Efficient Statistics, in High Dimensions, from Truncated Samples

Constantinos Daskalakis
EECS and CSAIL
MIT
Cambridge, USA
costis@csail.mit.edu

Themis Gouleakis
EECS and CSAIL
MIT
Cambridge, USA
tgoule@mit.edu

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

Manolis Zampetakis
EECS and CSAIL
MIT
Cambridge, USA
mzampet@mit.edu

Abstract—We provide an efficient algorithm for the classical problem, going back to Galton, Pearson, and Fisher, of estimating, with arbitrary accuracy the parameters of a multivariate normal distribution from truncated samples. Truncated samples from a d-variate normal means a samples is only revealed if it falls in S of the d-dimensional Euclidean space; otherwise the samples are hidden and their count in proportion to the revealed samples is also hidden. We show that the mean and covariance matrix can be estimated with arbitrary accuracy in polynomial-time, as long as we have oracle access to S, and S has non-trivial measure under the unknown d-variate normal distribution. Additionally we show that without oracle access to S, any non-trivial estimation is impossible.

Keywords—efficient statistics; truncated samples; normal distribution; high dimension;

I. INTRODUCTION

A classical challenge in Statistics is estimation from truncated or censored samples. Truncation occurs when samples falling outside of some subset S of the support of the distribution are not observed, and their count in proportion to the observed samples is also not observed. Censoring is similar except the fraction of samples falling outside of S is given. Truncation and censoring of samples have myriad manifestations in business, economics, manufacturing, engineering, quality control, medical and biological sciences, management sciences, social sciences, and all areas of the physical sciences. As a simple illustration, the values that insurance adjusters observe are usually left-truncated, right-censored, or both. Indeed, clients usually only report losses that are over their deductible, and may report their loss as equal to the policy limit when their actual loss exceeds the policy limit as this is the maximum that the insurance company would pay.

Statistical estimation under truncated or censored samples has had a long history in Statistics, going back to at least the work of Daniel Bernoulli who used it to demonstrate the efficacy of smallpox vaccination in 1760 [1]. In 1897, Galton analyzed truncated samples corresponding to registered speeds of American trotting horses [2]. His samples consisted of running times of horses that qualified for registration by trotting around a one-mile course in not more than 2 minutes and 30 seconds while harnessed to a two-wheeled cart carrying a weight of not less than 150 pounds including the driver. No records were kept for the slower, unsuccessful trotters, and their number thus remained unknown. Assuming that the running times prior to truncation were normal, Galton applied simple estimation procedures to estimate their mean and standard deviation. Dissatisfaction with Galton’s estimates led Pearson [3] and later Pearson and Lee [4] and Lee [5] to use the method of moments in order to estimate the mean and standard deviation of a univariate normal distribution from truncated samples. A few years later, Fisher used the method of maximum likelihood to estimate univariate normal distributions from truncated samples [6].

Following the early works of Galton, Pearson, Lee and Fisher, there has been a large volume of research devoted to estimating the truncated normal or other distributions; see e.g. [7], [8], [9] for an overview of this work. However, estimation methods, based on moments or maximum likelihood estimation, are intractable for high-dimensional data and are only known to be consistent in the limit, as the number of samples tends to infinity, even for normal distributions.

With infinitely many samples, it seems intuitive that given the density of a multivariate normal \( N(\mu, \Sigma) \) conditioned on a measurable set \( S \), the “local shape” of the density inside \( S \) should provide enough information to reconstruct the density everywhere. Indeed, results of Hotelling [10] and Tukey [11] prove that the conditional mean and variance on any measurable \( S \) are in one-to-one correspondence with the un-conditional parameters. When the sample is finite, however, it is not clear what features of the sample to exploit to estimate the parameters, and in particular it is unclear how sensitive to error is the correspondence between conditional and unconditional parameters. To illustrate, in Figure 1, we show one thousand samples from two bi-variate normals, which are far in total variation distance, truncated to a box. Distinguishing between the two Gaussians is not immediate despite the large total variation distance between these normals.

In this paper, we revisit this classical problem of multivariate normal estimation from truncated samples to obtain polynomial time and sample algorithms, while also accom-
modulating a very general truncation model. We suppose that samples, $X_1, X_2, \ldots$ from an unknown $d$-variate normal $\mathcal{N}(\mu, \Sigma)$ are only revealed if they fall into some subset $S \subseteq \mathbb{R}^d$; otherwise the samples are hidden and their count in proportion to the revealed samples is also hidden. We make no assumptions about $S$, except that its measure with respect to the unknown distribution is non-trivial, say $\alpha = 1\%$, and that we are given oracle access to this set, namely, given a point $x$ the oracle outputs $1_{x \in S}$. Otherwise, set $S$ can be any measurable set, and in particular need not be convex.

In contrast, to the best of our knowledge, prior work only considers sets $S$ that are boxes, while still not providing computational tractability, or finite sample bounds for consistent estimation. We provide the first time and sample efficient estimation algorithms even for simple truncation sets, but we also accommodate very general sets. This, in turn, enables statistical estimation in settings where set $S$ is determined by a complex set of rules, as it happens in many important applications, especially in high-dimensional settings. Revisiting our earlier example, insurance policies on a collection of risks may be complex, so the adjustor’s observation set $S$ may be determined by a complex function on the loss vector $X$.

Our main result is that the mean vector $\mu$ and covariance matrix $\Sigma$ of an unknown $d$-variate normal can be estimated to arbitrary accuracy in polynomial-time from a truncated sample. In particular,

**Theorem 1.** Given oracle access to a measurable set $S$, whose measure under some unknown $d$-variate normal $\mathcal{N}(\mu, \Sigma)$ is at least some constant $\alpha > 0$, and samples $X_1, X_2, \ldots$ from $\mathcal{N}(\mu, \Sigma)$ that are truncated to this set, there exists a polynomial-time algorithm that recovers estimates $\hat{\mu}$ and $\hat{\Sigma}$. In particular, for all $\epsilon > 0$, the algorithm uses $\tilde{O}(d^2/\epsilon^2)$ truncated samples and queries to the oracle and produces estimates that satisfy the following with probability at least $99\%$:

\[
\left\| \Sigma^{-1/2}(\mu - \hat{\mu}) \right\|_2 \leq \epsilon; \quad \text{and} \quad (1)
\]

\[
\left\| J - \Sigma^{-1/2} \Sigma \Sigma^{-1/2} \right\|_F \leq \epsilon. \quad (2)
\]

Note that under the above conditions the total variation distance between $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$ is $O(\epsilon)$, and the number of samples used by the algorithm is optimal, even when there is no truncation, i.e. when $S = \mathbb{R}^d$.

It is important to note that the measure $\alpha$ assigned by the unknown distribution on $S$ can be arbitrarily small, yet the accuracy of estimation can be driven to arbitrary precision. Moreover, we note that without oracle access to the indicator $1_{z \in S}$, it is information-theoretically impossible to even attain a crude approximation to the unknown normal, even in one dimension, namely,

**Theorem 2.** For all $\alpha > 0$, given infinitely many samples from a univariate normal $\mathcal{N}(\mu, \sigma^2)$, which are truncated to an unknown set $S$ of measure $\alpha$, it is impossible to estimate parameters $\hat{\mu}$ and $\hat{\sigma}^2$ such that the distributions $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ are guaranteed to be within $\frac{1}{1-\alpha}$.

**Overview of the Techniques:** The proofs of Theorems 1 and 2 are sketched in Sections III and IV respectively. Here we provide an overview of our proof of Theorem 1. Our algorithm shown in Figure 2 is (Projected) Stochastic Gradient Descent (SGD) on the negative log-likelihood of the truncated samples. Notice that we cannot write a closed-form of the log-likelihood as the set $S$ is arbitrary and unknown to us. Indeed, we only have oracle access to this set and can thus not write down a formula for the measure of $S$ under different estimates of the parameters. While we cannot write a closed-form expression for the negative log-likelihood, we still show that it is convex for arbitrary sets $S$, as long as we re-parameterize our problem in terms of $v = \Sigma^{-1/2} \mu$ and $T = \Sigma^{-1}$ (see Lemma 1). Using anti-concentration of polynomials of the Gaussian measure, we show that the negative log-likelihood is in fact strongly convex, with a constant that depends on the measure of $S$ under the current estimate $(v, T)$ of the parameters (see Lemma 4). In particular, to maintain strong convexity, SGD must remain within a region of parameters $(v, T)$ that assign non-trivial measure to $S$. We show that the pair of parameters $(v, T)$ corresponding to the conditional (on $S$) mean and covariance, which can be readily estimated from the truncated sample, satisfies this property (see Corollary 1). So we use these as our initial estimation of the parameters. Moreover, we define a convex set of parameters $(v, T)$ that all assign non-trivial measure to $S$. This set contains both our initialization and the ground truth (see Lemmas 7 and 6), and we can also efficiently project on it (see Lemma 8). So
we run our Projected Gradient Descent procedure on this set. As we have already noted, we have no closed-form for the log-likelihood or its gradient. Nevertheless, we show that, given oracle access to set $S$, we can get unbiased samples for the gradient of the log-likelihood function using rejection sampling from the normal distribution defined our current estimate of the parameters $(\nu, T)$ (see Lemma 9). For this additional reason, it is important to keep the invariant that SGD remains within a set of parameters that all assign nontrivial measure to $S$.

**Related work.** We have already discussed work on censored and truncated statistical estimation. More broadly, this problem falls in the realm of robust statistics, where there has been a strand of recent works studying robust estimation and learning in high dimensions. A celebrated result by Candes et al. [12] computes the PCA of a matrix, even when a constant fraction of its entries is adversarially corrupted, but it requires the locations of the corruptions to be relatively evenly distributed. Related work of Xu et al. [13] provides a robust PCA algorithm for arbitrary corruption locations, requiring at most 50% of the points to be corrupted.

Closer to our work, [14], [15], [16], [17] perform robust estimation of the parameters of multi-variate Gaussian distributions in the presence of corruptions to a small $\varepsilon$ fraction of the samples, allowing for both deletions of samples and additions of samples that can also be chosen adaptively (i.e. after seeing the sample generated by the Gaussian). The authors in [18] show that corruptions of an arbitrarily large fraction of samples can be tolerated as well, as long as we allow “list decoding” of the parameters of the Gaussian. In particular, they design learning algorithms that work when an $(1 - \alpha)$-fraction of the samples can be adversarially corrupted, but output a set of $\text{poly}(1/\alpha)$ answers, one of which is guaranteed to be accurate.

Similar to [18], we study a regime where only an arbitrarily small constant fraction of the samples from a normal distribution can be observed. In contrast to [18], however, there is a fixed set $S$ on which the samples are revealed without corruption, and we have oracle access to this set. The upshot is that we can provide a single accurate estimation of the normal rather than a list of candidate answers as in [18], while accommodating a much larger number of deletions of samples compared to [14], [17].

Other robust estimation works include robust linear regression [19] and robust estimation algorithms under sparsity assumptions [20], [21]. In [22], the authors study robust subspace recovery having both upper and lower bounds that give a trade-off between efficiency and robustness. Some general criteria for robust estimation are formulated in [23].

## II. Preliminaries

### A. Notation

We use small bold letters $x$ to refer to real vectors in finite dimension $\mathbb{R}^d$ and capital bold letters $A$ to refer to matrices in $\mathbb{R}^{d \times d}$. Similarly, a function with image in $\mathbb{R}^d$ is represented by a small and bold letter $f$. Let $(x, y)$ be the inner product of $x, y \in \mathbb{R}^d$. We use $I_d$ to refer to the identity matrix in $d$ dimensions. We may skip the subscript when the dimensions are clear. We use $E_{i,j}$ to refer to the all zero matrix with one 1 at the $(i, j)$ entry. Let $A \in \mathbb{R}^{d \times d}$, we define $A^\flat \in \mathbb{R}^{d^2}$ to be the standard vectorization of $A$. We define $\sharp$ to be the inverse operator, i.e. $(A^\sharp)^\sharp = A$. Let also $Q_d$ be the set of all the symmetric $d \times d$ matrices. $\mathbb{E}[X]$ is the expected value of the random variable $X$ and $\text{Var}[X]$ is the variance of $X$. The covariance matrix between two random variables $X, Y$ is $\text{Cov}[X, Y]$.

**Vector and Matrix Norms.** We define the $\ell_p$-norm of $x \in \mathbb{R}^d$ to be $\|x\|_p = (\sum_i x_i^p)^{1/p}$ and the $\ell_\infty$-norm of $x$ to be $\|x\|_\infty = \max_i |x_i|$. We also define the spectral norm of a matrix $A$ to be equal to

$$\|A\|_2 = \max_{x \in \mathbb{R}^d, x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$  

It is well known that $\|A\|_2 = \lambda_\text{max} A$, where $\lambda_i$‘s are the eigenvalues of $A$.

The Frobenius norm of a matrix $A$ is defined as follows:

$$\|A\|_F = \|A^\flat\|_2.$$  

The Mahalanobis distance between two vectors $x, y$ given a positive semidefinite matrix $\Sigma$ is defined as:

$$\|x - y\|_\Sigma = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$  

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

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

Also, let $\mathcal{N}(\mu, \Sigma; S)$ denote the probability mass of a set $S$ under this Gaussian measure.

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

$$\mathcal{N}(\mu, \Sigma, S; x) = \begin{cases} \mathcal{N}(\mu, \Sigma; x) & x \in S \\ \frac{\mathcal{N}(\mu, \Sigma; y) dy}{\int_S \mathcal{N}(\mu, \Sigma; y) dy} & 0 \ x \notin S \end{cases}.$$  

We will assume that the covariance matrix $\Sigma$ is full rank. The case where $\Sigma$ is not full rank can be easily detected by drawing $d$ samples and testing whether they are not linearly
independent. In that case, one can solve the estimation problem in the subspace that those samples span.

**Membership Oracle of a Set.** Let \( S \subseteq \mathbb{R}^d \) be a subset of the \( d \)-dimensional Euclidean space. A membership oracle of \( S \) is an efficient procedure \( M_S \) that computes the characteristic function of \( S \), i.e. \( M_S(x) = \mathbb{1}_{x \in S} \).

### 2. Useful Concentration Results

The following lemma is useful in cases where one wants to show concentration of a weighted sum of squares of i.i.d random variables.

**Theorem 3** (Lemma 1 of Laurent and Massart [24]). Let \( X_1, \ldots, X_n \) independent identically distributed random variables following \( \mathcal{N}(0,1) \) and let \( a \in \mathbb{R}^d \). Then, the following inequalities hold for any \( t \in \mathbb{R}^+ \):

\[
P \left( \sum_{i=1}^{d} a_i (X_i^2 - 1) \geq 2 \|a\|_2 \sqrt{t} + 2 \|a\|_\infty t \right) \leq \exp(-t),
\]

\[
P \left( \sum_{i=1}^{d} a_i (X_i^2 - 1) \leq -2 \|a\|_2 \sqrt{t} \right) \leq \exp(-t).
\]

The following matrix concentration result is also useful in order to show that the empirical covariance matrix of samples drawn from an identity covariance distribution is itself close to identity in the Frobenius norm.

**Theorem 4** (Corollary 4.12 of Diakonikolas et’al [25]). Let \( \rho, \tau > 0 \) and \( X_1, \ldots, X_n \) be i.i.d samples from \( \mathcal{N}(0, I) \). There is a \( \delta_2 = O(\rho \log(1/\rho)) \), such that if we draw \( n = \Omega(\frac{d^2+\log(1/\tau)}{\delta_2^2}) \), we get:

\[
P \left( \exists T \subseteq [n] : |T| \leq \rho n \text{ and } \left\| \frac{1}{|T|} \sum_{i \in T} X_i X_i^T - I \right\|_F \geq O \left( \delta_2 \frac{n}{|T|} \right) \right) \leq \tau
\]

Using the well known fact that the squared Frobenius norm of a symmetric matrix is equal to the sum of squares of its eigenvalues, we can obtain a bound on the \( l_2 \) distance of the eigenvalue vector to the all ones vector.

### 3. SGD for Learning Truncated Normals

In this section, we present and analyze our main algorithm for estimating the true mean and covariance matrix of the normal distribution from which the truncated samples are drawn. Our algorithm is (Projected) Stochastic Gradient Descent (SGD) on the negative log-likelihood of the truncated samples.

**A. Strong-convexity of the negative log-likelihood**

Let \( S \subseteq \mathbb{R}^d \) be a subset of the \( d \)-dimensional Euclidean space. We assume that we have access to \( n \) samples \( x_i \) from \( \mathcal{N}(\mu^*, \Sigma^*, S) \). We start by showing that the negative log-likelihood of the truncated samples is strongly convex as long as we re-parameterize our problem in terms of \( v = \Sigma^{-1/2} \mu \) and \( T = \Sigma^{-1} \).

1) **Log-Likelihood for a Single Sample:** Given one vector \( x \in \mathbb{R}^d \), the negative log-likelihood that \( x \) is a sample of the form \( \mathcal{N}(\mu, \Sigma, S) \) is given by:

\[
\ell(x, \Sigma; x) = \frac{1}{2} (x - \mu)\Sigma^{-1}(x - \mu) + \log \left( \int_S \exp \left( -\frac{1}{2} (z - \mu)^T \Sigma^{-1} (z - \mu) \right) dz \right)
\]

\[
= \frac{1}{2} x^T \Sigma^{-1} x - x^T \mu
\]

\[
+ \log \left( \int_S \exp \left( -\frac{1}{2} z^T \Sigma^{-1} z + z^T \mu \right) dz \right)
\]

Here we will define a different parametrization of the problem with respect to the variables \( T = \Sigma^{-1} \) and \( \nu = \Sigma^{-1/2} \mu \), where \( T \in \mathbb{R}^{d \times d} \) and \( \nu \in \mathbb{R}^d \). Then the likelihood function with respect to \( \nu \) and \( T \) is equal to:

\[
\ell(x, T; x) = \frac{1}{2} x^T T x - x^T \nu
\]

\[
+ \log \left( \int_S \exp \left( -\frac{1}{2} z^T T z + z^T \nu \right) dz \right)
\]

We now compute the gradient of \( \ell(x, T; x) \) with respect to the set of variables \( (T, \nu) \).

\[
\nabla \ell(x, T; x) = \left( \begin{array}{c}
- \frac{1}{2} x x^T \nu \\
- \frac{1}{2} x x^T z \\
\end{array} \right)
\]

\[
\int_S \left( \begin{array}{c}
- \frac{1}{2} x x^T \nu \\
- \frac{1}{2} x x^T z \\
\end{array} \right) \exp \left( -\frac{1}{2} z^T T z + z^T \nu \right) dz
\]

\[
= \left( \begin{array}{c}
- \frac{1}{2} x x^T \nu \\
- \frac{1}{2} x x^T z \\
\end{array} \right)
\]

\[
\int_S \left( \begin{array}{c}
- \frac{1}{2} x x^T \nu \\
- \frac{1}{2} x x^T z \\
\end{array} \right) \exp \left( -\frac{1}{2} z^T T z + z^T \nu \right) dz
\]

\[
= \left( \begin{array}{c}
- \frac{1}{2} x x^T \nu \\
- \frac{1}{2} x x^T z \\
\end{array} \right) + \mathbb{E}_{x \sim \mathcal{N}(T^{-1} \nu, T^{-1} S)} \left[ \left( \begin{array}{c}
- \frac{1}{2} x x^T \nu \\
- \frac{1}{2} x x^T z \\
\end{array} \right) \right]
\]

Finally, with similar calculations we compute the Hessian \( H_t \) of the log-likelihood function.
\( H_{l}(\nu, T) = \) \( \text{Cov}_{z \sim N(T^{-1}\nu, T^{-1}, S)} \left[ \left( -\frac{1}{2}zz^T \right)^{\lambda}, \left( -\frac{1}{2}zz^T \right)^{\lambda} \right] \).

Since the covariance matrix of a random variable is always positive semidefinite, we conclude that \( H_{l}(\nu, T) \) is positive semidefinite everywhere and hence, we have the following lemma.

**Lemma 1.** The function \( \ell(\nu, T^{-1}; x) \) is convex with respect to \( T^\nu \) for all \( x \in \mathbb{R}^d \).

2) Log-Likelihood in the Population Model: The negative log-likelihood function in the population model is equal to

\[
\tilde{\ell}(\nu, T) = \mathbb{E}_{z \sim N(\mu^*, \Sigma^*, S)} \left[ \frac{1}{2} z^T T x - x^T \nu \right] - \log \left( \int_S \exp \left( -\frac{1}{2} z^T T z + z^T \nu \right) dz \right)
\]

Also, using (8) we have that

\[
\nabla \tilde{\ell}(\nu, T) = -\mathbb{E}_{z \sim N(\mu^*, \Sigma^*, S)} \left[ \left( -\frac{1}{2} z^T \Sigma \right) \nu \right] + \mathbb{E}_{z \sim N(T^{-1}\nu, T^{-1}, S)} \left[ \left( -\frac{1}{2} z^T \Sigma \right) \nu \right].
\]

Hence from Lemma 1 we have that \( \tilde{\ell} \) is a convex function with respect to \( \nu \) and \( T \). Also, from (11) we get that the gradient \( \nabla \tilde{\ell}(\nu, T) \) is 0, when \( \mu^* = T^{-1}\nu \) and \( \Sigma^* = T^{-1} \).

From this observation together with the convexity of \( \ell \) we conclude that the true parameters, \( \nu^* = \Sigma^{*-1}\mu^* \) and \( T^* = \Sigma^{*-1} \), maximize the log-likelihood function.

**Lemma 2.** Let \( \nu^* = \Sigma^{*-1}\mu^* \), \( T^* = \Sigma^{*-1} \), then for any \( \nu \in \mathbb{R}^d \), \( T \in \mathbb{R}^{d \times d} \) it holds that

\[
\tilde{\ell}(\nu^*, T^*) \leq \tilde{\ell}(\nu, T).
\]

Finally, observe that the Hessian of \( \ell \) is the same as the Hessian of \( \tilde{\ell} \).

One property of the log likelihood that will be important later is its strong convexity.

**Definition 1 (Strong Convexity).** Let \( g : \mathbb{R}^d \rightarrow \mathbb{R} \), and let \( H_{g} \) be the Hessian of \( g \). We say that \( g \) is \( \lambda \)-strongly convex if \( H_{g} \succeq \lambda I \).

Our goal it to prove that \( \tilde{\ell} \) is strongly convex for any \( S \) such that \( N(\mu^*, \Sigma^*, S) \) is at least a constant \( \alpha \). We first prove strong concavity for the case \( S = \mathbb{R}^d \). For this, we need some definitions.

**Definition 2 (Minimum Eigenvalue of Fourth Moment Tensor).** Let \( \Sigma \in \mathbb{R}^{d \times d} \) and also \( \Sigma \succeq 0 \) with eigenvalues \( \lambda_1, \ldots, \lambda_n \), then we define the minimum eigenvalue \( \lambda_m(\Sigma) \) of the fourth moment tensor of \( N(0, \Sigma) \) as

\[
\lambda_m(\Sigma) = \min \left\{ \min_{i,j \in [d]} \lambda_i, \min_{i \in [d]} \lambda_i \right\}.
\]

**Lemma 3 (Strong Convexity without Truncation).** Let \( H_{\ell} \) be the Hessian of the negative log likelihood function \( \ell(\nu, T; X) \), when there is no truncation, i.e. \( S = \mathbb{R}^d \), then \( H_{\ell} \) as an operator on \( (\nu, T) \) with \( T \) being a symmetric matrix, satisfies

\[
H_{\ell} \succeq \lambda_m(T^{-1}) \cdot I.
\]

The proof of Lemma 3 can be found in the full version of the paper [26].

To prove strong convexity in the presence of truncation, we use the following anticoncentration bound of the Gaussian measure on sets characterized by polynomial threshold functions.

**Theorem 5 (Theorem 8 of [27]).** Let \( q, \gamma \in \mathbb{R}_+, \mu \in \mathbb{R}^d \), \( \Sigma \in \mathcal{Q}_d \) and \( p : \mathbb{R}^d \rightarrow \mathbb{R} \) be a multivariate polynomial of degree at most \( k \), we define

\[
S = \{ x \in \mathbb{R}^d \mid |p(x)| \leq \gamma \},
\]

then there exists an absolute constant \( C \) such that

\[
N(\mu, \Sigma; S) \leq \frac{Cq\gamma^{1/k}}{\left( \mathbb{E}_{z \sim N(\mu, \Sigma)} \left[ |p(z)|^q/k \right] \right)^{1/q}}.
\]

**Lemma 4 (Strong Convexity with Truncation).** Let \( H_{\ell} \) be the Hessian of the negative log likelihood function \( \ell(\nu, T; X) \), with the presence of arbitrary truncation \( S \). Let \( \mu = T^{-1}\nu \), \( \Sigma = T^{-1} \) and \( N(\mu, \Sigma; S) \geq \beta \), then \( H_{\ell} \) as an operator on \( (\nu, T) \) with \( T \) being a symmetric matrix, satisfies

\[
H_{\ell} \succeq \frac{1}{2^{13}} \left( \frac{\beta}{\gamma} \right)^4 \lambda_m(\Sigma) \cdot I,
\]

where \( C \) is the universal constant guaranteed by Theorem 5.

The proof of Lemma 4 combines Lemma 3 with Theorem 5 and can be found in the full version of the paper [26].

**B. Initialization with the conditional mean and covariance**

In order to efficiently optimize the negative log-likelihood and maintain its strong-convexity we need to search over a set of parameters that assign significant measure to the truncation set we consider. In addition, we need that the initial point of our algorithm lies in that set and satisfies this condition.

We will prove that a good such initialization is the sample mean and sample covariance, i.e. the empirical mean and covariance of the truncated distribution \( N(\mu, \Sigma, S) \).
We begin by showing that only few truncated samples suffice to obtain accurate estimates $\hat{\mu}_S$ and $\hat{\Sigma}_S$ of the mean and covariance.

**Lemma 5.** Let $(\mu_S, \Sigma_S)$ be the mean and covariance of the truncated Gaussian $N(\mu, \Sigma, S)$ for a set $S$ such that $N(\mu, \Sigma; S) = \alpha$. Using $O(\frac{d^2 \log(1/\alpha)}{\delta^2} \log(1/\delta))$ samples, we can compute estimates $\hat{\mu}_S$ and $\hat{\Sigma}_S$ such that

$$
\|\Sigma^{-1/2}(\hat{\mu}_S - \mu_S)\|_2 \leq \epsilon \quad \text{and} \\
(1 - \epsilon)\Sigma_S \preceq \hat{\Sigma}_S \preceq (1 + \epsilon)\Sigma_S
$$

with probability $1 - \delta$.

**Proof:** Let $\{x_i\}_{i=1}^n$ be the i.i.d samples drawn from the truncated Gaussian, $\mu_S = \frac{1}{n} \sum_{i=1}^n x_i$ and $\Sigma_S = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_S)(x_i - \hat{\mu}_S)^T$.

We assume w.l.o.g. that $(\mu, \Sigma) = (0, I)$. Since the $n$ samples $x_i$ are drawn from $N(0, I, S)$, which can be seen as drawing $O(n/\alpha)$ samples $N(0, I)$ and keeping only those that follows in the set $S$. This implies that with probability at least $1 - \delta$, for all samples $i$ and coordinates $j$, $|x_{ij}| \leq \log \left(\frac{nd}{\alpha \delta}\right)$.

By Hoeffding’s inequality, we get that:

$$
Pr[|\hat{\mu}_{S,j} - \mu_{S,j}| \geq \epsilon \sqrt{d}] \leq 2 \exp\left(-\frac{d \epsilon^2}{2 \log(1/\delta)}\right)
$$

Therefore, if $n \geq \Omega\left(\frac{d \log(nd/\alpha \delta) \log(1/\delta)}{\epsilon^2}\right)$, $n$ samples are sufficient to learn $\mu_S$ within $\epsilon$ with probability $1 - \delta$.

Similarly, we can use a matrix concentration inequality ([28], Corollary 5.52) to get that with probability $1 - \delta$:

$$
(1 - \epsilon)\Sigma_S \preceq \hat{\Sigma}_S \preceq (1 + \epsilon)\Sigma_S
$$

if $n \geq \Omega\left(d/\epsilon^2 \log(nd/\alpha \delta) \log(1/\delta)\right)$.

Thus setting $n = \Theta(d/\epsilon^2 \log^2(1/\alpha \delta))$ gives the result.

We also show that these estimates $\hat{\mu}_S$ and $\hat{\Sigma}_S$ are sufficiently close to the parameters $\mu, \Sigma$ of the true (un-truncated) distribution. We do this by first considering the real conditional mean and covariance parameters $\mu_S$ and $\Sigma_S$.

**Lemma 6.** Let $(\mu_S, \Sigma_S)$ be the mean and covariance matrix of the truncated Gaussian $N(\mu, \Sigma, S)$ with $N(\mu, \Sigma; S) = \alpha$. Then it holds that

1) $\|\mu_S - \mu\|_\Sigma \leq O(\sqrt{\log \frac{1}{\alpha}})$ and
2) $\Sigma_S \succeq \Omega(\alpha^2)\Sigma$,
3) $\|\Sigma^{-1/2}\Sigma_S\Sigma^{-1/2} - I\|_F \leq O(\log \frac{1}{\alpha}).$

**Proof:** First observe that we can assume without loss of generality that $\mu = 0$, $\Sigma = I$ and that $\Sigma_S$ is a diagonal matrix with entries $\lambda_1 \leq \cdots \leq \lambda_d$.

We will show that

$$
\|\mu_S\|_2 \leq \sqrt{2 \log \frac{1}{\alpha} + 1}
$$

which implies 1. for arbitrary $\mu, \Sigma$ after applying the standard transformation.

Consider the direction of the sample mean $\hat{\mu}_S$. The worst case subset $S \subset \mathbb{R}^d$ of mass at least $\alpha$ that would maximize $\|\mu_S\|_2$ is $\{\frac{\hat{E}_{x \sim N(0, I)}[x] \mu_S^T}{\hat{E}_{x \sim N(0, I)}[\|x\|^2]}\}$ is the following:

$$
S = \{x^T \hat{\mu}_S > F^{-1}(1 - \alpha)\}
$$

where $F$ is the CDF of the standard normal distribution. Since $\alpha = 1 - F(t) \leq e^{-t^2/2}$, we have that $t \leq \sqrt{2 \log \frac{1}{\alpha}}$. The bound follows since $\hat{E}_{x \sim N(0, I)}[\|x\|] \leq 1 + t$.

2) We want to bound the following expectation: $\lambda_1 = \hat{E}_{x \sim N(0, I)}(x^2 - \mu_S^2) = Var_{x \sim N(0, I)}[g_1]$, where $g_1$ denotes the marginal distribution of $g$ along the direction of $e_1$.

Therefore, if $\lambda \geq \lambda_1$, the worst case set (i.e. the one that minimizes $\text{Var}[g_1]$) is the one that has mass as close as possible to the hyperplane $x_1 = \mu_S$. However, the maximum mass that a gaussian places at the set $\{x_1 : |x_1 - \mu_S| < \epsilon\}$ is at most $2\epsilon$ as the density of the univariate gaussian is at most 1. Thus the $\hat{E}_{x \sim N(0, I)}(x_1 - \mu_S)^2$ is at least the variance of the uniform distribution $U[-\alpha/2, \alpha/2]$ which is $\alpha^2/12$. Thus $\lambda_1 \geq \lambda_1 \geq \alpha^2/12$.

3) Finally, case 3, follows from Theorem 4. Consider any large set of $n$ samples from $N(\mu, \Sigma)$. Theorem 4, implies that with probability $1 - \alpha(n)/1$, for all $T \subseteq [n]$ with $|T| = \Theta(\alpha(n))$, we have that $\left\|\sum_{i \in T} \frac{1}{|T|} X_i X_i^T - I\right\|_F \geq O(\log(1/\alpha))$. In particular, the same is true for the set of $\Theta(\alpha(n))$ samples that lie in the set $S$. As $n \to \infty$, $\sum_{i \in T} \frac{1}{|T|} X_i X_i^T$ converges to $\Sigma_S + \mu_S \mu_S^T$.

We thus obtain that $\left\|\Sigma_S + \mu_S \mu_S^T - I\right\|_F \leq \sqrt{O(\log(1/\alpha))}$, which implies that $\|\Sigma_S - I\|_F \leq \sqrt{O(\log(1/\alpha))} + \mu_S \mu_S^T \leq O(\log(1/\alpha))$.

Combining the two Lemmas and Theorem 4, we get that

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

1) $\|\hat{\mu}_S - \mu\|_\Sigma \leq O(\sqrt{\log \frac{1}{\alpha \delta}})$ and
2) $\hat{\Sigma}_S \succeq \Omega(\alpha^2)\Sigma$,
3) $\left\|\Sigma^{-1/2}\hat{\Sigma}_S\Sigma^{-1/2} - I\right\|_F \leq O(\log \frac{1}{\alpha \delta})$.

**Proof:** The first two properties follow by applying Lemma 5 with $\epsilon = 1/2$ to Lemma 6. The last one follows by Theorem 4, using an identical argument to the proof of Lemma 6. Note that the required sample complexity has a quadratic dependence on $d$ as it is necessary for closeness in Frobenius norm, and is required by Theorem 4.
C. A set of parameters with non-trivial mass in the truncation set

Corollary 1 shows that the empirical mean \( \hat{\mu}_S \) and the empirical covariance matrix \( \hat{\Sigma}_S \) of the truncated distribution \( N(\mu^*, \Sigma^*, S) \) using \( n = \tilde{O}(d^2) \) samples are close to the true parameters \( \mu^*, \Sigma^* \).

We show that these guarantees are sufficient to show that the Normal \( N(\hat{\mu}_S, \hat{\Sigma}_S) \) assigns constant mass to the set \( S \). Lemma 7 gives us the necessary conditions to show this statement.

Lemma 7. Consider two Gaussians \( N(\mu_1, \Sigma_1) \) and \( N(\mu_2, \Sigma_2) \), such that

1. \( \|I - \Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2}\|_F \leq B \) and
2. \( 1/B \leq \|\Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2}\| \leq B \),
3. \( \|\Sigma_2^{-1/2} (\mu_1 - \mu_2)\| \leq B \).

Suppose that for a set \( S \) we have that \( N(\mu_1, \Sigma_1; S) \geq \alpha \). Then \( N(\mu_2, \Sigma_2; S) \geq (\alpha/12)^{30B^2} \).

The proof of Lemma 7 can be found in the full version of the paper [26].

We will apply Lemma 7 to bound the measure assigned to \( S \) by \( N(\hat{\mu}_S, \hat{\Sigma}_S) \). For this, we need to convert the bounds of Corollary 1 to those required by Lemma 7.

Proposition 1. It holds that

- \( \|I - \Sigma^{1/2} \Sigma_S^{-1} \Sigma^{1/2}\|_F \leq O(\log(1/\alpha) / \alpha^2) \) and
- \( \|I - \Sigma_S^{1/2} \Sigma^{1/2} \Sigma_S^{-1}\|_F \leq O(\log(1/\alpha) / \alpha^2) \),
- \( \Omega(\alpha^2) \leq \|\Sigma_S^{1/2} \Sigma_S \Sigma_S^{-1/2}\|_2 \leq O(1/\alpha^2) \) and
- \( \Omega(\alpha^2) \leq \|\Sigma_S^{1/2} \Sigma_S^{-1}\|_2 \leq O(1/\alpha^2) \),
- \( \|\Sigma_S^{-1/2} (\hat{\mu}_S - \mu^*)\| \leq O(\log(1/\alpha^2) / \alpha^2) \) and
- \( \|\Sigma_S^{-1/2} (\hat{\mu}_S - \mu^*)\| \leq O(\log(1/\alpha^2) / \alpha^2) \).

Proof: Without loss of generality \( \mu = 0 \), \( \Sigma = I \) and \( \Sigma_S = \text{diag}(\lambda_1, \ldots, \lambda_d) \). From Corollary 1, we have that

\[
\|I - \Sigma^{1/2} \Sigma_S^{-1} \Sigma^{1/2}\|_F = \sum_i (1 - 1/\lambda_i)^2 \leq \frac{1}{\alpha^2} \sum_i (1 - 1/\lambda_i)^2 \leq O\left(\log(1/\alpha^2) / \alpha^2\right).
\]

Moreover, we have that \( \|\Sigma_S^{-1} \Sigma^{1/2} (\mu_1 - \mu_2)\| = \sum_i \frac{1}{\lambda_i} \|\mu_1 - \mu_2\| \leq O\left(\log(1/\alpha^2) / \alpha^2\right) \).

Similarly, we have that

\[
\|I - \Sigma_S^{1/2} \Sigma^{1/2} \Sigma_S^{-1}\|_F = \sum_i (1 - 1/\lambda_i)^2 \leq O\left(\log(1/\alpha) / \alpha^2\right).
\]

Moreover, we have that \( \|\Sigma_S^{-1} \Sigma^{1/2} (\mu_1 - \mu_2)\| = \sum_i \lambda_i^2 \|\mu_1 - \mu_2\| \leq O\left(\log(1/\alpha) / \alpha^2\right) \).

Proposition 1 implies that Lemma 7 can be invoked with \( B = O\left(\log(1/\alpha^2) / \alpha^2\right) \) to obtain the following corollaries:

**Corollary 2.** Consider a truncated normal distribution \( N(\mu^*, \Sigma^*, S) \) with \( N(\mu^*, \Sigma^*, S) \geq \alpha > 0 \). The estimates \( (\hat{\mu}_S, \hat{\Sigma}_S) \) obtained by Corollary 1 satisfy \( N(\hat{\mu}_S, \hat{\Sigma}_S; S) \geq c_\alpha \) for some constant \( c_\alpha \) that depends only on the constant \( \alpha > 0 \).

**Corollary 3.** Consider a truncated normal distribution \( N(\mu^*, \Sigma^*, S) \) with \( N(\mu^*, \Sigma^*, S) \geq \alpha > 0 \). Let \( (\mu_S, \Sigma_S) \) be the estimate obtained by Corollary 1 and let \( (\mu, \Sigma) \) be any estimate that satisfies

- \( \|I - \Sigma_S^{1/2} \Sigma_S^{-1} \Sigma_S^{1/2}\|_F \leq O(\log(1/\alpha) / \alpha^2) \),
- \( \|I - \Sigma_S^{1/2} \Sigma_S^{-1} \Sigma_S^{1/2}\|_F \leq O(\log(1/\alpha) / \alpha^2) \),
- \( \|\Sigma_S^{-1/2} (\mu_S - \mu^*)\| \leq O(\log(1/\alpha^2) / \alpha^2) \),
- \( \|\Sigma_S^{-1/2} (\mu_S - \mu^*)\| \leq O(\log(1/\alpha^2) / \alpha^2) \).

Then, \( N(\mu, \Sigma; S) \geq c_\alpha \) for some constant \( c_\alpha \) that depends only on the constant \( \alpha > 0 \).

D. Analysis of SGD: Proof of Theorem 1

We define \( t = T^b \) and \( \omega = \left(\begin{array}{c} t \\ \omega \end{array}\right) \) for simplicity. Our goal is to run projected stochastic gradient descent to optimize \( \tilde{E} \) with respect to the parameters \( \nu, T \) as we describe in detail in Figure 2. This algorithm iterates over the estimation \( \tilde{E} \) of the true parameters \( \nu^* \). Let also \( O \) the sample oracle from the unknown distribution \( N(\nu^*, \Sigma^*) \).

The initialization step of our algorithm computes the empirical mean \( \hat{\mu} \) and the empirical covariance matrix \( \hat{\Sigma} \) of the truncated distribution \( N(\mu^*, \Sigma^*, S) \) using \( n = \tilde{O}(d^2) \) samples \( x_1, \ldots, x_n \).

\[
\hat{\mu}_S = \frac{1}{n} \sum_{i=1}^n x_i, \hat{\Sigma}_S = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_S) (x_i - \hat{\mu}_S)^T.
\]

Then, we transform our space such that \( \hat{\mu}_S \rightarrow 0 \) and \( \hat{\Sigma}_S \rightarrow I \).

We are now ready to describe the domain \( D \) of our projected stochastic gradient ascent in the transformed space where \( \mu_S = 0 \) and \( \Sigma_S = I \). The domain is parameterized by the positive number \( r \) and is defined as follows

\[
D_r = \{ (\nu, T) \mid \|\nu\|_2 \leq r, \|I - T\|_F \leq r, \|T^{-1}\|_2 \leq r \}.
\]

Observe that \( D_r \) is a convex set as the constraint \( \|T^{-1}\|_2 \leq r \) an infinite set of linear, with respect to \( T \), constraints of the form \( x^T T x \geq 1/r \) for all \( x \in \mathbb{R}^d \). Moreover, in Algorithm 3 we present an efficient procedure to project any point in our space to \( D_r \). The next Lemma 8 shows the missing details and proves the correctness of Algorithm 3.

**Lemma 8.** Given \( (\nu^*, T^*) \), there exists an efficient algorithm that solves the following problem which corresponds to projecting \( (\nu, T) \) to the set \( D_r \):

\[
\arg\min_{(\nu, T) \in D_r} \|\nu - \nu^*\|_2 + \|T - T^*\|_F^2.
\]
Proof of Lemma 8: Because of the form of $D_r$ and the objective function of our program, observe that we can project $\nu$ and $T$ separately. The projection of $\nu'$ to $D_r$ is the solution of the following problem \( \arg\min_{\nu'} \|\nu - \nu'\|_2^2 \) which has a closed form. So we focus on the projection of $T$.

To project $T$ to $D_r$, we need to solve the following program.

\[
\arg\min_{T} \|T - T'\|_F^2 \\
\text{s.t. } \|T - I\|_F^2 \leq r^2 \\
T \succeq \frac{1}{r} I
\]

Equivalently, we can perform binary search over the Lagrange multiplier $\lambda$ and at each step solve the following program.

\[
\arg\min_{T} \|T - T'\|_F^2 + \lambda \|T - I\|_F^2 \\
\text{s.t. } T \succeq \frac{1}{r} I
\]

After completing the squares on the objective, we can set $H = I - \frac{1}{r} T'$ and make the change of variables $R = I - T$ and solve the program.

\[
\arg\min_{T} \|R - H\|_F^2 \\
\text{s.t. } R \preceq \left(1 - \frac{1}{r}\right) I
\]

Observe now that without loss of generality $H$ is diagonal. If this is not the case we can compute the singular value decomposition of $H$ and change the base of the space so that $H$ is diagonal. Then, after finding the answer to this case we transform back the space to get the correct $R$. Then $H$ is diagonal the solution to this problem is very easy and it even has a closed form. ■

Apart from efficient projection (Lemma 8) and strong concavity (Lemma 4) we also need a bound on the square of the norm of the gradient vector in order to prove theoretical guarantees for our SGD algorithm.

Lemma 9. Let $\nu^{(i)}$ the gradient of the log likelihood function at step $i$ as computed in the line 6 of Algorithm 1. Let $\nu, T$ be the guesses of the parameters after step $i - 1$ according to which the gradient is computed with $\mu = T^{-1} \nu$ and $\Sigma = T^{-1}$. Let $\mu^*, \Sigma^*$ be the parameters we want to recover, with $\nu^* = \Sigma^*^{-1} \mu^*$ and $T^* = \Sigma^*^{-1}$. Define $\mathcal{N}(\mu, \Sigma; S)$ to be a normal distribution on $S$ with parameters $\mu$ and $\Sigma$.

\[
\mathcal{N}(\mu, \Sigma; S) \geq \beta, \mathcal{N}(\mu^*, \Sigma^*; S) \geq \beta.
\]

Then, we have that

\[
\mathbb{E} \left[ \|\nu^{(i)}\|_2^2 \right] \leq \frac{100}{\beta} d^2 r^2.
\]

The proof of Lemma 9 can be found in the full version of the paper [26].

Now we have all the ingredients to use the basic tools for analysis of projected stochastic gradient descent. The formulation we use is from Chapter 14 of [29].

Theorem 6 (Theorem 14.11 of [29]). Let $f = -\bar{\ell}$. Assume that $f$ is $\lambda$-strongly convex, that $\mathbb{E} \left[ \|\nu^{(i)}\|_2^2 \right] \leq \frac{\lambda}{\lambda^2}$. Let $\bar{w}^* \in \arg\min_{w \in D_r} f(w)$ be an optimal solution. Then,

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

where $\bar{w}$ is the output of Algorithm 1.

We also state a simple lemma for strong convexity functions that follows easily from the definition of strong convexity.

Lemma 10 (Lemma 13.5 of [29]). If $f$ is $\lambda$-strongly convex and $\bar{w}^*$ is a minimizer of $f$, then, for any $w$ it holds that

\[
f(w) - f(\bar{w}^*) \geq \frac{\lambda}{2} \|w - \bar{w}^*\|_2^2.
\]

Using Theorem 6, together with Lemmata 4, 9 and 10 we can get our first theorem that bounds the expected cost of Algorithm 1. Then we can also use Markov’s inequality to get and our first result in probability.

Lemma 11. Let $\mu^*, \Sigma^*$ be the underline parameters of our model, $f = -\bar{\ell}$, $r = O \left(\frac{\log(1/\delta)}{\alpha^2}\right)$ and also

\[
\beta_* = \min_{(\nu, T) \in D_r} N(T^{-1} \nu, T^{-1}; S) \text{ and also}
\]

\[
\lambda_* = \min_{(\nu, T) \in D_r} \lambda_m(T^{-1}) \geq r.
\]

then there exists a universal constant $C > 0$ such that

\[
\mathbb{E} \left[ f(\bar{w}) \right] - \mathbb{E} \left[ f(\bar{w}^*) \right] \leq \frac{C \cdot r}{\beta_*^2} \frac{d^2}{M} \left(1 + \log(M)\right),
\]

where $\bar{w}$ is the output of Algorithm 1.

Proof of Lemma 11: This result follows directly from Theorem 6, if our initial estimate $\bar{w}^{(0)}$ belongs to $D_r$. To ensure that this is the case we set $r = O \left(\frac{\log(1/\delta)}{\alpha^2}\right)$ and we apply Proposition 1 and the Lemma follows. ■

We are now ready to prove our main Theorem 1.

Proof of Theorem 1: Using Lemma 11 and applying Markov’s inequality we get that

\[
P \left( f(\bar{w}) - f(\bar{w}^*) \geq \frac{2C \cdot r}{\beta_*^2} \frac{d^2}{M} \left(1 + \log(M)\right) \right) \leq \frac{1}{2^4},
\]

(14)

We can easily amplify the probability of success to $1 - \delta$ by repeating $\log(1/\delta)$ from scratch the optimization procedure and keeping the estimation that achieves the maximum log-likelihood value. The high probability result enables the use of Lemma 10 to get closeness in parameter space.

To get our estimation we first repeat the SGD procedure $K = \log(1/\delta)$ times independently, with parameter $M$ each time. We then get the set of estimates $\mathcal{E} = \{\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_K\}$. Our final estimate is

\[
\bar{w} = \arg\min_{w \in \mathcal{E}} \bar{\ell}(w)
\]
By (14) and the independence of all the $K$ estimates in $\mathcal{E}$ we get that

$$\mathbb{P} \left( f(\hat{w}) - f(w^*) \geq \frac{2c\cdot r}{\beta^2} \frac{d^2}{M} (1 + \log(M)) \right) \leq \delta.$$  

(15)

Hence we can condition on the event $f(\hat{w}) - f(w^*) \leq \frac{2c\cdot r}{\beta^2} \frac{d^2}{M} (1 + \log(M))$ and we only lose probability at most $\delta$. Now remember that for Lemma 11 to apply we have $r = O \left( \frac{\log(1/\alpha)}{\alpha^2} \right)$. Also, using Corollary 3 we get that $\beta^* \geq \epsilon_\alpha$ where $\epsilon_\alpha$ is a constant that depends only on the constant $\alpha$. Hence, with probability at least $1 - \delta$ we have that

$$f(\hat{w}) - f(w^*) \leq \epsilon_\alpha \frac{d^2}{M} (1 + \log(M)), $$

where $\epsilon_\alpha$ is a constant that depends only on $\alpha$. Now we can use Lemma 10 to get that

$$\|\hat{w} - w^*\|_2 \leq \epsilon_\alpha \sqrt{\frac{d^2}{M} (1 + \log(M))}. \quad (16)$$

Also, it holds that

$$\|\hat{w} - w^*\|_2^2 = \|\nu - \nu^*\|_2^2 + \|T - T^*\|_F^2 = \|\Sigma^{-1}\mu - \Sigma^{-1}\mu^*\|_2^2 + \|\Sigma^{-1} - \Sigma^{-1}\|_F^2. $$

Hence, for $M \geq \tilde{O} \left( \frac{d^2}{\epsilon^2} \right)$ and using (16) we have that

$$\|\Sigma^{-1}\mu - \Sigma^{-1}\mu^*\|_2^2 \leq \|\Sigma^{-1} - \Sigma^{-1}\|_F^2 \leq \epsilon.$$

The number of samples is $O(KM)$ and the running time is $\text{poly}(K,M,1/\epsilon, \delta)$. Hence for $K = \log(1/\delta)$ and $M \geq \tilde{O} \left( \frac{d^2}{\epsilon^2} \right)$ our theorem follows.

Algorithm 1 Projected Stochastic Gradient Descent. Given access to samples from $\mathcal{N}(\mu^*, \Sigma^*, S)$.

1. procedure SGD($M, \lambda$) \triangleright $M$: number of steps, $\lambda$: parameter
2. $w(0) \leftarrow \begin{pmatrix} \Sigma_{S}^{\frac{1}{2}} \\ \mu_{S} \end{pmatrix}$
3. for $i = 1, \ldots, M$ do
4. Sample $x^{(i)}$ from $O$
5. $\eta_i \leftarrow \frac{1}{\sqrt{i}}$
6. $\nu^{(i)} \leftarrow \text{GRADIENTESTIMATION}(x^{(i)}, w^{(i-1)})$
7. $r^{(i)} \leftarrow w^{(i-1)} - \eta_i \nu^{(i)}$
8. $w^{(i)} \leftarrow \text{PROJECTTODOMAIN}(r^{(i)})$
9. return $\hat{w} \leftarrow \frac{1}{M} \sum_{i=1}^{M} w^{(i)}$ \triangleright Output the average.

Figure 2: Description of the Stochastic Gradient Descent (SGD) algorithm for estimating the parameters of a truncated Normal.

IV. IMPOSSIBILITY OF ESTIMATION WITH AN UNKNOWN TRUNCATION SET

We have showed that if one assumes query access to the truncation set, the estimation problem for truncated Normals can be efficiently solved with very few queries and samples.

If the truncation set is unknown, as we show in this section, it is information theoretically impossible to produce an estimate that is closer than a constant in total variation distance to the true distribution even for single-dimensional truncated Gaussians.

To do this, we consider two Gaussian distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ with

$$\text{dist}_{TV}(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = \alpha.$$
We show that there exist a distribution over truncation sets $S_1$ with $\mathcal{N}(\mu_1, \sigma_1^2; S_1)$, $D_1$, and a distribution over truncation sets $S_2$ with $\mathcal{N}(\mu_2, \sigma_2^2; S_2)$, $D_2$, such that a random instance $\{S_1, \mathcal{N}(\mu_1, \sigma_1^2)\}$ with $S_1$ drawn from $D_1$ is indistinguishable from a random instance $\{S_2, \mathcal{N}(\mu_2, \sigma_2^2)\}$ with $S_2$ drawn from $D_2$.

**Lemma 12** (Indistinguishability with unknown set). Consider two single-dimensional Gaussian distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ with $\text{dist}_{TV}(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = 1 - \alpha$. The truncated Gaussian $\mathcal{N}(\mu_1, \sigma_1^2, S_1)$ with an unknown set $S_1$ such that $\mathcal{N}(\mu_1, \sigma_1^2; S_1) = \alpha$ is indistinguishable from the distribution $\mathcal{N}(\mu_2, \sigma_2^2; S_2)$ with unknown set $S_2$ such that $\mathcal{N}(\mu_2, \sigma_2^2; S_2) = \alpha$.

**Proof:** We define the randomized family of sets $D_i$: The random set $S_i \sim D_i$ is constructed by adding every point $x \in \mathbb{R}$ with probability $\min\{N(\mu_1, \sigma_1^2; x)/N(\mu_1, \sigma_1^2; x), 1\}$.

Now consider the distribution $p$ with density $\frac{1}{\pi} \min\{N(\mu_1, \sigma_1^2; x), N(\mu_2, \sigma_2^2; x)\}$. This is a proper distribution as $\text{dist}_{TV}(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = 1 - \alpha$. Note that samples from $p$ can be generated by performing the following rejection sampling process: Pick a sample from the distribution $\mathcal{N}(\mu_1, \sigma_1^2)$ and reject it with probability $\min\{N(\mu_3 - i, \sigma_3^2; x)/N(\mu_1, \sigma_1^2; x), 1\}$.

We now argue that samples from the distribution $\mathcal{N}(\mu_1, \sigma_1^2; S_1)$ for a random $S_1 \sim D_1$, are indistinguishable from $p$. This is because an alternative way of sampling the distribution $\mathcal{N}(\mu_1, \sigma_1^2; S_1)$ can be sampled as follows. Draw a sample $x$ from $\mathcal{N}(\mu_1, \sigma_1^2)$ and then check if $x \in S_1$. By the principle of deferred randomness, we may not commit to a particular set $S_1$ only decide whether $x \in S_1$ after drawing $x$ as long as the selection is consistent. That is, every time we draw the same $x$ we must output the same answer. This sampling process is identical to the sampling process of $p$ until the point where an $x_i$ is sampled twice. As the distributions are continuous and every time a sample is accepted with probability $\alpha > 0$ no collisions will be found in this process.

The following corollary completes the proof of Theorem 2.

**Corollary 4.** For all $\alpha > 0$, given infinitely many samples from a univariate normal $\mathcal{N}(\mu, \sigma^2)$, which are truncated to an unknown set $S$ of measure $\alpha$, it is impossible to estimate parameters $\hat{\mu}$ and $\hat{\sigma}^2$ such that the distributions $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ are guaranteed to be within $\frac{1 - \alpha}{2}$.

To see why the corollary is true, note that since it is impossible to distinguish between the truncated Gaussian distributions $\mathcal{N}(\mu_1, \sigma_1^2, S_1)$ and $\mathcal{N}(\mu_2, \sigma_2^2, S_2)$ with $\text{dist}_{TV}(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = 1 - \alpha$, any estimated $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ will satisfy either $\text{dist}_{TV}(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)) > \frac{1 - \alpha}{2}$ or $\text{dist}_{TV}(\mathcal{N}(\mu_2, \sigma_2^2), \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)) > \frac{1 - \alpha}{2}$.

**Remark 1.** The construction in Lemma 12 uses random sets $S_1$ and $S_2$ that select each point on the real line with some probability. One may use coarser sets by including all points within some range $\varepsilon$ of the randomly chosen points. In this case the collision probability is no longer 0 and depends on $\varepsilon$. Given that no collisions are seen the two cases are again indistinguishable. Moreover, for very small $\varepsilon$, an extremely large number of samples are needed to see a collision.

**ACKNOWLEDGMENTS**

The authors were supported by NSF CCF-1551875, CCF-1617730, CCF-1650733, CCF-1733808, CCF-1740751 and IIS-1741137.

**REFERENCES**


