Efficient Statistics, in High Dimensions, from Truncated Samples

Constantinos Daskalakis * Themis Gouleakis † Christos Tzamos ‡ Manolis Zampetakis §

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 $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means a samples is only revealed if it falls in 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 show that the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ 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.

1 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 [Ber60]. In 1897, Galton analyzed truncated samples corresponding to registered speeds of American trotting horses [Gal97]. 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

*Affiliation: MIT, email: costis@csail.mit.edu
†Affiliation: MIT, email: tgoule@mit.edu
‡Affiliation: University of Wisconsin-Madison, email: tzamos@wisc.edu
§Affiliation: MIT, email: mzampet@mit.edu

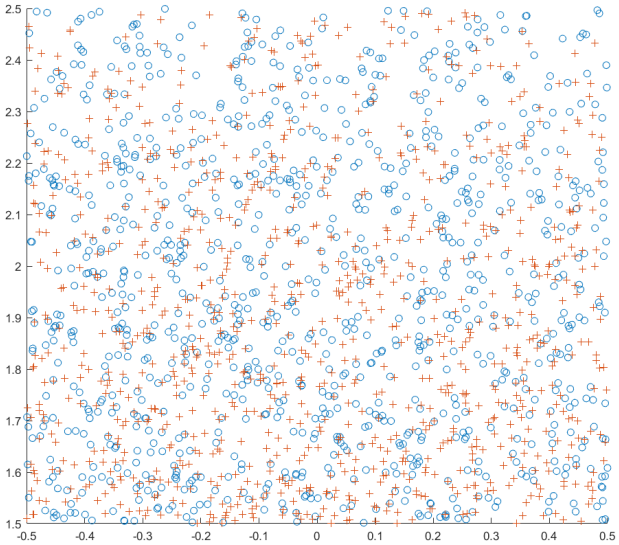

Figure 1: One thousand samples from $\mathcal{N}((0,1), I)$ and from $\mathcal{N}((0,0), 4I)$ truncated into the $[-0.5, 0.5] \times [1.5, 2.5]$ box. We leave it to the reader to guess which sample comes from which.

deviation. Dissatisfaction with Galton's estimates led Pearson [Pea02] and later Pearson and Lee [PL08] and Lee [Lee14] 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 [Fis31].

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. [Sch86, Coh16, BC14] 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 $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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 [Hot48] and Tukey [Tuk49] 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 accommodating a very general truncation model. We suppose that samples, $X_1, X_2, \ldots$, from an unknown d-variate

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

$$\left\| \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) \right\|_2 \leq \varepsilon; \quad \text{and} \quad \left\| I - \boldsymbol{\Sigma}^{-1/2} \hat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}^{-1/2} \right\|_F \leq \varepsilon. \tag{1.1}$$

Note that under the above conditions the total variation distance between $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ is $O(\varepsilon)$, 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 α 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 $\mathbf{1}_{x \in S}$, 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 α, 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 $\frac{1-\alpha}{2}$.*

Overview of the Techniques. The proofs of Theorems 1 and 2 are provided in Sections 3 and 4 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 $v = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{T} = \boldsymbol{\Sigma}^{-1}$ (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 $(\boldsymbol{v}, \boldsymbol{T})$ of the parameters (see Lemma 4). In particular, to maintain strong convexity, SGD must remain within a region of parameters $(\boldsymbol{v}, \boldsymbol{T})$ that assign non-trivial measure to S. We show that the pair of parameters $(\boldsymbol{v}, \boldsymbol{T})$ 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 $(\boldsymbol{v}, \boldsymbol{T})$ 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 un-biased samples for the gradient of the log-likelihood function using rejection sampling from the normal distribution defined our current estimate of the parameters $(\boldsymbol{v}, \boldsymbol{T})$ (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 non-trivial 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. [CLMW11] computes the PCA of a matrix, even when a constant fraction of its entries to be adversarially corrupted, but it requires the locations of the corruptions to be relatively evenly distributed. Related work of Xu et al. [XCM10] provides a robust PCA algorithm for arbitrary corruption locations, requiring at most 50% of the points to be corrupted.

Closer to our work, [DKK+16a, LRV16, DKK+17, DKK+18] perform robust estimation of the parameters of multi-variate Gaussian distributions in the presence of 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 (i.e. after seeing the sample generated by the Gaussian). The authors in [CSV17] 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 $poly(1/\alpha)$ answers, one of which is guaranteed to be accurate.

Similar to [CSV17], we study a regime where only an arbitrarily small constant fraction of the samples from a normal distribution can be observed. In contrast to [CSV17], 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 [CSV17], while accommodating a much larger number of deletions of samples compared to [DKK+16a, DKK+18].

Other robust estimation works include robust linear regression [BJK15] and robust estimation algorithms under sparsity assumptions [Li17, BDLS17]. In [HM13], 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 [SCV18].

2 Preliminaries

2.1 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 $\boldsymbol{f}$. Let $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ be the inner product of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. We use $\boldsymbol{I}_d$ to refer to the identity matrix in d dimensions. We may skip the subscript when the dimensions are clear. We use $\boldsymbol{E}_{i,j}$ to refer to the all zero matrix with one 1 at the (i,j) entry. 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 $\boldsymbol{A}$. We define $\sharp$ to be the inverse operator, i.e. $\left(\boldsymbol{A}^\flat\right)^\sharp = \boldsymbol{A}$. Let also $\mathcal{Q}_d$ be the set of all the symmetric $d \times d$ matrices.
$\mathbb{E}[\boldsymbol{X}]$ is the expected value of the random variable $\boldsymbol{X}$ and $\mathrm{Var}[\boldsymbol{X}]$ is the variance of $\boldsymbol{X}$. The covariance matrix between two random variables $\boldsymbol{X}, \boldsymbol{Y}$ is $\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}]$.

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

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

It is well known that $\|\boldsymbol{A}\|_2 = \max\{\lambda_i\}$, where λ_i's are the eigenvalues of $\boldsymbol{A}$.

The *Frobenius norm* of a matrix $\boldsymbol{A}$ is defined as follows:

$$\|\boldsymbol{A}\|_F = \left\|\boldsymbol{A}^\flat\right\|_2.$$

The *Mahalanobis distance* between two vectors $\boldsymbol{x}, \boldsymbol{y}$ given a covariance matrix Σ is defined as:

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

Truncated 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

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) = \frac{1}{\sqrt{2\pi \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{2.2}$$

Also, let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S)$ 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* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ 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

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S; \boldsymbol{x}) = \begin{cases} \dfrac{1}{\int_S \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{y})d\boldsymbol{y}} \cdot \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) & \boldsymbol{x} \in S \\ 0 & \boldsymbol{x} \notin S \end{cases}. \tag{2.3}$$

We will assume that the covariance matrix Σ is full rank. The case where Σ 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. $M_S(\boldsymbol{x}) = \mathbb{1}_{\boldsymbol{x}\in S}$.

2.2 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 [LM00].). *Let $X_1, \ldots, X_n$ independent identically distributed random variables following $\mathcal{N}(0,1)$ and let $\boldsymbol{a} \in \mathbb{R}_+^d$. Then, the following inequalities hold for any $t \in \mathbb{R}_+$.*

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

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

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 [DKK+16b]). *Let $\rho, \tau > 0$ and $X_1, \ldots, X_n$ be i.i.d samples from $\mathcal{N}(0, \boldsymbol{I})$. 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:*

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

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 Stochastic Gradient Descent 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.

3.1 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 $\boldsymbol{x}_i$ from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$. 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 $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{T} = \boldsymbol{\Sigma}^{-1}$.

3.1.1 Log-Likelihood for a Single Sample

Given one vector $\boldsymbol{x} \in \mathbb{R}^d$, the negative log-likelihood that $\boldsymbol{x}$ is a sample of the form $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ is

$$\begin{aligned}
\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) =& \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) + \log\left(\int_S \exp\left(-\frac{1}{2}(z - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(z - \boldsymbol{\mu})\right) dz\right) \\
=& \frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \log\left(\int_S \exp\left(-\frac{1}{2} z^T \boldsymbol{\Sigma}^{-1} z + z^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\right) dz\right) \qquad (3.4)
\end{aligned}$$

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

$$\ell(\boldsymbol{\nu}, \boldsymbol{T}; \boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{T}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{\nu} + \log\left(\int_S \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{T}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\nu}\right) d\boldsymbol{z}\right) \tag{3.5}$$

We now compute the gradient of $\ell(\boldsymbol{\nu}, \boldsymbol{T}; \boldsymbol{x})$ with respect to the set of variables $\begin{pmatrix}\boldsymbol{T}^\flat \\ \boldsymbol{\nu}\end{pmatrix}$.

$$\begin{aligned}
\nabla\ell(\boldsymbol{\nu}, \boldsymbol{T}; \boldsymbol{x}) &= -\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^T\right)^\flat \\ \boldsymbol{x}\end{pmatrix} + \frac{\int_S \begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{T}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\nu}\right) d\boldsymbol{z}}{\int_S \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{T}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\nu}\right) d\boldsymbol{z}} \\
&= -\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^T\right)^\flat \\ \boldsymbol{x}\end{pmatrix} + \frac{\int_S \begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{T}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\nu} - \left\|\boldsymbol{T}^{-1}\boldsymbol{\nu}\right\|_2^2\right) d\boldsymbol{z}}{\int_S \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{T}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{B}\boldsymbol{x} - \left\|\boldsymbol{T}^{-1}\boldsymbol{\nu}\right\|_2^2\right) d\boldsymbol{z}} \\
&= -\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^T\right)^\flat \\ \boldsymbol{x}\end{pmatrix} + \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{T}^{-1}\boldsymbol{\nu}, \boldsymbol{T}^{-1}, S)}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix}\right]
\end{aligned} \tag{3.6}$$

Finally, we compute the Hessian $\mathbf{H}_\ell$ of the log-likelihood function.

$$\begin{aligned}
\mathbf{H}_\ell(\boldsymbol{\nu}, \boldsymbol{T}) =& \frac{\int_S \begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix} \begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix}^T \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{T}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\nu}\right) d\boldsymbol{z}}{\int_S \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{T}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\nu}\right) d\boldsymbol{z}} \\
&- \frac{\int_S \begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{T}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\nu}\right) d\boldsymbol{z}}{\int_S \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{T}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\nu}\right) d\boldsymbol{z}} \cdot \\
&\cdot \frac{\int_S \begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix}^T \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{T}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\nu}\right) d\boldsymbol{z}}{\int_S \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{T}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\nu}\right) d\boldsymbol{z}} \\
=& \mathrm{Cov}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{T}^{-1}\boldsymbol{\nu}, \boldsymbol{T}^{-1}, S)}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix}, \begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix}\right].
\end{aligned} \tag{3.7}$$

Since the covariance matrix of a random variable is always positive semidefinite, we conclude that $\mathbf{H}_\ell(\boldsymbol{\nu}, \boldsymbol{T})$ is positive semidefinite everywhere and hence, we have the following lemma.

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

3.1.2 Log-Likelihood in the Population Model

The negative log-likelihood function in the population model is equal to

$$\bar{\ell}(\boldsymbol{\nu}, \boldsymbol{T}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)}\left[\frac{1}{2}\boldsymbol{x}^T\boldsymbol{T}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{\nu}\right] - \log\left(\int_S \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{T}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{\nu}\right)d\boldsymbol{z}\right) \tag{3.8}$$

Also, using (3.6) we have that

$$\nabla\bar{\ell}(\boldsymbol{\nu}, \boldsymbol{T}) = -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^T\right)^\flat \\ \boldsymbol{x}\end{pmatrix}\right] + \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{T}^{-1}\boldsymbol{\nu},\boldsymbol{T}^{-1},S)}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix}\right]. \tag{3.9}$$

Hence from Lemma 1 we have that $\bar{\ell}$ is a convex function with respect to $\boldsymbol{\nu}$ and $\boldsymbol{T}$. Also, from (3.9) we get that the gradient $\nabla\bar{\ell}(\boldsymbol{\nu}, \boldsymbol{T})$ is $\mathbf{0}$, when $\boldsymbol{\mu}^* = \boldsymbol{T}^{-1}\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}^* = \boldsymbol{T}^{-1}$. From this observation together with the convexity of $\bar{\ell}$ we conclude that the true parameters, $\boldsymbol{\nu}^* = \boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*$ and $\boldsymbol{T}^* = \boldsymbol{\Sigma}^{*-1}$, *maximize the log-likelihood function*.

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

$$\bar{\ell}(\boldsymbol{\nu}^*, \boldsymbol{T}^*) \leq \bar{\ell}(\boldsymbol{\nu}, \boldsymbol{T}).$$

Finally, observe that the Hessian of ℓ is the same as the Hessian of $\bar{\ell}$.
One property of the log likelihood that will be important later is its *strong convexity*.

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

Our goal it to prove that $\bar{\ell}$ is strongly convex for any S such that $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; S)$ is at least a constant α. 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 $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ and also $\Sigma \succeq \mathbf{0}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$, then we define the minimum eigenvalue $\lambda_m(\boldsymbol{\Sigma})$ of the fourth moment tensor of $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ as*

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

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

$$\mathbf{H}_\ell \succeq \lambda_m(\boldsymbol{T}^{-1}) \cdot \boldsymbol{I}.$$

Proof of Lemma 3: Let $\boldsymbol{\mu} = \boldsymbol{T}^{-1}\boldsymbol{\nu}$ and $\boldsymbol{\Sigma} = \boldsymbol{T}^{-1}$, we have that

$$\mathbf{H}_\ell = \mathrm{Cov}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix}, \begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix}\right].$$

Now let $\boldsymbol{v} \in \mathbb{R}^d$ and $\boldsymbol{U} \in \mathcal{Q}_d$ with $\|\boldsymbol{v}\|_2^2 + \|\boldsymbol{U}\|_F^2 = 1$. It is easy to prove the following inequality

$$\begin{pmatrix}\left(\boldsymbol{U}^\flat\right)^T & \boldsymbol{v}^T\end{pmatrix}\mathrm{Cov}_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix}, \begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat \\ \boldsymbol{z}\end{pmatrix}\right]\begin{pmatrix}\boldsymbol{U}^\flat \\ \boldsymbol{v}\end{pmatrix} \geq 1, \tag{3.10}$$

from which we easily get that

$$\begin{aligned}\left(\left(\boldsymbol{U}^\flat\right)^T \quad \boldsymbol{v}^T\right)&\mathrm{Cov}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix},\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix}\right]\begin{pmatrix}\boldsymbol{U}^\flat\\ \boldsymbol{v}\end{pmatrix}\\ &\geq \left(\left(\boldsymbol{U}^\flat\right)^T \quad \boldsymbol{v}^T\right)\begin{pmatrix}\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Sigma}\end{pmatrix}\begin{pmatrix}\boldsymbol{U}^\flat\\ \boldsymbol{v}\end{pmatrix}\\ &\geq \lambda_m(\boldsymbol{\Sigma}).\end{aligned} \tag{3.11}$$

Hence the lemma follows. ■

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 [CW01]). *Let $q, \gamma \in \mathbb{R}_+$, $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} \in \mathcal{Q}_d$ and $p : \mathbb{R}^d \to \mathbb{R}$ be a multivariate polynomial of degree at most k, we define*

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

then there exists an absolute constant C such that

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

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

$$\mathbf{H}_\ell \succeq \frac{1}{2^{13}}\left(\frac{\beta}{C}\right)^4 \lambda_m(\boldsymbol{\Sigma}) \cdot \boldsymbol{I},$$

where C is the universal constant guaranteed by Theorem 5.

Proof of Lemma 4: We have that

$$\begin{aligned}\boldsymbol{R} &= \mathrm{Cov}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix},\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix}\right]\\ &= \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\left(\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix} - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix}\right]\right)\cdot\right.\\ &\qquad\left.\cdot\left(\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix} - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix}\right]\right)^T\right].\end{aligned}$$

We also define

$$\begin{aligned}\boldsymbol{R}' = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}&\left[\left(\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix} - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix}\right]\right)\cdot\right.\\ &\left.\cdot\left(\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix} - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix}\right]\right)^T\right],\end{aligned}$$

and

$$\boldsymbol{R}^* = \mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\left(\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix} - \mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix}\right]\right)\cdot \right.$$
$$\left.\cdot\left(\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix} - \mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T\right)^\flat\\ \boldsymbol{z}\end{pmatrix}\right]\right)^T\right],$$

Now let $\boldsymbol{v}\in\mathbb{R}^d$ and $\boldsymbol{U}\in\mathcal{Q}_d$ with $\|\boldsymbol{v}\|_2^2+\|\boldsymbol{U}\|_F^2=1$. We have that

$$\left(\left(\boldsymbol{U}^\flat\right)^T \quad \boldsymbol{v}^T\right)\boldsymbol{R}\begin{pmatrix}\boldsymbol{U}^\flat\\ \boldsymbol{v}\end{pmatrix} = \mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[p_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\right]$$

$$\left(\left(\boldsymbol{U}^\flat\right)^T \quad \boldsymbol{v}^T\right)\boldsymbol{R}'\begin{pmatrix}\boldsymbol{U}^\flat\\ \boldsymbol{v}\end{pmatrix} = \mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[p_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\right]$$

$$\left(\left(\boldsymbol{U}^\flat\right)^T \quad \boldsymbol{v}^T\right)\boldsymbol{R}^*\begin{pmatrix}\boldsymbol{U}^\flat\\ \boldsymbol{v}\end{pmatrix} = \mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[p^*_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\right]$$

where $p_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})$, $p^*_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})$ are polynomial of degree at most 4 whose coefficients depend on $\boldsymbol{U}$ and $\boldsymbol{v}$. Also, observe that for every $\boldsymbol{z}\in\mathbb{R}^d$ we have that $p_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\geq 0$ and $p^*_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\geq 0$. Since the mean of any distribution minimizes the expectation $\mathbb{E}[(x-a)^2]$ with respect to a, we have that $\mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[p_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\right] \geq \mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[p^*_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\right]$. From Lemma 3 $\mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[p^*_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\right] \geq \lambda_m(\boldsymbol{\Sigma})$ and hence

$$\mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[p_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\right] \geq \lambda_m(\boldsymbol{\Sigma}). \tag{3.12}$$

What is left is to bound $\mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[p_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\right]$ with respect to $\mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[p_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\right]$. For this purpose we are going to use Theorem 5 and the fact that $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};S)=\beta$. We define

$$\gamma = \left(\frac{1}{C}\frac{\beta}{8}\right)^4\lambda_m(\boldsymbol{\Sigma}) \quad \text{and} \quad \bar{S} = \{\boldsymbol{x}\in\mathbb{R}^d \mid p_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{x})\leq\gamma\}$$

and applying Theorem 5 together with (3.12) we get that $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma};\bar{S}) \leq \frac{\beta}{2}$. Therefore, when calculating the expectation $\mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[p_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\right]$ at least the half of the mass of the mass of $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)$ is in points $\boldsymbol{z}$ such that $\boldsymbol{z}\notin\bar{S}$. This implies that

$$\mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[p_{(\boldsymbol{U},\boldsymbol{v})}(\boldsymbol{z})\right] \geq \frac{1}{2}\gamma = \frac{1}{2^{13}}\left(\frac{\beta}{C}\right)^4\lambda_m(\boldsymbol{\Sigma}),$$

and the lemma follows. ■

3.2 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 $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$.

We begin by showing that only few truncated samples suffice to obtain accurate estimates $\hat{\boldsymbol{\mu}}_S$ and $\hat{\boldsymbol{\Sigma}}_S$ of the mean and covariance.

Lemma 5. *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 $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S) = \alpha$. Using $\tilde{O}(\frac{d}{\varepsilon^2}\log(1/\alpha)\log^2(1/\delta))$ samples, we can compute estimates $\hat{\boldsymbol{\mu}}_S$ and $\hat{\boldsymbol{\Sigma}}_S$ such that*

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

with probability $1-\delta$.

Proof. Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be the i.i.d samples drawn from the truncated Gaussian, $\hat{\boldsymbol{\mu}}_S = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ and $\hat{\boldsymbol{\Sigma}}_S = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_S)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_S)^T$.

We assume w.l.o.g. that $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{0}, \boldsymbol{I})$. Since the n samples $\boldsymbol{x}_i$ are drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}, S)$, which can be seen as drawing $O(n/\alpha)$ samples $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 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, $|x_{i,j}| \leq \log\left(\frac{nd}{\alpha\delta}\right)$.

By Hoeffding's inequality, we get that:

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

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

Similarly, we can use a matrix concentration inequality ([Ver10], Corollary 5.52) to get that with probability $1-\delta$:

$$(1-\varepsilon)\boldsymbol{\Sigma}_S \preceq \hat{\boldsymbol{\Sigma}}_S \preceq (1+\varepsilon)\boldsymbol{\Sigma}_S$$

if $n \geq \Omega(d/\varepsilon^2\log(nd/\alpha\delta)\log(d/\delta))$.

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

□

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

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

1. $\|\boldsymbol{\mu}_S - \boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} \leq O(\sqrt{\log\frac{1}{\alpha}})$ *and*
2. $\boldsymbol{\Sigma}_S \succeq \Omega(\alpha^2)\boldsymbol{\Sigma}$,

3. $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_S\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{I}\|_F \leq O(\log \frac{1}{\alpha})$.

Proof. First observe that we can assume without loss of generality that $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = \boldsymbol{I}$ and that $\boldsymbol{\Sigma}_S$ is a diagonal matrix with entries $\lambda_1 \leq \cdots \leq \lambda_d$.

1. We will show that
$$\|\boldsymbol{\mu}_S\|_2 \leq \sqrt{2\log\frac{1}{\alpha}} + 1$$
which implies 1. for arbitrary $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ after applying the standard transformation.
Consider the direction of the sample mean $\hat{\boldsymbol{\mu}}_S$. The worst case subset $S \subset \mathbb{R}^d$ of mass at least α that would maximize $\|\boldsymbol{\mu}_S\| = \frac{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\mathbf{1}_{\boldsymbol{x}\in S}\boldsymbol{x}^T\hat{\boldsymbol{\mu}}_S\right]}{\mathbb{E}_{\boldsymbol{x}\sim N(\mathbf{0},\boldsymbol{I})}[\mathbf{1}_{\boldsymbol{x}\in S}]}$ is the following:
$$S = \left\{\boldsymbol{x}^T\hat{\boldsymbol{\mu}}_S > F^{-1}(1-\alpha)\right\}$$
where F is the CDF of the standard normal distribution. Since $\alpha = 1 - F(t) \leq e^{-\frac{t^2}{2}}$, we have that $t \leq \sqrt{2\log(\frac{1}{\alpha})}$. The bound follows since $\mathbb{E}_{x\sim N(0,1)|x\geq t}[x] \leq 1 + t$.
2. We want to bound the following expectation: $\lambda_1 = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I},S)}[(x_1 - \mu_{S,1})^2] = \mathrm{Var}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I},S)}[g_1]$, where g_1 denotes the marginal distribution of g along the direction of $\boldsymbol{e_1}$.
Since $\mathcal{N}(\mathbf{0}, \boldsymbol{I}; S) = \alpha$, the worst case set (i.e the one that minimizes $\mathrm{Var}[g_1]$) is the one that has α mass as close as possible to the hyperplane $x_1 = \mu_{S,1}$. However, the maximum mass that a gaussian places at the set $\{x_1 : |x_1 - \mu_{S,1}| < c\}$ is at most $2c$ as the density of the univariate gaussian is at most 1. Thus the $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I},S)}[(x_1 - \mu_{S,1})^2]$ is at least the variance of the uniform distribution $U[-\alpha/2, \alpha/2]$ which is $\alpha^2/12$. Thus $\lambda_i \geq \lambda_1 \geq \alpha^2/12$.
3. Finally, case 3, follows from Theorem 4. Consider any large set of n samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Theorem 4, implies that with probability $1 - o(1/n)$, for all $T \subseteq [n]$ with $|T| = \Theta(\alpha n)$, we have that $\left\|\sum_{i\in T}\frac{1}{|T|}\boldsymbol{X_i}\boldsymbol{X_i^T} - \boldsymbol{I}\right\|_F \geq O\left(\log(1/\alpha)\right)$. In particular, the same is true for the set of $\Theta(\alpha n)$ samples that lie in the set S. As $n \to \infty$, $\sum_{i\in T}\frac{1}{|T|}\boldsymbol{X_i}\boldsymbol{X_i^T}$ converges to $\boldsymbol{\Sigma}_S + \boldsymbol{\mu}_S\boldsymbol{\mu}_S^T$.
We thus obtain that $\left\|\boldsymbol{\Sigma}_S + \boldsymbol{\mu}_S\boldsymbol{\mu}_S^T - \boldsymbol{I}\right\|_F \leq \sqrt{O\left(\log(1/\alpha)\right)}$, which implies that $\|\boldsymbol{\Sigma}_S - \boldsymbol{I}\|_F \leq \sqrt{O\left(\log(1/\alpha)\right)} + \boldsymbol{\mu}_S\boldsymbol{\mu}_S^T \leq O\left(\log(1/\alpha)\right)$.

□

Combining the two Lemmas and Theorem 4, we get that

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

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

Proof. The first two properties follow by applying Lemma 5 with $\varepsilon = 1/2$ 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. □

3.3 A set of parameters with non-trivial mass in the truncation set

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

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

Lemma 7. *Consider two Gaussians $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, such that*

1. $\|\boldsymbol{I} - \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2}\|_F \leq B$ *and*
2. $1/B \leq \|\boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2}\|_2 \leq B$,
3. $\|\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\| \leq B$.

Suppose that for a set S we have that $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1; S) \geq \alpha$. Then $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2; S) \geq (\alpha/12)^{30B^5}$.

Proof. To prove the statement, we note that $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2; S) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)}[\mathbf{1}_{\boldsymbol{x}\in S}\frac{\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2;x)}{\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1;x)}]$.

Without loss of generality, we may assume that $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = (\mathbf{0}, \boldsymbol{I})$ and that $\boldsymbol{\Sigma}_2 = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. Also, for simplicity we denote $\boldsymbol{\mu}_2$ simply by $\boldsymbol{\mu}$. The given bounds can be rewritten as $\sum_i (1 - 1/\lambda_i)^2 < B^2$, $1/B < \lambda_i < B$ and $\sum_i \mu_i^2/\lambda_i^2 < B$. The last two bounds imply that $\sum_i \mu_i^2 < B^3$, while the first two imply that $\sum_i (1-\lambda_i)^2 < B^4$.

We have that

$$\frac{\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2; x)}{\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1; x)} = \exp(-\sum_i [(x_i - \mu_i)^2/\lambda_i - x_i^2 + \log\lambda_i]) \tag{3.13}$$

We will show that for a random $\boldsymbol{x}$ with high probability (higher than $1 - \alpha/2$) the above ratio is larger than some bound T. This will imply that $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2; S) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)}[\mathbf{1}_{\boldsymbol{x}\in S}\frac{\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2;x)}{\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1;x)}] \geq \frac{\alpha}{2}T$. To obtain the bound T, we note that:

1. if $\lambda_i < 1/2$, then $(x_i - \mu_i)^2/\lambda_i - x_i^2 + \log\lambda_i < B(x_i - \mu_i)^2 < Bx_i^2 - 2B\lambda_i x + \mu_i^2$.
2. if $\lambda_i > 2$, then $(x_i - \mu_i)^2/\lambda_i - x_i^2 + \log\lambda_i < \mu_i^2 - 2x_i\mu_i + \log B$.
3. f $\lambda_i \in [1/2, 2]$, then $(x_i - \mu_i)^2/\lambda_i - x_i^2 + \log\lambda_i < (1/\lambda_i - 1)x_i^2 + 2\mu_i^2 - 4x_i\mu_i + \log\lambda_i$.

We observe that since $\|\boldsymbol{\lambda} - \mathbf{1}\|_2^2 \leq B^4$, the total number of coordinates with $\lambda_i \notin [1/2, 2]$ is at most $4B^4$.

For case 1, we have that $\Pr_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\sum_{i:\lambda_i<1/2} Bx_i^2 > 8B^5\log(8/\alpha)\right] < \alpha/8$. This is true by a tail bound on the norm of a Gaussian vector $\boldsymbol{x}$ in B^4 dimensions.

For cases 1,2, we have that $\Pr_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[-\sum_{i:\lambda_i\notin[1/2,2]} x_i\mu_i > \|\mu\|\log(8/\alpha)\right] < \alpha/8$. This is true by a tail bound on the projection of the Gaussian in the direction $\boldsymbol{\mu}$ which is also normally distributed.

For case 3, we first restrict our attention to all $\lambda_i \in [1, 2]$ and use Theorem 3 to obtain the bound $\Pr_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\sum_{i:\lambda_i\in[1,2]}(1 - 1/\lambda_i)(x_i^2 - 1) < -2\sqrt{\sum_i (1/\lambda_i - 1)^2\log(16/\alpha)}\right] \leq \alpha/16$.

For all $\lambda_i \in [1/2, 1]$, we again apply Theorem 3 to obtain the bound

$$\Pr_{\boldsymbol{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\sum_{i:\lambda_i\in[1,2]}(1/\lambda_i - 1)(x_i^2 - 1) < 2\sqrt{\sum_i (1/\lambda_i - 1)^2\log(16/\alpha) + 4\log(16/\alpha)}\right] \leq \alpha/16.$$

We thus have that with probability $1-\alpha/8$, $\sum_{i:\lambda_i\in[1/2,2]}(1/\lambda_i-1)x_i^2 \le \sum_{i:\lambda_i\in[1/2,2]}(1/\lambda_i-1) + 2\sqrt{\sum_i(1/\lambda_i-1)^2\log(16/\alpha)} + 4\log(16/\alpha) \le \sum_{i:\lambda_i\in[1/2,2]}(1/\lambda_i-1) + 8\log(16/\alpha)$

However, for all $\lambda_i \in [1/2,2]$, we have that $(1/\lambda_i-1)+\log\lambda_i \le 4(\lambda_i-1)^2$. Thus overall, we have that with probability at least $1-\alpha/4$

$$\sum_{i:\lambda_i\in[1/2,2]}(x_i-\mu_i)^2/\lambda_i - x_i^2 + \log\lambda_i \le \sum_{i:\lambda_i\in[1/2,2]}(4(\lambda_i-1)^2 + 2\mu_i^2 - 4x_i\mu_i) + 8\log(16/\alpha) \le 16B^3\log(12/\alpha)$$

where we worked as in case 1 and 2 to bound the contribution of $2\mu_i^2 - 4x_i\mu_i$.

Overall, we obtain that with probability at least $1-\alpha/2$, $\exp(-\sum_i[(x_i-\mu_i)^2/\lambda_i - x_i^2 + \log\lambda_i]) > \exp(-30B^5\log(12/\alpha))$.

This implies that the probability that $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ assigns to the set S is at least $(\alpha/12)^{30B^5}$. □

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

Proposition 1. *It holds that*

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

Proof. W.l.o.g. $\boldsymbol{\mu} = \boldsymbol{0}$, $\boldsymbol{\Sigma} = \boldsymbol{I}$ and $\hat{\boldsymbol{\Sigma}}_S = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. From Corollary 1, we have that

$$\|\boldsymbol{I} - \boldsymbol{\Sigma}^{*1/2}\hat{\boldsymbol{\Sigma}}_S{}^{-1}\boldsymbol{\Sigma}^{*1/2}\|_F = \sum_i(1-1/\lambda_i)^2 \le \frac{1}{\alpha^2}\sum_i(1-\lambda_i)^2 \le O\left(\frac{\log(1/\alpha)}{\alpha^2}\right).$$

Moreover, we have that $\|\hat{\boldsymbol{\Sigma}}_S{}^{-1}\boldsymbol{\Sigma}^{*1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\| = \sum_i \frac{1}{\lambda_i^2}\mu_i^2 \le O\left(\frac{\log(1/\alpha)}{\alpha^2}\right)$

Similarly, we have that

$$\|\boldsymbol{I} - \hat{\boldsymbol{\Sigma}}_S{}^{*1/2}\boldsymbol{\Sigma}^{*-1}\hat{\boldsymbol{\Sigma}}_S{}^{*1/2}\|_F = \sum_i(1-\lambda_i)^2 \le O\left(\log(1/\alpha)\right).$$

Moreover, we have that $\|\hat{\boldsymbol{\Sigma}}_S{}^{-1}\boldsymbol{\Sigma}^{*1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\| = \sum_i \lambda_i^2\mu_i^2 \le O\left(\frac{\log(1/\alpha)}{\alpha^2}\right)$ □

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

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

Corollary 3. *Consider a truncated normal distribution $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$ with $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; S) \ge \alpha > 0$. Let $(\hat{\boldsymbol{\mu}}_S, \hat{\boldsymbol{\Sigma}}_S)$ be the estimate obtained by Corollary 1 and let $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be any estimate that satisfies*

- $\|\boldsymbol{I} - \hat{\boldsymbol{\Sigma}}_S{}^{1/2}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}_S{}^{1/2}\|_F \le O(\frac{\log(1/\alpha)}{\alpha^2})$,
- $\Omega(\alpha^2) \le \|\hat{\boldsymbol{\Sigma}}_S{}^{1/2}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}_S{}^{1/2}\|_2 \le O(1/\alpha^2)$,
- $\|\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}_S{}^{1/2}(\hat{\boldsymbol{\mu}}_S - \boldsymbol{\mu})\| \le O(\frac{\log(1/\alpha)}{\alpha^2})$.

Then, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S) \ge c_\alpha$ for some constant c_α that depends only on the constant $\alpha > 0$.

3.4 Analysis of SGD: Proof of Theorem 1

We define $\boldsymbol{t} = \boldsymbol{T}^\flat$ and $\boldsymbol{w} = \begin{pmatrix} \boldsymbol{t} \\ \boldsymbol{v} \end{pmatrix}$ for simplicity. Our goal is to run *projected stochastic gradient descent* to optimize $\bar{\ell}$ with respect to the parameters $\boldsymbol{v}, \boldsymbol{t}$ as we describe in detail in Figure 2. This algorithm iterates over the estimation $\boldsymbol{w}$ of the true parameters $\boldsymbol{w}^*$. Let also $\mathcal{O}$ the sample oracle from the unknown distribution $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$.

The initialization step of our algorithm computes the empirical mean $\hat{\boldsymbol{\mu}}$ and the empirical covariance matrix $\hat{\boldsymbol{\Sigma}}$ of the truncated distribution $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)$ using $n = \tilde{O}\left(\frac{d^2}{\varepsilon^2}\right)$ samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

$$\hat{\boldsymbol{\mu}}_S = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \qquad \hat{\boldsymbol{\Sigma}}_S = \frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_S\right)\left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_S\right)^T. \tag{3.14}$$

Then, we transform our space such that $\hat{\boldsymbol{\mu}}_S \mapsto \boldsymbol{0}$ and $\hat{\boldsymbol{\Sigma}}_S \mapsto \boldsymbol{I}$.

We are now ready to describe the domain $\mathcal{D}$ of our projected stochastic gradient ascent in the transformed space where $\hat{\boldsymbol{\mu}}_S = \boldsymbol{0}$ and $\hat{\boldsymbol{\Sigma}}_S = \boldsymbol{I}$. The domain is parameterized by the positive number r and is defined as follows

$$\mathcal{D}_r = \left\{ (\boldsymbol{v}, \boldsymbol{T}) \;\middle|\; \|\boldsymbol{v}\|_2 \le r,\ \|\boldsymbol{I} - \boldsymbol{T}\|_F \le r,\ \left\|\boldsymbol{T}^{-1}\right\|_2 \le r \right\}. \tag{3.15}$$

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

Lemma 8. *Given $(\boldsymbol{v}', \boldsymbol{T}')$, there exists an efficient algorithm that solves the following problem which corresponds to projecting $(\boldsymbol{v}, \boldsymbol{T})$ to the set $\mathcal{D}_r$*

$$\arg \min_{(\boldsymbol{v},\boldsymbol{T}) \in \mathcal{D}_r} \left\|\boldsymbol{v} - \boldsymbol{v}'\right\|_2^2 + \left\|\boldsymbol{T} - \boldsymbol{T}'\right\|_F^2.$$

Proof of Lemma 8: Because of the form of $\mathcal{D}_r$ and the objective function of our program, observe that we can project $\boldsymbol{v}$ and $\boldsymbol{T}$ separetely. The projection of $\boldsymbol{v}'$ to $\mathcal{D}_r$ is the solution of the following problem $\arg\min_{\boldsymbol{v} | \|\boldsymbol{v}\|_2 \le r} \|\boldsymbol{v} - \boldsymbol{v}'\|_2^2$ which has a closed form. So we focus on the projection of $\boldsymbol{T}$. To project $\boldsymbol{T}$ to $\mathcal{D}_r$ we need to solve the following program.

$$\begin{aligned} \arg\min_{\boldsymbol{T}} & \left\|\boldsymbol{T} - \boldsymbol{T}'\right\|_F^2 \\ \text{s.t.} \quad & \|\boldsymbol{T} - \boldsymbol{I}\|_F^2 \le r^2 \\ & \boldsymbol{T} \succeq \frac{1}{r}\boldsymbol{I} \end{aligned}$$

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

$$\begin{aligned} \arg\min_{\boldsymbol{T}} & \left\|\boldsymbol{T} - \boldsymbol{T}'\right\|_F^2 + \lambda \|\boldsymbol{T} - \boldsymbol{I}\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{T} \succeq \frac{1}{r}\boldsymbol{I} \end{aligned}$$

After completing the squares on the objective, we can set $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{1+\lambda}\left(\boldsymbol{T}' + \lambda \boldsymbol{I}\right)$ and make the change of variables $\boldsymbol{R} = \boldsymbol{I} - \boldsymbol{T}$ and solve the program.

$$\begin{aligned} &\arg\min_{\boldsymbol{T}} \|\boldsymbol{R} - \boldsymbol{H}\|_F^2 \\ &\text{s.t.} \quad \boldsymbol{R} \preceq \left(1 - \frac{1}{r}\right) \boldsymbol{I} \end{aligned}$$

Observe now that without loss of generality $\boldsymbol{H}$ is diagonal. If this is not the case we can compute the singular value decomposition of $\boldsymbol{H}$ and change the base of the space so that $\boldsymbol{H}$ is diagonal. Then, after finding the answer to this case we transform back the space to get the correct $\boldsymbol{R}$. When $\boldsymbol{H}$ 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 $\boldsymbol{v}^{(i)}$ the gradient of the log likelihood function at step i as computed in the line 6 of Algorithm 1. Let $\boldsymbol{\nu}$, $\boldsymbol{T}$ be the guesses of the parameters after step $i-1$ according to which the gradient is computed with $\boldsymbol{\mu} = \boldsymbol{T}^{-1}\boldsymbol{\nu}$ and $\boldsymbol{\Sigma} = \boldsymbol{T}^{-1}$. Let $\boldsymbol{\mu}^*$, $\boldsymbol{\Sigma}^*$ be the parameters we want to recover, with $\boldsymbol{\nu}^* = \boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*$ and $\boldsymbol{T}^* = \boldsymbol{\Sigma}^{*-1}$. Define We assume that $(\boldsymbol{\nu}, \boldsymbol{T}) \in \mathcal{D}_r$, $(\boldsymbol{\nu}^*, \boldsymbol{T}^*) \in \mathcal{D}_r$ and that $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S) \geq \beta$, $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; S) \geq \beta$. Then, we have that*

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

Proof. Let $\boldsymbol{\mu} = \boldsymbol{T}^{-1}\boldsymbol{\nu}$ and $\boldsymbol{\Sigma} = \boldsymbol{T}^{-1}$. According to Algorithm 2 and equation (3.6) we have that

$$\begin{aligned} \mathbb{E}\left[\left\|\boldsymbol{v}^{(i)}\right\|_2^2\right] &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)}\left[\mathbb{E}_{\boldsymbol{y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\left\|\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^T\right)^\flat \\ \boldsymbol{x}\end{pmatrix} - \begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{y}\boldsymbol{y}^T\right)^\flat \\ \boldsymbol{y}\end{pmatrix}\right\|_2^2\right]\right] \\ &\leq 3\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*,S)}\left[\left\|\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^T\right)^\flat \\ \boldsymbol{x}\end{pmatrix}\right\|_2^2\right] + 3\mathbb{E}_{\boldsymbol{y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\left\|\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{y}\boldsymbol{y}^T\right)^\flat \\ \boldsymbol{y}\end{pmatrix}\right\|_2^2\right]. \end{aligned}$$

In order to bound each of these terms we use two facts: (1) that the parameters $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ belong in $\mathcal{D}_r$, (2) the measure of S is greater than β for both sets of parameters, i.e. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S) \geq \beta$ and $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; S) \geq \beta$. Hence, we will show how to get an upper bound for the second term and the upper bound of the first term follows the same way.

$$\begin{aligned} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\left\|\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^T\right)^\flat \\ \boldsymbol{x}\end{pmatrix}\right\|_2^2\right] &= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\left\|\boldsymbol{x}\boldsymbol{x}^T\right\|_F^2\right] + \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\|\boldsymbol{x}\|_2^2\right] \\ &\leq \frac{1}{2}\frac{1}{\beta}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\left\|\boldsymbol{x}\boldsymbol{x}^T\right\|_F^2\right] + \frac{1}{\beta}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\|\boldsymbol{x}\|_2^2\right] \end{aligned}$$

We define $\boldsymbol{\rho} = \begin{pmatrix}\sigma_{11} & \sigma_{22} & \cdots & \sigma_{dd}\end{pmatrix}^T$, by straightforward calculations of the last two expressions we get that

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\left\|\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^T\right)^\flat\\ \boldsymbol{x}\end{pmatrix}\right\|_2^2\right] \leq \frac{2}{\beta}\left(\|\boldsymbol{\Sigma}\|_F^2+\|\boldsymbol{\rho}\|_2^4+\|\boldsymbol{\rho}\|_2^2\|\boldsymbol{\mu}\|_2^2+2\boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}+\|\boldsymbol{\mu}\|_2^4+\|\boldsymbol{\rho}\|_2^2+\|\boldsymbol{\mu}\|_2^2\right)$$

But from the fact that $(\boldsymbol{\nu},\boldsymbol{T})\in\mathcal{D}_r$ we can get the following bounds

$$\|\boldsymbol{\Sigma}\|_F^2\leq d\cdot r,\quad \|\boldsymbol{\rho}\|_2^2\leq d\cdot r,\quad \boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}\leq r,\quad \|\boldsymbol{\mu}\|_2^2\leq r$$

From which we get that

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma},S)}\left[\left\|\begin{pmatrix}\left(-\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^T\right)^\flat\\ \boldsymbol{x}\end{pmatrix}\right\|_2^2\right] \leq \frac{16}{\beta}\cdot d^2\cdot r^2$$

Finallly, we apply these bounds to the first bound for $\mathbb{E}\left[\left\|\boldsymbol{v}^{(i)}\right\|_2^2\right]$ and the lemma follows. □

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 [SSBD14].

Theorem 6 (Theorem 14.11 of [SSBD14].)**.** *Let $f=-\bar{\ell}$. 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 $\boldsymbol{w}^*\in\arg\min_{\boldsymbol{w}\in\mathcal{D}_r}f(\boldsymbol{w})$ be an optimal solution. Then,*

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

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

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

Lemma 10 (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*

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

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

$$\beta_*=\min_{(\boldsymbol{\nu},\boldsymbol{T})\in\mathcal{D}_r}\mathcal{N}(\boldsymbol{T}^{-1}\boldsymbol{\nu},\boldsymbol{T}^{-1};S)\quad\text{and}\quad\lambda_*=\min_{(\boldsymbol{\nu},\boldsymbol{T})\in\mathcal{D}_r}\lambda_m(\boldsymbol{T}^{-1})\geq r.$$

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

$$\mathbb{E}\left[f(\bar{\boldsymbol{w}})\right]-f(\boldsymbol{w}^*)\leq\frac{C\cdot r}{\beta_*^5}\cdot\frac{d^2}{M}\left(1+\log(M)\right),$$

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

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

$$\mathbb{P}\left(f(\bar{\boldsymbol{w}}) - f(\boldsymbol{w}^*) \geq \frac{3C \cdot r}{\beta_*^5} \cdot \frac{d^2}{M}\left(1 + \log(M)\right)\right) \leq \frac{1}{3}. \tag{3.16}$$

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 $\mathcal{E} = \{\bar{\boldsymbol{w}}_1, \bar{\boldsymbol{w}}_2, \ldots, \bar{\boldsymbol{w}}_K\}$. Our ideal final estimate would be

$$\hat{\boldsymbol{w}} = \arg\min_{\bar{\boldsymbol{w}} \in \mathcal{E}} \bar{\ell}(\boldsymbol{w})$$

but we don't have access to the exact value of $\bar{\ell}(\boldsymbol{w})$. Because of (3.16) 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 $\bar{\ell}(\boldsymbol{w}) - \bar{\ell}(\boldsymbol{w}^*) \leq \eta$ where $\eta = \frac{3C \cdot r}{\beta_*^5} \cdot \frac{d^2}{M}\left(1 + \log(M)\right)$. Moreover we will prove later that $\bar{\ell}(\boldsymbol{w}) - \bar{\ell}(\boldsymbol{w}^*) \leq \eta$ 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 condition on the event $f(\hat{\boldsymbol{w}}) - f(\boldsymbol{w}^*) \leq \frac{2C \cdot r}{\beta_*^5} \cdot \frac{d^2}{M}\left(1 + \log(M)\right)$ and we only lose probability at most δ. 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 $\beta^* \geq c_\alpha$ where is a constant c_α that depends only on the constant α. Hence, with probability at least $1-\delta$ we have that

$$f(\hat{\boldsymbol{w}}) - f(\boldsymbol{w}^*) \leq c'_\alpha \cdot \frac{d^2}{M}\left(1 + \log(M)\right),$$

where c'_α is a constant that depends only on α. Now we can use Lemma 10 to get that

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2 \leq c''_\alpha \sqrt{\frac{d^2}{M}\left(1 + \log(M)\right)}. \tag{3.17}$$

Also, it holds that

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

Hence, for $M \geq \tilde{O}\left(\frac{d^2}{\varepsilon^2}\right)$ and using (3.17) we have that

$$\left\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*\right\|_2^2 + \left\|\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{*-1}\right\|_F^2 \leq \varepsilon.$$

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

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

procedure SGD(M, λ) ▷ M: number of steps, λ: parameter
  $\boldsymbol{w}^{(0)} \leftarrow \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_S^\flat \\ \hat{\boldsymbol{\mu}}_S \end{pmatrix}$
  for $i = 1, \ldots, M$ **do**
    Sample $\boldsymbol{x}^{(i)}$ from $\mathcal{O}$
    $\eta_i \leftarrow \frac{1}{\lambda \cdot i}$
    $\boldsymbol{v}^{(i)} \leftarrow \text{GRADIENTESTIMATION}(\boldsymbol{x}^{(i)}, \boldsymbol{w}^{(i-1)})$
    $\boldsymbol{r}^{(i)} \leftarrow \boldsymbol{w}^{(i-1)} - \eta_i \boldsymbol{v}^{(i)}$
    $\boldsymbol{w}^{(i)} \leftarrow \text{PROJECTTODOMAIN}(\boldsymbol{r}^{(i)})$
  return $\bar{\boldsymbol{w}} \leftarrow \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{w}^{(i)}$ ▷ Output the average.

Algorithm 2 The function to estimate the gradient of log-likelihood as in (3.6).

function GRADIENTESTIMATION($\boldsymbol{x}, \boldsymbol{w}$) ▷ $\boldsymbol{x}$: sample from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$
  $\begin{pmatrix} \boldsymbol{T}^\flat \\ \boldsymbol{v} \end{pmatrix} \leftarrow \boldsymbol{w}$
  $\boldsymbol{\mu} \leftarrow \boldsymbol{T}^{-1}\boldsymbol{v}$
  $\boldsymbol{\Sigma} \leftarrow \boldsymbol{T}^{-1}$
  repeat
    Sample $\boldsymbol{y}$ from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  until $M_S(\boldsymbol{y}) = 1$ ▷ M_S is the membership oracle of the set S.
  return $-\begin{pmatrix} \left(-\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^T\right)^\flat \\ \boldsymbol{x} \end{pmatrix} + \begin{pmatrix} \left(-\frac{1}{2}\boldsymbol{y}\boldsymbol{y}^T\right)^\flat \\ \boldsymbol{y} \end{pmatrix}$

Algorithm 3 The function that projects a current guess back to the domain $\mathcal{D}$.

function PROJECTTODOMAIN($\boldsymbol{r}$) ▷ let r be the parameter of the domain $\mathcal{D}_r$
  $\begin{pmatrix} \boldsymbol{T}^\flat \\ \boldsymbol{v} \end{pmatrix} \leftarrow \boldsymbol{r}$
  $\boldsymbol{v}' \leftarrow \arg\min_{\boldsymbol{b}:\|\boldsymbol{b}\| \le r_1} \|\boldsymbol{b} - \boldsymbol{v}\|_2^2$
  repeat binary search over λ
    solve the projection problem

$$\boldsymbol{T}' \leftarrow \arg\min_{\boldsymbol{T}' \succeq r_3 \boldsymbol{I}} \left\|\boldsymbol{T} - \boldsymbol{T}'\right\|_F^2 + \lambda \left\|\boldsymbol{I} - \boldsymbol{T}'\right\|_F^2$$

  until a $\boldsymbol{T}'$ is found with minimum objective value and $\|\boldsymbol{I} - \boldsymbol{T}'\|_F^2 \le r_2^2$
  return $(\boldsymbol{v}', \boldsymbol{T}')$

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

4 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 $\mathcal{N}(\mu_1, \sigma_1^2; S_1)$, $\mathcal{D}_1$, and a distribution over truncation sets S_2 with $\mathcal{N}(\mu_2, \sigma_2^2; S_2)$, $\mathcal{D}_2$, such that a random instance $\{S_1, \mathcal{N}(\mu_1, \sigma_1^2)\}$ with S_1 drawn from $\mathcal{D}_1$ is indistinguishable from a random instance $\{S_2, \mathcal{N}(\mu_2, \sigma_2^2)\}$ with S_2 drawn from $\mathcal{D}_2$.

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 $\mathcal{D}_i$: The random set $S_i \sim \mathcal{D}_i$ is constructed by adding every point $x \in \mathbb{R}$ with probability $\min\{\mathcal{N}(\mu_{3-i}, \sigma_{3-i}^2; x)/\mathcal{N}(\mu_i, \sigma_i^2; x), 1\}$.

Now consider the distribution p with density $\frac{1}{\alpha}\min\{\mathcal{N}(\mu_1, \sigma_1^2; x), \mathcal{N}(\mu_2, \sigma_2^2; x)\}$. 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 $\mathcal{N}(\mu_i, \sigma_i^2)$ and reject it with probability $\min\{\mathcal{N}(\mu_{3-i}, \sigma_{3-i}^2; x)/\mathcal{N}(\mu_i, \sigma_i^2; x), 1\}$.

We now argue that samples from the distribution $\mathcal{N}(\mu_i, \sigma_i^2; S_i)$ for a random $S_i \sim \mathcal{D}_i$ are indistinguishable from p. This is because an alternative way of sampling the distribution $\mathcal{N}(\mu_i, \sigma_i^2; S_i)$ can be sampled as follows. Draw a sample x from $\mathcal{N}(\mu_i, \sigma_i^2)$ and then check if $x \in S_i$. By the principle of deferred randomness, we may not commit to a particular set S_i only decide whether $x \in S_i$ 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 α, 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

$$\text{dist}_{TV}(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)) > \frac{1-\alpha}{2} \quad \text{or} \quad \text{dist}_{TV}(\mathcal{N}(\mu_2, \sigma_2^2), \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)) > \frac{1-\alpha}{2}$$

.

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 ε of the randomly chosen points. In this case the collision probability is no longer 0 and depends on ε. Given that no collisions are seen the two cases are again indistinguishable. Moreover, for very small ε, an extremely large number of samples are needed to see a collision.*

Acknowledgements

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

References

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

[BDLS17] Sivaraman Balakrishnan, Simon S. Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 169–212, 2017.

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

[BJK15] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

[CLMW11] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[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.

[DKK+16a] 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+16b] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Zheng Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. *CoRR*, abs/1604.06443, 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.

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

[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.

[HM13] Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Conference on Learning Theory*, pages 354–375, 2013.

[Hot48] Harold Hotelling. Fitting generalized truncated normal distributions. In *Annals of Mathematical Statistics*, volume 19, pages 596–596, 1948.

[Lee14] Alice Lee. Table of the Gaussian "Tail" Functions; When the "Tail" is Larger than the Body. *Biometrika*, 10(2/3):208–214, 1914.

[Li17] Jerry Li. Robust sparse estimation tasks in high dimensions. *CoRR*, abs/1702.05860, 2017.

[LM00] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[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.

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

[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.

[SCV18] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 45:1–45:21, 2018.

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

[Tuk49] John W Tukey. Sufficiency, truncation and selection. *The Annals of Mathematical Statistics*, pages 309–311, 1949.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[XCM10] Huan Xu, Constantine Caramanis, and Shie Mannor. Principal component analysis with contaminated data: The high dimensional case. *arXiv preprint arXiv:1002.4658*, 2010.