Learning High-dimensional Gaussians from Censored Data

Arnab Bhattacharyya
The University of Warwick

Constantinos Daskalakis
Massachusetts Institute of Technology

Themis Gouleakis
Nanyang Technological University

Yuhao Wang
National University of Singapore

Abstract

We provide efficient algorithms for the problem of distribution learning from high-dimensional Gaussian data where in each sample, some of the variable values are missing. We suppose that the variables are *missing not at random (MNAR)*. The *missingness model*, denoted by $\mathbb{S}(\mathbf{y})$, is the function that maps any point $\mathbf{y} \in \mathbb{R}^d$ to the subsets of its coordinates that are seen. In this work, we assume that it is known. We study the following two settings:

(i) [**Self-censoring**] An observation $\mathbf{x}$ is generated by first sampling the true value $\mathbf{y}$ from a d-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ with unknown $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$. For each coordinate i, there exists a set $S_i \subseteq \mathbb{R}^d$ such that $x_i = y_i$ if and only if $y_i \in S_i$. Otherwise, x_i is missing and takes a generic value (e.g "?").
We design an algorithm that learns $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ up to TV distance ε, using poly $(d, 1/\varepsilon)$ samples, assuming only that each pair of coordinates is observed with sufficiently high probability.

(ii) [**Linear thresholding**] An observation $\mathbf{x}$ is generated by first sampling $\mathbf{y}$ from a d-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ with unknown $\boldsymbol{\mu}^*$ and known $\boldsymbol{\Sigma}$, and then applying the missingness model $\mathbb{S}$ where $\mathbb{S}(\mathbf{y}) = \{i \in [d] : \mathbf{v}_i^T \mathbf{y} \leq b_i\}$ for some $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^d$ and $b_1, \ldots, b_d \in \mathbb{R}$. We design an efficient mean estimation algorithm, assuming that none of the possible missingness patterns is very rare conditioned on the values of the observed coordinates and that any small subset of coordinates is observed with sufficiently high probability.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

Missing data is a quite prevalent factor contributing to bias in statistical inference. It arises from various causes, such as limitations in instruments leading to unreliable data, incomplete data collection resulting in missing relevant information, societal biases influencing the suppression of observations, behavioral biases leading to subjects dropping out of studies or avoiding survey questions, ethical, legal, or privacy considerations restricting the utilization of collected data, and other similar factors. Unfortunately training models without consideration of missing data can lead to models that incorporate biases in the training data and make incorrect predictions, which may in turn reinforce those biases when the models are deployed.

Since the early days of statistics, missing data has been a well-known challenge in statistical inference, which occurs in a variety of domains, such as biology, physics, clinical trial design, genetics, economics, survey research, and the social sciences. It has motivated a vast effort towards developing methodologies that are more robust to missing data. As example, we refer the reader to some of the early works in statistics Galton (1898); Pearson (1902); Pearson and Lee (1908); Lee (1914); Fisher (1931), some standard references in statistics and econometrics Tobin (1958); Amemiya (1973); Hausman and Wise (1977); Heckman (1979); Hajivassiliou and McFadden (1998); Little and Rubin (2019), works targeting missing data in specific domains Warga (1992); Brick and Kalton (1996); Troyanskaya et al. (2001); Armitage et al. (2008); Honaker and King (2010), books overviewing this literature Maddala (1986); Breen et al. (1996); Balakrishnan and Cramer (2014), and finallly some recent work in computer science Mohan et al. (2013); Daskalakis et al. (2018, 2019, 2020, 2021a,b); Kontonis et al. (2019); Fotakis et al. (2020); Plevrakis (2021).

The effect that data misssingness has on statistical inference depends heavily on the missingness model. In general, missingness models in which the value of some datapoint influences whether or not it will be

missing from the dataset are harder to deal with compared to models in which this happens randomly. Techniques that have been extensively researched in scenarios where missingness either does not depend on the data or only depends on the observed data are referred to as missing completely at random (MCAR) and missing at random (MAR) respectively (Rubin, 1976; Tsiatis, 2006; Little and Rubin, 2019). In problems where missing entries depend on the underlying values which are themselves censored, known as missing not at random (MNAR), is substantially more difficult and less explored (Robins and Gill, 1997; Rotnitzky and Robins, 1997; Scharfstein et al., 1999; Shpitser et al., 2015; Adak et al., 2020). The MNAR model is quite often relevant in practical applications. For example, the depression registry for mental health status is more likely to have missing questionnaires leading to the self-censoring missingness (Carreras et al., 2021). Data are missing by design due to the limitations of measurement resources, or the treatment discontinuation when participants go off-control due to the lack of tolerability (Little et al., 2012).

The goal of this work is to advance our understanding of density estimation in the non-asymptotic sample regime when data is missing not at random. In particular, we consider the standard task of high-dimensional Gaussian distribution estimation, albeit in settings where every sample of the Gaussian may have a subset of its coordinates censored and which subset this is depends on the sample itself. We consider two models for how the censoring may depend on the sample:

- Self-censoring model (see Section 1.1.1): in this model, a sample $\mathbf{y}$ is drawn from an underlying Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, and each coordinate y_i of this sample is censored (i.e. replaced with a '?') depending on whether or not it satisfies a coordinate-specific Boolean predicate, i.e. whether $S_i(y_i) = 1$ or not.
- Linear thresholding model (see Section 1.1.2): in this more challenging model, whether or not each coordinate is seen depends on whether the **whole** sample satisfies a coordinate-specific predicate.

Our goal in both cases is to identify conditions on the predicates and the underlying distribution under which their parameters can be estimated computationally and sample-efficiently in each of the aforedescribed models. Our work advances prior research on Gaussian estimation in the presence of MNAR data in the non-asymptotic sample regime along the following axes:

- Gaussian estimation under censoring (see e.g. the classical works of Galton (1898); Pearson (1902); Pearson and Lee (1908); Fisher (1931) and the ensuing literature): Prior work on this problem in the non-asymptotic sample regime studies the "all-or-nothing setting," where either all coordinates or no-coordinate can be observed (Daskalakis et al., 2018). They also require that some absolute constant fraction of the Gaussian can be observed. In comparison to this work, we allow heavily corrupted data where no such constant fraction exists where no coordinate is missing. However, we allow the predicate determining the censoring of coordinate i to either be very general but only dependent on this coordinate (self-censoring model), or depend on all coordinates but be simpler, namely a hyperplane (linear-thresholding model).
- Gaussian estimation under self-selection (see e.g. the classical work of Roy (1951) and the ensuing literature): Prior work on this problem in the non-asymptotic sample regime (Cherapanamjeri et al., 2023) studies specific selection mechanisms (in particular hiding all but the maximum coordinate of each sample) and also assumes independence among the coordinates. In comparison to this work, we allow correlations among coordinates and more general masking mechanisms. However, we focus on Gaussian distributed coordinates while they can accommodate non-parametric distributions.

1.1 OUR CONTRIBUTIONS

In this work, we are interested in recovering the "uncorrupted" Gaussian distribution given samples from a "corrupted" distribution (according to our missingness model). We use a population maximum likelihood approach as the estimation algorithm, and apply projected stochastic gradient descent on the likelihood function. We give theoretical proof of fast convergence in the parameter space.

A *missingness model* is defined by a function $\mathbb{S} : \mathbb{R}^d \to 2^{[d]}$. For an underlying d-dimensional vector $\mathbf{y}$, $\mathbb{S}(\mathbf{y})$ is interpreted as the set of coordinates of $\mathbf{y}$ that are not missing. An observation is a pair $(A, \mathbf{x})$, where $A = \mathbb{S}(\mathbf{y})$ and $\mathbf{x} = \mathbf{y}_A$ for an underlying sample $\mathbf{y} \in \mathbb{R}^d$.

1.1.1 DISTRIBUTION LEARNING UNDER THE SELF-CENSORING MECHANISM

Self-censoring Missingness Model. The Self-censoring mechanism is commonly encountered in practice. In this model the missingness of an outcome is affected by its underlying value. For example, smokers are not willing to report their smoking behavior in insurance applications. Voters holding particular beliefs may not disclose their political preferences in election surveys. The Self-censoring model is of significant interest because the model is: (i) conceptually well-motivated, and (ii) can be considered as a baseline for

other more complex missingness models. We say that $\mathbb{S}$ is a *self-censoring missingness model* if there exist sets $S_1, \ldots, S_d$ such that $\mathbb{S}(\mathbf{y}) = \{i \in [d] : y_i \in S_i\}$. Our result in this setting rests on the following hypothesis:

Assumption 1.1. *For any pair of coordinates $i, j \in [d]$: $\Pr_{\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*)}[y_i \in S_i, y_j \in S_j] \geq \alpha$.*

Theorem 1.2. *Suppose we can observe samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ censored through a self-censoring missingness model $\mathbb{S}$. If Assumption 1.1 is satisfied for some constant value of the parameter α, there exists a polynomial-time algorithm that recovers estimated $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$ with arbitrary accuracy. Specifically, for all $\varepsilon > 0$, and given that the eigenvalues of $\boldsymbol{\Sigma}^*$ lie in the interval $[\lambda_{min}, \lambda_{max}]$, the algorithm uses $\tilde{\mathcal{O}}\left(\frac{d^2(\lambda_{max}/\lambda_{min})^2}{\alpha\varepsilon^2}\right)$ samples and produces estimates that satisfy the following:*

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

Note that the sample complexity is proportional to $1/\alpha$. Furthermore, under the above conditions, we have $\mathrm{d_{TV}}(\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})) \leq \mathcal{O}(\varepsilon)$.

1.1.2 MEAN ESTIMATION UNDER LINEAR THRESHOLDING MISSINGNESS

We say $\mathbb{S}$ is a *linear thresholding missingness model* if there exist $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^d$ and $b_1, \ldots, b_d \in \mathbb{R}$ such that $\mathbb{S}(\mathbf{y}) = \{i \in [d] : \mathbf{v}_i^T\mathbf{y} \leq b_i\}$. For instance, if for a pair of coordinates x_i, x_j, we can only observe the maximum of the two, this can be modeled by a linear thresholding model where the i'th coordinate is observed if $x_j - x_i \leq 0$ and the j'th coordinate is observed if $x_i - x_j \leq 0$.

Our main algorithmic result rests on the following two data hypotheses:

Assumption 1.3. *There exist some $\alpha, \beta > 0$ such that for any set $A \subseteq [d]$ of size at most βd,*

$$\Pr_{\mathbf{y}\sim N(\boldsymbol{\mu}^*,\boldsymbol{\Sigma})}[A \subseteq \mathbb{S}(\mathbf{y})] \geq \alpha.$$

Note that we only consider the case where βd is a positive integer without loss of generality.

Assumption 1.4. *(Informal) There exists an* anchoring set *of coordinates C such that (i) C is observed in every sample, and (ii) conditioned on the values at C, each missingness pattern occurs with probability* 0 *or at least γ.*

The second assumption states that the anchoring subset is compulsorily observed and values at these coordinates determine the missingness pattern (the set of coordinates observed) almost fully. This is in analogy to the anchor topic modeling used in natural language processing, which is a variation of probabilistic topic modeling that incorporates a set of predefined "anchor words" to guide the topic modeling process. Our anchored missingness is similar to the "anchor words" assumption. For instance, the anchoring subset might be a set of questions in a questionnaire that are mandatory to answer and whose values are very indicative of the respondant's behavior. We now informally state our main algorithmic result here:

Theorem 1.5. *For a known covariance matrix $\boldsymbol{\Sigma}$, suppose we can observe samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ censored through a linear thresholding missingness model $\mathbb{S}$. If Assumption 1.3 and Assumption 1.4 are satisfied, there exists a polynomial-time algorithm that recovers estimated $\boldsymbol{\mu}^*$ with arbitrary accuracy. Specifically, for all $\varepsilon > 0$, the algorithm uses $\mathrm{poly}(d, 1/\alpha, 1/\beta, 1/\gamma, \lambda_{\max}(\boldsymbol{\Sigma})/\lambda_{\min}(\boldsymbol{\Sigma}), 1/\varepsilon, \log(1/\delta))$ samples and running time, and with probability at least $1-\delta$, produces an estimate $\hat{\boldsymbol{\mu}}$ such that $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|_\Sigma \leq \varepsilon$.*

1.2 OUR TECHNIQUES

Self Censoring In the self censoring model, we show that the problem can be reduced to solving truncation problems in each of the 2-dimensional subspaces spanned by pairs e_i, e_j of the basis vectors. This allows us to use a 2-dimensional version of the algorithm in Daskalakis et al. (2018) as subroutine for our algorithm in order to extract the information about pairwise correlation of coordinates needed to reconstruct the true covariance matrix $\boldsymbol{\Sigma}^*$. The mean is reconstructed in a more straightforward way via solving 1-dimensional truncation problems for each coordinate.

In particular, to estimate the diagonal entries of $\boldsymbol{\Sigma}^*$, we use 1-dimensional subproblems and for each off-diagonal entry Σ_{ij} we solve the 2×2 subproblem on the coordinates i and j, but we only care about the off-diagonal entries of the result (ignoring the rest) and we use exactly that value as the estimate for the entry. To achieve the required guarantees, assuming the condition number of the true covariance matrix $\boldsymbol{\Sigma}^*$ is constant, one needs to run the subproblems with $\varepsilon' = c \cdot \varepsilon/d$ for some constant $c < 1$ (since the Frobenious distance can be at most a factor d larger than the maximum entry-wise difference), which would imply a sample complexity of $O(1/\varepsilon'^2)$ for each subproblem. Therefore, we will need $O(1/\varepsilon'^2) = O(d^2/\varepsilon^2)$ samples in which both coordinates i and j are present for each of the $O(d^2)$ possible coordinate pairs. According to our Assumption 1.3, this happens with probability at least α for a particular pair of coordinates. Therefore, if we draw $O(\frac{d^2\log(1/\delta)}{\alpha\varepsilon^2})$ samples, we get that for each pair of coordinates, we have the required amount of samples

with probability at least $1-\delta$. Since we need $\delta < \frac{1}{d^2}$ to apply a union bound, we have that $O(\frac{d^2 \log d}{\alpha\varepsilon^2}) = \tilde{O}(\frac{d^2}{\alpha\varepsilon^2})$ samples are sufficient for $\mathbf{\Sigma}^*$ with constant condition number.

Linear-thresholding Model For the linear thresholding model, the above reduction does not work because the problem can no longer be "decomposed" into 2-dimensional ones. The reason is that whether or not some pair of coordinates (x_i, x_j) appears can now be affected by the value of $\boldsymbol{x}$ in coordinates different than i and j. Therefore, we design a projected stochastic gradient descent (PSGD) algorithm that, given the covariance matrix $\mathbf{\Sigma}$ is known, yet arbitrary, it will give us an estimate of the true mean $\boldsymbol{\mu}^*$ of the original distribution, which can be arbitrarily close to it with the right choice of parameters. The algorithm first uses an empirical estimator for the initialization of the estimate. We show that its distance to the true mean is bounded as a function of $\mathbf{\Sigma}$ and the parameters of Assumption 1.3. Subsequently, we run a PSGD algorithm, whose projection step maintains this property. The gradient sampling step is non-trivial as a straightforward rejection sampling approach would run in exponential time. Therefore, we resort to a *Langevin Monte Carlo algorithm* which yields an approximately unbiased sample of the gradient. The projection set in this algorithm ensures that the centralized second moment of the gradient estimator is bounded, while its bias is also kept small. Combining this with our lower bound on the convexity parameter of the strongly convex likelihood function $\ell(\boldsymbol{\mu})$, we are able to show the result.

1.3 RELATED WORK

High-dimensional distribution learning (Kearns et al., 1994) initiated a systematic investigation of the computational complexity of distribution learning. Since then, there has been a large volume of works devoted to the parameter and distribution learning from a wide range of distributions in both low and high dimensions (Dasgupta, 1999; Sanjeev and Kannan, 2001; Chan et al., 2013; Ge et al., 2015; Diakonikolas et al., 2019; Bakshi et al., 2022). Broadly, this problem falls into the realm of robust statistics. Following the pioneering works by (Tukey, 1960; Huber, 1992), other recent works on high dimensional robust distribution learning can be found at (Charikar et al., 2017; Rekatsinas et al., 2017; Diakonikolas et al., 2018; Khosravi et al., 2019; Diakonikolas et al., 2020; Kane, 2021). We will be particularly interested in robustly estimating mean and covariance from high-dimensional data with partially-reliable data samples (Baranchik, 1964; Szatrowski, 1980; Stein, 1981; Boldea and Magnus, 2009; Belkin and Sinha, 2010; Pascal et al., 2013; Lai et al., 2016; Diakonikolas and Kane, 2019; Diakonikolas et al., 2019; Lei et al., 2020; Cheng et al., 2020; Cherapanamjeri et al., 2020; Hopkins et al., 2020). Settings similar to ours are studied in (Liu et al., 2021; Hu and Reingold, 2021) regarding robust mean estimation with coordinate-level corruptions. In this paper, we obtain stronger guarantees for the mean estimation, yet incomparable to (Liu et al., 2021) due to their stronger corruption model.

Learning from truncated or censored samples Distribution learning under censored, truncated mechanisms has had a long history. Censoring happens when the events can be detected, but the measurements (the values) are completely unknown, while truncation occurs when an object falling outside some subset are not observed, and their count in proportion to the observed samples is also not known, see (Deemer Jr and Votaw Jr, 1955; Cohen, 1957; Dixon, 1960; Haas and Scheff, 1990; Cohen, 1991; Barr and Sherrill, 1999; Cha et al., 2013; Charikar et al., 2017) for an overview of the related works in estimating the censored or truncated normal or other type of distributions. (Daskalakis et al., 2018, 2019, 2020) developed computationally and statistically efficient algorithms under the assumption that the truncation set is known. Furthermore, (Wu et al., 2019) considered the problem of estimating the parameters of a d-dimensional rectified Gaussian distribution from i.i.d. samples. This can be seen as a special case of the self-censoring truncation, where the truncation happens due to the ReLU generative model. (Shpitser et al., 2015) explored the identification and estimations conditions when data are missing not-at-random. While (Bhattacharya et al., 2020; Nabi et al., 2020; Malinsky et al., 2021) explored the necessary and sufficient graphical conditions to recover the full data distribution under no self-censoring condition.

Learning from general missingness More broadly, self-selection models fall under the literature of regression with MNAR in the outcomes (Rotnitzky and Robins, 1995; Rotnitzky et al., 1998; Tchetgen et al., 2018). Unlike self-censoring, this project doesn't restrict the form of the representation. Two most popular methods are the expectation-maximization algorithm (Dempster et al., 1977) and Gibbs sampling (Geman and Geman, 1984) under MAR. Despite the long history and the application of missing data models, most of the existing methods with regard to robust learning (Ramoni and Sebastiani, 2001) are consistent in the asymptotic sample regime. For example, likelihood method (Enders and Bandalos, 2001), multiple imputation (Allison, 2000), semiparametric estimation with influence function (Robins et al., 2000), inverse probability weighted complete-case estimator (Wooldridge, 2007; Seaman and White, 2013), and double/debiased machine learning (Chernozhukov et al., 2018a). See textbook (Tchetgen, 2006; Tsiatis, 2006; Van Buuren,

2018) for more introductions and further applications in this field. Recently, there are several finite sample guarantees for the double robust estimator when data are MNAR (Chernozhukov et al., 2018b, 2021) and high-dimensional (Quintas-Martinez, 2022). In addition to the works discussed, there has been significant research on detecting truncation (De et al., 2023, 2024) and estimation under unknown truncation (Kontonis et al., 2019; Diakonikolas et al., 2024).

2 NOTATIONS AND PRELIMINARIES

Throughout, let $d \geq 1$ denote the dimension of the underlying domain. For a d-dimensional vector $\mathbf{u}$ and a subset $A \subseteq [d]$, let $\mathbf{u}_A \in \mathbb{R}^{|A|}$ denote the restriction of $\mathbf{u}$ to the coordinates in A. A *missingness model* is defined by a function $\mathbb{S} : \mathbb{R}^d \to 2^{[d]}$. Given a distribution $\mathcal{D}$ on $\mathbb{R}^d$, an *observation of $\mathcal{D}$ censored by $\mathbb{S}$* is a pair $(A, \mathbf{x}) \in 2^{[d]} \times \mathbb{R}^{|A|}$, generated by first sampling $\mathbf{y} \sim \mathcal{D}$ and then setting $A = \mathbb{S}(\mathbf{y})$ and $\mathbf{x} = \mathbf{y}_A$. The interpretation is that y_i is seen for every $i \in \mathbb{S}(\mathbf{y})$ while y_i is missing for every $i \notin \mathbb{S}(\mathbf{y})$. We denote the resulting distribution on pairs by $\mathcal{D}^{\mathbb{S}}$. If the density function of $\mathcal{D}$ is f, then the density function of $\mathcal{D}^{\mathbb{S}}$ is $f^{\mathbb{S}}$ defined as:

$$f^{\mathbb{S}}(A, \mathbf{x}) = \int_{\mathbf{y} \in \mathbb{R}^d} \mathbf{1}[\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) f(\mathbf{y}) d\mathbf{y}. \quad (1)$$

Note that $\sum_{A \subseteq [d]} \int_{\mathbf{x} \in \mathbb{R}^{|A|}} f^{\mathbb{S}}(A, \mathbf{x}) = 1$, as desired. We say that $\mathbb{S}$ is a *self-censoring missingness model* if there exist sets $S_1, \ldots, S_d$ such that $\mathbb{S}(\mathbf{y}) = \{i \in [d] : y_i \in S_i\}$. We say $\mathbb{S}$ is a *linear threshold missingness model* if there exist $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^d$ and $b_1, \ldots, b_d \in \mathbb{R}$ such that $\mathbb{S}(\mathbf{y}) = \{i \in [d] : \mathbf{v}_i^T \mathbf{y} \leq b_i\}$.

Fact 2.1. *Let $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ be two $d \times d$ matrices such that $\forall i, j : |a_{ij} - b_{ij}| \leq \delta$. Then, $\|A - B\|_F \leq \delta \cdot d$.*

3 DISTRIBUTION LEARNING UNDER SELF-CENSORING MISSINGNESS

The problem of learning a distribution from truncated samples was studied in (Daskalakis et al., 2018). Their guarantee, as presented in Theorem 3.1, is given under the assumption that a fraction α of all the samples is fully observed across all dimensions.

Theorem 3.1 (adapted from (Daskalakis et al., 2018))**.** *Given oracle access to a measurable set T, whose measure under some unknown d-variate normal $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ is at least some constant $\alpha > 0$, and samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$ 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 $\varepsilon > 0$, the algorithm uses $\tilde{\mathcal{O}}(d^2/\varepsilon^2)$ truncated samples and queries to the oracle and produces estimates that satisfy the following with probability at least 99%.*

$$\|(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})\|_2 \leq \varepsilon \sqrt{\lambda_{max}}; \quad \text{and} \quad \|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\|_F \leq \varepsilon \lambda_{max}.$$

This simplifies the problem because with enough samples, part of the shape of the Gaussian distribution can be observed, allowing for simultaneous estimation of the mean and covariance. In contrast, the self-censoring missingness only allows us to observe a subset of samples, making the estimation problem more challenging. The goal is to recover $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ under minimal assumptions on the censoring mechanism. In this section, we present and analyze our algorithm for estimating the true mean and covariance of the multivariate normal distribution under self-censoring missingness.

The main idea behind our algorithm for self-censoring missingness is to use the solutions to 1-dimensional and 2-dimensional subproblems as subroutines and subsequently combine them appropriately to obtain the solution. These subproblems are either the restriction of our problem to a single coordinate or a pair of coordinates. Assumption 1.1 guarantees the existence of sufficiently many samples for these problems and allows us to use the $1D$ and $2D$ versions of Algorithm 1 in (Daskalakis et al., 2018) as our `Univariate_SGD_truncation` and `Bivariate_SGD_truncation` estimator respectively.

Even though these subroutines can give us accurate estimates for each coordinate of the true mean and the correlations between pairs of sample coordinates it is unfortunately not straightforward to provide an estimate of the $d \times d$ covariance matrix satisfying our desired guarantees. We explain below how to get around this issue.

We reconstruct the covariance matrix by only considering pairs of coordinates. For each $i \neq j$, we apply the 2-dimensional version of the algorithm in Daskalakis et al. (2018) (`Bivariate_SGD_truncation`) on the ith and j'th coordinates to obtain the 2×2-matrix $\hat{\boldsymbol{\Sigma}}^{ij}$. We will show that the $d \times d$ matrix $\hat{\Sigma}$ whose off diagonal entries ($\hat{\Sigma}_{ij}$) are given by the off diagonal entries ($\hat{\Sigma}^{ij}_{12}$) of the corresponding 2×2 matrix is a good approximation for the true Σ.

We are now ready to describe Algorithm 1 demonstrating our distribution learning approach under self-censoring missingness mechanism.

Mean Estimation We show in Lemma 3.2 the finite sample bound with a consistent mean estimation up to a bounded error of $\mathcal{O}(\varepsilon)$. The proof is deferred to the appendix.

Lemma 3.2. *Let $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ be the normal distribution with mean $\boldsymbol{\mu}^*$ and covariance matrix $\boldsymbol{\Sigma}^*$. Suppose that*

Algorithm 1: [Truncation_PSGD] Mean and covariance recovery algorithm with oracle access that generates samples with incomplete data.

Input : Data $\mathbf{x} \in \mathbb{R}^{n \times d}$, where $n = \frac{1}{\alpha \varepsilon'^2}$

1 **for** $i \leftarrow 1$ **to** d **do**

2 $\quad \hat{\mu}_i, \hat{\Sigma}_{ii} \leftarrow$ Uni_SGD_trunc$(\mathbf{x}_i, S_i)$;

3 **for** $i \leftarrow 1$ **to** $d-1$ **do**

4 $\quad$ **for** $j \leftarrow i+1$ **to** d **do**

5 $\quad\quad \hat{\mathbf{\Sigma}}^{ij}_{12} \leftarrow$ Biv_SGD_trunc$(\mathbf{x}_i, \mathbf{x}_j, S_i \times S_j)$;

6 $\hat{\boldsymbol{\mu}} \leftarrow [\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_d]$;

7 **for** $i \leftarrow 1$ **to** $d-1$ **do**

8 $\quad$ **for** $j \leftarrow i+1$ **to** d **do**

9 $\quad\quad \hat{\mathbf{\Sigma}}_{ij} \leftarrow \hat{\mathbf{\Sigma}}^{ij}_{12}; \qquad \hat{\mathbf{\Sigma}}_{ji} \leftarrow \hat{\mathbf{\Sigma}}^{ij}_{12};$

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

Assumption 1.1 holds for some constant $\alpha > 0$, and let $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_d)$ be the estimated mean from the censored Gaussian in Line 6 of Algorithm 1. For all $\varepsilon > 0$, using $\tilde{\mathcal{O}}(\frac{d^2}{\alpha \varepsilon^2})$ samples[1] we have that:

$$\forall i \in [d] : |\mu^*_i - \hat{\mu}_i| \leq (\varepsilon/d)\sigma_i \leq (\varepsilon/d)\sqrt{\lambda_{max}(\mathbf{\Sigma})}$$

where σ_i denotes the standard deviation of coordinate i (i.e $\sigma_i = \sqrt{\Sigma^*_{ii}}$, where Σ^*_{ii} is the i-th diagonal entry of the covariance matrix $\mathbf{\Sigma}^*$).

Covariance Estimation In Lemma 3.3 below, we show that if for each pair of coordinates we are given enough samples in which this particular pair is seen, we are able to obtain an accurate estimation of $\mathbf{\Sigma}^*$. In particular, we will run the 2D version of the problem for each of the $\binom{d}{2}$ pairs of coordinates and require that the estimate has accuracy ε_2. By applying Theorem 3.1 for $d = 2$, and error $\delta = \frac{1}{100\binom{d}{2}}$, we conclude that $\tilde{\mathcal{O}}(1/\varepsilon_2^2)$ samples are sufficient to achieve 99% success probability via a union bound. We will show that ε_2 doesn't need to be too small.

Lemma 3.3. *Let $\hat{\mathbf{\Sigma}}$ be the matrix with entries $\hat{\Sigma}_{ij} = \hat{\Sigma}^{ij}_{12}$, where $\hat{\Sigma}^{ij}_{12}$ denotes the value of the off diagonal entries of the 2×2 matrix $\hat{\mathbf{\Sigma}}^{\boldsymbol{ij}}$.By $\hat{\mathbf{\Sigma}}^{ij}$ we denote the estimation of a 2×2 covariance matrix that we get when we restrict the input data to coordinates i and j. Then the following holds: Using $\tilde{\mathcal{O}}(\frac{d^2}{\alpha \varepsilon^2})$ samples to get the above estimates,we have that:*

$$\|\mathbf{\Sigma}^* - \hat{\mathbf{\Sigma}}\|_F \leq \varepsilon \lambda_{max}$$

where λ_{max} is the maximum eigenvalue of Σ^*

[1] We note that the $\tilde{\mathcal{O}}_\alpha$ notation here hides both $\log d$ and $\log(1/\delta)$ factors.

Based on the above results, we summarize our main results under the self-censoring missingness mechanism in Theorem 1.2.

Theorem 1.2. *Suppose we can observe samples from $\mathcal{N}(\boldsymbol{\mu}^*, \mathbf{\Sigma}^*)$ censored through a self-censoring missingness model $\mathbb{S}$. If Assumption 1.1 is satisfied for some constant value of the parameter α, there exists a polynomial-time algorithm that recovers estimated $\boldsymbol{\mu}^*, \mathbf{\Sigma}^*$ with arbitrary accuracy. Specifically, for all $\varepsilon > 0$, and given that the eigenvalues of $\mathbf{\Sigma}^*$ lie in the interval $[\lambda_{min}, \lambda_{max}]$, the algorithm uses $\tilde{\mathcal{O}}\left(\frac{d^2(\lambda_{max}/\lambda_{min})^2}{\alpha \varepsilon^2}\right)$ samples and produces estimates that satisfy the following:*

$$\left\|\mathbf{\Sigma}^{*-1/2}(\boldsymbol{\mu}^* - \hat{\boldsymbol{\mu}})\right\|_2 \leq \mathcal{O}(\varepsilon);$$

$$\text{and } \left\|\boldsymbol{I} - \mathbf{\Sigma}^{*-1/2}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{*-1/2}\right\|_F \leq \mathcal{O}(\varepsilon).$$

Note that the sample complexity is proportional to $1/\alpha$. Furthermore, under the above conditions, we have $\mathrm{d_{TV}}(\mathcal{N}(\boldsymbol{\mu}^*, \mathbf{\Sigma}^*), \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}})) \leq \mathcal{O}(\varepsilon)$.

With the following lemma we will show a lower bound, which shows that even if the TV distance is large, in which case the distributions are easily distinguishable in the classical sampling model, the censoring model makes the distribution hard to distinguish.

Lemma 3.4. *Given $m = o(1/\sqrt{\lambda_{min}})$ censored samples according to the missingness model $\mathbb{S}$ and $\varepsilon = \Omega(\sqrt{\lambda_{min}})$. No algorithm can estimate the true mean with accuracy $O(\varepsilon)$ and probability larger than $2/3$.*

Note that, for $\varepsilon = \Omega(\sqrt{\lambda_{min}})$ the TV distance between the distributions P_λ and Q_λ is $\Omega(1)$, yet the $\Omega(1/\sqrt{\lambda_{min}})$ censored samples are necessary.

4 MEAN ESTIMATION UNDER LINEAR THRESHOLDING MISSINGNESS

In this section, we present sufficient conditions for mean estimation under linear thresholding missingness. As earlier, we let $\mathbb{S}$ denote the missingness model, and $\mathcal{N}(\boldsymbol{\mu}^*, \mathbf{\Sigma})$ denote the ground truth distribution. Our observations are drawn from $\mathcal{N}(\boldsymbol{\mu}^*, \mathbf{\Sigma})^{\mathbb{S}}$.

We will make the following two assumptions on the missingness mechanism and the ground truth distribution. Our first assumption ensures that any small subset of coordinates is observed simultaneously with non-negligible probability.

Assumption 1.3. *There exist some $\alpha, \beta > 0$ such that for any set $A \subseteq [d]$ of size at most βd,*

$$\Pr_{\mathbf{y} \sim N(\boldsymbol{\mu}^*, \mathbf{\Sigma})}[A \subseteq \mathbb{S}(\mathbf{y})] \geq \alpha.$$

Note that we only consider the case where βd is a positive integer without loss of generality.

This is a stronger version of Assumption 1.1. Our second assumption postulates existence of an "anchoring" subset.

Definition 4.1 (Anchored missingness)**.** *A subset $C \subseteq [d]$ is γ-*anchoring *if*

(i) $C \subseteq \mathbb{S}(\mathbf{y})$ for any $\mathbf{y}$, and

(ii) for any $A \subseteq [d]$, $\Pr_{\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})}[\mathbb{S}(\mathbf{y}) = A \mid \mathbf{y}_C]$ is either 0 or at least γ.

Assumption 1.4. *There exists a γ-anchoring subset C for the true distribution $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ in combination with the missingness model $\mathbb{S}$.*

Given the assumptions above, we will prove the following result showing that we can accurately and efficiently recover the mean of the distribution using censored samples:

Theorem 1.5. *For a known covariance matrix $\boldsymbol{\Sigma}$, suppose we can observe samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ censored through a linear thresholding missingness model $\mathbb{S}$. If Assumption 1.3 and Assumption 1.4 are satisfied, there exists a polynomial-time algorithm that recovers estimated $\boldsymbol{\mu}^*$ with arbitrary accuracy. Specifically, for all $\varepsilon > 0$, the algorithm uses $\mathrm{poly}(d, 1/\alpha, 1/\beta, 1/\gamma, \lambda_{\max}(\boldsymbol{\Sigma})/\lambda_{\min}(\boldsymbol{\Sigma}), 1/\varepsilon, \log(1/\delta))$ samples and running time, and with probability at least $1-\delta$, produces an estimate $\hat{\boldsymbol{\mu}}$ such that $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|_\Sigma \le \varepsilon$.*

General outline In this section, we present and analyze our mean estimation algorithm of `MissingDescent` under anchor missingness models. As a high-level overview, our approach involves running a Projected Stochastic Gradient Descent (PSGD) algorithm on a negative log-likelihood function whose optimal value coincides with the true mean. The steps of proof are as follows:

- We develop an appropriate objective function in Section 4.1 and design an efficient mean estimation algorithm `MissingDescent` in Algorithm 3, assuming that any small subset of coordinates is observed with sufficiently high probability (Assumption 1.3), and the observed missingness pattern is not very rare conditioned on the values of the observed coordinates (Assumption 1.4).
- We show that our objective function is strongly convex with respect to the correct parameterization and hence the optimum is unique. Furthermore, it is equal to the true mean.
- We analyze our `MissingDescent` algorithm in Section 4.3 by showing that Algorithm 3 approximately optimizes ℓ with bounds on the runtime and sample complexity.
- Specifically, we show in Algorithm 2 in Section 4.2 that we can use the `Initialize` algorithm to efficiently compute an initial feasible point to start the optimization.
- In the `SampleGradient` algorithm in Algorithm 5, we demonstrate that it is possible to obtain an estimate of $\Delta\ell(\boldsymbol{\mu})$ that is approximately unbiased by sampling from the conditional distribution. Additionally, we use the `ProjectToDomain` algorithm in Algorithm 6 to project a current guess back onto the domain.

4.1 NEGATIVE LOG-LIKELIHOOD OBJECTIVE FUNCTION WITH ANCHOR MISSINGNESS

We will approach the mean estimation problem via optimization of the population log-likelihood with respect to a given parameter estimate μ for the true mean μ^*. Define $g_{\boldsymbol{\mu}}$ to be the density function of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$g_{\boldsymbol{\mu}}(\mathbf{y}) = (2\pi)^{\frac{-d}{2}} |\boldsymbol{\Sigma}|^{\frac{-1}{2}} \exp\left(-\frac{(\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}-\boldsymbol{\mu})}{2}\right).$$

Recall the notation $g^{\mathbb{S}}_{\boldsymbol{\mu}}$ defined in Section 2 to be the density function of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ censored by $\mathbb{S}$. We can then write down the population negative log-likelihood ℓ as:

$$\begin{aligned}\ell(\boldsymbol{\mu}) &= \mathop{\mathbb{E}}_{(A,\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})^{\mathbb{S}}} [-\log g^{\mathbb{S}}_{\boldsymbol{\mu}}(A, \mathbf{x})] \\ &= \mathop{\mathbb{E}}_{(A,\mathbf{x})} \left[-\log \int_{\mathbf{y}} \mathbf{1}[\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) \cdot g_{\boldsymbol{\mu}}(\mathbf{y}) d\mathbf{y}\right].\end{aligned}$$

In the second equality, and everywhere later, $(A, \mathbf{x})$ is an observation sampled from the censored version of the true distribution: $\mathcal{N}(\boldsymbol{\mu}^*, \Sigma)^{\mathbb{S}}$. The integral above marginalizes over all $\mathbf{y}$ for which the missingness model would yield the observation $(A, \mathbf{x})$.

The gradient with respect to $\boldsymbol{\mu}$ of $\boldsymbol{\nabla}\ell(\boldsymbol{\mu})$ can be expressed as

$$\mathop{\mathbb{E}}_{(A,\mathbf{x})} \left[-\frac{\int_{\mathbf{y}} \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \cdot \mathbf{1}[\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) g_{\boldsymbol{\mu}}(\mathbf{y}) d\mathbf{y}}{\int_{\mathbf{y}} \mathbf{1}[\mathbb{S}(\mathbf{y}) = A] \cdot \delta(\mathbf{y}_A - \mathbf{x}) \cdot g_{\boldsymbol{\mu}}(\mathbf{y}) d\mathbf{y}}\right] \tag{2}$$

$$= -\mathop{\mathbb{E}}_{(A,\mathbf{x})} [\mathop{\mathbb{E}}_{\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)} [\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}]] \tag{3}$$

Lemma 4.2. *For any $\boldsymbol{\mu} \in \mathbb{R}^d$, it holds that: $\ell(\boldsymbol{\mu}) \ge \ell(\boldsymbol{\mu}^*)$.*

Lemma 4.3 (Strong Convexity with Missing Entries)**.** *Given our missingness model and Assumption 1.3 with $\beta = \frac{c}{d}$ for some integer $c \in \{1, \dots d\}$, we have that the function with general covariance $\ell(\boldsymbol{\mu})$ is λ-strongly convex for $\lambda = \alpha\beta/\lambda_{\max}(\boldsymbol{\Sigma})$.*

Remark. Convexity may not hold if the missignnness pattern is not linear thresholding. For example, even for $d = 1$, if $\mu^* = 0$ and $\mathbb{S}(y) = \{1\}$ if $y \in [2, 4]$ and $\emptyset$ otherwise, the function $\ell(\mu)$ is not convex.

4.2 ALGORITHM

Initialization Our first step for efficiently optimizing the negative log-likelihood function is finding a good initial point for the PSGT. Specifically, we take the empirical mean $\hat{\mu}$. This is a biased estimate, but we show below that this is good enough for initialization: the distance of the empirical estimates and true mean μ^* is a constant that depends only on the constant β, mass α and λ_{max} of the known $\boldsymbol{\Sigma}$. The pseudocode for `Initialize` appears in Algorithm 2.

Algorithm 2: [`Initialize`] Initialization for the main algorithm.

Input : Access to data generator $\mathcal{O}$, parameter $\beta = \frac{c}{d}$ for some integer $c \in \{1, \dots, d\}$, number of samples M_{init}

1 $\mathbf{w} \leftarrow$ empty array of length d
2 $\mathbf{X} \leftarrow$ matrix with M_{init} rows, each an independent sample from $\mathcal{O}$
3 **for** $i \leftarrow 0$ **to** $\lceil 1/\beta \rceil - 1$ **do**
4 $s \leftarrow i\beta d + 1$
5 $t \leftarrow \min\{(i+1)\beta d, d\}$
6 $\mathbf{Y}_i \leftarrow$ submatrix of $\mathbf{X}$ consisting of columns $s, s+1, \dots, t$
7 Remove all rows of $\mathbf{Y}_i$ containing at least one $*$
8 $\hat{\boldsymbol{\mu}}_i \leftarrow$ average of the rows of $\mathbf{Y}_i$
9 $\mathbf{w}[s, s+1, \dots, t] \leftarrow \hat{\boldsymbol{\mu}}_i$
10 **return** $\mathbf{w}$

By Assumption 1.3, we have that after line 7 in `Initialize`, each $\mathbf{Y}_i$ is the truncation of a βd-dimensional gaussian where the truncation set has mass at least α. Using Lemma 6 of (Daskalakis et al., 2018), the mean of such a truncated gaussian is $O(\sqrt{\log(1/\alpha)})$ distance away from the untruncated mean. Hence, we have $\|\mathbb{E}[\mathbf{w}] - \boldsymbol{\mu}^*\|_2^2 = \sum_i \|\mathbb{E}[\hat{\boldsymbol{\mu}}_i] - \boldsymbol{\mu}[i\beta d+1, \dots, (i+1)\beta d]\|_2^2 \le \lambda_{\max} \sum_i \|\mathbb{E}[\hat{\boldsymbol{\mu}}_i] - \boldsymbol{\mu}[i\beta d+1, \dots, (i+1)\beta d]\|_{\boldsymbol{\Sigma}}^2 \le \mathcal{O}(\frac{\lambda_{\max}}{\beta}\log(1/\alpha))$.[2] Therefore, $\|\mathbb{E}[\mathbf{w}] - \boldsymbol{\mu}^*\|_2 \le \mathcal{O}(\sqrt{\frac{\lambda_{\max}}{\beta}\log(1/\alpha)})$. Later in Section 4.3, we analyze the number of samples M_{init} required for $\|\mathbf{w} - \boldsymbol{\mu}^*\|_2$ to satisfy this bound with high probability.

[2] Define $\boldsymbol{x} = \mathbb{E}[\hat{\boldsymbol{\mu}}_i] - \boldsymbol{\mu}[i\beta d + 1, \dots, (i+1)\beta d]$, and the eigenvalue decomposition of $\boldsymbol{\Sigma}^{-1}$ as $\boldsymbol{\Sigma}^{-1} = Q^\top D^{-1} Q$. The first inequality holds because $\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}}^2 = \|\boldsymbol{x}^\top Q^\top D^{-1} Q\boldsymbol{x}\|_2 = \|D^{-1/2}Q\boldsymbol{x}\|_2^2 \ge \frac{1}{\lambda_{\max}}\|Q\boldsymbol{x}\|_2^2 = \frac{1}{\lambda_{\max}}\|\boldsymbol{x}\|_2^2$. Therefore, we have $\|\boldsymbol{x}\|_2^2 \le \lambda_{\max}\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}}^2$

Algorithm 3: [`MissingDescent`] Mean recovery algorithm given access to an oracle that generates samples with incomplete data.

Input : Access to data generator $\mathcal{O}$, parameters $\beta, \lambda_{\text{sgd}}, \eta_{\text{lmc}}, R_{\text{lmc}}, r_{\text{proj}}, M_{\text{init}}, M_{\text{sgd}}, M_{\text{grad}}$

1 $\boldsymbol{\mu}^{(0)} \leftarrow$ `Initialize`$(\mathcal{O}, \beta, M_{\text{init}})$
2 **for** $i \leftarrow 1$ **to** M_{sgd} **do**
3 Sample $(A^{(i)}, \mathbf{x}^{(i)})$ from $\mathcal{O}$
4 $\eta_i \leftarrow \frac{1}{\lambda_{\text{sgd}} \cdot i}$
5 $\mathbf{g}^{(i)} \leftarrow$ `SampleGradient`$((A^{(i)}, \mathbf{x}^{(i)}), \boldsymbol{\mu}^{(i-1)}, \eta_{\text{lmc}}, R_{\text{lmc}}, M_{\text{grad}})$
6 $\mathbf{v}^{(i)} \leftarrow \boldsymbol{\mu}^{(i-1)} - \eta_i \mathbf{g}^{(i)}$
7 $\boldsymbol{\mu}^{(i)} \leftarrow$ `ProjectToDomain`$(\boldsymbol{\mu}^{(0)}, \mathbf{v}^{(i)}, r_{\text{proj}})$
8 $\bar{\boldsymbol{\mu}} \leftarrow \frac{1}{M_{\text{sgd}}} \sum_{i=1}^{M_{\text{sgd}}} \boldsymbol{\mu}^{(i)}$
9 **return** $\bar{\boldsymbol{\mu}}$

Note that, in each iteration of SGD in `MissingDescent` (Algorithm 3), we choose a projection set, to make sure that PSGD converges. Specifically, we project a current guess back to a $\mathcal{B}_{\boldsymbol{\Sigma}}$ ball scaled by r_{proj} and centered at $\boldsymbol{\mu}^{(0)}$ as shown below:

Algorithm 4: [`ProjectToDomain`] The function that projects a current guess back to the domain onto the $\mathcal{B}_{\boldsymbol{\Sigma}}$ ball.

Input : $\boldsymbol{\mu}^{(0)}, \mathbf{v}$, parameter r_{proj}

1 **return** $\boldsymbol{\mu}^{(0)} + \min\{r_{proj}, \|(\mathbf{v} - \boldsymbol{\mu}^{(0)})\|_\Sigma\} \cdot \frac{(\mathbf{v} - \boldsymbol{\mu}^{(0)})}{(\|\mathbf{v} - \boldsymbol{\mu}^{(0)})\|_\Sigma}$

Our goal is to minimize the population negative log-likelihood ℓ via (projected) stochastic gradient descent while maintaining its strong-convexity. Specifically, Algorithm 3 above describes this strategy. In order to apply Algorithm 3 to our log-likelihood objective function, we need to solve the following three algorithmic problems:

- **Initialization**: efficiently compute an initial feasible point from which to start the optimization. The pseudocode for `Initialize` appears in Algorithm 2;
- **Gradient estimation**: design a nearly unbiased sampler for $\nabla\ell(\boldsymbol{\mu})$ using Langevin sampling. The `SampleGradient` pseudocode appears in Algorithm 5;

- **Efficient projection**: perform an efficient projection into a set of feasible points to make sure that PSGD converges. The pseudocode presents in Algorithm 6.

Algorithm 5: [SampleGradient] Sampler for $\nabla \ell(\boldsymbol{\mu})$.

Input : $(A, \mathbf{x})$, $\boldsymbol{\mu}$, parameters η, R, M

$a \leftarrow |A|$

Compute $\boldsymbol{\mu}_{\text{cond}}$ and $\boldsymbol{\Sigma}_{\text{cond}}$ as in (15) and (16)

Let W be such that $\boldsymbol{\Sigma}_{\text{cond}} = WW^\top$

Compute $\mathcal{L} = (W^{-1}\mathcal{K}) \cap \mathcal{B}_\Sigma(W^{-1}\boldsymbol{\mu}_{\text{cond}}, R)$

$\mathbf{z}^{(0)} \leftarrow$ a point in $\mathcal{L}$

for $t = 0$ **to** $M - 1$ **do**

  Sample $\boldsymbol{\zeta}^{(t)}$ from $\mathcal{N}(0, I_{d-a})$

  $\mathbf{z}^{(t+1)} \leftarrow$

  $\Pi_{\mathcal{L}}\left(\mathbf{z}^{(t)} - \eta(\mathbf{z}^{(t)} - W^{-1}\boldsymbol{\mu}_{\text{cond}}) + \sqrt{\eta} \cdot \boldsymbol{\zeta}^{(t)}\right)$

return $-\boldsymbol{\Sigma}^{-1}(\mathbf{x} \circ (W\mathbf{z}^{(M)}) - \boldsymbol{\mu})$

4.3 ANALYSIS OF MissingDescent

We show in this section that Algorithm 3 approximately optimizes ℓ with bounds on the runtime and sample complexity. The following lemma describes the ingredients necessary to obtain such bounds:

Lemma 4.4 (Lemma 6 in Cherapanamjeri et al. (2022)). *Let $f : \mathbb{R}^k \to \mathbb{R}$ be a convex function, $K \subseteq \mathbb{R}^k$ a convex set, and fix an initial estimate $\mathbf{x}^{(0)} \in K$. Now, let $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}$ be the iterates generated by running T steps of projected SGD using gradient estimates $\mathbf{g}^{(1)}, \ldots, \mathbf{g}^{(T)}$ satisfying $\mathbb{E}[\mathbf{g}^{(i)} \mid \mathbf{x}^{(i-1)}] = \nabla f(\mathbf{x}^{(i-1)}) + \mathbf{b}^{(i)}$. Let $\mathbf{x}_* = \arg\min_{\mathbf{x} \in K} f(\mathbf{x})$ be a minimizer of f. Then, if we assume:*

(i) ***Bounded step variance****: $\mathbb{E}[\|\mathbf{g}^{(i)}\|_2^2] \le \rho^2$,*

(ii) ***Strong convexity****: f is λ-strongly convex, and*

(iii) ***Bounded gradient bias****: $\|\mathbf{b}^{(i)}\|^2 \le \frac{\rho^2}{2\lambda \cdot diam(K) \cdot i}$,*

then the average iterate $\widehat{\mathbf{x}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}^{(T)}$ satisfies $\mathbb{E}[f(\widehat{\mathbf{x}}) - f(\mathbf{x}_*)] \le \frac{\rho^2}{\lambda T}(1 + \log(T))$.

We study each of the three conditions in Lemma 4.4 above, before wrapping up with the overall analysis.

4.3.1 STRONG CONVEXITY

To show convergence of stochastic gradient descent on ℓ, we require *strong convexity* such that the optimum of ℓ is unique. Specifically, we need to show: $\boldsymbol{\nabla}^2 \ell(\boldsymbol{\mu}) \succeq \beta I$ for some parameter $\beta > 0$ such that the probability mass is at least a constant. Once we proved the strong convexity, we can apply projected stochastic gradient descent (PSGD) to recover the parameter μ.

Lemma 4.5 (Strong Convexity with Missing Entries)**.** *Given our missingness model and Assumption 1.3 with $\beta = \frac{c}{d}$ for some integer $c \in \{1, \ldots d\}$, we have that the function with general covariance $\ell(\boldsymbol{\mu})$ is λ-strongly convex for $\lambda = \alpha\beta/\lambda_{\max}(\boldsymbol{\Sigma})$.*

4.3.2 BOUNDED STEP VARIANCE AND GRADIENT BIAS

In this section, we analyze Algorithm 5, SampleGradient with an illustration of the relationship between convex sets appeared in this section is available in Fig. 2. We first study the distribution of $\mathbf{z}^{(M)}$, which then will allow us to show an additive approximation guarantee for the output of the algorithm.

Theorem 4.6. *Assume $\|\boldsymbol{\mu}^* - \boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} \le S$. For $R = \tilde{O}(\sqrt{d} + S + \log(1/\gamma\varepsilon))$, if $M = \text{poly}\,(d, S, 1/\gamma, 1/\varepsilon)$ and $\eta = \tilde{\Theta}(R^2/M)$, then*

$$\mathrm{d_{TV}}(\mathbf{z}^{(M)}, \mathcal{N}(W^{-1}\boldsymbol{\mu}_{cond}, I)) \le \varepsilon.$$

Fix $(A, \mathbf{x})$. Without loss of generality, assume $\boldsymbol{\mu}^* = \mathbf{0}$.

Corollary 4.7. *Let $\hat{\mathbf{g}}$ be the output of Algorithm 5 with inputs $\tilde{\mathbf{x}}$ and $\boldsymbol{\mu}$ and parameters R, M, η as in Theorem 4.6. Also, let $\mathbf{g} = -\mathbb{E}_{\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}]$. Then, we have that:*

$$\|\mathbb{E}[\hat{\mathbf{g}}] - \mathbf{g}\|_2 \le \varepsilon \cdot \text{poly}\,(S, d, 1/\gamma, 1/\varepsilon, \lambda_{\max}, 1/\lambda_{\min}) \quad (4)$$

Furthermore, we have the following bound

$$\mathbb{E}[\|\hat{\mathbf{g}}\|_2^2] \le \text{poly}\,(d, 1/\gamma, S, 1/\lambda_{\min}) \quad (5)$$

5 DISCUSSION AND FUTURE WORK

In the context of linear-thresholding missingness with a known covariance matrix, we can obtain the mean by initially observing that the set $\mathbf{y} : \mathbb{S}(\mathbf{y}) = A \wedge \mathbf{y}_A = \mathbf{x}$ is convex for any set A and any $\mathbf{x} \in \mathbb{R}^{|A|}$. By leveraging the fact (Corollary 2.1 of Kanter and Proppe (1977)) that the variance of a Gaussian decreases when conditioned on a convex set, we can establish that the Hessian of our likelihood function is positive definite. This property ensures that our objective function is strongly convex and thus we can learn the distribution from a MNAR model. However, in scenarios where the covariance matrix is unknown, recovering the distribution becomes much more challenging as the Hessian of our likelihood function incorporates a fourth moment. Thus, we leave this as our future work.

References

Adak, M. F., Lieberzeit, P., Jarujamrus, P., and Yumusak, N. (2020). Classification of alcohols obtained by qcm sensors with different characteristics using abc based neural network. *Engineering Science and Technology, an International Journal*, 23(3):463–469. 2

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological methods & research*, 28(3):301–309. 4

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica: Journal of the Econometric Society*, pages 997–1016. 1

Armitage, P., Berry, G., and Matthews, J. N. S. (2008). *Statistical methods in medical research*. John Wiley & Sons. 1

Bakshi, A., Diakonikolas, I., Jia, H., Kane, D. M., Kothari, P. K., and Vempala, S. S. (2022). Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1234–1247. 4

Balakrishnan, N. and Cramer, E. (2014). The art of progressive censoring. *Statistics for industry and technology*. 1

Baranchik, A. J. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution. Technical report, STANFORD UNIV CALIF. 4

Barr, D. R. and Sherrill, E. T. (1999). Mean and variance of truncated normal distributions. *The American Statistician*, 53(4):357–361. 4

Belkin, M. and Sinha, K. (2010). Polynomial learning of distribution families. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 103–112. IEEE. 4

Bhattacharya, R., Nabi, R., Shpitser, I., and Robins, J. M. (2020). Identification in missing data models represented by directed acyclic graphs. In *Uncertainty in Artificial Intelligence*, pages 1149–1158. PMLR. 4

Boldea, O. and Magnus, J. R. (2009). Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association*, 104(488):1539–1549. 4

Breen, R. et al. (1996). *Regression models: Censored, sample selected, or truncated data*, volume 111. Sage. 1

Brick, J. M. and Kalton, G. (1996). Handling missing data in survey research. *Statistical methods in medical research*, 5(3):215–238. 1

Bubeck, S., Eldan, R., and Lehec, J. (2018). Sampling from a log-concave distribution with projected langevin monte carlo. *Discrete & Computational Geometry*, 59(4):757–783. 20, 22

Carreras, G., Miccinesi, G., Wilcock, A., Preston, N., Nieboer, D., Deliens, L., Groenvold, M., Lunder, U., van der Heide, A., and Baccini, M. (2021). Missing not at random in end of life care studies: multiple imputation and sensitivity analysis on data from the action study. *BMC medical research methodology*, 21(1):1–12. 2

Cha, J., Cho, B. R., and Sharp, J. L. (2013). Rethinking the truncated normal distribution. *International Journal of Experimental Design and Process Optimisation*, 3(4):327–363. 4

Chan, S.-O., Diakonikolas, I., Sun, X., and Servedio, R. A. (2013). Learning mixtures of structured distributions over discrete domains. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1380–1394. SIAM. 4

Charikar, M., Steinhardt, J., and Valiant, G. (2017). Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60. 4

Cheng, Y., Diakonikolas, I., Ge, R., and Soltanolkotabi, M. (2020). High-dimensional robust mean estimation via gradient descent. In *International Conference on Machine Learning*, pages 1768–1778. PMLR. 4

Cherapanamjeri, Y., Daskalakis, C., Ilyas, A., and Zampetakis, M. (2022). What makes a good fisherman? linear regression under self-selection bias. *arXiv preprint arXiv:2205.03246*. 9, 21, 22, 23

Cherapanamjeri, Y., Daskalakis, C., Ilyas, A., and Zampetakis, M. (2023). What makes a good fisherman? linear regression under self-selection bias. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1699–1712. 2

Cherapanamjeri, Y., Mohanty, S., and Yau, M. (2020). List decodable mean estimation in nearly linear time. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 141–148. IEEE. 4

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. 4

Chernozhukov, V., Newey, W., and Singh, R. (2018b). De-biased machine learning of global and local parameters using regularized riesz representers. *arXiv preprint arXiv:1802.08667*. 5

Chernozhukov, V., Newey, W. K., and Singh, R. (2021). A simple and general debiased machine learning the-

orem with finite sample guarantees. *arXiv preprint arXiv:2105.15197*. 5

Cohen, A. C. (1957). On the solution of estimating equations for truncated and censored samples from normal populations. *Biometrika*, 44(1/2):225–236. 4

Cohen, A. C. (1991). *Truncated and censored samples: theory and applications*. CRC press. 4

Dasgupta, S. (1999). Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE. 4

Daskalakis, C., Gouleakis, T., Tzamos, C., and Zampetakis, M. (2018). Efficient statistics, in high dimensions, from truncated samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–649. IEEE. 1, 2, 3, 4, 5, 8, 22

Daskalakis, C., Gouleakis, T., Tzamos, C., and Zampetakis, M. (2019). Computationally and statistically efficient truncated regression. In *Conference on Learning Theory*, pages 955–960. PMLR. 1, 4

Daskalakis, C., Kontonis, V., Tzamos, C., and Zampetakis, E. (2021a). A statistical taylor theorem and extrapolation of truncated densities. In *Conference on Learning Theory*, pages 1395–1398. PMLR. 1

Daskalakis, C., Rohatgi, D., and Zampetakis, E. (2020). Truncated linear regression in high dimensions. *Advances in Neural Information Processing Systems*, 33:10338–10347. 1, 4

Daskalakis, C., Stefanou, P., Yao, R., and Zampetakis, E. (2021b). Efficient truncated linear regression with unknown noise variance. *Advances in Neural Information Processing Systems*, 34. 1

De, A., Li, H., Nadimpalli, S., and Servedio, R. A. (2024). Detecting low-degree truncation. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1027–1038. 5

De, A., Nadimpalli, S., and Servedio, R. A. (2023). Testing convex truncation. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4050–4082. SIAM. 5

Deemer Jr, W. L. and Votaw Jr, D. F. (1955). Estimation of parameters of truncated or censored exponential distributions. *The Annals of Mathematical Statistics*, 26(3):498–504. 4

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22. 4

Diakonikolas, I., Hopkins, S. B., Kane, D., and Karmalkar, S. (2020). Robustly learning any clusterable mixture of gaussians. *arXiv preprint arXiv:2005.06417*. 4

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2019). Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864. 4

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2018). Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. SIAM. 4

Diakonikolas, I. and Kane, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*. 4

Diakonikolas, I., Kane, D. M., Pittas, T., and Zarifis, N. (2024). Statistical query lower bounds for learning truncated gaussians. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1336–1363. PMLR. 5

Dixon, W. J. (1960). Simplified estimation from censored normal samples. *The Annals of Mathematical Statistics*, pages 385–391. 4

Enders, C. K. and Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural equation modeling*, 8(3):430–457. 4

Fisher, R. (1931). Properties and applications of hh functions. *Mathematical tables*, 1:815–852. 1, 2

Fotakis, D., Kalavasis, A., and Tzamos, C. (2020). Efficient parameter estimation of truncated boolean product distributions. In *Conference on Learning Theory*. 1

Galton, F. (1898). 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. 1, 2

Ge, R., Huang, Q., and Kakade, S. M. (2015). Learning mixtures of gaussians in high dimensions. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 761–770. 4

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741. 4

Haas, C. N. and Scheff, P. A. (1990). Estimation of averages in truncated samples. *Environmental science & technology*, 24(6):912–919. 4

Hajivassiliou, V. A. and McFadden, D. L. (1998). The method of simulated scores for the estimation of ldv models. *Econometrica*, pages 863–896. 1

Hausman, J. A. and Wise, D. A. (1977). Social experimentation, truncated distributions, and efficient estimation. *Econometrica: Journal of the Econometric Society*, pages 919–938. 1

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161. 1

Honaker, J. and King, G. (2010). What to do about missing values in time-series cross-section data. *American journal of political science*, 54(2):561–581. 1

Hopkins, S., Li, J., and Zhang, F. (2020). Robust and heavy-tailed mean estimation made simple, via regret minimization. *Advances in Neural Information Processing Systems*, 33:11902–11912. 4

Hu, L. and Reingold, O. (2021). Robust mean estimation on highly incomplete data with arbitrary outliers. In *International Conference on Artificial Intelligence and Statistics*, pages 1558–1566. PMLR. 4

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer. 4

Kane, D. M. (2021). Robust learning of mixtures of gaussians. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1246–1258. SIAM. 4

Kanter, M. and Proppe, H. (1977). Reduction of variance for gaussian densities via restriction to convex sets. *Journal of Multivariate Analysis*, 7(1):74–81. 9, 19

Kearns, M., Mansour, Y., Ron, D., Rubinfeld, R., Schapire, R. E., and Sellie, L. (1994). On the learnability of discrete distributions. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 273–282. 4

Khosravi, P., Liang, Y., Choi, Y., and Broeck, G. V. d. (2019). What to expect of classifiers? reasoning about logistic regression with missing features. *arXiv preprint arXiv:1903.01620*. 4

Kontonis, V., Tzamos, C., and Zampetakis, M. (2019). Efficient truncated statistics with unknown truncation. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1578–1595. IEEE. 1, 5

Lai, K. A., Rao, A. B., and Vempala, S. (2016). Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE. 4

Lee, A. (1914). Table of the gaussian" tail" functions; when the" tail" is larger than the body. *Biometrika*, 10(2/3):208–214. 1

Lei, Z., Luh, K., Venkat, P., and Zhang, F. (2020). A fast spectral algorithm for mean estimation with subgaussian rates. In *Conference on Learning Theory*, pages 2598–2612. PMLR. 4

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360. 2

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons. 1, 2

Liu, Z., Park, J. H., Rekatsinas, T., and Tzamos, C. (2021). On robust mean estimation under coordinate-level corruption. In *International Conference on Machine Learning*, pages 6914–6924. PMLR. 4

Maddala, G. S. (1986). *Limited-dependent and qualitative variables in econometrics*. Number 3. Cambridge university press. 1

Malinsky, D., Shpitser, I., and Tchetgen Tchetgen, E. J. (2021). Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association*, pages 1–9. 4

Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. *Advances in neural information processing systems*, 26. 1

Nabi, R., Bhattacharya, R., and Shpitser, I. (2020). Full law identification in graphical models of missing data: Completeness results. In *International Conference on Machine Learning*, pages 7153–7163. PMLR. 4

Pascal, F., Bombrun, L., Tourneret, J.-Y., and Berthoumieu, Y. (2013). Parameter estimation for multivariate generalized gaussian distributions. *IEEE Transactions on Signal Processing*, 61(23):5960–5971. 4

Pearson, K. (1902). On the systematic fitting of curves to observations and measurements. *Biometrika*, 1(3):265–303. 1, 2

Pearson, K. and Lee, A. (1908). On the generalised probable error in multiple normal correlation. *Biometrika*, 6(1):59–68. 1, 2

Plevrakis, O. (2021). Learning from censored and dependent data: The case of linear dynamics. In *Conference on Learning Theory*, pages 3771–3787. PMLR. 1

Quintas-Martinez, V. (2022). Finite-sample guarantees for high-dimensional dml. *arXiv preprint arXiv:2206.07386*. 5

Ramoni, M. and Sebastiani, P. (2001). Robust learning with missing data. *Machine Learning*, 45:147–170. 4

Rekatsinas, T., Chu, X., Ilyas, I. F., and Ré, C. (2017). Holoclean: Holistic data repairs with probabilistic inference. *arXiv preprint arXiv:1702.00820*. 4

Robins, J. M. and Gill, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in medicine*, 16(1):39–56. 2

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. *IMA VOLUMES IN MATHEMATICS AND ITS APPLICATIONS*, 116:1–94. 4

Rotnitzky, A. and Robins, J. (1997). Analysis of semiparametric regression models with non-ignorable nonresponse. *Statistics in medicine*, 16(1):81–102. 2

Rotnitzky, A. and Robins, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82(4):805–820. 4

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, 93(444):1321–1339. 4

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford economic papers*, 3(2):135–146. 2

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592. 2

Sanjeev, A. and Kannan, R. (2001). Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. 4

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120. 2

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295. 4

Shpitser, I., Mohan, K., and Pearl, J. (2015). Missing data as a causal and probabilistic problem. Technical report, CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE. 2, 4

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151. 4

Szatrowski, T. H. (1980). Necessary and sufficient conditions for explicit solutions in the multivariate normal estimation problem for patterned means and covariances. *The Annals of Statistics*, pages 802–810. 4

Tchetgen, E. J. T. (2006). *Statistical methods for robust inference in causal and missing data models*. Harvard University. 4

Tchetgen, E. J. T., Wang, L., and Sun, B. (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *Statistica Sinica*, 28(4):2069. 4

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36. 1

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525. 1

Tsiatis, A. A. (2006). Semiparametric theory and missing data. 2, 4

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485. 4

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press. 4

Warga, A. (1992). Bond returns, liquidity, and missing data. *Journal of Financial and Quantitative Analysis*, 27(4):605–617. 1

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, 141(2):1281–1301. 4

Wu, S., Dimakis, A. G., and Sanghavi, S. (2019). Learning distributions generated by one-layer relu networks. *Advances in neural information processing systems*, 32. 4

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A DEFERRED PROOFS FROM Section 2

Below we provide the proof of Fact Fact 2.1 for completeness.

Fact 2.1. *Let $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ be two $d \times d$ matrices such that $\forall i, j : |a_{ij} - b_{ij}| \leq \delta$. Then, $\|A - B\|_F \leq \delta \cdot d$.*

Proof. By definition of the Frobenious norm we have:

$$\|A - B\|_F^2 = \sum_{i=1}^{d} \sum_{j=1}^{d} (a_{ij} - b_{ij})^2 \leq d^2 \delta^2$$

Thus,

$$\|A - B\|_F \leq \delta \cdot d$$

□

B DEFERRED PROOFS FROM Section 3

This section provides the formal proofs that were deferred in favor of readability. For convenience, we will restate the statements before proving them.

Lemma 3.2. *Let $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ be the normal distribution with mean $\boldsymbol{\mu}^*$ and covariance matrix $\boldsymbol{\Sigma}^*$. Suppose that Assumption 1.1 holds for some constant $\alpha > 0$, and let $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_d)$ be the estimated mean from the censored Gaussian in Line 6 of Algorithm 1. For all $\varepsilon > 0$, using $\tilde{\mathcal{O}}(\frac{d^2}{\alpha \varepsilon^2})$ samples[3] we have that:*

$$\forall i \in [d] : |\mu_i^* - \hat{\mu}_i| \leq (\varepsilon/d)\sigma_i \leq (\varepsilon/d)\sqrt{\lambda_{max}(\boldsymbol{\Sigma})}$$

where σ_i denotes the standard deviation of coordinate i (i.e $\sigma_i = \sqrt{\Sigma_{ii}^*}$, where Σ_{ii}^* is the i-th diagonal entry of the covariance matrix $\boldsymbol{\Sigma}^*$).

Proof. Fix a coordinate $i \in [d]$. If coordinate i appears in a censored sample, then the value would follow the distribution $\mathcal{N}(\mu_i^*, \Sigma_{ii}^*) = \mathcal{N}(\mu_i^*, \sigma_i^2)$. In order to apply Theorem 3.1 for $d = 1$ and $\varepsilon' = \varepsilon/d$, we need coordinate i to be present in at least $\tilde{O}(1/\varepsilon^2)$ censored samples for every $i \in [d]$. Assumption 1.1 implies that coordinate i is present in each censored sample with probability at least α. Since in every batch of $\mathcal{O}(1/\alpha(\varepsilon')^2)$ samples, there is a constant probability that the required number of $\mathcal{O}(1/(\varepsilon')^2)$ appearances of coordinate i is met, the error probability can be reduced to $1/d^2$ a the cost of an extra log factor in the sample complexity. Therefore, by union bound over all d coordinates, the statement holds with probability at least $1 - 1/d$ using $\mathcal{O}(\log d/\alpha(\varepsilon')^2) = \tilde{\mathcal{O}}(\frac{d^2}{\alpha \varepsilon^2})$ samples. □

Lemma 3.3. *Let $\hat{\boldsymbol{\Sigma}}$ be the matrix with entries $\hat{\Sigma}_{ij} = \hat{\Sigma}_{12}^{ij}$, where $\hat{\Sigma}_{12}^{ij}$ denotes the value of the off diagonal entries of the 2×2 matrix $\hat{\boldsymbol{\Sigma}}^{\mathbf{ij}}$.By $\hat{\boldsymbol{\Sigma}}^{ij}$ we denote the estimation of a 2×2 covariance matrix that we get when we restrict the input data to coordinates i and j. Then the following holds: Using $\tilde{\mathcal{O}}(\frac{d^2}{\alpha \varepsilon^2})$ samples to get the above estimates,we have that:*

$$\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\|_F \leq \varepsilon \lambda_{max}$$

where λ_{max} is the maximum eigenvalue of Σ^*

Proof. Consider any pair of coordinates $i, j \in [d]$. Given the sample size of $\frac{1}{\alpha \varepsilon'^2}$, Assumption 1.1 and Fact 2.1, there will be at least $\frac{1}{\varepsilon'^2}$ samples with non-censored entries in both coordinates i, j for any such pair. Therefore, we can apply Theorem 3.1 for $d = 2$ and $\varepsilon' = \varepsilon/d$ to get that:

$$\|\boldsymbol{\Sigma}^{*(ij)} - \hat{\boldsymbol{\Sigma}}^{(\mathbf{ij})}\|_F \leq \varepsilon' \lambda_{max}\left(\boldsymbol{\Sigma}^{*(ij)}\right) = \varepsilon \lambda_{max}\left(\boldsymbol{\Sigma}^{*(ij)}\right)/d \leq \varepsilon \lambda_{max}\left(\boldsymbol{\Sigma}^*\right)/d$$

Therefore, all the entries of the 2×2 matrix on the lhs have absolute value at most ε/d and thus this is also an upper bound on the maximum difference of corresponding off diagonal entries of the $d \times d$ matrices $\boldsymbol{\Sigma}^*$ and $\hat{\boldsymbol{\Sigma}}$ as constructed by Algorithm 1 (see line 9). The same upper bound holds for the diagonal entries, since they are more accurately estimated using 1d subproblems (line 2 of Algorithm 1). Therefore, the maximum entry-wise difference overall between the $d \times d$ matrices $\boldsymbol{\Sigma}^*$ and $\hat{\boldsymbol{\Sigma}}$ as constructed by Algorithm 1 is $\varepsilon \lambda_{max}/d$. We can now apply Fact 2.1 to get $\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\|_F \leq \varepsilon \lambda_{max}$. □

[3] We note that the $\tilde{\mathcal{O}}_\alpha$ notation here hides both $\log d$ and $\log(1/\delta)$ factors.

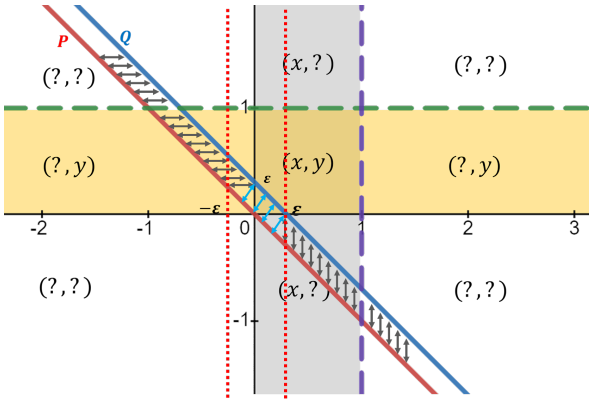

Figure 1: In this example, the self-censoring missingness mechanism is as follows: Each coordinate x, y of the sample is seen if and only if $x \in [0, 1]$ or $y \in [0, 1]$ respectively.

Lower bound on the sample complexity We now explain why some dependence on the eigenvalues of $\mathbf{\Sigma}^*$ is necessary. To see this, consider the case where $\lambda_{\min}(\mathbf{\Sigma}^*) = 0$. In this case, all the samples come from a subspace of dimension at most $d - 1$. Consider one such subspace and its translation, by an infinitesimal amount ε, along the eigenvector corresponding to the 0 eigenvalue. It is not hard to see that if we take two such parallel subspaces arbitrarily close to each other, then with high probability no finite amount of censored samples can distinguish between these 2 cases, which would be required for keeping this Mahalanobis error bounded.

An illustrative example is in Fig. 1: The censoring mechanism is as follows:

$$P^{\mathbb{S}} : \mathcal{N}(\mathbf{0}, \mathbf{I})|_{x+y=0}$$
$$Q^{\mathbb{S}} : \mathcal{N}(\mathbf{0}, \mathbf{I})|_{x+y=\varepsilon}$$

We will now present the following lemma, which formally justifies the above remark. It also acts as a warm-up for our subsequent lower bound.

Lemma B.1. *For any sufficiently small value of $\varepsilon > 0$, given $m = o(1/\varepsilon)$ censored samples according to the missingness model $\mathbb{S}$ from either $P^{\mathbb{S}}$ or $Q^{\mathbb{S}}$, no algorithm can distinguish whether the samples are coming from $P^{\mathbb{S}}$ or $Q^{\mathbb{S}}$ with probability larger than $2/3$.*

Proof. We will define a coupling between the distributions $P^{\mathbb{S}}$ and $Q^{\mathbb{S}}$ such that with probability at least $1 - O(\varepsilon)$, the same missingness pattern and the same values for the seen coordinates are observed. This implies our lower bound of $\Omega(1/\varepsilon)$ samples for distinguishing the two distributions with constant probability.

More specifically, we define the following coupling matching the probability mass that points on the line $\{x+y = 0\}$ have due to $P^{\mathbb{S}}$ to the mass that $Q^{\mathbb{S}}$ imposes on the line $\{x + y = \varepsilon\}$:

$$(x, -x) \leftrightarrow \begin{cases} (x + \varepsilon, -x) & x \le -\varepsilon & \text{observed: } (?, -x) \\ (x + \varepsilon/2, -x + \varepsilon/2) & -\varepsilon/2 \le x \le \varepsilon/2 & \\ (x, -x + \varepsilon) & x \ge \varepsilon & \text{observed: } (x, ?) \end{cases} \tag{6}$$

For the segments of the support of $P^{\mathbb{S}}$ unaccounted for in the above equation, we match their mass arbitrarily in a valid way to finish the coupling. Note that for the first and third branch of Eq. (6), the point $(x, -x)$ has strictly larger probability density from $P^{\mathbb{S}}$ than $(x + \varepsilon, -x)$ and $(x, -x + \varepsilon)$ respectively have from $Q^{\mathbb{S}}$, while for any $x \in \mathbb{R}$, the point $(x + \varepsilon/2, -x + \varepsilon/2)$ has the same density in $Q^{\mathbb{S}}$ as $(x, -x)$ has in $P^{\mathbb{S}}$.

To see this, consider the coordinate system: $\left\{w := \frac{x-y}{\sqrt{2}}, z := \frac{x+y}{\sqrt{2}}\right\}$ (which is a rotation of the original one by $\pi/4$ rad). Since the distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is isotropic, can now write the distributions $P^{\mathbb{S}}$ and $Q^{\mathbb{S}}$ as follows:

$$P^{\mathbb{S}} : \mathcal{N}(\mathbf{0}, \mathbf{I})|_{z=0}$$
$$Q^{\mathbb{S}} : \mathcal{N}(\mathbf{0}, \mathbf{I})|_{z=\varepsilon/\sqrt{2}}$$

Due to the fact that the marginals over w and z are independent, we have that $w \sim \mathcal{N}(0,1)$ in both of the cases above.

Given the above coupling, it follows that whenever the true sample has first coordinate in the set $(-\infty, -\varepsilon) \cup (\varepsilon, +\infty)$ the observed sample would be exactly the same (see Eq. (6)) either with $P^{\mathbb{S}}$ or $Q^{\mathbb{S}}$.

Therefore, with probability at least $1 - 2\,\mathrm{erf}(\varepsilon\sqrt{2}) \geq 1 - 2\sqrt{2}\,\mathrm{erf}(\varepsilon)$, the censored sample that we get will not give us any information for our task distinguishing $P^{\mathbb{S}}$ from $Q^{\mathbb{S}}$. Thus, any algorithm that is able to distinguish $P^{\mathbb{S}}$ from $Q^{\mathbb{S}}$ need to draw $\Omega(\frac{1}{\mathrm{erf}(\varepsilon)})$ samples. This is $\Omega(1/\varepsilon)$ samples as ε gets arbitrarily close to 0. □

Lemma 3.4. *Given $m = o(1/\sqrt{\lambda_{min}})$ censored samples according to the missingness model $\mathbb{S}$ and $\varepsilon = \Omega(\sqrt{\lambda_{min}})$. No algorithm can estimate the true mean with accuracy $O(\varepsilon)$ and probability larger than $2/3$.*

Proof. We now consider two families of distributions parameterized by $\lambda \in [0,1]$. We define the two families in the rotated (z,w) coordinate system, as in Lemma B.1, as follows:

$$\begin{aligned} P_\lambda &: (z,w) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \lambda & 0 \\ 0 & 1 \end{bmatrix}\right) \\ Q_\lambda &: (z,w) \sim \mathcal{N}\left(\begin{bmatrix} \varepsilon \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda & 0 \\ 0 & 1 \end{bmatrix}\right) \end{aligned} \tag{7}$$

Note that the total variation distance between P_λ and Q_λ is:

$$\begin{aligned} d_{TV}(P_\lambda, Q_\lambda) &= d_{TV}(P_\lambda|_{w=0}, Q_\lambda|_{w=0}) \\ &= 2\Phi\left(\frac{\varepsilon}{2\sqrt{\lambda}}\right) - 1 = \mathrm{erf}\left(\frac{\varepsilon}{2\sqrt{2}\sqrt{\lambda}}\right) = O\left(\frac{\varepsilon}{\sqrt{\lambda}}\right). \end{aligned} \tag{8}$$

Consider the distributions P_λ and Q_λ defined in Eq. (7) for $\lambda = \lambda_{min}$. The main idea of the proof is that we can apply Lemma B.1 for the case where $\varepsilon = \Theta(\sqrt{\lambda})$. Note that the distance between the means of these two distributions is $\varepsilon/\sqrt{2}$. Thus, any algorithm that can estimate the mean with accuracy at most $\varepsilon\sqrt{2}/4$, should be able to distinguish them.

We will use the same coupling as in Eq. (6) between the two distributions P_λ and Q_λ and use it to bound the probability that we will observe different censored samples p_c and q_c respectively. We observe that (similarly to the setting of Lemma B.1) any sample from P_λ falling outside the band: $B = \{(x,y) : -\varepsilon \leq x \leq \varepsilon\}$, has an identical censored sample to the censored sample of the corresponding point in Q_λ via the coupling.

We now upper bound the probability that sample from P_λ falls in the band B:

$$\Pr_{(x,y)\sim P_\lambda}[-\varepsilon \leq x \leq \varepsilon] \leq \Pr_{(z,w)\sim P_\lambda}\left[-\varepsilon\sqrt{2} \leq w \leq \varepsilon\sqrt{2}\right] = O(\varepsilon) \tag{9}$$

Thus, we have:

$$\Pr[p_c \neq q_c] \leq \Pr_{(x,y)\sim P_\lambda}[-\varepsilon \leq x \leq \varepsilon] = O(\varepsilon) \tag{10}$$

In addition, due to Eq. (8), there exists a different coupling for which the following holds:

$$\Pr[p_c \neq q_c] \leq \Pr_{p\sim P_\lambda, q\sim Q_\lambda}[p \neq q] = O\left(\frac{\varepsilon}{\sqrt{\lambda}}\right). \tag{11}$$

By Eq. (10) and Eq. (11), we get that no algorithm with $o(\max\{1/\varepsilon, \sqrt{\lambda}/\varepsilon\})$ samples can distinguish P_λ from Q_λ with probability at least $2/3$.

This implies the statement.

□

Theorem 1.2. *Suppose we can observe samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ censored through a self-censoring missingness model $\mathbb{S}$. If Assumption 1.1 is satisfied for some constant value of the parameter α, there exists a polynomial-time algorithm that recovers estimated $\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*$ with arbitrary accuracy. Specifically, for all $\varepsilon > 0$, and given that the eigenvalues of $\boldsymbol{\Sigma}^*$ lie in the interval $[\lambda_{min}, \lambda_{max}]$, the algorithm uses $\tilde{\mathcal{O}}\left(\frac{d^2(\lambda_{max}/\lambda_{min})^2}{\alpha\varepsilon^2}\right)$ samples and produces estimates that satisfy the following:*

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

Note that the sample complexity is proportional to $1/\alpha$. Furthermore, under the above conditions, we have $\mathrm{d_{TV}}(\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})) \leq \mathcal{O}(\varepsilon)$.

Proof of Theorem 1.2. By Lemma 3.2 and Lemma 3.3, we conclude that we can use $\tilde{O}(\frac{d^2}{\alpha\varepsilon^2})$ samples to get the following guarantees:

$$\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2 \leq \varepsilon\sqrt{\lambda_{max}}$$

and

$$\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\|_F \leq \varepsilon\lambda_{max}$$

Thus,we get the following:

$$\left\|\boldsymbol{\Sigma}^{*-1/2}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})\right\|_2 \leq \frac{1}{\sqrt{\lambda_{min}}}\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2 \leq \varepsilon\sqrt{\lambda_{max}/\lambda_{min}}$$

$$\begin{aligned}\left\|\boldsymbol{I} - \boldsymbol{\Sigma}^{*-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{*-1/2}\right\|_F &= \left\|\boldsymbol{\Sigma}^{*-1/2}(\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{*-1/2}\right\|_F \\ &\leq \frac{1}{\lambda_{min}}\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\|_F \\ &\leq \varepsilon\lambda_{max}/\lambda_{min}\end{aligned}$$

Thus, by substituting $\varepsilon'' = \varepsilon\lambda_{max}/\lambda_{min}$, we get sample complexity of $\tilde{O}(\frac{d^2(\lambda_{max}/\lambda_{min})^2}{\alpha\varepsilon^2})$ for the following guarantees:

$$\left\|\boldsymbol{\Sigma}^{*-1/2}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})\right\|_2 \leq \mathcal{O}(\varepsilon)$$
$$\left\|\boldsymbol{I} - \boldsymbol{\Sigma}^{*-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{*-1/2}\right\|_F \leq \mathcal{O}(\varepsilon)$$

□

C Section 4 OMITTED PROOFS

Lemma 4.2. *For any $\boldsymbol{\mu} \in \mathbb{R}^d$, it holds that: $\ell(\boldsymbol{\mu}) \geq \ell(\boldsymbol{\mu}^*)$.*

Proof. We first verify that the gradient vanishes at $\boldsymbol{\mu} = \boldsymbol{\mu}^*$. First, observe that

$$\nabla\ell(\boldsymbol{\mu}) = -\sum_{A\subseteq[d]}\int_{\mathbf{x}\in\mathbb{R}^{|A|}}\frac{\int_{\mathbf{y}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\cdot\mathbf{1}[\mathbb{S}(\mathbf{y})=A]\cdot\delta(\mathbf{y}_A-\mathbf{x})g_{\boldsymbol{\mu}}(\mathbf{y})d\mathbf{y}}{g^{\mathbb{S}}_{\boldsymbol{\mu}}(A,\mathbf{x})}g^{\mathbb{S}}_{\boldsymbol{\mu}^*}(A,\mathbf{x})d\mathbf{x}$$

Hence:

$$\begin{aligned}\nabla\ell(\boldsymbol{\mu}^*) &= -\sum_{A\subseteq[d]}\int_{\mathbf{x}\in\mathbb{R}^{|A|}}\int_{\mathbf{y}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}^*)\cdot\mathbf{1}[\mathbb{S}(\mathbf{y})=A]\cdot\delta(\mathbf{y}_A-\mathbf{x})g_{\boldsymbol{\mu}^*}(\mathbf{y})d\mathbf{y}d\mathbf{x} \\ &= -\int_{\mathbf{y}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}^*)\cdot\sum_{A\subseteq[d]}\mathbf{1}[\mathbb{S}(\mathbf{y})=A]\cdot\int_{\mathbf{x}\in\mathbb{R}^{|A|}}\delta(\mathbf{y}_A-\mathbf{x})d\mathbf{x}\cdot g_{\boldsymbol{\mu}^*}(\mathbf{y})d\mathbf{y} \\ &= -\int_{\mathbf{y}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}^*)\cdot g_{\boldsymbol{\mu}^*}(\mathbf{y})d\mathbf{y} = 0.\end{aligned}$$

One can also show this by using (3) for the gradient and then using the law of total expectation. We next prove ℓ is convex by showing that $\nabla^2\ell$ is positive semidefinite for any value of μ.

$$\begin{aligned}
&\nabla^2\ell(\boldsymbol{\mu}) \\
&= \mathop{\mathbb{E}}_{(A,\mathbf{x})}\left[-\frac{\int_{\mathbf{y}}(-\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1})\mathbf{1}[\mathbb{S}(\mathbf{y})=A]\cdot\delta(\mathbf{y}_A-\mathbf{x})g_{\boldsymbol{\mu}}(\mathbf{y})d\mathbf{y}}{g^{\mathbb{S}}_{\boldsymbol{\mu}}(A,\mathbf{x})}\right]+ \\
&\qquad \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma})}\left[\left(\frac{\int_{\mathbf{y}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\mathbf{1}[\mathbb{S}(\mathbf{y})=A]\cdot\delta(\mathbf{y}_A-\mathbf{x})g_{\boldsymbol{\mu}}(\mathbf{y})d\mathbf{y}}{\int_{\mathbf{y}}\mathbf{1}[\mathbb{S}(\mathbf{y})=A]\cdot\delta(\mathbf{y}_A-\mathbf{x})g_{\boldsymbol{\mu}}(\mathbf{y})d\mathbf{y}}\right)^2\right] \\
&= \mathop{\mathbb{E}}_{(A,\mathbf{x})}\left[\boldsymbol{\Sigma}^{-1}-\mathop{\mathbf{Cov}}_{\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\mid\mathbb{S}(\mathbf{y})=A,\mathbf{y}_A=\mathbf{x}\right]\right].
\end{aligned} \tag{12}$$

Observe that for a linear thresholding missingness pattern, the set $\{\mathbf{y}:\mathbb{S}(\mathbf{y})=A,\mathbf{y}_A=\mathbf{x}\}$ is convex for any set A and any $\mathbf{x}\in\mathbb{R}^{|A|}$. Using the fact (Corollary 2.1 of Kanter and Proppe (1977)) that the variance of a Gaussian is non-increasing when restricted to a convex set:

$$\mathop{\mathbf{Cov}}_{\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\mid\mathbb{S}(\mathbf{y})=A,\mathbf{y}_A=\mathbf{x}]\preceq\mathop{\mathbf{Cov}}_{\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})]=\boldsymbol{\Sigma}^{-1}.$$

Plugging into (12), we get that $\nabla^2\ell(\boldsymbol{\mu})\succeq 0$ for any $\boldsymbol{\mu}$. □

Lemma 4.5 (Strong Convexity with Missing Entries)**.** *Given our missingness model and Assumption 1.3 with $\beta=\frac{c}{d}$ for some integer $c\in\{1,\dots d\}$, we have that the function with general covariance $\ell(\boldsymbol{\mu})$ is λ-strongly convex for $\lambda=\alpha\beta/\lambda_{\max}(\boldsymbol{\Sigma})$.*

Proof. Equivalently, we need to show that the minimum eigenvalue of the Hessian of the function $\ell(\boldsymbol{\mu})$ is at least λ. From (12), we have

$$\nabla^2\ell(\boldsymbol{\mu})=\mathop{\mathbb{E}}_{(A,\mathbf{x})}\left[\boldsymbol{\Sigma}^{-1}-\mathop{\mathbf{Cov}}_{\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\mid\mathbb{S}(\mathbf{y})=A,\mathbf{y}_A=\mathbf{x}\right]\right] \tag{13}$$

Fix a unit vector $\mathbf{v}\in\mathbb{R}^d$. Let $H\subseteq[d]$ denote the set of indices that contains the $c=\beta d$ highest v_i^2 values. Therefore, $\sum_{i\in H}v_i^2\geq\beta$. Using $\mathbf{v}$ as a test vector: enecccfckgvkkjlfnlclcggguvjldtdurlrlltnecnid

$$\mathbf{v}^\top\nabla^2(\ell(\boldsymbol{\mu}))\mathbf{v}=\mathbf{v}^\top\boldsymbol{\Sigma}^{-1}\mathbf{v}-\mathbf{v}^\top\cdot\left(\mathop{\mathbb{E}}_{(A,\mathbf{x})}\mathop{\mathbf{Cov}}_{\mathbf{y}}(\boldsymbol{\Sigma}^{-1}\mathbf{y}\mid\mathbb{S}(\mathbf{y})=A,\mathbf{y}_A=\mathbf{x})\right)\cdot\mathbf{v}$$

With probability $1-\alpha(H)$, some coordinate in H is not in A. Under this event, we upper-bound $\mathbf{Cov}_{\mathbf{y}}(\boldsymbol{\Sigma}^{-1}\mathbf{y}\mid\mathbb{S}(\mathbf{y})=A,\mathbf{y}_A=\mathbf{x})$ by $\mathbf{Cov}_{\mathbf{y}}(\boldsymbol{\Sigma}^{-1}\mathbf{y})=\boldsymbol{\Sigma}^{-1}$ using the facts that the missingness is linear thresholding and the aforementioned Corollary 2.1 of Kanter and Proppe (1977). Under the complement event that $H\subseteq A$ is fully observed, we use the upper-bound $\mathbf{Cov}_{\mathbf{y}}(\boldsymbol{\Sigma}^{-1}\mathbf{y}\mid\mathbf{y}_H=\mathbf{x}_H)$, again by the same facts. By Assumption 1.3, $\alpha(A)\geq\alpha$ and so:

$$\begin{aligned}
\mathbf{v}^\top\nabla^2(\ell(\boldsymbol{\mu}))\mathbf{v}&\geq\alpha\left(\mathbf{v}^\top\boldsymbol{\Sigma}^{-1}\mathbf{v}-\mathbf{v}^\top\cdot\mathop{\mathbf{Cov}}_{\mathbf{y}}(\boldsymbol{\Sigma}^{-1}\mathbf{y}\mid\mathbf{y}_H=\mathbf{x}_H)\cdot\mathbf{v}\right) \\
&=\alpha\mathbf{v}^\top\left(I-\boldsymbol{\Sigma}^{-1}\mathop{\mathbf{Var}}_{\mathbf{y}}(\mathbf{y}\mid\mathbf{y}_H=\mathbf{x}_H)\right)\boldsymbol{\Sigma}^{-1}\mathbf{v}
\end{aligned} \tag{14}$$

We use the standard facts that for any set A:

(i) $\mathbf{Var}_{\mathbf{y}}(\mathbf{y}\mid\mathbf{y}_A=\mathbf{x}_A)=\begin{bmatrix}0 & 0\\ 0 & \boldsymbol{\Sigma}_{\bar{A}\bar{A}}-\boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}}\end{bmatrix}$

(ii) $\boldsymbol{\Sigma}^{-1}=\begin{bmatrix}\boldsymbol{\Sigma}_{AA}^{-1}+\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}}(\boldsymbol{\Sigma}_{\bar{A}\bar{A}}-\boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}})^{-1}\boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1} & -\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}}(\boldsymbol{\Sigma}_{\bar{A}\bar{A}}-\boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}})^{-1}\\ -(\boldsymbol{\Sigma}_{\bar{A}\bar{A}}-\boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}})^{-1}\boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1} & (\boldsymbol{\Sigma}_{\bar{A}\bar{A}}-\boldsymbol{\Sigma}_{\bar{A}A}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{A\bar{A}})^{-1}\end{bmatrix}$

Using them repeatedly to simplify (14), we get:

$$\begin{aligned}\mathbf{v}^\top \nabla^2(\ell(\boldsymbol{\mu}))\mathbf{v} &\geq \alpha \mathbf{v}^\top \left(I - \begin{bmatrix} 0 & -\boldsymbol{\Sigma}_{HH}^{-1}\boldsymbol{\Sigma}_{H\bar{H}} \\ 0 & I \end{bmatrix}\right) \boldsymbol{\Sigma}^{-1}\mathbf{v} \\ &= \alpha \mathbf{v}^\top \begin{bmatrix} I & \boldsymbol{\Sigma}_{HH}^{-1}\boldsymbol{\Sigma}_{H\bar{H}} \\ 0 & 0 \end{bmatrix} \boldsymbol{\Sigma}^{-1}\mathbf{v} \\ &= \alpha \mathbf{v}^\top \begin{bmatrix} \boldsymbol{\Sigma}_{HH}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \mathbf{v} \\ &= \alpha \mathbf{v}_H^\top \boldsymbol{\Sigma}_{HH}^{-1}\mathbf{v}_H\end{aligned}$$

Finally, we use our choice of H and that $\lambda_{\min}(\boldsymbol{\Sigma}_{HH}^{-1}) = 1/\lambda_{\max}(\boldsymbol{\Sigma}_{HH}) \geq 1/\lambda_{\max}(\boldsymbol{\Sigma})$ to obtain our claim. □

We now describe our solutions to the following three problems as outlined in Section 4.

- **Initialization**: efficiently compute an initial feasible point from which to start the optimization. The pseudocode for `Initialize` appears in Algorithm 2;
- **Gradient estimation**: design a nearly unbiased sampler for $\nabla\ell(\boldsymbol{\mu})$ using Langevin sampling. The `SampleGradient` pseudocode appears in Algorithm 5;
- **Efficient projection**: perform an efficient projection into a set of feasible points to make sure that PSGD converges. The pseudocode presents in Algorithm 6.

C.1 GRADIENT ESTIMATION

Recall from the gradient expression in (2) that the main obstacle in computing $\nabla\ell(\boldsymbol{\mu})$ is the term $\mathbb{E}_{(A,\mathbf{x})}\mathbb{E}_{\mathbf{y}}[\mathbf{y} \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}]$. Here, $(A, \mathbf{x})$ is an observation generated from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})^{\mathbb{S}}$, while $\mathbf{y}$ is sampled from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for some $\boldsymbol{\mu} \in \mathbb{R}^d$. So, to implement `SampleGradient`, we need an approximately unbiased estimator.

The most straightforward way is to apply rejection sampling: for an $(A, \mathbf{x})$ generated by $\mathcal{O}$, keep sampling $\mathbf{y}$ from the conditional[4] distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid \mathbf{y}_A = \mathbf{x})$ until $\mathbb{S}(\mathbf{y}) = A$. If $\boldsymbol{\mu}^* = \boldsymbol{\mu}$, then the expected cost of the rejection sampling is $O(1/\gamma)$ by Assumption 1.4, as $\mathbf{y}_A = \mathbf{x}$ implies $\mathbf{y}_C = \mathbf{x}_C$. The issue that arises is that the probability of $\mathbb{S}(\mathbf{y}) = A$ can decrease exponentially in the distance between $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^*$, and so, rejection sampling becomes infeasible.

To sample the gradient when $\boldsymbol{\mu}$ is far from $\boldsymbol{\mu}^*$, we use the projected Langevin Monte Carlo algorithm (Bubeck et al., 2018). For an observation $(A, \mathbf{x})$, suppose $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\boldsymbol{\mu}_{\text{cond}}$ and $\boldsymbol{\Sigma}_{\text{cond}}$ be the mean and covariance of $\mathbf{y}_{\bar{A}}$ conditioned on $\mathbf{y}_A = \mathbf{x}$. It is well-known that the conditional distribution is Gaussian with:

$$\boldsymbol{\mu}_{\text{cond}} = \boldsymbol{\mu}_{\bar{A}} + \boldsymbol{\Sigma}_{\bar{A},A}\boldsymbol{\Sigma}_{A,A}^{-1}(\mathbf{x} - \boldsymbol{\mu}_A) \tag{15}$$

$$\boldsymbol{\Sigma}_{\text{cond}} = \boldsymbol{\Sigma}_{\bar{A},\bar{A}} - \boldsymbol{\Sigma}_{\bar{A},A}\boldsymbol{\Sigma}_{A,A}^{-1}\boldsymbol{\Sigma}_{A,\bar{A}} \tag{16}$$

where $\bar{A}$ represents $[d] \setminus A$. Let[5] $\mathcal{K} = \{\mathbf{z} \in \mathbb{R}^{d-|A|} \mid \mathbb{S}(\mathbf{x} \circ \mathbf{z}) = A\}$. The iteration of the projected Langevin Monte Carlo algorithm takes the following form:

$$\mathbf{z}^{(t+1)} = \Pi_{\mathcal{K}\cap\mathcal{B}_\Sigma(\boldsymbol{\mu},R)}\left(\mathbf{z}^{(t)} - \frac{\eta}{2}\boldsymbol{\Sigma}_{\text{cond}}^{-1}(\mathbf{z}^{(t)} - \boldsymbol{\mu}_{\text{cond}}) + \sqrt{\eta}\cdot\boldsymbol{\zeta}^{(t)}\right) \tag{17}$$

where η is a step-size parameter, R is an appropriate radius parameter, and $\boldsymbol{\zeta}^{(0)}, \boldsymbol{\zeta}^{(1)}, \ldots$ are i.i.d. samples from the standard normal distribution in $(d - |A|)$-dimensions. We implicitly make the reasonable assumption here that Mahalanobis projection to the convex set $\Pi_{\mathcal{K}\cap\mathcal{B}_\Sigma(\boldsymbol{\mu},R)}(\cdot)$ can be performed efficiently. The pseudocode for `SampleGradient` appears in Algorithm 5.

C.2 PROJECTION TO FEASIBLE DOMAIN

Next, in each iteration of SGD in `MissingDescent` (Algorithm 3), we need to choose a projection set to make sure that PSGD converges. Specifically, we project a current guess back to a $\mathcal{B}_{\boldsymbol{\Sigma}}$ ball centered at $\boldsymbol{\mu}^{(0)}$.

[4] We need to sample from the conditional distribution because $\mathbf{y} \equiv \mathbf{x}$ occurs with probability 0 in the $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ measure.

[5] $\mathbf{x}_A \circ \mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^{d-|A|}$, denotes the vector $\mathbf{y} \in \mathbb{R}^d$ where $\mathbf{y}_A = \mathbf{x}_A$ and $\mathbf{y}_{\bar{A}} = \mathbf{z}$.

Algorithm 6: [ProjectToDomain] The function that projects a current guess back to the domain onto the $\mathcal{B}_{\mathbf{\Sigma}}$ ball.

Input : $\boldsymbol{\mu}^{(0)}, \mathbf{v}$, parameter r_{proj}

return $\boldsymbol{\mu}^{(0)} + \min\{r_{proj}, \|(\mathbf{v} - \boldsymbol{\mu}^{(0)})\|_{\Sigma}\} \cdot \frac{(\mathbf{v}-\boldsymbol{\mu}^{(0)})}{(\|\mathbf{v}-\boldsymbol{\mu}^{(0)})\|_{\Sigma}}$

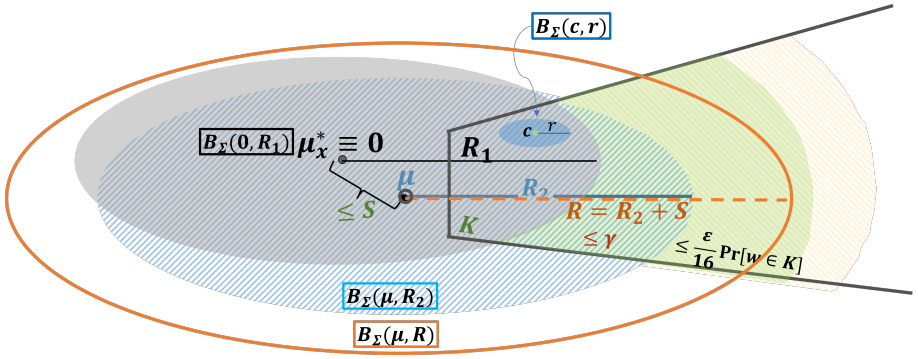

Figure 2: An illustration of convex sets in Section 4.3.2.

C.3 BOUNDED STEP VARIANCE AND GRADIENT BIAS

Lemma C.1. *$\mathcal{K} \cap \mathcal{B}_{\mathbf{\Sigma}}(\mathbf{0}, R_1) \supseteq \mathcal{B}_{\mathbf{\Sigma}}(\mathbf{c}, r)$ for some $\mathbf{c}$, where $R_1 = \sqrt{d} + O(\sqrt{\log(1/\gamma)}$ and $r = \Omega(\gamma/d^2)$*

Lemma C.2. *For $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\|\boldsymbol{\mu}\|_{\mathbf{\Sigma}} \leq S$,*

$$\Pr[\mathbf{w} \notin \mathcal{B}_{\mathbf{\Sigma}}(\boldsymbol{\mu}, R_2) \cap K] \leq \frac{\varepsilon}{16} \cdot \Pr[\mathbf{w} \in K]$$

where $R_2 = \text{poly}\,(d, S, \log(1/\gamma), \log(1/\varepsilon)) > R_1$.

Proof. Suppose $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. By standard concentration of gaussians:

$$\Pr[\|\mathbf{v}\|_{\mathbf{\Sigma}} > \sqrt{d} + O(\sqrt{\log(1/\delta)})] \leq \delta$$

Setting $\delta = \gamma/2$, we get that:

$$\Pr[\mathbf{v} \in \mathcal{B}_{\mathbf{\Sigma}}(\mathbf{0}, R_1) \cap \mathcal{K}] \geq \gamma/2. \tag{18}$$

Invoking Lemma 12 of Cherapanamjeri et al. (2022), we get that there exists $\mathbf{c}$ such that $\mathcal{B}_{\mathbf{\Sigma}}(\mathbf{0}, R_1) \cap \mathcal{K}$ contains $\mathcal{B}_{\mathbf{\Sigma}}(\mathbf{c}, r)$ for $r = \Omega(\gamma/d^2)$. □

Lemma C.2. *For $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\|\boldsymbol{\mu}\|_{\mathbf{\Sigma}} \leq S$,*

$$\Pr[\mathbf{w} \notin \mathcal{B}_{\mathbf{\Sigma}}(\boldsymbol{\mu}, R_2) \cap K] \leq \frac{\varepsilon}{16} \cdot \Pr[\mathbf{w} \in K]$$

where $R_2 = \text{poly}\,(d, S, \log(1/\gamma), \log(1/\varepsilon)) > R_1$.

Proof. Using (18):

$$\Pr[\mathbf{w} \in \mathcal{K}] \geq \Pr[\mathbf{w} \in \mathcal{B}_{\mathbf{\Sigma}}(\mathbf{0}, R_1) \cap \mathcal{K}] \geq \frac{\gamma}{2} \exp\left(-\frac{\|\boldsymbol{\mu}\|_{\mathbf{\Sigma}}^2 + 2R_1}{2}\right) \geq \frac{\gamma}{2} \exp\left(-\frac{S^2}{2} - R_1\right)$$

Call the lower-bound on the right γ'. Note that γ' may be exponentially smaller than γ for large S.

We define R_2 large enough so that $\Pr[\mathbf{w} \notin \mathcal{B}_{\mathbf{\Sigma}}(\boldsymbol{\mu}, R_2)] \leq \frac{\varepsilon}{16}\gamma'$. By concentration of gaussians, it suffices to take $R_2 = \sqrt{d} + O(\sqrt{\log(1/\gamma')}) = \sqrt{d} + O\left(\sqrt{\log \frac{1}{\gamma\varepsilon} + S^2 + R_1}\right)$. Note that the claim about R_2 follows. □

Theorem 4.6. *Assume* $\|\boldsymbol{\mu}^* - \boldsymbol{\mu}\|_{\boldsymbol{\Sigma}} \le S$. *For* $R = \tilde{O}(\sqrt{d} + S + \log(1/\gamma\varepsilon))$, *if* $M = \mathrm{poly}\,(d, S, 1/\gamma, 1/\varepsilon)$ *and* $\eta = \tilde{\Theta}(R^2/M)$, *then*

$$\mathrm{d_{TV}}(\mathbf{z}^{(M)}, \mathcal{N}(W^{-1}\boldsymbol{\mu}_{cond}, I)) \le \varepsilon.$$

Proof of Theorem 4.6. With R_2 as in Lemma C.2, set $R = R_2 + S$, so that $\mathcal{L} = \mathcal{B}_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}, R) \cap \mathcal{K}$. Since $R_2 > R_1$, Lemma C.1 implies that $\mathcal{L}$ contains a ball of radius r. On the other hand, by Lemma C.2, $\Pr[\mathbf{w} \notin \mathcal{L}] \le \frac{\varepsilon}{4}\Pr[\mathbf{w} \in \mathcal{K}]$ for $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which implies that the truncation of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to $\mathcal{K}$ and to $\mathcal{L}$ are at most $\varepsilon/2$ far from each other in TV distance.

We can now use the main result of Bubeck et al. (2018) to approximately sample from the truncated gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathcal{L})$ with TV error $\varepsilon/2$. This work analyzes the projected Langevin Monte Carlo algorithm for sampling from a distribution μ on $\mathbb{R}^d$ whose density is proportional to $\exp(-f(\mathbf{x})) \cdot 1[\boldsymbol{x} \in \mathcal{M}]$ where $\mathcal{M}$ is a convex body containing the origin. Suppose for all $\mathbf{x}, \mathbf{y} \in \mathcal{M}$, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le \beta\|\mathbf{x} - \mathbf{y}\|$ and $\|\nabla f(\mathbf{x})\| \le \ell$. Consider the Langevin dynamics with $\bar{\mathbf{X}}_0 = \mathbf{0}$ and:

$$\bar{\mathbf{X}}_{k+1} = \Pi_{\mathcal{M}}\left(\bar{\mathbf{X}}_k - \frac{\eta}{2}\nabla f(\bar{\mathbf{X}}_k) + \sqrt{\eta}\boldsymbol{\zeta}_k\right)$$

where $\boldsymbol{\zeta}_0, \boldsymbol{\zeta}_1, \ldots$ are i.i.d. standard normal variables. If $\mathcal{M}$ contains a Euclidean ball of radius 1 and is contained in a Euclidean ball of radius R_{out}, then Theorem 1 of Bubeck et al. (2018) claims that $\mathrm{d_{TV}}(\bar{\mathbf{X}}_N, \mu) \le \varepsilon$ if $\eta = \tilde{\Theta}(R_{\mathrm{out}}^2/N)$ and $N = \tilde{\Omega}(R_{\mathrm{out}}^6 \max(d, R_{\mathrm{out}}\ell, R_{\mathrm{out}}\beta)^{12}/\varepsilon^{12})$.

In our context, Algorithm 5 already transforms by a Cholesky decomposition of $\boldsymbol{\Sigma}$ so as to transform Mahalanobis distance to Euclidean distance. We can then scale by r (from Lemma C.1) to ensure that a Euclidean unit ball is contained inside the transformed $\mathcal{L}$. The radius of the outer ball is then $R_{\mathrm{out}} \le R/r \le \tilde{O}(d^3 S/\gamma)$. We can bound the parameters β and ℓ (similarly to Section B.3 of Cherapanamjeri et al. (2022)):

$$\beta = O\left(\frac{\gamma^2}{d^4}\right) \qquad \ell = \tilde{O}\left(\frac{\gamma S}{d^{1.5}}\right)$$

So, invoking the result of Bubeck et al. (2018), for any particular $\mathbf{x}$, the running time of `SampleGradient` is $\mathrm{poly}\,(d, S, 1/\gamma, 1/\varepsilon)$. □

Corollary 4.7. *Let* $\hat{\mathbf{g}}$ *be the output of Algorithm 5 with inputs* $\tilde{\mathbf{x}}$ *and* $\boldsymbol{\mu}$ *and parameters* R, M, η *as in Theorem 4.6. Also, let* $\mathbf{g} = -\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}]$. *Then, we have that:*

$$\|\mathbb{E}[\hat{\mathbf{g}}] - \mathbf{g}\|_2 \le \varepsilon \cdot \mathrm{poly}\,(S, d, 1/\gamma, 1/\varepsilon, \lambda_{\max}, 1/\lambda_{\min}) \tag{4}$$

Furthermore, we have the following bound

$$\mathbb{E}[\|\hat{\mathbf{g}}\|_2^2] \le \mathrm{poly}\,(d, 1/\gamma, S, 1/\lambda_{\min}) \tag{5}$$

Proof. We first show (4):

$$\begin{aligned}
\|\mathbb{E}[\hat{\mathbf{g}}] - \mathbf{g}\|_2 &= \left\|\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_A \circ \mathbb{E}[W\mathbf{z}^{(M)}] - \tilde{\mathbf{x}}_A \circ \mathbb{E}[\mathbf{y}_{\bar{A}} \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}]\right)\right\| \\
&\le \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})}\|\mathbb{E}[W\mathbf{z}^{(M)}] - \mathbb{E}[\mathbf{y}_{\bar{A}} \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}]\| \\
&\le \frac{\varepsilon\sqrt{\lambda_{\max}(\boldsymbol{\Sigma})}}{\lambda_{\min}(\boldsymbol{\Sigma})}(R + O(\sqrt{\log(1/\gamma)}))
\end{aligned}$$

The last inequality holds by the guarantee of Theorem 4.6, as well as the fact that $\mathbf{z}^{(M)}$ is contained within $\mathcal{B}_{\boldsymbol{\Sigma}}(\mu_{\mathrm{cond}}, R)$ while $\mathbb{E}[\mathbf{y}_{\bar{A}} \mid \mathbf{y}_A = \mathbf{x}_A, \mathcal{A}(\mathbf{y}) = A]$ is within $\mathcal{B}_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}_{\mathrm{cond}}, O(\sqrt{\log(1/\gamma)}))$ by Assumption 1.4 and Lemma 6 of Daskalakis et al. (2018).

Given the above and the existence of the projection step in the gradient estimator, we can get the following bound on the centralized second moment of the gradient estimator:

$$\mathbb{E}[\|\hat{\mathbf{g}}\|_2^2] \le \mathbb{E}[\|\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\|^2 \mid \mathbb{S}(\mathbf{y}) = A, \mathbf{y}_A = \mathbf{x}] + \varepsilon\left(\frac{R}{\sqrt{\lambda_{\min}}}\right)^2 \le \mathrm{poly}\,(d, 1/\gamma, S, 1/\lambda_{\min})$$

□

C.3.1 Bound on the Initialization

Lemma C.3 (Empirical Parameters vs True Parameters)**.** *The empirical mean* $\mathbb{E}[w]$ *computed using* $\tilde{\mathcal{O}}(\frac{d\log(nd/\alpha\beta\delta)\log(1/\delta\beta)}{\varepsilon^2})$ *samples by sampling from the general missingness model with probability at least* $1-\delta$ *satisfy* $\|\mathbf{w}-\boldsymbol{\mu}^*\|_2 \leq \mathcal{O}\left(\sqrt{\frac{\lambda_{\max}}{\beta}\log(1/\alpha)}\right)$.

Proof. For each iteration of the for loop in Algorithm 2 uses $\tilde{\mathcal{O}}(\frac{\beta d\log(nd/\alpha\delta')\log(1/\delta')}{\varepsilon^2})$ samples due to the sample complexity bound in Lemma 5 of (Cherapanamjeri et al., 2022) by applying Lemma 5 on Lemma 6 (1) using triangle inequality, and the bound holds with probability $1-\delta'$. Since the for loop makes $\lceil\frac{1}{\beta}\rceil$ iterations, we conclude that using $\tilde{\mathcal{O}}(\frac{d\log(nd/\alpha\delta')\log(1/\delta')}{\varepsilon^2})$ samples, our algorithm satisfies with probability at least $1-\delta'\cdot\lceil\frac{1}{\beta}\rceil$ using union bound. Let $\delta=\delta'\cdot\lceil\frac{1}{\beta}\rceil$. We have $\delta'=O(\delta\beta)$. Therefore, with $\tilde{\mathcal{O}}(\frac{d\log(nd/\alpha\beta\delta)\log(1/\delta\beta)}{\varepsilon^2})$ samples, with probability at least $1-\delta$, the output of Algorithm 2 satisfies that $\|\mathbf{w}-\boldsymbol{\mu}^*\|_2 \leq \mathcal{O}\left(\sqrt{\frac{\lambda_{\max}}{\beta}\log(1/\alpha)}\right)$. □