Efficient Statistics for Sparse Graphical Models from Truncated Samples

Arnab Bhattacharya
School of Computing
NUS
arnabb@nus.edu.sg

Rathin Desai
Computer Science
Boston University
desairathin18@gmail.com

Sai Ganesh Nagarajan
ISTD & ESD
SUTD
sganesh.22@gmail.com

Ioannis Panageas
ISTD
SUTD
ioannis@sutd.edu.sg

Abstract

In this paper, we study high-dimensional estimation from truncated samples. We focus on two fundamental and classical problems: (i) inference of sparse graphical Gaussian models and (ii) support recovery of sparse linear models.

(i) For Gaussian graphical models, suppose d-dimensional samples $\mathbf{x}$ are generated from a Gaussian $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and observed only if they belong to a subset $S \subseteq \mathbb{R}^d$. We show that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be estimated with error ε in the Frobenius norm, using[1] $\tilde{O}\left(\frac{\mathrm{nz}(\boldsymbol{\Sigma}^{-1})}{\varepsilon^2}\right)$ samples from a truncated $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and having access to a membership oracle for S. The set S is assumed to have non-trivial measure under the unknown distribution but is otherwise arbitrary.

(ii) For sparse linear regression, suppose samples $(\mathbf{x}, y)$ are generated where $y = {\boldsymbol{\Omega}^*}^\top \mathbf{x} + \mathcal{N}(0,1)$ and $(\mathbf{x}, y)$ is seen only if y belongs to a truncation set $S \subseteq \mathbb{R}$. We consider the case that $\boldsymbol{\Omega}^*$ is sparse with a support set of size k. Our main result is to establish precise conditions on the problem dimension d, the support size k, the number of observations n, and properties of the samples and the truncation that are sufficient to recover the support of $\boldsymbol{\Omega}^*$. Specifically, we show that under some natural assumptions, only $O(k^2 \log d)$ samples are needed to estimate $\boldsymbol{\Omega}^*$ in the ℓ_∞-norm up to a bounded error.

For both problems, the estimator is obtained by minimizing the sum of the empirical negative log-likelihood function and an ℓ_1-regularization term.

1 Introduction

Sparse high-dimensional models are a mainstay of modern statistics and machine learning. In this work, we consider two different sparse linear models that have been the subject of intensive study.

– **Sparse Gaussian Graphical Models.** Graphical models are used to represent the probabilistic relationships between a collection of variables. These models are used in a huge number of different domains, such as statistical physics, computational biology, finance, and machine learning; the books [23, 27, 37, 21] give an indication of the breadth of this area. We focus on Gaussian graphical models in which the d variables $X_1, \ldots, X_d$ are distributed

[1] $\mathrm{nz}(A)$ denotes the number of non-zero entries of matrix A.

according to a d-dimensional Gaussian. Specifically, the distribution is described in terms of a density function $p(\mathbf{X})$ where $\mathbf{X} = (X_1, \ldots, X_d)$ and

$$p(\mathbf{X}) = (2\pi)^{-d/2} \cdot (\det \mathbf{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right).$$

Here, $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ correspond to the mean and variance of the distribution respectively.
It is convenient to reparametrize the density function in terms of the inverse covariance matrix or the *precision matrix*, $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$:

$$p(\mathbf{X}) = (2\pi)^{-d/2} \cdot \exp\left(\boldsymbol{\mu}^T \mathbf{\Theta} \mathbf{X} - \frac{1}{2}\mathbf{X}^\top \mathbf{\Theta} \mathbf{X} - \frac{1}{2}\boldsymbol{\mu}^\top \mathbf{\Theta} \boldsymbol{\mu} + \frac{1}{2}\log\det(\mathbf{\Theta})\right)$$

Note that the exponent is a quadratic polynomial in which the coefficient of $X_i X_j$ is $\Theta_{i,j}$. The symmetric matrix $\mathbf{\Theta}$ naturally defines an undirected graph G on d vertices in which $(i,j) \in E(G)$ iff $\Theta_{i,j} \neq 0$. The graph G also admits a very nice probabilistic interpretation: X_i and X_j are independent conditioned on all other variables if and only if $\Theta_{i,j} = 0$. Thus, for natural systems, it is quite reasonable to assume that the degree of each node in G is small, as this corresponds to assuming that each variable is "directly" dependent on a small number of variables. Note that even if $\mathbf{\Theta}$ is sparse, $\mathbf{\Sigma}$ could be dense; in fact, in many typical systems, any pair of variables is correlated even though they are not directly dependent.
The problem of learning sparse high-dimensional Gaussian graphical models (in terms of the precision matrix) has a rich history. Popular approaches include the graphical Lasso [17, 40, 3, 12, 32, 31], neighborhood-based methods [4, 26, 35], and CLIME [6] which have been proved to work under different sets of assumptions.

– **Sparse Linear Regression.** A fundamental problem in data science is to solve the following inverse problem. Given pairs $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$, find the "best" choice of $\mathbf{\Omega} \in \mathbb{R}^d$ so that $Y_i - \mathbf{\Omega}^\top \mathbf{X_i}$ is small in some norm. It is natural to want $\mathbf{\Omega}$ to be sparse so that the prediction can be made based on a small number of variables.
Consider the model $Y = \mathbf{\Omega^*}^\top \mathbf{X} + \varepsilon$ where ε is a Gaussian random variable and $\mathbf{\Omega}^*$ is a sparse vector. There has been a huge amount of work on this problem. In the high-dimensional setting, a very popular approach is using ℓ_1-regularization, leading to the Lasso algorithm [33]. By now, we have an almost complete understanding of the necessary and sufficient conditions needed for Lasso to recover $\mathbf{\Omega}^*$; see the discussion and references in Chapter 7 of [36].

In our work, we study the above two problems in the setting where the samples are subject to *truncation*. Truncation is also a classic challenge in statistics, occurring whenever the observation process is dependent on the drawn sample. Following early work by Galton [18], there has been a sustained history of research on truncated distributions, in particular, truncated Gaussians (see the citations in [9]) and truncated linear regression [34, 1, 19, 5]. We pick up the thread at [10] who developed a computationally and statistically efficient algorithm to learn a multivariate Gaussian given truncated samples and assuming that the truncation set is known. A follow-up work, [11], extended the analysis to the linear regression problem where only those samples $(\mathbf{X}_i, Y_i)$ are seen in which $Y_i \in S$, the truncation set.

To the best of our knowledge, ours is the first work that examines the problems of learning sparse Gaussian graphical models and linear models with truncated samples. We state our results next.

Statement of the results The first contribution of the paper is the following theorem on learning Gaussian graphical models up to small Frobenius norm error. The sampling process is that samples from an unknown d-variate Gaussian are only revealed if they belong to a subset $S \subseteq \mathbb{R}^d$; otherwise, the samples are completely hidden.

Theorem 1.1 (Frobenius norm)**.** *Suppose that we are given oracle access to a measurable set S, so that $\int_S \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}^*, \mathbf{\Sigma}^*) d\mathbf{x} = \alpha > 0$ for some d-variate $\mathcal{N}(\boldsymbol{\mu}^*, \mathbf{\Sigma}^*)$ and constant[2] $\alpha > 0$. There exists an estimator $\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{\Sigma}}$ that uses $\tilde{O}\left(\frac{nz(\mathbf{\Sigma}^{*-1})}{\varepsilon^2}\right)$ samples from the truncated distribution $\mathcal{N}(\boldsymbol{\mu}^*, \mathbf{\Sigma}^*, S)$ so that with probability at least 99%*

$$\left\| I - \mathbf{\Sigma}^{*-1/2} \tilde{\mathbf{\Sigma}} \mathbf{\Sigma}^{*-1/2} \right\|_F \leqslant \varepsilon \text{ and } \left\| \mathbf{\Sigma}^{*-1/2} (\boldsymbol{\mu}^* - \tilde{\boldsymbol{\mu}}) \right\|_2 \leqslant \varepsilon.$$

[2] Think of α like 1%.

The second contribution of the paper solves the variable selection problem for linear models, under certain assumptions. The sampling process is as follows: each covariate $\mathbf{x}^{(i)} \in \mathbb{R}^d$ is picked arbitrarily, and the value $y_i = \mathbf{\Omega^*}^\top \mathbf{X}_i + \varepsilon_i$ is revealed only if $y_i \in S$. Here, $\varepsilon_i \sim \mathcal{N}(0, 1)$, the standard normal distribution.

Theorem 1.2 (Linear regression, informal)**.** *Suppose that we are given oracle access to a measurable set S. Let $\mathbf{X}$ denote a design matrix consisting of n samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^d$. Let K denote the unknown support of $\mathbf{\Omega}^*$, and let $k = |K|$. Assume that:*

*(i) (**Survival probability**) For each observed $\mathbf{x}^{(i)}$, the probability that $\mathbf{\Omega^*}^\top \mathbf{x}^{(i)} + w_i$ survives the truncation is not too small.*
*(ii) (**Minimum eigenvalue**) The vector $\mathbf{\Omega}^*$ is identifiable if its support K was known a priori.*
*(iii) (**Mutual incoherence**) Covariates not in the support set K form columns in $\mathbf{X}$ that are approximately orthogonal to the space spanned by the columns corresponding to K.*
*(iv) (**Normalization**) Each entry of $\mathbf{X}$ is small in magnitude.*

Then, with only $n = O(k^2 \log d)$ samples $(\mathbf{x}^{(i)}, y_i)$ from the truncated distribution, one can recover a vector $\hat{\mathbf{\Omega}}$ such that with high probability:

(a) The support of $\hat{\mathbf{\Omega}}$ is contained in K.
(b) If for some $j \in K$, Ω_j^* is larger than a threshold τ (which depends on the problem parameters but not d), then $\hat{\Omega}_j \neq 0$.

Our techniques We first discuss the ideas behind Theorem 1.1. In [10], it was shown that using $n = \tilde{O}\left(\frac{d^2}{\varepsilon^2}\right)$ samples from a d-variate truncated Gaussian distribution with truncation set S of measure some constant $\alpha > 0$, the mean $\boldsymbol{\mu}^*$ and the covariance $\mathbf{\Sigma}^*$ of the untruncated distribution can be estimated with ε error in ℓ_2 and Frobenius norm respectively. The crux of their proof involves proving that the infinite population negative log-likelihood is κ-strongly convex in a neighborhood U ($U \subseteq \mathcal{S}_{d\times d} \times \mathbb{R}^d$) of the true parameters where the radius of U and κ are functions of α[3]. Moreover, they run projected Stochastic Gradient Descent (SGD) with an efficient projection procedure in that neighborhood U. SGD requires a sample from the true truncated distribution in every iteration, so the sample complexity of this approach is at least as much as the number of iterations of SGD. Due to variance reasons, for SGD to converge, the number of samples needed is $\Omega\left(\frac{d^2}{\varepsilon^2}\right)$.

To improve up on their sample complexity, our estimator is the minimizer of a different function - denoted by L_n - which is the *finite* population negative log-likelihood plus a regularization term (see Equation (3.4)). The regularization term is the sum of the absolute values of the entries of the precision matrix (excluding the diagonal entries). This approach is the well-known Graphical Lasso.

One first easy observation is that the finite population negative log-likelihood and the infinite population negative log-likelihood have the same Hessian (thus same convexity properties, see Equation (3.6)). Moreover, since the extra regularization term does not change the convexity properties of the finite population negative log-likelihood, we get for free from [10] that the function L_n is κ-strongly convex in a neighborhood U of the true parameters (same κ and U as before). The crucial part now is that for the Lasso approach to work, we need that the empirical mean and the empirical covariance (from the truncated distribution) is close in ℓ_∞ and max-norm respectively (and not in ℓ_2 and Frobenius norm). The only requirement for the proof to go through is that the number of samples gives the statistical guarantee for Lasso to work (see Lemma 3.2).

For the problem of sparse linear regression with truncated samples, we again consider the Lasso objective, i.e., the sum of the finite population negative log-likelihood plus $\lambda\|\mathbf{\Omega}\|_1$. This objective function is globally convex. Suppose we already know the support K of $\mathbf{\Omega}^*$, the true k-sparse coefficient vector. In this case, we can solve the Lasso objective restricted to the variables in K and hope that it is strongly convex so that the minimum is unique. For the untruncated case, the minimum eigenvalue assumption (Assumption (ii) in Theorem 1.2) implies global strong convexity. In the truncated case, we can only guarantee strong convexity in a neighborhood around $\mathbf{\Omega}^*$. By tuning the

[3]Think of the radius r as $O\left(\frac{\log(1/\alpha)}{\alpha^2}\right)$ and κ to be $O(\alpha^{cr^5})$ where c some constant. U is a subset of $\mathcal{S}_{d\times d} \times \mathbb{R}^d$ where $\mathcal{S}_{d\times d}$ denotes the symmetric matrices of size $d \times d$.

regularization parameter λ, we can ensure that the minimum of the restricted Lasso objective will be in this neighborhood, and hence, is uniquely defined.

The main challenge in proving Theorem 1.2 is to extend the above ideas to when K is not known. To this end, we use the *primal-dual witness method* that has proven very useful for studying many Lasso-type algorithms [35, 29, 20, 7, 28, 30, 24, 38, 39]. We identify a strict dual feasibility condition that implies uniqueness of the Lasso solution and then demonstrate for a set of parameters that the condition holds. In contrast to the untruncated case, we are not able to drive the ℓ_∞-error to zero as n grows to infinity. Also, we require a stronger normalization on the entries of the design matrix. We leave as an interesting open problem the question of overcoming these deficiencies in our analysis.

Other related works Our work comes under the purview of robust statistics where the body of work relating to [14, 15, 13, 22, 8] provided guarantees for computationally efficient robust estimators in the presence of corruptions of an ε fraction of the data, when the samples are drawn from a multivariate Gaussian distribution. In addition, [16] provide statistical query lower bounds on estimation problems related to multivariate Gaussians such as learning mixtures of high dimensional Gaussians. These works generally talk about the seemingly inherent trade-off between increasing the sample complexity for computational tractability. As a result, an important assumption about the underlying problem or the statistical model is that of sparsity. Aside from the works related to estimation in sparse models in classical statistics such as sparse linear regression (LASSO) and sparse PCA [41] to mention a few, there is a line of work related to robust estimation in sparse models, such as robust sparse mean estimation when the covariance matrix is identity and then detection of rank 1 sparse shifts of high dimensional covariances of Gaussian distributions when the mean is zero, using the spiked covariance model as studied in [25, 2].

2 Preliminaries

2.1 Definitions and Notations

Notation We use bold faces to denote vectors and matrices. By $\mathbf{x}_{-j}$ we denote the vector $\mathbf{x}$ that involves all coordinates but j. We use $\text{vec}(A)$ to denote the standard vectorization of matrix A. Moreover, we use $\|\text{vec}(A)\|_{1,\text{off}}$ to denote the ℓ_1 norm of $\text{vec}(A)$ by excluding the diagonal entries of matrix A and $\text{nz}(A)$ for the number of non-zero entries of matrix A. We denote by $\mathcal{S}_{d\times d}$ the set of symmetric matrices.

Norms For a $d \times d$ matrix A,

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2\,,\ \|A\|_\infty = \max_{j\in[n]} \sum_{i=1}^{n} |A_{ij}|,\ \|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}^2}.$$

When A is a symmetric matrix we have that $\|A\|_2 \leqslant \|A\|_\infty \leqslant \|A\|_F \leqslant \sqrt{n}\,\|A\|_2 \leqslant \sqrt{n}\,\|A\|_\infty$. For a vector $\mathbf{x} \in \mathbb{R}^d$ we also have,

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{d} \mathbf{x}_i^2},\ \|\mathbf{x}\|_\infty = \max_{j\in[d]} |\mathbf{x}_j|,\ \|\mathbf{x}\|_1 = \sum_{i=1}^{d} |\mathbf{x}_i|.$$

It holds that ℓ_1 is the dual of ℓ_∞ and for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ one can have $\mathbf{x}^T\mathbf{y} \leqslant \|\mathbf{x}\|_1 \|\mathbf{y}\|_\infty$ (Holder's inequality).

Truncated Gaussian Distribution A truncated Gaussian distribution for a measurable set S with parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ is defined as follows

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S; \mathbf{x}) \stackrel{\text{def}}{=} \begin{cases} \dfrac{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x})}{\int_S \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x}) d\mathbf{x}}\,, & \mathbf{x} \in S \\ 0\,, \ \mathbf{x} \notin S \end{cases} \tag{2.1}$$

where

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

Definition 2.1 (Membership oracle). *Let $S \subset \mathbb{R}^d$ be a measurable set. A membership oracle of S is a function that given an arbitrary $\mathbf{x} \in \mathbb{R}^d$, it returns yes if it belongs to the set, otherwise no (i.e., it implements the indicator function of S). We assume oracle access to the indicator of S.*

Precision matrix and sparsity Let $G = (V, E)$ be an undirected graph with $V = [d]$. A random vector $\mathbf{X} \in \mathbb{R}^d$ is said to be distributed according to (undirected) Gaussian Graphical model with graph G if $\mathbf{X}$ has a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\left(\boldsymbol{\Sigma}^{-1}\right)_{ij} = 0 \;\; \forall\, (i, j) \notin E, \tag{2.2}$$

$\boldsymbol{\Sigma}^{-1}$ which we denote by $\boldsymbol{\Theta}$ is known as the precision matrix. In our results, the sample complexity depends on the number of non-zero entries of $\boldsymbol{\Sigma}^{-1}$, i.e., $\mathrm{nz}(\boldsymbol{\Sigma}^{-1})$.

3 Statistics for Frobenius norm

3.1 Graphical Lasso and finite population Likelihood

The infinite population negative log-likelihood for a truncated Gaussian $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ with variables $(\boldsymbol{\Theta}, \mathbf{v})$ where $\boldsymbol{\Theta}$ captures $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{v} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is given by (see [10] for calculations)

$$\bar{l}(\boldsymbol{\Theta}, \mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)}\left[\frac{1}{2}\mathbf{x}^T\boldsymbol{\Theta}\mathbf{x} - \mathbf{x}^T\mathbf{v}\right] - \log\left(\int_S \exp(-\frac{1}{2}\mathbf{z}^T\boldsymbol{\Theta}\mathbf{z} + \mathbf{z}^T\mathbf{v})d\mathbf{z}\right). \tag{3.1}$$

Moreover, the gradient of the function above $\bar{l}(\boldsymbol{\Theta}, \mathbf{v})$ is given by

$$\nabla\bar{l}(\boldsymbol{\Theta}, \mathbf{v}) := -\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, S)}\left[\begin{pmatrix} \mathrm{vec}(-\frac{1}{2}\mathbf{x}\mathbf{x}^T) \\ \mathbf{x} \end{pmatrix}\right] + \mathbb{E}_{\mathbf{z}\sim\mathcal{N}\left(\boldsymbol{\Theta}^{-1}\mathbf{v}, \boldsymbol{\Theta}^{-1}, S\right)}\left[\begin{pmatrix} \mathrm{vec}(-\frac{1}{2}\mathbf{z}\mathbf{z}^T) \\ \mathbf{z} \end{pmatrix}\right] \tag{3.2}$$

and its Hessian is

$$\nabla^2\bar{l}(\boldsymbol{\Theta}, \mathbf{v}) := \mathrm{Cov}_{\mathbf{z}\sim\mathcal{N}\left(\boldsymbol{\Theta}^{-1}\mathbf{v}, \boldsymbol{\Theta}^{-1}, S\right)}\left[\begin{pmatrix} \mathrm{vec}(-\frac{1}{2}\mathbf{z}\mathbf{z}^T) \\ \mathbf{z} \end{pmatrix}, \begin{pmatrix} \mathrm{vec}(-\frac{1}{2}\mathbf{z}\mathbf{z}^T) \\ \mathbf{z} \end{pmatrix}\right]. \tag{3.3}$$

We define the following score objective with parameter $\lambda > 0$ to be chosen later

$$L_n(\boldsymbol{\Theta}, \mathbf{v}) := l_n(\boldsymbol{\Theta}, \mathbf{v}) + \lambda \left\|\mathrm{vec}(\boldsymbol{\Theta})\right\|_{1,\text{off}}, \tag{3.4}$$

where

$$l_n(\boldsymbol{\Theta}, \mathbf{v}) = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\mathbf{x}_i^T\boldsymbol{\Theta}\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^T\mathbf{v} - \log\left(\int_S \exp(-\frac{1}{2}\mathbf{z}^T\boldsymbol{\Theta}\mathbf{z} + \mathbf{z}^T\mathbf{v})d\mathbf{z}\right), \tag{3.5}$$

given i.i.d samples $\mathbf{x}_1, ..., \mathbf{x}_n$ from the true truncated distribution i.e., it is the finite population negative log-likelihood.

From Lemma A.4 (one of the main Lemmas of [10]), we know that $\bar{l}(\mathbf{v}, \boldsymbol{\Theta})$ is strongly convex in some neighborhood $U \subseteq \mathcal{S}_{d\times d} \times \mathbb{R}^d$ of the true parameters. We can conclude that $L_n(\boldsymbol{\Theta}, \mathbf{v})$ is also strongly convex in the same neighborhood because the term $\lambda \left\|\mathrm{vec}(\boldsymbol{\Theta})\right\|_{1,\text{off}}$ is a convex function and

$$\nabla^2 l_n(\boldsymbol{\Theta}, \mathbf{v}) = \nabla^2\bar{l}(\boldsymbol{\Theta}, \mathbf{v}) \text{ i.e., they have same strong-convexity properties.} \tag{3.6}$$

The following lemma indicates that the minimizer $\tilde{\boldsymbol{\Theta}}, \tilde{\mathbf{v}}$ of function L_n does not put too much weight on the coordinates ij of $\tilde{\boldsymbol{\Theta}}$ for which $\boldsymbol{\Theta}^*_{ij} = 0$, where $(\boldsymbol{\Theta}^*, \mathbf{v}^*)$ denote the true parameters.

Lemma 3.1 (Lasso guarantee). *Let $(\tilde{\boldsymbol{\Theta}}, \tilde{\mathbf{v}})$ be the minimum of L_n and $(\boldsymbol{\Theta}^*, \mathbf{v}^*)$ be the true parameters. Assume that $\lambda \geqslant 2 \left\|\nabla_{\boldsymbol{\Theta}} l_n(\boldsymbol{\Theta}^*, \mathbf{v}^*)\right\|_{\infty,\mathit{off}}$ and $\Delta = \tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*, \delta = \tilde{\mathbf{v}} - \mathbf{v}^*$ then it holds*

$$\frac{1}{3}\left\|vec(\Delta_{\tilde{T}})\right\|_1 - \frac{1}{3}\left\|\delta\right\|_1 \leqslant \left\|vec(\Delta_T)\right\|_1,$$

where T denotes the support of $\boldsymbol{\Theta}^*$ and $\tilde{T}$ denotes the complement. Moreover, we may assume that $\tilde{\boldsymbol{\Theta}}$ is symmetric.

From Lemma 3.1 and Cauchy-Schwarz inequality we conclude that

$$\|\mathrm{vec}(\Delta)\|_1 + \|\delta\|_1 \leqslant \|\mathrm{vec}(\Delta_T)\|_1 + 3\|\mathrm{vec}(\Delta_T)\|_1 + 2\|\delta\|_1 \leqslant 4\sqrt{\mathrm{nz}(\mathbf{\Theta}^*) + d}(\|\Delta\|_F + \|\delta\|_2), \tag{3.7}$$

We can now prove using Lemma 3.1 that for an appropriate choice of λ, the minimizer $(\tilde{\mathbf{\Theta}}, \tilde{\mathbf{v}})$ of L_n will be close to the true parameters $(\mathbf{\Theta}^*, \mathbf{v}^*)$.

Lemma 3.2 ($(\tilde{\mathbf{\Theta}}, \tilde{\mathbf{v}})$ are close to the true parameters)**.** *Let L_n be κ-strong convex in a neighborhood of the true parameters. By choosing λ to be $O\left(\frac{\kappa\cdot\varepsilon}{\sqrt{nz(\mathbf{\Theta}^*)}}\right)$ and moreover $\lambda \geqslant 2\|\nabla l_n(\mathbf{\Theta}^*, \mathbf{v}^*)\|_\infty$ then $\left\|\tilde{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_F + \|\tilde{\mathbf{v}} - \mathbf{v}^*\|_2 \leqslant \varepsilon$.*

We finish this section with a concentration lemma about how close the empirical mean and covariance is from the truncated mean and covariance in terms of ℓ_∞ and max norm respectively.

Lemma 3.3 (Concentration of gradient)**.** *Assume that n is $\Omega\left(\frac{\log d \log(1/\delta)}{t^2}\right)$ It holds that*

$$\mathbb{P}\left[\|\nabla l_n(\mathbf{\Theta}^*, \mathbf{v}^*)\|_\infty \geqslant \frac{t}{2}\right] \leqslant \delta.$$

3.2 Proof of Theorem 1.1

We choose λ to be $\tilde{O}\left(\frac{\varepsilon}{12\sqrt{\mathrm{nz}(\mathbf{\Theta}^*)+d}}\right)$ and consider the estimator $(\tilde{\mathbf{\Theta}}, \tilde{\mathbf{v}}) := \arg\min_{\mathbf{\Theta},\mathbf{v}} L_n(\mathbf{\Theta}, \mathbf{v})$. Notice that we need the oracle access to the truncation set S so that we can compute the term $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{\Theta}^{-1}\mathbf{v},\mathbf{\Theta}^{-1},S)}\left[\begin{pmatrix}\mathrm{vec}(-\frac{1}{2}\mathbf{z}\mathbf{z}^T)\\ \mathbf{z}\end{pmatrix}\right]$ which *does not involve* the true parameters $(\mathbf{\Theta}^*, \mathbf{v}^*)$, (so we can approximate to arbitrary accuracy).

We will prove that $(\tilde{\mathbf{\Theta}}, \tilde{\mathbf{v}})$ satisfies the statement of Theorem 1.1.

From Lemma 3.3 we conclude that if n is $\tilde{O}\left(\frac{(\mathrm{nz}(\mathbf{\Theta}^*)+d)\log(1/\delta)}{\varepsilon^2}\right)$ we get that $\lambda \geqslant 2\|\nabla l_n(\mathbf{\Theta}^*, \mathbf{v}^*)\|_\infty$ with probability $1-\delta$. Therefore the assumptions of Lemma 3.2 hold and is guaranteed that the minimizer $(\tilde{\mathbf{\Theta}}, \tilde{\mathbf{v}})$ of L_n satisfies

$$\left\|\tilde{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_F \leqslant \varepsilon \text{ and } \|\tilde{\mathbf{v}} - \mathbf{v}^*\|_2 \leqslant \varepsilon. \tag{3.8}$$

4 Sparse Linear Regression

Recall the model described in the Introduction for the linear regression problem. The probability of obtaining a sample $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ is:

$$\frac{\exp\left(-\frac{1}{2}(y - \mathbf{\Omega}^{*\top}\mathbf{x})^2\right)}{\int \exp\left(-\frac{1}{2}(z - \mathbf{\Omega}^{*\top}\mathbf{x})^2\right) S(z)dz}$$

The infinite population negative log-likelihood function with n samples is then:

$$\bar{\ell}(\mathbf{\Omega}) = \frac{1}{n}\sum_{i=1}^n \mathop{\mathbb{E}}_{y\sim\mathcal{N}(\mathbf{\Omega}^{*\top}\mathbf{x}^{(i)},1,S)}\left[\frac{1}{2}y^2 - y\cdot\mathbf{\Omega}^\top\mathbf{x}^{(i)} - \log\int\exp\left(-\frac{1}{2}z^2 + z\cdot\mathbf{\Omega}^\top\mathbf{x}^{(i)}\right)dz\right] \tag{4.1}$$

As in the last section, we instead work with the finite sample negative log-likelihood, which is based on n samples $(\mathbf{x}^{(1)}, y^{(i)}), \ldots, (\mathbf{x}^{(n)}, y^{(i)})$ with each $y^{(i)}$ being drawn from the distribution $\mathcal{N}(\mathbf{\Omega}^{*\top}\mathbf{x}^{(i)}, 1, S)$:

$$\ell_n(\mathbf{\Omega}) = \frac{1}{n}\sum_{i=1}^n\left(\frac{1}{2}{y^{(i)}}^2 - y^{(i)}\mathbf{\Omega}^\top\mathbf{x}^{(i)} + \log\int\exp\left(-\frac{1}{2}z^2 + z\mathbf{\Omega}^\top\mathbf{x}^{(i)}\right)S(z)dz\right).$$

Note that ℓ_n is a random variable. We add a regularizer to the sample negative log-likelihood to obtain the *truncated Lasso estimator*:

$$\hat{\mathbf{\Omega}} \in \arg\min_{\mathbf{\Omega} \in \mathbb{R}^d} \left\{ \ell_n(\mathbf{\Omega}) + \lambda \|\mathbf{\Omega}\|_1 \right\}. \tag{4.2}$$

In the following, let $\mathbf{X}$ denote the n-by-d design matrix whose i'th row corresponds to the i'th sample $\mathbf{x}^{(i)}$. Also, we let $\mathbf{x}_j \in \mathbb{R}^n$ denote the j'th column of $\mathbf{X}$.

4.1 Assumptions

We now formally state the assumptions under which our result holds. For vectors $\mathbf{\Omega}$ and $\mathbf{x}$, let $\alpha(\mathbf{\Omega}, \mathbf{x}) \overset{\text{def}}{=} \mathbb{E}_{y \sim \mathcal{N}(\mathbf{\Omega}^\top \mathbf{x}, 1)}[S(y)]$. Also, in the following, let $K \subseteq [d]$ denote the support of $\mathbf{\Omega}^*$, and let $k = |K|$.

Our first assumption states that for every observed $\mathbf{x}^{(i)}$, there is a significant probability that the corresponding response variable $y^{(i)}$ is not truncated.

Assumption 4.1 (Survival Probability)**.** *There exists a constant $\alpha > 0$ such that for every $i \in [n]$, $\alpha(\mathbf{\Omega}^*, \mathbf{x}^{(i)}) \geqslant \alpha$.*

Our second assumption is quite mild. It ensures that the model is identifiable when the support set S is known in advance.

Assumption 4.2 (Lower Eigenvalue)**.** *There exists a constant $\sigma_{\min} > 0$ such that*

$$\frac{1}{n} \mathbf{X}_K^\top \mathbf{X}_K \succeq \sigma_{\min} \cdot \mathbf{I}.$$

Our third assumption ensures that the covariates corresponding to the support set are sufficiently prominent. More precisely, the mutual incoherence assumption below requires that if $j \notin K$, then $\mathbf{x}_j$ is approximately orthogonal to the span of the submatrix $\mathbf{X}_K$ corresponding to the covariates in K.

Assumption 4.3 (Mutual incoherence)**.** *There exists a constant $\beta \in (0, 1)$ such that:*

$$\max_{j \notin K} \|\mathbf{x}_j^\top \mathbf{X}_K (\mathbf{X}_K^\top \mathbf{X}_K)^{-1}\|_1 \leqslant \beta.$$

Mutual incoherence is known to hold, for example, with high probability when $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are drawn i.i.d. from $N(0, \mathbf{I}_{d \times d})$ as long as $n \geqslant \Omega\,(k \log d)$.

Our last assumption puts a bound on each entry of $\mathbf{X}$.

Assumption 4.4 (Normalization)**.** *There exists a parameter C such that $\max_{i \in [n]} \|\mathbf{x}^{(i)}\|_\infty = \max_{j \in [d]} \|\mathbf{x}_j\|_\infty \leqslant C$.*

4.2 Support Recovery

We formally state the main theorem of this section.

Theorem 4.5. *Consider a k-sparse linear regression model for which Assumptions 4.1, 4.2, 4.3, and 4.4 are all satisfied. Moreover, suppose that $\frac{C^2 k}{\alpha^5 \sigma_{\min}(1-\beta)}$ is a sufficiently small constant. Then, if*

$$n \geqslant \Omega\left(\frac{C^4 k^2 \log d}{(1-\beta)^2 \sigma_{\min}^2 \alpha^9}\right) \qquad \text{and} \qquad \lambda = \Theta\left(\frac{\alpha^4 \sigma_{\min}}{Ck}\right),$$

any solution $\hat{\mathbf{\Omega}}$ to the objective (4.2) satisfies the following properties with high probability.

(a) ***Uniqueness**: There is a unique solution $\hat{\mathbf{\Omega}}$.*
(b) ***No false inclusion**: $\mathrm{supp}(\hat{\mathbf{\Omega}}) \subseteq \mathrm{supp}(\mathbf{\Omega}^*)$.*
(c) ***ℓ_∞-bounds**: The error $\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*$ satisfies*

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty \leqslant O\left(\sqrt{\frac{\log(1/\alpha)}{\sigma_{\min}}} + \frac{\alpha^4}{C\sqrt{k}}\right).$$

In other words, if the non-zero entries of $\mathbf{\Omega}^*$ are greater than a particular threshold τ (which is independent of d), then the support of $\hat{\mathbf{\Omega}}$ exactly matches with the support of $\mathbf{\Omega}^*$.

In the untruncated setting, it is known (see Chapter 7 of [36]) that λ can be made to scale as $\sim \frac{1}{\sqrt{n}}$, and the ℓ_∞ error is the sum of two terms, one proportional to λ and the other to $\frac{1}{\sqrt{n}}$. Hence, by making n large, the ℓ_∞ error can be made arbitrarily small. In contrast, in our analysis, we cannot make λ arbitrarily small; so, above, we fix it in terms of the other problem parameters.

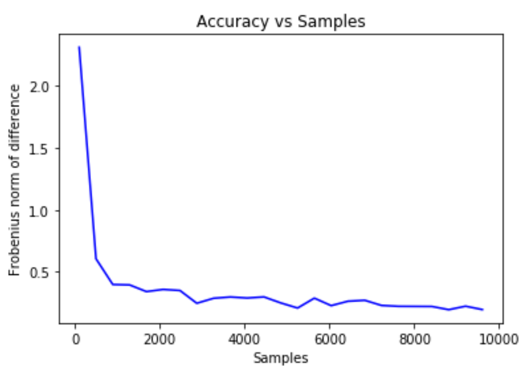
The other notable aspect of Theorem 4.5 is the hypothesis that $\frac{C^2 k}{\alpha^5 \sigma_{\min}(1-\beta)}$ is small, which is also absent from the untruncated setting. The hypothesis can be satisfied if C is mildly decreasing in d (e.g., $\sim 1/\log(d)$), and d is very large.

5 Experiments

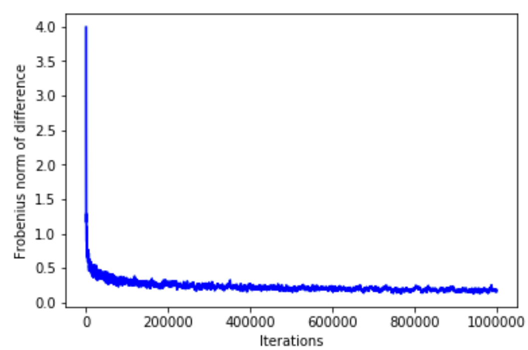
In this section we provide two experiments. The reason of the first experiment is to understand how the error ε (between the true parameters and the estimates) varies as a function of the number of iterations of (proposed algorithm) projected SGD for a fixed number of samples from the true distribution. The second experiment focuses on how the error varies with the number of samples (fixing the number of iterations of SGD to be 10^6.) Note that the reason we use projected SGD is to ensure that $\mathbf{\Theta}$ corresponds to a positive definite matrix at every iteration.

Setup. We generate samples from a 10-dimensional Gaussian distribution with each co-ordinate truncated on a support $(-2, 2)$. The mean of the distribution $\boldsymbol{\mu}$ is set to be $(0, 0, \cdots, 0)$. Moreover, we set the precision matrix $\mathbf{\Theta}$ to be the identity matrix plus 0.2's entries in the the upper and lower diagonal, thus implying the nonzero entries of the precision matrix to be 30 (out of 100 entries).

Algorithm. We run projected SGD on the negative finite sample log-likelihood plus ℓ_1-regularizer, which is function L_n, see (3.4). In every iteration, we require a fresh sample from a truncated distribution using the current estimates as its parameters to get an unbiased estimate of the gradient of L_n.



(a) Error vs Number of Samples



(b) Error vs Number of Iterations of SGD

Evaluation. It seems that projected SGD performs rather poorly (note that the function we optimize L_n is locally strong convex and the initialization is not necessarily in a small neighborhood of the true parameters). In Figure 1a we see that the number of samples scale like $1/\varepsilon^2$ w.r.t the error ε.

6 Conclusion

We studied the problem of parameter estimation for sparse Gaussian Graphical models and the problem of sparse linear regression, given samples that are prone to truncation. We provided sample efficient estimators for both aforementioned problems. One interesting future direction is the computational efficiency, i.e., to come up with polynomial time algorithms that compute estimators for the abovementioned problems with same sample complexity guarantees as in our claims.

Broader Impact

Research in parameter estimation from truncated or corrupted data is a fundamental topic in Statistics and Machine Learning more broadly. Our work is theoretical and we do not believe it has any ethical or future societal consequences.

References

[1] Takeshi Amemiya. Regression analysis when the dependent variable is truncated normal. *Econometrica*, pages 997–1016, 1973.

[2] Sivaraman Balakrishnan, Simon S. Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 169–212, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

[3] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

[4] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

[5] Richard Breen. *Regression Models: Censored, Sample Selected, or Truncated Data*. SAGE Publications, Inc, 1 edition, 1996.

[6] Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[7] Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613. IEEE, 2010.

[8] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data, 2016.

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

[10] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 639–649, 2018.

[11] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Computationally and statistically efficient truncated regression. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 955–960, 2019.

[12] Alexandre d'Aspremont, Onureena Banerjee, and Laurent El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008.

[13] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability, 2016.

[14] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical, 2017.

[15] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently, 2017.

[16] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

[17] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[18] Francis Galton. An examination into the registered speeds of american trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London*, 62(379-387):310–315, 1898.

[19] Jerry A Hauseman and David A Wise. Social experimentation, truncated distributions, and efficient estimation. *Econometrica*, pages 919–938, 1977.

[20] Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep K Ravikumar. A dirty model for multi-task learning. In *Advances in neural information processing systems*, pages 964–972, 2010.

[21] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[22] Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance, 2016.

[23] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[24] Jason D Lee, Yuekai Sun, and Jonathan E Taylor. On model selection consistency of m-estimators with geometrically decomposable penalties. *arXiv preprint arXiv:1305.7477*, 241, 2013.

[25] Jerry Li. Robust sparse estimation tasks in high dimensions, 2017.

[26] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

[27] Marc Mezard, Marc Mezard, and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[28] Sahand N Negahban and Martin J Wainwright. Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization. *IEEE Transactions on Information Theory*, 57(6):3841–3863, 2011.

[29] Guillaume Obozinski, Martin J Wainwright, and Michael I Jordan. Union support recovery in high-dimensional multivariate regression. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 21–26. IEEE, 2008.

[30] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[31] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[32] Adam J Rothman, Peter J Bickel, Elizaveta Levina, Ji Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[33] Robert Tibshirani. Regression shrinkage and selection via the lasso. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 58:267–288, 1994.

[34] James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, pages 24–36, 1958.

[35] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, May 2009.

[36] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

[37] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[38] Weiguang Wang, Yingbin Liang, and Eric P Xing. Collective support recovery for multi-design multi-response linear regression. *IEEE Transactions on Information Theory*, 61(1):513–534, 2014.

[39] Min Xu, Minhua Chen, and John Lafferty. Faithful variable screening for high-dimensional convex regression. *The Annals of Statistics*, 44(6):2624–2660, 2016.

[40] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[41] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

A Useful Lemmas

Lemma A.1 (folklore, see p.309 of [36])**.** *Suppose that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is κ-strongly convex in the sense that*

$$f(\mathbf{y}) \geqslant f(\mathbf{x}) + (\mathbf{y}-\mathbf{x})^T \nabla f(\mathbf{x}) + \frac{\kappa}{2} \|\mathbf{y}-\mathbf{x}\|_2^2 \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{A.1}$$

It holds that

$$(\mathbf{y}-\mathbf{x})^T \left(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\right) \geqslant \kappa \|\mathbf{y}-\mathbf{x}\|_2^2 \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{A.2}$$

Lemma A.2. *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is a twice differentiable, convex function that is locally κ-strongly convex around $\mathbf{x}$, in the sense that the lower bound A.1 holds for all vectors $\mathbf{z}$ in the ball $\mathbb{B}_2 = \{\mathbf{z} : \|\mathbf{z}-\mathbf{x}\|_2 \leqslant \rho\}$. It holds that*

$$(\mathbf{y}-\mathbf{x})^T \left(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\right) \geqslant \rho\kappa \|\mathbf{y}-\mathbf{x}\|_2 \text{ for all } \mathbf{y} \in \mathbb{R}^d \backslash \mathbb{B}_2. \tag{A.3}$$

Proof. Let $\mathbf{y} \in \mathbb{R}^d \backslash \mathbb{B}_2$ and $\mathbf{x}_t = t(\mathbf{y}-\mathbf{x}) + \mathbf{x}$ and $g(t) = \nabla f(\mathbf{x}_t)$. The derivative is given by $g'(t) = \nabla^2 f(\mathbf{x}_t)(\mathbf{y}-\mathbf{x})$. Let $0 \leqslant b \leqslant 1$ be such that $\|\mathbf{x}_b - \mathbf{x}\|_2 = \rho$ and observe that $\rho = b\|\mathbf{y}-\mathbf{x}\|_2$. From fundamental theorem of calculus we get

$$\begin{aligned}
(\mathbf{y}-\mathbf{x})^T(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})) &= (\mathbf{y}-\mathbf{x})^T(g(1) - g(0)) \\
&= \int_0^1 (\mathbf{y}-\mathbf{x})^T \nabla^2 f(\mathbf{x}_t)(\mathbf{y}-\mathbf{x}) dt \\
&\geqslant \int_0^b (\mathbf{y}-\mathbf{x})^T \nabla^2 f(\mathbf{x}_t)(\mathbf{y}-\mathbf{x}) dt \text{ since } f \text{ is convex} \\
&\geqslant \int_0^b \kappa \|\mathbf{y}-\mathbf{x}\|_2^2 \, dt \text{ since } f \text{ is } \kappa-\text{strongly convex} \\
&= b\kappa \|\mathbf{y}-\mathbf{x}\|_2^2 = \kappa\rho \|\mathbf{y}-\mathbf{x}\|_2 ,
\end{aligned}$$

and the claim follows. □

Lemma A.3 (Lemma 6.26 in [36])**.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be an i.i.d. sequence of d-dimensional zero-mean random vectors with covariance matrix $\mathbf{\Sigma}$, and suppose that each component x_{ij} is a sub-Gaussian with parameter at most σ. If $n > \log d$, then for any $\delta > 0$ we have*

$$\mathbb{P}\left[\left\|vec(\mathbf{\Sigma} - \bar{\mathbf{\Sigma}})\right\|_\infty \geqslant t\sigma^2\right] \leqslant 8e^{-\frac{n}{16}\min(t,t^2) + 2\log d} \text{ for all } t > 0, \tag{A.4}$$

where $\bar{\mathbf{\Sigma}}$ is the empirical covariance.

Lemma A.4 (Lemma 4 and Lemma 7 in [10])**.** *Let H be the Hessian of the negative log-likelihood function $\bar{l}(\mathbf{\Theta}, \mathbf{v})$, with the presence of arbitrary truncation S of measure α in the true truncated distribution. Assume that*

1. $\left\|I - \mathbf{\Sigma}^{*1/2}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}^{*1/2}\right\|_F \leqslant B$.

2. $1/B \leqslant \left\| \mathbf{\Sigma}^{*-1/2} \mathbf{\Sigma} \mathbf{\Sigma}^{*-1/2} \right\|_2 \leqslant B$.

3. $\left\| \mathbf{\Sigma}^{-1} \mathbf{\Sigma}^{*1/2} (\boldsymbol{\mu}^* - \boldsymbol{\mu}) \right\|_2 \leqslant B$.

It holds that there exists a constant C so that

$$H \succeq C \left(\frac{\alpha}{12} \right)^{120B^5} \lambda_{\min}(\mathbf{\Sigma}^*)\mathbf{I}.$$

B can be chosen to be $O\left(\frac{\log(1/\alpha)}{\alpha^2}\right)$.

We also record for later use a convenient lemma from [11].

Lemma A.5 (Lemma 6 in [11])**.** *For a set $S \subseteq \mathbb{R}$ and vectors $\mathbf{\Omega}, \mathbf{\Omega}', \mathbf{x} \in \mathbb{R}^d$, :*

$$\alpha(\mathbf{\Omega}, \mathbf{x}) \geqslant \alpha(\mathbf{\Omega}', \mathbf{x})^2 \cdot \exp\left(-|(\mathbf{\Omega} - \mathbf{\Omega}')^\top \mathbf{x}|^2 - 2 \right).$$

Lemma A.6 (Theorem 6.5 in [36])**.** *There are universal constants c_1, c_2, c_3 such that for any row-wise σ-sub-Gaussian random matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, the sample covariance matrix $\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$ satisfies:*

$$\mathbb{P}\left[\frac{\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2}{\sigma^2} \geqslant c_1 \left\{ \sqrt{\frac{d}{n}} + \frac{d}{n} \right\} + \delta \right] \leqslant c_2 e^{-c_3 n \min(\delta, \delta^2)}$$

where $\mathbf{\Sigma}$ is the covariance matrix $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$.

Lemma A.7. *Suppose $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$ is a random vector in $\mathbb{R}^d$, and $F : \mathbb{R}^d \to \{0, 1\}$ is a random function such that $\mathbb{E}_{\mathbf{X}}[\mathbb{P}[F(\mathbf{X}) = 1]] = \alpha$. Let $\boldsymbol{\mu}_F = \mathbb{E}[\mathbf{X} \cdot F(\mathbf{X})]$ and $\mathbf{\Sigma}_F = \mathbb{E}[\mathbf{X}\mathbf{X}^\top \cdot F(\mathbf{X})]$, and denote by $\hat{\boldsymbol{\mu}}_F$ and $\hat{\mathbf{\Sigma}}_F$ their respective empirical counterparts using $n = \tilde{O}(d\varepsilon^{-2} \log \alpha^{-1} \log^2 \delta^{-1})$. Then, with probability $1 - \delta$:*

$$\|\mathbf{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}}_F - \boldsymbol{\mu}_F)\|_2 \leqslant \varepsilon \quad \text{and} \quad (1 - \varepsilon)\mathbf{\Sigma}_F \preceq \hat{\mathbf{\Sigma}}_F \preceq (1 + \varepsilon)\mathbf{\Sigma}_F.$$

Proof. The proof is exactly that of Lemma 5 in [10]. The only difference is that F is a random function here whereas F is deterministic (indicator function of a subset) in [10]. However, the proof remains unchanged. ☐

Lemma A.8. *Suppose $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$, and define $\boldsymbol{\mu}_F$ and $\mathbf{\Sigma}_F$ as in Lemma A.7. Then:*

$$\|\boldsymbol{\mu}_F - \boldsymbol{\mu}\|_{\mathbf{\Sigma}} \leqslant O(\sqrt{\log \alpha^{-1}}) \quad \text{and} \quad O(\alpha^{-2})\mathbf{\Sigma} \succeq \mathbf{\Sigma}_F \succeq \Omega(\alpha^2)\mathbf{\Sigma}.$$

Lemma A.9 (Lemma 14 in [11])**.** *For $\mathbf{x}, \mathbf{w} \in \mathbb{R}^k$, if $z \sim N(\mathbf{w}^\top \mathbf{x}, 1, S)$, then:*

$$\mathbb{E}[(z - \mathbb{E}[z])^2] \geqslant \frac{1}{12} \left(\int_S N(\mathbf{w}^\top \mathbf{x}, 1)(y) \cdot S(y) \cdot dy \right)^2.$$

B Missing Proofs in Section 3

B.1 Proof of Lemma 3.1

Proof. It holds that

$$\left\| \text{vec}(\hat{\mathbf{\Theta}}) \right\|_{1,\text{off}} - \|\text{vec}(\mathbf{\Theta}^*)\|_{1,\text{off}} = \|\text{vec}(\mathbf{\Theta}^*) + \text{vec}(\Delta_T)\|_{1,\text{off}} + \|\text{vec}(\Delta_{\tilde{T}})\|_{1,\text{off}} - \|\text{vec}(\mathbf{\Theta}^*)\|_{1,\text{off}} \tag{B.1}$$

$$\geqslant \|\text{vec}(\Delta_{\tilde{T}})\|_{1,\text{off}} - \|\text{vec}(\Delta_T)\|_{1,\text{off}} \tag{B.2}$$

Observe that

$$L_n(\mathbf{\Theta}^*, \mathbf{v}^*) - L_n(\hat{\mathbf{\Theta}}, \hat{\mathbf{v}}) \geqslant 0. \tag{B.3}$$

Moreover by convexity of l_n we get that

$$\begin{aligned} l_n(\hat{\mathbf{\Theta}}, \hat{\mathbf{v}}) - l_n(\mathbf{\Theta}^*, \mathbf{v}^*) &\geqslant (\Delta^T \ \delta^T) \nabla l_n(\mathbf{\Theta}^*, \mathbf{v}^*) \\ &\geqslant -\left(\|\text{vec}(\Delta)\|_{1,\text{off}} + \|\delta\|_1 \right) \cdot \|\nabla l_n(\mathbf{\Theta}^*, \mathbf{v}^*)\|_{\infty,\text{off}}, \end{aligned}$$

where the last inequality comes from Holder's inequality. Assuming that $\lambda \geqslant 2\left\|\nabla l_n(\mathbf{\Theta}^*, \mathbf{v}^*)\right\|_{\infty,\text{off}}$

$$l_n(\hat{\mathbf{\Theta}}, \hat{\mathbf{v}}) - l_n(\mathbf{\Theta}^*, \mathbf{v}^*) \geqslant -\frac{\lambda}{2}(\|\text{vec}(\Delta_T)\|_{1,\text{off}} + \|\text{vec}(\Delta_{\tilde{T}})\|_{1,\text{off}} + \|\delta\|_1) \tag{B.4}$$

We multiply (B.1) by λ and add it with (B.3) and (B.4). It follows that

$$0 \geqslant \frac{\lambda}{2}\left\|\text{vec}(\Delta_{\tilde{T}})\right\|_{1,\text{off}} - \frac{3\lambda}{2}\left\|\text{vec}(\Delta_T)\right\|_{1,\text{off}} - \frac{\lambda}{2}\left\|\delta\right\|_1 \tag{B.5}$$

Therefore $\|\text{vec}(\Delta_T)\|_1 \geqslant \|\text{vec}(\Delta_T)\|_{1,\text{off}} \geqslant \frac{1}{3}\left\|\text{vec}(\Delta_{\tilde{T}})\right\|_{1,\text{off}} - \frac{1}{3}\|\delta\|_1 = \frac{1}{3}\left\|\text{vec}(\Delta_{\tilde{T}})\right\|_1 - \frac{1}{3}\|\delta\|_1$ and the claim follows. To show that $\hat{\mathbf{\Theta}}$ is symmetric observe that if $(X, \hat{\mathbf{v}})$ is a minimum of L_n by symmetry, so is $(X^T, \hat{\mathbf{v}})$. But $L\left(\frac{X+X^T}{2}, \hat{\mathbf{v}}\right) \leqslant \frac{1}{2}(L(X, \hat{\mathbf{v}}) + L(X^T, \hat{\mathbf{v}}))$ by the triangle inequality for ℓ_1 and the claim follows. ☐

B.2 Proof of Lemma 3.2

Proof. We set $\Delta = \hat{\mathbf{\Theta}} - \mathbf{\Theta}^*$, $\delta = \hat{\mathbf{v}} - \mathbf{v}^*$. By the optimality of $(\hat{\mathbf{\Theta}}, \hat{\mathbf{v}})$

$$(\text{vec}(\Delta)^T\ \delta^T)(\nabla l_n(\hat{\mathbf{\Theta}}, \hat{\mathbf{v}}) + \lambda \text{vec}(\mathbf{Z})) = 0, \tag{B.6}$$

where $\text{vec}(\mathbf{Z})$ is a subgradient for $\|\mathbf{\Theta}\|_{1,\text{off}}$ computed at $\hat{\mathbf{\Theta}}$. Hence by Holder's inequality and taking absolute value for the subgradient part we get

$$(\text{vec}(\Delta)^T\ \delta^T)(\nabla l_n(\hat{\mathbf{\Theta}}, \hat{\mathbf{v}}) - \nabla l_n(\mathbf{\Theta}^*, \mathbf{v}^*)) \leqslant (\|\text{vec}(\Delta)\|_1 + \|\delta\|_1)\left\|\nabla l_n(\mathbf{\Theta}^*, \mathbf{v}^*)\right\|_\infty - \lambda \text{vec}(\Delta)^T(\text{vec}(\mathbf{Z})) \tag{B.7}$$

$$\leqslant \lambda|\text{vec}(\Delta)^T(\text{vec}(\mathbf{Z}))| + \frac{\lambda}{2}(\|\text{vec}(\Delta)\|_1 + \|\delta\|_1) \tag{B.8}$$

$$\leqslant \frac{3\lambda}{2}(\|\text{vec}(\Delta)\|_1 + \|\delta\|_1). \tag{B.9}$$

Assume that $\|\Delta\|_F^2 + \|\delta\|_2^2 > r'^2$ where the ball $B\begin{pmatrix}\text{vec}(\mathbf{\Theta}^*) \\ \mathbf{v}^* \end{pmatrix}, r'\Big)$ is a subset of D_r as defined in [10]. Observe that r' is a function of α and l_n is strongly convex with parameter κ in the ball $B\begin{pmatrix}\text{vec}(\mathbf{\Theta}^*) \\ \mathbf{v}^* \end{pmatrix}, r'\Big)$ (see Lemma A.4). From Lemma A.2 we get that

$$(\text{vec}(\Delta)^T\ \delta^T)(\nabla l_n(\hat{\mathbf{\Theta}}, \hat{\mathbf{v}}) - \nabla l_n(\mathbf{\Theta}^*, \mathbf{v}^*)) \geqslant \kappa r' \sqrt{\|\Delta\|_F^2 + \|\delta\|_2^2} \geqslant \frac{\kappa r'}{2}(\|\Delta\|_F + \|\delta\|_2) \tag{B.10}$$

Combining (B.8) and (B.10) along with (3.7) we get that

$$\frac{3\lambda}{2}(\|\text{vec}(\Delta)\|_1 + \|\delta\|) \geqslant \frac{\kappa r'}{2}(\|\Delta\|_F + \|\delta\|_2) \geqslant \frac{\kappa r'}{8\sqrt{\text{nz}(\mathbf{\Theta}^*) + d}}(\|\text{vec}(\Delta)\|_1 + \|\delta\|_1).$$

Therefore, if we choose $\lambda < \frac{\kappa r'}{12\sqrt{\text{nz}(\mathbf{\Theta}^*)+d}}$, we conclude that $\sqrt{\|\Delta\|_F^2 + \|\delta\|_2^2} \leqslant r'$.

Now, by strong convexity of l_n in that ball and Lemma A.1 it follows that

$$(\text{vec}(\Delta)^T\ \delta^T)(\nabla l_n(\hat{\mathbf{\Theta}}, \hat{\mathbf{v}}) - \nabla l_n(\mathbf{\Theta}^*, \mathbf{v}^*)) \geqslant \kappa\left(\|\Delta\|_F^2 + \|\delta\|_2^2\right) \geqslant \frac{\kappa}{2}\left(\|\Delta\|_F + \|\delta\|_2\right)^2. \tag{B.11}$$

Hence again by combining the above with Equations (B.8) and (3.7) we get that

$$6\lambda\sqrt{\text{nz}(\mathbf{\Theta}^*) + d}(\|\Delta\|_F + \|\delta\|_2) \geqslant \frac{\kappa}{2}\left(\|\Delta\|_F + \|\delta\|_2\right)^2.$$

Thus we choose $\lambda = \min\left(\frac{\kappa r'}{12\sqrt{\text{nz}(\mathbf{\Theta}^*)+d}}, \frac{\kappa\varepsilon}{12\sqrt{\text{nz}(\mathbf{\Theta}^*)+d}}\right)$ and we conclude that $\|\Delta\|_F + \|\delta\|_2 \leqslant \varepsilon$ and the proof is complete. ☐

B.3 Proof of Lemma 3.3

Proof. Observe that

$$\begin{aligned} \nabla_{\boldsymbol{\Theta}} l_n(\boldsymbol{\Theta}^*, \mathbf{v}^*) &= \text{vec}\left(\tfrac{1}{2}\left(\bar{\boldsymbol{\Sigma}}_S + \bar{\boldsymbol{\mu}}_S \bar{\boldsymbol{\mu}}_S^T - \boldsymbol{\Sigma}_S - \boldsymbol{\mu}_S \boldsymbol{\mu}_S^T\right)\right) \\ \nabla_{\mathbf{v}} l_n(\boldsymbol{\Theta}^*, \mathbf{v}^*) &= \text{vec}\left(\boldsymbol{\mu}_S - \bar{\boldsymbol{\mu}}_S\right) \end{aligned} \tag{B.12}$$

where $\bar{\boldsymbol{\mu}}_S, \bar{\boldsymbol{\Sigma}}_S$ are the empirical mean, covariance from n samples from the true truncated distribution and $\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S$ are the mean and covariance matrix of the true truncated Gaussian.

By an easy exact argument to Lemma 5 in [10] and using Hoeffding's inequality, we get that with n at least $\Omega\left(\frac{\log(nd/\alpha\delta)\log(1/\delta)}{t^2}\right)$ it holds that

$$\mathbb{P}\left[\|\bar{\boldsymbol{\mu}}_S - \boldsymbol{\mu}_S\|_\infty > \frac{t}{24}\right] \leqslant \frac{\delta}{2}.$$

It is also clear that if $\|\bar{\boldsymbol{\mu}}_S - \boldsymbol{\mu}_S\|_\infty \leqslant C$ and $C \leqslant 1$ then $\left\|\text{vec}(\bar{\boldsymbol{\mu}}_S \bar{\boldsymbol{\mu}}_S^T - \boldsymbol{\mu}_S \boldsymbol{\mu}_S^T)\right\|_\infty \leqslant C(\|\bar{\boldsymbol{\mu}}_S\|_\infty + \|\boldsymbol{\mu}_S\|_\infty) \leqslant 3C\max(1, \|\boldsymbol{\mu}_S\|_\infty)$. Moreover, by a union bound argument one can show Lemma A.3, thus if n at least $\Omega\left(\frac{\log d \log(1/\delta)}{t^2}\right)$ it holds that

$$\mathbb{P}\left[\left\|\bar{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\right\|_\infty > \frac{t}{4}\right] \leqslant \frac{\delta}{2}.$$

By adding the error probabilities and triangle inequality the claim follows. □

C Proof of Theorem 4.5

Standard calculations (see, e.g., [10, 11]) show that the gradient and Hessian of the empirical log-likelihood can be written as:

$$\nabla \ell_n(\boldsymbol{\Omega}) = -\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}^{(i)} y^{(i)} - \underset{z^{(i)} \sim \mathcal{N}(\boldsymbol{\Omega}^\top \mathbf{x}^{(i)}, 1, S)}{\mathbb{E}}[\mathbf{x}^{(i)} z^{(i)}]\right) \tag{C.1}$$

$$H(\boldsymbol{\Omega}) \stackrel{\text{def}}{=} \nabla^2 \ell_n(\boldsymbol{\Omega}) = \frac{1}{n}\sum_{i=1}^{n} \underset{z^{(i)} \sim \mathcal{N}(\boldsymbol{\Omega}^\top \mathbf{x}^{(i)}, 1, S)}{\mathbf{Cov}}[\mathbf{x}^{(i)} z^{(i)}, \mathbf{x}^{(i)} z^{(i)}] \tag{C.2}$$

From (C.2), since $H(\boldsymbol{\Omega}) \succeq 0$, it is clear that ℓ_n is convex for arbitrary choices of $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$. Let:

$$F(\boldsymbol{\Omega}) = \ell_n(\boldsymbol{\Omega}) + \lambda\|\boldsymbol{\Omega}\|_1.$$

Clearly, F is also convex since ℓ_n is. Now, any optimum $\hat{\boldsymbol{\Omega}} \in \arg\min F(\boldsymbol{\Omega})$ has to satisfy the zero-subgradient condition:

$$\nabla \ell_n(\hat{\boldsymbol{\Omega}}) + \lambda \hat{\mathbf{W}} = 0 \tag{C.3}$$

where $\hat{\mathbf{W}} \in \partial\|\hat{\boldsymbol{\Omega}}\|_1$ and $\nabla \ell_n$ is as in (C.1).

Recall that $K \subseteq [d]$ denotes the support of $\boldsymbol{\Omega}^*$ and $|K| = k$. Let

$$\check{\boldsymbol{\Omega}} \in \arg \min_{\boldsymbol{\Omega} \in \mathbb{R}^{d-1}: \forall j \notin K, \Omega_j = 0} F(\boldsymbol{\Omega}) \tag{C.4}$$

The rest of the proof goes as follows. First, we establish strong convexity of the restricted likelihood function:

$$\ell_n^K(\boldsymbol{\Omega}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}{y^{(i)}}^2 - \boldsymbol{\Omega}^\top \mathbf{x}_K^{(i)} y^{(i)} - \ln \int \exp\left(-\frac{1}{2}z^2 + \boldsymbol{\Omega}\mathbf{x}_K^{(i)} z\right) S(z) dz\right),$$

when $\boldsymbol{\Omega}$ lies in a neighborhood of $\boldsymbol{\Omega}^*$. We show that our choice of λ implies that $\check{\boldsymbol{\Omega}}$ falls in this neighborhood, and hence, there is a unique choice of $\check{\boldsymbol{\Omega}}$ in (C.4). Next, we use the primal-dual witness method to show that if strict dual feasibility holds, then $\hat{\boldsymbol{\Omega}}$ must in fact equal $\check{\boldsymbol{\Omega}}$. Then, we verify that strict dual feasibility holds under our choice of parameters and assumptions, establishing parts (a) and (b) of Theorem 4.5. Finally, it remains to bound the ℓ_∞-distance between $\check{\boldsymbol{\Omega}}$ and $\boldsymbol{\Omega}^*$, proving part (c).

Local strong convexity: We first show that ℓ_n^K is strongly convex for Ω within a ball centered at Ω^*. Below, $\mathbf{\Omega}_K^*$ is the restriction of $\mathbf{\Omega}^*$ to the coordinates in K.

Lemma C.1. *There exists $\kappa \geqslant \Omega(\alpha^4 \cdot \sigma_{\min}) > 0$ such that $\nabla^2 \ell_n^K(\mathbf{\Omega}) \succeq \kappa \cdot \mathbf{I}$ for all $\mathbf{\Omega} \in \mathbb{R}^k$ satisfying $\|\mathbf{\Omega} - \mathbf{\Omega}_K^*\|_2 \leqslant \frac{1}{C\sqrt{k}}$.*

Proof. Computing the Hessian of ℓ_n^K:

$$\begin{aligned}
\nabla^2 \ell_n^K(\mathbf{\Omega}) &= \frac{1}{n}\sum_{i=1}^n \underset{z^{(i)} \sim \mathcal{N}(\mathbf{\Omega}^\top \mathbf{x}_K^{(i)}, 1, S)}{\mathbf{Cov}} [\mathbf{x}_K^{(i)} z^{(i)}, \mathbf{x}_K^{(i)} z^{(i)}] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{x}_K^{(i)} \mathbf{x}_K^{(i)\top} \cdot \underset{z^{(i)} \sim \mathcal{N}(\mathbf{\Omega}^\top \mathbf{x}_K^{(i)}, 1, S)}{\mathbb{E}} [(z^{(i)} - \mathbb{E}\, z^{(i)})^2] \\
&\succeq \frac{1}{12n}\sum_{i=1}^n \mathbf{x}_K^{(i)} \mathbf{x}_K^{(i)\top} \cdot (\alpha(\mathbf{\Omega}, \mathbf{x}^{(i)}))^2 \\
&\succeq \frac{1}{12e^2 n}\sum_{i=1}^n \mathbf{x}_K^{(i)} \mathbf{x}_K^{(i)\top} (\alpha(\mathbf{\Omega}^*, \mathbf{x}^{(i)}))^4 \cdot \exp\left(-2|(\mathbf{\Omega} - \mathbf{\Omega}_K^*)^\top \mathbf{x}_K^{(i)}|^2 - 4\right)
\end{aligned}$$

where the third line[4] follows from Lemma A.9 and the fourth line follows from Lemma A.5. Recall that by Assumption 4.4, $\|\mathbf{x}_K^{(i)}\|_\infty \leqslant C$, and hence, $\|\mathbf{x}_K^{(i)}\|_2 \leqslant C\sqrt{k}$. Plugging into the above, and again using Assumptions 4.1 and 4.2, we obtain:

$$\nabla^2 \ell_n^K(\mathbf{\Omega}) \succeq \frac{1}{12e^2} \cdot \alpha^4 \cdot e^{-6} \cdot \sigma_{\min}\mathbf{I}.$$

□

Bounding the ℓ_2-distance between $\check{\mathbf{\Omega}}$ and $\mathbf{\Omega}^*$: Let $r = \frac{1}{C\sqrt{k}}$ be the radius of the ball B_r around $\mathbf{\Omega}_K^*$ inside which ℓ_n^K is strongly convex (by Lemma C.1). This means that $\ell_n^K(\mathbf{\Omega}) + \lambda\|\mathbf{\Omega}\|_1$ is also locally strongly convex in B_r. The next proposition shows that for an appropriate choice of λ, $\check{\mathbf{\Omega}}$, defined in (C.4) as the minimizer of $\ell_n^K(\mathbf{\Omega}) + \lambda\|\mathbf{\Omega}\|_1$, is in the ball B_r.

Lemma C.2. *If $2\|\nabla \ell_n^T(\mathbf{\Omega}_K^*)\|_\infty < \lambda < \frac{2\kappa r}{3\sqrt{k}}$, where κ and r are as above, then $\check{\mathbf{\Omega}} \in B_r$. Moreover, our choice of λ satisfies these conditions with high probability.*

Proof. The first part follows exactly the same steps as the proof of Lemma 3.2, and so we omit it. For the second part, note that:

$$\nabla \ell_n^K(\mathbf{\Omega}_K^*) = -\frac{1}{n}\sum_{i=1}^n \left(\mathbf{x}_K^{(i)} y^{(i)} - \underset{z^{(i)} \sim \mathcal{N}(\mathbf{\Omega}_K^{*\top} \mathbf{x}_K^{(i)}, 1, S)}{\mathbb{E}} [\mathbf{x}_K^{(i)} z^{(i)}]\right).$$

Consider the i'th summand $\mathbf{x}_K^{(i)} \cdot \left(y^{(i)} - \mathbb{E}_{z^{(i)} \sim \mathcal{N}(\mathbf{\Omega}_K^{*\top} \mathbf{x}_K^{(i)}, 1, S)}[z^{(i)}]\right)$. For every i, each coordinate of the i'th summand has mean zero. Also, for every i, each coordinate of the i'th summand is bounded by $O(C\sqrt{\log(n/\alpha)})$ with high probability because:

(i) $\|\mathbf{x}_K^{(i)}\|_\infty \leqslant C$ for all i by Assumption 4.4.

(ii) $y^{(i)} - \Omega_K^* \mathbf{x}_K^{(i)}$ is distributed as a standard normal truncated to a set of volume at least α. The maximum of n/α standard normal variables is $O(\sqrt{\log(n/\alpha)})$ with high probability.

(iii) $\Omega_K^* \mathbf{x}_K^{(i)} - \mathbb{E}[\mathcal{N}(\mathbf{\Omega}_K^{*\top} \mathbf{x}_K^{(i)}, 1, S)] = \mathbb{E}[\mathcal{N}(\mathbf{\Omega}_K^{*\top} \mathbf{x}_K^{(i)}, 1)] - \mathbb{E}[\mathcal{N}(\mathbf{\Omega}_K^{*\top} \mathbf{x}_K^{(i)}, 1, S)]$ is bounded by $O(\sqrt{\log(1/\alpha)})$ by Lemma 6 in [10].

By the Hoeffding bound, $\|\nabla \ell_K(\mathbf{\Omega}_K^*)\|_\infty < t$ with high probability if $n \geqslant \Omega(C^2 \log k \log(n/\alpha)/t^2)$. Plugging in $t = O(\kappa r/\sqrt{k})$ and the values of κ and r from Lemma C.1, we see that the constraint is non-vacuous if $n \geqslant \tilde{\Omega}(C^2/(\kappa^2 r^2/k)) = \tilde{\Omega}(C^4 k^2/\alpha^8 \sigma_{\min}^2)$. The condition on n is satisfied by the hypotheses of Theorem 4.5. □

[4] We abuse notation in the third line. When we say $\alpha(\mathbf{\Omega}, \mathbf{x}^{(i)})$, we mean by $\mathbf{\Omega}$ the extension of the vector to $\mathbb{R}^d$ where the coordinates not in K are set to zero.

The above two lemmas imply that $\check{\mathbf{\Omega}}$ is uniquely defined. Now, the goal is to relate it to the structure of $\hat{\mathbf{\Omega}}$.

Unique global minimum under strict dual feasibility: We construct a vector $\check{\mathbf{W}}$ so that $(\check{\mathbf{\Omega}}, \check{\mathbf{W}})$ satisfy the zero subgradient condition (C.3). Note that by definition, $\check{\mathbf{\Omega}}$ satisfies the restricted zero-subgradient condition on the coordinates in K:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_K^{(i)} y^{(i)} - \mathop{\mathbb{E}}_{z^{(i)} \sim \mathcal{N}(\check{\mathbf{\Omega}}^\top \mathbf{x}_K^{(i)}, 1, S))} [\mathbf{x}_K^{(i)} z^{(i)}] - \lambda \check{\mathbf{W}}_K = 0$$

for some $\check{\mathbf{W}}_K \in \partial \|\check{\mathbf{\Omega}}_K\|_1$. We set $\check{\mathbf{W}}$ so that it restricts to $\check{\mathbf{W}}_K$ on K and satisfies the zero-subgradient condition (C.3) on all other coordinates as well.

Lemma C.3. *If* $\|\check{\mathbf{W}}_{-K}\|_\infty < 1$*, then* $\check{\mathbf{\Omega}} = \hat{\mathbf{\Omega}}$ *is the unique minimizer of* F*.*

Proof. Note that $\langle \check{\mathbf{W}}, \check{\mathbf{\Omega}} \rangle = \|\check{\mathbf{\Omega}}\|_1$ because outside K, $\check{\mathbf{\Omega}}$ is zero, and $\check{\mathbf{W}}_K \in \partial \|\mathbf{\Omega}_K\|_1$. Then, since $F(\hat{\mathbf{\Omega}}) \leqslant F(\check{\mathbf{\Omega}})$:

$$\begin{aligned}
\lambda \|\hat{\mathbf{\Omega}}\|_1 &\leqslant \ell_n(\check{\mathbf{\Omega}}) + \lambda \|\check{\mathbf{\Omega}}\|_1 - \ell_n(\hat{\mathbf{\Omega}}) \\
&= \ell_n(\check{\mathbf{\Omega}}) - \ell_n(\hat{\mathbf{\Omega}}) + \langle \lambda \check{\mathbf{W}}, \check{\mathbf{\Omega}} \rangle \\
&= \ell_n(\check{\mathbf{\Omega}}) - \ell_n(\hat{\mathbf{\Omega}}) - \langle \nabla \ell_n(\check{\mathbf{\Omega}}), \check{\mathbf{\Omega}} \rangle
\end{aligned}$$

where we used the fact that $\check{\mathbf{\Omega}}$ and $\check{\mathbf{W}}$ satisfy the zero-subgradient condition (C.3). Invoking the convexity of ℓ_n:

$$\lambda \|\hat{\mathbf{\Omega}}\|_1 \leqslant -\langle \nabla \bar{\ell_n}(\check{\mathbf{\Omega}}), \hat{\mathbf{\Omega}} \rangle = \lambda \langle \check{\mathbf{W}}, \hat{\mathbf{\Omega}} \rangle$$

Hence, if $|\check{\mathbf{W}}_K| < 1$ for any s, then $\hat{\mathbf{\Omega}}_K = 0$. The claim then follows since we have already established that $\check{\mathbf{\Omega}}$ is defined uniquely. □

Verifying strict dual feasibility: We next confirm that the $\check{\mathbf{W}}$ vector constructed above satisfies the condition of Lemma C.3.

Lemma C.4. *With high probability,* $\|\check{\mathbf{W}}_{-K}\|_\infty < 1$.

Proof. Since $(\check{\mathbf{\Omega}}, \check{\mathbf{W}})$ satisfy the zero subgradient condition (C.3), we can solve for $\check{\mathbf{W}}_{-K}$:

$$\begin{aligned}
\check{\mathbf{W}}_{-K} &= \frac{1}{n\lambda} \sum_{i=1}^{n} \mathbf{x}_{-K}^{(i)} \left(y^{(i)} - \mathop{\mathbb{E}}_{z^{(i)} \sim \mathcal{N}(\check{\mathbf{\Omega}} \mathbf{x}^{(i)}, 1, S)} z^{(i)} \right) \\
&= \frac{1}{n\lambda} \sum_{i} \mathbf{x}_{-K}^{(i)} \left({\mathbf{x}_K^{(i)}}^\top (\mathbf{\Omega}^* - \check{\mathbf{\Omega}}) + z^{*(i)} - \mathbb{E}\, \check{z}^{(i)} \right)
\end{aligned} \tag{C.5}$$

where $z^{*(i)} \sim N(0, 1, S(\cdot + \mathbf{\Omega}^* \mathbf{x}^{(i)}))$ and $\check{z}^{(i)} \sim N(0, 1, S(\cdot + \check{\mathbf{\Omega}} \mathbf{x}^{(i)}))$. Similarly, we solve for $\check{\mathbf{W}}_K$:

$$\begin{aligned}
\check{\mathbf{W}}_K &= \frac{1}{n\lambda} \sum_{i=1}^{n} \mathbf{x}_K^{(i)} \left(y^{(i)} - \mathop{\mathbb{E}}_{z^{(i)} \sim \mathcal{N}(\check{\mathbf{\Omega}}^\top \mathbf{x}^{(i)}, 1, S)} z^{(i)} \right) \\
&= \frac{1}{n\lambda} \sum_{i} \mathbf{x}_K^{(i)} \left({\mathbf{x}_K^{(i)}}^\top (\mathbf{\Omega}^* - \check{\mathbf{\Omega}}) + z^{*(i)} - \mathbb{E}\, \check{z}^{(i)} \right)
\end{aligned}$$

Let $\mathbf{u}$ denote the random vector whose i'th coordinate is $z^{*(i)} - \mathbb{E}\, \check{z}^{(i)}$; note that the components of $\mathbf{u}$ are independent. We can rewrite the above as:

$$\mathbf{\Omega}^* - \check{\mathbf{\Omega}} = \lambda n (\mathbf{X}_K^T \mathbf{X}_K)^{-1} \check{\mathbf{W}}_K - (\mathbf{X}_K^\top \mathbf{X}_K)^{-1} \mathbf{X}_K^\top \mathbf{u} \tag{C.6}$$

Substituting back into (C.5), we get:

$$\mathbf{W}_{-K} = \mathbf{X}_{-K}^\top \mathbf{X}_K (\mathbf{X}_K^\top \mathbf{X}_K)^{-1} \check{\mathbf{W}}_K + \frac{1}{\lambda n} \mathbf{X}_{-K}^\top (\mathbf{I} - \mathbf{X}_K (\mathbf{X}_K^\top \mathbf{X}_K)^{-1} \mathbf{X}_K^\top) \mathbf{u}. \tag{C.7}$$

We analyze the contribution of each of the two terms above separately.

Claim C.5. $\|\mathbf{X}_{-K}^\top \mathbf{X}_K(\mathbf{X}_K^\top \mathbf{X}_K)^{-1}\check{\mathbf{W}}_K\|_\infty \leqslant \beta$.

Proof. For any $j \notin K$,

$$|\mathbf{x}_j^\top \mathbf{X}_K(\mathbf{X}_K^\top \mathbf{X}_K)^{-1}\check{\mathbf{W}}_K| \leqslant \|\mathbf{x}_j^\top \mathbf{X}_K(\mathbf{X}_K^\top \mathbf{X}_K)^{-1}\|_1 \leqslant \beta$$

by Assumption 4.3. □

Claim C.6. *With high probability,* $\|\frac{1}{\lambda n}\mathbf{X}_{-K}^\top(\mathbf{I} - \mathbf{X}_K(\mathbf{X}_K^\top \mathbf{X}_K)^{-1}\mathbf{X}_K^\top)\mathbf{u}\|_\infty \leqslant \frac{1-\beta}{2}$.

Proof. For any $j \notin K$, let $\mathbf{v}_j^\top = \mathbf{x}_j^\top(\mathbf{I} - \mathbf{X}_K(\mathbf{X}_K^\top \mathbf{X}_K)^{-1}\mathbf{X}_K^\top)$. Note that $\mathbf{I} - \mathbf{X}_K(\mathbf{X}_K^\top \mathbf{X}_K)^{-1}\mathbf{X}_K^\top$ is an orthogonal projection matrix, and hence, by Assumption 4.4, $\|\mathbf{v}_j\|_2 \leqslant \|\mathbf{x}_j\|_2 \leqslant C\sqrt{n}$ and $\|\mathbf{v}_j\|_1 \leqslant Cn$.

Write $\mathbf{u} = \mathbf{z}^* - \mathbb{E}[\check{\mathbf{z}}]$. By the triangle inequality, it suffices to bound: (i) $\frac{1}{\lambda n}|\mathbf{v}_j^\top(\mathbf{z}^* - \mathbb{E}[\mathbf{z}^*])|$ and (ii) $\frac{1}{\lambda n}|\mathbf{v}_j^\top(\mathbb{E}[\mathbf{z}^*] - \mathbb{E}[\check{\mathbf{z}}])|$.

We first bound (ii).

$$\frac{1}{\lambda n}|\mathbf{v}_j^\top(\mathbb{E}[\mathbf{z}^*] - \mathbb{E}[\check{\mathbf{z}}])| \leqslant \frac{1}{\lambda n}\|\mathbf{v}_j\|_1 \|\mathbb{E}[\mathbf{z}^*] - \mathbb{E}[\check{\mathbf{z}}])\|_\infty \leqslant \frac{C}{\lambda} \cdot \max_{i\in[n]}(|\mathbb{E}[z^{*(i)}]| + |\mathbb{E}[\check{z}^{(i)}]|)$$

Since each $z^{*(i)}$ is a standard normal truncated to a set of volume at least α, Lemma 6 of [10] implies $\mathbb{E}[z^{*(i)}] = O(\sqrt{\log(1/\alpha)})$. Similarly, $\mathbb{E}[\check{z}^{(i)}] = O\left(\sqrt{\log(1/\alpha(\check{\mathbf{\Omega}}_K, \mathbf{x}_K^{(i)}))}\right)$. Invoking the fact that $\|\mathbf{\Omega}^* - \check{\mathbf{\Omega}}\|_2 \leqslant \frac{1}{C\sqrt{k}}$ and Lemma A.5, we obtain that $\alpha(\check{\mathbf{\Omega}}, \mathbf{x}^{(i)}) = \Omega(\alpha^2)$. Hence, we get:

$$\frac{1}{\lambda n}|\mathbf{v}_j^\top(\mathbb{E}[\mathbf{z}^*] - \mathbb{E}[\check{\mathbf{z}}])| \leqslant \frac{C}{\lambda} \cdot O(\sqrt{\log(1/\alpha)}) \leqslant O\left(\frac{C^2 k}{\alpha^5 \sigma_{\min}}\right) < \frac{1-\beta}{4}.$$

where we used the assumption in the Theorem.

Now, we turn to (i). Observe that $\boldsymbol{\zeta}^* = \mathbf{z}^* - \mathbb{E}[\mathbf{z}^*]$ is a zero-mean vector with independent components $\zeta^{*(1)}, \ldots, \zeta^{*(n)}$. In order to bound $\mathbf{v}_j^\top \boldsymbol{\zeta}^* = \sum_{i\in[n]} v_j^{(i)} \zeta^{*(i)}$, we use Bernstein's inequality. Fix an $i \in [n]$, and let ζ denote $\zeta^{*(i)}$. For Bernstein's Lemma, we need bounds on $\mathbb{E}[\zeta^p]$ for $p > 1$. It is easy to see that these quantities are maximized when $\zeta \sim \mathcal{N}(0, 1, S_q)$ where $S_q = \{x : x^2 \geqslant q\}$ for some q chosen such that $\mathcal{N}(0, 1, S_q) = \alpha$. Routine calculations (Lemma 13 of [11]) show that $\mathbb{E}[\zeta^2] \leqslant 2 + 2\log(2/\alpha)$ and $\mathbb{E}[\zeta^p] \leqslant p!(2 + 2\log(2/\alpha))^p$. Applying Bernstein, we get that with probability $1 - \exp(-t^2)$, $\frac{1}{\lambda n}|\mathbf{v}_j^T \boldsymbol{\zeta}^*| \leqslant O\left(\frac{Ct\log(1/\alpha)}{\lambda\sqrt{n}}\right)$ for $t \leqslant O(\sqrt{n/\log(1/\alpha)})$. Setting $t = \Omega(\sqrt{\log d})$, we get that with high probability, for all $j \notin K$:

$$\frac{1}{\lambda n}|\mathbf{v}_j^T \boldsymbol{\zeta}^*| \leqslant O\left(\frac{C\sqrt{\log d}\log(1/\alpha)}{\lambda\sqrt{n}}\right) = O\left(\frac{C^2 k\sqrt{\log d}\log(1/\alpha)}{\alpha^4 \sigma_{\min}\sqrt{n}}\right).$$

If we take $n \geqslant \Omega((C^4 k^2 \log d)/((1-\beta)^2 \sigma_{\min}^2 \alpha^9))$, then the above is less than $(1-\beta)/4$. This proves the claim. □

Bounding the ℓ_∞-error: The max-error can be bounded using (C.6).

$$\|\check{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty \leqslant \lambda \left|\left|\left|\left(\frac{\mathbf{X}_K^\top \mathbf{X}_K}{n}\right)^{-1}\right|\right|\right|_\infty + \left\|\left(\frac{\mathbf{X}_K^\top \mathbf{X}_K}{n}\right)^{-1} \mathbf{X}_K^\top \frac{\mathbf{u}}{n}\right\|_\infty$$

where for a matrix A with r rows, $|||A|||_\infty = \max_{i\in[r]} \|A^{(i)}\|_1$ is the matrix ℓ_∞-norm.

The first term is deterministic and can be bounded as follows:

$$\lambda \left|\left|\left|\left(\frac{\mathbf{X}_K^\top \mathbf{X}_K}{n}\right)^{-1}\right|\right|\right|_\infty \leqslant \lambda\sqrt{k} \cdot \lambda_{\max}\left(\left(\frac{\mathbf{X}_K^\top \mathbf{X}_K}{n}\right)^{-1}\right) \leqslant \frac{\lambda\sqrt{k}}{\sigma_{\min}} = O\left(\frac{\alpha^4}{C\sqrt{k}}\right)$$

To analyze the second term, define for $j \in K$, the vector $\mathbf{w}_j^\top = \mathbf{e}_j^\top \left(\frac{\mathbf{X}_K^\top \mathbf{X}_K}{n}\right)^{-1} \frac{\mathbf{X}_K^\top}{n}$. We need to bound $\max_{j \in K} |\mathbf{w}_j^\top \mathbf{u}|$. Note that $\|\mathbf{w}_j\|_2^2 = \frac{1}{n}\mathbf{e}_j^\top \left(\frac{\mathbf{X}_K^\top \mathbf{X}_K}{n}\right)^{-1} \mathbf{e}_j \leqslant \frac{1}{n\sigma_{\min}}$. Similar to the analysis in the proof of Claim C.6, we write $\mathbf{u}$ as $(\mathbf{z}^* - \mathbb{E}[\mathbf{z}^*]) + (\mathbb{E}[\mathbf{z}^*] - \mathbb{E}[\check{\mathbf{z}}])$. The quantity $\max_{j \in K} |\mathbf{w}_j^\top (\mathbf{z}^* - \mathbb{E}[\mathbf{z}^*])|$ can be shown using Bernstein's inequality to be at most $O\left(\frac{\sqrt{\log k}\log(1/\alpha)}{\sqrt{\sigma_{\min}}n}\right)$ with high probability. The other term can be bounded as:

$$\max_{j \in K} |\mathbf{w}_j^\top (\mathbb{E}[\mathbf{z}^*] - \mathbb{E}[\check{\mathbf{z}}])| \leqslant \max_{j \in K} \|\mathbf{w}_j\|_1 O(\sqrt{\log(1/\alpha)}) \leqslant O\left(\sqrt{\frac{\log(1/\alpha)}{\sigma_{\min}}}\right)$$

So, putting everything together:

$$\|\check{\mathbf{\Omega}} - \mathbf{\Omega}^*\|_\infty \leqslant O\left(\sqrt{\frac{\log(1/\alpha)}{\sigma_{\min}}} + \frac{\alpha^4}{C\sqrt{k}} + \frac{\sqrt{\log k}\log(1/\alpha)}{\sqrt{\sigma_{\min}}n}\right).$$

The last term is negligible because of the lower bound on n. $\square$