**Submitted to *Operations Research***
**manuscript (Please, provide the manuscript number!)**



# Interplay Between Statistical Accuracy and Running Time Cost: A Framework and Cases

Ran Chen
Laboratory for Information & Decision Systems, Massachusetts Institute of Technology, MA 02139, ran1chen@mit.edu

*Key words*: optimization, high-dimensional statistics, computational efficiency, statistical accuracy, convex optimization

## 1. Introduction

With the advent of iterative methods and the increasing scale of data, computational cost has become a great concern in addition to statistical accuracy. Approaches from different angles have been proposed, including categorizing different methods with the triple of sample size, computation time and statistical error (Chandrasekaran and Jordan 2013), computational-theoretical approach that differentiates between regions of parameters where the problem is polynomial-time computable or not polynomial-time computable (Wang et al. 2016, Berthet and Rigollet 2013), reducing the effective sample size (Shender and Lafferty 2013, Horev et al. 2015, Sussman et al. 2015, Kpotufe and Verma 2017), and separately investigating both optimization running time and statistical accuracy, when the problem enjoys good properties like a certain form of strong convexity, smoothness or isotropic property (Loh and Wainwright 2015, Wang et al. 2017, Chen and Wainwright 2015, Bottou and Bousquet 2011).

Our approach is to provide theoretically guaranteed iterative optimization algorithm and precise quantification of how iteration number affects the statistical accuracy for a class of problems that admits estimators of a certain general form without imposing artificial or hard-to-verify conditions.

Our approach is different from the computational-theoretical approach in that we quantify the affects of running time on statistical accuracy on a continuous scale rather than a binary answer of polynomial time computability.

Compared with literature that deals with only statistical problem, only statistically rooted optimization problem, or both optimization and statistical aspects of a statistical problem, our approach provide theoretically guaranteed optimization procedure; our approach provides refined optimization-wise convergence rate that considers the dimension of the statistical problem as a changing quantity rather than a constant; and our approach combines optimization and statistic in a more intrinsic way so that we do not need artificial hard-to-verify conditions to give theoretical guarantee for our optimization procedure in terms of its influence on statistical accuracy.

To further illustrate this, we digress a little into the existing works.

Existing literature usually treats statistical properties and optimization properties separately. Statistical properties (i.e. statistical convergence rate) are usually established for a perfect solution of an optimization problem. And optimization convergence rates are established targeting the perfect solution for a certain method. Literature attempting to consider both aspects jointly also follow this style.

But this separation has three undesired consequences. It requires assumptions that facilitates convergence rate in the sense of conventional optimization. It gives convergence results in the sense of conventional optimization. It deals with problems that's considered interesting in the sense of conventional optimization.

Those assumptions include strongly convex in some form for the objective function and the uniqueness of the solution, among others. However, for the original statistical problems, these assumptions are hard to verify or invalid. For example, strong convexity type condition is hard

to verify and always violated in statistical problems, and solutions to the optimization problem can be multiple in over-parametrized settings like neural network and robust Principle Component Analysis (RPCA).

One of our key observations is that these assumptions are not necessary for producing statistically well behaved computed estimators, as we do not need to solve the optimization problem well in the conventional way to guarantee its statistical performance — there is an alternative way of characterizing how well the optimization problem is solved in terms of solution's statistical performance. Further, solving it well in the conventional way does not give additional help to statistical analysis.

The convergence results in the sense of conventional optimization are also not enough for statistical consideration. In high dimensional statistics, we are essentially dealing with a class of optimization problems with changing dimensions, and we need to know how optimization-induced statistical error changes in terms of both iteration number and the dimension. Conventional optimization results usually view dimension related quantity as a constant intrinsic to the optimization problem.

Many statistically rooted optimization problems are not considered general enough or interesting enough under conventional optimization sense, but the statistical problems are important from statistical perspective. Therefore, many heuristic optimization methods widely used in statistical literature are not nearly well understood. Many statistically good estimators also lack optimization algorithms. And some optimization results targeting statistically rooted optimization problem generalize the problem in the way making it no longer useful for the root statistical problem.

Our approach is free from all these problems. We propose a framework consists of three parts. We incorporate the consideration of optimization error into the statistical analysis through an approximate optimization problem rather than an approximate optimization solution. We provide a template optimization algorithm. We show its convergence in terms of converging to the optimization problem. Our convergence results takes the possibly growing dimension and other changing geometry quantities into consideration in addition to the iteration number. All three added

together, we have a theoretically guaranteed algorithm and a precise quantification of statistical accuracy given iteration number.

In two examples, 1-bit matrix completion (Davenport et al. 2014) and causal inference for panel data (Athey et al. 2021), we apply our framework, which yields novel results for both problems. And our framework can also be applied to network analysis, robust principle analysis, kernel ridge regression, SVM, simple neural networks, LASSO, etc. We take LASSO for an example. LASSO in (high dimensional sparse) linear regression is a simpler and degenerate case for our framework. Through it, we show that our framework automatically adapt to the setting where stronger assumptions are satisfied (e.g. restricted strong convexity).

In addition to our framework, our statistical analysis of causal inference for panel data using matrix completion is also sharper and yields better statistical convergence rate in the special case that the solution is perfect, which is the case considered in the literature.

### 1.1. Our framework

Our framework deals with statistical problems where the most promising estimator can be written as a solution to an optimization problem of the form

$$\begin{aligned} \min_{X} \quad & f(X) + g(X) \\ \text{s.t.} \quad & X \in C_1 \cap C_2 \cap \cdots C_J, \end{aligned} \tag{1}$$

where $X$ is an $m \times n$ parameter matrix, with vector being a special case by taking $n = 1$, $f$ is an $L(\epsilon)$-smooth (optimization wise) and $L_f(\epsilon)$-Lipschitz convex function on the constraint set and its $\epsilon$ neighborhood (with $L(\epsilon), L_f(\epsilon) > 0$), $g$ is a possibly non-smooth but $L_g(\epsilon)$-Lipschitz convex function on the same area (with $L_g(\epsilon) > 0$), $C_1$ to $C_J$ are convex constraint sets that are easy to project on. Note that $f$ and $g$ here are usually data dependent.

In some cases $f$ is data dependent. Examples include negative log likelihood, sum of least squares in high-dimensional linear regression, or the objective function in principle component analysis (PCA). In these cases $g$ can be penalty term or 0. In some cases, $g$ is data-dependent and $f$ is the

regularization term. Examples include soft support vector machine and neural network with Relu activation function.

So this general form includes a wide range of estimators, including constrained maximum log likelihood estimators, penalized maximum log likelihood estimator, support vector machine, etc.. This wide range of estimators have proved their power by achieving minimax optimality for many statistical problems, especially in high dimensional statistics, or by achieving good empirical performances, especially in machine learning.

Note that we do not require strong convexity, restricted strong convexity or strong convexity of any form for $f(X)$, which is almost a conventional assumption in the literature considering both optimization and statistical properties. We will see later that the absence of strong convexity is indeed very common in reality.

A specific example fitting this general form is the 1-bit matrix completion with constrained maximum log likelihood estimators. It's helpful to see how this concrete example fits the general framework.

EXAMPLE 1 (1-BIT MATRIX COMPLETION). The statistical setting for 1-bit matrix completion is as follows (Davenport et al. 2014). Given the true parameter matrix $M \in \mathbb{R}^{d_1 \times d_2}$, a random subset of indices $\Omega \subset [d_1] \times [d_2]$ indicating the elements we observe, and a differentiable link function $l : \mathcal{D} \to [0,1]$, where $\mathcal{D} \subset \mathbb{R}$, the observation is a matrix $Y \in \mathbb{R}^{d_1 \times d_2}$ defined as follows. Entries of $Y$ are independent.

For $(i,j) \in \Omega$,

$$Y_{i,j} = \begin{cases} +1 \text{ with probability } l(M_{i,j}) \\ -1 \text{ with probability } 1 - l(M_{i,j}) \end{cases}. \tag{2}$$

For $(i,j) \notin \Omega$, $Y_{i,j} = 0$. The assumptions are as follows. $M$ is nuclear norm bounded ( $\|M\|_* \leq \alpha\sqrt{rd_1d_2}$) and element wise bounded ( $\|M\|_\infty \leq \alpha$). The random subset of indices satisfies $\mathbb{E}|\Omega| = n$ with each entry being chosen with probability $\frac{n}{d_1 \times d_2}$ independently.

Then the log-likelihood function of this problem is

$$\mathcal{L}_{\Omega,Y}(X) = \sum_{(i,j)\in\Omega} (\mathbb{1}\{Y_{i,j} = 1\} \log(l(X_{i,j})) + \mathbb{1}\{Y_{i,j} = -1\} \log(1 - l(X_{i,j}))). \tag{3}$$

Davenport et al. (2014) show that the minimax optimal estimator $\hat{M}$ is a solution of the following optimization problem

$$\begin{aligned} \min_{X} \quad & -\mathcal{L}_{\Omega,Y}(X) \\ s.t. \quad & \|X\|_* \le \alpha\sqrt{rd_1d_2} \text{ and } \|X\|_\infty \le \alpha. \end{aligned} \tag{4}$$

If we further assume twice differentiability of the link function, which is true for all link function examples in Davenport et al. (2014), this estimator satisfies our general formulation (1), with

$$\begin{aligned} & f(X) = -\mathcal{L}_{\Omega,Y}(X), \qquad g(X) = 0, \\ & C_1 = [-\alpha,\alpha]^{d_1\times d_2}, \qquad C_2 = \{M \in \mathbb{R}^{d_1\times d_2} | \|M\|_* \le \alpha\sqrt{rd_1d_2}\}, \\ & L_f(\epsilon) = \sup_{|x|\le \epsilon+\alpha} \frac{|l'(x)|}{l(x)(1-l(x))}, \qquad L_g(\epsilon) = 0, \text{ and} \\ & L(\epsilon) = \sup_{|x|\le \epsilon+\alpha} \max\{\frac{|l''(x)l(x)-(l'(x))^2|}{l(x)^2}, \frac{|l''(x)(1-l(x))+(l'(x))^2|}{(1-l(x))^2}\}. \end{aligned} \tag{5}$$

Remark 1. Note that in Example 1, $-\mathcal{L}_{\Omega,Y}(X)$ in most cases is not strongly convex, or restricted strongly convex (Loh and Wainwright 2015, Wang et al. 2017), hence the approach of establishing convergence in parameter space (the space of $X$) for the optimization problem separated from the statistical problem, which is adopted in the literature, is not a good, if possible, approach.

Remark 2. In the original work by Davenport et al. (2014), where Example 1 arises, they only have a heuristic algorithm computing the solution of optimization problem (4) with no theoretical guarantee.

Remark 3. Causal inference for panel data (Athey et al. 2021) also satisfies the general formulation (1). We discuss it in detail in Section 4, where we not only develop a theoretically guaranteed optimization algorithm and provide a precise quantification of how iteration number comes in the statistical accuracy based on our framework, but also give a sharper upper bound on statistical accuracy than that in Athey et al. (2021) when the solution is exact, all of which are interesting results on their own.

Remark 4. Lasso for linear regression is another example satisfying our framework. But it is a severely degenerate case: it is for parameter vector; it does not have constraints; it admits restricted strong convexity in high dimensional sparse setting. We discuss it in detail in Section 5.

REMARK 5. More examples fit into our framework. For the reason of space, we do not give detailed discussion in this dissertation.

In our framework, to be free from strong convexity of any form or other artificial conditions, we consider $\tilde{X}$ that has small violations on both constraints and minimum objective function value. We analyze statistical property of $\tilde{X}$. The analysis of $\tilde{X}$ is independent from any optimization procedure and it is an interface between statistical property and optimization error, so we call this step *Statistical-Optimization Interplay*. Then we develop an optimization algorithm and analyze its convergence in terms of those small vanishing violations. Therefore, we can give a precise quantification of how number of iterations in our algorithm translates into statistical accuracy. Given that the number of iteration is the key bottleneck for running time and can not be reduced through parallel computing, this shows how running time could buy statistical accuracy until the statistical limit is reached.

**1.1.1. Statistical-Optimization Interplay** The first step of our framework is to integrate the optimization error into statistical analysis before solving the optimization problem.

Given the data, functions $f$ and $g$ in optimization problem (1) are decided. The target estimator $X^*$ is a solution to the optimization problem (1). But the exact solution of optimization problem (1) can be computationally infeasible and only approximate solutions can be computed. We need to consider the statistical property of this approximate solution.

Instead of considering the convergence rate of the computed solution $\tilde{X}$ to the target estimator $X^*$, we move the consideration of optimization to the start of statistical analysis. We consider an approximate estimator $\tilde{X}$ satisfying the approximate conditions in (6) and investigate its statistical properties. Basically, approximate conditions mean both constraints and optimal objective function can be violated a little ($\delta, \delta_0, \delta_1, \cdots, \delta_J \geq 0$).

$$\begin{aligned} & f(\tilde{X}) + g(\tilde{X}) \leq f(X^*) + g(X^*) + \delta, \\ & \inf_{Z \in C_i} \|Z - \tilde{X}\|_2 \leq \delta_i, \text{ for all } 1 \leq i \leq J, \\ & \inf_{Z \in C_1 \cap C_2 \cap \cdots C_J} \|Z - \tilde{X}\|_2 \leq \delta_0. \end{aligned} \tag{6}$$

Note that the target estimator $X^*$, the optimizer of Optimization Problem (1), satisfies these inequalities with $\delta = \delta_0 = \cdots = \delta_J = 0$. When $\delta, \delta_0, \cdots, \delta_J \to 0^+$, the approximate conditions are infinitely close to the original Optimization Problem (1), and when $\delta = \delta_0 = \cdots = \delta_J = 0$, the approximate conditions define an equivalent optimization problem as the original one. So this is a way of characterizing how close the computed estimator $\tilde{X}$ is to the target estimator $X^*$. An interesting observation is that the statistical analysis of, or the tools used in the statistical analysis of most constrained $M-$estimators, a kind of estimators satisfying the conditions of our framework, can be carried to this approximate version estimator relatively easily. We concrete the idea in three examples, 1-bit matrix completion, causal panel data analysis and LASSO. 1-bit matrix completion problem is analyzed as a representation for constrained log-likelihood estimator. Causal panel data is analyzed as a representation for constrained penalized log-likelihood estimator. Lasso is a representation of a degenerate case for our framework, where we show that the statistical-optimization interplay automatically adapt to simpler settings to give strong results in the simpler setting. For causal panel data, we also sharpened the backbone statistical analysis. And our framework is applied to the sharpened statistical analysis.

Note that in this step, we do not yet have an optimization procedure and the analysis is entirely irrelevant to the optimization procedure. Yet the slightly violated conditions fully characterize the statistical property of computed solution $\tilde{X}$ in the sense that non-violated version conditions are the starting point for any statistical analysis for the exact solution. So we do not need the optimization procedure to have a traditional optimization convergence.

Existing work on considering both optimization error and statistical error (e.g. Bottou and Bousquet (2011), Loh and Wainwright (2015)) usually considers the optimization error after the statistical problem is fully analyzed. They consider the optimization wise convergence rate of the computed solution to the true solution. But this approach does not work when the true solution is hard or unable to computed well. One of such setting is when the optimization problem has multiple solutions. Examples include neural network, which is usually over-parametrized, and principal

component analysis (PCA). People deals with the problem of multiple solution in PCA through defining a distance that implicitly equalize the solutions, partly leading to a huge volume of literature on non-convex optimization, see Chi et al. (2019) for a review. Another situation that the true solution is hard to be computed well is when the objective function does not enjoy good properties in the sense of optimization, e.g. strong convexity of some form.

**1.1.2. Optimization Algorithm and Convergence Analysis** The second step is to develop an optimization procedure with theoretical guarantees in terms of convergence to an estimator satisfying inequalities (6).

We adopt a double-loop optimization procedure where the outer loop is proximal gradient descent and the inner loop is 3-block ADMM.

We give convergence rate of the optimization procedure that considers both iteration number and statistically important quantity (e.g. dimension). This includes the convergence rate for inexact proximal gradient descent, convergence rate for our inner loop (3-block ADMM), and a bound for a dimension-related geometric quantity involved in the convergence rate.

There can be variants to our optimization procedure ( e.g. using accelerated proximal gradient descent for outer loop, using 2-block ADMM for inner loop when reducible). But our analysis for outer loop can be easily carried to accelerated version. And our analysis of the geometric quantity can also be easily carried to 2-block ADMM. Another reason for taking 3-block ADMM is that in addition to fitting our two examples, the 3-block ADMM can serve as a building block for a general number of constraints, as is in our general framework.

## 1.2. Related Literature

In addition to the literature mentioned at the beginning of this sections. The problems considered in this paper is also connected to the following problems and literature.

Computational issue for low-rank matrix completion has been studied through a matrix factorization approach which leads to nonconvex optimization problem. See, for example, Wang et al.

(2017), Jain et al. (2013), Chen and Wainwright (2015), and the overview paper, Chi et al. (2019). In this line of research, 1-bit matrix completion problem is also correctly studied by Chen and Wainwright (2015). However, this approach requires the exact low rank assumption, the knowledge of the rank, and also at least one other conditions like RIP condition (Jain et al. 2013), restricted convexity (Wang et al. 2017), and incoherence Chen and Wainwright (2015), which are strong and hard-to-verify condition in many settings. Further, the convergence rate for 1-bit matrix problem in Chen and Wainwright (2015) depends on the mostly unknown incoherence, which varies greatly, and the worst case different from the best by order.

Computational issue for $M-$estimator is also considered in Loh and Wainwright (2015), where they consider Lasso type estimator. Their work deal with vectors (instead of matrices) with restricted strong convexity (RSC) requirement. Our framework is designed for the more general case: matrix without RSC condition. This includes the simpler setting (vector with RSC condition). And as shown in our third example, our framework automatically adapts to the simpler setting and provides stronger results under stronger conditions.

Schmidt et al. (2011) studied convergence rate for inexact proximal gradient and inexact accelerated proximal gradient when the non-smooth part is finite. But in our setting, the existence of constraints dictates the infinity of the non-smooth part. Jiang et al. (2012) studied inexact accelerated proximal gradient descent, but it is for linearly constrained convex SDP.

Literature on the convergence of 3-block ADMM includes, for example, Cai et al. (2017), Lin et al. (2018), Hong and Luo (2017), Lin et al. (2016). But they either are not applicable to our setting (Hong and Luo 2017, Lin et al. 2018), or establishes convergence rate on Lagrange Functions (Cai et al. 2017), or establishes convergence rate on objective function with results weaker than ours in its applicable setting (Lin et al. 2016). Tibshirani (2017) considers projection on intersection of convex sets, but it is for coordinate descent and for vectors instead of matrices, thus not applicable to our setting.

### 1.3. Organization of the Chapter

In Section 2 we introduce our general framework and give general results. In Section 3 we discuss the results of applying our framework to 1 bit matrix completion example, where we get interesting new results. In section 4 we discuss the results for causal panel data example, where in addition to applying our framework we provide tighter back-bone statistical analysis. In Section 5, we discuss applying our framework to (high dimensional) linear regression and compare with the results in literature for this degenerated setting. In section 6, we discuss some directions for future work. For the reason of space, the proofs are given in the appendix Section EC.1.

### 1.4. Notation and Definition

Both $\|\cdot\|$ and $\|\cdot\|_F$ stand for Frobeneous norm. $\|\cdot\|_F$ is to give special emphasis for matrices when there might be confusion. $\|\cdot\|_*$ stands for nuclear norm. We use $|\mathcal{O}|$ to denote the number of elements in $\mathcal{O}$ when $\mathcal{O}$ is a set. We use $D(A\|B) = \frac{1}{d_1 d_2}\sum_{i,j} D(A_{i,j}\|B_{i,j})$ to denote average KL divergence between $d_1$ by $d_2$ probability matrix A and B for 1-bit matrix completion, where $D(a\|b) = a\log(\frac{a}{b}) + (1-a)\log(\frac{1-a}{1-b})$. We use $\mathfrak{T}\{A\}$ to denote the function where it takes 0 if $A$ holds and $\infty$ if $A$ does not hold. We use $\mathcal{R}(\varepsilon, C)$ to denote the $\varepsilon$ neighborhood of convex set C: $\mathcal{R}(\varepsilon, C) = \{X : \inf_{Z\in C}\|X - Z\| \le \varepsilon\}$. We use $B_d(x)$ to denote a ball centered at x with radius d under Frobeneous norm. We use $\vee$ to denote taking max: $a \vee b = \max\{a, b\}$. We use $\text{Proj}_C(P)$ to denote the projection point of $P$ on convex set $C$, the projection is in terms of Euclidean distance.

Now we introduce the definition of smoothness in optimization sense.

Definition 1 (Optimization-wise Smoothness). A convex function f(X) is said to be $L$-smooth if for any $X$ in the domain, there is a subgradient $\partial f(X)$ at $X$ such that for all $Y$ in the domain,

$$f(Y) \le f(X) + \partial f(X)(Y - X) + \frac{L}{2}\|X - Y\|^2. \tag{7}$$

## 2. General Framework

In this section, we introduce the general framework. The general framework has three parts: statistical-optimization interplay, optimization-template algorithm, and optimization convergence analysis.

### 2.1. Statistical-Optimization Interplay

In statistical-optimization interplay, we integrate the optimization consideration into the statistical analysis by considering the statistical accuracy of an estimator coming from an approximate optimization problem instead of just an approximate solution.

Recall that the target estimator $X^*$ is the solution in (1). To consider the optimization-induced statistical error, we consider the statistical property of approximate estimator $\tilde{X}$ satisfying Inequalities (6). The measurements for how well the optimization problem is eventually solved are $\delta, \delta_0, \delta_1, \cdots, \delta_J$.

Suppose one of the true parameters of the statistical model is $X_t$.

The key ingredient for statistical-optimization interplay is an interesting but natural observation on statistical analysis of estimator of the form (1). The statistical analysis for $X^*$ usually starts with the inequality

$$f(X^*) + g(X^*) \leq f(X_t) + g(X_t). \tag{8}$$

This inequality is usually reduced to simpler form with or without using the constraint conditions. And then the simpler form becomes a solvable inequality for the statistical error or the simpler form is further reduced. Typical tools for further reducing the inequality includes empirical process, which is also where the constraints in (1) usually comes in.

A reflection on this whole procedure gives that the additive nature of (8) is untouched, so are the constraints in (1).

These characteristics of the analysis mean that similar analysis can go through for approximate solution $\tilde{X}$, as it adds in the optimization errors (e.g.$\delta, \delta_0, \cdots, \delta_j$) in an additive way. Specifically, the analysis for $\tilde{X}$ starts with

$$f(\tilde{X}) + g(\tilde{X}) \leq f(X_t) + g(X_t) + \delta. \tag{9}$$

Constraints also enter the analysis with an additional error term.

In this way, the focus is shifted from the final approximate solution $\tilde{X}$ to the approximate optimization problem (6). We do not need strong convexity or uniqueness of the solution or other

conditions to ensure the fast proximity of the solutions. We only need proximity of the problems, which is the only thing relevant to the statistical accuracy while being much relaxed in terms of optimization.

As statistical analysis varies from problem to problem. We will concrete the idea of analyzing solution satisfying the approximate optimization problem through examples in Section 3, Section 4 and Section 5.

REMARK 6. In our framework, we consider problems with constraints, but it is also applicable to the setting where there is no constraints. The problem with no constraints is a degenerated case where we do not need to consider projection in optimization part. We show in Section 5 that in a degenerate case, (high dimensional) linear regression, our framework automatically adapts to the simpler setting and stronger conditions to give stronger results.

REMARK 7. Statistical-Optimization Interplay, the interface building optimization error into statistical analysis before solving the optimization problem, can work alone. That is, the optimization procedures and analysis can be replaced when needed. Further, the statistical-optimization interplay can also be extended to Z-estimators and other type of estimators coming from equation/inequality system, which is in my future work.

## 2.2. Template Algorithm

The second step of the framework is to have an algorithm finding $\tilde{X}$ satisfying (6). Our template algorithm is a double-loop algorithm, where the outer loop is inexact proximal gradient descent and inner loop is a 3-block ADMM to approximately compute quantities in the outer loop. Our inner loop algorithm can be replaced and generalized to fit arbitrary number of constraints, but to avoid unnecessary complexity while being sufficient for our examples, we elaborate on 3-block ADMM and remark on generalized algorithm.

**2.2.1. Outer Loop** Note that optimization problem (1) is equivalent to minimizing the following function.

$$F(X) = f(X) + \left(g(X) + \mathfrak{T}\{X \in C_1\} + \mathfrak{T}\{X \in C_2\} + \cdots + \mathfrak{T}\{X \in C_J\}\right). \tag{10}$$

To minimize $F(X)$, we do proximal gradient descent but with an "approximate" proximal step, as shown in algorithm 2.1.

Outer Loop: Inexact Proximal Gradient Descent Starting point is $X_0 \in C_1 \cap C_2 \cap \cdots \cap C_J$. Step size is $\eta > 0$. For $k \geq 0$,

$$X_{k+0.5} = X_k - \eta \nabla f(X_k), \qquad X_{k+1} = \widetilde{\text{Prox}}_{\eta(g(X)+\mathfrak{T}\{C_1\cap C_2\cap\cdots\cap C_J\})}(X_{k+0.5}), \tag{11}$$

where $\widetilde{\text{Prox}}_{\eta(g(X)+\mathfrak{T}\{X\in C_1\cap C_2\cap\cdots\cap C_J\})}(X_{k+0.5})$ is a close approximation of

$$\begin{aligned} &\text{Prox}_{\eta(g(X)+\mathfrak{T}\{X\in C_1\cap C_2\cap\cdots\cap C_J\})}(X_{k+0.5}) = \\ &\arg\min_X \left(\frac{1}{2}\|X - X_{k+0.5}\|_F^2 + \eta\Big(g(X) + \mathfrak{T}\{X \in C_1 \cap C_2 \cap \cdots \cap C_J\}\Big)\right). \end{aligned} \tag{12}$$

$\text{Prox}_{\eta(g(X)+\mathfrak{T}\{X\in C_1\cap C_2\cap\cdots\cap C_J\})}(\cdot)$ is called a *proximal operator*. However, we do not have an exact solution to (12) to give $\text{Prox}_{\eta(g(X)+\mathfrak{T}\{X\in C_1\cap C_2\cap\cdots\cap C_J\})}(X_{k+0.5})$. We only have an approximate proximal $\widetilde{\text{Prox}}_{\eta(g(X)+\mathfrak{T}\{C_1\cap C_2\cap\cdots\cap C_J\})}(X_{k+0.5})$ in the outer loop by approximately solving the optimization problem corresponding to it, which is our inner loop.

Before we proceed to inner loop, we conclude with a remark that the approximate proximal gradient can be replaced by its accelerated version for outer loop. But given the commonly seen phenomenon that accelerated version of algorithms are usually less robust to errors along the computation, we do not discuss the accelerated version for our setting. Similar discussion can be given for accelerated version.

**2.2.2. Inner Loop** The optimization problem that inner loop aims to solve is

$$\min_X \left(\frac{1}{2}\|X - X_{k+0.5}\|_F^2 + \eta\Big(g(X) + \mathfrak{T}\{X \in C_1 \cap C_2 \cap \cdots \cap C_J\}\Big)\right). \tag{13}$$

We can write it as

$$\min_P \left(\|P - P_0\|_F^2 + \Big(h_1(P) + h_2(P) + \cdots + h_m(P)\Big)\right), \tag{14}$$

where $P_0$ equals to $X_{k+0.5}$ in (13), and $h_i(\cdot)$ are convex functions not necessarily smooth and potentially take infinity value. In the case $J \geq 1$, at least one $h_i(\cdot)$ takes infinity value.

We first consider the case that $m=2$. In this case, Optimization Problem (14) is equivalent to the following problem:

$$\begin{aligned} &\min_{W,Z,P} \|P-P_0\|_F^2 + h_1(W) + h_2(Z), \\ &s.t. W = P, Z = P. \end{aligned} \tag{15}$$

We take 3-block ADMM to solve this problem. The Augmented Lagrange Function for this 3-block ADMM is

$$\mathcal{L}_\beta(W,Z,P,\Lambda_1,\Lambda_2) = \|P-P_0\|_F^2 + h_1(W) + h_2(Z) + \frac{\beta}{2}(\|W-P+\frac{\Lambda_1}{\beta}\|_F^2 + \|Z-P+\frac{\Lambda_2}{\beta}\|_F^2), \tag{16}$$

where $\beta>0$ is the dual step size and $\mathbf{\Lambda}=(\Lambda_1,\Lambda_2)$ is the dual variable.

The optimization procedure for this 3 block ADMM is in algorithm 2.2.

INNER LOOP: 3 BLOCK ADMM The starting points are $P^0=P_0$, $\Lambda_1^0=\mathbf{0}$, $\Lambda_2^0=\mathbf{0}$. The dual step size is $\beta>0$. For $k\geq 0$, the iteration steps are

$$\begin{aligned} W^{k+1} &= \arg\min_W \mathcal{L}_\beta(W,Z^k,P^k,\Lambda_1^k,\Lambda_2^k) = \arg\min_W h_1(W) + \frac{\beta}{2}\|W-P^k+\frac{\Lambda_1^k}{\beta}\|_F^2, \\ Z^{k+1} &= \arg\min_Z \mathcal{L}_\beta(W^k,Z,P^k,\Lambda_1^k,\Lambda_2^k) = \arg\min_Z h_2(Z) + \frac{\beta}{2}\|Z-P^k+\frac{\Lambda_2^k}{\beta}\|_F^2, \\ P^{k+1} &= \arg\min_P \mathcal{L}_\beta(W^k,Z^k,P,\Lambda_1^k,\Lambda_2^k) \\ &= \arg\min_P \|P-P_0\|_F^2 + \frac{\beta}{2}(\|W^{k+1}-P+\frac{\Lambda_1^k}{\beta}\|_F^2 + \|Z^{k+1}-P+\frac{\Lambda_2^k}{\beta}\|_F^2), \\ \Lambda_1^{k+1} &= \Lambda_1^k + \beta(W^{k+1}-P^{k+1}), \\ \Lambda_2^{k+1} &= \Lambda_2^k + \beta(Z^{k+1}-P^{k+1}). \end{aligned} \tag{17}$$

Note that when $h_1(\cdot)$ comes from a constraint function, the update step for $W$ is a projection step. Analogous result holds for $h_2(\cdot)$.

Usually, 3-block ADMM is enough for solving most of the problems encountered in statistics, including our two examples, as $m$ in (14) is usually not very large. In the case that 3-block ADMM is not enough (i.e. $m\geq 2$), the reason for $m\geq 2$ is that the number of constraints is large. Then a natural way is to do recursive ADMM. For example, if we have $g=0$ and four constraints $C_1$, $C_2$, $C_3$, $C_4$, we can do a 3-block ADMM for $\arg\min_P \|P-P_0\|_F^2 + \mathfrak{T}\{P\in C_1\cap C_2\} + \mathfrak{T}\{P\in C_3\cap C_4\}$,

where in each projection step, say on $C_1 \cap C_2$, we can do another 3-block ADMM. This could be costly, but do-able.

Another remark is that in some cases, Optimization Problem (14) can be reduced to 2-block ADMM. But less blocks sometimes may lead to worse performance (Lin et al. 2018) and it's not generalizable to more blocks, we rest with 3-block ADMM.

### 2.3. Optimization Convergence Analysis

In this section we give theoretical analysis for the algorithm-template we introduced in Section 2.2.

For outer loop, we have the convergence result for inexact proximal gradient descent in Theorem 1.

THEOREM 1 **(Inexact Proximal Gradient Descent)**. *We take the inexact proximal gradient descent algorithm 2.1. Suppose the inner loop (approximation of the proximal) satisfies*

$$\left|\widetilde{Prox}_{\eta(g(x)+\mathfrak{T}\{x\in C_1\cap C_2\cap\cdots\cap C_J\})}(X) - Prox_{\eta(g(x)+\mathfrak{T}\{x\in C_1\cap C_2\cap\cdots\cap C_J\})}(X)\right| \leq \delta_0 \tag{18}$$

*for all $X \in \mathcal{R}(\delta_0, C_1\cap C_2\cap\cdots\cap C_J)$. Suppose on $\mathcal{R}(\delta_0, C_1\cap C_2\cap\cdots\cap C_J)$, $f$ is $L$ smooth and $L_f$ Lipschitz,and $g$ is $L_g$ Lipschitz. We let step size $\eta \leq \frac{1}{L}$. Suppose $\tilde{X}$ has the smallest $f(X)+g(X)$ value among $X_0, X_1, \cdots, X_K$, the starting point and the results of first $K$ iterations. Then we have*

$$\begin{aligned} f(\tilde{X}) + g(\tilde{X}) - f(X^*) - g(X^*) \leq & \frac{1}{2K\eta}\|X_0 - X^*\|^2 + \\ & (L_f + L_g)\delta_0 + \frac{L}{2}\delta_0^2 + \frac{\delta_0 D}{\eta} + \frac{\delta_0^2}{2\eta}, \end{aligned} \tag{19}$$

*where $D$ is the diameter of $C_1\cap C_2\cap\cdots\cap C_J$.*

REMARK 8. Schmidt et al. (2011) studied the convergence of inexact proximal gradient descent when the non-smooth part is finite. But in the presence of constraints, although function $g$ is finite, the optimization problem (14) in our setting is always infinite.

REMARK 9. The Lipschitz conditions needed for $f$ and $g$ are natural conditions satisfied in most setting. For most non-smooth penalties, $g$ satisfies Lipschitz condition on the entire space. For most problems, the constraint set is compact (or contained in a compact set), thus the smoothness and convexity of f dictates Lipschitz condition.

Now we turn to the convergence of the inner loop, 3-block ADMM.

Let $W^*, Z^*, P^*$ be true primal variables and $\mathbf{\Lambda}^* = (\Lambda_1, \Lambda_2)$ be the true dual variable, i.e. solution to the optimization problem

$$\max_{\Lambda_1,\Lambda_2} \min_{W,Z,P} \mathcal{L}_\beta(W, Z, P, \Lambda_1, \Lambda_2).$$

We have Proposition 2.1 for the convergence rate of the 3-block ADMM.

3 BLOCK ADMM CONVERGENCE RATE Suppose we take algorithm 2.2 with dual step size $\beta \leq \frac{6}{17}$, suppose $\overline{P}^t = \frac{1}{t}\sum_{j=1}^t P^j$, then we have

$$\|\overline{P}^t - P^*\|^2 \leq \frac{1}{2\beta t}\left(\beta^2\|Z^1 - P^*\|^2 + 2\beta^2\|P^1 - P^*\|^2 + \|\mathbf{\Lambda}^1 - \mathbf{\Lambda}^*\|^2 + \frac{20}{3}\beta^2\|P^1 - P_0\|^2\right). \tag{20}$$

REMARK 10. In general, convergence for multi-block ADMM with more than two blocks does not hold (Chen et al. 2016). Convergence in some specific settings has been studied. But to our knowledge, no convergence rate has been established for direct 3-block ADMM applied to our setting. In the most closely related literature, Cai et al. (2017) does not have convergence rate; the requirement on constraints in Lin et al. (2018) or Hong and Luo (2017) does not fit our setting; Lin et al. (2016) has strict requirement on dual step size and slower rate based on their requirement.

Note that $\|\mathbf{\Lambda}^1 - \mathbf{\Lambda}^*\|$ is involved in the convergence rate. $\mathbf{\Lambda}^1$ depends explicitly on $\beta$, $P_0$, $h_1(\cdot)$ and $h_2(\cdot)$, which can usually be easily studied and bounded, and it's usually relatively small in our setting. $\mathbf{\Lambda}^*$, however, can be very large (in terms of norm) and depends implicitly on the geometry of $h_1(\cdot)$ and $h_2(\cdot)$, which is dimension-dependent. But optimization literature does not deal with it, as it is considered as a constant for a single optimization problem. This issue is not particular to 3-block ADMM. 2-block ADMM also involves true dual variable in the convergence rate, which is treated as constant in the literature.

We bound $\|\Lambda^*\|$, a geometry related quantity, by easy-to-compute geometry quantities.

To understand the involvement of geometry intuitively, figure 1 takes the projection on the intersection of two convex sets as an example for illustration. If the point to be taken projection on, say $P_0$, satisfies $\text{Proj}_{C_1\cap C_2}(P_0) = A$, the number of iterations needed to get enough close to

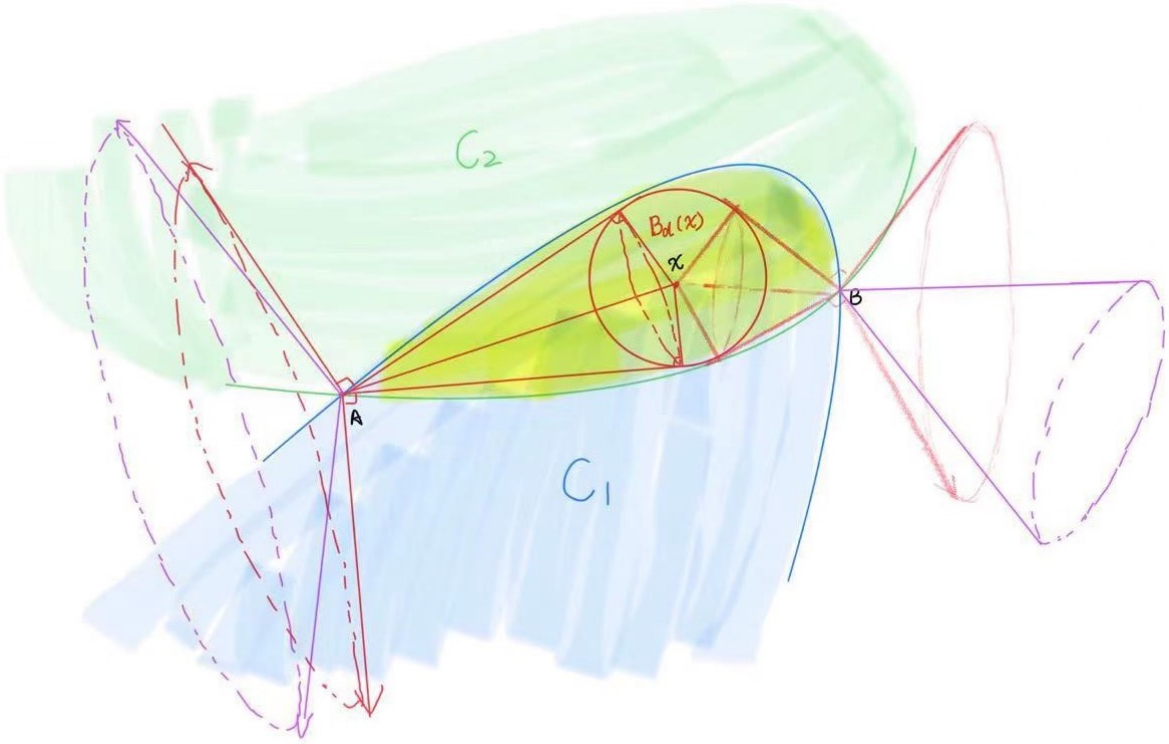


**Figure 1** Illustration of geometry of dual variable

$C_1 \cap C_2$ would be relatively large, as $P_k$ can stay far from $C_1 \cap C_2$ while it's already close to both $C_1$ and $C_2$ separately. On the other hand, when $\text{Proj}_{C_1 \cap C_2}(P_0) = B$, it would take less iterations to get enough close to $B$. Simple calculation show that $-\Lambda_1^*$ and $-\Lambda_2^*$ are subgradients for $\mathfrak{T}\{X \in C_1\}$ and $\mathfrak{T}\{X \in C_2\}$ at $\text{Proj}_{C_1 \cap C_2}(P_0)$, satisfying $-\Lambda_1^* - \Lambda_2^* = 2(P_0 - P^*)$. In figure 1, the purple cones at $A$ and $B$ show the region $\Lambda_1^*$ and $\Lambda_2^*$ can take value in at $A$ and $B$ respectively. We find bound for $\|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2$ by finding bound for "the maximum angle" the purple cones. The purple cone (smaller cone) at $A$ can be considered as polar cone (Chandrasekaran and Jordan 2013) of the smallest cone containing $C_1 \cap C_2$ with $A$ considered as origin, which at least contains the ball $B_d(x)$. Thus we can bound the purple cone by red cone. Same logic applies to purple cone (smaller cone) at $B$. Lemma 1 gives the precise description of this intuition.

LEMMA 1 **(Geometry Bound)**. *We define the generalized polar cone of convex set $C$ at point $P$ to be $N_C(P) = \{\mathbf{a} : \langle \mathbf{a}, P - x \rangle \geq 0$ for all $x \in C\}$. Define the maximum angle of two convex sets $C_1$*

*and* $C_2$ *to be*

$$\theta(C_1, C_2) = \sup_{P \in \partial(C_1 \cap C_2)} \sup_{\lambda_1 \in N_{C_1}(P), \lambda_2 \in N_{C_2}(P)} \arccos\left(\langle \lambda_1, \lambda_2 \rangle\right),$$

*where* $\partial(C_1 \cap C_2)$ *is the boundary of* $C_1 \cap C_2$*. We define a quantity based on maximum angle of* $C_1$ *and* $C_2$ *to be* $C(C_1, C_2) = \frac{1}{2\cos^2(\frac{\theta(C_1,C_2)}{2})}$*, then we have*

$$C(C_1, C_2) \leq \frac{D^2}{2d^2},$$

*where* $D = \sup_{x,y \in C_1 \cap C_2} \|x - y\|_2^2$*,* $d = \sup\{d : \exists x \in C_1 \cap C_2 \text{ such that } B_d(x) \subset C_1 \cap C_2\}$*. Further, suppose* $\mathbf{\Lambda}^*$ *and* $P^*$ *are the true dual variable and primal variable of the Augmented Lagrange function* (16)*. Then when* $h_1(W) = \mathfrak{T}\{W \in C_1\}$ *and* $h_2(Z) = \mathfrak{T}\{Z \in C_2\}$*, we have*

$$\|\mathbf{\Lambda}^*\|_2^2 \leq \max\{4, 4C(C_1, C_2)\}\|P_0 - P^*\|^2.$$

### 2.4. Remark

With the statistical-optimization interplay, algorithm-template, and optimization analysis, we are ready to provide theoretically guaranteed algorithm for a large class of estimator for a wide class of problems, and produce a precise analysis of how running time affects statistical accuracy.

## 3. Application to 1 Bit Matrix Completion

In this section we apply the framework introduced in Section 2 to the 1 bit matrix completion example we introduced in Section 1.1, which yields novel results and also further illustrates our framework.

### 3.1. Statistical-Optimization Interplay

Suppose a solution to optimization problem (4) is $X^*$. The approximation optimization conditions (6) of the computed estimator $\tilde{X}$ in 1 bit matrix completion setting becomes

$$\begin{gathered} -\mathcal{L}_{\Omega,Y}(\tilde{X}) \leq -\mathcal{L}_{\Omega,Y}(X^*) + \delta, \\ \|\tilde{X}\|_\infty \leq \alpha + \delta_1, \|\tilde{X}\|_* \leq \alpha\sqrt{r d_1 d_2} + \delta_2, \inf_{Z \in C_1 \cap C_2} \|Z - \tilde{X}\|_2 \leq \delta_0, \end{gathered} \tag{21}$$

where $C_1 = [-\alpha, \alpha]^{d_1 \times d_2}$ and $C_2 = \{M \in \mathbb{R}^{d_1 \times d_2} | \|M\|_* \leq \alpha\sqrt{r d_1 d_2}\}$.

Our goal is to understand the statistical behavior of $\tilde{X}$. Applying statistical-optimization interplay step of our framework to the statistical analysis in Davenport et al. (2014), where $X^*$ is $\hat{M}$ and $\tilde{X}$ is $\tilde{M}$, gives Theorem 2, which describes how optimization-induced error affects the statistical accuracy before solving the optimization problem.

THEOREM 2. *Consider 1 bit matrix completion problem introduced in Example 1. Let $\hat{M}$ be a solution to optimization problem (4). Suppose $\tilde{M}$ satisfies $-\mathcal{L}_{\Omega,Y}(\tilde{M}) \le -\mathcal{L}_{\Omega,Y}(\hat{M}) + \delta$, $\|\tilde{M}\|_* \le \alpha\sqrt{rd_1d_2} + \delta_2$, $\|\tilde{M}\|_\infty \le \alpha + \delta_1$. Recall that $D(A\|B)$ is the average KL divergence between matrix A and B. Denote*

$$L_\gamma = \sup_{|x|\le\gamma} \frac{|l'(x)|}{l(x)(1-l(x))} \tag{22}$$

*for $\gamma > 0$ such that $l(x) \in (0,1)$ for $|x| \le \gamma$. Then we have, with probability at least $1 - \frac{c_1}{d_1+d_2}$,*

$$D(l(M)\|l(\tilde{M})) \le c_0 L_{\alpha+\delta_1}(\alpha\sqrt{rd_1d_2} + \delta_2)\sqrt{\frac{d_1+d_2}{nd_1d_2}}\sqrt{1 + \frac{(d_1+d_2)\log d_1d_2}{n}} + \frac{\delta}{n}, \tag{23}$$

*$c_0$, $c_1$ are absolute constants that can be explicitly written out.*

REMARK 11. Note that in the formulation of example 1 we require link function $l$ to be twice differentiable in addition to mere differentiability in the original work (Davenport et al. 2014) for fitting into our framework. But for statistical-optimization interplay, twice differentiability is not necessary, as Theorem 2 still holds with only differentiability.

REMARK 12. Note that when $\delta = 0$, $\delta_1 = 0$ and $\delta_2 = 0$, $\tilde{M}$ in Theorem 2 is exactly the target estimator and the rate is of the same order with that in Davenport et al. (2014). In the view of approximate optimization, the target exact solution is a special case.

REMARK 13. 1 bit matrix completion is a representative example for constrained M-estimator, or more precisely, constrained maximum likelihood estimator with no penalty term or optimization-wise smooth penalty term. Other constrained M-estimator includes constrained kernel ridge regression and constrained version of sparse principle component analysis.

### 3.2. Optimization Algorithm

Note that in Davenport et al. (2014), they use a heuristic method without theoretical guarantee. Here we apply our optimization template algorithm to 1 bit matrix completion and give results on its convergence in terms of the approximate optimization conditions.

Note that the proximal operator $\text{Prox}_{\eta(g(X)+\mathfrak{T}\{X\in C_1\cap C_2\cap\cdots\cap C_J\})}(\cdot)$ in optimization template algorithm becomes projection operator $\text{Proj}_{C_1\cap C_2}(\cdot)$ for 1 bit matrix completion, which gives the outer loop in Algorithm 3.1.

1-bit Matrix Completion Outer Loop: Inexact Projected Gradient Descent

Starting point is $X_0 = \mathbf{0}$. Step size $\eta > 0$. For $k \geq 0$, the iteration steps are

$$X_{k+0.5} = X_k - \eta\nabla(-\mathcal{L}_{\Omega,Y}(X_k)), \qquad X_{k+1} = \widetilde{\text{Proj}}_{C_1\cap C_2}(X_{k+0.5}), \tag{24}$$

where $\widetilde{\text{Proj}}_{C_1\cap C_2}(X_{k+0.5})$ is a close approximation of *projection point*

$$\begin{aligned}&\text{Proj}_{C_1\cap C_2}(X_{k+0.5}) = \\ &\arg\min_X \left(\|X - X_{k+0.5}\|_F^2 + \mathfrak{T}\{X \in C_1 \cap C_2\}\right).\end{aligned} \tag{25}$$

To compute approximate projection point $\widetilde{\text{Proj}}_{C_1\cap C_2}(P_0))$, we apply the template algorithm inner loop. We know that the Augmented Lagrange Function for this 3-block ADMM is:

$$\begin{aligned}\mathcal{L}_\beta(W,Z,P,\Lambda_1,\Lambda_2) =& \mathfrak{T}\{W\in C_1\} + \mathfrak{T}\{Z\in C_2\} + \|P - P_0\|_F^2 + \\ &\frac{\beta}{2}(\|W - P + \frac{\Lambda_1}{\beta}\|_2^2 + \|Z - P + \frac{\Lambda_2}{\beta}\|_2^2),\end{aligned} \tag{26}$$

where $\Lambda_1$ and $\Lambda_2$ are dual variables and $\beta$ is the dual update step size.

Applying the inner loop template algorithm, Algorithm 2.2, to 1 bit matrix completion, gives the inner loop steps for 1 bit matrix completion in Algorithm 3.2.

1-bit Matrix Completion Inner Loop: 3-block ADMM The starting points are $P^0 = P_0, \Lambda_1^0 = \mathbf{0}, \Lambda_2^0 = \mathbf{0}$. For $k \geq 0$, the iterative steps are

$$\begin{aligned} &W^{k+1} = \mathrm{Proj}_{C_1}(P^k - \frac{1}{\beta}\Lambda_1^k), Z^{k+1} = \mathrm{Proj}_{C_2}(P^k - \frac{1}{\beta}\Lambda_2^k), \\ &P^{k+1} = \frac{1}{\beta+1}\left(P_0 + \Lambda_1^k + \Lambda_2^k + \frac{\beta}{2}(W^{k+1} + Z^{k+1})\right), \\ &\Lambda_1^{k+1} = \Lambda_1^k + \beta\left(W^{k+1} - P^{k+1}\right), \\ &\Lambda_2^{k+1} = \Lambda_2^k + \beta\left(Z^{k+1} - P^{k+1}\right). \end{aligned} \tag{27}$$

Take the average $\overline{P}^k = \frac{1}{k}\sum_{i=1}^k P^i$ for the output if we end it at k-th iteration.

### 3.3. Optimization Convergence

In this section, we establish convergence rate for optimization algorithm introduced in Section 3.2 in terms of the approximate optimization conditions. We apply results in Section 2.3 to 1 bit matrix completion setting with appropriate modifications.

In this section, we need the assumption that the link function $l$ for 1-bit matrix completion is twice differentiable, as introduced in section 1.1. So in addition to Lipschitz constant defined in (22), we have well defined smoothness constant for 1 bit matrix completion example, defined as

$$\tilde{L}_\gamma = \sup_{|x|\leq\gamma}\{\frac{|l''(x)l(x) - (l'(x))^2|}{l(x)^2}, \frac{|l''(x)(1-l(x)) + (l'(x))^2|}{(1-l(x))^2}\}, \tag{28}$$

for $\gamma > 0$ such that $l(x) \in (0,1)$ for $|x| \leq \gamma$.

For the convergence of the outer loop, we apply Theorem 1 to 1 bit matrix completion setting, which gives Proposition 3.1.

Outer loop for 1 bit matrix completion Suppose we take projected gradient descent, Algorithm 3.1, for outer loop, and the projection error in all steps satisfies

$$\|\widetilde{\mathrm{Proj}}_{C_1\cap C_2}(X) - \mathrm{Proj}_{C_1\cap C_2}(X)\| \leq \delta_0$$

. Suppose the link function $l(x)$ is twice differentiable. Let $\tilde{L}_\gamma$ be defined in (28). Suppose $\tilde{L}_{\alpha+\delta_0} \leq L$. Let $L_\gamma$ be defined in (22). Let $X^*$ be a solution of optimization problem (4). Take step size $\eta = \frac{1}{L}$, we have

$$\min_{0\leq k\leq K} -\mathcal{L}_{\Omega,Y}(X_k) \leq -\mathcal{L}_{\Omega,Y}(X^*) + \frac{\alpha^2 L d_1 d_2}{2K} + \delta_0(2\alpha L\sqrt{d_1 d_2} + L_{\alpha+\delta_0} + L\delta_0). \tag{29}$$

To investigate the convergence for inner loop, we apply Proposition 2.1 and Lemma 1 in the general framework to 1 bit matrix completion example. Proposition 3.2 gives the convergence for inner loop for 1 bit matrix completion.

CONVERGENCE OF INNER LOOP FOR 1 BIT MATRIX COMPLETION Suppose $P^* = \text{Proj}_{C_1 \cap C_2}(P_0)$. Taking Algorithm 3.2, with dual step size $\beta \leq \frac{6}{17}$, we have

$$\|\overline{P}^t - P^*\|^2 \leq \frac{1}{2\beta t}\left(7\beta^2 + \max\{4, 8C(C_1, C_2)\} + \frac{20}{3}\frac{\beta^4}{(\beta+1)^2}\right)\|P_0 - P^*\|^2, \tag{30}$$

where $C(C_1, C_2) \leq \frac{d_1 d_2}{2}$.

Combing the inner loop result, Proposition 3.2, and outer loop result, Proposition 3.1, we have that Theorem 3 showing the overall optimization convergence in terms of approximate conditions.

THEOREM 3 **(Optimization: 1 bit matrix completion)**. *Suppose we take projected gradient descent, Algorithm 3.1, for outer loop, and 3-block ADMM, Algorithm 3.2, for inner loop, where* $P_0$ *in the inner loop is* $X_{k+0.5}$ *in the outer loop. Let* $L_\alpha$ *be defined in* (22). *Let* $\tilde{L}_\alpha$ *be defined in* (28). *If we take step size* $\eta = \frac{1}{2\tilde{L}_\alpha}$, *dual step size* $\beta \leq \frac{6}{17}$, *the number of iterations of inner loop* $t \geq t_0$, *and take* $T$ *iterations for outer loop, then* $\tilde{X} = \arg\min_{X \in \{X_0, X_1, \cdots, X_T\}} -\mathcal{L}_\Omega(X)$ *satisfies the approximate conditions* (21) *with*

$$\begin{aligned} &\delta \leq \frac{\alpha^2 \tilde{L}_\alpha d_1 d_2}{T} + (4\alpha\tilde{L}_\alpha\sqrt{d_1 d_2} + 2L_\alpha)\sqrt{\frac{1}{t}}\sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} + 2\tilde{L}_\alpha\frac{1}{t}\left(q(\beta) + \frac{2d_1 d_2}{\beta}\right), \\ &\max\{\delta_1, \delta_2, \delta_0\} \leq \sqrt{\frac{1}{t}}\sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}}, \end{aligned} \tag{31}$$

*where* $q(\beta) = \frac{7\beta}{2} + \frac{10}{3}\frac{\beta^3}{(\beta+1)^2}$, $u_0 = \max\{u : L_{\alpha+u} \leq 2L_\alpha, \tilde{L}_{\alpha+u} \leq 2\tilde{L}_\alpha\}$, *and* $t_0 = \frac{1}{2\beta}\left(7\beta^2 + 4d_1 d_2 + \frac{20}{3}\frac{\beta^4}{(\beta+1)^2}\right)(1 + \frac{L_\alpha}{u_0\tilde{L}_\alpha} + \frac{L_\alpha}{\tilde{L}_\alpha})^2$.

### 3.4. Overall Result

In this section, we are ready to show how the running time affects the statistical accuracy, as shown in Theorem 4.

THEOREM 4. *For 1 bit matrix completion introduced in Section 1.1, suppose the link function $l(x)$ is twice differentiable. Let $\tilde{L}_\alpha$ be define in (28). Let $L_\alpha$ be defined in (22). Suppose we take projected gradient descent, Algorithm 3.1, for outer loop with step size $\eta = \frac{1}{2\tilde{L}_\alpha}$ and $T$ iterations, and 3-block ADMM, Algorithm 3.2, for inner loop, where $P_0$ in the inner loop is $X_{k+0.5}$ in the outer loop. For inner loop, Algorithm 3.1, we take dual step size $\beta \le \frac{6}{17}$ and iteration number $t \ge t_0$, where $t_0$ is specified later. Let $\tilde{M}$ be among the starting point and resulting points in first $T$ iterations of the outer loop, $\{X_0, X_1, \cdots, X_T\}$, such that it has the smallest $-\mathcal{L}_{\Omega,Y}(\cdot)$ value. Then with probability at least $1 - \frac{c_1}{d_1+d_2}$, we have*

$$\begin{aligned} & D(l(M)\|l(\tilde{M})) \\ & \le 2c_0 L_\alpha \left( \alpha\sqrt{r d_1 d_2} + \sqrt{\frac{1}{t}} \sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} \right) \sqrt{\frac{d_1+d_2}{n d_1 d_2}} \sqrt{1 + \frac{(d_1+d_2)\log(d_1 d_2)}{n}} \\ & + \frac{\alpha^2 \tilde{L}_\alpha d_1 d_2}{Tn} + \frac{4\alpha \tilde{L}_\alpha \sqrt{d_1 d_2} + 2L_\alpha}{n} \sqrt{\frac{1}{t}} \sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} + \frac{2\tilde{L}_\alpha}{n} \frac{1}{t} \left( q(\beta) + \frac{2d_1 d_2}{\beta} \right). \end{aligned} \tag{32}$$

*where $c_0$, $c_1$ are absolute constants, and $q(\beta), t_0$ is defined as follows.*

$$\begin{aligned} & q(\beta) = \frac{7\beta}{2} + \frac{10}{3} \frac{\beta^3}{(\beta+1)^2}, u_0 = \max\{u : L_{\alpha+u} \le 2L_\alpha, \tilde{L}_{\alpha+u} \le 2\tilde{L}_\alpha\}, \\ & t_0 = \frac{1}{2\beta} \left( 7\beta^2 + 4d_1 d_2 + \frac{20}{3} \frac{\beta^4}{(\beta+1)^2} \right) (1 + \frac{L_\alpha}{u_0 \tilde{L}_\alpha} + \frac{L_\alpha}{\tilde{L}_\alpha})^2. \end{aligned}$$

Note that, when the computing resource in terms of running time is unlimited, meaning $t \to \infty$ and $T \to \infty$, the rate is the same with that established in Davenport et al. (2014). Also note that Theorem 4 gives better understanding of the roles the iteration number T and t play. The running-time-induced statistical error is of the order $O\left(\sqrt{\frac{1}{t}} \cdot \left( \frac{L_\alpha}{\sqrt{min\{d_1,d_2\}}} + \alpha\tilde{L}_\alpha \right)\right) + O(\frac{\alpha^2 \tilde{L}_\alpha}{T})$. The running time for inner loop plays a crucial role, which is reasonable as the inner-loop-error propagates down the outer loop.

There are flexibility in the choice of step sizes $\eta$, similar results can be given for other legitimate choices of step sizes. The heuristic algorithm in Davenport et al. (2014) is a 2-block ADMM. Our framework can also be adapted to 2-block ADMM, the change in the down-stream-convergence-analysis is to replace the 3-block convergence rate with 2-block convergence rate and analyze the dimension-dependent geometric quantity involved there with the insights provided by Lemma 1.

## 4. Application to Causal Inference for Panel Data

In this section, we apply our framework to the causal inference for panel data. Athey et al. (2021) proposed an estimator of the general form (1) for causal inference for panel data. Their statistical analysis, however, is not tight, and they do not have an optimization procedure targeting their estimator. We provide an improved statistical analysis and apply our framework based on our improved analysis, resulting in a theoretically guaranteed algorithm with precise quantification of the statistical accuracy after certain running time of user's choice.

We take the statistical model in the work by Athey et al. (2021). The model is for panel data. There are N items, which can stand for companies. The time period is T. For each item i, there is an adoption time $t_i$, after which item $i$ is treated all the way to time T, and this adoption time is set to $T$ if never treated. They take Rubin's potential outcome framework. And the complete potential outcome matrix when all are assigned to the control group is $Y^{full}$,

$$Y^{full} = \mathbf{L}^* + \boldsymbol{\varepsilon}, \quad \text{where } \mathbb{E}(\boldsymbol{\varepsilon}|\mathbf{L}^*) = 0. \tag{33}$$

The assumptions on $\boldsymbol{\varepsilon}$ are as follows. $\boldsymbol{\varepsilon}$ is independent from $\mathbf{L}^*$ and the elements of $\boldsymbol{\varepsilon}$ are $\sigma$-sub-Gaussian and independent of each other.

$\mathcal{O}$ is the observation-pair set indicating whether a unit (an item at a certain time) is treated. If we let $W$ to be defined as

$$W_{it} = \begin{cases} 1, & \text{for } (i,t) \notin \mathcal{O} \\ 0, & \text{for } (i,t) \in \mathcal{O} \end{cases}. \tag{34}$$

The assumptions for $\mathcal{O}$ and thus $W$ are as follows. For each row, suppose row i, there is an adoption time $t_i$, such that $W_{it} = 1$ for all $t_i < t \leq T$, $t_i = T$ if the unit never adopt the treatment. The rows of $W$ are independent. Condition on $\mathbf{L}^*$, the adoption time $t_i$ are independent of each other and $\boldsymbol{\varepsilon}$. Also, $|\mathbf{L}^*|_\infty \leq L_{max}$, where $L_{max}$ is a positive real number.

Then under this model, the observed controls are $Y_{it} = Y_{it}^{full}, (i,t) \in \mathcal{O}$. For treated elements, i.e. $(i,t) \notin \mathcal{O}$ , $Y_{it}^{full}$ is missing and we let $Y_{it} = 0$. The goal is to estimate $\mathbf{L}^*$.

We introduce some quantities here. For item i, the probability that it's not treated through out is $\pi_T^{(i)} = \mathbb{E}(\mathfrak{T}\{t_i = T\})$. The minimum of this "probability of control" over N items is $p_c = \min_{1\le i\le N} \pi_T^{(i)}$. We use $\mathbf{P}_{\mathcal{O}}$ to denote an operator mapping N by T matrix to N by T matrix, with each elements defined as $\mathbf{P}_{\mathcal{O}}(B)_{(i,t)} = B_{(i,t)}$ if $(i,t) \in \mathcal{O}$, and 0 if $(i,t) \notin \mathcal{O}$.

Note that in this setting, the matrix $W$ do not have independence for columns, which renders RIP condition and restricted strong convexity invalid. The targeted estimator (Athey et al. 2021) is

$$\hat{\mathbf{L}} = \underset{|\mathbf{L}|_\infty \le L_{\max}}{\arg\min} \{\frac{1}{|\mathcal{O}|}\|\mathbf{P}_{\mathcal{O}}(Y - \mathbf{L})\|_F^2 + \lambda\|\mathbf{L}\|_*\} \tag{35}$$

So causal inference for panel data example fits our general framework (1). The smooth convex function $f$, the convex-but-not-necessarily-smooth function $g$ and the constraint set in the general framework become follows in causal panel data setting.

$$f(\mathbf{L}) = \frac{1}{|\mathcal{O}|}\|\mathbf{P}_{\mathcal{O}}(Y - \mathbf{L})\|_F^2, g(\mathbf{L}) = \lambda\|\mathbf{L}\|_*, C_1 = [-L_{\max}, L_{\max}]^{N*T}. \tag{36}$$

Applying our framework to it are two sub-problem as follows.

The first sub-problem is to investigating the statistical behavior of an estimator $\tilde{\mathbf{L}}$ satisfying conditions (37).

$$\begin{gathered}\frac{1}{|\mathcal{O}|}\|\mathbf{P}_{\mathcal{O}}(Y - \tilde{\mathbf{L}})\|_F^2 + \lambda\|\tilde{\mathbf{L}}\|_* \le \frac{1}{|\mathcal{O}|}\|\mathbf{P}_{\mathcal{O}}(Y - \hat{\mathbf{L}})\|_F^2 + \lambda\|\hat{\mathbf{L}}\|_* + \delta, \\ |\tilde{\mathbf{L}}|_\infty \le L_{\max} + \delta_1,\end{gathered} \tag{37}$$

where $\hat{\mathbf{L}}$ is defined in (35).

The second sub-problem is developing theoretically guaranteed algorithm finding an $\tilde{\mathbf{L}}$ satisfying (37) and analyzing its convergence rate in terms of $\delta$ and $\delta_1$ in (37). Athey et al. (2021) does not have an algorithm for $\hat{\mathbf{L}}$ in (35) and the heuristic algorithm used there is for another target estimator.

### 4.1. Statistical-Optimization Interplay

We start with the first sub-problem.

The statistical property of the approximate estimator $\tilde{\mathbf{L}}$ satisfying (37) is shown in Theorem 5, which describes how optimization induced error affects statistical error before solving the optimization problem.

THEOREM 5. *Consider statistical model for causal inference of panel data. Suppose the true parameter matrix* $\mathbf{L}^*$ *has rank at most R, and the penalty parameter*

$$\lambda = \frac{13\sigma \max\{\sqrt{N\log(N+T)}, 8\sqrt{T}\log^{\frac{3}{2}}(N+T)\}}{|\mathcal{O}|}$$

*. Let* $\hat{\mathbf{L}}$ *be defined in* (35)*. Suppose the computed estimator* $\tilde{\mathbf{L}}$ *satisfies* $f(\tilde{\mathbf{L}})+g(\tilde{\mathbf{L}}) \le f(\hat{\mathbf{L}})+g(\hat{\mathbf{L}})+\delta$ *and* $|\tilde{\mathbf{L}}|_\infty \le L_{max}+\delta_1$*. Then with probability at least* $1-\frac{2}{(N+T)^2}$*, we have*

$$\begin{aligned}\frac{\|\tilde{\mathbf{L}}-\mathbf{L}^*\|_F^2}{NT} \le \max\Bigg\{ & q_0\frac{R\sigma^2}{p_c^2}\frac{(N+T)\log^3(N+T)}{NT} + \frac{72}{p_c}\delta + q_1\frac{\delta(L_{max}+\delta_1)}{\sigma p_c}\frac{1}{NT} \\ & + q_2\frac{R(L_{max}+\delta_1)^2}{p_c^2}\frac{N+T}{NT}, \\ & \frac{132(L_{max}+\delta_1)^2}{p_c}\frac{\log(N+T)}{N}\Bigg\},\end{aligned} \tag{38}$$

*where* $q_0$*,* $q_1$*,* $q_2$ *are constants that can be explicitly written out.*

REMARK 14. Note that when $\delta = 0$ and $\delta_1 = 0$, the estimator becomes the original exact estimator (i.e. $\hat{\mathbf{L}}$ in (35)), and our rate becomes of order

$$\max\{\sigma^2 R(\frac{N+T}{NT})log^3(N+T)\frac{1}{p_c}, L_{\max}\frac{\log(N+T)}{N}\}.$$

This is a faster rate than that in Athey et al. (2021), which is because we sharpen the statistical analysis of the original estimator and we apply our framework to our own analysis of the statistical performance of the original exact estimator. If we apply our framework directly to the analysis in Athey et al. (2021), we expect the same rate when $\delta$ and $\delta_1$ are set to 0.

REMARK 15. causal inference for panel data is a representation for constrained penalized M-estimator, or more precisely, constrained penalized maximum likelihood estimator, where the penalty term is not smooth (optimization wise). Other constrained non-smoothly-penalized M-estimator includes Lasso with constraints, Danzig selector, elastic net, SVM, sparse principle component analysis in the penalized form, neural network with Relu activation function.

### 4.2. Optimization Algorithm

In this Section, we apply our algorithm template to causal inference of panel data, which gives theoretically guaranteed optimization algorithm for causal inference of panel data.

To standardize the optimization problem for fitting into our optimization template better, the target optimization problem can be written as

$$\min_{\mathbf{L}} \frac{1}{2}\|\mathbf{P}_{\mathcal{O}}(Y-\mathbf{L})\|_F^2 + \frac{1}{2}\lambda|\mathcal{O}|\|\mathbf{L}\|_* + \mathfrak{T}\{|\mathbf{L}|_\infty \le L_{\max}\}. \tag{39}$$

Applying general outer loop, Algorithm 2.1, to causal inference for panel data gives Algorithm 4.1.

Causal Inference for Panel Data Outer Loop: Inexact Proximal Gradient Descent

Start from point $\mathbf{L}_0 = \mathbf{0}$. Step size is $\eta > 0$. For $k \ge 0$,

$$\begin{aligned} \mathbf{L}_{k+0.5} &= \mathbf{L}_k - \eta\nabla(\|\mathbf{P}_{\mathcal{O}}(Y-\mathbf{L}_k)\|_F^2), \\ \mathbf{L}_{k+1} &= \widetilde{\text{Prox}}_{\eta(\frac{1}{2}\lambda|\mathcal{O}|\|\mathbf{L}\|_* + \mathfrak{T}\{|\mathbf{L}|_\infty \le L_{\max}\})}(\mathbf{L}_{k+0.5}), \end{aligned} \tag{40}$$

where $\widetilde{\text{Prox}}$ is an approximate proximal algorithm aiming at finding the proximal of $\mathbf{L}_{k+0.5}$,

$$\begin{aligned} &\text{Prox}_{\eta(\frac{1}{2}\lambda|\mathcal{O}|\|\mathbf{L}\|_* + \mathfrak{T}\{|\mathbf{L}|_\infty \le L_{\max}\})}(\mathbf{L}_{k+0.5}) = \\ &\arg\min_{\mathbf{L}} \left( \frac{1}{2}\|\mathbf{L}-\mathbf{L}_{k+0.5}\|^2 + \eta\left( \frac{\lambda|\mathcal{O}|\|\mathbf{L}\|_*}{2} + \mathfrak{T}\{|\mathbf{L}|_\infty \le L_{\max}\} \right) \right). \end{aligned} \tag{41}$$

We abbreviate the approximate proximal and proximal in equation (40) and (41) as $\widetilde{\text{Prox}}_\eta(\mathbf{L}_{k+0.5})$ and $\text{Prox}_\eta(\mathbf{L}_{k+0.5})$, respectively, when there is no confusion.

For the inner loop (i.e. computing approximate proximal point $\widetilde{\text{Prox}}_\eta(\mathbf{L}_{k+0.5})$), we apply the template-algorithm, Algorithm 2.2.

In this setting, the Augmented Lagrange Function for 3-block ADMM with dual step size $\beta$ and $\mathbf{L}_{k+0.5}$ replaced by $P_0$ is

$$\mathcal{L}_\beta(W,Z,P) = \mathfrak{T}\{W \in C_1\} + \|Z\|_*\lambda|\mathcal{O}| + \|P-P_0\|_F^2 + \frac{\beta}{2}(\|W-P+\frac{\Lambda_1}{\beta}\|_2^2 + \|Z-P+\frac{\Lambda_2}{\beta}\|_2^2), \tag{42}$$

where $\Lambda_1$ and $\Lambda_2$ are dual variables.

The template inner loop, Algorithm 2.2, in this setting becomes Algorithm 4.2.

3 BLOCK ADMM FOR CAUSAL INFERENCE FOR PANEL DATA The starting points are $P^0 = P_0$, $\Lambda_1^0 = \mathbf{0}$, $\Lambda_2^0 = \mathbf{0}$. Dual step size is $\beta > 0$. For $k \geq 0$, the iterative steps are

$$\begin{aligned} W^{k+1} &= \text{Proj}_{C_1}(P^k - \frac{1}{\beta}\Lambda_1^k), Z^{k+1} = thresh(P^k - \frac{1}{\beta}\Lambda_2^k, \frac{\lambda|\mathcal{O}|}{\beta}), \\ P^{k+1} &= \frac{1}{\beta+1}\left(P_0 + \Lambda_1^k + \Lambda_2^k + \frac{\beta}{2}(W^{k+1} + Z^{k+1})\right), \\ \Lambda_1^{k+1} &= \Lambda_1^k + \beta\left(W^{k+1} - P^{k+1}\right), \\ \Lambda_2^{k+1} &= \Lambda_2^k + \beta\left(Z^{k+1} - P^{k+1}\right), \end{aligned} \tag{43}$$

where $thresh(P, b)$ is defined as follows. Suppose the Singular value decomposition of $P$ is $P = UDV$, then $thresh(P, b) = U(D - diag(b))_+ V$. We take the average $\overline{P}^k = \frac{1}{k}\sum_{i=1}^k P^k$ for the output if we end it at k-th iteration.

### 4.3. Optimization Convergence

In this section, we establish convergence rate for our optimization algorithm introduced in Section 4.2 in terms of approximate optimization conditions. We apply results in Section 2.3 to our causal inference for panel data setting with appropriate modifications.

Applying theorem 1 to causal inference for panel data, we have Proposition 4.1.

OUTER LOOP FOR CAUSAL INFERENCE FOR PANEL DATA Suppose we take the gradient proximal algorithm, Algorithm 4.1, for outer loop with $\eta = 1$. Suppose the proximal error satisfies

$$|\widetilde{\text{Prox}}_{\frac{\lambda|\mathcal{O}|}{2}\|\mathbf{L}\|_* + \mathfrak{T}\{\mathbf{L}\in C_1\}}(X) - \text{Prox}_{\frac{\lambda|\mathcal{O}|}{2} + \mathfrak{T}\{\mathbf{L}\in C_1\}}(X)| \leq \delta_0$$

for all $X \in \mathcal{R}(\delta_0, C_1)$. $C_1$ is defined in (36) and $\delta_0$ is a positive real number. Let $\hat{\mathbf{L}}$ be the target estimator define in (35). Then we have

$$\begin{aligned} \min_{0\leq k\leq K} \frac{1}{2}\|\mathbf{P}_\mathcal{O}(Y - \mathbf{L}_k)\|_F^2 + \frac{\lambda|\mathcal{O}|}{2}\|\mathbf{L}_k\|_* &\leq \frac{1}{2}\|\mathbf{P}_\mathcal{O}(Y - \hat{\mathbf{L}})\|_F^2 + \frac{\lambda|\mathcal{O}|}{2}\|\hat{\mathbf{L}}\|_* \\ + \frac{1}{2K}\|\mathbf{L}_0 - \hat{\mathbf{L}}\|^2 + \delta_0^2 + 2\delta_0 L_{max}\sqrt{NT} + C(Y)\delta_0 + min\{\sqrt{N}, \sqrt{T}\}\frac{\lambda|\mathcal{O}|}{2}\delta_0, \end{aligned} \tag{44}$$

where $C(Y) = \sup_{\mathbf{L}\in C_1} \|\mathbf{P}_\mathcal{O}(Y - \mathbf{L})\|$.

For the inner loop, we have the convergence result in Proposition 4.2.

CONVERGENCE OF INNER LOOP FOR CAUSAL INFERENCE FOR PANEL DATA Taking algorithm 4.2, with dual step size $\beta \le \frac{6}{17}$, after k iterations, we have

$$\begin{aligned}\|\overline{P}^k - P^*\|^2 \le & \frac{1}{\beta k}\left((3\beta^2+8)\|P_0-P^*\|^2 + \left(5+\frac{8}{3}(\frac{\beta}{1+\beta})^2\right)(\lambda|\mathcal{O}|)^2\min\{N,T\}\right.\\ & \left. + \left(\beta^2+\frac{8}{3}(\frac{\beta^2}{1+\beta})^2\right)\|P_0-\mathrm{Proj}_{C_1}(P_0)\|^2\right).\end{aligned} \tag{45}$$

Combing the inner loop result, Proposition 4.2, and outer loop result, Proposition 4.1, we have Theorem 6 showing the overall convergence in terms of approximate conditions.

THEOREM 6 **(optimization : causal inference for panel data)**. *Suppose we take proximal gradient descent, Algorithm 4.1 with $\eta = 1$, for outer loop, and 3-block ADMM algorithm 4.2 with dual step size $\beta \le \frac{6}{17}$ for inner loop, where $P_0$ in the inner loop is $\mathbf{L}_{k+0.5}$ in the outer loop. Define four constants depending on $\beta$ only, $q_0(\beta), q_1(\beta), q_2(\beta), q_3(\beta)$, which we will explicitly write out later. Suppose the number of iterations for inner loop $k \ge q_0(\beta)$. Suppose we take $K$ iterations for outer loop and $\tilde{\mathbf{L}} = \arg\min_{0\le i\le K}\{\frac{1}{|\mathcal{O}|}\|\mathbf{P}_{\mathcal{O}}(Y-\mathbf{L}_i)\|_F^2 + \lambda\|\mathbf{L}_i\|_*\}$. Define a quantity $\delta(k)$ as*

$$\delta(k) = \sqrt{\frac{q_1(\beta)(\lambda|\mathcal{O}|)^2\min\{N,T\} + q_2(\beta)C(Y)^2 + q_3(\beta)\left(\|Y\|^2 + 2(NT-|\mathcal{O}|)L_{\max}^2\right)}{k-q_0(\beta)}}. \tag{46}$$

*Then we have $\tilde{\mathbf{L}}$ satisfies the polluted conditions* (37) *with*

$$\begin{aligned}&\delta_1 \le \delta(k),\\ &\delta \le \frac{NTL_{\max}^2}{|\mathcal{O}|K} + \frac{2\delta(k)^2}{|\mathcal{O}|} + \delta(k)\left(\frac{4L_{\max}\sqrt{NT}}{|\mathcal{O}|} + \frac{2C(Y)}{|\mathcal{O}|} + \min\{\sqrt{N},\sqrt{T}\}\lambda\right),\end{aligned} \tag{47}$$

*where $C(Y) = \sup_{\mathbf{L}\in C_1}\|\mathbf{P}_{\mathcal{O}}(Y-\mathbf{L})\|$. The $\beta$ dependent constants are*

$$\begin{aligned}q_0(\beta) &= \left(\frac{1}{\beta}\left(6\beta^2+16+2\beta^2+\frac{16}{3}(\frac{\beta^2}{1+\beta})^2\right)\right), q_3(\beta) = \frac{1}{\beta}(3\beta^2+8),\\ q_1(\beta) &= \frac{1}{\beta}\left(5+\frac{8}{3}(\frac{\beta}{1+\beta})^2\right), q_2(\beta) = \beta\left(2+\frac{16}{3}(\frac{\beta}{1+\beta})^2\right).\end{aligned}$$

### 4.4. Overall Results

In this section, we are ready to show how the running time influences the statistical accuracy, as shown in Theorem 7.

THEOREM 7. *Suppose* $\mathbf{L}^*$ *has rank at most* $R$*, and the penalty parameter*

$$\lambda = \frac{13\sigma \max\{\sqrt{N\log(N+T)}, 8\sqrt{T}\log^{\frac{3}{2}}(N+T)\}}{|\mathcal{O}|}.$$

*Suppose we take proximal gradient descent, Algorithm* (4.1) *with* $\eta = 1$*, for outer loop and 3-block ADMM, Algorithm 4.2, with dual step size* $\beta \le \frac{6}{17}$*, for inner loop, where* $P_0$ *in the inner loop is* $\mathbf{L}_{k+0.5}$ *in the outer loop. There are constants depending on* $\beta$ *only, namely,* $q_0(\beta)$*,* $\widetilde{q_1(\beta)}$*,* $\widetilde{q_2(\beta)}$*,* $\widetilde{q_3(\beta)}$ *such that for iteration number of inner loop* $k > q_0(\beta)$*, the error for inner loop is upper bounded by*

$$\delta(k) = \sqrt{\frac{\widetilde{q_1(\beta)}\sigma^2 NT\log^3(N+T) + \widetilde{q_2(\beta)}\|Y\|^2 + \widetilde{q_3(\beta)}NTL_{\max}^2}{k - q_0(\beta)}}. \tag{48}$$

*Denote* $\tilde{\mathbf{L}}$ *to be the outcome in* $K$ *iterations in outer loop that has the minimum* $f(\tilde{\mathbf{L}}) + g(\tilde{\mathbf{L}})$*. There are absolute constants* $q_0$*,* $q_1$*,* $q_2$ *such that with probability at least* $1 - \frac{2}{(N+T)^2}$*,*

$$\begin{aligned}
&\frac{\|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F^2}{NT} \\
\le &\max\Bigg\{ q_0 \frac{\sigma^2 R}{p_c^2}\left(\frac{N+T}{NT}\right)\log^3(N+T) + \Bigg[\frac{NTL_{\max}^2}{|\mathcal{O}|K} + \frac{2\delta(k)^2}{|\mathcal{O}|} + \\
&\delta(k)\left(\frac{8L_{\max}\sqrt{NT} + 2\|Y\|}{|\mathcal{O}|} + \min\{\sqrt{N},\sqrt{T}\}\lambda\right)\Bigg]\left(\frac{72}{p_c} + q_1\frac{\big(L_{\max} + \delta(k)\big)}{\sigma p_c NT}\right) \\
&+ q_2\left(\frac{N+T}{NT}\right)\frac{\sqrt{R}\big(L_{\max} + \delta(k)\big)^2}{p_c^2}, \\
&132\big(L_{\max} + \delta(k)\big)^2\frac{\log(N+T)}{Np_c}\Bigg\}.
\end{aligned} \tag{49}$$

Note that the optimization error induced statistical error increase is of the order $O(\frac{L_{\max}^2}{K}) + O(\frac{L_{\max}+\sigma}{\sqrt{k}})$, meaning that inner loop can be the bottle neck in terms of convergence rate to the limit statistical accuracy. Also, note that when the computing resource is infinity, i.e. $k \to \infty$ and $K \to \infty$, our results is stronger than that in the work by Athey et al. (2021). This is because our statistical analysis is stricter and we apply our framework based on our analysis. Our framework can also be applied directly to the problem in terms of the part of statistical analysis of the approximate estimator (i.e. statistical-optimization interplay) based on their original work (Athey et al. 2021), then it would lead to the same rate in the case of infinity computing resource as that in Athey et al. (2021).

## 5. Application to Linear Regression (LASSO)

Our framework is designed for problems considering general matrices with constraints, but it is also applicable to vector setting without constraints, which can be considered as a degenerate case. In this section, we show that linear regression with LASSO is such a setting.

We show that analysis and template optimization algorithm in our framework are applicable to (high dimensional sparse) linear regression with LASSO. The optimization algorithm converges to the target LASSO estimator and we give a quantification of how iteration number affects the statistical accuracy of the computed estimator. Further, under restricted strong convexity condition, which holds with high probability and is considered by Loh and Wainwright (2015), our template algorithm applied to LASSO actually has linear convergence rate in a certain range, which matches the optimization rate in Loh and Wainwright (2015). Compared with Loh and Wainwright (2015), we pose less conditions, our optimization algorithm is fully convergent to the target estimator (theirs is not), and in the range that their optimization method performs well, ours is equally well.

Consider the linear model

$$y = \mathbf{X}\theta^* + w, \tag{50}$$

where we observe the vector-matrix pair $(y, \mathbf{X}) \in \mathbb{R}^n \times \mathbb{R}^{n\times d}$. $d$-dimensional vector $\theta^*$ is the unknown true parameter and $w$ is the noise vector. Each row of $\mathbf{X}$, $x_i$, is i.i.d. drawn from $N(\mathbf{0}, \Sigma)$. Noise $w$ is independent of $\mathbf{X}$. Each element of $w$, $w_i$, is i.i.d drawn from $N(0, \sigma^2)$. The goal is to estimate $\theta^*$. LASSO estimator is given by

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2 + \lambda_n\|\theta\|_1, \tag{51}$$

for a chosen $\lambda_n$.

Under our framework (1), the smooth convex function $f(\cdot)$ is $f(\theta) = \|y - \mathbf{X}\theta\|_2^2$, the convex-not-necessarily-smooth function $g(\cdot)$ is $g(\theta) = \lambda_n\|\theta\|_1$. And we do not have constraints.

The first sub-problem becomes investigating the statistical behavior of $\tilde{\theta}$ satisfying

$$\frac{1}{2n}\|y - \mathbf{X}\tilde{\theta}\|_2^2 + \lambda_n\|\tilde{\theta}\|_1 \le \frac{1}{2n}\|y - \mathbf{X}\hat{\theta}\|_2^2 + \lambda_n\|\hat{\theta}\|_1 + \delta. \tag{52}$$

And the second sub-problem is the optimization problem shown in (51). Our optimization template algorithm in Section 2.2 degenerates into the ordinary proximal gradient descent algorithm.

### 5.1. Statistical-Optimization Interplay

LASSO has been intensively analyzed in the literature and the statistical behavior of $\hat{\theta}$ in Equation (51) is well understood. The analysis procedures of $\hat{\theta}$ is consistent with our observation of the analysis of estimators following the general form (1), specifically summarized as follows. Those analysis start with

$$\frac{1}{2n}\|y-\mathbf{X}\hat{\theta}\|_2^2+\lambda_n\|\hat{\theta}\|_1 \leq \frac{1}{2n}\|y-\mathbf{X}\theta^*\|_2^2+\lambda_n\|\theta^*\|_1. \tag{53}$$

Then with proper conditions on $\lambda_n$, this inequality can be easily reduced to

$$0 \leq \frac{1}{2n}\|\mathbf{X}\left(\hat{\theta}-\theta^*\right)\|_2^2 \leq \frac{\lambda_n}{2}\left(\|\hat{\theta}-\theta^*\|_1+2\|\theta^*\|_1-1\|\hat{\theta}\|_1\right). \tag{54}$$

Given that the middle part is essentially a quadratic form of $\hat{\theta}-\theta^*$ and the right hand side is essentially of linear order for $\hat{\theta}-\theta^*$, Inequality (54) implies $\|\hat{\theta}-\theta^*\|$ is upper bounded. This is the key idea in the analysis of LASSO estimator. A careful reflection on this procedure gives the key observation that the additive nature of the inequality (53) is never touched throughout the analysis, which is in align with the mechanism of our framework, meaning that analysis of LASSO estimator can be relatively easily carried to its approximate version solution, i.e. $\tilde{\theta}$ satisfying (52).

Theorem 8 describes the statistical behavior of $\tilde{\theta}$, where we can see how the optimization-induced error affects statistical error before solving the optimization problem.

THEOREM 8. *Let $\rho^2(\Sigma)$ be the maximum diagonal entry of the covariance matrix $\Sigma$. Under the linear regression model* (50)*, for any sparse index set $S$ such that the cardinal of $S$, $|S|=s$, denote $\theta^*_{S^c}$ to be the vector keeping elements not in $S$ the same and setting those in $S$ to be 0. Suppose $c_1\kappa \geq 64s\cdot c_2\rho^2(\Sigma)\frac{\log d}{n}$, where $c_1, c_2$ are constants and can be taken as $c_1=1/8, c_2=50$, and $\kappa$ is the smallest singular value of $\Sigma$. For $\lambda_n \geq 4\sigma\rho(\Sigma)\sqrt{1+\frac{\log d}{n}}\sqrt{\frac{\log 2(n+d)}{n}}$, $\tilde{\theta}$ satisfying* (52) *satisfies the following inequality with probability at least $1-\frac{\exp(-n/32)}{1-\exp(-n/32)}-\exp(-\frac{n}{2})-\frac{1}{2(n+d)}$.*

$$\|\tilde{\theta}-\theta^*\|_2 < \frac{\delta}{2\lambda_n\sqrt{s}}+\frac{\|\theta^*_{S^c}\|_1}{\sqrt{s}}+(2+4\sqrt{s}+\frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}. \tag{55}$$

REMARK 16. The error bound in Theorem 8 has three terms. The first corresponds to optimization error. The second corresponds to approximation error (how different from an s sparse vector). The third term corresponds to estimation error associated with $s$ unknown coefficients. Till now, we do not need an optimization algorithm that guarantee $\|\tilde{\theta}-\theta^*\|$ or $\delta$ in Inequality (52) to be small. All we need is Inequality (52) for some $\delta$. So the optimization convergence rate for $\delta$ in Inequality (52) is possibly faster than general optimization convergence with additional strong convexity or restricted strong convexity conditions. We will show that this is indeed the case, which shows that the first two parts of our framework (i.e. statistical-optimization interplay and optimization template algorithm) automatically adapts to additional stronger conditions.

### 5.2. Optimization Algorithm and Convergence

In the absence of the constraints, our template optimization method degenerates into the ordinary proximal gradient descent as shown in Algorithm 5.1.

ALGORITHM 5.1. Starting point is $\theta_0=\mathbf{0}$. Step size is $\eta>0$. For $k\geq 0$,

$$\begin{aligned}\theta_{k+0.5}&=\theta_k-\eta\nabla_\theta\left(\frac{1}{2n}\|y-\mathbf{X}\theta\|_2^2\right),\\ \theta_{k+1}&=\arg\min_\theta\left(\frac{1}{2}\|\theta-\theta_{k+0.5}\|^2+\eta\lambda_n\|\theta\|_1\right).\end{aligned}\tag{56}$$

Note that $\theta_{k+1}=\arg\min_\theta\left(\frac{1}{2}\|\theta-\theta_{k+0.5}\|^2+\eta\lambda_n\|\theta\|_1\right)$ has explicit expression: the $i$-th element of $\theta_{k+1}$ is $(\theta_{k+1})_i=sign((\theta_{k+0.5})_i)\cdot(|(\theta_{k+0.5})_i|-\eta\lambda_n)_+$, where $sign(x)=-1$ for $x<0$, $sign(x)=0$ for $x=0$ and $sign(x)=1$ for $x>0$.

From the convergence results of our template optimization method, i.e. Theorem 1, we have the optimization convergence rate for Algorithm 5.1 in Theorem 9.

THEOREM 9 **(Optimization Convergence Rate)**. *Let $\|\frac{\mathbf{X}^T\mathbf{X}}{n}\|_s$ be the spectral norm of $\frac{\mathbf{X}^T\mathbf{X}}{n}$. Let step size $\eta\leq\|\frac{n}{\mathbf{X}^T\mathbf{X}}\|_s$ for Algorithm 5.1. Suppose $\tilde{\theta}$ is among $\theta_0,\theta_1,\cdots,\theta_T$ and has the smallest $\frac{1}{2n}\|y-\mathbf{X}\theta\|_2^2+\lambda_n\|\theta\|_1$ value. Then we have that*

$$\frac{1}{2n}\|y-\mathbf{X}\tilde{\theta}\|_2^2+\lambda_n\|\tilde{\theta}\|_1\leq\frac{1}{2n}\|y-\mathbf{X}\hat{\theta}\|_2^2+\lambda_n\|\hat{\theta}\|_1+\frac{1}{2T\eta}\|\hat{\theta}\|^2,\tag{57}$$

*where $\hat{\theta}$ is defined in* (51).

Theorem 9 gives fully converging sub-linear convergence rate, which does not require strong convexity of any form.

Loh and Wainwright (2015) exploits restricted strong convexity, which holds with high probability in high dimensional sparse linear regression, and gives an algorithm with linear convergence rate in certain region. But their convergence result is not fully converging, i.e. optimization error does not converge to 0. We show that, under restricted strong convexity condition, our fully converging optimization algorithm also has linear convergence rate in certain region. Theorem 10 shows how our optimization algorithm performs under different conditions.

THEOREM 10. *Under the linear regression model* (50)*, let* $S$ *be an index set with* $s$ *elements. Suppose* $\lambda_n \geq 2\|\frac{\mathbf{X}^T w}{n}\|_\infty$*, and*

$$\frac{\|\mathbf{X}\theta\|_2^2}{n} \geq a_1\|\theta\|_2^2 - a_2\|\theta\|_1^2, \text{ for all } \theta \in \mathbb{R}^d, \tag{58}$$

*with* $a_2 \leq \frac{1}{64s}a_1$*. Set the step size* $\eta = \frac{n}{\|\mathbf{X}^T\mathbf{X}\|_s}$ *in Algorithm 5.1. Denote* $F(\theta) = \frac{1}{2n}\|\mathbf{X}\theta\|_2^2 + \lambda_n\|\theta\|_1$*. Suppose,* $F(\theta_K) - F(\hat{\theta}) \leq \varepsilon_K$*, where* $\hat{\theta}$ *is defined in Equation* (51)*. Then we have for* $k \geq K$*,*

$$\begin{aligned} F(\theta_k) - F(\hat{\theta}) \leq & \left(1 - \frac{a_1}{8\frac{\|\mathbf{X}^T\mathbf{X}\|_s}{n}}\right)_+^{k-K} \varepsilon_K + 128a_2 s \cdot \left(\frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2 \\ & + 8a_2\frac{\varepsilon_K^2}{\lambda_n^2}, \end{aligned} \tag{59}$$

*where* $\|\cdot\|_s$ *is spectral norm,* $\kappa$ *is the smallest singular value of* $\Sigma$*, and* $\theta^*_{S^c}$ *is* $\theta^*$ *taking only elements in* $S^c$ *to be the same and setting others to 0.*

*Without above conditions except for step size* $\eta = \frac{n}{\|\mathbf{X}^T\mathbf{X}\|_s}$ *in Algorithm 5.1 and using the same notation, we have for* $k \geq 1$*,*

$$\varepsilon_k \leq \frac{\frac{\|\mathbf{X}^T\mathbf{X}\|_s}{n}}{2k}\|\hat{\theta}\|_2^2. \tag{60}$$

Inequality (59) in Theorem 10 has similar form with Theorem 3 in Loh and Wainwright (2015), but our optimization procedure is unconstrained and does not require a pre-specified bound for $\|\theta^*\|_1$. We explain the results in details in remarks. In addition to Inequality (59), we have Inequality (60), a fully converging convergence result without restricted strong convexity requirement, which parallels Theorem (9).

REMARK 17. Note that Inequality (59) is only meaningful for $\varepsilon_K < \frac{\lambda_n^2}{8a_2}$. This means the algorithm needs to start with a close enough initial point or the algorithm can get into this region after some iterations. Similar issue exists for that considered in Loh and Wainwright (2015). Loh and Wainwright (2015) dealt with it by posing hard constraints on $\|\theta\|_1$, which leads to a constrained optimization. However, this constraint is not necessary for Lasso. As shown in Inequality (60) in Theorem 10, $\varepsilon_K$ goes to zero with a rate at least $\frac{1}{K}$, so the algorithm will get into the region $\varepsilon_K < \frac{\lambda_n^2}{8a_2}$ after some iterations. Also, without the knowledge of $\|\theta^*\|_1$, hand-choosing constraint will likely miss the target.

REMARK 18. Note that the right hand side Inequality (59) is larger than or equal to $128a_2 s \cdot \left(\frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2$. Hence this convergence result has a limit and does not go to 0 with iteration number going to $\infty$. It also implies another requirement for Inequality (59) to be meaningful: $\varepsilon_K > 128a_2 s \cdot \left(\frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2$. So Inequality (59) does not show fully convergence of the algorithm. Result in Loh and Wainwright (2015) has similar issue, and they established that this optimization limit is smaller than the statistical limit as $n$ is relatively large. Similar logic applies to our case. This optimization limit highly depends on $a_2$. In fact, condition (58) holds with high probability for $a_1 = c_1\kappa$ and $a_2 = c_2\rho^2(\Sigma)\frac{\log d}{n}$. The optimization limit in our case is also a shrinking quantity (with respect to $n$) times the statistical accuracy. We will see this more clearly in Theorem 11. We now examine how large a region Inequality (59) applies to. We need $\varepsilon_K$ to satisfy

$$128a_2 s \cdot \left(\frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2 \leq \varepsilon_K \leq \frac{\lambda_n^2}{8a_2}. \tag{61}$$

Note that $\lambda_n$ in Theorem 10 needs to satisfy a lower bound condition (i.e. $\lambda_n \geq 2\|\frac{\mathbf{X}^T w}{n}\|_\infty$). In fact, for $\lambda_n \sim \rho(\Sigma)\sigma\sqrt{\frac{\log(n+d)}{n}}$, the lower bound holds with high probability. As $\lambda_n \sim \rho(\Sigma)\sigma\sqrt{\frac{\log(n+d)}{n}}$, $a_2 \sim \rho^2(\Sigma)\frac{\log d}{n}$, we have (62), which shows that the left hand side of Inequality (61) is significantly smaller than the right hand side of Inequality (61) when the dimension is not extremely high.

$$128a_2 s \cdot \left(\frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2 \sim \max\{\frac{\log d}{n}\|\theta^*_{S^c}\|_1, \frac{s^2(\log d)^2}{n^2}\frac{\rho^2(\Sigma)}{\kappa}\}\frac{\lambda_n^2}{8a_2}. \tag{62}$$

REMARK 19. Inequality (59) in Theorem 10 implies the block-wise linear convergence rate within range $[k_0, k_1]$, where

$$\varepsilon_{k_0} \leq \frac{\lambda_n^2}{48a_2} \text{ and } \varepsilon_{k_1} \geq 6 \cdot 128 a_2 s \cdot \left( \frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa} \right)^2.$$

If $a_1 < 8\|\mathbf{X}^T\mathbf{X}/n\|_s$, for $k \geq k_0$, let $T_k = \lfloor (k-k_0)/\lceil \frac{\log 1/6}{\log(1 - \frac{a_1}{8\|\mathbf{X}^T\mathbf{X}/n\|_s})} \rceil \rfloor$. If $a_1 \geq 8\|\mathbf{X}^T\mathbf{X}/n\|_s$, for $k \geq k_0$, let $T_k = k - k_0$. We have

$$F(\theta_k) - F(\hat{\theta}) \leq \max\{2^{-T_k}\varepsilon_{k_0}, 6 \cdot 128 a_2 s \cdot \left( \frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa} \right)^2\}. \tag{63}$$

A more detailed proof of this statement is given in the proof of Theorem 11. In fact, Theorem 10 implies the conventional linear convergence within the range discussed in (61) with properly chosen decay factor. But that involves much more tedious details without giving additional insight, so we do not make that a formal assertion here.

### 5.3. Overall Results

With Theorem 8 and optimization convergence results in Theorem 10, we have Theorem 11 describing how iteration number affects the statistical accuracy.

THEOREM 11. *Let $\rho^2(\Sigma)$ be the maximum diagonal entry of the covariance matrix $\Sigma$. Under the linear regression model* (50)*, for any sparse index set $S$ such that the cardinal of $S$, $|S| = s$, denote $\theta^*_{S^c}$ to be the vector keeping elements not in $S$ the same and setting those in $S$ to be 0. Suppose $c_1\kappa \geq 64 s \cdot c_2 \rho^2(\Sigma) \frac{\log d}{n}$, where $c_1, c_2$ are constants and can be taken as $c_1 = 1/8, c_2 = 50$, and $\kappa$ is the smallest singular value of $\Sigma$. Suppose $\lambda_n \geq 4\rho(\Sigma)\sqrt{1 + \frac{\log d}{n}}\sqrt{\frac{\log 2(n+d)}{n}}$. Use Algorithm 5.1 with step size $\eta = \frac{\|\mathbf{X}^T\mathbf{X}\|_s}{n}$. Let*

$$K_0 = \lceil \frac{48 c_2 \rho^2(\Sigma) \frac{\log d}{n} \left( \|\theta^*\|_2 + \frac{\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa} \right)^2 \|\mathbf{X}^T\mathbf{X}/n\|_s}{2\lambda_n^2} \rceil. \tag{64}$$

*Let*

$$T_k = \begin{cases} \lfloor (k-k_0)/\lceil \dfrac{\log 1/6}{\log(1 - \frac{c_1\kappa}{8\|\mathbf{X}^T\mathbf{X}/n\|_s})} \rceil \rfloor, & \text{when } c_1\kappa < 8\|\mathbf{X}^T\mathbf{X}/n\|_s \\ k - k_0, & \text{otherwise} \end{cases}. \tag{65}$$

*Let*

$$\begin{aligned}\delta_k = &\\ \min\Bigg\{ &\frac{\|\mathbf{X}^T\mathbf{X}/n\|_s}{2k}\left(\frac{\|\theta^*_{S^c}\|_1}{\sqrt{s}}+(2+4\sqrt{s}+\frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}+\|\theta^*\|_2\right)^2,\\ &\max\left\{2^{-T_k}\frac{\lambda_n^2}{48c_2\rho^2(\Sigma)\frac{\log d}{n}},\rho^2(\Sigma)\frac{\log d}{n}s\cdot\left(\frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}}+(2+4\sqrt{s}+\frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2\cdot 768c_2\right\}\\ &+\mathbb{1}\{k\le K_0\}\frac{\|y\|_2^2}{2n}\Bigg\}.\end{aligned} \tag{66}$$

*Then with probability at least* $1-\frac{\exp(-n/32)}{1-\exp(-n/32)}-\exp(-\frac{n}{2})-\frac{1}{2(n+d)}$*, the following statements holds.*

$$\|\theta_k-\theta^*\|_2<\frac{\delta_k}{2\lambda_n\sqrt{s}}+\frac{\|\theta^*_{S^c}\|_1}{\sqrt{s}}+(2+4\sqrt{s}+\frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}. \tag{67}$$

REMARK 20. Theorem 11 shows how the number of iteration affects the statistical accuracy of the computed estimator. It shows that the error caused by optimization goes to zero with the iteration number goes to infinity. Recall that $\lambda_n\sim\sqrt{\frac{\log(n+d)}{n}}$ when $n\ge\log d$, which is satisfied as we do not consider extreme high dimensional case. Note that when the computation resource is infinity, $\|\theta_k-\theta^*\|_2\sim\frac{\|\theta^*_{S^c}\|_1}{\sqrt{s}}+\sqrt{s}\sqrt{\frac{\log(n+d)}{n}}$. When the true vector $\theta^*$ is indeed $s-$sparse, $\|\theta_k-\theta^*\|_2\sim\sqrt{s}\sqrt{\frac{\log(n+d)}{n}}$, which is the optimal rate for high dimensional linear regression.

REMARK 21. From the expression of $\delta_k$ in Inequality (66) and the role of $\delta_k$ on statistical accuracy shown in Inequality (67), the convergence rate of error caused by optimization, $\frac{F(\theta_k)-F(\hat{\theta})}{2\lambda_n\sqrt{s}}$, has convergence rate $\sim\frac{1}{k}$ when

$$\frac{F(\theta_k)-F(\hat{\theta})}{2\lambda_n\sqrt{s}}\ge\frac{\lambda_n}{\sqrt{s}\cdot 96c_2\rho^2(\Sigma)\frac{\log d}{n}}\sim\frac{\sigma}{\rho(\Sigma)}\frac{\sqrt{n\log(n+d)}}{\sqrt{s}\log d},$$

or when

$$\begin{aligned}\frac{F(\theta_k)-F(\hat{\theta})}{2\lambda_n\sqrt{s}}\le&\frac{768c_2\rho^2(\Sigma)\frac{\log d}{n}s}{2\lambda_n\sqrt{s}}\left(\frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}}+(2+4\sqrt{s}+\frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2\\ \sim&\rho^2(\Sigma)\frac{\log d}{n}s\left(\frac{\|\theta^*_{S^c}\|_1^2}{s\sqrt{s}\lambda_n}+\sqrt{s}\frac{\lambda_n}{\kappa^2}\right).\end{aligned}$$

Otherwise, the optimization algorithm has linear convergence rate. Considering the case where $\theta^*_{S^c}=\mathbf{0}$, which is the conventional setting in high dimensional sparse linear regression, we have

that the upper and lower bound for the range where $\frac{F(\theta_k)-F(\hat{\theta})}{2\lambda_n\sqrt{s}}$ has linear convergence are of the order $\frac{n}{s\log d}\frac{\kappa}{\rho^2(\Sigma)}\Delta_{stat}$ and $\frac{s\log d}{n}\frac{\rho^2(\Sigma)}{\kappa}\Delta_{stat}$ respectively, where $\Delta_{stat}=(2+4\sqrt{s}+\frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}$ is the limit statistical accuracy. Therefore, our algorithm performs as well as that in literature (e.g. Loh and Wainwright (2015)) under the classical setting, and is fully convergent in general or in special cases (i.e. sparsity and RSC conditions), which is not shown in Loh and Wainwright (2015) for any cases. This shows that our framework, including statistical-optimization interplay and the template algorithm, automatically adapts to the special cases that has simpler setting admitting stronger assumptions. The optimization convergence results for the general framework, however, need to be further crafted when additional conditions are satisfied.

REMARK 22. Note that the results has $\mathbf{X}$, $y$, and $\theta^*$ involved. $\mathbf{X}$ and $y$ are observable, so we can adjust iteration number accordingly to guarantee the desired accuracy in terms of $\theta^*$. For $\theta^*$, usually we can have a conservative upper bound for $\|\theta^*\|_2$, hence we adjust our iteration number accordingly for the guaranteed accuracy.

## 6. Discussion

In the present work, we proposed a framework for considering the influence of the running time on the statistical accuracy and applied the framework to three examples: 1-bit matrix completion and causal inference for panel data and high dimensional sparse linear regression. We get novel interesting novel results for the first two examples and show that our framework adapts to the degenerate case in the third example. Our backbone statistical analysis for causal panel data is also sharper than that in the literature. It would be interesting to see what results can be derived when our framework is applied to other applicable problems, like kernel ridge regression, SVM, network analysis, neural network, and more intensively studied problems like Danzig selector and elastic net to see how the results compare.

Our framework focuses on estimators that are matrices (and vectors as a special case), but our way of integrating optimization consideration into statistical accuracy before solving the optimization problem can be easily carried to tensors. It would be interesting to see how a parallel tensor version framework performs.

Our framework provides a new perspective of the relationship between computational cost and statistical accuracy, where we quantify the value of computing resource in terms of how much statistical accuracy it can buy, precisely and on a continuous scale. This perspective makes it possible to be used in equilibrium in economic problems, e.g. the computing resource invested is the cost and statistical accuracy generates revenue. It would be interesting to see how it works in those equilibrium and it would also be interesting to further investigate the interplay along this perspective.

Our optimization template algorithm can fill in the blank of theoretically guaranteed optimization algorithm for estimators in a large class of statistical problems that fit in the general form of our framework.

The optimization convergence analysis in our framework provides a pipeline for analyzing an optimization problem to the level meeting statistical needs. It would be interesting to investigate the unanalyzed heuristic algorithms or finer the analysis of other statistic-induced optimization problem to make the constants free from dimension or other statistically important quantities. Also, for our inner loop, we exploited and analyzed the convergence rate of 3-block ADMM, which usually meets the need for statistical problems encountered and can serve as building stone for more blocks, but it would be interesting to investigate the convergence rate for direct multi-block ADMM or its variant under reasonable assumptions.

## Acknowledgments

## References

Athey S, Bayati M, Doudchenko N, Imbens G, Khosravi K (2021) Matrix completion methods for causal panel data models. *Journal of the American Statistical Association* 116:1716–1730.

Beck A (2017) *First-order methods in optimization* (SIAM).

Berthet Q, Rigollet P (2013) Computational lower bounds for sparse pca. *arXiv preprint arXiv:1304.0828* .

Bottou L, Bousquet O (2011) 13 the tradeoffs of large-scale learning. *Optimization for machine learning* 351.

Boucheron S, Lugosi G, Massart P (2013) *Concentration inequalities: A nonasymptotic theory of independence* (Oxford university press).

Cai X, Han D, Yuan X (2017) On the convergence of the direct extension of admm for three-block separable convex minimization models with one strongly convex function. *Computational Optimization and Applications* 66(1):39–73.

Chandrasekaran V, Jordan MI (2013) Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences* 110(13):E1181–E1190.

Chen C, He B, Ye Y, Yuan X (2016) The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming* 155(1):57–79.

Chen Y, Wainwright MJ (2015) Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025* .

Chi Y, Lu YM, Chen Y (2019) Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing* 67(20):5239–5269.

Davenport MA, Plan Y, Van Den Berg E, Wootters M (2014) 1-bit matrix completion. *Information and Inference: A Journal of the IMA* 3(3):189–223.

Hong M, Luo ZQ (2017) On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming* 162(1-2):165–199.

Horev I, Nadler B, Arias-Castro E, Galun M, Basri R (2015) Detection of long edges on a computational budget: A sublinear approach. *SIAM Journal on Imaging Sciences* 8(1):458–483.

Jain P, Netrapalli P, Sanghavi S (2013) Low-rank matrix completion using alternating minimization. *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, 665–674.

Jiang K, Sun D, Toh KC (2012) An inexact accelerated proximal gradient method for large scale linearly constrained convex sdp. *SIAM Journal on Optimization* 22(3):1042–1064.

Kpotufe S, Verma N (2017) Time-accuracy tradeoffs in kernel prediction: controlling prediction quality. *The Journal of Machine Learning Research* 18(1):1443–1471.

Ledoux M, Talagrand M (1991) *Probability in Banach Spaces: isoperimetry and processes*, volume 23 (Springer Science & Business Media).

Lin T, Ma S, Zhang S (2016) Iteration complexity analysis of multi-block admm for a family of convex minimization without strong convexity. *Journal of Scientific Computing* 69(1):52–81.

Lin T, Ma S, Zhang S (2018) Global convergence of unmodified 3-block admm for a class of convex minimization problems. *Journal of Scientific Computing* 76(1):69–88.

Loh PL, Wainwright MJ (2015) Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research* 16(1):559–616.

Schmidt M, Roux NL, Bach F (2011) Convergence rates of inexact proximal-gradient methods for convex optimization. *arXiv preprint arXiv:1109.2415* .

Seginer Y (2000) The expected norm of random matrices. *Combinatorics, Probability and Computing* 9(2):149–166.

Shender D, Lafferty J (2013) Computation-risk tradeoffs for covariance-thresholded regression. *International Conference on Machine Learning*, 756–764 (PMLR).

Sussman DL, Volfovsky A, Airoldi EM (2015) Analyzing statistical and computational tradeoffs of estimation procedures. *arXiv preprint arXiv:1506.07925* .

Tibshirani RJ (2017) Dykstra's algorithm, admm, and coordinate descent: Connections, insights, and extensions. *arXiv preprint arXiv:1705.04768* .

Tropp JA (2012) User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics* 12(4):389–434.

Tropp JA (2015) An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571* .

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

Wang L, Zhang X, Gu Q (2017) A unified computational and statistical framework for nonconvex low-rank matrix estimation. *Artificial Intelligence and Statistics*, 981–990 (PMLR).

Wang T, Berthet Q, Samworth RJ (2016) Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics* 44(5):1896–1930.

# Proofs of Statements

This supplement gives all the proofs of the results in Chapter **??**. We start with proving three overall results for our examples using statistical-optimization-interplay results and optimization results, which are proved later. Next we prove the statistical-optimization interplay results for our examples. Then we prove optimization results for our general optimization template. In the end, we prove the optimization results for our examples.

## EC.1. Proof of Theorem 4

Recall Theorem 2, Theorem 3.

According to Theorem 3, we have

$$\begin{aligned}
&\delta \le \frac{\alpha^2 \tilde{L}_\alpha d_1 d_2}{T} + (4\alpha \tilde{L}_\alpha \sqrt{d_1 d_2} + 2L_\alpha)\sqrt{\frac{1}{t}}\sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} + 2\tilde{L}_\alpha \frac{1}{t}\left(q(\beta) + \frac{2d_1 d_2}{\beta}\right), \\
&\max\{\delta_1, \delta_2, \delta_0\} \le \sqrt{\frac{1}{t}}\sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} \le u_0,
\end{aligned} \tag{EC.1}$$

Therefore, $L_{\alpha+\delta_1} \le 2L_\alpha$.

Combing with Theorem 2 through plugging in the bounds of $\delta, \delta_1, \delta_2$, we have the following holds with probability at least $1 - \frac{c_1}{d_1+d_2}$.

$$\begin{aligned}
&D(l(M)\|l(\tilde{M})) \\
&\le 2c_0 L_\alpha \left(\alpha\sqrt{r d_1 d_2} + \sqrt{\frac{1}{t}}\sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}}\right)\sqrt{\frac{d_1+d_2}{n d_1 d_2}}\sqrt{1 + \frac{(d_1+d_2)\log(d_1 d_2)}{n}} \\
&+ \frac{\alpha^2 \tilde{L}_\alpha d_1 d_2}{Tn} + \frac{4\alpha\tilde{L}_\alpha\sqrt{d_1 d_2} + 2L_\alpha}{n}\sqrt{\frac{1}{t}}\sqrt{q(\beta) + \frac{2d_1 d_2}{\beta}} + \frac{2\tilde{L}_\alpha}{n}\frac{1}{t}\left(q(\beta) + \frac{2d_1 d_2}{\beta}\right).
\end{aligned} \tag{EC.2}$$

## EC.2. Proof of Theorem 7

Note that according to Theorem 6, we have

$$\begin{aligned}
&\delta_1 \le \\
&\sqrt{\frac{q_1(\beta)(\lambda|\mathcal{O}|)^2 \min\{N,T\} + q_2(\beta)C(Y)^2 + q_3(\beta)\left(\|Y\|^2 + 2(NT - |\mathcal{O}|)L_{\max}^2\right)}{k - q_0(\beta)}},
\end{aligned} \tag{EC.3}$$

where $q_0(\beta), q_1(\beta), q_2(\beta), q_3(\beta)$ are defined in Theorem 6.

Noting that

$$C(Y) = \sup_{L \in C_1} \|\mathbf{P}_{\mathcal{O}}(Y - L)\| \leq \|Y\| + L_{\max}\sqrt{|\mathcal{O}|} \leq \sqrt{2\|Y\|^2 + 2L_{\max}^2|\mathcal{O}|}, \tag{EC.4}$$

we have

$$\begin{aligned}
&\delta_1 \leq \\
&\sqrt{\frac{q_1(\beta)(\lambda|\mathcal{O}|)^2 \min\{N,T\} + (2q_2(\beta) + q_3(\beta))\,\|Y\|^2 + (2|\mathcal{O}|(q_2(\beta) - q_3(\beta)) + 2NTq_3(\beta))\,L_{\max}^2}{k - q_0(\beta)}} \\
&\leq \sqrt{\frac{q_1(\beta)(\lambda|\mathcal{O}|)^2 \min\{N,T\} + (2q_2(\beta) + q_3(\beta))\,\|Y\|^2 + 2NT\max\{q_2(\beta), q_3(\beta)\}\,L_{\max}^2}{k - q_0(\beta)}}.
\end{aligned} \tag{EC.5}$$

Note that we have $\lambda|\mathcal{O}| \leq 13 \times 8\sigma \max\{\sqrt{N}, \sqrt{T}\} \log^{\frac{3}{2}}(N+T)$, we have

$$\begin{aligned}
&\delta_1 \leq \\
&\sqrt{\frac{104^2 q_1(\beta)\sigma^2 NT \log^3(N+T) + (2q_2(\beta) + q_3(\beta))\,\|Y\|^2 + 2NT\max\{q_2(\beta), q_3(\beta)\}\,L_{\max}^2}{k - q_0(\beta)}}.
\end{aligned} \tag{EC.6}$$

Let $\widetilde{q_1(\beta)} = 104^2 q_1(\beta)$, $\widetilde{q_2(\beta)} = 2q_2(\beta) + q_3(\beta)$, $\widetilde{q_3(\beta)} = 2\max\{q_2(\beta), q_3(\beta)\}$, then we have

$$\delta_1 \leq \sqrt{\frac{\widetilde{q_1(\beta)}\sigma^2 NT \log^3(N+T) + \widetilde{q_2(\beta)}\|Y\|^2 + \widetilde{q_3(\beta)}NTL_{\max}^2}{k - q_0(\beta)}}. \tag{EC.7}$$

In the proof of Theorem 6, we derive the bound for $\delta_1$ through that of $\delta_0$, the $L_2$ distance between the resulting approximate solution of inner loop and the target exact solution of the inner loop. So the bound in Inequality (EC.7) also holds for $\delta_0$. We set the upper bound for inner loop error $\delta_0$ at iteration number $k$ as

$$\delta(k) = \sqrt{\frac{\widetilde{q_1(\beta)}\sigma^2 NT \log^3(N+T) + \widetilde{q_2(\beta)}\|Y\|^2 + \widetilde{q_3(\beta)}NTL_{\max}^2}{k - q_0(\beta)}}.$$

Invoking outer loop convergence rate, Proposition 4.1, similarly to the proof in Theorem 6, we have the optimization error for objective function as defined in (37) is upper bounded as follows.

$$\delta \leq \frac{NTL_{\max}^2}{|\mathcal{O}|K} + \frac{2\delta(k)^2}{|\mathcal{O}|} + \delta(k)\left(\frac{4L_{\max}\sqrt{NT}}{|\mathcal{O}|} + \frac{2C(Y)}{|\mathcal{O}|} + \min\{\sqrt{N}, \sqrt{T}\}\lambda\right). \tag{EC.8}$$

Using

$$C(Y) \leq \|Y\| + L_{\max}\sqrt{|\mathcal{O}|} \leq \|Y\| + L_{\max}\sqrt{NT} \tag{EC.9}$$

and invoking Theorem 5 we get the statement of Theorem 4.

## EC.3. Proof of Theorem 11

Denote $F(\theta)=\frac{\|\mathbf{X}\theta\|_2^2}{2n}-\lambda_n\|\theta\|_1^2$.

From Inequality (EC.145), Lemma (EC.7), Theorem (8), we know that with probability at least $1-\frac{\exp(-n/32)}{1-\exp(-n/32)}-\exp(-\frac{n}{2})-\frac{1}{2(n+d)}$ the following holds.

$$\|\mathbf{X}^T w\|_\infty < 4\rho(\Sigma)\sqrt{1+\frac{\log d}{n}}\sqrt{\frac{\log 2(n+d)}{n}}, \tag{EC.10}$$

$$\frac{\|\mathbf{X}\theta\|_2^2}{n} \geq c_1\kappa\|\theta\|_2^2 - c_2\rho^2(\Sigma)\frac{\log d}{n}\|\theta\|_1^2, \tag{EC.11}$$

$$\|\theta-\theta^*\| \leq \frac{F(\theta)-F(\hat{\theta})}{2\lambda_n\sqrt{s}} + \frac{\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2+4\sqrt{s}+\frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}, \tag{EC.12}$$

where $c_1, c_2$ are constants and can be taken as $c_1=1/8, c_2=50$.

Therefore, the condition in Theorem 10 is satisfied with

$$a_1 = c_1\kappa, a_2 = c_2\rho^2(\Sigma)\frac{\log d}{n}. \tag{EC.13}$$

We only need to prove that under these conditions $F(\theta_k)-F(\hat{\theta})\leq\delta_k$ holds.

By Inequality (60) in Theorem 10 and Inequality (EC.12) we have

$$F(\theta_k)-F(\hat{\theta}) \leq \frac{\|\mathbf{X}^T\mathbf{X}/n\|_s}{2k}\|\hat{\theta}\|_2^2 \leq \frac{\|\mathbf{X}^T\mathbf{X}/n\|_s}{2k}\left(\|\theta^*\|_2 + \frac{\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2+4\sqrt{s}+\frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2. \tag{EC.14}$$

According to Inequality (EC.160) we know that

$$F(\theta_k)-F(\hat{\theta}) \leq F(\theta_0)-F(\hat{\theta}) \leq \frac{\|y\|_2^2}{2n}. \tag{EC.15}$$

For $k\geq K_0$, Inequality (60) and Inequality (EC.160) gives

$$F(\theta_k)-F(\hat{\theta}) \leq F(\theta_{K_0})-F(\hat{\theta}) \leq \frac{\lambda_n^2}{48c_2\rho(\Sigma)\frac{\log d}{n}}. \tag{EC.16}$$

Now we are only left to prove for $k\geq K_0$,

$$\begin{aligned} &F(\theta_k)-F(\hat{\theta}) \leq \\ &\max\left\{2^{-T_k}\frac{\lambda_n^2}{48c_2\rho^2(\Sigma)\frac{\log d}{n}}, \rho^2(\Sigma)\frac{\log d}{n}s\cdot\left(\frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}}+(2+4\sqrt{s}+\frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2\cdot 768c_2\right\}, \end{aligned} \tag{EC.17}$$

which is also Inequality (63) in Remark 19.

To prove this, we only need to prove that for $k_1 \geq k_0$ and $k$ satisfying

$$k \geq \begin{cases} k_1 + \lceil \dfrac{\log 1/6}{\log\left(1 - \frac{c_1\kappa}{8\|\mathbf{X}^T\mathbf{X}/n\|_s}\right)} \rceil, & \text{when } c_1\kappa < 8\|\mathbf{X}^T\mathbf{X}/n\|_s \\ k_1 + 1, & \text{otherwise} \end{cases}, \tag{EC.18}$$

the following holds

$$\begin{aligned} &F(\theta_k) - F(\hat{\theta}) \leq \\ &\max\{\frac{1}{2}\left(F(\theta_{k_1}) - F(\hat{\theta})\right), \rho^2(\Sigma)\frac{\log d}{n}s \cdot \left(\frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2 \cdot 768c_2\}. \end{aligned} \tag{EC.19}$$

If

$$F(\theta_k) - F(\hat{\theta}) \geq \rho^2(\Sigma)\frac{\log d}{n}s \cdot \left(\frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2 \cdot 768c_2, \tag{EC.20}$$

then $F(\theta_{k_1}) - F(\hat{\theta}) \geq \rho^2(\Sigma)\frac{\log d}{n}s \cdot \left(\frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2 \cdot 768c_2$.

Also, since $k_1 \geq K_0$, we have

$$F(\theta_{k_1}) - F(\hat{\theta}) \leq F(\theta_{K_0}) - F(\hat{\theta}) \leq \frac{\lambda_n^2}{48c_2\rho^(\Sigma)\frac{\log d}{n}}. \tag{EC.21}$$

By Inequality (59), we have

$$\begin{aligned} F(\theta_k) - F(\hat{\theta}) \leq &\frac{1}{6}\left(F(\theta_{k_1}) - F(\hat{\theta})\right) \\ &+ \rho^2(\Sigma)\frac{\log d}{n}s \cdot \left(\frac{2\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}\right)^2 \cdot 128c_2 \\ &+ \frac{8c_2\rho^2(\Sigma)\frac{\log d}{n}}{\lambda_n^2}\frac{\lambda_n^2}{48c_2\rho^(\Sigma)\frac{\log d}{n}}\left(F(\theta_{k_1}) - F(\hat{\theta})\right) \\ \leq &\frac{1}{2}\left(F(\theta_{k_1}) - F(\hat{\theta})\right). \end{aligned} \tag{EC.22}$$

Thus we concludes the proof.

## EC.4. Proof of Theorem 2

The structure of the proof is similar to the proof of Theorem 2 in Davenport et al. (2014), but to show how the statistical-optimization interface work, we will show in details how the optimization error terms get into the statistical accuracy.

Let

$$\bar{\mathcal{L}}_{\Omega,Y}(X) = \mathcal{L}_{\Omega,Y}(X) - \mathcal{L}_{\Omega,Y}(\mathbf{0}). \tag{EC.23}$$

Then we know that

$$-\bar{\mathcal{L}}_{\Omega,Y}(\tilde{M}) \le -\bar{\mathcal{L}}_{\Omega,Y}(\hat{M}) + \delta \le -\bar{\mathcal{L}}_{\Omega,Y}(M) + \delta. \tag{EC.24}$$

We also know that

$$\|\tilde{M}\|_* \le \alpha\sqrt{rd_1d_2} + \delta_2, \|\tilde{M}\|_\infty \le \alpha + \delta_1. \tag{EC.25}$$

We have the following lemma, which we will proof later in this section.

LEMMA EC.1. *Let* $G \in \mathbb{R}^{d_1\times d_2}$ *be*

$$G = \{X \in R^{d_1\times d_2} : \|X\|_* \le \alpha\sqrt{rd_1d_2} + \delta_2, \|\tilde{M}\|_\infty \le \alpha + \delta_1\} \tag{EC.26}$$

*for some* $r \le \min\{d_1, d_2\}$ *and* $\alpha \ge 0$. *Then*

$$\begin{aligned}&\mathbb{P}\left(\sup_{X\in G} |\bar{\mathcal{L}}_{\Omega,Y}(X) - \mathbb{E}\bar{\mathcal{L}}_{\Omega,Y}(X)| \ge \tilde{c}_0 L_{\alpha+\delta_1}\left(\alpha\sqrt{rd_1d_2} + \delta_2\right)\sqrt{\frac{n(d_1+d_2)}{d_1d_2} + \log(d_1d_2)}\right)\\ &\le \frac{c_1}{d_1+d_2},\end{aligned} \tag{EC.27}$$

*where* $\tilde{c}_0, c_1$ *are absolute constants and the probability and the expectation are both over the choice of* $\Omega$ *and draw of* $Y$.

Note that for any $X$ we have

$$\begin{aligned}&\mathbb{E}\left(\bar{\mathcal{L}}_{\Omega,Y}(X) - \bar{\mathcal{L}}_{\Omega,Y}(M)\right)\\ =&\frac{n}{d_1d_2}\sum_{i,j}\left(l(M_{i,j})\log\left(\frac{l(X_{i,j})}{l(M_{i,j})}\right) + \log\left(\frac{1-l(X_{i,j})}{1-l(M_{i,j})}\right)\right)\\ =&-nD\left(l(M)\|l(X)\right).\end{aligned} \tag{EC.28}$$

Therefore, we have

$$\begin{aligned}&-\delta\\ &\le \bar{\mathcal{L}}_{\Omega,Y}(\tilde{M}) - \bar{\mathcal{L}}_{\Omega,Y}(M)\\ &= \mathbb{E}\left(\bar{\mathcal{L}}_{\Omega,Y}(\tilde{M}) - \bar{\mathcal{L}}_{\Omega,Y}(M)\right) + \left(\bar{\mathcal{L}}_{\Omega,Y}(\tilde{M}) - \mathbb{E}\left(\bar{\mathcal{L}}_{\Omega,Y}(\tilde{M})\right)\right) - \left(\bar{\mathcal{L}}_{\Omega,Y}(M) - \mathbb{E}\left(\bar{\mathcal{L}}_{\Omega,Y}(M)\right)\right)\\ &\le \mathbb{E}\left(\bar{\mathcal{L}}_{\Omega,Y}(\tilde{M}) - \bar{\mathcal{L}}_{\Omega,Y}(M)\right) + 2\sup_{X\in G}\left|\bar{\mathcal{L}}_{\Omega,Y}(X) - \mathbb{E}\left(\bar{\mathcal{L}}_{\Omega,Y}(X)\right)\right|\\ &= -nD\left(l(M)\|l(\tilde{M})\right) + 2\sup_{X\in G}\left|\bar{\mathcal{L}}_{\Omega,Y}(X) - \mathbb{E}\left(\bar{\mathcal{L}}_{\Omega,Y}(X)\right)\right|,\end{aligned} \tag{EC.29}$$

where $G$ is defined in (EC.26).

Applying Lemma EC.1, we have that with probability at least $1-\frac{c_1}{d_1+d_2}$

$$\begin{aligned}
&D\left(l(M)\|l(\tilde{M})\right)\\
&\leq \frac{2}{n}\tilde{c}_0 L_{\alpha+\delta_1}\left(\alpha\sqrt{rd_1d_2}+\delta_2\right)\sqrt{\frac{n(d_1+d_2)}{d_1d_2}+\log\left(d_1d_2\right)}+\frac{\delta}{n} \qquad \text{(EC.30)}\\
&\leq 2\tilde{c}_0 L_{\alpha+\delta_1}\left(\alpha\sqrt{rd_1d_2}+\delta_2\right)\sqrt{\frac{d_1+d_2}{nd_1d_2}}\sqrt{1+\frac{(d_1+d_2)\log\left(d_1d_2\right)}{n}}+\frac{\delta}{n}.
\end{aligned}$$

Let $c_0=2\tilde{c}_0$ we have the theorem.

### EC.4.1. Proof of Lemma EC.1

Noting that

$$\begin{aligned}
&\bar{\mathcal{L}}_{\Omega,Y}(X)=\\
&\sum_{(i,j)}\mathbb{1}\{(i,j)\in\Omega\}\left(\mathbb{1}\{Y_{i,j}=1\}\log\left(\frac{l(X_{i,j})}{l(0)}\right)+\mathbb{1}\{Y_{i,j}=-1\}\log\left(\frac{1-l(X_{i,j})}{1-l(0)}\right)\right),
\end{aligned} \qquad \text{(EC.31)}$$

by symmetrization (i.e Lemma 6.3 in Ledoux and Talagrand (1991)) we have

$$\begin{aligned}
\mathbb{E}\left(\sup_{X\in G}|\bar{\mathcal{L}}_{\Omega,Y}(X)-\mathbb{E}\bar{\mathcal{L}}_{\Omega,Y}(X)|^h\right)\leq 2^h\mathbb{E}\Bigg(\sup_{X\in G}\Bigg|\sum_{(i,j)}\zeta_{i,j}\mathbb{1}\{(i,j)\in\Omega\}\\
\left(\mathbb{1}\{Y_{i,j}=1\}\log\left(\frac{l(X_{i,j})}{l(0)}\right)+\mathbb{1}\{Y_{i,j}=-1\}\log\left(\frac{1-l(X_{i,j})}{1-l(0)}\right)\right)\Bigg|^h\Bigg),
\end{aligned} \qquad \text{(EC.32)}$$

where $\zeta_{i,j}$ are i.i.d. Rademacher random variables and the expectation is with respect to $\Omega$, $Y$ and $\zeta_{i,j}$. Next is to apply the contraction principle (i.e. Theorem 4.12 in Ledoux and Talagrand (1991)). By the definition of $L_{\alpha+\delta_1}$ and definition of $G$, we know that

$$\frac{1}{L_{\alpha+\delta_1}}\log\left(\frac{l(x)}{l(0)}\right)\text{ and }\frac{1}{L_{\alpha+\delta_1}}\log\left(\frac{1-l(x)}{1-l(0)}\right)$$

are contractions that vanish at 0 within the domain of any $X_{i,j}$ such that $X\in G$. Invoking contraction principle gives

$$\begin{aligned}
&\mathbb{E}\left(\sup_{X\in G}|\bar{\mathcal{L}}_{\Omega,Y}(X)-\mathbb{E}\bar{\mathcal{L}}_{\Omega,Y}(X)|^h\right)\\
&\leq 2^h(2L_{\alpha+\delta_1})^h\mathbb{E}\left(\sup_{X\in G}\left|\sum_{(i,j)}\zeta_{i,j}\mathbb{1}\{(i,j)\in\Omega\}\left(\mathbb{1}\{Y_{i,j}=1\}X_{i,j}-\mathbb{1}\{Y_{i,j}=-1\}X_{i,j}\right)\right|^h\right) \qquad \text{(EC.33)}\\
&\leq (4L_{\alpha+\delta_1})^h\mathbb{E}\left(\sup_{X\in G}|\langle\Delta_\Omega\circ Z\circ Y,X\rangle|^h\right),
\end{aligned}$$

where $Z$ denotes the matrix with $(i,j)$th element being $\zeta_{i,j}$, $\Delta_\Omega$ denotes the indicator matrix for $\Omega$ such that elements are zero when not in $\Omega$ and 1 when in $\Omega$, and $\circ$ denotes Hadamard product. Observing that $Z \circ Y$ has the same distribution with $Z$, $(Z, Z\circ Y) \perp\!\!\!\perp \Omega$ and $\langle A, B\rangle \le \|A\|_{op}\|B\|_*$, we have

$$\begin{aligned}\mathbb{E}\left(\sup_{X\in G}|\langle \Delta_\Omega \circ Z\circ Y, X\rangle|^h\right) &= \mathbb{E}\left(\sup_{X\in G}|\langle \Delta_\Omega \circ Z, X\rangle|^h\right)\\ \le \mathbb{E}\left(\sup_{X\in G}\|\Delta_\Omega\circ Z\|_{op}^h\|X\|_*^h\right) &= \left(\alpha\sqrt{rd_1d_2}+\delta_2\right)^h\mathbb{E}\left(\|\Delta_\Omega\circ Z\|_{op}^h\right).\end{aligned} \tag{EC.34}$$

Observe that $Z\circ\Delta_\Omega$ is a matrix with i.i.d. symmetric random variables, so according to Theorem 1.1 in Seginer (2000) there is absolute constant C such that for $h \le 2\log(\max\{d_1,d_2\})$ we have

$$\mathbb{E}\left(\|Z\circ\Delta_\Omega\|^h\right) \le C\left(\mathbb{E}\left(\max_{1\le i\le d_1}\left(\sum_{j=1}^{d_2}\Delta_{i,j}\right)^{h/2}\right)+\mathbb{E}\left(\max_{1\le j\le d_2}\sum_{i=1}^{d_1}\Delta_{i,j}\right)^{h/2}\right). \tag{EC.35}$$

Note that $\left(\mathbb{E}(|f|^{h/2})\right)^{2/h}$ is a norm for $h\ge 2$ and $(a+b)^{1/h}\le a^{1/h}+b^{1/h}$, so we have

$$\begin{aligned}&\left(\|Z\circ\Delta_\Omega\|_{op}^h\right)^{1/h}\\ &\le C^{1/h}\left(\left(\mathbb{E}\left[\left(\max_{1\le j\le d_1}\sum_{i=1}^{d_2}\Delta_{i,j}\right)^{h/2}\right]\right)^{1/h}+\left(\mathbb{E}\left[\left(\max_{1\le j\le d_2}\sum_{i=1}^{d_1}\Delta_{i,j}\right)^{h/2}\right]\right)^{1/h}\right)\\ &\le C^{1/h}\left(\mathbb{E}\left[\left(\max_{1\le j\le d_1}\left|\sum_{i=1}^{d_2}\left(\Delta_{i,j}-\frac{n}{d_1d_2}\right)\right|+\frac{n}{d_1}\right)^{h/2}\right]\right)^{1/h}+\\ &\quad C^{1/h}\left(\mathbb{E}\left[\left(\max_{1\le j\le d_2}\left|\sum_{i=1}^{d_1}\left(\Delta_{i,j}-\frac{n}{d_1d_2}\right)\right|+\frac{n}{d_2}\right)^{h/2}\right]\right)^{1/h}\\ &\le C^{1/h}\left(\sqrt{\frac{n}{d_1}}+\sqrt{\frac{n}{d_2}}\right)+C^{1/h}\left(\mathbb{E}\left[\left(\max_{1\le j\le d_1}\left|\sum_{i=1}^{d_2}\left(\Delta_{i,j}-\frac{n}{d_1d_2}\right)\right|\right)^{h/2}\right]\right)^{1/h}+\\ &\quad C^{1/h}\left(\mathbb{E}\left[\left(\max_{1\le j\le d_2}\left|\sum_{i=1}^{d_1}\left(\Delta_{i,j}-\frac{n}{d_1d_2}\right)\right|\right)^{h/2}\right]\right)^{1/h}.\end{aligned} \tag{EC.36}$$

Using Bernstein's inequality, we have for $t>0$

$$\mathbb{P}\left(\left|\sum_{j=1}^{d_2}\left(\Delta_{i,j}-\frac{n}{d_1d_2}\right)\right|>t\right)\le 2\exp\left(\frac{-\frac{t^2}{2}}{\frac{n}{d_1}+\frac{t}{3}}\right). \tag{EC.37}$$

For $t\ge\frac{6n}{d_1}$, for each i, we have

$$\mathbb{P}\left(\left|\sum_{j=1}^{d_2}\left(\Delta_{i,j}-\frac{n}{d_1d_2}\right)\right|>t\right)\le 2\exp(-t)=2\mathbb{P}(W_i>t), \tag{EC.38}$$

where $W_1, \ldots, W_{d_1}$ are i.i.d. exponential random variables.

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\left(\max_{1\le j\le d_2}\left|\sum_{i=1}^{d_1}\left(\Delta_{i,j}-\frac{n}{d_1d_2}\right)\right|\right)^{h/2}\right]\\
&=\int_0^\infty \mathbb{P}\left(\max_{1\le i\le d_1}\left|\sum_{j=1}^{d_2}\left(\Delta_{i,j}-\frac{n}{d_1d_2}\right)\right|^h\right)dt\\
&\le\left(\frac{6n}{d_1}\right)^h+2\int_{\left(\frac{6n}{d_1}\right)^h}^\infty \mathbb{P}\left(\max_{1\le i\le d_1} W_i^h\ge t\right)dt\\
&\le\left(\frac{6n}{d_1}\right)^h+2\mathbb{E}\left[\left(\max_{1\le i\le d_1} W_i\right)^h\right].
\end{aligned}\tag{EC.39}$$

Note that for i.i.d. exponential random variables $W_1, \ldots, W_{d_1}$ we have

$$\begin{aligned}
\mathbb{E}\left[\left(\max_{1\le i\le d_1} W_i\right)^h\right]&\le\mathbb{E}\left[\left(\max_{1\le i\le d_1} W_i^h-\log d_1\right)_+^h\right]+\log\left(d_1\right)^h\\
&\le 2h!+\log^h\left(d_1\right).
\end{aligned}\tag{EC.40}$$

Therefore, we have

$$\begin{aligned}
&\left(\mathbb{E}\left[\left(\max_{1\le j\le d_2}\left|\sum_{i=1}^{d_1}\left(\Delta_{i,j}-\frac{n}{d_1d_2}\right)\right|\right)^{h/2}\right]\right)^{1/h}\\
&\le(1+\sqrt{6})\sqrt{\frac{n}{d_1}}+2^{1/2h}\left(\sqrt{d_1}+2^{1/2h}\sqrt{h}\right)\\
&\le(1+\sqrt{6})\sqrt{\frac{n}{d_1}}+\left(2+\sqrt{2}\right)\sqrt{\log\left(d_1+d_2\right)},
\end{aligned}\tag{EC.41}$$

where in the last inequality we use $h=\log\left(d_1+d_2\right)\ge 1$. It's easy to check that this choice of $h$ satisfies the condition required for getting Inequality (EC.35).

Using similar argument to bound the third term in the right hand side of the last inequality in Inequality (EC.36), we have

$$\begin{aligned}
\left(\mathbb{E}\left[\left\|\Delta_\Omega\circ Z\right\|_{op}^h\right]\right)^{1/h}&\le C^{1/h}\left((1+\sqrt{6})\left(\sqrt{\frac{n}{d_1}}+\sqrt{\frac{n}{d_2}}\right)+(4+2\sqrt{2})\sqrt{\log\left(d_1+d_2\right)}\right)\\
&\le C^{1/h}\sqrt{\frac{n}{d_1}+\frac{n}{d_2}+\log\left(d_1+d_2\right)}\sqrt{(1+\sqrt{6})^2+4+2\sqrt{2}}\\
&<9C^{1/h}\sqrt{\frac{n}{d_1}+\frac{n}{d_2}+\log\left(d_1+d_2\right)}
\end{aligned}\tag{EC.42}$$

Combing Inequality (EC.33),(EC.34),(EC.42), we have

$$\begin{aligned}&\left(\mathbb{E}\left(\sup_{X\in G}|\bar{\mathcal{L}}_{\Omega,Y}(X)-\mathbb{E}\bar{\mathcal{L}}_{\Omega,Y}(X)|^h\right)\right)^{1/h}\\&\le 4L_{\alpha+\delta_1}\left(\alpha\sqrt{rd_1d_2}+\delta_2\right)\times 9C^{1/h}\sqrt{\frac{n}{d_1}+\frac{n}{d_2}+\log(d_1+d_2)}.\end{aligned}\tag{EC.43}$$

Let $t=4L_{\alpha+\delta_1}\left(\alpha\sqrt{rd_1d_2}+\delta_2\right)\times 9\sqrt{\frac{n}{d_1}+\frac{n}{d_2}+\log(d_1+d_2)}\times e$. Then we know that

$$\begin{aligned}&\mathbb{P}\left(\sup_{X\in G}|\bar{\mathcal{L}}_{\Omega,Y}(X)-\mathbb{E}\bar{\mathcal{L}}_{\Omega,Y}(X)|\ge t\right)\\&\le C\exp(-h)=\frac{C}{d_1+d_2}.\end{aligned}\tag{EC.44}$$

Set $\tilde{c}_0=4\times 9\times e$, $c_1=C$, we have the lemma.

## EC.5. Proof of Theorem 5

Denote $\mathbf{A}_{it}$ to be the matrix with element $(i,t)$ being 1 and others being 0. Denote $\varepsilon_{it}$ to the $(i,t)$-th element of $\boldsymbol{\varepsilon}$. Let $\mathfrak{E}=\sum_{(i,t)\in\mathcal{O}}\varepsilon_{it}\mathbf{A}_{it}$. And $\|\cdot\|_{op}$ denotes the operator norm (i.e. the largest singular value).

The overall structure of the proof is similar to that in Athey et al. (2021), we have three main lemmas, which we will prove later. The first two lemmas primarily show how the optimization error comes in, and for the third lemma, we do the statistical analysis differently and have improved rate than that in Athey et al. (2021). The three lemmas are as follows.

LEMMA EC.2. *For all* $\lambda\ge 3\|\mathfrak{E}\|_{op}/|\mathcal{O}|$,

$$\sum_{(i,t)\in\mathcal{O}}\frac{\langle\mathbf{A}_{it},\mathbf{L}^*-\tilde{\mathbf{L}}\rangle^2}{|\mathcal{O}|}\le 10\sqrt{2R}\lambda\|\mathbf{L}^*-\tilde{\mathbf{L}}\|_F+6\delta.\tag{EC.45}$$

LEMMA EC.3. *With probability at least* $1-\frac{1}{(N+T)^2}$, *we have*

$$\|\mathfrak{E}\|_{op}\le 4\sigma\max\{\sqrt{N\log(N+T)},8\sqrt{T}\log^{\frac{3}{2}}(N+T)\}+\sigma.\tag{EC.46}$$

LEMMA EC.4. *Suppose* $\lambda\ge 3\|\mathfrak{E}\|_{op}/|\mathcal{O}|$.

*Then when* $\|\tilde{\mathbf{L}}-\mathbf{L}^*\|_F^2\ge 132(L_{max}+\delta_1)^2\times T\log(N+T)\frac{1}{p_c}$,

$$\begin{aligned}\mathbb{P}_\pi\Bigg(&\frac{\|\tilde{\mathbf{L}}-\mathbf{L}^*\|_F^2p_c}{6}>\sum_{(i,t)\in\mathcal{O}}\langle A_{it},\tilde{\mathbf{L}}-\mathbf{L}^*\rangle^2+3648\frac{72R}{p_c}(\sqrt{N}+\sqrt{T})^2(4(L_{max}+\delta_1)^2)\\&+\frac{432\delta(L_{max}+\delta_1)}{\lambda}(\sqrt{N}+\sqrt{T})\Bigg)\le\frac{1}{(N+T)^3}\end{aligned}\tag{EC.47}$$

Therefore, when $\lambda \geq \frac{12\sigma \max\{\sqrt{N\log(N+T)}, 8\sqrt{T}\log^{\frac{3}{2}}(N+T)\}+3\sigma}{|\mathcal{O}|}$, if $\|\tilde{\mathbf{L}}-\mathbf{L}^*\|_F^2 \geq 132(L_{max}+\delta_1)^2 \times T\log(N+T)\frac{1}{p_c}$, then with probability at least $1-\frac{2}{(N+T)^2}$,

$$\begin{aligned}\frac{\|\tilde{\mathbf{L}}-\mathbf{L}^*\|_F^2 p_c}{6} \leq & \sum_{(i,t)\in\mathcal{O}} \langle A_{it}, \tilde{\mathbf{L}}-\mathbf{L}^*\rangle^2 + 3648\frac{72R}{p_c}(\sqrt{N}+\sqrt{T})^2(4(L_{max}+\delta_1)^2) \\ & + \frac{432\delta(L_{max}+\delta_1)}{\lambda}(\sqrt{N}+\sqrt{T}) \\ \leq & 10\sqrt{2R}(\lambda|\mathcal{O}|)\|\tilde{\mathbf{L}}-\mathbf{L}^*\|_F + 6\delta|\mathcal{O}| + 3648\frac{72R}{p_c}(\sqrt{N}+\sqrt{T})^2(4(L_{max}+\delta_1)^2) \\ & + \frac{432\delta(L_{max}+\delta_1)}{\lambda}(\sqrt{N}+\sqrt{T}).\end{aligned} \tag{EC.48}$$

Note that

$$10\sqrt{2R}(\lambda|\mathcal{O}|)\|\mathbf{L}^*-\tilde{\mathbf{L}}\|_F \leq \frac{12\times 200R\left(\lambda|\mathcal{O}|\right)^2}{p_c} + \frac{\|\tilde{\mathbf{L}}-\mathbf{L}^*\|_F^2 p_c}{12}, \tag{EC.49}$$

and $|\mathcal{O}| \leq NT$.

We take $\lambda = \frac{13\sigma\max\{\sqrt{N\log(N+T)}, 8\sqrt{T}\log^{\frac{3}{2}}(N+T)\}}{|\mathcal{O}|}$.

Move the $\frac{\|\tilde{\mathbf{L}}-\mathbf{L}^*\|_F^2 p_c}{12}$ term from the right hand side to the left hand side and then divide both sides with $\frac{p_c NT}{12}$, we have there are constants $q_0$, $q_1$, $q_2$, such that

$$\begin{aligned}\frac{\|\tilde{\mathbf{L}}-\mathbf{L}^*\|_F^2}{NT} \leq & q_0\frac{R\sigma^2}{p_c^2}\frac{(N+T)\log^3(N+T)}{NT} + \frac{72}{p_c}\delta + q_1\frac{\delta(L_{max}+\delta_1)}{\sigma p_c}\frac{1}{NT} \\ & + q_2\frac{R(L_{max}+\delta_1)^2}{p_c^2}\frac{N+T}{NT}.\end{aligned} \tag{EC.50}$$

## EC.5.1. Proof of Lemma EC.2

By the definition of $\tilde{\mathbf{L}}$, $\hat{\mathbf{L}}$, $\mathbf{L}^*$, we have

$$\begin{aligned}& \sum_{(i,t)\in\mathcal{O}} \frac{\langle Y_{it}-\tilde{\mathbf{L}}\rangle^2}{|\mathcal{O}|} + \lambda|\tilde{\mathbf{L}}|_* \\ \leq & \sum_{(i,t)\in\mathcal{O}} \frac{\langle Y_{it}-\hat{\mathbf{L}}\rangle^2}{|\mathcal{O}|} + \lambda|\hat{\mathbf{L}}|_* + \delta \\ \leq & \sum_{(i,t)\in\mathcal{O}} \frac{\langle Y_{it}-\mathbf{L}^*\rangle^2}{|\mathcal{O}|} + \lambda|\mathbf{L}^*|_* + \delta.\end{aligned} \tag{EC.51}$$

Therefore, we have

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \mathbf{L}^*-\tilde{\mathbf{L}}, \mathbf{A}_{it}\rangle^2}{|\mathcal{O}|} + 2\sum_{(i,t)\in\mathcal{O}} \frac{\varepsilon_{it}\langle \mathbf{L}^*-\tilde{\mathbf{L}}, \mathbf{A}_{it}\rangle}{|\mathcal{O}|} \leq \lambda\|\mathbf{L}^*\|_* - \lambda\|\tilde{\mathbf{L}}\|_* + \delta. \tag{EC.52}$$

Denoting $\boldsymbol{\Delta} = \mathbf{L}^* - \tilde{\mathbf{L}}$, Inequality (EC.52) becomes

$$\begin{aligned}\sum_{(i,t)\in\mathcal{O}} \frac{\langle \boldsymbol{\Delta}, \mathbf{A}_{it}\rangle^2}{|\mathcal{O}|} &\le -\frac{2}{|\mathcal{O}|}\langle \boldsymbol{\Delta}, \mathfrak{E}\rangle + \lambda\|\mathbf{L}^*\|_* - \lambda\|\tilde{\mathbf{L}}\|_* + \delta \\ &\le \frac{2}{|\mathcal{O}|}\|\boldsymbol{\Delta}\|_*\|\mathfrak{E}\|_{op} + \lambda\|\mathbf{L}^*\|_* - \lambda\|\tilde{\mathbf{L}}\|_* + \delta \\ &\le \frac{5}{3}\lambda\|\boldsymbol{\Delta}\|_* + \delta,\end{aligned} \tag{EC.53}$$

the inequalities in which are due to the duality of operator norm and nuclear norm, and the range of $\lambda$.

Now we state the following lemma, which is proved later in this section.

LEMMA EC.5. *Let $\boldsymbol{\Delta} = \mathbf{L}^* - \tilde{\mathbf{L}}$ for $\lambda \ge 3\|\mathfrak{E}\|_{op}/|\mathcal{O}|$ Then there exist a decomposition $\boldsymbol{\Delta} = \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2$ such that*

1. $\langle \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2\rangle = 0$,
2. $rank(\boldsymbol{\Delta}_1) \le 2R$,
3. $\|\boldsymbol{\Delta}_2\|_* \le 5\|\boldsymbol{\Delta}_1\|_* + \frac{3\delta}{\lambda}$.

Now, invoking the decomposition $\boldsymbol{\Delta} = \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2$, we have

$$\|\boldsymbol{\Delta}\|_* \le 6\|\boldsymbol{\Delta}_1\|_* + \frac{3\delta}{\lambda} \le 6\sqrt{2R}\|\boldsymbol{\Delta}_1\|_F + \frac{3\delta}{\lambda} \le 6\sqrt{2R}\|\boldsymbol{\Delta}\|_F + \frac{3\delta}{\lambda}. \tag{EC.54}$$

Plugging Inequality (EC.54) back to Inequality (EC.53), we have

$$\sum_{(i,t)\in\mathcal{O}} \frac{\langle \boldsymbol{\Delta}, \mathbf{A}_{it}\rangle^2}{|\mathcal{O}|} \le 10\sqrt{2R}\lambda\|\boldsymbol{\Delta}\|_F + 6\delta. \tag{EC.55}$$

*Proof of Lemma EC.5.* Let $\mathbf{L}^* = \mathbf{U}_{N\times R}\mathbf{S}_{R\times R}\left(\mathbf{V}_{T\times R}\right)^T$ be the singular value decomposition for the at most rank R matrix $\mathbf{L}^*$. Let $\mathbf{P_U} = \mathbf{U}\mathbf{U}^T$, $\mathbf{P}_{\mathbf{U}^\perp} = \mathbf{U}^\perp(\mathbf{U}^\perp)^T$, $\mathbf{P_V} = \mathbf{V}\mathbf{V}^T$, $\mathbf{P}_{\mathbf{V}^\perp} = \mathbf{V}^\perp(\mathbf{V}^\perp)^T$. Let $\boldsymbol{\Delta}_2 = \mathbf{P}_{\mathbf{U}^\perp}\boldsymbol{\Delta}\mathbf{P}_{\mathbf{V}^\perp}$, $\boldsymbol{\Delta}_1 = \boldsymbol{\Delta} - \boldsymbol{\Delta}_2$.

It's easy to see that $\mathbf{P_U} + \mathbf{P}_{\mathbf{U}^\perp} = \mathbf{I_N}$ and $\mathbf{P_V} + \mathbf{P}_{\mathbf{V}^\perp} = \mathbf{I_T}$.

Now we check the three claims for Lemma EC.5.

$$\begin{aligned}\langle \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2\rangle &= \langle \boldsymbol{\Delta} - \mathbf{P}_{\mathbf{U}^\perp}\boldsymbol{\Delta}\mathbf{P}_{\mathbf{V}^\perp}, \mathbf{P}_{\mathbf{U}^\perp}\boldsymbol{\Delta}\mathbf{P}_{\mathbf{V}^\perp}\rangle \\ &= \langle \mathbf{P_U}\boldsymbol{\Delta} + \mathbf{P}_{\mathbf{U}^\perp}\boldsymbol{\Delta}\mathbf{P_V}, \mathbf{P}_{\mathbf{U}^\perp}\boldsymbol{\Delta}\mathbf{P}_{\mathbf{V}^\perp}\rangle \\ &= 0.\end{aligned} \tag{EC.56}$$

$$rank(\mathbf{\Delta}_1) = rank(\mathbf{P}_{\mathbf{U}}\mathbf{\Delta} + \mathbf{P}_{\mathbf{U}^\perp}\mathbf{\Delta}\mathbf{P}_{\mathbf{V}}) \le rank(\mathbf{P}_{\mathbf{U}}\mathbf{\Delta}) + rank(\mathbf{P}_{\mathbf{U}^\perp}\mathbf{\Delta}\mathbf{P}_{\mathbf{V}}) \le 2R. \tag{EC.57}$$

For the third one, note that

$$\begin{aligned} \langle \mathbf{\Delta}_2, \mathbf{L}^* \rangle &= \langle \mathbf{P}_{\mathbf{U}^\perp}\mathbf{\Delta}\mathbf{P}_{\mathbf{V}^\perp}, \mathbf{U}_{N\times R}\mathbf{S}_{R\times R}\left(\mathbf{V}_{T\times R}\right)^T \rangle \\ &= 0. \end{aligned} \tag{EC.58}$$

And Inequality (EC.53) implies that

$$\begin{aligned} \lambda\left(\|\tilde{\mathbf{L}}\|_* - \|\mathbf{L}^*\|_*\right) &\le \frac{2}{|\mathcal{O}|}\|\mathbf{\Delta}\|_*\|\mathfrak{E}\|_{op} + \delta \\ &\le \frac{2}{3}\lambda\|\mathbf{\Delta}\|_* + \delta \le \frac{2}{3}\lambda\left(\|\mathbf{\Delta}_1\|_* + \|\mathbf{\Delta}_2\|_*\right) + \delta. \end{aligned} \tag{EC.59}$$

The main part of the left hand sided is lower bound by

$$\begin{aligned} \|\tilde{\mathbf{L}}\|_* - \|\mathbf{L}^*\|_* &= \|\mathbf{L}^* - \mathbf{\Delta}_1 - \mathbf{\Delta}_2\|_* - \|\mathbf{L}^*\|_* \ge \|\mathbf{L}^* - \mathbf{\Delta}_1\|_* - \|\mathbf{\Delta}_2\|_* - \|\mathbf{L}^*\|_* \\ &= \|\mathbf{L}^*\|_* + \|\mathbf{\Delta}_1\|_* - \|\mathbf{\Delta}_2\|_* - \|\mathbf{L}^*\|_* = \|\mathbf{\Delta}_1\|_* - \|\mathbf{\Delta}_2\|_*. \end{aligned} \tag{EC.60}$$

Combining Inequality (EC.59) and (EC.60), we have

$$\|\mathbf{\Delta}_2\|_* \le 5\|\mathbf{\Delta}_1\|_* + \frac{3\delta}{\lambda}. \tag{EC.61}$$

**EC.5.1.1. Proof of Lemma EC.3** The proof is very similar to that of lemma 2 in Athey et al. (2021), but our task is to write out the constants explicitly and have the bound as tight as possible.

Although the major parts are very similar, we still write out all the steps for completeness.

The goal is to invoke matrix version Bernstein inequality, a proof of which is in Tropp (2012). Proposition EC.6.1 states the matrix version Bernstein inequality.

MATRIX BERNSTEIN INEQUALITY Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_N$ be independent matrices in $\mathbb{R}^{d_1\times d_2}$ such that $\mathbb{E}[\mathbf{Z}_i] = \mathbf{0}$ and $\|\mathbf{Z}_i\|_{op} \le D$ almost surely for all $i \in [N]$. Let $\sigma_Z$ be such that

$$\sigma_Z^2 \ge \max\left\{\left\|\sum_{i=1}^N \mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i^T]\right\|_{op}, \left\|\sum_{i=1}^N \mathbb{E}[\mathbf{Z}_i^T\mathbf{Z}_i]\right\|_{op}\right\}.$$

Then, for any $\alpha \ge 0$,

$$\mathbb{P}\left\{\left\|\sum_{i=1}^N \mathbf{Z}_i\right\|_{op} \ge \alpha\right\} \le (d_1+d_2)\exp\left[\frac{-\alpha^2}{2\sigma_Z^2 + (2D\alpha)/3}\right]. \tag{EC.62}$$

Same as the notations in Athey et al. (2021), define in dependent random matrices $\mathbf{B}_1, \ldots, \mathbf{B}_N$ as follows. For $1 \le i \le N$, define

$$\mathbf{B}_i = \sum_{t=1}^{t_i} \varepsilon_{it} \mathbf{A}_{it}.$$

Then, $\mathfrak{E} = \sum_{i=1}^{N} \mathbf{B}_i$ and $\mathbb{E}[\mathbf{B}_i] = 0$. Define the bound $D = C_2 \sigma \sqrt{log(N+T)}$ for a constant $C_2$ that we will specify later. For each $(i,t) \in \mathcal{O}$, let $\bar{\varepsilon}_{it} = \varepsilon_{it} \mathbb{1}\{|\varepsilon_{it}| \le D\}$. For $1 \le i \le N$, let $\overline{\mathbf{B}}_i = \sum_{t=1}^{t_i} \bar{\varepsilon}_{it} \mathbf{A}_{it}$.

The $\sigma$-sub-Gaussian implies

$$\begin{aligned} \mathbb{P}(|\varepsilon_{it} \ge t|) &= 2 \frac{1}{\sqrt{2\pi}} \int_t^{\infty} \frac{1}{\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \\ &\le \frac{2\sigma}{\sqrt{2\pi}} \int_{\frac{t^2}{2\sigma^2}}^{\infty} \exp(-x) dx = \frac{2\sigma}{\sqrt{2\pi} t} \exp\left(-\frac{t^2}{2\sigma^2}\right). \end{aligned} \tag{EC.63}$$

Therefore, for $\alpha > 0$,

$$\begin{aligned} \mathbb{P}\left\{\|\mathfrak{E}\|_{op} \ge \alpha\right\} \le & \mathbb{P}\left\{\left\|\sum_{i=1}^{B} \overline{\mathbf{B}}_i\right\|_{op} \ge \alpha\right\} + \sum_{(i,t)\in\mathcal{O}} \mathbb{P}(|\varepsilon_{it}| \ge D) \\ \le & \mathbb{P}\left\{\left\|\sum_{i=1}^{B} \overline{\mathbf{B}}_i\right\|_{op} \ge \alpha\right\} + |\mathcal{O}| \times \frac{2\sigma}{\sqrt{2\pi} D} \exp\left(-\frac{D^2}{2\sigma^2}\right) \\ \le & \mathbb{P}\left\{\left\|\sum_{i=1}^{B} \overline{\mathbf{B}}_i\right\|_{op} \ge \alpha\right\} + \sqrt{\frac{2}{\pi}} \frac{NT}{C_2 \sqrt{\log(N+T)}} (N+T)^{-\frac{C_2^2}{2}}. \end{aligned} \tag{EC.64}$$

For $1 \le i \le N$, define $\mathbf{Z}_i = \overline{\mathbf{B}}_i - \mathbb{E}[\overline{\mathbf{B}}_i]$. Then,

$$\begin{aligned} \left\|\sum_{i=1}^{N} \overline{\mathbf{B}}_i\right\|_{op} &\le \left\|\sum_{i=1}^{N} \mathbf{Z}_i\right\|_{op} + \left\|\mathbb{E}\left[\sum_{i=1}^{N} \overline{\mathbf{B}}_i\right]\right\|_{op} \\ &\le \left\|\sum_{i=1}^{N} \mathbf{Z}_i\right\|_{op} + \left\|\mathbb{E}\left[\sum_{i=1}^{N} \overline{\mathbf{B}}_i\right]\right\|_{F} \le \left\|\sum_{i=1}^{N} \mathbf{Z}_i\right\|_{op} + \sqrt{NT} \left|\mathbb{E}\left[\sum_{i=1}^{N} \overline{\mathbf{B}}_i\right]\right|_{\infty}. \end{aligned} \tag{EC.65}$$

Further,

$$\begin{aligned} |\mathbb{E}[\bar{\varepsilon}_{it}]| &= |\mathbb{E}[\varepsilon_{it} \mathbb{1}\{|\varepsilon_{it}| \le D\}]| = |\mathbb{E}[\varepsilon_{it} \mathbb{1}\{|\varepsilon_{it}| \ge D\}]| \le \sqrt{\mathbb{E}[\varepsilon_{it}^2] \mathbb{P}(|\varepsilon_{it}| \ge D)} \\ &\le \sigma \sqrt{\sqrt{\frac{2}{\pi}} \frac{1}{C_2 \sqrt{\log(N+T)}} (N+T)^{-\frac{C_2^2}{2}}}. \end{aligned} \tag{EC.66}$$

Therefore,

$$\sqrt{NT} \left|\mathbb{E}\left[\sum_{i=1}^{N} \overline{\mathbf{B}}_i\right]\right|_{\infty} \le \sigma \sqrt{\sqrt{\frac{2}{\pi}} \frac{NT}{C_2 \sqrt{\log(N+T)}} (N+T)^{-\frac{C_2^2}{2}}}. \tag{EC.67}$$

Note that $\|\mathbf{Z}_i\|_{op} \le 2D\sqrt{T}$ for all $1 \le i \le N$. The only step left for invoking Proposition EC.6.1 is to calculation $\sigma_Z$ in there.

Recall that $\mathbb{E}[(\bar{\varepsilon}_{it} - \mathbb{E}[\bar{\varepsilon}_{it}])^2] \le \sigma^2$.

We have

$$\begin{aligned}\left\|\sum_{i=1}^{N}\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i^T]\right\|_{op} &\le \max_{1\le i\le N}\left(\mathbb{E}\left(\sum_{t:(i,t)\in\mathcal{O}}\mathbb{E}[(\bar{\varepsilon}_{it}-\mathbb{E}[\bar{\varepsilon}_{it}])^2]\right)\right)\\ &\le \sigma^2 T,\end{aligned} \tag{EC.68}$$

and

$$\begin{aligned}\left\|\sum_{i=1}^{N}\mathbb{E}[\mathbf{Z}_i^T\mathbf{Z}_i]\right\|_{op} &\le \sigma^2 \max_{1\le i\le T}\sum_{j=1}^{N}\mathbb{P}((j,i)\in\mathcal{O})\\ &\le \sigma^2 N.\end{aligned} \tag{EC.69}$$

The first inequality in Inequality (EC.69) is due to $\mathbb{E}\left\{(\bar{\varepsilon}_{it}-\mathbb{E}[\bar{\varepsilon}_{it}])(\bar{\varepsilon}_{js}-\mathbb{E}[\bar{\varepsilon}_{js}])\middle|\mathcal{O}\right\} = 0$ for $(i,t) \ne (j,s)$.

Therefore $\sigma_Z^2 = \sigma^2\max\{N,T\}$ is a possible choice. Invoking Proposition EC.6.1, we have

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{N}\mathbf{Z}_i\right\|_{op} \ge \alpha\right\} \le (N+T)\exp\left[\frac{-\alpha^2}{2\sigma^2\max\{N,T\} + (4C_2\sigma\sqrt{\log(N+T)T}\alpha)/3}\right]. \tag{EC.70}$$

Taking $C_2 = 3$, $\alpha = \max\{4\sigma\sqrt{\max\{N,T\}}\sqrt{\log(N+T)}, 32T^{\frac{1}{2}}(\log(N+T))^{\frac{3}{2}}\sigma\}$.

Combing Inequalities (EC.64), (EC.65), (EC.67), (EC.70), we have with probability at least $1 - \frac{1}{2(N+T)^2} - \frac{1}{2(N+T)^3}$

$$\|\mathfrak{E}\|_{op} \le 4\sigma\max\{\sqrt{\max\{N,T\}}\sqrt{\log N + T}, 8T^{\frac{1}{2}}(\log(N+T))^{\frac{3}{2}}\} + \sigma. \tag{EC.71}$$

**EC.5.1.2. Proof of Lemma EC.4** We define some additional notation here, which are similar to the additional notation in Athey et al. (2021). Given observation set $\mathcal{O}$, for every $N$ by $T$ matrix $\mathbf{M}$, define $\mathcal{X}_{\mathcal{O}}(\mathbf{M})$ and $\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})$ as follows.

$$\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M}) = [\langle\mathbf{A}_{i1},\mathbf{M}\rangle,\cdots,\langle\mathbf{A}_{it},\mathbf{M}\rangle]^T, \tag{EC.72}$$

$$\mathcal{X}_{\mathcal{O}}(\mathbf{M}) = \begin{bmatrix} \mathcal{X}_{\mathcal{O}}^{(1)}(M) \\ . \\ . \\ . \\ \mathcal{X}_{\mathcal{O}}^{(N)}(M) \end{bmatrix}. \tag{EC.73}$$

Define a $L^2_{(\Pi)}$ norm of $\mathbf{M}$ as

$$\|\mathbf{M}\|_{L^2_{(\Pi)}} = \sqrt{\mathbb{E}_\pi\left(\|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|_2^2\right)}, \tag{EC.74}$$

where $\mathbb{E}_\pi$ is taking expectation with respect to the distribution of $\mathcal{O}$.

Define the constraint set as

$$\mathcal{C}(\theta,\eta) = \left\{ \mathbf{M} \in \mathbb{R}^{N\times T} \Big| \|\mathbf{M}\|_\infty \le 1, \|\mathbf{M}\|^2_{L^2_{(\Pi)}} \ge \theta, \|\mathbf{M}\|_* \le \sqrt{\eta}\|\mathbf{M}\|_F + \frac{3\delta}{2\lambda(L_{max}+\delta_1)} \right\}. \tag{EC.75}$$

Then according to Lemma EC.3, we know that either

$$\frac{\tilde{\mathbf{L}} - \mathbf{L}^*}{2(L_{max}+\delta_1)} \in \mathcal{C}(\theta, \left(6\sqrt{2R}\right)^2)$$

or

$$\|\frac{\tilde{\mathbf{L}} - \mathbf{L}^*}{2(L_{max}+\delta_1)}\|^2_{L^2_{(\Pi)}} \le \theta.$$

Observe that $\|\frac{\tilde{\mathbf{L}}-\mathbf{L}^*}{2(L_{max}+\delta_1)}\|^2_{L^2_{(\Pi)}} \le \theta$ implies $\|\tilde{\mathbf{L}} - \mathbf{L}^*\|_F^2 \le \frac{4(L_{max}+\delta_1)^2\theta}{p_c}$.

We set $\theta = 33T\log(N+T)$.

Let $\xi > 1$ be a number that we will specify later. Define

$$\mathcal{C}(\theta,\eta,\rho) = \left\{ \mathbf{M} \in \mathcal{C}(\theta,\eta) \Big| \rho \le \|\mathbf{M}\|^2_{L^2_{(\Pi)}} \le \rho\xi \right\}. \tag{EC.76}$$

We state a lemma that we will prove later in this section.

Lemma EC.6. *Suppose $\xi > 1$. Let*

$$Z_\rho = \frac{1}{T} \sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)} \{\|\mathbf{M}\|^2_{L^2_{(\Pi)}} - \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2\}, \tag{EC.77}$$

*then for* $t>0$,

$$P\left(Z_\rho \geq \frac{48}{T}\left(\sqrt{\frac{\eta\rho\xi}{p_c}}+\frac{3\delta}{2\lambda(L_{max}+\delta_1)}\right)(\sqrt{N}+\sqrt{T})+t\right)\leq \\ \exp\left(-\frac{t}{4}\log\left(1+2\log\left(1+\frac{t}{\frac{96}{T}\left(\sqrt{\frac{\eta\rho\xi}{p_c}}+\frac{3\delta}{2\lambda(L_{max}+\delta_1)}\right)(\sqrt{N}+\sqrt{T})+\frac{\rho\xi}{T}}\right)\right)\right). \tag{EC.78}$$

According to Lemma EC.6, if we set

$$t_0=\frac{1}{4T}\left(96\left(\sqrt{\frac{\eta\rho\xi}{p_c}}+\frac{3\delta}{2\lambda(L_{max}+\delta_1)}\right)(\sqrt{N}+\sqrt{T})+\rho\xi\right), \\ t=\frac{1}{T}\left(\frac{\rho\xi}{4}+\frac{\rho}{4}+\frac{4*144\eta\xi}{p_c}(\sqrt{N}+\sqrt{T})^2+\frac{72\delta}{2\lambda(L_{max}+\delta_1)}(\sqrt{N}+\sqrt{T})\right), \tag{EC.79}$$

then we know that $t_0\leq t$, so we have

$$\mathbb{P}\Big(\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho),\|\mathbf{M}\|^2_{L^2_{(\Pi)}}\geq\|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2+48\|\mathbf{M}\|_{L^2_{(\Pi)}}\sqrt{\frac{\eta\xi}{p_c}}(\sqrt{N}+\sqrt{T}) \\ +\frac{144\delta}{2\lambda(L_{max}+\delta_1)}(\sqrt{N}+\sqrt{T})+ \\ \frac{\|\mathbf{M}\|^2_{L^2_{(\Pi)}}}{4}+\frac{\|\mathbf{M}\|^2_{L^2_{(\Pi)}}\xi}{4}+\frac{576\xi\eta}{p_c}(\sqrt{N}+\sqrt{T})^2+\frac{72\delta}{2\lambda(L_{max}+\delta_1)}(\sqrt{N}+\sqrt{T})\Big) \\ \leq\exp\left(-\frac{1}{22T}\rho(\xi+1)-\frac{10\eta\xi}{p_c}\right). \tag{EC.80}$$

Given that

$$\bigcup_{i=0}^{\infty}\mathcal{C}(\theta,\eta,\theta\xi^i)=\mathcal{C}(\theta,\eta) \tag{EC.81}$$

we have

$$\mathbb{P}\Big(\mathbf{M}\in\mathcal{C}(\theta,\eta),\|\mathbf{M}\|^2_{L^2_{(\Pi)}}\geq\|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2+48\|\mathbf{M}\|_{L^2_{(\Pi)}}\sqrt{\frac{\eta\xi}{p_c}}(\sqrt{N}+\sqrt{T}) \\ +\frac{144\delta}{2\lambda(L_{max}+\delta_1)}(\sqrt{N}+\sqrt{T})+ \\ \frac{\|\mathbf{M}\|^2_{L^2_{(\Pi)}}}{4}+\frac{\|\mathbf{M}\|^2_{L^2_{(\Pi)}}\xi}{4}+\frac{576\xi\eta}{p_c}(\sqrt{N}+\sqrt{T})^2+\frac{72\delta}{2\lambda(L_{max}+\delta_1)}(\sqrt{N}+\sqrt{T})\Big) \\ \leq\exp\left(-\frac{\theta}{11T}-10\eta\xi\right)\frac{1}{1-\exp\left(\frac{-\theta(\xi-1)}{22T}\right)}. \tag{EC.82}$$

Note that $48\|\mathbf{M}\|_{L^2_{(\Pi)}}\sqrt{\frac{\eta\xi}{p_c}}(\sqrt{N}+\sqrt{T}) \le \frac{\|\mathbf{M}\|^2_{L^2_{(\Pi)}}}{4} + 2304\frac{\eta}{p_c}(\sqrt{N}+\sqrt{T})^2$, and $\|\mathbf{M}\|^2_{L^2_{(\Pi)}} \ge p_c\|\mathbf{M}\|^2_F$, if we set $\xi = \frac{4}{3}$, we have

$$\begin{aligned}
&\mathbb{P}\Big(\frac{p_c\|\mathbf{M}\|^2_F}{6} \ge \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2 + 3648\frac{\eta}{p_c}(\sqrt{N}+\sqrt{T})^2 + \frac{216\delta}{2\lambda(L_{max}+\delta_1)}(\sqrt{N}+\sqrt{T})\Big) \\
&\le P\Big(\frac{\|\mathbf{M}\|^2_{L^2_{(\Pi)}}}{6} \ge \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2 + 3648\frac{\eta}{p_c}(\sqrt{N}+\sqrt{T})^2 + \frac{216\delta}{2\lambda(L_{max}+\delta_1)}(\sqrt{N}+\sqrt{T})\Big) \\
&\le \exp\left(-\frac{\theta}{11T}\right)\frac{\exp(-10\eta)}{1-\exp(-\frac{\theta}{66T})}.
\end{aligned} \tag{EC.83}$$

Note that we set $\theta = 33T\log(N+T)$ and we have $\frac{\tilde{\mathbf{L}}-\mathbf{L}^*}{2(L_{max}+\delta_1)} \in \mathcal{C}(\theta,\eta)$ for $\eta = 72R$ according to Lemma EC.2, so we have the Lemma EC.4.

*Proof of Lemma EC.6* The goal here is to invoke theorem 12.9 of Boucheron et al. (2013).

Note that $\|\mathbf{M}\|^2_{L^2_{(\Pi)}} - \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2$ has its rows independent and

$$\mathbb{E}_\pi\left(\mathbb{E}_\pi(\|\mathcal{X}^{(i)}_{\mathcal{O}}(\mathbf{M})\|^2) - \|\mathcal{X}^{(i)}_{\mathcal{O}}(\mathbf{M})\|^2\right) = 0$$

for all $1 \le i \le N$. Although theorem 12.9 in Boucheron et al. (2013) requires countability of the index set, given that $\mathcal{C}(\theta,\eta,\rho)$ is bounded, compact, and $\|\mathbf{M}\|^2_{L^2_{(\Pi)}} - \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2$ is uniformly continuous for all $\mathcal{O}$, theorem 12.9 is applicable to our setting. The next steps are to find a bound for $\mathbb{E}(Z_\rho)$ and

$$\sigma^2 = \frac{1}{T^2}\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\sum_{i=1}^N Var(\|\mathcal{X}^{(i)}_{\mathcal{O}}(\mathbf{M})\|^2).$$

For $\sigma^2$, we have

$$\begin{aligned}
\sigma^2 &\le \frac{1}{T^2}\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\sum_{i=1}^N \mathbb{E}(\|\mathcal{X}^{(i)}_{\mathcal{O}}(\mathbf{M})\|_2^4) \\
&\le \frac{1}{T}\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\sum_{i=1}^N \mathbb{E}(\|\mathcal{X}^{(i)}_{\mathcal{O}}(\mathbf{M})\|_2^2) \le \frac{\rho\xi}{T}.
\end{aligned} \tag{EC.84}$$

For $\mathbb{E}(Z_\rho)$, suppose $\zeta_i$ $(i=1,\cdots,N)$ are i.i.d. Rademacher variable, then we have, for any $\tau$

$$\begin{aligned}
\mathbb{E}(Z_\rho) &\overset{\text{(i)}}{\leq} \frac{1}{T}\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\left\{\left|\|\mathbf{M}\|_{L^2_{(\Pi)}}^2 - \|\mathcal{X}_{\mathcal{O}}(\mathbf{M})\|^2\right|\right\} \\
&\overset{\text{(ii)}}{\leq} \frac{2}{T}\mathbb{E}\left[\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\left|\sum_{i=1}^N \zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2\right|\right] \\
&\overset{\text{(iii)}}{\leq} \frac{4}{T}\mathbb{E}\left[\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\sum_{i=1}^N \zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2\right] \\
&\overset{\text{(iv)}}{\leq} \frac{4}{T}\left(2\tau^2 + 2\log N(\tau,\theta,\eta,\rho) + 2\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\mathbb{E}(\sum_{i=1}^N \zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2)\right) \\
&= \frac{8}{T}\left(\tau^2 + \log N(\tau,\theta,\eta,\rho)\right),
\end{aligned} \tag{EC.85}$$

where Inequality ii is due to lemma 6.3 of Ledoux and Talagrand (1991), Inequality iii is due to

$$\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\sum_{i=1}^N \zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2 \geq 0 \tag{EC.86}$$

and

$$\begin{aligned}
\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\left|\sum_{i=1}^N \zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2\right| &= \sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\left|\sum_{i=1}^N -\zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2\right| \\
\sum_{i=1}^N \zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2 &= -\sum_{i=1}^N -\zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2.
\end{aligned} \tag{EC.87}$$

$N(\tau,\theta,\eta,\rho)$ in Inequality iv is the $\tau$ covering number (Wainwright 2019) of $\mathcal{C}(\theta,\eta,\rho)$, and Inequality iv is due to typical arguments bounding empirical process that we list as follows. Let $\mathfrak{N} = N(\tau,\theta,\eta,\rho)$. Suppose $\mathbf{M}_1,\ldots,\mathbf{M}_{\mathfrak{N}}$ is the $\tau-$ cover. Then we have

$$\begin{aligned}
&\mathbb{E}\left(\sup_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\sum_{i=1}^N \zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M})\|_2^2\right) \\
&\leq \mathbb{E}\left(2\sup_{1\leq j\leq\mathfrak{N}}\sum_{i=1}^N \zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M}_j)\|_2^2 + 2\sup_{1\leq j\leq\mathfrak{N}}\inf_{\mathbf{M}\in\mathcal{C}(\theta,\eta,\rho)}\|\mathbf{M}_j-\mathbf{M}\|_2^2\right) \\
&= 2\log\left(\exp\left(\mathbb{E}\left(\sup_{1\leq j\leq\mathfrak{N}}\sum_{i=1}^N \zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M}_j)\|_2^2\right)\right)\right) + 2\tau^2 \\
&\leq 2\log\left(\sum_{j=1}^{\mathfrak{N}}\exp\left\{\mathbb{E}\left(\sum_{i=1}^N \zeta_i\|\mathcal{X}_{\mathcal{O}}^{(i)}(\mathbf{M}_j)\|_2^2\right)\right\}\right) + 2\tau^2 \\
&= 2\log\mathfrak{N} + 2\tau^2.
\end{aligned} \tag{EC.88}$$

Readers interested in more details on covering number can take Wainwright (2019) as a reference.

Now we proceed with Inequality (EC.85) with bounding $\log N(\tau,\theta,\eta,\rho)$.

Suppose $G$ is a $\mathbb{R}^{N\times T}$ matrix with i.i.d. $N(0,1)$ entries. Let $B_1(R)=\{\Delta\in\mathbb{R}^{N\times T}\big|\|\Delta\|_*\le R\}$. Then $\mathcal{C}(\theta,\eta,\rho)\subset B_1(\sqrt{\frac{\eta\rho\xi}{p_c}}+\frac{3\delta}{2\lambda(L_{max}+\delta_1)})$. Let $\tilde{N}(\tau,R)$ be the $\tau$-covering number of $B_1(R)$. By Sudakov minoration (Theorem 5.20 in Wainwright (2019)), and the fact that packing number is no smaller than covering number, we have

$$\begin{aligned}\sqrt{\log N(\tau,\theta,\eta,\rho)}&\le\sqrt{\tilde{N}(\tau,\sqrt{\frac{\eta\rho\xi}{p_c}}+\frac{3\delta}{2\lambda(L_{max}+\delta_1)})}\\&\le\frac{3}{\tau}\mathbb{E}\left[\sup_{\|\Delta\|_*\le\sqrt{\frac{\eta\rho\xi}{p_c}}+\frac{3\delta}{2\lambda(L_{max}+\delta_1)}}\langle G,\Delta\rangle\right]\\&\le\frac{3(\sqrt{\frac{\eta\rho\xi}{p_c}}+\frac{3\delta}{2\lambda(L_{max}+\delta_1)})}{\tau}\mathbb{E}(\|G\|_{op}).\end{aligned}\tag{EC.89}$$

By (4.2.5) in Tropp (2015), we have

$$\mathbb{E}(\|G\|_{op})\le\sqrt{N}+\sqrt{T}.\tag{EC.90}$$

Therefore, taking $\tau=\sqrt{\frac{3(\sqrt{\frac{\eta\rho\xi}{p_c}}+\frac{3\delta}{2\lambda(L_{max}+\delta_1)})}{\tau}(\sqrt{N}+\sqrt{T})}$, we have

$$\mathbb{E}(Z_\rho)\le\frac{48}{T}\left(\sqrt{\frac{\eta\rho\xi}{p_c}}+\frac{3\delta}{2\lambda(L_{max}+\delta_1)}\right)(\sqrt{N}+\sqrt{T}).\tag{EC.91}$$

Now invoking theorem 12.9 of Boucheron et al. (2013) with Inequalities (EC.91) and (EC.84), we have, for $t>0$,

$$\begin{aligned}&P\left(Z_\rho\ge\frac{48}{T}\left(\sqrt{\frac{\eta\rho\xi}{p_c}}+\frac{3\delta}{2\lambda(L_{max}+\delta_1)}\right)(\sqrt{N}+\sqrt{T})+t\right)\le\\&\exp\left(-\frac{t}{4}\log\left(1+2\log\left(1+\frac{t}{\frac{96}{T}\left(\sqrt{\frac{\eta\rho\xi}{p_c}}+\frac{3\delta}{2\lambda(L_{max}+\delta_1)}\right)(\sqrt{N}+\sqrt{T})+\frac{\rho\xi}{T}}\right)\right)\right).\end{aligned}\tag{EC.92}$$

## EC.6. Proof of Theorem 1

Write the $F$ in Equation (10) in the following form

$$F(X)=f(X)+g(X)+\mathfrak{T}\{X\in C_1\cap C_2\cap\cdots\cap C_J\}.\tag{EC.93}$$

For ease of notation, denote $\mathcal{C}=C_1\cap C_2\cap\cdots\cap C_J$.

Recalling that

$$\begin{aligned} X_{k+0.5} &= X_k - \eta \nabla f(X_k) \\ X_{k+1} &= \widetilde{\text{Prox}}_{\eta(g(X)+\mathfrak{T}\{\mathcal{C}\})}(X_{k+0.5}), \end{aligned} \tag{EC.94}$$

we denote

$$\begin{aligned} G(X_k) &= \frac{X_k - \text{Prox}_{\eta(g(X)+\mathfrak{T}\{\mathcal{C}\})}(X_{k+0.5})}{\eta} \\ \tilde{G}(X_k) &= \frac{X_k - \widetilde{\text{Prox}}_{\eta(g(X)+\mathfrak{T}\{\mathcal{C}\})}(X_{k+0.5})}{\eta}. \end{aligned} \tag{EC.95}$$

Then it's clear that

$$\begin{aligned} X_{k+1} &= X_k - \eta \tilde{G}(X_k) \\ \text{Prox}_{\eta(g(X)+\mathfrak{T}\{\mathcal{C}\})}(X_{k+0.5}) &= X_k - \eta G(X_k). \end{aligned} \tag{EC.96}$$

Recalling the definition of $\text{Prox}_{\eta(g(X)+\mathfrak{T}\{\mathcal{C}\})}(X_{k+0.5})$,

$$\text{Prox}_{\eta(g(X)+\mathfrak{T}\{\mathcal{C}\})}(X_{k+0.5}) = \arg\min_X \left\{ \frac{1}{2\eta}\|X - X_{k+0.5}\|^2 + g(X) + \mathfrak{T}\{X \in \mathcal{C}\} \right\}, \tag{EC.97}$$

we know that

$$\mathbf{0} \in X - X_{k+0.5} + \eta \partial g(X) + \eta \partial \mathfrak{T}\{X \in \mathcal{C}\}\Big|_{X = X_k - \eta G(X_k)}. \tag{EC.98}$$

In the later part of this proof, we choose $\partial g(X_k - \eta G(X_k))$ and $\partial \mathfrak{T}\{X_k - \eta G(X_k) \in \mathcal{C}\}$ such that

$$\partial g(X_k - \eta G(X_k)) + \partial \mathfrak{T}\{X_k - \eta G(X_k) \in \mathcal{C}\} + \nabla f(X_k) - G(X_k) = \mathbf{0}.$$

We have

$$\begin{aligned} & f(X_k - \eta \tilde{G}(X_k)) + g(X_k - \eta \tilde{G}(X_k)) \\ & \le f(X_k - \eta G(X_k)) + \langle \nabla f(X_k - \eta G(X_k)), (X_k - \eta \tilde{G}(X_k)) - (X_k - \eta G(X_k)) \rangle + \\ & \quad \frac{L}{2}\|\eta \tilde{G}(X_k) - \eta G(X_k)\|^2 + g(X_k - \eta G(X_k)) + \langle \partial g(X_k - \eta \tilde{G}(X_k)), \eta G(X_k) - \eta \tilde{G}(X_k) \rangle \\ & \le f(X_k - \eta G(X_k)) + g(X_k - \eta G(X_k)) + L_f \delta_0 + L_g \delta_0 + \frac{L}{2}\delta_0^2. \end{aligned} \tag{EC.99}$$

To further bound the first two terms in the right hand side, we have for any $y \in \mathbb{R}^{n\times m}$,

$$\begin{aligned}
&f(X_k - \eta G(X_k)) + g(X_k - \eta G(X_k)) + \mathfrak{T}\{X_k - \eta G(X_k) \in \mathcal{C}\} \\
&\le f(X_k) + \langle \nabla f(X_k), -\eta G(X_k)\rangle + \frac{L}{2}\|\eta G(X_k)\|^2 + \\
&\quad g(y) + \langle \partial g(X_k - \eta G(X_k)), X_k - \eta G(X_k) - y\rangle \\
&\quad + \mathfrak{T}\{y \in \mathcal{C}\} + \langle \partial I X_k - \eta G(X_k), X_k - \eta G(X_k) - y\rangle \\
&\le f(y) + \langle \nabla f(X_k), X_k - y - \eta G(X_k)\rangle + \frac{L}{2}\|\eta G(X_k)\|^2 + g(y) + \mathfrak{T}\{y \in \mathcal{C}\} + \\
&\quad \langle \partial g(X_k - \eta G(X_k)) + \partial \mathfrak{T}\{X_k - \eta G(X_k) \in \mathcal{C}\}, X_k - \eta G(X_k) - y\rangle \\
&= f(y) + \langle G(X_k), X_k - y - \eta G(X_k)\rangle + \frac{L}{2}\|\eta G(X_k)\|^2 + g(y) + I(y),
\end{aligned} \tag{EC.100}$$

where the last equality is due to (EC.98).

If we further let $y = X^*$, we have

$$\begin{aligned}
&f(X_k - \eta G(X_k)) + g(X_k - \eta G(X_k)) + \mathfrak{T}\{X_k - \eta G(X_k) \in \mathcal{C}\} \\
&\le f(X^*) + g(X^*) + \mathfrak{T}\{X^* \in \mathcal{C}\} + \langle G(X_k), X_k - X^* - \frac{\eta G(X_k)}{2}\rangle \\
&\quad + \left(\frac{L}{2}\eta^2 - \frac{\eta}{2}\right)\|G(X_k)\|^2 \\
&= f(X^*) + g(X^*) + \mathfrak{T}\{X^* \in \mathcal{C}\} + \frac{1}{2\eta}\left(\|X_k - X^*|^2 - \|X_k - \eta G(X_k) - X^*\|^2\right) \\
&\quad + \frac{\eta}{2}(L\eta - 1)\|G(X_k)\|^2 \\
&\le f(X^*) + g(X^*) + \mathfrak{T}\{X^* \in \mathcal{C}\} + \frac{1}{2\eta}\left(\|X_k - X^*|^2 - \|X_k - \eta \tilde{G}(X_k) - X^*\|^2\right) \\
&\quad + \frac{\delta_0^2}{2\eta} + \frac{\delta_0 D}{\eta} + \frac{\eta}{2}(L\eta - 1)\|G(X_k)\|^2,
\end{aligned} \tag{EC.101}$$

where $D$ is the diameter of $\mathcal{C}$, and the last Inequality is due to

$$\begin{aligned}
&\|X_k - \eta\tilde{G}(X_k) - X^*\|^2 - \|X_k - \eta G(X_k) - X^*\|^2 \\
&= \|X_k - \eta\tilde{G}(X_k) - X^* - (X_k - \eta G(X_k) - X^*)\|^2 + \\
&\quad 2\langle\left(X_k - \eta\tilde{G}(X_k)\right) - \left(X_k - \eta G(X_k)\right), X_k - \eta G(X_k) - X^*\rangle \\
&\le \delta_0^2 + 2\delta_0 D.
\end{aligned} \tag{EC.102}$$

If we further let $\eta \leq \frac{1}{L}$ in Inequality (EC.101), combing with Inequality (EC.99), and noting that $X_k - \eta G(X_k), X^* \in \mathcal{C}$, we have

$$\begin{aligned} f(X_{k+1}) + g(X_{k+1}) \leq & f(X^*) + g(X^*) + \frac{1}{2\eta}\left(\|X_k - X^*\|^2 - \|X_{k+1} - X^*\|^2\right) \\ & + \frac{\delta_0^2}{2\eta} + \frac{\delta_0 D}{\eta} + \frac{L}{2}\delta_0^2 + (L_f + L_g)\delta_0. \end{aligned} \tag{EC.103}$$

Adding up $k = 0 \cdots K-1$ for Inequality (EC.103), we have

$$\frac{1}{K}\sum_{j=1}^{K}(f(X_j) + g(X_j)) \leq f(X^*) + g(X^*) + \frac{1}{2\eta}\|X_0 - X^*\|^2 + \frac{\delta_0^2}{2\eta} + \frac{\delta_0 D}{\eta} + \frac{L}{2}\delta_0^2 + (L_f + L_g)\delta_0. \tag{EC.104}$$

This proves the theorem. But now, we also give a variant of the theorem. Suppose $\bar{X}^K = \frac{1}{K}\sum_{j=1}^{K} X_j$, then the convexity of $f$ and $g$ implies that the left hand side of Inequality (EC.104) is larger equal to $f(\bar{X}^K) + g(\bar{X}^K)$.

## EC.7. Proof of Proposition 2.1

Define the following averages:

$$\overline{W}^t = \frac{1}{t-1}\sum_{i=1}^{t} W^i, \overline{Z}^t = \frac{1}{t-1}\sum_{i=1}^{t} Z^i, \overline{P}^t = \frac{1}{t-1}\sum_{i=1}^{t} P^i. \tag{EC.105}$$

Writing the constraints of optimization problem (15) in matrix form, we have

$$\begin{pmatrix} \mathbf{0} & -\boldsymbol{I}_{nm} & \boldsymbol{I}_{nm} \\ -\boldsymbol{I}_{nm} & \mathbf{0} & \boldsymbol{I}_{nm} \end{pmatrix} \begin{pmatrix} vec(W) \\ vec(Z) \\ vec(P) \end{pmatrix} = \mathbf{0}. \tag{EC.106}$$

Note that the coefficient matrix blocks corresponding to $vec(Z)$ and $vec(P)$ in the linear constraint (EC.106) are full column rank matrices. It suffices the conditions of Theorem 4.1 in Cai et al. (2017). Applying Inequality (4.3) in Cai et al. (2017) to our setting with $\theta_1(x) = h_1(x)$, $\theta_2(x) = h_2(x)$, $\theta_3(x) = \|x - P_0\|^2$, $x_1' = W^*$, $x_2' = Z^*$, $x_3' = P^*$, we have, for $\beta \leq \frac{6}{17}$,

$$
\begin{aligned}
&2\beta t\Bigg\{\left[h_1(\overline{W}^t)+h_2(\overline{Z}^t)+\|\overline{P}^t-P_0\|^2+\langle\Lambda_1^*,(\overline{W}^t-\overline{P}^t)\rangle+\langle\Lambda_2^*,(\overline{Z}^t-\overline{P}^t)\rangle\right]\\
&\quad-\left[h_1(W^*)+h_2(Z^*)+\|P^*-P_0\|^2+\langle\Lambda_1^*,(W^*-P^*)\rangle+\langle\Lambda_2^*,(Z^*-P^*)\rangle\right]\Bigg\} \qquad \text{(EC.107)}\\
&\le\beta^2\|Z^1-Z^*\|^2+2\beta^2\|P^1-P^*\|^2+\|\boldsymbol{\Lambda}^1-\boldsymbol{\Lambda}^*\|^2+\frac{10}{3}\beta^2*2\|P^1-P^0\|^2.
\end{aligned}
$$

For the left hand side, we define a function

$$
U(W,Z,P)=h_1(W)+h_2(Z)+\|P-P_0\|_F^2+\langle\Lambda_1^*,W\rangle+\langle\Lambda_2^*,Z\rangle-\langle(\Lambda_1^*+\Lambda_2^*),P\rangle. \qquad \text{(EC.108)}
$$

Given that $(W^*,Z^*,P^*),(\Lambda_1^*,\Lambda_2^*)$ is a solution to

$$
\max_{\Lambda_1,\Lambda_2}\min_{W,Z,P} h_1(W)+h_2(Z)+\|P-P_0\|_F^2+\langle\Lambda_1,W\rangle+\langle\Lambda_2,Z\rangle-\langle(\Lambda_1+\Lambda_2),P\rangle,
$$

we have

$$
0=\frac{\partial U(W,Z,P)}{\partial P}\bigg|_{W=W^*,Z=Z^*,P=P^*}=2(P^*-P_0)-(\Lambda_1^*+\Lambda_2^*). \qquad \text{(EC.109)}
$$

Further, since $U(W,Z,P)$ is separable with respect to $W$, $Z$, $P$, we have

$$
\begin{aligned}
&U(W,Z,P)-U(W^*,Z^*,P^*)\\
&\ge U(W^*,Z^*,P)-U(W^*,Z^*,P^*)\\
&=\|P-P_0\|^2-\|P^*-P_0\|^2-(\Lambda_1^{*T}+\Lambda_2^{*T})(P-P^*) \qquad \text{(EC.110)}\\
&=\|P-P^*\|^2+\langle P-P^*,2(P^*-P_0)-\Lambda_1^*-\Lambda_2^*\rangle\\
&=\|P-P^*\|^2.
\end{aligned}
$$

Combining Equation (EC.109) and (EC.110), we have

$$
\|\overline{P}^t-P^*\|^2\le\frac{1}{2\beta t}\left(\beta^2\|Z^1-Z^*\|^2+2\beta^2\|P^1-P^*\|^2+\|\boldsymbol{\Lambda}^1-\boldsymbol{\Lambda}^*\|^2+\frac{20}{3}\beta^2\|P^1-P^0\|^2\right). \qquad \text{(EC.111)}
$$

## EC.8. Proof of Lemma 1

We begin with bounding $C(C_1,C_2)$.

$$
C(C_1,C_2)=\frac{1}{2\cos^2(\frac{\theta(C_1,C_2)}{2})}=\frac{1}{\cos(\theta(C_1,C_2))+1}. \qquad \text{(EC.112)}
$$

Observe that $B_d(x) \subset C_1 \cap C_2$, we have

$$\begin{aligned}
\cos(\theta(C_1, C_2)) &= \inf_{P \in \partial(C_1 \cap C_2)} \cos\Big( \sup_{\boldsymbol{\lambda_1} \in N_{C_1}(P), \boldsymbol{\lambda_2} \in N_{C_2}(P)} \arccos\left(\langle \boldsymbol{\lambda_1}, \boldsymbol{\lambda_2} \rangle\right)\Big) \\
&\geq \inf_{P \in \partial(C_1 \cap C_2)} \cos\Big( \sup_{\boldsymbol{\lambda_1} \in N_{B_d(x)}(P), \boldsymbol{\lambda_2} \in N_{B_d(x)}(P)} \arccos\left(\langle \boldsymbol{\lambda_1}, \boldsymbol{\lambda_2} \rangle\right)\Big) \\
&= \inf_{P \in \partial(C_1 \cap C_2)} -(2\frac{\|P - x\|^2 - d^2}{\|P - x\|^2} - 1) \\
&\geq -1 + \frac{2d^2}{\tilde{D}^2},
\end{aligned} \tag{EC.113}$$

where $\tilde{D} = \sup_{P \in \partial(C_1 \cap C_2)} \|P - x\|_F$.

Therefore,

$$C(C_1, C_2) \leq \frac{\tilde{D}^2}{2d^2} \leq \frac{D^2}{2d^2}, \tag{EC.114}$$

where $D = \sup_{P_1, P_2 \in \partial(C_1 \cap C_2)} \|P_1 - P_2\|_F$.

Now we continue with bounding dual variable $\boldsymbol{\Lambda}^*$ in the case that $h_1(X) = \mathfrak{T}\{X \in C_1\}, h_2(X) = \mathfrak{T}\{X \in C_2\}$.

From Equation (EC.109), we know that

$$\begin{aligned}
4\|P^* - P_0\|^2 &= \|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2 + 2\langle \Lambda_1^*, \Lambda_2^* \rangle \\
&\geq \|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2 + 2\cos(\theta(C_1, C_2))\|\Lambda_1^*\|\|\Lambda_2^*\| \\
&\geq \|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2 + \min\{0, 2\cos(\theta(C_1, C_2))\}\frac{\|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2}{2} \\
&\geq \min\{1, \frac{1}{C(C_1, C_2)}\}(\|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2).
\end{aligned} \tag{EC.115}$$

Therefore, we have

$$\|\boldsymbol{\Lambda}^*\|_F^2 \leq \max\{4, 4C(C_1, C_2)\}\|P^* - P_0\|^2. \tag{EC.116}$$

## EC.9. Proof of Proposition 3.1

To apply Proposition 1 to 1 bit completion matrix problem, we only need to find the $L, L_f, L_g, D$ and a bound for $\|X_0 - X^*\|$ in Proposition 1 in 1 bit matrix completion setting and bound.

Since $g = 0$ in this case, we have $L_g = 0$. Since $C_1 = [-\alpha, \alpha]^{d_1 \times d_2}$, we have $D \leq 2\alpha\sqrt{d_1 d_2}$.

Easy calculation also shows $\sup_{|x| \leq \alpha + \delta_0} \frac{|l'(x)|}{l(x)(1-l(x))}$ is the Lipschitz constant for the smooth objective function $-\mathcal{L}_{\Omega,Y}(X)$.

Easy calculation also show that

$$\sup_{|x|\leq\alpha+\delta_0}\{\frac{|l''(x)l(x)-(l'(x))^2|}{l(x)^2},\frac{|l''(x)(1-l(x))+(l'(x))^2|}{(1-l(x))^2}\}$$

is the smoothness parameter for the smooth objective function $-\mathcal{L}_{\Omega,Y}(X)$.

Also, given that $X_0=\mathbf{0}$ and $X^*\in[-\alpha,\alpha]^{d_1\times d_2}$, we have $\|X_0-X^*\|^2\leq\alpha^2 d_1 d_2$.

With the step size set to be the inverse of smoothness parameter, we completes the proof of the Proposition.

## EC.10. Proof of Proposition 3.2

Note that when $X\in\mathbb{R}^{d_1\times d_2}$ satisfies $\|X\|_F\leq\alpha$, we have $\|X\|_*\leq\sqrt{rank(X)}\|X\|_F\leq\sqrt{\min\{d_1,d_2\}}\alpha\leq\alpha\sqrt{rd_1d_2}$, and $\|X\|_\infty\leq\alpha$. Therefore, we have $d\geq\alpha$.

Note that when $X\in[-\alpha,\alpha]^{d_1\times d_2}$, we have $\|X\|_F\leq\alpha\sqrt{d_1d_2}$. Therefore, $\tilde{D}\leq\alpha\sqrt{d_1d_2}$, where $\tilde{D}$ is defined after Inequality (EC.113).

According to the proof of Lemma 1, when we take $x$ in the $B_d(x)$ there to be $\mathbf{0}$, we have $C(C_1,C_2)\leq\frac{d_1d_2}{2}$.

We continue with bounding the terms in right hand side of Inequality (20) in Proposition 2.1.

Recall the steps we take in Algorithm 3.2, then we have

$$
\begin{aligned}
&\|Z^1 - Z^*\| = \|\mathrm{Proj}_{C_2}(P_0) - P^*)\| \le \|P_0 - P^*\|,\\
&\|P^1 - P^*\| = \|\frac{\beta}{2(\beta+1)}(\mathrm{Proj}_{C_1}(P_0) - P^* + \mathrm{Proj}_{C_2}(P_0) - P^*)\\
&\qquad\qquad + \frac{1}{\beta+1}(P_0 - P^*)\| \le \|P_0 - P^*\|,\\
&\|\mathbf{\Lambda}^1 - \mathbf{\Lambda}^*\|^2 \le 2\|\mathbf{\Lambda}^1\|^2 + 2\|\mathbf{\Lambda}^*\|^2\\
&\qquad\qquad \le 2\beta^2\left\|\frac{1}{\beta+1}\left(P_0 + \frac{\beta}{2}\mathrm{Proj}_{C_2}(P_0) - (1+\frac{\beta}{2})\mathrm{Proj}_{C_1}(P_0)\right)\right\|^2\\
&\qquad\qquad + 2\beta^2\left\|\frac{1}{\beta+1}\left(P_0 + \frac{\beta}{2}\mathrm{Proj}_{C_1}(P_0) - (1+\frac{\beta}{2})\mathrm{Proj}_{C_2}(P_0)\right)\right\|^2\\
&\qquad\qquad + \max\{4, 8C(C_1,C_2)\}\|P_0 - P^*\|^2\\
&\qquad\qquad \le 4\beta^2\|P_0 - P^*\|^2 + \max\{4, 8C(C_1,C_2)\}\|P_0 - P^*\|^2,\\
&\|P^1 - P_0\| \le \frac{\beta}{2(\beta+1)}\|\mathrm{Proj}_{C_1}(P_0) - P_0 + \mathrm{Proj}_{C_2}(P_0) - P_0\| \le \frac{\beta}{\beta+1}\|P_0 - P^*\|.
\end{aligned}
\tag{EC.117}
$$

Some of the inequalities in Inequality (EC.117) are due to $\|P_0 - \mathrm{Proj}_{C_i}(P_0)\| \le \|P_0 - P^*\|$, $\|\mathrm{Proj}_{C_1}(P_0) - \mathrm{Proj}_{C_2}(P_0)\| \le \sum_{i=1}^2 \|P_0 - \mathrm{Proj}_{C_i}(P_0)\|$.

Plugging Inequality EC.117 back to Proposition 2.1, we have

$$
\begin{aligned}
\|\overline{P}^t - P^*\|^2 &\le \frac{1}{2\beta t}(7\beta^2 + \max\{4, 8C(C_1,C_2)\} + \frac{20}{3}\frac{\beta^4}{(\beta+1)^2})\|P_0 - P^*\|^2\\
&\le \frac{1}{2\beta t}(7\beta^2 + 4d_1d_2 + \frac{20}{3}\frac{\beta^4}{(\beta+1)^2})\|P_0 - P^*\|^2.
\end{aligned}
\tag{EC.118}
$$

## EC.11. Proof of Theorem 3

First, we will show that for $t \ge t_0$, $\delta_0 \le \min\{u_0, 1\}$.

We prove this by mathematical induction. For $X_0$, $X_0 \in C_1 \cap C_2$, therefore $\delta_0 \le u_0$ holds for k=0. One thing to note is that $L_{\alpha+u_0} \le 2L_\alpha, \tilde{L}_{\alpha+u_0} \le 2\tilde{L}_\alpha$. Also, recall that $\eta = \frac{1}{2\tilde{L}_\alpha}$. Suppose $delta_0 \le \min\{u_0, 1\}$ holds for $k \le H$, then for $k = H+1$, we have

$$
\begin{aligned}
\|X_k - \eta\nabla f(X_k) - \mathrm{Prox}_{C_1\cap C_2}(X_k)\| &\le \|X_k - \mathrm{Prox}_{C_1\cap C_2}(X_k)\| + |\eta\nabla f(X_k)|\\
&\le u_0 + \frac{1}{2\tilde{L}_\alpha}2L_\alpha.
\end{aligned}
\tag{EC.119}
$$

Therefore,

$$\|X_k - \eta\nabla f(X_k) - \mathrm{Prox}_{C_1\cap C_2}(X_k - \eta\nabla f(X_k))\| \le u_0 + \frac{L_\alpha}{\tilde{L}_\alpha}. \tag{EC.120}$$

According to Proposition 3.2, for

$$t \ge \frac{1}{2\beta}\left(7\beta^2 + 4d_1d_2 + \frac{20}{3}\frac{\beta^4}{(\beta+1)^2}\right)(1 + \frac{L_\alpha}{u_0\tilde{L}_\alpha} + \frac{L_\alpha}{\tilde{L}_\alpha})^2,$$

we have

$$\|X_{k+1} - \mathrm{Prox}_{C_1\cap C_2}(X_{k+0.5})\|^2 \le \min\{u_0^2, 1\}. \tag{EC.121}$$

So $\delta_0 \le \{u_0, 1\}$ also holds for $k = H+1$.

Therefore, $\delta_0 \le \{u_0, 1\}$ for all k. So the Lipschitz constant $L_f \le 2L_\alpha$, and the smooth parameter $L \le 2\tilde{L}_\alpha$ for the objective function on $u_0$ neighbor of $C_1 \cap C_2$.

Further, we have,

$$\delta_0 \le \sqrt{\frac{1}{t}}\sqrt{\frac{1}{2\beta}\left(7\beta^2 + 4d_1d_2 + \frac{20}{3}\frac{\beta^4}{(\beta+1)^2}\right)}. \tag{EC.122}$$

According to Proposition 3.1, we have

$$\begin{aligned}\delta \le{}& \frac{\alpha^2\tilde{L}_\alpha d_1 d_2}{T} + 4\alpha\tilde{L}_\alpha\sqrt{d_1d_2}\sqrt{\frac{1}{t}}\sqrt{\frac{1}{2\beta}\left(7\beta^2 + 4d_1d_2 + \frac{20}{3}\frac{\beta^4}{(\beta+1)^2}\right)} \\ &+ 2L_\alpha\sqrt{\frac{1}{t}}\sqrt{\frac{1}{2\beta}\left(7\beta^2 + 4d_1d_2 + \frac{20}{3}\frac{\beta^4}{(\beta+1)^2}\right)} + 2\tilde{L}_\alpha\frac{1}{t}\frac{1}{2\beta}\left(7\beta^2 + 4d_1d_2 + \frac{20}{3}\frac{\beta^4}{(\beta+1)^2}\right).\end{aligned} \tag{EC.123}$$

## EC.12. Proof of Proposition 4.1

To apply Proposition 1 to causal inference for panel data, we only need to find the $L, L_f, L_g, D$ and a bound for $\|X_0 - X^*\|$ in Proposition 1 in causal inference for panel data.

Since $C_1 = [-L_{\max}, L_{\max}]^{N\times T}$, we have $D = 2L_{\max}\sqrt{NT}$.

Since $g(\mathbf{L}) = \frac{\lambda|\mathcal{O}|}{2}|\mathbf{L}|$, we have $\|\partial g\| \le \frac{\lambda|\mathcal{O}|}{2}\sqrt{\min\{N,T\}}$.

Since $f(\mathbf{L}) = \frac{1}{2}\|\mathbf{P}_{\mathcal{O}}(Y_{\mathbf{L}})\|_F^2$, we have the smooth parameter $L \le 1$, the Lipschitz constant $L_f \le \max_{L\in C_1}\|Y - L\|_F$.

Also, we have $\|\mathbf{L}_0 - \hat{\mathbf{L}}\| \le L_{\max}\sqrt{NT}$. Recall that $\eta = 1$.

Plugging in the quantities into Proposition 1, we have

$$\begin{aligned}\min_{0\le k\le K}\frac{1}{2}\|\mathbf{P}_{\mathcal{O}}(Y-\mathbf{L}_k)\|_F^2+\frac{\lambda|\mathcal{O}|}{2}\|\mathbf{L}_k\|_* \le \frac{1}{2}\|\mathbf{P}_{\mathcal{O}}(Y-\hat{\mathbf{L}})\|_F^2+\frac{\lambda|\mathcal{O}|}{2}\|\hat{\mathbf{L}}\|_* \\ +\frac{1}{2K}\|\mathbf{L}_0-\hat{\mathbf{L}}\|^2+\left(\frac{\lambda|\mathcal{O}|}{2}\sqrt{\min\{N,T\}}+\max_{L\in C_1}\|\mathbf{P}_{\mathcal{O}}(Y-L)\|_F\right)\delta_0 \\ +\delta_0^2+2L_{\max}\sqrt{NT}\delta_0.\end{aligned} \tag{EC.124}$$

## EC.13. Proof of Proposition 4.2

We continue with bounding the terms in right hand side of Inequality (20) in Proposition 2.1. Recall the steps we take in Algorithm 43, we have

$$\begin{aligned}\|Z^1-Z^*\| &= \|thresh(P_0,\frac{\lambda|\mathcal{O}|}{\beta})-P^*\| \le \|P_0-P^*\|+\frac{\lambda|\mathcal{O}|}{\beta}\sqrt{\min\{N,T\}},\\ \|P^1-P^*\| &\le \|P_0-P^*\|+\|P^1-P_0\|,\\ \|\mathbf{\Lambda}^1-\mathbf{\Lambda}^*\|^2 &\le 2\left(\|\Lambda_1^1\|^2+\|\Lambda_2^1\|^2+\|\Lambda_1^*\|^2+\|\Lambda_2^*\|^2\right)\\ &\le 2(\frac{\beta}{1+\beta})^2\left(\left\|P_0-\mathrm{Proj}_{C_1}(P_0)+\frac{\beta}{2}\left(thresh(P_0,\frac{\lambda|\mathcal{O}|}{\beta})-\mathrm{Proj}_{C_1}(P_0)\right)\right\|^2\right.\\ &\quad\left.+\left\|P_0-thresh(P_0,\frac{\lambda|\mathcal{O}|}{\beta})-\frac{\beta}{2}\left(thresh(P_0,\frac{\lambda|\mathcal{O}|}{\beta})-\mathrm{Proj}_{C_1}(P_0)\right)\right\|^2\right)\\ &\quad+2\|\Lambda_1^*\|+2\|\Lambda_2^*\|,\\ \|P^1-P_0\| &= \left\|\frac{\beta}{2(\beta+1)}\left(\mathrm{Proj}_{C_1}(P_0)-P_0+thresh(P_0,\frac{\lambda|\mathcal{O}|}{\beta})-P_0\right)\right\|\\ &\le \frac{\beta}{2(\beta+1)}\|\mathrm{Proj}_{C_1}(P_0)-P_0\|+\frac{\beta}{2(\beta+1)}\frac{\lambda|\mathcal{O}|}{\beta}\sqrt{\min\{N,T\}}.\end{aligned} \tag{EC.125}$$

We continue with bounding the two terms in the right hand side for $\|\mathbf{\Lambda}^1-\mathbf{\Lambda}^*\|^2$. We start with the first term.

$$\begin{aligned}
&\left\|P_0 - \mathrm{Proj}_{C_1}(P_0) + \frac{\beta}{2}\left(thresh(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - \mathrm{Proj}_{C_1}(P_0)\right)\right\|^2 \\
&+ \left\|P_0 - thresh(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - \frac{\beta}{2}\left(thresh(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - \mathrm{Proj}_{C_1}(P_0)\right)\right\|^2 \\
&= \|P_0 - \mathrm{Proj}_{C_1}(P_0)\|^2 + \|P_0 - thresh(P_0, \frac{\lambda|\mathcal{O}|}{\beta})\|^2 \\
&\quad + (\beta + \frac{\beta^2}{2})\|thresh(P_0, \frac{\lambda|\mathcal{O}|}{\beta}) - \mathrm{Proj}_{C_1}(P_0)\|^2 \\
&\leq (1+\beta)^2\left(\|P_0 - \mathrm{Proj}_{C_1}(P_0)\|^2 + \|P_0 - thresh(P_0, \frac{\lambda|\mathcal{O}|}{\beta})\|^2\right) \\
&\leq (1+\beta)^2\left(\|P_0 - \mathrm{Proj}_{C_1}(P_0)\|^2 + \min\{N,T\}(\frac{\lambda|\mathcal{O}|}{\beta})^2\right).
\end{aligned} \tag{EC.126}$$

We proceed with bounding $\|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2$.

According to Equation (EC.109), we have

$$\begin{aligned}
\|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2 &= \|2(P^* - P_0) - \Lambda_2^*\|^2 + \|\Lambda_2^*\|^2 \\
&\leq 8\|P^* - P_0\|^2 + 3\|\Lambda_2^*\|^2.
\end{aligned} \tag{EC.127}$$

Taking derivative with respect to $Z$ for function $U(W,Z,P)$ at point $(W^*, Z^*, P^*)$, we have

$$\mathbf{0} = \frac{\partial U(W,Z,P)}{\partial Z}\bigg|_{W=W^*, Z=Z^*, P=P^*} = \partial h_2(Z^*) + \Lambda_2^*. \tag{EC.128}$$

Observe that $\partial h_2(Z^*) \leq \lambda|\mathcal{O}|\sqrt{\min\{N,T\}}$, continuing with Inequality (EC.127), we have

$$\|\Lambda_1^*\|^2 + \|\Lambda_2^*\|^2 \leq 8\|P^* - P_0\|^2 + 3(\lambda|\mathcal{O}|)^2\min\{N,T\}. \tag{EC.129}$$

Putting together Inequalities (EC.125), (EC.126), (EC.129), together with Proposition 2.1, we have

$$\begin{aligned}
&\|\overline{P}^k - P^*\|^2 \\
&\le \frac{1}{2\beta k}\Bigg(2\beta^2\|P_0 - P^*\|^2 + 2(\lambda|\mathcal{O}|)^2\min\{N,T\} + 4\beta^2\|P_0 - P^*\|^2 + \\
&\quad 4\beta^2\|P^1 - P_0\|^2 + 2\beta^2\|P_0 - \mathrm{Proj}_{C_1}(P_0)\|^2 + 2\min\{N,T\}\left(\lambda|\mathcal{O}|\right)^2 \\
&\quad + 16\|P^* - P_0\|^2 + 6(\lambda|\mathcal{O}|)^2\min\{N,T\} + \frac{20}{3}\beta^2\|P^1 - P_0\|^2\Bigg) \\
&\le \frac{1}{2\beta k}\Bigg((6\beta^2+16)\|P_0 - P^*\|^2 + \left(10 + (2+\frac{10}{3})(\frac{\beta}{1+\beta})^2\right)(\lambda|\mathcal{O}|)^2\min\{N,T\} \\
&\quad + \left(2\beta^2 + (2+\frac{10}{3})\left(\frac{\beta^2}{1+\beta}\right)^2\right)\|P_0 - \mathrm{Proj}_{C_1}(P_0)\|^2\Bigg) \\
&= \frac{1}{\beta k}\Bigg((3\beta^2+8)\|P_0 - P^*\|^2 + \left(5 + \frac{8}{3}(\frac{\beta}{1+\beta})^2\right)(\lambda|\mathcal{O}|)^2\min\{N,T\} \\
&+ \left(\beta^2 + \frac{8}{3}(\frac{\beta^2}{1+\beta})^2\right)\|P_0 - \mathrm{Proj}_{C_1}(P_0)\|^2\Bigg).
\end{aligned} \tag{EC.130}$$

## EC.14. Proof of Theorem 6

Suppose $\inf_{\mathbf{L}\in C_1}\|\mathbf{L}_j - \mathbf{L}\| \le \delta_0$ for $j \le k$, where $k \ge 0$.

Recall that

$$\mathbf{L}_{k+0.5} = \mathbf{L}_k + \mathbf{P}_{\mathcal{O}}(Y - \mathbf{L}_k), \tag{EC.131}$$

we have

$$\|\mathrm{Proj}_{C_1}(\mathbf{L}_{k+0.5}) - \mathbf{L}_{k+0.5}\|^2 \le \|\mathrm{Proj}_{C_1}(\mathbf{L}_k) - \mathbf{L}_{k+0.5}\|^2 \le (C(Y) + \delta_0)^2 \le 2C(Y)^2 + 2\delta_0^2. \tag{EC.132}$$

Recalling that $\mathrm{Prox}_{\frac{\lambda|\mathcal{O}|}{2}\|\mathbf{L}\|_* + \mathfrak{T}\{\mathbf{L}\in C_1\}}(\mathbf{L}_{k+0.5})$ is defined as

$$\arg\min_{\mathbf{L}} \|\mathbf{L} - \mathbf{L}_{k+0.5}\|^2 + \lambda|\mathcal{O}|\|\mathbf{L}\|_* + \mathfrak{T}\{\mathbf{L}\in C_1\}, \tag{EC.133}$$

we have

$$\begin{aligned}
&\|\mathrm{Prox}(\mathbf{L}_{k+0.5}) - \mathbf{L}_{k+0.5}\|^2 + \lambda|\mathcal{O}|\|\mathrm{Prox}(\mathbf{L}_{k+0.5})\|_* + \mathfrak{T}\{\mathrm{Prox}(\mathbf{L}_{k+0.5})\in C_1\} \\
&\le \|\mathbf{0} - \mathbf{L}_{k+0.5}\|^2 + \lambda|\mathcal{O}|\|\mathbf{0}\|_* + \mathfrak{T}\{\mathbf{0}\in C_1\} = \|\mathbf{L}_{k+0.5}\|^2 \\
&\le \|Y\|_2^2 + (\sqrt{NT - |\mathcal{O}|}L_{\max} + \delta_0)^2.
\end{aligned} \tag{EC.134}$$

Combing Proposition 4.1 and Proposition 4.2, we have for $\beta \le \frac{6}{17}$, then

$$\begin{aligned}
\|\mathbf{L}_{k+1} - \mathrm{Prox}(\mathbf{L}_{k+0.5})\|^2 \le \frac{1}{\beta k}\Bigg( & (3\beta^2+8)\left(\|Y\|_2^2 + (\sqrt{NT-|\mathcal{O}|}L_{\max}+\delta_0)^2\right) \\
& + \left(5+\frac{8}{3}(\frac{\beta}{1+\beta})^2\right)(\lambda|\mathcal{O}|)^2\min\{N,T\} \\
& + \left(\beta^2+\frac{8}{3}(\frac{\beta^2}{1+\beta})^2\right)(2C(Y)^2+2\delta_0^2)\Bigg) \\
\le \frac{1}{k}\Bigg( & \delta_0^2\left(\frac{1}{\beta}\left(6\beta^2+16+2\beta^2+\frac{16}{3}(\frac{\beta^2}{1+\beta})^2\right)\right) + \\
& \frac{1}{\beta}(3\beta^2+8)\left(\|Y\|^2+2(NT-|\mathcal{O}|)L_{\max}^2\right) + \\
& \frac{1}{\beta}\left(5+\frac{8}{3}(\frac{\beta}{1+\beta})^2\right)(\lambda|\mathcal{O}|)^2\min\{N,T\} + \\
& \beta\left(2+\frac{16}{3}(\frac{\beta}{1+\beta})^2\right)C(Y)^2\Bigg).
\end{aligned} \tag{EC.135}$$

Let

$$\begin{aligned}
q_0(\beta) &= \left(\frac{1}{\beta}\left(6\beta^2+16+2\beta^2+\frac{16}{3}(\frac{\beta^2}{1+\beta})^2\right)\right), \\
q_1(\beta) &= \frac{1}{\beta}\left(5+\frac{8}{3}(\frac{\beta}{1+\beta})^2\right), \\
q_2(\beta) &= \beta\left(2+\frac{16}{3}(\frac{\beta}{1+\beta})^2\right), \\
q_3(\beta) &= \frac{1}{\beta}(3\beta^2+8), \\
\delta(k) &= \sqrt{\frac{q_1(\beta)(\lambda|\mathcal{O}|)^2\min\{N,T\} + q_2(\beta)C(Y)^2 + q_3(\beta)\left(\|Y\|^2+2(NT-|\mathcal{O}|)L_{\max}^2\right)}{k-q_0(\beta)}}
\end{aligned} \tag{EC.136}$$

We show next that when $k \ge q_0(\beta)$, $\inf_{\mathbf{L}\in C_1}\|\mathbf{L}_k - \mathbf{L}\| \le \delta(k)$ and $\|\mathbf{L}_{k+1} - \mathrm{Prox}(\mathbf{L}_{k+0.5})\| \le \delta(k)$ for all $k \ge 0$. For $k=0$, $\mathbf{L}_0 \in C_1$, the first part claim holds. Suppose the first part of claim holds for $k \le k_0$, where $k_0 \ge 0$, then for $k = k_0+1$,

$$\begin{aligned}
& \|\mathbf{L}_{k_0+1} - \mathrm{Prox}(\mathbf{L}_{k_0+0.5})\|^2 \\
& \le \frac{1}{k}\Bigg(\delta(k)^2 q_0(\beta) + q_1(\beta)(\lambda|\mathcal{O}|)^2\min\{N,T\} + q_2(\beta)C(Y)^2 \\
& \qquad + q_3(\beta)\left(\|Y\|^2+2(NT-|\mathcal{O}|)L_{\max}^2\right)\Bigg) \\
& = \delta(k)^2.
\end{aligned} \tag{EC.137}$$

Since $\text{Prox}(\mathbf{L}_{k_0+0.5}) \in C_1$, the first part of claim holds for $k = k_0 + 1$. So the first part of the holds for all $k \geq 0$. Since Inequality EC.137 is based on $\|\mathbf{L}_{k_0} - \text{Proj}_{C_1}(\mathbf{L}_{k_0})\| \leq \delta(k)$, it holds for all $k_0 \geq 0$.

Therefore, for $k \geq q_0(\beta)$, we have $\delta_0 \leq \delta(k)$. Therefore, we know that $\delta_1 \leq \delta(k)$.

Now we proceed with bounding $\delta$. According to Proposition 4.1, we have

$$\begin{aligned}\delta &\leq \frac{2}{|\mathcal{O}|}\left(\frac{1}{2K}\|\mathbf{L}_0 - \hat{\mathbf{L}}\|^2 + \delta(k)^2 + \left(2L_{\max}\sqrt{NT} + C(Y) + \min\{\sqrt{N},\sqrt{T}\}\frac{\lambda|\mathcal{O}|}{2}\right)\delta(k)\right) \\ &\leq \frac{NTL_{\max}^2}{K|\mathcal{O}|} + \frac{2\delta(k)^2}{|\mathcal{O}|} + \left(\frac{4L_{\max}\sqrt{NT}}{|\mathcal{O}|} + \frac{2C(Y)}{|\mathcal{O}|} + \min\{\sqrt{N},\sqrt{T}\}\lambda\right)\delta(k).\end{aligned} \tag{EC.138}$$

This finishes the proof.

## EC.14.1. Proof of Theorem 8

Recall that we use $\rho^2(\Sigma)$ to denote the maximum diagonal entry of the covariance matrix $\Sigma$.

It suffices to prove the following two results

PROPOSITION EC.14.1. Under the linear regression model (50), for any sparse index set $S$ such that the cardinal of $S$, $|S| = s$, denote $\theta^*_{S^c}$ to be the vector keeping elements not in $S$ the same and setting those in $S$ to be 0. Suppose $c_1\kappa \geq 64s \cdot c_2\rho^2(\Sigma)\frac{\log d}{n}$, where $c_1, c_2$ are constants and can be taken as $c_1 = 1/8, c_2 = 50$, and $\kappa$ is the smallest singular value of $\Sigma$. For $\lambda_n \geq \frac{2\|\mathbf{X}^T w\|_\infty}{n}$, $\tilde{\theta}$ satisfying (52) has the following property

$$P(\|\Delta\|_2 < \frac{\delta}{2\lambda_n\sqrt{s}} + \frac{\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}) \geq 1 - \frac{\exp(-n/32)}{1 - \exp(-n/32)}. \tag{EC.139}$$

LEMMA EC.7. *For the random matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, in which each row $x_i$ is drawn i.i.d. from a $N(0,\Sigma)$ distribution, its columns $\tilde{x}_k$ satisfies the following with probability at least $1 - \exp(-\frac{n}{4\rho^2(\Sigma)}\epsilon)$,*

$$\max_{1\leq k\leq d}\frac{\|\tilde{x}_k\|_2^2}{n} \leq 2\log 2 \cdot \rho^2(\Sigma) + \frac{4\rho^2(\Sigma)}{n}\log d + \epsilon. \tag{EC.140}$$

*For $w$ with $w_i \overset{i.i.d.}{\sim} N(0,\sigma^2)$ and $w$ independent with $\mathbf{X}$, we have that*

$$P_{\mathbf{X},w}\left(\|\frac{\mathbf{X}^T w}{n}\|_\infty < 2\rho(\Sigma)\sqrt{(\frac{\log d}{n} + 1)}\sigma\sqrt{\frac{2\log(2d)}{n} + \mu}\right) \geq 1 - \exp(-\frac{n}{2}) - \exp(-\frac{n\mu}{2}). \tag{EC.141}$$

**EC.14.1.1. Proof of Proposition EC.15.1** From Inequality (52), we have

$$\|y-\mathbf{X}\tilde{\theta}\|_2^2+\lambda_n\|\tilde{\theta}\|_1\le\|y-\mathbf{X}\hat{\theta}\|_2^2+\lambda_n\|\hat{\theta}\|_1+\delta\le\|y-\mathbf{X}\theta^*\|_2^2+\lambda_n\|\theta^*\|_1+\delta. \tag{EC.142}$$

Denote $\Delta=\tilde{\theta}-\theta^*$.

Therefore, we have that

$$\begin{aligned}0\le\frac{1}{2n}\|\mathbf{X}\Delta\|^2\le&\|\frac{\mathbf{X}^Tw}{n}\|_\infty\|\Delta\|_1+\lambda_n\left(\|\theta^*\|_1-\|\tilde{\theta}\|_1\right)+\delta\\ \le&\frac{\lambda_n}{2}\left(\|\Delta\|_1+2\|\theta^*\|_1-2\|\tilde{\theta}\|_1\right)+\delta\\ \le&\frac{\lambda_n}{2}\left(3\|\Delta_S\|_1+\|\Delta_{S^c}\|_1+2\|\theta^*_{S^c}\|_1-2\|\theta^*_{S^c}+\Delta_{S^c}\|_1\right)+\delta\\ \le&\frac{\lambda_n}{2}\left(3\|\Delta_S\|_1-\|\Delta_{S^c}\|_1+4\|\theta^*_{S^c}\|_1\right)+\delta.\end{aligned} \tag{EC.143}$$

Therefore, we have

$$\|\Delta\|_1=\|\Delta_S\|_1+\|\Delta_{S^c}\|_1\le4\|\Delta_S\|_1+4\|\theta^*_{S^c}\|_1+\frac{2\delta}{\lambda_n}\le4\sqrt{s}\|\Delta\|_2+4\|\theta^*_{S^c}\|_1+\frac{2\delta}{\lambda_n}. \tag{EC.144}$$

On the other hand, according Theorem 7.16 in Wainwright (2019), we have that with probability at $1-\frac{\exp(-n/32)}{1-\exp(-n/32)}$,

$$\frac{\|\mathbf{X}\Delta\|_2^2}{n}\ge c_1\|\sqrt{\Sigma}\Delta\|_2^2-c_2\rho^2(\Sigma)\frac{\log d}{n}\|\Delta\|_1^2, \tag{EC.145}$$

where $c_1$, $c_2$ are absolute constants and can be taken as $c_1=1/8$, $c_2=50$.

Note that $\|\sqrt{\Sigma}\Delta\|_2^2\ge\kappa\|\Delta\|_2^2$, going back to Inequality (EC.143), we have

$$\begin{aligned}c_1\kappa\|\Delta\|_2^2\le&c_2\rho^2(\Sigma)\frac{\log d}{n}\|\Delta\|_1^2+\lambda_n(3\|\Delta_S\|_1-\|\Delta_{S^c}\|_1+4\|\theta^*_{S^c}\|_1)+2\delta\\ \le&c_2\rho^2(\Sigma)\frac{\log d}{n}\left(4\sqrt{s}\|\Delta\|_2+4\|\theta^*_{S^c}\|_1+\frac{2\delta}{\lambda_n}\right)^2\\ &+\lambda_n(3\|\Delta_S\|_1-\|\Delta_{S^c}\|_1+4\|\theta^*_{S^c}\|_1)+2\delta\\ \le&c_1\kappa\left(\frac{\|\Delta\|_2}{2}+\frac{\delta}{4\lambda_n\sqrt{s}}+\frac{\|\theta^*_{S^c}\|_1}{2\sqrt{s}}\right)^2+\lambda_n(3\|\Delta_S\|_1-\|\Delta_{S^c}\|_1+4\|\theta^*_{S^c}\|_1)+2\delta\\ \le&c_1\kappa\left(\frac{\|\Delta\|_2}{2}+\frac{\delta}{4\lambda_n\sqrt{s}}+\frac{\|\theta^*_{S^c}\|_1}{2\sqrt{s}}\right)^2+\lambda_n(3\sqrt{s}\|\Delta\|_2+4\|\theta^*_{S^c}\|_1)+2\delta.\end{aligned} \tag{EC.146}$$

Solving the Inequality for $\|\Delta\|_2$, we have

$$\|\Delta\|_2<\frac{\delta}{2\lambda_n\sqrt{s}}+\frac{\|\theta^*_{S^c}\|_1}{\sqrt{s}}+(2+4\sqrt{s}+\frac{1}{\sqrt{s}})\frac{\lambda_n}{c_1\kappa}. \tag{EC.147}$$

**EC.14.1.2. Proof of Lemma EC.7** Denote $\nu_k = \frac{\|\tilde{x}_k\|_2^2}{n}$.

For $\frac{n}{2\rho^2(\Sigma)} > \lambda > 0$,

$$\mathbb{E}(\exp(\lambda \max\{\nu_k : 1 \le k \le d\})) \le \sum_{k=1}^{d} \mathbb{E}(\exp(\nu_k \lambda)) \le d(\frac{1}{1 - \frac{2\lambda\rho^2(\Sigma)}{n}})^{\frac{n}{2}}. \tag{EC.148}$$

Therefore, for $\Delta > 0$

$$P(\max\{\nu_k : 1 \le k \le d\} > \Delta) \le d(\frac{1}{1 - \frac{2\lambda\rho^2(\Sigma)}{n}})^{\frac{n}{2}} \exp(-\lambda\Delta). \tag{EC.149}$$

Take $\lambda = \frac{n}{4\rho^2(\Sigma)}$, and $\Delta = 2\rho^2(\Sigma)\log 2 + \frac{4\rho^2(\Sigma)}{n}\log d + \epsilon$, we have

$$P(\max\{\nu_k : 1 \le k \le d\} > \Delta) \le \exp(-\frac{n}{4\rho^2(\Sigma)}\epsilon). \tag{EC.150}$$

Therefore, the proof of the first statement is concluded.

For the second statement, suppose $\max\{\nu_k : 1 \le k \le d\} \le C_\nu$. Then we have for $\lambda > 0$,

$$\mathbb{E}\left(\exp(\lambda \max\{|\frac{\tilde{X}_k^T w}{n}| : 1 \le k \le d\})\right) \le 2d \exp(\frac{\lambda^2}{n} C_\nu^2 \sigma^2/2). \tag{EC.151}$$

Therefore, for $\Delta > 0$,

$$P(\|\frac{\mathbf{X}w}{n}\|_\infty > \Delta) \le \exp(\log(2d) + \frac{\lambda^2}{n} C_\nu^2 \sigma^2/2 - \lambda\Delta). \tag{EC.152}$$

Take $\lambda = \frac{n\Delta}{C_\nu^2\sigma^2}$, we have

$$P(\|\frac{\mathbf{X}w}{n}\|_\infty > \Delta) \le \exp(\log(2d) - \frac{n\Delta^2}{2C_\nu^2\sigma^2}). \tag{EC.153}$$

Setting

$$\Delta = C_\nu \sigma \sqrt{\frac{2\log(2d)}{n} + \mu}, \tag{EC.154}$$

and note that $C_\nu \le \sqrt{4\rho^2(\Sigma) + 4\rho^2(\Sigma)\frac{\log d}{n}}$ with probability at least $1 - \exp(-\frac{n}{2})$, we have the statement of second inequality of the lemma.

## EC.15. Proof of Theorem 9

It's easy to check that

$$\frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2 \tag{EC.155}$$

is $\|\frac{\mathbf{X}^T\mathbf{X}}{n}\|_s$-smooth.

By Theorem 1, and take $\delta_0 = 0$ gives the result.

## EC.16. Proof of Theorem 10

It's easy to see that $\frac{1}{2n}\|\mathbf{X}\theta\|_2^2$ is $\frac{\|\mathbf{X}^T\mathbf{X}\|_s}{n}-$smooth, where $\|\cdot\|_s$ denotes the spectral norm.

Denote $L=\frac{\|\mathbf{X}^T\mathbf{X}\|_s}{n}$.

Note that we have an alternative expression for $\theta_{k+1}$ for $k\geq 0$:

$$\theta_{k+1}=\arg\min_{\theta}\frac{1}{2n}\|\mathbf{X}\theta_k\|_2^2+\langle\frac{\mathbf{X}^T\mathbf{X}\theta_k}{n},\theta-\theta_k\rangle+\frac{L}{2}\|\theta-\theta_k\|_2^2+\lambda_n\|\theta\|_1. \tag{EC.156}$$

For simplicity we define

$$\phi_k(\theta)=\frac{1}{2n}\|\mathbf{X}\theta_k\|_2^2+\langle\frac{\mathbf{X}^T\mathbf{X}\theta_k}{n},\theta-\theta_k\rangle+\frac{L}{2}\|\theta-\theta_k\|_2^2+\lambda_n\|\theta\|_1. \tag{EC.157}$$

Theorem 10.16 in Beck (2017) gives that

$$F(\theta)-F(\theta_{k+1})\geq\frac{L}{2}\|\theta-\theta_{k+1}\|_2^2-\frac{L}{2}\|\theta-\theta_k\|_2^2+D(\theta,\theta_k), \tag{EC.158}$$

where

$$D(\theta,\theta_k)=\frac{1}{2n}\|\mathbf{X}\theta\|_2^2-\frac{1}{2n}\|\mathbf{X}\theta_k\|_2^2-\langle\frac{\mathbf{X}^T\mathbf{X}\theta_k}{n},\theta-\theta_k\rangle. \tag{EC.159}$$

Taking $\theta=\theta_k$ gives

$$F(\theta_k)\geq F(\theta_{k+1})+\frac{L}{2}\|\theta_k-\theta_{k+1}\|_2^2. \tag{EC.160}$$

Taking $\theta=\theta^*$ gives

$$F(\theta^*)-F(\theta_{k+1})\geq\frac{L}{2}\|\theta^*-\theta_{k+1}\|_2^2-\frac{L}{2}\|\theta^*-\theta_k\|_2^2+D(\theta^*,\theta_k). \tag{EC.161}$$

Adding up the inequality from 1 to $k+1$ gives

$$\frac{L}{2}\|\theta^*\|_2^2\geq\sum_{j=1}^{k+1}F(\theta_j)-F(\theta^*)\geq(k+1)(F(\theta_{k+1})-F(\theta^*)). \tag{EC.162}$$

Taking $\theta=\hat{\theta}$, gives

$$F(\hat{\theta})-F(\theta_{k+1})\geq\frac{L}{2}\|\hat{\theta}-\theta_{k+1}\|_2^2-\frac{L}{2}\|\hat{\theta}-\theta_k\|_2^2+D(\hat{\theta},\theta_k). \tag{EC.163}$$

Adding up the inequality from 1 to $k+1$ gives

$$F(\theta_{k+1})-F(\hat{\theta})\leq\frac{1}{k+1}\frac{L}{2}\|\hat{\theta}\|_2^2. \tag{EC.164}$$

This gives the second statement of the theorem.

Recalling Inequality (EC.143), we have that

$$0 \leq 3\| \left(\theta_k - \theta^*\right)_S \|_1 - \| \left(\theta_k - \theta^*\right)_{S^c} \|_1 + 4\|\theta^*_{S^c}\|_1 + \frac{2(F(\theta_k) - F(\theta^*))}{\lambda_n}. \tag{EC.165}$$

This gives

$$\|\theta_k - \theta^*\|_1 \leq 4\sqrt{s}\|\theta_k - \theta^*\|_2 + 4\|\theta^*_{S^c}\|_1 + \frac{2(F(\theta_k) - F(\theta^*))}{\lambda_n}. \tag{EC.166}$$

Therefore

$$\|\hat{\theta} - \theta_k\|_1 \leq \|\theta_k - \theta^*\|_1 + \|\hat{\theta} - \theta^*\|_1 \leq 4\sqrt{s}\|\hat{\theta} - \theta_k\|_2 + 8\sqrt{s}\|\hat{\theta} - \theta^*\|_2 + 8\|\theta^*_{S^c}\|_1 + \frac{2(F(\theta_k) - F(\theta^*))}{\lambda_n}. \tag{EC.167}$$

Recall the definition of $\phi_k(\theta)$ in Equation (EC.157). For $0 < \alpha < 1$, we have

$$\begin{aligned} &F(\theta_{k+1}) \leq \phi_k(\theta_{k+1}) \leq \phi_k(\alpha\hat{\theta} + (1-\alpha)\theta_k) \\ &\leq \frac{1}{2n}\|\mathbf{X}\theta_k\|_2^2 + \alpha\langle\frac{\mathbf{X}^T\mathbf{X}\theta_k}{n}, \hat{\theta} - \theta_k\rangle + \frac{L\alpha^2}{2}\|\hat{\theta} - \theta_k\|_2^2 + \alpha\lambda_n\|\hat{\theta}\|_1 + (1-\alpha)\lambda_n\|\theta_k\|_1 \\ &\leq \alpha F(\hat{\theta}) + (1-\alpha)F(\theta_k) + \frac{L\alpha^2}{2}\|\theta_k - \hat{\theta}\|_2^2. \end{aligned} \tag{EC.168}$$

Now we will bound $\|\theta_k - \hat{\theta}\|_2^2$.

Note that $\hat{\theta}$ is the minimizer of $F(\theta)$, we have

$$\begin{aligned} &F(\theta_k) - F(\hat{\theta}) \\ &= F(\theta_k) - F(\hat{\theta}) - \langle \partial F(\hat{\theta}), \theta_k - \hat{\theta}\rangle \geq D(\theta_k, \hat{\theta}) \geq \frac{a_1}{2}\|\hat{\theta} - \theta_k\|_2^2 - \frac{a_2}{2}\|\hat{\theta} - \theta_k\|_1^2 \\ &\geq \frac{a_1}{2}\|\hat{\theta} - \theta_k\|_2^2 - \frac{a_2}{2}\left(4\sqrt{s}\|\theta_k - \hat{\theta}\|_2 + 8\sqrt{s}\|\hat{\theta} - \theta^*\|_2 + 8\|\theta^*_{S^c}\|_1 + \frac{2(F(\theta_k) - F(\theta^*))}{\lambda_n}\right)_+^2 \end{aligned} \tag{EC.169}$$

Since $a_1 \geq 64s \cdot a_2$, we have

$$\frac{a_1}{4}\|\hat{\theta} - \theta_k\|_2^2 \leq F(\theta_k) - F(\hat{\theta}) + a_2\left(8\sqrt{s}\|\hat{\theta} - \theta^*\|_2 + 8\|\theta^*_{S^c}\|_1 + \frac{2(F(\theta_k) - F(\theta^*))}{\lambda_n}\right)_+^2. \tag{EC.170}$$

Therefore

$$\begin{aligned} \|\hat{\theta} - \theta_k\|_2^2 \leq &\frac{4}{a_1}\left(F(\theta_k) - F(\hat{\theta})\right) + \frac{4a_2}{a_1} \cdot 128\left(\sqrt{s}\|\hat{\theta} - \theta^*\|_2 + \|\theta^*_{S^c}\|_1\right)^2 \\ &+ \frac{32a_2}{a_1}\left(\frac{F(\theta_k) - F(\theta^*)}{\lambda_n}\right)_+^2. \end{aligned} \tag{EC.171}$$

Let $\alpha = \frac{a_1}{4L}$ in Inequality (EC.168), we have that

$$F(\theta_{k+1}) - F(\hat{\theta}) \leq (1 - \frac{a_1}{8L}) \left( F(\theta_k) - F(\hat{\theta}) \right) + \frac{a_1}{4L} \cdot 64 a_2 s \cdot \left( \|\hat{\theta} - \theta^*\|_2 + \frac{\|\theta^*_{S^c}\|_1}{\sqrt{s}} \right)^2 + \frac{a_1 \cdot 64 a_2 s}{64 L \cdot s} \left( \frac{F(\theta_k) - F(\theta^*)}{\lambda_n} \right)_+^2. \tag{EC.172}$$

From Theorem 8 we have that

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{\|\theta^*_{S^c}\|_1}{\sqrt{s}} + (2 + 4\sqrt{s} + \frac{1}{\sqrt{s}}) \frac{\lambda_n}{c_1 \kappa}. \tag{EC.173}$$

Plug in Inequality (EC.173) into Inequality (EC.172) and note that $F(\theta_k) - F(\theta^*) \leq F(\theta_k) - F(\hat{\theta}) \leq F(\theta_K) - F(\hat{\theta})$ for $K \leq k$ gives the statement.