimetric inequality in euclidean and gauss space. *Bulletin des sciences mathématiques*, 118(6):485–510, 1994.

[Lee14] Alice Lee. Table of the gaussian" tail" functions; when the" tail" is larger than the body. *Biometrika*, 10(2/3):208–214, 1914.

[LRV16] Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 665–674, 2016.

[Naz03] Fedor Nazarov. *On the Maximal Perimeter of a Convex Set in $\mathbb{R}^n$ with Respect to a Gaussian Measure*, pages 169–187. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

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

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

[Pis86] Gilles Pisier. Probabilistic methods in the geometry of banach spaces. In *Probability and analysis*, pages 167–241. Springer, 1986.

[PL08] Karl Pearson and Alice Lee. On the generalised probable error in multiple normal correlation. *Biometrika*, 6(1):59–68, 1908.

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

[SJ66] SM Shah and MC Jaiswal. Estimation of parameters of doubly truncated normal distribution from first four sample moments. *Annals of the Institute of Statistical Mathematics*, 18(1):107–111, 1966.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[Sze67] G. Szegö. *Orthogonal Polynomials*. Number τ. 23 in American Mathematical Society colloquium publications. American Mathematical Society, 1967.