Linear Regression and M-estimator for Diverging p

Yajie Bao*

March 16, 2020

1 Preliminaries on random matrix and random vector

The root of high-dimensional statistics is dating back to work on random matrix theory and high-dimensional testing problems (Negahban et al. [2012]). To develop theoretical results on linear regression and M-estimator in the diverging dimension case, we need to introduce some important spectral norm concentration inequalities of random matrix. It's worthy to mention that the "High-dimensional" in this article means that

$$p = n^{\alpha}, \quad \alpha \in (0, 1).$$

1.1 Concentration inequalities on random matrix norm

For simple normal case, here we states Lemma 9 without proof in Wainwright [2009]:

Lemma 1.1 *For $k \leq n$, let $X \in \mathbb{R}^{n\times k}$ have i.i.d rows $X_i \sim N(0, \Lambda)$ and $\delta(n, k, t) := 2(\sqrt{\frac{k}{n}} + t) + (\sqrt{\frac{k}{n}} + t)^2$*

1. *If the covariance matrix Λ has maximum eigenvalue $C_{max} < \infty$, then for all $t > 0$, we have*
$$\mathbb{P}\left[\left\|\frac{1}{n}X^T X - \Lambda\right\|_2 \geq C_{\max}\delta(n, k, t)\right] \leq 2\exp\left(-nt^2/2\right). \tag{1.1}$$
2. *If the covariance matrix Λ has minimum eigenvalue $C_{min} > 0$, then for all $t > 0$, we have*
$$\mathbb{P}\left[\left\|\left(\frac{X^T X}{n}\right)^{-1} - \Lambda^{-1}\right\|_2 \geq \frac{\delta(n, k, t)}{C_{\min}}\right] \leq 2\exp\left(-nt^2/2\right). \tag{1.2}$$

Next we will generalize the concentration inequality to sub-gaussian case. Recall the operator norm or spectral norm of $m \times n$ matrix A is defined by

$$\|A\|_2 := \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x\in S^{n-1}} \|Ax\|_2,$$

which is the largest singular value of A. For symmetric matrix, the spectral norm is the largest eigenvalue.

*Department of Mathematics, Shanghai Jiao Tong University, Eamil: baoyajie2019stat@sjtu.edu.cn

Lemma 1.2 *The covering numbers of the unit Euclidean sphere S^{n-1} satisfy the following for any $\varepsilon > 0$,*

$$\mathcal{N}\left(S^{n-1}, \varepsilon\right) \leq \left(\frac{2}{\varepsilon} + 1\right)^n.$$

Lemma 1.3 *Let A be an $m \times n$ matrix and $\delta > 0$. Suppose that*

$$\left\|A^\top A - I_n\right\| \leq \max\left(\delta, \delta^2\right),$$

then

$$(1-\delta)\|x\|_2 \leq \|Ax\|_2 \leq (1+\delta)\|x\|_2 \quad \textit{for all } x \in \mathbb{R}^n.$$

Proof: W.L.O.G, let $\|x\|_2 = 1$.Using the assumption we have

$$\max\left(\delta, \delta^2\right) \geq \left|\left\langle\left(A^\top A - I_n\right)x, x\right\rangle\right| = \left|\|Ax\|_2^2 - 1\right|.$$

Applying the elementary inequality,

$$\max\left(|z-1|, |z-1|^2\right) \leq \left|z^2 - 1\right|, \quad z \geq 0$$

for $z = \|Ax\|_2$, we concluded that $\|Ax\|_2 - 1| \leq \delta$. ∎

Then we introduce the two-sided bounds on the entire spectrum of $m \times n$ matrix A (see Vershynin [2018], page 97).

Theorem 1.4 (Two-sided spectral norm bounds) *Let A be an $m \times n$ matrix whose rows A_i are independent, mean zero, sub-gaussian isotropic random vectors in $\mathbb{R}^n$. Then for any $t > 0$ we have*

$$\sqrt{m} - CK^2(\sqrt{n} + t) \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + CK^2(\sqrt{n} + t) \tag{1.3}$$

with probability at least $1 - 2\exp\left(-t^2\right)$. Here $K = \max_i \|A_i\|_{\psi_2}$.

Proof: Using Lemma 1.3, it suffices to show

$$\left\|\frac{1}{m}A^\top A - I_n\right\| \leq K^2 \max\left(\delta, \delta^2\right) \quad \text{where} \quad \delta = C\left(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}\right).$$

By Lemma 1.2, we can find an $\frac{1}{4}$−net $\mathcal{N}$ of the unit sphere S^{n-1} with cardinality $|\mathcal{N}| \leq 9^n$. Then we can evaluate operator norm on the $\mathcal{N}$,

$$\left\|\frac{1}{m}A^\top A - I_n\right\| \leq 2\max_{x\in\mathcal{N}}\left|\left\langle\left(\frac{1}{m}A^TA - I_n\right)x, x\right\rangle\right| = 2\max_{x\in\mathcal{N}}\left|\frac{1}{m}\|Ax\|_2^2 - 1\right|. \tag{1.4}$$

Let $X_i = x^T A_i$ which is indpendent sub-gaussian random variables, note that

$$\frac{1}{m}\|Ax\|_2^2 - 1 = \frac{1}{m}\sum_{i=1}^m\left[(x^TA_i)^2 - 1\right] = \frac{1}{m}\sum_{i=1}^m(X_i^2 - 1),$$

Using the fact that A_i are isotropic and $\|x\|_2 = 1$, $\|X_i\|_{\phi_2} \le K$. Then $X_i^2 - 1$ is sub-exponential random variables satisfying that $\|X_i^2 - 1\|_{\phi_1} \le CK$. By Bernstein inequality and we obtain

$$\begin{aligned}
\mathbb{P}\left\{\left|\frac{1}{m}\|Ax\|_2^2 - 1\right| \ge \frac{\varepsilon}{2}\right\} &= \mathbb{P}\left\{\left|\frac{1}{m}\sum_{i=1}^{m} X_i^2 - 1\right| \ge \frac{\varepsilon}{2}\right\} \\
&\le 2\exp\left[-c_1 \min\left(\frac{\varepsilon^2}{K^4}, \frac{\varepsilon}{K^2}\right) m\right] \\
&= 2\exp\left[-c_1 \delta^2 m\right] \\
&\le 2\exp\left[-c_1 C^2 \left(n + t^2\right)\right],
\end{aligned}$$

where the second equality follows that $\frac{\varepsilon}{K^2} = \max\left(\delta, \delta^2\right)$ and the last inequality follows that $(a+b)^2 \ge (a^2 + b^2)$. Using (1.4) we have

$$\begin{aligned}
\mathbb{P}\left(\left\|\frac{1}{m}A^\top A - I_n\right\| \ge K^2 \max\left(\delta, \delta^2\right)\right) &\le \mathbb{P}\left(2\max_{x\in\mathcal{N}}\left|\frac{1}{m}\|Ax\|_2^2 - 1\right| > K^2 \max\left(\delta, \delta^2\right)\right) \\
&\le 2 \cdot 9^n \exp\left[-c_1 C^2 \left(n + t^2\right)\right].
\end{aligned}$$

Choose sufficiently large C and the result follows. ∎

After proving this conclusion, we can apply this to covariance matrix estimation.

Theorem 1.5 *Let X be a $p-$dimensional multivariate sub-gaussian random variables with covariance matrix Σ and mean $\mathbf{0}$, and there exists $K \ge 1$ such that*

$$\|\langle X, x\rangle\|_{\psi_2} \le K x^T \Sigma x \text{ for any } x \in \mathbb{R}^p. \tag{1.5}$$

Then for sample covariance matrix $\widehat{\Sigma}_n$ we have

$$\|\Sigma_n - \Sigma\| \le C\lambda_{max}(\Sigma) K^2 \left(\sqrt{\frac{p+t^2}{n}} + \frac{p+t^2}{n}\right) \tag{1.6}$$

holds with probability at least $1 - \exp(-t^2/2)$.

Proof: Let $Z_i = \Sigma^{-1/2} X_i$, then Z_i are independent isotropic sub-gaussian random vector. Using (1.5) we have

$$\|Z_i\|_{\phi_2} = \sup_{x\in S^{p-1}} \|\langle Z_i, x\rangle\|_{\psi_2} \le K. \tag{1.7}$$

Then note that,

$$\|\Sigma_n - \Sigma\| = \left\|\Sigma^{1/2} R_n \Sigma^{1/2}\right\| \le \|R_n\| \, \|\Sigma\|,$$

where

$$R_n := \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^\top - I_p.$$

Let A be the $n \times p$ matrix with rows Z_i, then apply Theorem 1.4 we obtain that

$$\|\Sigma_n - \Sigma\| \le K^2 \|\Sigma\| \max\left(\delta, \delta^2\right)$$

holds with at least probability $1 - 2\exp(-t^2/2)$. Moreover,

$$\max\left(\delta, \delta^2\right) \le \delta + \delta^2 \le C\left(\sqrt{\frac{p+t^2}{n}} + \frac{p+t^2}{n}\right).$$

Thus the proof is completed. ∎

Remark. The theorem above implies that for low dimensional setting, i.e., $p < n$

$$\|\Sigma_n - \Sigma\| = O_p\left(\sqrt{\frac{p}{n}}\right). \tag{1.8}$$

Using the fact that

$$\left\|\Sigma_n^{-1} - \Sigma^{-1}\right\| = \Omega_p\left(\|\Sigma_n - \Sigma\|\right),$$

then if $\lambda_{min}(\Sigma) > 0$ we have

$$\left\|\Sigma_n^{-1} - \Sigma^{-1}\right\| = O_p\left(\sqrt{\frac{p}{n}}\right). \tag{1.9}$$

1.2 Concentration inequalities on random vextor norm

We start with the definitions of subGaussian random vectors and norm-subGaussian random vectors.

Definition 1.6 *A random vector $\boldsymbol{X} \in \mathbb{R}^d$ is subGaussian, if there exists $\sigma \in \mathbb{R}$ so that*

$$\mathbb{E}e^{\langle \mathbf{v}, \mathbf{X} - \mathbb{E}\mathbf{X}\rangle} \le e^{\frac{\|\mathbf{v}\|\mathbf{1}^2\sigma^2}{2}}, \quad \forall \mathbf{v} \in \mathbb{R}^d. \tag{1.10}$$

Definition 1.7 *A random vector $\boldsymbol{X} \in \mathbb{R}^d$ is norm-subGaussian ($\mathrm{nSG}(\sigma)$), if there exists $\sigma \in \mathbb{R}$ so that*

$$\mathbb{P}(\|\mathbf{X} - \mathbb{E}\mathbf{X}\| \ge t) \le 2e^{-\frac{t^2}{2\sigma^2}}, \quad \forall t \in \mathbb{R}. \tag{1.11}$$

Norm-subGaussian random vectors is proposed by Jin et al. [2019], which includes both subGaussian (with a smaller σ parameter) and bounded norm random vectors as special cases.

Lemma 1.8 *There exists absolute constant c so that following random vectors are all $\mathrm{nSG}(c\cdot\sigma)$*

1. *A bounded random vector $\mathbf{X} \in \mathbb{R}^d$ so that $\|\mathbf{X}\| \le \sigma$.*
2. *A random vector $\mathbf{X} \in \mathbb{R}^d$ where $\mathbf{X} = \xi\mathbf{e}_1$ and random variable $\xi \in \mathbb{R}$ is $\sigma-$subGaussian.*
3. *A random vector $\mathbf{X} \in \mathbb{R}^d$ that is $(\sigma/\sqrt{d})-$subGaussian.*

Theorem 1.9 (Jin et al. [2019]) *There exists an absolute constant c such that if $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ are independent zero-mean $\mathrm{nSG}(\sigma)$ random vectors. Then for any $\delta > 0$, with probability at least $1 - \delta$*

$$\left\|\sum_{i=1}^n \mathbf{X}_i\right\| \le c \cdot \sqrt{\sum_{i=1}^n \sigma_i^2 \log\frac{2d}{\delta}}. \tag{1.12}$$

From Theorem 1.9, we can obtain that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \right\| = O_p\left(\sqrt{\frac{\log d}{n}} \right).$$

And in section 2, we will prove that the random vectors $\boldsymbol{X}_i$ with sub-gaussian coordinates assumption has the following convergence rate

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \right\| = O_p\left(\sqrt{\frac{d \log d}{n}} \right).$$

In section 3, we assume that the random vectors $\boldsymbol{X}_i$ with bounded expectation of norm, i.e., $\mathbb{E}\left(\|X_i\|_2^2\right) \leq M$, which leads

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \right\| = O_p\left(\sqrt{\frac{1}{n}} \right).$$

2 Linear regression

Now consider the following linear regression model with random ensembles:

$$y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^* + e_i, \quad i = 1, 2, ..., n \tag{2.1}$$

where $e_i, i = 1, 2, ..., n$ are independent sub-gaussion random variables with mean 0 and parameter σ and $\boldsymbol{\beta}^* \in \mathbb{R}^p$. We have known that the LSE of $\boldsymbol{\beta}^*$ is

$$\widehat{\boldsymbol{\beta}} = \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} y_i \boldsymbol{X}_i \right). \tag{2.2}$$

Theorem 2.1 (Consistence) *For linear regression model (2.1), suppose that X_i are independent sub-gaussion random vectors with same mean $\mathbf{0}$ and covariance matrix Σ and X_i are independent with e_i. Assume that $\lambda_{min}(\Sigma) = \lambda_0 > 0$ and $\|X_i\|_{\psi_2} \leq K$, then*

$$\left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|_2 = O_p\left(\sqrt{\frac{p \log p}{n}} \right). \tag{2.3}$$

Proof: By (2.1),

$$\begin{aligned} \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|_2 &= \left\| \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i e_i \right) \right\|_2 \\ &= \left\| \widehat{\Sigma}_n^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i e_i \right) \right\|_2 \\ &\leq \|\widehat{\Sigma}_n^{-1} - \Sigma^{-1}\|_2 \left\| \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i e_i \right) \right\|_2 + \|\Sigma^{-1}\|_2 \left\| \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i e_i \right) \right\|_2 . \end{aligned} \tag{2.4}$$

All we need to do is bounding the term $\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i e_i\right)\right\|_2$, let $Z_{ij} = X_{ij}e_i$. Using the basic inequality $|ab| \leq \frac{a^2+b^2}{2}$ and $s^2 e^s \leq e^{2s}$, for $\eta > 0$ we have

$$\begin{aligned}\mathbb{E}\left(Z_{ij}^2 e^{\eta|Z_{ij}|}\right) &\leq \mathbb{E}\left(\eta^{-2}\exp(2\eta|Z_{ij}|)\right) \\ &\leq \eta^2\mathbb{E}\left[\exp\left(2\eta X_{ij}^2\right)\exp\left(2\eta e_i^2\right)\right] \\ &\leq \eta^2\sqrt{\mathbb{E}\left[\exp\left(2\eta X_{ij}^2\right)\right]\mathbb{E}\left[\exp\left(2\eta e_i^2\right)\right]}.\end{aligned}$$

Then by the property of sub-gaussian random variable, there exists some $M > 0$, such that

$$\mathbb{E}\left[\exp\left(2\eta X_{ij}^2\right)\right] \leq M, \mathbb{E}\left[\exp\left(2\eta e_i^2\right)\right] \leq M.$$

Next use the exponential inequality in Cai et al. [2011], we set $\bar{B}_n^2 = nM\eta^{-2}$

$$\begin{aligned}\mathbb{P}\left(\max_j^p\left|\frac{1}{n}\sum_{i=1}^{n}Z_{ij}\right| > C\sqrt{\frac{\log p}{n}}\right) &\leq \sum_{j=1}^{p}\mathrm{P}\left(\left|\sum_{i=1}^{n}Z_{ij}\right| > C\sqrt{n\log p}\right) \\ &= \sum_{j=1}^{p}\mathrm{P}\left(\sum_{i=1}^{n}|Z_{ij}| > C\bar{B}_n M^{-1}\eta\sqrt{\log p}\right) \\ &= p^{-\gamma}.\end{aligned}$$

And if we choose sufficiently large C, we can obtain that

$$\max_j^p\left|\frac{1}{n}\sum_{i=1}^{n}Z_{ij}\right| = O_p\left(\sqrt{\frac{\log p}{n}}\right).$$

The proof is completed by (2.4) and Theorem 1.5. ∎

The theorem above implies that if $p\log p = o(n)$, LSE is consistent. Next we will give the central limt theorem for LSE.

Theorem 2.2 (Asymptotic Normality) *Under the condition of Theorem 2.1, and assume that covariates $\boldsymbol{X}$ and noise e are independent. We have*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\Sigma^{-1}\right) \tag{2.5}$$

Proof: Note that,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i e_i\right). \tag{2.6}$$

By law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T \xrightarrow{p} \Sigma.$$

And using the independence, we have $\mathbb{E}\left(\boldsymbol{X}_i e_i\right) = 0$ and

$$\mathbb{E}\left(\boldsymbol{X}_i e_i\right)\left(\boldsymbol{X}_i e_i\right)^T = \sigma^2\Sigma.$$

Thus by multivariate central limt theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i e_i \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\Sigma\right)$$

Then the result follows from Slutsky's Lemma. ∎

3 M estimator

Given sample $\{X_i, i = 1, 2, ..., n\} \in \mathcal{X}_n$ is drawn independently according to some distribution $\mathbb{P}$. And in the well-specified case the distribution $\mathcal{P}$ is a member of parameterized family $\{\mathbb{P}_\theta, \theta \in \Omega\}$, where $\boldsymbol{\Omega}$ is the parameter space, then the goal is to estimate parameter $\boldsymbol{\theta}^*$. For mis-specified models, in which case the target parameter $\boldsymbol{\theta}^*$ is defined as the minimizer of the population lost function (see Wainwright [2019]).

A function $\mathcal{L}_n : \boldsymbol{\Omega} \times \mathcal{X}_n$ used to measure the goodness of estimation using sample $\boldsymbol{X}_n$, which is called *lost function*. The population lost function is defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}\left(\mathcal{L}_n\left(\boldsymbol{\theta}, \boldsymbol{X}_n\right)\right), \tag{3.1}$$

where

$$\mathcal{L}_n\left(\boldsymbol{\theta}, \boldsymbol{X}_n\right) = \frac{1}{n}\sum_{i=1}^{n} L(\boldsymbol{\theta}, X_i).$$

Next we define the *target parameter* as the minimum of the population lost function

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Omega}} \mathcal{L}(\boldsymbol{\theta}). \tag{3.2}$$

For example, the negative log-likelihood function is a lost function. Our overall estimator is based on solving the optimization problem

$$\widehat{\theta} \in \arg\min_{\theta\in\Omega}\left\{\mathcal{L}_n\left(\theta; Z_1^n\right) + \lambda_n \Phi(\theta)\right\}, \tag{3.3}$$

where $\lambda_n > 0$ is regularization parameter and $\Phi(\theta) : \boldsymbol{\Omega} \to \mathbb{R}$ is the penalty function. The estimator (3.3) is called **M estimator**, where the "M" stands for minimization (or maximization). We begin with no-penalty problem, and the following assumptions is needed to estabilish theory results, and these assumptions can be found in Zhang et al. [2013] and Jordan et al. [2019].

Assumption 3.1 (Parameter space) *The parameter space Θ is a compact and convex subset of $\mathbb{R}^p$. Moreover, $\theta^* \in \operatorname{int}(\Theta)$ and $R := \sup_{\theta\in\Theta} \|\theta - \theta^*\|_2 > 0$.*

Assumption 3.2 (Local convexity) *The lost function $L(X_i, \boldsymbol{\theta})$ is twice differentiable with respective to $\boldsymbol{\theta}$, and the Hessian matrix $I(\boldsymbol{\theta}) = \nabla^2\mathcal{L}(\boldsymbol{\theta})$ of the population lost function $\mathcal{L}(\boldsymbol{\theta})$ is invertible at $\boldsymbol{\theta}^*$. Moreover, there exists two positive constants $\mu_- < \mu_+$ such that $\mu_- I_d \preceq I(\boldsymbol{\theta}) \preceq \mu_+ I_d$.*

Assumption 3.3 (Smoothness) *There exists some positive constant (G, L) and positive integers (k_0, k_1), such that*

$$\mathbb{E}\left[\|\nabla L(\boldsymbol{\theta}, X)\|_2^{k_0}\right] \le G^{k_0}, \quad \mathbb{E}\left[\|\nabla^2 L(\boldsymbol{\theta}, X) - \nabla^2\mathcal{L}(\boldsymbol{\theta})\|_2^{k_1}\right] \le L^{k_1}. \tag{3.4}$$

Moreover, for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in U(\boldsymbol{\theta}^*, \rho)$ (a ball around the truth $\boldsymbol{\theta}^*$ with radius $\rho > 0$) there exists some positive constant M and some positive integer k_2 such that

$$\left\|\nabla^2\mathcal{L}(\boldsymbol{\theta}_1, X) - \nabla^2\mathcal{L}(\boldsymbol{\theta}_2, X)\right\|_2 \le M(X)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2, \tag{3.5}$$

and $\mathbb{E}[M(X)^{k_2}] \le M^{k_2}$.

Before bound the ℓ_2 error between the optimization solution $\widehat{\boldsymbol{\theta}}$ and ture parameter $\boldsymbol{\theta}^*$, we state the following Lemma.

Lemma 3.4 *For convex function $f(x)$, x^* is the global minimizer of $f(x)$. If for any $x \in \{x : |x - \tilde{x}|^2 = a\}$, s.t., $f(x) \geq f(\tilde{x})$, then*

$$|x^* - \tilde{x}| \leq a.$$

Proof: If there exists $x^{'}$ such that $|x^{'} - \tilde{x}|^2 > a$ and $f(x^{'}) \leq f(x^*)$. By the convexity of f, we have

$$f(\alpha x^{'} + (1-\alpha)\tilde{x}) \leq \alpha f(x^{'}) + (1-\alpha) f(\tilde{x}) < f(\tilde{x}),$$

where $0 < \alpha < 1$. Note that

$$|\alpha x^{'} + (1-\alpha)\tilde{x} - \tilde{x}| = \alpha |x^{'} - \tilde{x}|,$$

let $\alpha = |x^{'} - \tilde{x}| / |x^* - \tilde{x}|$, then $|\alpha x^{'} + (1-\alpha)\tilde{x} - \tilde{x}| = a$. But

$$f(\alpha x^{'} + (1-\alpha)\tilde{x}) < f(\tilde{x}),$$

which is a contradiction. ∎

Next we state Lemma 7 in Zhang et al. [2013] without proof as following:

Lemma 3.5 *Under Assumption 3.3, there exist some constants C_1 and C_2 (dependent only on the moments k_0 and k_1 respectively) such that*

$$\mathbb{E}\left[\|\nabla \mathcal{L}_n(\boldsymbol{\theta}^*)\|_2^{k_0}\right] \leq C_1 \frac{G^{k_0}}{n^{k_0/2}}, \tag{3.6}$$

$$\mathbb{E}\left[\left\|\nabla^2 \mathcal{L}_n(\boldsymbol{\theta}^*, X) - \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)\right\|_2^{k_1}\right] \leq C_2 \frac{\log^{k_1/2}(2p) L^{k_1}}{n^{k_1/2}}. \tag{3.7}$$

Theorem 3.6 *Under Assumption 3.2 and Assumption 3.3,*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{3.8}$$

Proof: According to Lemma 3.4, it suffices to show that for any $\boldsymbol{\theta}$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 = O\left(\frac{1}{\sqrt{n}}\right)$ such that

$$\mathcal{L}_n(\boldsymbol{\theta}) \geq \mathcal{L}_n(\boldsymbol{\theta}^*).$$

Taking Taylor expansion for $\mathcal{L}_n(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$,

$$\mathcal{L}_n(\boldsymbol{\theta}) = \mathcal{L}_n(\boldsymbol{\theta}^*) + \nabla \mathcal{L}_n(\boldsymbol{\theta}^*)^T (\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \nabla^2 \mathcal{L}_n(\tilde{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \boldsymbol{\theta}^*), \tag{3.9}$$

where $\tilde{\boldsymbol{\theta}}$ is some point between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. Define the following three events:

$$\mathcal{E}_0 := \left\{\frac{1}{n}\sum_{i=1}^{n} M(X_i) \leq 2M\right\},$$

$$\mathcal{E}_1 := \left\{\left\|\nabla^2 \mathcal{L}_n(\boldsymbol{\theta}^*, X) - \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*)\right\|_2 \leq \frac{\mu_-}{2}\right\},$$

$$\mathcal{E}_2 := \left\{\|\nabla \mathcal{L}_n(\boldsymbol{\theta}^*)\|_2 \leq \frac{C_0}{\sqrt{n}}\right\}.$$

Using Assumption 3.2, Assumption 3.3 and Markov inequality

$$\mathrm{P}\left(\mathcal{E}_0^c \cup \mathcal{E}_1^c\right) \leq \frac{C_3}{n^{k_2/2}} + \frac{C_4 \log^{k_1/2}(2p)}{n^{k_1/2}}.$$

Since $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2 = O\left(\frac{1}{\sqrt{n}}\right)$, there exists some positive constant C such that

$$\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2 = \frac{C^{'}\mu_-}{2\sqrt{n}}$$

Under event $\mathcal{E}_0 \cap \mathcal{E}_1$, we can bound $\nabla^2\mathcal{L}_n(\tilde{\boldsymbol{\theta}})$ by

$$\begin{aligned}
\lambda_{min}\left(\nabla^2\mathcal{L}_n(\tilde{\boldsymbol{\theta}})\right) &\geq \lambda_{min}\left(I(\boldsymbol{\theta}^*)\right) - \|\nabla^2\mathcal{L}_n(\boldsymbol{\theta}^*) - I(\boldsymbol{\theta}^*)\|_2 - \|\nabla^2\mathcal{L}_n(\tilde{\boldsymbol{\theta}}) - \nabla^2\mathcal{L}_n(\boldsymbol{\theta}^*)\|_2 \\
&\geq \mu_- - \frac{\mu_-}{2} - 2M\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2 \\
&= (1-\frac{2MC^{'}}{\sqrt{n}})\frac{\mu_-}{2}.
\end{aligned}$$

Using (3.6) and Jessen inequlity, we have

$$\begin{aligned}
\mathbb{E}\left[\|\nabla\mathcal{L}_n(\boldsymbol{\theta}^*)\|_2\right] = \mathbb{E}\left[\left(\|\nabla\mathcal{L}_n(\boldsymbol{\theta}^*)\|_2^{k_0}\right)^{1/k_0}\right] &\leq \left(\mathbb{E}\left[\|\nabla\mathcal{L}_n(\boldsymbol{\theta}^*)\|_2^{k_0}\right]\right)^{1/k_0} \\
&\leq \frac{C_1 G}{\sqrt{n}}.
\end{aligned}$$

Then event $\mathcal{E}_2$ happens with high probability, which follows from $O_p(Y_n) = O(\mathbb{Y}_\ltimes)$. Therefore under event $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2$ we have

$$\begin{aligned}
\mathcal{L}_n(\boldsymbol{\theta}) - \mathcal{L}_n(\boldsymbol{\theta}^*) &\geq \nabla\mathcal{L}_n(\boldsymbol{\theta}^*)^T(\boldsymbol{\theta}-\boldsymbol{\theta}^*) + (1-\frac{2MC^{'}}{\sqrt{n}})\frac{\mu_-}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 \\
&\geq -\|\nabla\mathcal{L}_n(\boldsymbol{\theta}^*)\|_2\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2 + (1-\frac{2MC^{'}}{\sqrt{n}})\frac{\mu_-}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2 \\
&\geq -\frac{C^{'}\mu_-}{2\sqrt{n}}\frac{C_0}{\sqrt{n}} + (1-\frac{2MC^{'}}{\sqrt{n}})\frac{\mu_-}{2}\frac{(C^{'}\mu_-)^2}{4n}.
\end{aligned}$$

If we choose sufficiently large $C^{'}$, $\mathcal{L}_n(\boldsymbol{\theta}) - \mathcal{L}_n(\boldsymbol{\theta}^*) \geq 0$ holds with high probability. ∎

Remark. Note that, if we substitute moment condition for gradient in (3.4) by

$$\mathbb{E}\left[\|\nabla L(\boldsymbol{\theta}, X)\|_2^{k_0}\right] \leq p^{k_0/2}G^{k_0},$$

we can obtain the new convergence rate

$$\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\| = O_p\left(\sqrt{\frac{p}{n}}\right).$$

The following asymptotic result can help us conduct statistical inference, such as interval estimation and hypothesis testing.

Theorem 3.7 *Under Assumption 3.2 and Assumption 3.3,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right) \xrightarrow{d} \mathcal{N}\left(0,\widetilde{\Sigma}\right), \tag{3.10}$$

where

$$\widetilde{\Sigma} = I(\boldsymbol{\theta}^*)^{-1}\mathbb{E}\left[\nabla L(\boldsymbol{\theta}^*,X)^T\nabla L(\boldsymbol{\theta}^*,X)\right] I(\boldsymbol{\theta}^*)^{-1}.$$

Proof: First we perform Taylor expansion for $\nabla\mathcal{L}_n(\widehat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}^*$,

$$0 = \nabla\mathcal{L}_n(\widehat{\boldsymbol{\theta}}) = \nabla\mathcal{L}_n(\boldsymbol{\theta}^*) + \nabla^2\mathcal{L}_n(\boldsymbol{\theta}^*)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right) + uO_p(\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2),$$

where $u \in \mathbb{R}^p$ is the unit vector. Then taking simple linear algebra we obtain

$$\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^* = -\nabla^2\mathcal{L}_n(\boldsymbol{\theta}^*)^{-1}\nabla\mathcal{L}_n(\boldsymbol{\theta}^*) + \frac{C}{n}\nabla^2\mathcal{L}_n(\boldsymbol{\theta}^*)^{-1}u.$$

Using law of large numbers, multivariate central limt theorem and Slutsky's lemma, we have

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right) &= \left(\frac{1}{n}\sum_{i=1}^n \nabla^2 L(\boldsymbol{\theta}^*,X_i)\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla L(\boldsymbol{\theta}^*,X_i)\right) + \frac{C}{\sqrt{n}}\nabla^2\mathcal{L}_n(\boldsymbol{\theta}^*)^{-1}u \\ &\xrightarrow{d} \mathcal{N}\left(0,\widetilde{\Sigma}\right).\end{aligned}$$

■

Remark. The following plug-in estimator is a consistent estimator for $\widetilde{\Sigma}$,

$$\left(\frac{1}{n}\sum_{i=1}^n \nabla^2 L(\widehat{\boldsymbol{\theta}},X_i)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \nabla L(\widehat{\boldsymbol{\theta}},X_i)L(\widehat{\boldsymbol{\theta}},X_i)^T\right)\left(\frac{1}{n}\sum_{i=1}^n \nabla^2 L(\widehat{\boldsymbol{\theta}},X_i)\right)^{-1} \tag{3.11}$$

More generally, by Assumption 3.3 we set $\rho \in (0,1)$, then choosing the potentially smaller radius $\delta_\rho = \min\{\rho, \rho\mu_-/4L\}$. We can define the following good events

$$\begin{aligned}\mathcal{E}_0 &:= \left\{\frac{1}{n}\sum_{i=1}^n M\left(X_i\right) \le 2M\right\}, \\ \mathcal{E}_1 &:= \left\{\left\|\nabla^2\mathcal{L}_n\left(\boldsymbol{\theta}^*,X\right) - \nabla^2\mathcal{L}\left(\boldsymbol{\theta}^*\right)\right\|_2 \le \frac{\rho\mu_-}{2}\right\}, \\ \mathcal{E}_2 &:= \left\{\|\nabla\mathcal{L}_n\left(\boldsymbol{\theta}^*\right)\|_2 \le \frac{(1-\rho)\mu_-\delta_\rho}{2}\right\}.\end{aligned}$$

The following lemma is Lemma 6 in Zhang et al. [2013].

Lemma 3.8 *Under the events $\mathcal{E}_0$, $\mathcal{E}_1$ and $\mathcal{E}_2$, we have*

$$\|\theta_1-\theta^*\|_2 \le \frac{2\left\|\nabla F_1\left(\theta^*\right)\right\|_2}{(1-\rho)\mu_-}, \quad \textit{and} \quad \nabla^2 F_1(\theta) \succeq (1-\rho)\mu_- I_{p\times p}. \tag{3.12}$$

We can assume that $\|\hat{\theta}-\theta^*\|_2 \leq R$, then make decomposition as

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\theta}-\theta^*\right\|_2^k\right] &= \mathbb{E}\left[1_{(\mathcal{E})}\left\|\hat{\theta}-\theta^*\right\|_2^k\right] + \mathbb{E}\left[1_{(\mathcal{E}^c)}\left\|\hat{\theta}-\theta^*\right\|_2^k\right] \\
&\leq \frac{2^k \mathbb{E}\left[1_{(\mathcal{E})}\left\|\nabla \mathcal{L}_n\left(\theta^*\right)\right\|_2^k\right]}{(1-\rho)^k \lambda^k} + \mathbb{P}\left(\mathcal{E}^c\right) R^k \\
&\leq \frac{2^k \mathbb{E}\left[\left\|\nabla \mathcal{L}_n\left(\theta^*\right)\right\|_2^k\right]}{(1-\rho)^k \lambda^k} + \mathbb{P}\left(\mathcal{E}^c\right) R^k.
\end{aligned}$$

Using Assumption 3.2, Assumption 3.3 and Lemma 3.4, we can prove

$$\mathbb{P}\left(\mathcal{E}^c\right) \leq C_2 \frac{1}{n^{k_2/2}} + C_1 \frac{\log^{k_1/2}(2d) H^{k_1}}{n^{k_1/2}} + C_0 \frac{G^{k_0}}{n^{k_0/2}},$$

for some universal constants $C_0,\ C_1,\ C_2$. Therefore for any $k \in \mathbb{N}$ with $k \leq \min\left\{k_0, k_1, k_2\right\}$ we have

$$\mathbb{E}\left[\left\|\theta_1-\theta^*\right\|_2^k\right] = \mathcal{O}\left(n^{-k/2} \cdot \frac{G^k}{(1-\rho)^k \lambda^k} + n^{-k_0/2} + n^{-k_1/2} + n^{-k_2/2}\right) = \mathcal{O}\left(n^{-k/2}\right). \quad (3.13)$$

We can also obtain the ℓ_2 error bound $\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2 = O_p\left(\frac{1}{\sqrt{n}}\right)$ form (3.13). There are two very useful concentration inequlities for random vector and random matrix, which is used to prove Lemma 3.5 (Lemma 7 in Zhang et al. [2013]).

Lemma 3.9 (De Acosta et al. [1981]) *Let $k \geq 2$ and X_i be a sequence of independent random vectors in a separable Banach space with norm $\|\cdot\|$ and $\mathbb{E}\left[\|X_i\|^k\right] < \infty$. There exists a finite constant C_k such that*

$$\mathbb{E}\left[\left|\left\|\sum_{i=1}^n X_i\right\| - \mathbb{E}\left[\left\|\sum_{i=1}^n X_i\right\|\right]\right|^k\right] \leq C_k\left[\left(\sum_{i=1}^n \mathbb{E}\left[\|X_i\|^2\right]\right)^{k/2} + \sum_{i=1}^n \mathbb{E}\left[\|X_i\|^k\right]\right]. \quad (3.14)$$

Lemma 3.10 (Chen et al. [2012]) *Let $X_i \in \mathbb{R}^{d\times d}$ be independent and symmetrically distributed Hermitian matrices. Then*

$$\mathbb{E}\left[\left\|\sum_{i=1}^n X_i\right\|^k\right]^{1/k} \leq \sqrt{2e\log d}\left\|\left(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right]\right)^{1/2}\right\| + 2e\log d\left(\mathbb{E}\left[\max_i \|X_i\|^k\right]\right)^{1/k}. \quad (3.15)$$

4 Newton Raphson algorithm

For optimization probelm (3.3), there are no analytic solutions usually. And Newton Raphson algorithm use iteration method to approximate solution $\widehat{\boldsymbol{\theta}}$,

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_{t-1})^{-1} \nabla \mathcal{L}_n(\boldsymbol{\theta}_{t-1}), \quad (4.1)$$

where $\eta \in (0,1)$ is step size. According to optimal condition we have

$$\begin{aligned}\boldsymbol{\theta}_t - \widehat{\boldsymbol{\theta}} &= \boldsymbol{\theta}_{t-1} - \widehat{\boldsymbol{\theta}} - \eta \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_{t-1})^{-1} \nabla \mathcal{L}_n(\boldsymbol{\theta}_{t-1}) \\ &= \boldsymbol{\theta}_{t-1} - \widehat{\boldsymbol{\theta}} - \eta \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_{t-1})^{-1} \left(\nabla \mathcal{L}_n(\boldsymbol{\theta}_{t-1}) - \nabla \mathcal{L}_n(\widehat{\boldsymbol{\theta}}) \right) \\ &= \boldsymbol{\theta}_{t-1} - \widehat{\boldsymbol{\theta}} - \eta \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_{t-1})^{-1} \nabla^2 \mathcal{L}_n(\widetilde{\boldsymbol{\theta}}) \left(\boldsymbol{\theta}_{t-1} - \widehat{\boldsymbol{\theta}} \right) \\ &= \left(I_p - \eta \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_{t-1})^{-1} \nabla^2 \mathcal{L}_n(\widetilde{\boldsymbol{\theta}}) \right) \left(\boldsymbol{\theta}_{t-1} - \widehat{\boldsymbol{\theta}} \right),\end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}$ is some point between $\boldsymbol{\theta}_{t-1}$ and $\widehat{\boldsymbol{\theta}}$. Then we obtain

$$\left\| \boldsymbol{\theta}_t - \widehat{\boldsymbol{\theta}} \right\|_2 \leq \left\| I_p - \eta \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_{t-1})^{-1} \nabla^2 \mathcal{L}_n(\widetilde{\boldsymbol{\theta}}) \right\|_2 \left\| \boldsymbol{\theta}_{t-1} - \widehat{\boldsymbol{\theta}} \right\|_2,$$

if we assume that for some positive constant c so that

$$c \leq \lambda \left(\nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_{t-1})^{-1} \nabla^2 \mathcal{L}_n(\widetilde{\boldsymbol{\theta}}) \right) \leq c^{-1}, \tag{4.2}$$

then there exists some $\rho_\eta \in (0,1)$

$$\left\| \boldsymbol{\theta}_t - \widehat{\boldsymbol{\theta}} \right\|_2 \leq \rho_\eta \left\| \boldsymbol{\theta}_{t-1} - \widehat{\boldsymbol{\theta}} \right\|_2 \leq \cdots \leq \rho_\eta^t \left\| \boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}} \right\|_2,$$

which achives exponential convergence rate. Obviously, the error of Newton update can be bounded by

$$\left\| \boldsymbol{\theta}_t - \boldsymbol{\theta}^* \right\|_2 = O\left(\rho_\eta^t a_n \right) + O_p\left(\nabla \mathcal{L}_n(\boldsymbol{\theta}^*) \right),$$

where a_n is the initial estimation error bound $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*\|_2$. Condition (4.2) is quite rigorous, and the general Newton update convergence analysis can be found in Boyd and Vandenberghe [2004]. Next we give the convergence rate of Newton method in Bubeck [2014], which requires bound of initial error.

Lemma 4.1 (Theorem 5.3, Bubeck [2014]) *Assume that f has a Lipschitz Hessian, i.e., $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M\|x-y\|$. Let x^* be local minimum of f with strictly positive Hessian, that is, $\nabla^2 f(x^*) \succeq \mu \mathrm{I}_n, \mu > 0$. Suppose that the initial starting point x_0 of Newton's method is such that*

$$\|x_0 - x^*\| \leq \frac{\mu}{2M}.$$

Then Newton's method is well-defined and converges to x^* at a quadratic rate:

$$\|x_{k+1} - x^*\| \leq \frac{M}{\mu} \|x_k - x^*\|^2. \tag{4.3}$$

Proof: First note that,

$$\nabla f(x_k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^* + s(x_k - x^*))(x_k - x^*) ds.$$

Then using $\nabla f(x^*) = 0$, we have

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - \nabla^2 f(x_k)^{-1} \nabla f(x_k) \\ &= x_k - x^* - \nabla^2 f(x_k)^{-1} \int_0^1 \nabla^2 f(x^* + s(x_k - x^*))(x_k - x^*) ds \\ &= \nabla^2 f(x_k)^{-1} \left(\nabla^2 f(x_k)(x_k - x^*) - \int_0^1 \nabla^2 f(x^* + s(x_k - x^*))(x_k - x^*) ds \right) \end{aligned}$$

By Lipschitz Hessian, we have

$$\|x_{k+1} - x^*\| \leq \left\| \nabla^2 f(x_k)^{-1} \right\|_2 \frac{M}{2} \|x_{k+1} - x^*\|^2 ,$$

then using strong convexity assumption of f in x^* and $\|x_k - x^*\| \leq \frac{\mu}{2M}$,

$$\nabla^2 f(x_k) \succeq \nabla^2 f(x^*) - M \|x_k - x^*\| \mathrm{I}_n \succeq (\mu - M \|x_k - x^*\|) \mathrm{I}_n \succeq \frac{\mu}{2} \mathrm{I}_n .$$

Then the result follows. ■

References

S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

S. Bubeck. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 15, 2014.

T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

R. Y. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference: A Journal of the IMA*, 1(1):2–20, 2012.

A. De Acosta et al. Inequalities for b-valued random vectors with applications to the strong law of large numbers. *The Annals of Probability*, 9(1):157–161, 1981.

C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

M. I. Jordan, J. D. Lee, and Y. Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, B. Yu, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

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, 2009.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Y. Zhang, J. C. Duchi, and M. J. Wainwright. Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, 14(1):3321–3363, 2013.