

---

# Estimating Generalization Performance Along the Trajectory of Proximal SGD in Robust Regression

---

**Kai Tan**

Department of Statistics  
Rutgers University  
Piscataway, NJ 08854  
kai.tan@rutgers.edu

**Pierre C. Bellec**

Department of Statistics  
Rutgers University  
Piscataway, NJ 08854  
pierre.bellec@rutgers.edu

## Abstract

This paper studies the generalization performance of iterates obtained by Gradient Descent (GD), Stochastic Gradient Descent (SGD) and their proximal variants in high-dimensional robust regression problems. The number of features is comparable to the sample size and errors may be heavy-tailed. We introduce estimators that precisely track the generalization error of the iterates along the trajectory of the iterative algorithm. These estimators are provably consistent under suitable conditions. The results are illustrated through several examples, including Huber regression, pseudo-Huber regression, and their penalized variants with non-smooth regularizer. We provide explicit generalization error estimates for iterates generated from GD and SGD, or from proximal SGD in the presence of a non-smooth regularizer. The proposed risk estimates serve as effective proxies for the actual generalization error, allowing us to determine the optimal stopping iteration that minimizes the generalization error. Extensive simulations confirm the effectiveness of the proposed generalization error estimates.

## 1 Introduction

Consider the linear model:

$$\mathbf{y} = \mathbf{X}\mathbf{b}^* + \boldsymbol{\varepsilon}, \quad (1)$$

where  $\mathbf{y} \in \mathbb{R}^n$  is the response vector,  $\mathbf{X} \in \mathbb{R}^{n \times p}$  is the design matrix,  $\mathbf{b}^* \in \mathbb{R}^p$  is the unknown regression vector, and  $\boldsymbol{\varepsilon} \in \mathbb{R}^n$  is the noise vector that we assume independent of  $\mathbf{X}$ . The entries of  $\boldsymbol{\varepsilon}$  may be heavy-tailed, for instance our working assumptions allow for infinite second moment.

For the estimation of  $\mathbf{b}^*$ , we consider the following regularized optimization problem

$$\hat{\mathbf{b}} \in \arg \min_{\mathbf{b} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(y_i - \mathbf{x}_i^\top \mathbf{b}) + g(\mathbf{b}), \quad (2)$$

where  $\rho : \mathbb{R} \rightarrow \mathbb{R}$  is a data-fitting loss and  $g : \mathbb{R}^p \rightarrow \mathbb{R}$  is a regularization function. In the present robust regression setting, typical examples of  $\rho$  include the Huber [14] loss  $\rho(r; \delta) = \delta^2 \int_0^{|r|/\delta} \min(1, x) dx$ , the Pseudo-Huber loss  $\rho(r; \delta) = \delta^2(\sqrt{1 + (r/\delta)^2} - 1)$  or other Lipschitz loss functions to combat the possible heavy-tails of the additive noise. Typical examples of penalty functions include the L1/Lasso [27] penalty  $g(\mathbf{b}) = \lambda \|\mathbf{b}\|_1$ , group-Lasso penalty [29] for grouped variables, or their non-convex variants including for instance SCAD [12] or MCP [30].

In order to solve the optimization problem (2), practitioners resort to iterative algorithms, for instance gradient descent, accelerated gradient descent, stochastic gradient descent, and the corresponding proximal methods [20] in the presence of a non-smooth regularizer. Let the algorithm starts with

some initializer  $\hat{\mathbf{b}}^1 \in \mathbb{R}^p$  (typically  $\hat{\mathbf{b}}^1 = \mathbf{0}$ ) followed by consecutive iterates  $\hat{\mathbf{b}}^2, \hat{\mathbf{b}}^3, \dots$ , where  $\hat{\mathbf{b}}^t$  is typically obtained, for gradient descent and its variants as will be detailed below, from  $\hat{\mathbf{b}}^{t-1}$  and by applying an additive correction involving the gradient of the objective function. Our goal of this paper is to quantify the predictive performance of each iterate  $\hat{\mathbf{b}}^t$ .

We assume throughout that the covariance  $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top] = \Sigma$  of the feature vectors is finite. We measure the predictive performance of  $\hat{\mathbf{b}}^t$  using the out-of-sample error

$$\mathbb{E} \left[ \left( \mathbf{x}_{new}^\top \hat{\mathbf{b}}^t - \mathbf{x}_{new}^\top \mathbf{b}^* \right)^2 \mid (\mathbf{x}_i, y_i)_{i \in [n]} \right] = \|\Sigma^{1/2}(\hat{\mathbf{b}}^t - \mathbf{b}^*)\|^2$$

where  $\mathbf{x}_{new}$  is a new feature vector, independent of the data  $(\mathbf{x}_i, y_i)_{i \in [n]}$  and has the same distribution as  $\mathbf{x}_i$ . The above squared metric is used because the noise  $\varepsilon_i$  (and thus  $y_i$ ) is allowed to have infinite variance, and in this case the squared prediction error  $\mathbb{E}[(\mathbf{x}_{new}^\top \hat{\mathbf{b}}^t - y_{new})^2 \mid (\mathbf{x}_i, y_i)_{i \in [n]}] = +\infty$  irrespective of the value of  $\hat{\mathbf{b}}^t$ .

The paper proposes to estimate the out-of-sample error  $\|\Sigma^{1/2}(\hat{\mathbf{b}}^t - \mathbf{b}^*)\|^2$  of the  $t$ -th iterate using the right-hand side of the approximation

$$\|\Sigma^{1/2}(\hat{\mathbf{b}}^t - \mathbf{b}^*)\|^2 + \|\varepsilon\|^2/n \approx \left\| (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^t) + \sum_{s=1}^{t-1} w_{t,s} \mathbf{S}_s \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^s) \right\|^2/n, \quad (3)$$

where  $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$  is the derivative of  $\rho$  acting component-wise on each coordinate in  $\mathbb{R}^n$  and  $\mathbf{S}_s$  is a diagonal matrix of the form  $\mathbf{S}_s = \sum_{i \in I_s} \mathbf{e}_i \mathbf{e}_i^\top$  where  $I_s \subset [n]$  is the batch for the  $s$ -th stochastic gradient update and  $\mathbf{e}_i \in \mathbb{R}^n$  is the  $i$ -th canonical basis vector. Here the  $w_{t,s}$  are quantities, introduced in Section 3.3 below, that can be computed from data and do not require the knowledge of  $\Sigma$ . The approximation (3) is made rigorous in Theorem 3.6, where the right-hand side is proved to be consistent (i.e., the difference between the two sides of the inequality converges to 0 in probability) for a first set of weights  $(w_{s,t})_{s < t}$ , and in Theorem 3.7 where a second set of weights are proposed.

Because the right-hand side of (3) is observable from the data and the iterates  $(\hat{\mathbf{b}}^s)_{s \leq t}$  are computed from the iterative algorithm, the approximation (3) lets us compare the out-of-sample error of iterates  $\hat{\mathbf{b}}^t$  at different time  $t$  up to the additive term  $\|\varepsilon\|^2/n$  (which does not depend on  $t$  nor on the choice of the iterative scheme or the choice of loss and penalty). It also lets us compare different tuning parameters, for instance learning rate, multiplicative parameter of the penalty function, batch size in Stochastic Gradient Descent (SGD). The right-hand side of (3) can serve as the criteria to choose the iteration number or tuning parameters that achieves the smallest out-of-sample error.

## 1.1 Related literature

Estimation of prediction risk of regression estimates has received significant attention in the last few decades. One natural avenue to estimate the generalization performance is to use  $V$ -fold cross-validation or leave-one-out schemes. In the proportional regime of interest here, where dimension  $p$  and sample size  $n$  are of the same order,  $V$ -fold cross-validation with finite  $V$ , e.g.,  $V = 5, 10$  is known to fail at consistently estimate the risk of the estimator trained on the full dataset [23, Figure 1]; this is simply explained because training with the biased sample size  $n(V-1)/V$  may behave differently than training with the full dataset. Leave-one-out schemes, or drastically increasing  $V$ , requires numerous refitting and is thus computationally expensive.

This motivates computationally efficient estimates of the risk of an estimator trained on the full dataset without sample-splitting, including Approximate Leave-One-out (ALO) schemes [23] that do not rely on sample-splitting and refitting; see [1] and the references therein for recent developments. For ridge regression and other estimators constructed from the square loss, the Generalized Cross-Validation (GCV) [28] has been shown to be effective, and it avoids data-splitting and refitting; it only needs to fit the full data once and then adjust the training error by a multiplicative factor larger than 1. Beyond ridge regression, the extension of GCV using degrees-of-freedom has been studied for Lasso regression [2, 3, 18, 9], and alternatives were developed for robust M-estimators [3, 4]. While ALO or GCV and its extensions are good estimators of the predictive risk of a solution  $\hat{\mathbf{b}}$  to the optimization problem (2), they are not readily applicable to quantify the prediction risk of iterates  $\hat{\mathbf{b}}^t$  obtained by widely-used iterative algorithms such as gradient descent (GD), stochastic gradient

descent (SGD) or their proximal variants: ALO and GCV focus on estimating the final ( $t \rightarrow +\infty$ ) iterate of the algorithm, when a solution  $\hat{\mathbf{b}}$  in (2) is found. Our goal in the present paper is to develop risk estimation methodologies along the trajectory of the algorithm.

Luo et al. [17] developed methods to estimate the cross-validation error of iterates that solves an empirical risk minimization problem. Their approach requires the Hessian of the objective function to be well-conditioned (i.e., the smallest and largest eigenvalues are bounded) along all iterates. This condition is not satisfied for the regression problems we consider in this paper, such as high-dimensional robust regression with a Lasso penalty. In the context of least squares problems with both  $p$  and  $n$  being large, [7] studied the fundamental limits on the performance of first order methods, showing that these are dominated by a specific Approximate Message Passing algorithm. Paquette et al. [19] demonstrated that the dynamics of Stochastic Gradient Descent (SGD) become deterministic in the large sample and dimensional limit, providing explicit expressions for these dynamics when the design matrix is isotropic. Our work differs from [19] in two key ways: First, we address a more general regression problem incorporating a non-smooth regularizer, thereby considering both SGD and proximal SGD; second, we offer explicit risk estimates for each iteration, rather than focusing solely on the theoretical dynamics of the iterates. Celentano et al. [8] and [13] characterize the dynamical mean-field dynamics of iterative schemes, and identify that the limiting process involves a “memory” kernel, describing how the dynamics of early iterates affect later ones.

Most recently, and most closely related to the present paper, [5] proposed risk estimate for iterates  $\hat{\mathbf{b}}^t$  obtained by running gradient descent and proximal gradient descent methods for solving penalized least squares optimizations. However, [5] focuses exclusively on the square loss for  $\rho$  in (3), which is not readily applicable to robust regression with heavy tailed noise for which the Huber or other robust losses must be used. Bellec and Tan [5] is further restricted to gradient updates using the full dataset, which does not cover stochastic gradient descent. While several proof techniques used in the present paper are inspired by [5], we will explain in Remark 3.8 that directly generalizing [5] to SGD in robust regression leads to a poor risk estimate for small batch sizes. The proposal of the present paper leverages out of batch samples to overcome this issue.

For gradient descent for the square loss and without penalty, Patil et al. [21] demonstrates both the failure of GCV along the trajectory and the success of computationally expensive leave-one out schemes, and develops a proposal to reduce the computational cost. Finally, let us mention the works [10, 16] that characterize the dynamics of the iterates in phase retrieval and matrix sensing problems, assuming that a fresh batch of observations (independent of all previous updates) is used at each iteration. This is different from the usual SGD setting studied in the present paper where the observations used during a stochastic gradient update may be reused in future stochastic gradient updates, creating intricate probabilistic dependence between gradient updates at different iterations.

Robust regression is highly valuable in real data analysis due to its ability to handle heavy-tailed noise effectively, and we will see below that the use of stochastic gradient updates and data-fitting loss functions different from the square loss require estimates that have a drastically different structure than in the square loss case. The present paper develops generalization error estimates in situations where no consistent estimate have been proposed: (1) we develop generalization error estimates along the trajectory of iterative algorithms aimed at solving (2) for robust loss functions including the pseudo-Huber loss; (2) the estimates are applicable not only to gradient updates involving the full dataset (gradient descent and its variants), but also to SGD and proximal SGD where a random batch is used for each update.

## 2 Problem setup

The paper studies iterative algorithms aimed at solving the optimization problem (2). We consider the algorithm that generates iterates  $\hat{\mathbf{b}}^t$  for  $t = 1, 2, \dots, T$  according to the following iteration:

$$\hat{\mathbf{b}}^{t+1} = \phi_t \left( \hat{\mathbf{b}}^t + \frac{\eta_t}{|I_t|} \mathbf{X}^\top \mathbf{S}_t \psi(\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^t) \right), \quad (4)$$

where  $\mathbf{S}_t \in \mathbb{R}^{n \times n}$  is the diagonal matrix  $\mathbf{S}_t = \sum_{i \in I_t} \mathbf{e}_i \mathbf{e}_i^\top$  for  $I_t \subset [n]$  the  $t$ -th batch (independent of  $(\mathbf{X}, \mathbf{y})$ ), where  $\phi_t : \mathbb{R}^p \mapsto \mathbb{R}^p$  and  $\psi : \mathbb{R}^n \mapsto \mathbb{R}^n$  are two functions and  $\eta_t$  is the step size. Typically,  $\psi : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is the componentwise application of  $\rho'$  (where  $\rho$  is the data-fitting loss in (2)), and the matrix  $\mathbf{S}_t \in \mathbb{R}^{n \times n}$  is diagonal with elements in  $\{0, 1\}$  encoding the observations  $i \in [n]$

used in the  $t$ -th stochastic gradient update. The presence of  $S_t$  and possibly nonlinear function  $\psi$  is such that the above iteration scheme is not covered by previous related works including [5, 21], which only tackle  $S_t = I_n$  (full batch gradient updates) and  $\psi : \mathbb{R}^n \rightarrow \mathbb{R}^n$  the identity map ( $\rho$  in (2) restricted to be the square loss). The iterative scheme (4), on the other hand, covers SGD with robust loss functions.

In the next section, we first provide a few examples of algorithms encompassed in the general iteration (4). This includes Gradient Descent (GD), Stochastic Gradient Descent (SGD), and their corresponding proximal methods [20], Proximal GD and Proximal SGD. GD and SGD are widely used in practice, while the proximal methods are particularly useful for solving the optimization problem (2) with non-smooth regularizers.

## 2.1 Robust regression without penalty

If there are no penalties in (2), i.e.,  $g(\mathbf{b}) = 0$ , then the minimization problem becomes

$$\hat{\mathbf{b}} \in \arg \min_{\mathbf{b} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(y_i - \mathbf{x}_i^\top \mathbf{b}).$$

To solve this problem, provided  $\rho$  is differentiable, one may use gradient descent (GD) and stochastic gradient descent (SGD).

**Example 2.1 (GD).** The GD method consists of the following iteration:

$$\hat{\mathbf{b}}^{t+1} = \hat{\mathbf{b}}^t + \frac{\eta_t}{n} \mathbf{X}^\top \psi(\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^t), \quad (5)$$

where  $\psi$  is the derivative of  $\rho$  acting component-wise on its argument, and  $\eta_t$  is the step size (also known as learning rate). For the least squares loss  $\rho(x) = x^2/2$ , we have  $\psi(\mathbf{u}) = \mathbf{u}$ .

**Example 2.2 (SGD).** Suppose at  $t$ -th iteration, we use the batch  $I_t \subset [n]$  to compute the gradient,

$$\hat{\mathbf{b}}^{t+1} = \hat{\mathbf{b}}^t + \frac{\eta_t}{|I_t|} \sum_{i \in I_t} \mathbf{x}_i \psi(y_i - \mathbf{x}_i^\top \hat{\mathbf{b}}^t) = \hat{\mathbf{b}}^t + \frac{\eta_t}{|I_t|} \mathbf{X}^\top S_t \psi(\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^t), \quad (6)$$

where  $S_t = \sum_{i \in I_t} \mathbf{e}_i \mathbf{e}_i^\top$  and  $\mathbf{e}_i$  is the  $i$ -th canonical vector in  $\mathbb{R}^n$ . If  $I_t = [n]$  for each  $t$ , then  $|I_t| = n$  and  $S_t = I_n$ , hence this SGD method reduces to the GD method in (5).

## 2.2 Robust regression with Lasso penalty

Regularized regression is useful for high-dimensional regression problems where  $p$  is larger than  $n$ . We consider the Lasso penalty  $g(\mathbf{b}) = \lambda \|\mathbf{b}\|_1$  to fight for the curse of dimensionality and obtain sparse estimates (our working assumptions, on the other hand, do not assume that the ground truth  $\mathbf{b}^*$  is sparse). While GD and SGD are not directly applicable to solve the optimization problem (2) with Lasso penalty due to  $\|\cdot\|_1$  lacking differentiability at 0, Proximal Gradient Descent (Proximal GD) [20] and Stochastic Proximal Gradient Descent (Proximal SGD) can be used to solve this optimization with Lasso penalty.

**Example 2.3 (Proximal GD).** For  $g(\mathbf{b}) = \lambda \|\mathbf{b}\|_1$  in (2), the Proximal GD gives the following iterations:

$$\hat{\mathbf{b}}^{t+1} = \text{soft}_{\lambda \eta_t} \left( \hat{\mathbf{b}}^t + \frac{\eta_t}{n} \mathbf{X}^\top \psi(\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^t) \right),$$

where  $\text{soft}_\theta(\cdot)$  applies the soft-thresholding  $u \mapsto \text{sign}(u)(|u| - \theta)_+$  component-wise.

**Example 2.4 (Proximal SGD).** Similar to the Proximal GD, the Proximal SGD consists of the following iterations:

$$\hat{\mathbf{b}}^{t+1} = \text{soft}_{\lambda \eta_t} \left( \hat{\mathbf{b}}^t + \frac{\eta_t}{|I_t|} \mathbf{X}^\top S_t \psi(\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^t) \right).$$

Let  $\rho' : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be the function applies the derivative of  $\rho : \mathbb{R} \rightarrow \mathbb{R}$  to each of its component, i.e.,  $\rho'(\mathbf{u}) = (\rho'(u_1), \dots, \rho'(u_n))^\top$ . Then the above examples can be summarized in the following table with different definition of  $\psi$ ,  $\phi_t$ , and  $S_t$ .

Table 1: Specification of  $\psi, \phi_t, \mathbf{S}_t$  for each algorithm

|                      | GD                  | SGD                 | Proximal GD                               | Proximal SGD                              |
|----------------------|---------------------|---------------------|-------------------------------------------|-------------------------------------------|
| $\psi(\mathbf{u})$   | $\rho'(\mathbf{u})$ | $\rho'(\mathbf{u})$ | $\rho'(\mathbf{u})$                       | $\rho'(\mathbf{u})$                       |
| $\phi_t(\mathbf{v})$ | $\mathbf{v}$        | $\mathbf{v}$        | $\text{soft}_{\lambda\eta_t}(\mathbf{v})$ | $\text{soft}_{\lambda\eta_t}(\mathbf{v})$ |
| $\mathbf{S}_t$       | $\mathbf{I}_n$      | $\mathbf{S}_t$      | $\mathbf{I}_n$                            | $\mathbf{S}_t$                            |

To define the proposed estimators of the generalization error, we further define the following Jacobian matrices:

$$\mathbf{D}_t = \frac{\partial \psi(\mathbf{r})}{\partial \mathbf{r}} \Big|_{\mathbf{r}=\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^t} \in \mathbb{R}^{n \times n}, \quad \tilde{\mathbf{D}}_t = \frac{\partial \phi_t(\mathbf{v})}{\partial \mathbf{v}} \Big|_{\mathbf{v}=\hat{\mathbf{b}}^t + \frac{\eta_t}{|I_t|} \mathbf{X}^\top \mathbf{S}_t \psi(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^t)} \in \mathbb{R}^{p \times p}.$$

Then, we have  $\tilde{\mathbf{D}}_t = \mathbf{I}_p$  for GD and SGD, and  $\tilde{\mathbf{D}}_t = \sum_{j \in \hat{S}_t} \mathbf{e}_j \mathbf{e}_j^\top$  for Proximal GD and Proximal SGD based on soft-thresholding, where  $\hat{S}_t = \{j \in [p] : \mathbf{e}_j^\top \hat{\mathbf{b}}^{t+1} \neq 0\}$ .

### 3 Main results

**Assumption 3.1.** The design matrix  $\mathbf{X}$  has i.i.d. rows from  $\mathcal{N}_p(\mathbf{0}, \Sigma)$  for some positive definite matrix  $\Sigma$  satisfying  $0 < \lambda_{\min}(\Sigma) \leq 1 \leq \lambda_{\max}(\Sigma)$  and  $\|\Sigma\|_{\text{op}} \|\Sigma^{-1}\|_{\text{op}} \leq \kappa$ . We assume  $\text{Var}[\mathbf{x}_i^\top \mathbf{b}^*] \leq \delta^2$ , that is, the signal of the model (1) is bounded from above.

**Assumption 3.2.** The noise  $\varepsilon$  is independent of  $\mathbf{X}$  and has i.i.d. entries from a fixed distribution independent of  $n, p$ , with  $\mathbb{E}[|\varepsilon_i|] \leq \delta$ , that is, bounded first moment.

**Assumption 3.3.** The data fitting loss  $\rho : \mathbb{R} \rightarrow \mathbb{R}$  is convex, continuously differentiable and its derivative  $\psi$  is 1-Lipschitz and  $|\psi(x)| \leq \delta$  for all  $x \in \mathbb{R}$ . The function  $\phi_t$  is 1-Lipschitz and satisfies  $\phi_t(\mathbf{0}) = \mathbf{0}$ . The matrices  $\mathbf{S}_t = \sum_{i \in I_t} \mathbf{e}_i \mathbf{e}_i^\top$ , and  $|I_t| \geq c_0 n$  for some positive constant  $c_0 \in (0, 1]$ . Let  $\eta_{\max} = \max_{t \in [T]} \eta_t$ .

Huber loss and Psuedo-Huber loss all satisfy Assumption 3.3.

**Assumption 3.4.** The data fitting loss  $\rho$  is twice continuously differentiable with positive second derivative.

**Assumption 3.5.** The sample size  $n$  and feature dimension  $p$  satisfy  $p/n \leq \gamma$  for a constant  $\gamma \in (0, \infty)$ .

#### 3.1 Intuition regarding the estimates of the generalization error

This subsection provides the intuition behind the definition of the estimates define below. For the sake of clarify, and in this subsection only, assume that

$$\Sigma = \mathbf{I}_p, \quad \varepsilon = \mathbf{0}, \quad \eta_t/|I_t| = 1/n. \quad (7)$$

With the above working assumptions, the validity of the estimates defined below relies on the probabilistic approximation

$$\|\hat{\mathbf{b}}^t - \mathbf{b}^*\|^2 \approx \frac{1}{n} \sum_{i=1}^n \left( -\mathbf{x}_i^\top (\hat{\mathbf{b}}^t - \mathbf{b}^*) + \sum_{j=1}^p \mathbf{e}_j^\top \frac{\partial \hat{\mathbf{b}}^t}{\partial x_{ij}} \right)^2,$$

which was developed in [3] for risk estimation purposes, but outside the context of iterative algorithms. Above,  $\mathbf{e}_j \in \mathbb{R}^p$  is the  $j$ -th canonical basis vector. In the present noiseless case with  $\varepsilon = \mathbf{0}$ , the first term inside the squared norm in the right-hand side is equal to the residual  $y_i - \mathbf{x}_i^\top \hat{\mathbf{b}}^t$ , so that the above display resembles (3). Taking this probabilistic approximation for granted, to study the second term in the right-hand side, we must understand the derivatives of  $\hat{\mathbf{b}}^t$  with respect to the entries  $(x_{ij})_{i \in [n], j \in [p]}$  of  $\mathbf{X}$ . In (4), each iterate is a relatively simple function of the previous ones, with the simplifications (7) this is  $\hat{\mathbf{b}}^{t+1} = \phi_t(\hat{\mathbf{b}}^t + \mathbf{X}^\top \mathbf{S}_t \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^t)/n)$ . For  $t = 1$ , given that  $\hat{\mathbf{b}}^1$  is a constant initialization,  $\frac{\partial}{\partial x_{ij}} \hat{\mathbf{b}}^2 = \tilde{\mathbf{D}}_1 \mathbf{e}_j \mathbf{e}_i^\top \mathbf{S}_1 \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^1)/n - \tilde{\mathbf{D}}_1 \mathbf{X}^\top \mathbf{S}_1 \mathbf{D}_1 \mathbf{e}_i (\hat{\mathbf{b}}^1 - \mathbf{b}^*)_j/n$ . We find in the proof, that when summing these quantities over  $j \in [p]$ , the second term involving

$(\hat{\mathbf{b}}^1 - \mathbf{b}^*)_j$  is negligible, and the same negligibility holds at later iterations with terms involving  $(\hat{\mathbf{b}}^t - \mathbf{b}^*)_j$  (or any  $(\hat{\mathbf{b}}^s - \mathbf{b}^*)_j$ ,  $s \leq t$ ). By performing a similar simple calculation at the next iteration, and ignoring these terms, we find with  $f_i^1 = \mathbf{e}_i^\top \mathbf{S}_1 \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^1)$  and  $f_i^2 = \mathbf{e}_i^\top \mathbf{S}_2 \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^2)$  by the chain rule

$$\sum_{j=1}^p \mathbf{e}_j^\top \frac{\partial \hat{\mathbf{b}}^2}{\partial x_{ij}} \approx \underbrace{\text{Tr}\left[\frac{\tilde{\mathbf{D}}_1}{n}\right]}_{w_{2,1}} f_i^1, \quad \sum_{j=1}^p \mathbf{e}_j^\top \frac{\partial \hat{\mathbf{b}}^3}{\partial x_{ij}} \approx \underbrace{\text{Tr}\left[\frac{\tilde{\mathbf{D}}_2}{n}\right]}_{w_{3,2}} f_i^2 + \underbrace{\text{Tr}\left[\tilde{\mathbf{D}}_2(\mathbf{I}_p - \mathbf{X}^\top \mathbf{S}_2 \mathbf{D}_2 \mathbf{X}/n) \tilde{\mathbf{D}}_1/n\right]}_{w_{3,1}} f_i^1.$$

This reveals the weights  $(w_{s,t})_{s \leq t}$  in (3) at iteration  $t = 2$  and  $t = 3$ . We could continue this further by successive applications of the chain rule, although for later iterations this unrolling of the derivatives, capturing the interplay between the Jacobians  $\mathbf{D}_t$ ,  $\tilde{\mathbf{D}}_t$  and the stochastic gradient matrix  $\mathbf{S}_t$ , becomes increasingly complex. This recursive unrolling of the derivatives can be performed numerically at the same time as the computation of the iterates. On the other hand, for the mathematical proof, for the formal definition of the weights in (3) and for the proposed estimates of the generalization error, the matrix notation defined in the next subsection exactly captures this unrolling of the derivatives.

### 3.2 Formal matrix notation to capture recursive derivatives

We now set up the matrix notation that captures this recursive unrolling of the derivatives by the chain rule. Throughout,  $T$  is the final number of iterations. Define three block diagonal matrices  $\mathcal{D} \in \mathbb{R}^{nT \times nT}$ ,  $\tilde{\mathcal{D}} \in \mathbb{R}^{pT \times pT}$ , and  $\mathcal{S} \in \mathbb{R}^{nT \times nT}$  by  $\mathcal{D} = \sum_{t=1}^T ((\mathbf{e}_t \mathbf{e}_t^\top) \otimes \mathbf{D}_t)$ ,  $\tilde{\mathcal{D}} = \sum_{t=1}^T ((\mathbf{e}_t \mathbf{e}_t^\top) \otimes \tilde{\mathbf{D}}_t)$ , and  $\mathcal{S} = \sum_{t=1}^T ((\mathbf{e}_t \mathbf{e}_t^\top) \otimes \mathbf{S}_t)$ . Now we are ready to introduce the following matrices of size  $T \times T$ :

$$\mathbf{W} = \sum_{j=1}^p (\mathbf{I}_T \otimes \mathbf{e}_j^\top) (\mathbf{I}_T \otimes \Sigma^{1/2}) \Gamma (\mathbf{I}_T \otimes \Sigma^{1/2}) (\mathbf{I}_T \otimes \mathbf{e}_j), \quad (8)$$

$$\hat{\mathbf{A}} = \sum_{i=1}^n (\mathbf{I}_T \otimes \mathbf{e}_i^\top) \mathcal{D} (\mathbf{I}_T \otimes \mathbf{X}) \Gamma (\mathbf{I}_T \otimes \mathbf{X}^\top) (\mathbf{I}_T \otimes \mathbf{e}_i), \quad (9)$$

$$\hat{\mathbf{K}} = \sum_{t=1}^T \text{Tr}(\mathbf{D}_t) \mathbf{e}_t \mathbf{e}_t^\top - \sum_{i=1}^n (\mathbf{I}_T \otimes \mathbf{e}_i^\top) \mathcal{D} (\mathbf{I}_T \otimes \mathbf{X}) \Gamma (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{S} \mathcal{D} (\mathbf{I}_T \otimes \mathbf{e}_i), \quad (10)$$

where  $\Gamma = \mathcal{M}^{-1} \mathbf{L} (\mathbf{A} \otimes \mathbf{I}_p) \tilde{\mathcal{D}} \in \mathbb{R}^{pT \times pT}$ ,  $\mathbf{L} = \sum_{t=2}^T ((\mathbf{e}_t \mathbf{e}_{t-1}^\top) \otimes \mathbf{I}_p)$ ,  $\mathbf{A} = \sum_{t=1}^T \frac{\eta_t}{|\mathbf{I}_t|} \mathbf{e}_t \mathbf{e}_t^\top$ ,

$$\mathcal{M} = \begin{bmatrix} \mathbf{I}_p & & & \\ -\mathbf{P}_1 & \mathbf{I}_p & & \\ & \ddots & \ddots & \\ & & -\mathbf{P}_{T-1} & \mathbf{I}_p \end{bmatrix} \quad \text{and} \quad \mathbf{P}_t = \tilde{\mathbf{D}}_t \left( \mathbf{I}_p - \frac{\eta_t}{|\mathbf{I}_t|} \mathbf{X}^\top \mathbf{S}_t \mathbf{D}_t \mathbf{X} \right).$$

Although notationally involved, the purpose of these matrices is just to formalize the recursive computation of the derivatives by the chain rule mentioned in Section 3.1.

### 3.3 Main results: estimating the generalization error consistently

For each iterate  $\hat{\mathbf{b}}^t$ , define the target  $r_t$  (generalization error) and its estimate  $\hat{r}_t$  by

$$r_t \stackrel{\text{def}}{=} \|\Sigma^{1/2}(\hat{\mathbf{b}}^t - \mathbf{b}^*)\|^2 + \frac{\|\boldsymbol{\varepsilon}\|^2}{n}, \quad \hat{r}_t = \frac{1}{n} \left\| (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^t) + \sum_{s=1}^{t-1} w_{t,s} \mathbf{S}_s \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^s) \right\|^2, \quad (11)$$

where  $w_{t,s} := \mathbf{e}_t^\top \mathbf{W} \mathbf{e}_s$  and  $\mathbf{W} \in \mathbb{R}^{T \times T}$  is the matrix defined in (8). The following shows that  $|\hat{r}_t - r_t| \xrightarrow{P} 0$  (convergence to 0 in probability) under suitable assumptions.

**Theorem 3.6** (Proved in Appendix C.1). *Let Assumptions 3.1, 3.3 and 3.5 be fulfilled. Then  $\forall \epsilon > 0$ ,*

$$\mathbb{P}\left(|\hat{r}_t - r_t| > \epsilon\right) \leq \max\left\{1, \frac{C(T, \gamma, \eta_{\max}, c_0, \delta, \kappa)}{\epsilon}\right\} \left(\frac{1}{\sqrt{n}} + \mathbb{E}\left[\min\left\{1, \frac{\|\boldsymbol{\varepsilon}\|}{n}\right\}\right]\right). \quad (12)$$

*If additionally Assumption 3.2 holds then  $\mathbb{E}[\min\{1, \frac{\|\boldsymbol{\varepsilon}\|}{n}\}] \rightarrow 0$ , so that, as  $n, p \rightarrow +\infty$  while  $(T, \gamma, \eta_{\max}, c_0, \delta, \kappa, \epsilon)$  are held fixed, the right-hand side converges to 0 and  $\hat{r}_t - r_t$  converges to 0 in probability.*

This establishes that  $\hat{r}_t$  is consistent at estimating  $r_t$ . The statement  $\mathbb{E}[\min\{1, \|\varepsilon\|/n\}] \rightarrow 0$  is equivalent to  $\|\varepsilon\|^2/n^2 \xrightarrow{P} 0$  (convergence in probability), and is proved in [22] under the assumption that  $\mathbb{E}|\varepsilon_i| < +\infty$  with  $\varepsilon_i$  i.i.d. from a fixed distribution; this allows  $\text{Var}[\varepsilon_i] = +\infty$  as long as the first moment is finite. The expression of  $\mathbf{W}$  involves  $\Sigma$ , which is typically unknown in practice. Our next result provides a consistent estimate of  $\mathbf{W}$  using quantities that do not require the knowledge of  $\Sigma$ .

We propose to estimate  $\mathbf{W}$  by  $\widetilde{\mathbf{W}} \stackrel{\text{def}}{=} \widehat{\mathbf{K}}^{-1} \widehat{\mathbf{A}}$  where  $\widehat{\mathbf{K}}$  and  $\widehat{\mathbf{A}}$  are the  $T \times T$  matrices defined in (9)-(10). We define another estimate  $\tilde{r}_t$  by replacing  $\mathbf{W}$  in (11) with  $\widetilde{\mathbf{W}} = \widehat{\mathbf{K}}^{-1} \widehat{\mathbf{A}}$ :

$$\tilde{r}_t = \frac{1}{n} \left\| (\mathbf{y} - \mathbf{X} \widehat{\mathbf{b}}^t) + \sum_{s=1}^{t-1} \tilde{w}_{t,s} \mathbf{S}_s \psi(\mathbf{y} - \mathbf{X} \widehat{\mathbf{b}}^s) \right\|^2,$$

where  $\tilde{w}_{t,s} = \mathbf{e}_t^\top \widetilde{\mathbf{W}} \mathbf{e}_s$ .

**Theorem 3.7** (Proved in Appendix C.3). *Under Assumptions 3.1 and 3.3 to 3.5, for any  $\epsilon > 0$ ,*

$$\mathbb{P}\left(|\tilde{r}_t - r_t| > \epsilon\right) \leq 2e^{-n/18} + \max\left\{1, \frac{C(T, \gamma, \eta_{\max}, c_0, \delta, \kappa)}{\epsilon}\right\} \left[\frac{1}{\sqrt{n}} + \mathbb{E}[\min(1, \frac{\|\varepsilon\|}{n})]\right].$$

*If additionally Assumption 3.2 holds then  $\mathbb{E}[\min\{1, \frac{\|\varepsilon\|}{n}\}] \rightarrow 0$ , so that, as  $n, p \rightarrow +\infty$  while  $(T, \gamma, \eta_{\max}, c_0, \delta, \kappa, \epsilon)$  are held fixed, the right-hand side converges to 0 and  $\tilde{r}_t - r_t$  converges to 0 in probability.*

This establishes the consistency of  $\tilde{r}_t$ . The simulations presented next confirm that the two estimates  $\tilde{r}_t$  and  $\hat{r}_t$  both are accurate estimates of  $r_t$ . The estimate  $\tilde{r}_t$  has the advantage of not relying on the knowledge of  $\Sigma$  and are recommended in practice.

**Remark 3.8.** We highlight that directly generalizing the approach in [5] would lead to the approximation  $\tilde{\mathbf{A}} \approx \widetilde{\mathbf{K}} \mathbf{W}$ , where  $\tilde{\mathbf{A}}$  and  $\widetilde{\mathbf{K}}$  are given in (23) and (26), respectively. From  $\tilde{\mathbf{A}} \approx \widetilde{\mathbf{K}} \mathbf{W}$ , obtaining an estimate of  $\mathbf{W}$  requires inverting  $\widetilde{\mathbf{K}}$ . However, this inversion fails for SGD for small (but still very realistic) batch sizes of order  $0.1n$  in simulations (see Figure 4). The matrix  $\widetilde{\mathbf{K}}$  is lower triangular, and the reason for the lack of invertibility of  $\widetilde{\mathbf{K}}$  can be seen in the diagonal terms equal to  $\text{Tr}[\mathbf{S}_t \mathbf{D}_t]$  in (26), where  $\mathbf{S}_t \in \{0, 1\}^{n \times n}$  is the diagonal matrix with 1 in position  $(i, i)$  if and only if the  $i$ -th observation is used in the  $t$ -th batch. This diagonal element of  $\widetilde{\mathbf{K}}$  can easily be small (or even 0) for small batches, if the batch only contains observations such that  $(\mathbf{D}_t)_{ii}$  is 0 or small. Let  $\tilde{r}_t^{\text{sub}}$  denote the estimate of the same form as  $\tilde{r}_t$  but using the weight matrix  $\widetilde{\mathbf{K}}^{-1} \tilde{\mathbf{A}}$  instead. Simulation results in Figure 4 confirm that  $\tilde{r}_t^{\text{sub}}$  is suboptimal compared to our proposed  $\tilde{r}_t$ . For SGD and proximal SGD, we solved this issue regarding the invertibility of  $\widetilde{\mathbf{K}}$  by using out-of-batch samples in the construction of  $\widehat{\mathbf{K}}$  and  $\widehat{\mathbf{A}}$ , in order to avoid  $\mathbf{S}_t$  in the diagonal elements of  $\widetilde{\mathbf{K}}$  in equation (31). This is the key to making these estimators work for SGD and proximal SGD, and this use of out-of-batch samples is new compared to [5] (which only tackles the square loss with full-batch gradients).

**Remark 3.9.** The constant  $C(T, \gamma, \eta_{\max}, c_0, \delta, \kappa)$  in the above results is not explicit. Inspection of the proof reveals that the dependence of this constant in  $T$  is currently  $T^T$ , allowing  $T$  of order  $\log(n)/\log \log n$  before the bound becomes vacuous. Improving this dependence in  $T$  appears challenging and possibly out of reach of current tools, even for the well-studied Approximate Message Passing (AMP) algorithms. The papers [25, 24] feature for instance the same  $\log(n)/\log \log n$  dependence for approximating the risk of AMP. The preprint [15] offers the latest advances on the dependence on  $T$  in the bounds satisfied by AMP. It allows  $T \asymp \text{poly}(n)$  while still controlling certain AMP related quantities, although for the risk [15, equations (16)-(17)] the condition required on  $T$  is still logarithmic in  $n$ . This suggests that advances on this front are possible, at least for isotropic design and specific loss and regularizer such as those studied in [15]: Lasso or Robust M-estimation with no regularizer. Since these latest advances in [15] are obtained for specific estimates (Lasso or Robust M-estimation with no regularizer), it may be possible to follow a similar strategy and improve our bounds for specific examples of iterative algorithms closer to AMP, or algorithms featuring only separable losses and penalty. We leave such improvements for specific examples for future work, as the goal of the current paper is to cover a general framework allowing iterations of the form (4) with little restriction on the nonlinear functions except being Lipschitz.

## 4 Simulation

In this section, we present numerical experiments to assess the performance of the proposed risk estimates. All necessary code for reproducing these experiments is provided in the supplementary material and is publicly available in the GitHub repository <https://github.com/kaitan365/SGD-generalization-errors>. Our goal is to compare the performance of the proposed risk estimates with the true risk  $r_t$  for different regression methods and iterative algorithms.

We generate the dataset  $(\mathbf{X}, \mathbf{y})$  from the linear model (1), that is,  $\mathbf{y} = \mathbf{X}\mathbf{b}^* + \varepsilon$ . Here, the rows of  $\mathbf{X} \in \mathbb{R}^{n \times p}$  are sampled from a centered multivariate normal distribution with covariance matrix  $\Sigma = \mathbf{I}_p$ . The noise vector  $\varepsilon$  consists of i.i.d. entries drawn from a  $t$  distribution with two degrees of freedom so that the noise variance is infinite. The true regression vector  $\mathbf{b}^* \in \mathbb{R}^p$  is chosen with  $p/20$  nonzero entries, set to a constant value such that the signal strength  $\|\mathbf{b}^*\|^2$  equals 10.

We explore two scenarios of the  $(n, p)$  pairs and corresponding iterative algorithms:

- (i)  $(n, p) = (10000, 5000)$ : In this configuration, with  $n$  much larger than  $p$ , we examine Huber regression and Pseudo-Huber regression (without penalty or soft-thresholding). Both the gradient descent (GD) and stochastic gradient descent (SGD) algorithms are implemented for each type of regression.
- (ii)  $(n, p) = (10000, 12000)$ : Here, we investigate Huber regression and Pseudo-Huber regression with an L1 penalty,  $\lambda \|\mathbf{b}\|_1$  ( $\lambda = 0.002$ ) and corresponding soft-thresholding step. For each penalized regression, we employ the Proximal Gradient Descent (Proximal GD) and Stochastic Proximal Gradient Descent (Proximal SGD) algorithms.

In all algorithms, we start with the initial vector  $\hat{\mathbf{b}}^1 = \mathbf{0}_p$  and proceed with a fixed step size  $\eta = (1 + \sqrt{p/n_*})^{-2}$  where  $n_* = n$  for GD and proximal GD, and  $n_* = n/5$  for SGD and proximal SGD. We run each algorithm for  $T = 100$  steps. For SGD and Proximal SGD, batches  $I_t \subset \{1, 2, \dots, n\}$  are randomly sampled without replacement and independently of  $(\mathbf{X}, \mathbf{y}, (I_s)_{s \neq t})$ , each with cardinality  $|I_t| = \frac{n}{5}$ .

A crucial component of the proposed risk estimates  $\hat{r}_t$  and  $\tilde{r}_t$  involve the weight matrices  $\mathbf{W}$  and  $\tilde{\mathbf{W}}$ . The matrix  $\mathbf{W}$  is defined in Theorem 3.6, and  $\tilde{\mathbf{W}} = \hat{\mathbf{K}}^{-1} \hat{\mathbf{A}}$  is defined in Theorem 3.7. We employ Hutchinson's trace approximation to compute  $\mathbf{W}$ ,  $\hat{\mathbf{A}}$ , and  $\hat{\mathbf{K}}$ . This implementation is computationally efficient. We refer readers to [5, Section 4] for more details.

Recall that we have proposed two estimates for  $r_t = \|\Sigma^{1/2}(\hat{\mathbf{b}}^t - \mathbf{b}^*)\|^2 + \|\varepsilon\|^2/n$ , one is  $\hat{r}_t$  in Theorem 3.6 which requires knowing  $\Sigma = \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top]$ , and the other is  $\tilde{r}_t$  in Theorem 3.7 which does not need  $\Sigma$ . Since the quantity  $\|\varepsilon\|^2/n$  remains constant along the algorithm trajectory, we only focus on the estimation of  $\|\Sigma^{1/2}(\hat{\mathbf{b}}^t - \mathbf{b}^*)\|^2$ . We repeat each numerical experiment 100 times and present the aggregated results in Figures 1, 2 and 3.

In Figure 1, we focus on the scenario with  $(n, p) = (10000, 5000)$ , and plot the actual risk  $\|\Sigma^{1/2}(\hat{\mathbf{b}}^t - \mathbf{b}^*)\|^2$ , and its two estimates  $\hat{r}_t - \|\varepsilon\|^2/n$  and  $\tilde{r}_t - \|\varepsilon\|^2/n$  along with the 2 standard error bar for GD and SGD algorithms applied to both Huber and Pseudo Huber regression. In Figure 2, we focus on the scenario with  $(n, p) = (10000, 12000)$ , and present the risk curves for the Proximal GD and Proximal SGD algorithms applied to both L1-penalized Huber regression and Pseudo-Huber regression.

Figure 1 and Figure 2 confirm the three curves are in close agreement, indicating that the proposed estimates  $\hat{r}_t - \|\varepsilon\|^2/n$  and  $\tilde{r}_t - \|\varepsilon\|^2/n$  are consistent estimates of the actual risk  $\|\Sigma^{1/2}(\hat{\mathbf{b}}^t - \mathbf{b}^*)\|^2$ . The two estimates closely capture the risk  $\|\Sigma^{1/2}(\hat{\mathbf{b}}^t - \mathbf{b}^*)\|^2$  over the entire trajectory of the algorithms. For GD and Proximal GD, the risk curves exhibit a U-shape, first decreasing and then increasing, and the estimates  $\hat{r}_t$  and  $\tilde{r}_t$  closely capture this pattern. This suggests that the proposed estimates are reliable and can be used to monitor the risk of the iterates and find the optimal iteration (the iteration minimizing the generalization error) along the trajectory of the algorithm.

**Additional experiments: varying step sizes for different iterations.** We also conduct simulations to investigate the accuracy of the proposed risk estimates in a setting with varying step size. We consider two types of step sizes: 1).  $\eta_t = 1$  if  $t$  is odd, and  $\eta_t = 0$  if  $t$  is even; 2).  $\eta_t = 1$  if  $t$  is odd,



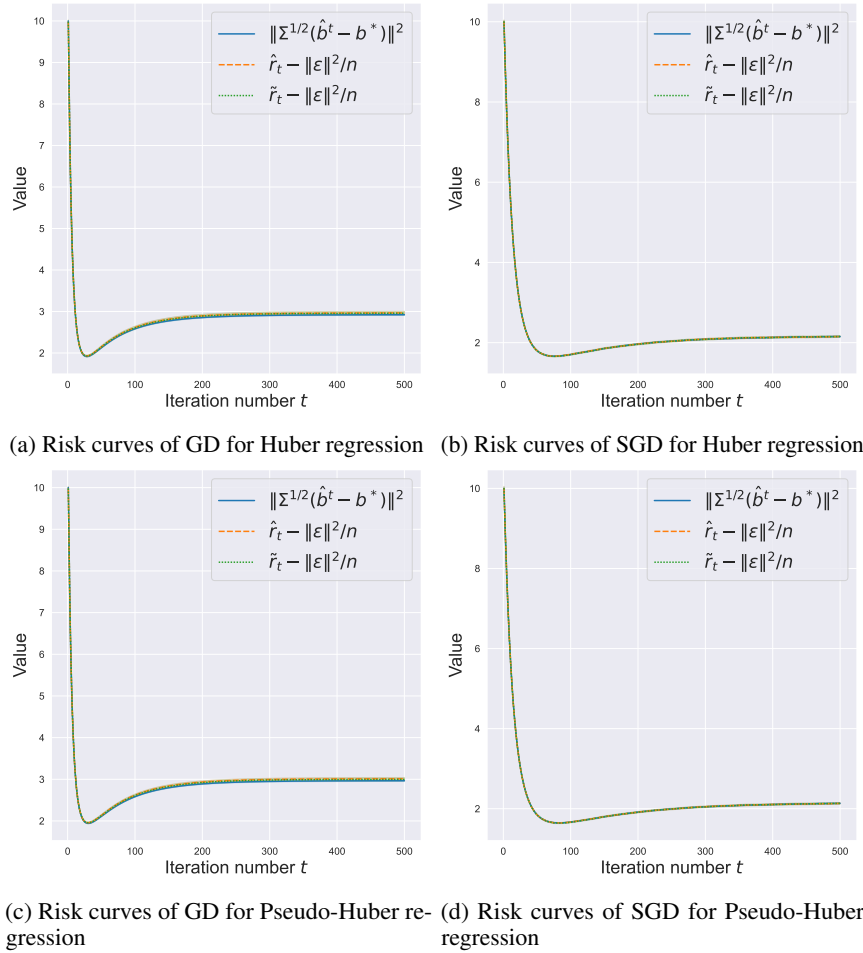


Figure 1: Risk curves for Huber and Pseudo-Huber regression with GD and SGD algorithms for the scenario  $(n, p) = (10000, 5000)$ . **Upper row:** Huber regression, **Lower row:** Pseudo-Huber regression. **Left column:** GD, **Right column:** SGD.

and  $\eta_t = 0.5$  if  $t$  is even. While the above choices of step size are not preferred in practice, here the goal is show that the proposed risk estimates is able to accurately capture the dynamics of the risk even when the step size changes along the trajectory of the algorithm. For instance, the first choice of step size should produce a risk curve that is flat when  $t$  is even. The results are presented in Figure 3, illustrating that the risk estimates accurately capture the flat segments of the true risk curve.

**Additional experiments: the estimate  $\tilde{r}_t^{\text{sub}}$  is suboptimal.** We compare the performance of  $\tilde{r}_t^{\text{sub}}$  with our proposed estimates in Huber regression with  $(n, p, T) = (4000, 1000, 20)$  and batch size  $|I_t| = n/10$  and  $\eta_t = 0.2$  for all  $t \in [T]$ . It is clear from Figure 4 that  $\tilde{r}_t$  is more accurate than the suboptimal estimator  $\tilde{r}_t^{\text{sub}}$ , especially when  $t$  increases.

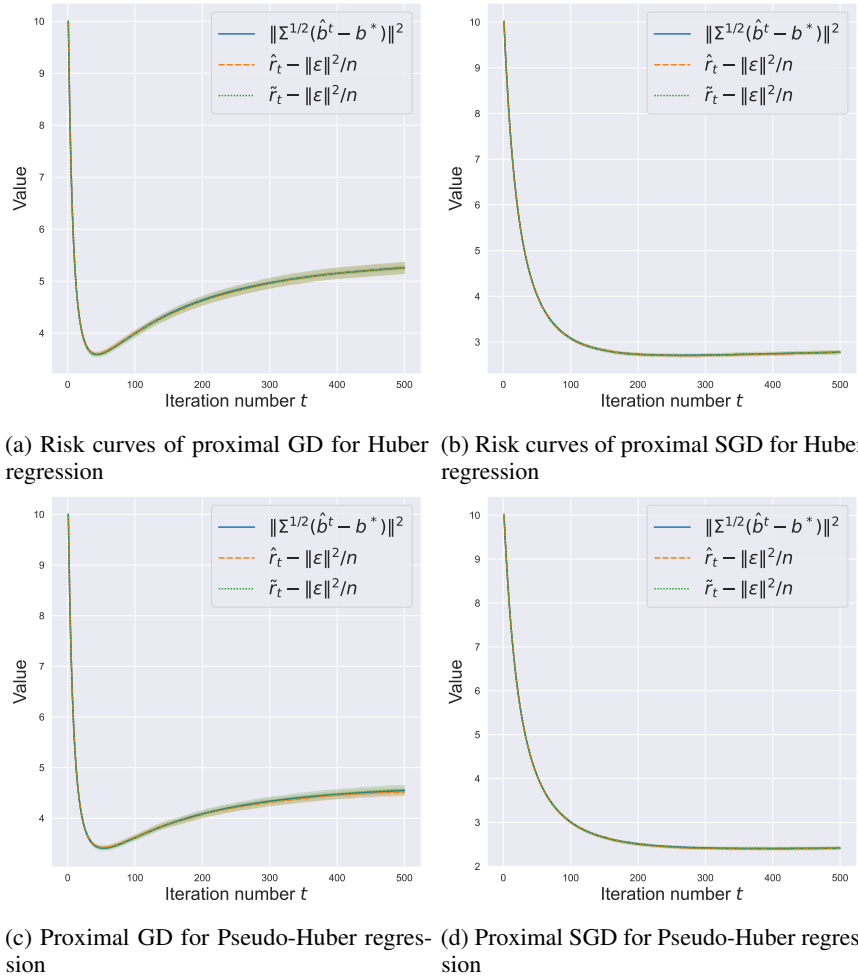


Figure 2: Risk curves for L1-penalized Huber and Pseudo-Huber regression with Proximal GD and Proximal SGD algorithms for the scenario  $(n, p) = (10000, 12000)$ . **Upper row:** L1-penalized Huber regression, **Lower row:** L1-penalized Pseudo-Huber regression. **Left column:** Proximal GD, **Right column:** Proximal SGD.

## 5 Discussion

This paper proposes a novel risk estimate for the generalization error of iterates generated by the proximal GD and proximal SGD algorithms in robust regression. The proposed risk estimates accurately capture the predictive risk of the iterates along the trajectory of the algorithms, and are provably consistent (Theorems 3.6 and 3.7). Three matrices in  $\mathbb{R}^{T \times T}$  in (8)-(10) reveal the interplay between the squared risk, the residuals and the gradients, so that the approximation (3) holds. This structure is different from the square loss case studied in [5] where only two matrices (inverse of each other) are sufficient.

Let us mention some open questions along with potential future research directions. The first question regards the probabilistic model: we currently assume Gaussian features  $x_i$ , and it would be of interest to study the extension in which our consistency results are universal, allowing non-Gaussian feature distributions. Second, it is of interest to extend the current estimates to more general optimization problems of the form (2) with non-smooth data-fitting loss, for instance the Least Absolute Deviation loss  $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_1$ . In this case the gradient does not exist at the origin, which calls for different algorithms than the GD and SGD variants presented here, for instance the Alternating Direction Method of Multipliers (ADMM) [6]. It is of independent interest to derive the risk estimates for iterates obtained by such primal-dual methods.

## Acknowledgments

P. C. Bellec acknowledges partial support from the NSF Grants DMS-1945428 and DMS-2413679. The authors acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here. The authors thank the anonymous reviewers for their valuable comments and suggestions that helped improve the presentation of the paper.

## References

- [1] Arnab Auddy, Haolin Zou, Kamiar Rahnamarad, and Arian Maleki. Approximate leave-one-out cross validation for regression with  $l_1$  regularizers. In *International Conference on Artificial Intelligence and Statistics*, pages 2377–2385. PMLR, 2024.
- [2] Mohsen Bayati, Murat A Erdogdu, and Andrea Montanari. Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- [3] Pierre C Bellec. Out-of-sample error estimate for robust m-estimators with convex penalty. *Information and Inference: A Journal of the IMA*, 12(4):2782–2817, 10 2023.
- [4] Pierre C Bellec and Yiwei Shen. Derivatives and residual distribution of regularized m-estimators with application to adaptive tuning. In *Conference on Learning Theory*, pages 1912–1947. PMLR, 2022.
- [5] Pierre C Bellec and Kai Tan. Uncertainty quantification for iterative algorithms in linear models with application to early stopping. *arXiv preprint arXiv:2404.17856*, 2024.
- [6] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.
- [7] Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In *Conference on Learning Theory*, pages 1078–1141. PMLR, 2020.
- [8] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.
- [9] Michael Celentano, Andrea Montanari, and Yuting Wei. The Lasso with general Gaussian designs with applications to hypothesis testing. *Ann. Statist.*, 51(5):2194–2220, 2023. ISSN 0090-5364.
- [10] Kabir Aladin Chandrasekher, Ashwin Pananjady, and Christos Thrampoulidis. Sharp global convergence guarantees for iterative nonconvex optimization: A gaussian process perspective. *arXiv preprint arXiv:2109.09859*, 2021.
- [11] Kenneth R Davidson and Stanislaw J Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.
- [12] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001. ISSN 0162-1459.
- [13] Cedric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborova. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM J. Math. Data Sci.*, 6(2):400–427, 2024.
- [14] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.
- [15] Gen Li and Yuting Wei. A non-asymptotic distributional theory of approximate message passing for sparse and robust regression. *arXiv preprint arXiv:2401.03923*, 2024.

- [16] Mengqi Lou, Kabir Aladin Verchand, and Ashwin Pananjady. Hyperparameter tuning via trajectory predictions: Stochastic prox-linear methods in matrix sensing. *arXiv preprint arXiv:2402.01599*, 2024.
- [17] Yuetian Luo, Zhimei Ren, and Rina Barber. Iterative approximate cross-validation. In *International Conference on Machine Learning*, pages 23083–23102. PMLR, 2023.
- [18] Léo Miolane and Andrea Montanari. The distribution of the Lasso: uniform control over sparse balls and adaptive parameter tuning. *Ann. Statist.*, 49(4):2313–2335, 2021. ISSN 0090-5364.
- [19] Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. Sgd in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Conference on Learning Theory*, pages 3548–3626. PMLR, 2021.
- [20] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.
- [21] Pratik Patil, Yuchen Wu, and Ryan Tibshirani. Failures and successes of cross-validation for early-stopped gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 2260–2268. PMLR, 2024.
- [22] Iosif Pinelis. Large deviations: Growth of empirical average of iid non-negative random variables with infinite expectations?, 2021. URL <https://mathoverflow.net/q/390939>. Accessed: 2024-05-21.
- [23] Kamiar Rahnama Rad and Arian Maleki. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 82(4):965–996, 2020. ISSN 1369-7412.
- [24] Cynthia Rush. An asymptotic rate for the lasso loss. In *International Conference on Artificial Intelligence and Statistics*, pages 3664–3673. PMLR, 2020.
- [25] Cynthia Rush and Ramji Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Trans. Inform. Theory*, 64(11):7264–7286, 2018.
- [26] Kai Tan, Gabriel Romon, and Pierre C. Bellec. Noise covariance estimation in multi-task high-dimensional linear models. *Bernoulli*, 30(3):1695 – 1722, 2024. doi: 10.3150/23-BEJ1644.
- [27] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.
- [28] Grace Wahba. A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.*, pages 1378–1402, 1985.
- [29] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006. ISSN 1369-7412.
- [30] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010. ISSN 0090-5364.

# Supplementary Material of “Estimating Generalization Performance Along the Trajectory of Proximal SGD in Robust Regression”

## A Additional simulation results

The following figures illustrate the proposed risk estimates accurately estimate the trajectory of the risk even when the step size changes at every step.

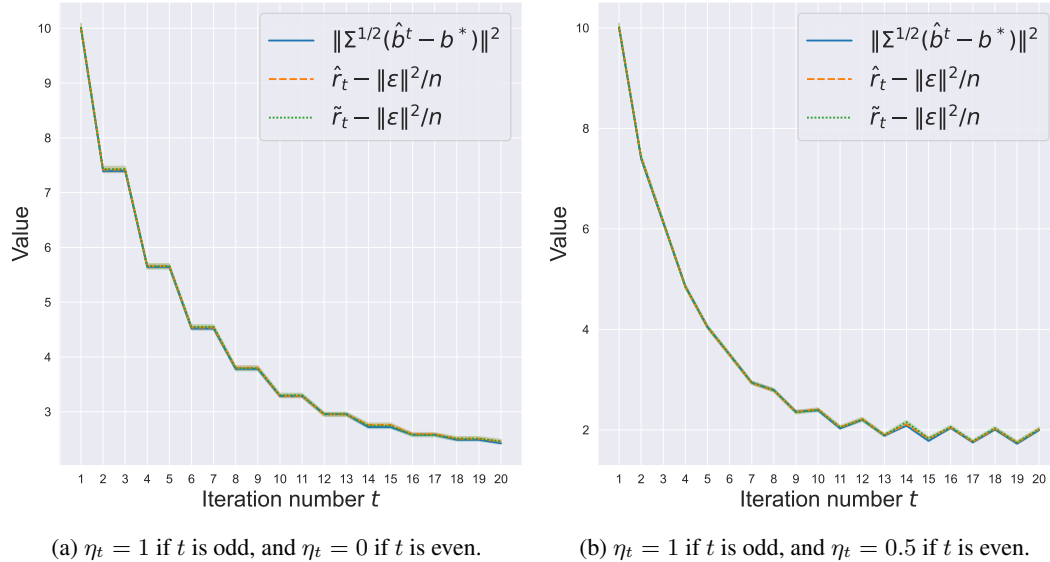


Figure 3: Risk curves for SGD applied to Huber regression with  $(n, p) = (3000, 1000)$  using different choices of step sizes. **Left panel:**  $\eta_t = 1$  if  $t$  is odd, and  $\eta_t = 0$  if  $t$  is even. **Right panel:**  $\eta_t = 1$  if  $t$  is odd, and  $\eta_t = 0.5$  if  $t$  is even.

Figure 4 compares the performance of the proposed estimators  $\hat{r}_t$ ,  $\tilde{r}_t$  and the estimator  $\tilde{r}_t^{\text{sub}}$  generalized directly from [5]. It confirms that our proposed estimators outperforms  $\tilde{r}_t^{\text{sub}}$ .

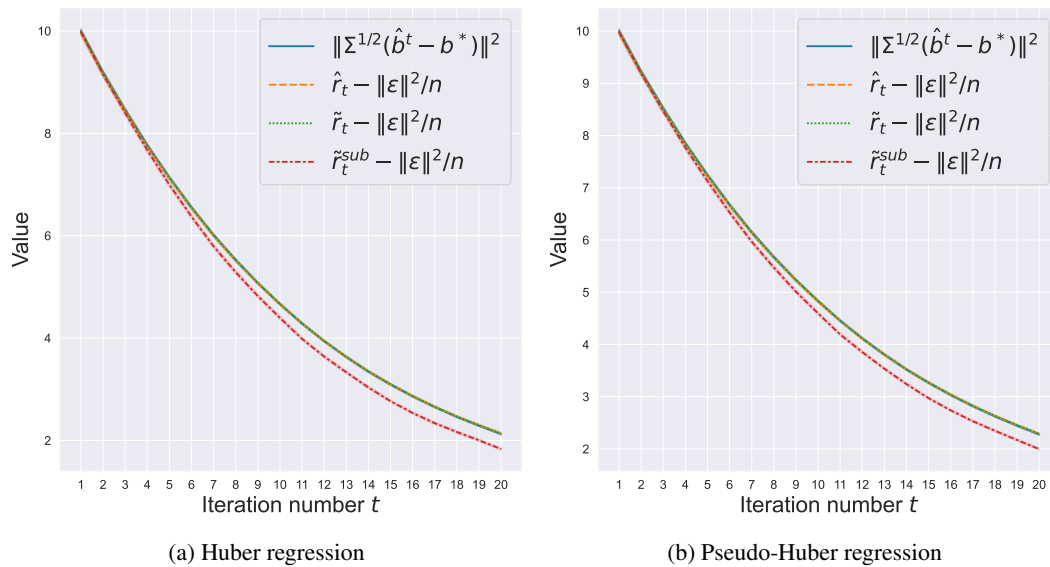


Figure 4: Risk curves for SGD applied to Huber and pseudo-Huber regression with  $(n, p, T) = (4000, 1000, 20)$ ,  $|I_t| = n/10$  and  $\eta_t = 0.2$  for all  $t$ .

## B Auxiliary Results

Throughout, we define

$$\mathbf{E} = [\varepsilon, \dots, \varepsilon] \in \mathbb{R}^{n \times T}, \quad \mathbf{F} = [S_1 \psi(\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^1), \dots, S_T \psi(\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^T)] \in \mathbb{R}^{n \times T}, \quad (13)$$

$$\mathbf{H} = \Sigma^{1/2} [\hat{\mathbf{b}}^1 - \mathbf{b}^*, \dots, \hat{\mathbf{b}}^T - \mathbf{b}^*] \in \mathbb{R}^{p \times T}, \quad \mathbf{R} = [\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^1, \dots, \mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^T] \in \mathbb{R}^{n \times T}. \quad (14)$$

Note that in the above  $\mathbf{E}, \mathbf{H}$  are not observable since  $\varepsilon$  and  $\mathbf{b}^*$  are unknown. However,  $\mathbf{F}$  and  $\mathbf{R}$  are observable and can be easily computed once the iterates  $(\hat{\mathbf{b}}^t)_{t \in [T]}$  are calculated.

### B.1 Change of variables

In this section, we conduct the change of variable to simplify the proof. Specifically, we view the linear model  $\mathbf{y} = \mathbf{X} \mathbf{b}^* + \varepsilon$  as a model with design matrix  $\mathbf{G}$  and the regression vector  $\boldsymbol{\theta}^*$ , i.e.

$$\mathbf{y} = \mathbf{X} \mathbf{b}^* + \varepsilon = \underbrace{\mathbf{X} \Sigma^{-1/2}}_{\mathbf{G}} \underbrace{\Sigma^{1/2} \mathbf{b}^*}_{\boldsymbol{\theta}^*} + \varepsilon.$$

This way, the design matrix  $\mathbf{G}$  has i.i.d. entries from standard normal distribution. Using the same argument in [5, Appendix D], we can show that the matrices  $\mathbf{H}, \mathbf{F}, \hat{\mathbf{A}}, \hat{\mathbf{K}}$  remains the same under the change of variable. Therefore, we can prove the main results using the model with design matrix  $\mathbf{G}$  and the regression vector  $\boldsymbol{\theta}^*$ . In other words, we assume without loss of generality that the design matrix  $\mathbf{X}$  has i.i.d.  $N(0, 1)$ , or equivalently that the independent rows of  $\mathbf{X}$  are normally distributed with covariance  $\Sigma = \mathbf{I}_p$ . We prove the main results using  $\Sigma = \mathbf{I}_p$ , and the results for general  $\Sigma$  follow by this change of variable with the constant  $C(T, \gamma, \eta_{\max}, c_0, \delta)$  appearing in the bounds depending additionally on  $\kappa$  (the upper bound of the condition number of  $\Sigma$  from Assumption 3.1).

### B.2 Derivative formulae

In this section, we present derivative formulae that will be useful in later proofs. The following formulas differ from [5] due to the use of robust loss functions and the application of SGD with random batches at each iteration. The formulae are also significantly more complex than in the case of regularized M-estimators [3, 4].

**Lemma B.1** (Proved in Appendix D.1). *Let  $(\hat{\mathbf{b}}^t)_{t \in [T]}$  be the iterates generated from the recursion (4) and the initial value  $\hat{\mathbf{b}}^1$  is independent of  $\mathbf{X}$ . Then the derivative of  $\hat{\mathbf{b}}^t$  with respect to  $\mathbf{X}$  is given by*

$$\frac{\partial \hat{\mathbf{b}}^t}{\partial x_{ij}} = (\mathbf{e}_t^\top \otimes \mathbf{I}_p) \boldsymbol{\Gamma} \left[ ((\mathbf{F}^\top \mathbf{e}_i) \otimes \mathbf{e}_j) - (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{SD}((\mathbf{H}^\top \mathbf{e}_j) \otimes \mathbf{e}_i) \right], \quad (15)$$

where  $\boldsymbol{\Gamma} = \mathcal{M}^{-1} \mathbf{L} (\boldsymbol{\Lambda} \otimes \mathbf{I}_p) \tilde{\mathcal{D}}$ ,  $\mathbf{L} = \sum_{t=2}^T ((\mathbf{e}_t \mathbf{e}_{t-1}^\top) \otimes \mathbf{I}_p)$ ,  $\boldsymbol{\Lambda} = \sum_{t=1}^T \frac{\eta_t}{|I_t|} \mathbf{e}_t \mathbf{e}_t^\top$ , and

$$\mathcal{M} = \begin{bmatrix} \mathbf{I}_p & & & & \\ -\mathbf{P}_1 & \mathbf{I}_p & & & \\ & \ddots & \ddots & & \\ & & -\mathbf{P}_{T-1} & \mathbf{I}_p & \end{bmatrix} \quad \text{where} \quad \mathbf{P}_t = \tilde{\mathcal{D}}_t (\mathbf{I}_p - \frac{\eta_t}{|I_t|} \mathbf{X}^\top \mathbf{S}_t \mathbf{D}_t \mathbf{X}).$$

Recall  $\mathbf{F} = [S_1 \psi(\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^1), \dots, S_T \psi(\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^T)]$ , we have  $F_{lt} = \mathbf{e}_l^\top \mathbf{S}_t \psi(\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^t)$ . The following two corollaries are a direct consequence of Lemma B.1.

**Lemma B.2** (Proved in Appendix D.2). *Under the same conditions of Lemma B.1. Let  $F_{lt} = \mathbf{e}_l^\top \mathbf{F} \mathbf{e}_t = \mathbf{e}_l^\top \mathbf{S}_t \psi(\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^t)$ , we have*

$$\frac{\partial F_{lt}}{\partial x_{ij}} = D_{ij}^{lt} + \Delta_{ij}^{lt}, \quad (16)$$

where

$$\begin{aligned} D_{ij}^{lt} &= -\mathbf{e}_l^\top \mathbf{S}_t \mathbf{D}_t \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} \mathbf{e}_t + ((\mathbf{e}_j^\top \mathbf{H}) \otimes \mathbf{e}_i^\top) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \boldsymbol{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{SD}(\mathbf{e}_t \otimes \mathbf{e}_l), \\ \Delta_{ij}^{lt} &= -((\mathbf{e}_i^\top \mathbf{F}) \otimes \mathbf{e}_j^\top) \boldsymbol{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{SD}(\mathbf{e}_t \otimes \mathbf{e}_l). \end{aligned}$$

**Lemma B.3** (Proved in Appendix D.3). Let  $\tilde{\mathbf{F}} = [\psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^1), \dots, \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^T)]$  and  $\tilde{\mathbf{F}}_{l,t} = \mathbf{e}_l^\top \tilde{\mathbf{F}} \mathbf{e}_t$ . Under the same conditions of Lemma B.1. We have

$$\frac{\partial \tilde{\mathbf{F}}_{l,t}}{\partial x_{ij}} = \tilde{D}_{ij}^{lt} + \tilde{\Delta}_{ij}^{lt}, \quad (17)$$

where

$$\begin{aligned} \tilde{D}_{ij}^{lt} &= -\mathbf{e}_l^\top \mathbf{D}_t \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} \mathbf{e}_t + ((\mathbf{e}_j^\top \mathbf{H}) \otimes \mathbf{e}_i^\top) \mathcal{D} \mathcal{S}(\mathbf{I}_T \otimes \mathbf{X}) \Gamma^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{D}(\mathbf{e}_t \otimes \mathbf{e}_l), \\ \tilde{\Delta}_{ij}^{lt} &= -((\mathbf{e}_i^\top \mathbf{F}) \otimes \mathbf{e}_j^\top) \Gamma^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{D}(\mathbf{e}_t \otimes \mathbf{e}_l). \end{aligned}$$

**Definition B.4.** Define the matrices  $\Upsilon_1 \in \mathbb{R}^{p \times T}$ ,  $\Upsilon_2 \in \mathbb{R}^{n \times T}$ ,  $\Upsilon_3 \in \mathbb{R}^{T \times T}$ ,  $\Upsilon_4 \in \mathbb{R}^{T \times T}$ ,  $\Upsilon_5 \in \mathbb{R}^{T \times T}$  by the identities

$$\forall j \in [p], \quad \sum_{i=1}^n \frac{\partial \mathbf{e}_i^\top \mathbf{F}}{\partial x_{ij}} = -\mathbf{e}_j^\top \mathbf{H} \tilde{\mathbf{K}}^\top - \mathbf{e}_j^\top \Upsilon_1, \quad (18)$$

$$\forall i \in [n], \quad \sum_{j=1}^p \frac{\partial \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} = \mathbf{e}_i^\top \mathbf{F} \mathbf{W}^\top - \mathbf{e}_i^\top \Upsilon_2, \quad (19)$$

$$\sum_{i=1}^n \sum_{j=1}^p \frac{\partial \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} = -\tilde{\mathbf{K}} \mathbf{H}^\top \mathbf{H} + \mathbf{F}^\top \mathbf{F} \mathbf{W}^\top - \Upsilon_3, \quad (20)$$

$$\sum_{i=1}^n \sum_{j=1}^p \frac{\partial \mathbf{H}^\top \mathbf{X}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} = (n \mathbf{I}_T - \tilde{\mathbf{A}}) \mathbf{H}^\top \mathbf{H} + \mathbf{H}^\top \mathbf{X}^\top \mathbf{F} \mathbf{W}^\top - \Upsilon_4, \quad (21)$$

$$\sum_{i=1}^n \sum_{j=1}^p \frac{\partial \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \tilde{\mathbf{F}}}{\partial x_{ij}} = -\tilde{\mathbf{K}} \mathbf{H}^\top \mathbf{X}^\top \tilde{\mathbf{F}} + p \mathbf{F}^\top \tilde{\mathbf{F}} - \mathbf{F}^\top \mathbf{F} \hat{\mathbf{A}}^\top - \Upsilon_5, \quad (22)$$

where the matrices  $\tilde{\mathbf{K}}, \hat{\mathbf{A}}, \tilde{\mathbf{A}}, \mathbf{W}$  are defined as follows

$$\tilde{\mathbf{A}} = \sum_{i=1}^n (\mathbf{I}_T \otimes \mathbf{e}_i^\top) (\mathbf{I}_T \otimes \mathbf{X}) \Gamma (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{D}(\mathbf{I}_T \otimes \mathbf{e}_i), \quad (23)$$

$$\hat{\mathbf{A}} = \sum_{i=1}^n (\mathbf{I}_T \otimes \mathbf{e}_i^\top) \mathcal{D}(\mathbf{I}_T \otimes \mathbf{X}) \Gamma (\mathbf{I}_T \otimes \mathbf{X}^\top) (\mathbf{I}_T \otimes \mathbf{e}_i), \quad (24)$$

$$\mathbf{W} = \sum_{j=1}^p (\mathbf{I}_T \otimes \mathbf{e}_j^\top) \Gamma (\mathbf{I}_T \otimes \mathbf{e}_j), \quad (25)$$

$$\tilde{\mathbf{K}} = \sum_{t=1}^T \text{Tr}(\mathbf{S}_t \mathbf{D}_t) \mathbf{e}_t \mathbf{e}_t^\top - \sum_{i=1}^n (\mathbf{I}_T \otimes \mathbf{e}_i^\top) \mathcal{D}(\mathbf{I}_T \otimes \mathbf{X}) \Gamma (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{D}(\mathbf{I}_T \otimes \mathbf{e}_i). \quad (26)$$

The matrices  $\Upsilon_1, \Upsilon_2, \dots$  are negligible in the sense that their Frobenius norms are of smaller orders compared to their preceding terms in (18)–(22). We provide the bounds in next lemma, which is obtained by deriving alternative expressions for  $\Upsilon_1, \dots, \Upsilon_5$  in Appendix D.4.

**Lemma B.5** (Proved in Appendix D.6). Under the same conditions of Theorem 3.6 with  $\Sigma = \mathbf{I}_p$ , we have

$$\begin{aligned} \max_{k \in \{1,3,4\}} \mathbb{E}[\|\Upsilon_k\|_{\text{op}}^2 \mid \varepsilon] &\leq C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2), \\ \mathbb{E}[\|\Upsilon_2\|_{\text{op}}^2 \mid \varepsilon] &\leq n^{-1} C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2), \\ \mathbb{E}[\|\Upsilon_5\|_{\text{op}}^2 \mid \varepsilon] &\leq n^2 C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2). \end{aligned}$$

We further define a few matrices of size  $T \times T$ :

$$\begin{aligned}
\Theta_1 &= \mathbf{F}^\top \mathbf{X} \mathbf{H} + \widetilde{\mathbf{K}} \mathbf{H}^\top \mathbf{H} - \mathbf{F}^\top \mathbf{F} \mathbf{W}^\top, \\
\Theta_2 &= n^{-1} [\mathbf{F}^\top \mathbf{X} \mathbf{X}^\top \widetilde{\mathbf{F}} + \widetilde{\mathbf{K}} \mathbf{H}^\top \mathbf{X}^\top \widetilde{\mathbf{F}} - p \mathbf{F}^\top \widetilde{\mathbf{F}} + \mathbf{F}^\top \mathbf{F} \widehat{\mathbf{A}}^\top], \\
\Theta_3 &= \mathbf{H}^\top \mathbf{X}^\top \mathbf{X} \mathbf{H} - (n \mathbf{I}_T - \widetilde{\mathbf{A}}) \mathbf{H}^\top \mathbf{H} - \mathbf{H}^\top \mathbf{X}^\top \mathbf{F} \mathbf{W}^\top, \\
\Theta_4 &= \frac{p}{n} \mathbf{F}^\top \widetilde{\mathbf{F}} - \frac{1}{n} (\widetilde{\mathbf{K}} \mathbf{H}^\top + \mathbf{F}^\top \mathbf{X}) (\widetilde{\mathbf{K}} \mathbf{H}^\top + \widetilde{\mathbf{F}}^\top \mathbf{X})^\top, \\
\Theta_5 &= n \mathbf{H}^\top \mathbf{H} - (\mathbf{W} \mathbf{F}^\top - \mathbf{H}^\top \mathbf{X}^\top) (\mathbf{W} \mathbf{F}^\top - \mathbf{H}^\top \mathbf{X}^\top)^\top, \\
\Theta_6 &= \|\mathbf{E}\|_{\text{F}}^{-1} (\mathbf{E}^\top \mathbf{X} \mathbf{H} - \mathbf{E}^\top \mathbf{F} \mathbf{W}^\top).
\end{aligned} \tag{27}$$

The next lemma provides the moment bounds for the Frobenius norm of these matrices.

**Lemma B.6** (Proved in Appendix D.7). *Under the same conditions of Theorem 3.6, we have*

$$\max_{k \in \{1, 2, 3\}} \mathbb{E}[\|\Theta_k\|_{\text{F}}^2 \mid \varepsilon] \leq nC(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2), \tag{28}$$

$$\max_{k \in \{4, 5\}} \mathbb{E}[\|\Theta_k\|_{\text{F}} \mid \varepsilon] \leq n^{1/2} C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2), \tag{29}$$

$$\mathbb{E}[\|\Theta_6\|_{\text{F}}^2 \mid \varepsilon] \leq C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2), \tag{30}$$

almost surely, where  $\mathbb{E}[\cdot \mid \varepsilon]$  is the conditional expectation given  $\varepsilon$ .

We are able to prove the main theorems using Lemma B.6.

## C Proof of main results

### C.1 Proof of Theorem 3.6

It suffices to prove this theorem for the case  $\Sigma = \mathbf{I}_p$ . When  $\Sigma \neq \mathbf{I}_p$ , the result can be derived using a change of variables argument, as outlined in Appendix B.1. By basic algebra, we have

$$\begin{aligned}
&\Theta_5 + \|\mathbf{E}\|_{\text{F}} (\Theta_6 + \Theta_6^\top) \\
&= n \mathbf{H}^\top \mathbf{H} - (\mathbf{X} \mathbf{H} - \mathbf{F} \mathbf{W}^\top)^\top (\mathbf{X} \mathbf{H} - \mathbf{F} \mathbf{W}^\top) + \mathbf{E}^\top (\mathbf{X} \mathbf{H} - \mathbf{F} \mathbf{W}^\top) + (\mathbf{X} \mathbf{H} - \mathbf{F} \mathbf{W}^\top)^\top \mathbf{E} \\
&= n \mathbf{H}^\top \mathbf{H} + \mathbf{E}^\top \mathbf{E} - (\mathbf{E} - \mathbf{X} \mathbf{H} + \mathbf{F} \mathbf{W}^\top)^\top (\mathbf{E} - \mathbf{X} \mathbf{H} + \mathbf{F} \mathbf{W}^\top) \\
&= n \mathbf{H}^\top \mathbf{H} + \mathbf{E}^\top \mathbf{E} - (\mathbf{R} + \mathbf{F} \mathbf{W}^\top)^\top (\mathbf{R} + \mathbf{F} \mathbf{W}^\top).
\end{aligned}$$

Notice that  $r_t = \|\widehat{\mathbf{b}}^t - \mathbf{b}^*\|^2 + \|\varepsilon\|^2/n$  is the  $t$ -th diagonal entry of  $(\mathbf{H}^\top \mathbf{H} + \mathbf{E}^\top \mathbf{E}/n)$ , and  $\hat{r}_t$  is the  $t$ -th diagonal entry of  $(\mathbf{R} + \mathbf{F} \mathbf{W}^\top)^\top (\mathbf{R} + \mathbf{F} \mathbf{W}^\top)/n$ . Since  $\|\mathbf{E}\|_{\text{F}} = \sqrt{T} \|\varepsilon\|$ , using the previous display that conditionally on  $\varepsilon$ , we have

$$\mathbb{E}[\hat{r}_t - r_t \mid \varepsilon] \leq n^{-1} \mathbb{E}[\|\Theta_5\|_{\text{F}} + 2\|\mathbf{E}\|_{\text{F}} \|\Theta_6\|_{\text{F}} \mid \varepsilon] = n^{-1} \mathbb{E}[\|\Theta_5\|_{\text{F}} + 2\sqrt{T} \|\varepsilon\| \|\Theta_6\|_{\text{F}} \mid \varepsilon].$$

Using the moment bounds of  $\Theta_5$  and  $\Theta_6$  in Lemma B.6, we have

$$\mathbb{E}[\hat{r}_t - r_t \mid \varepsilon] \leq \frac{C(T, \gamma, \eta_{\max}, c_0, \delta)}{\sqrt{n}} \left(1 + \frac{\|\varepsilon\|}{\sqrt{n}}\right).$$

Furthermore, if  $\mathbb{E}[\|\varepsilon_i\|]$  is finite, we have by [22] that  $\|\varepsilon\|/n \xrightarrow{P} 0$  (convergence in probability) if  $\varepsilon$  has i.i.d. entries with a fixed distribution independent of  $n$ . Under this assumption, the right-hand side of the previous display converges to 0 in probability. By enlarging the constant if necessary, assume  $C(T, \gamma, \eta_{\max}, c_0, \delta) \geq 1$ . To obtain a quantitative bound, by the conditional version of Markov's inequality, for any  $\epsilon > 0$ , and almost surely with respect to  $\varepsilon$  that

$$\begin{aligned}
\mathbb{P}(\hat{r}_t - r_t > \epsilon \mid \varepsilon) &\leq \min\left\{1, \frac{C(T, \gamma, \eta_{\max}, c_0, \delta)}{\epsilon} \left(\frac{1}{\sqrt{n}} + \frac{\|\varepsilon\|}{n}\right)\right\} \\
&\leq \max\left\{1, \frac{C(T, \gamma, \eta_{\max}, c_0, \delta)}{\epsilon}\right\} \min\left\{1, \frac{1}{\sqrt{n}} + \frac{\|\varepsilon\|}{n}\right\}.
\end{aligned}$$



Taking expectation with respect to  $\varepsilon$ , we obtain

$$\begin{aligned}\mathbb{P}\left(|\hat{r}_t - r_t| > \epsilon\right) &\leq \max\left\{1, \frac{C(T, \gamma, \eta_{\max}, c_0, \delta)}{\epsilon}\right\} \mathbb{E}\left[\min\left\{1, \frac{1}{\sqrt{n}} + \frac{\|\varepsilon\|}{n}\right\}\right] \\ &\leq \max\left\{1, \frac{C(T, \gamma, \eta_{\max}, c_0, \delta)}{\epsilon}\right\} \left(\frac{1}{\sqrt{n}} + \mathbb{E}\left[\min\left\{1, \frac{\|\varepsilon\|}{n}\right\}\right]\right)\end{aligned}$$

with  $\mathbb{E}[\min\{1, \frac{\|\varepsilon\|}{n}\}] \rightarrow 0$  (equivalently,  $\|\varepsilon\|/n \rightarrow^P 0$ ) if the entries of  $\varepsilon$  are i.i.d. with a fixed distribution independent of  $n$  with  $\mathbb{E}[|\varepsilon_i|] < +\infty$  by [22]. This finishes the proof of Theorem 3.6.

## C.2 Operator norm bound on $\widehat{K}$

We first recall the definition of  $\widehat{K}$  from (10) in the main text:

$$\widehat{K} = \sum_{t=1}^T \text{Tr}(\mathbf{D}_t) \mathbf{e}_t \mathbf{e}_t^\top - \sum_{i=1}^n (\mathbf{I}_T \otimes \mathbf{e}_i^\top) \mathcal{D}(\mathbf{I}_T \otimes \mathbf{X}) \Gamma(\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{SD}(\mathbf{I}_T \otimes \mathbf{e}_i). \quad (31)$$

Define two events:  $\Omega_1 = \{\mathbf{X} \in \mathbb{R}^{n \times p} : \|\mathbf{X}\|_{\text{op}}/\sqrt{n} \leq 2 + \sqrt{p/n}\}$  and  $\Omega_2 = \{|\{i \in [n] : |\varepsilon_i| \leq M\}| \geq \frac{2n}{3}\}$ , where  $M$  is a large enough constant such that  $\mathbb{P}(|\varepsilon_i| > M) \leq 1/6$ .

**Lemma C.1.** *Under the same conditions of Theorem 3.6 with  $\Sigma = \mathbf{I}_p$ , we have in the event  $\Omega_* = \Omega_1 \cap \Omega_2$  that*

$$n\|\widehat{K}^{-1}\|_{\text{op}} \leq C(T, \gamma, \eta_{\max}, c_0, \delta, \|\mathbf{b}^*\|).$$

Furthermore,  $\Omega_*$  has probability at least  $1 - e^{-n/18} - e^{-n/2}$ .

*Proof of Lemma C.1.* Under Assumptions 3.1 and 3.5, we know that  $\mathbb{P}(\Omega_1) \geq 1 - e^{-n/2}$  from [11]. In the event  $\Omega_1$ , we have by Lemma D.2 that

$$\|\mathbf{X}\mathbf{H}\|_{\text{F}}^2/n \leq C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2) := C_*.$$

Markov's inequality further implies

$$|\{i \in [n] : \|\mathbf{x}_i^\top \mathbf{H}\|^2 > 3C_*\}| \leq n/3.$$

In other words,  $|\{i \in [n] : \|\mathbf{x}_i^\top \mathbf{H}\|^2 \leq 3C_*\}| \geq \frac{2n}{3}$ . Recall that  $M$  is such that  $\mathbb{P}(|\varepsilon_i| > M) \leq 1/6$ . By Hoeffding's inequality, we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{|\varepsilon_i| > M\} \geq \mathbb{P}(|\varepsilon_i| > M) + a\right) \leq e^{-2na^2}.$$

Taking  $a = 1/6$ , we have  $\mathbb{P}(\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{|\varepsilon_i| > M\} \geq \mathbb{P}(|\varepsilon_i| > M) + 1/6) \leq e^{-n/18}$ . Furthermore, using  $|\{i \in [n] : |\varepsilon_i| > M\}| = \sum_{i=1}^n \mathbf{1}\{|\varepsilon_i| > M\}$ , we have

$$\begin{aligned}\left\{|\{i \in [n] : |\varepsilon_i| > M\}| \geq n/3\right\} &= \left\{\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{|\varepsilon_i| > M\} \geq 1/6 + 1/6\right\} \\ &\subseteq \left\{\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{|\varepsilon_i| > M\} \geq \mathbb{P}(|\varepsilon_i| > M) + 1/6\right\}.\end{aligned}$$

Therefore,

$$\mathbb{P}\left(|\{i \in [n] : |\varepsilon_i| > M\}| \geq n/3\right) \leq e^{-n/18}.$$

Equivalently, we have  $\mathbb{P}(|\{i \in [n] : |\varepsilon_i| \leq M\}| \geq \frac{2n}{3}) \geq 1 - e^{-n/18}$ . That is, at least  $\frac{2n}{3}$  of the entries of  $\varepsilon$  are bounded by  $M$  with probability at least  $1 - e^{-n/18}$ .

Recall that  $\Omega_2 = \{|\{i \in [n] : |\varepsilon_i| \leq M\}| \geq \frac{2n}{3}\}$ , then  $\mathbb{P}(\Omega_2) \geq 1 - e^{-n/18}$ . Hence,  $\mathbb{P}(\Omega_1 \cap \Omega_2) \geq 1 - e^{-n/18} - e^{-n/2}$ . In the event  $\Omega_1 \cap \Omega_2$ , the set

$$\hat{I} = \{i \in [n] : |\varepsilon_i| \leq M, \|\mathbf{x}_i^\top \mathbf{H}\|^2 \leq 3C_*\}$$

has size at least  $\frac{n}{3}$ . For any  $i \in \hat{I}$  and  $t \in [T]$ , we have

$$|y_i - \mathbf{x}_i^\top \hat{\mathbf{b}}^t| = |\boldsymbol{\varepsilon}_i - \mathbf{x}_i^\top \mathbf{H} \mathbf{e}_t| \leq |\boldsymbol{\varepsilon}_i| + |\mathbf{x}_i^\top \mathbf{H} \mathbf{e}_t| \leq M + \sqrt{3C_*}. \quad (32)$$

By the definition of  $\mathbf{D}_t$ , under Assumption 3.4, we have

$$\begin{aligned} \text{Tr}(\mathbf{D}_t) &= \sum_{i=1}^n \rho''(y_i - \mathbf{x}_i^\top \hat{\mathbf{b}}^t) \\ &> \sum_{i \in \hat{I}} \rho''(y_i - \mathbf{x}_i^\top \hat{\mathbf{b}}^t) && \text{since } \rho'' \geq 0 \text{ by convexity} \\ &\geq |\hat{I}| \min_{u: |u| \leq M + \sqrt{3C_*}} \rho''(u) && \text{due to (32)} \\ &\geq n/3 \min_{u: |u| \leq M + \sqrt{3C_*}} \rho''(u) := c_* n && \text{since } \hat{I} \text{ has size at least } n/3. \end{aligned}$$

Here  $c_*$  is a constant depending on  $\rho, M, C_*$  only.

By the definition of  $\widehat{\mathbf{K}}$  in (31),  $\widehat{\mathbf{K}}/n$  is a lower triangular matrix with diagonal entries equal to  $\text{Tr}(\mathbf{D}_t)/n$ . It is invertible if and only if all its diagonal entries are non-zero. Therefore, in the event  $\Omega_1 \cap \Omega_2$ , we have  $\widehat{\mathbf{K}}/n$  is invertible.

Let  $\widehat{\mathbf{\Lambda}} = \sum_{t=1}^T \text{Tr}(\mathbf{D}_t) \mathbf{e}_t \mathbf{e}_t^\top$ . Then it is diagonal,  $\|\widehat{\mathbf{\Lambda}}^{-1}\|_{\text{op}} = \max_{t \in [T]} \text{Tr}[\mathbf{D}_t]^{-1} \leq (c_* n)^{-1}$  and

$$\widehat{\mathbf{K}} = \widehat{\mathbf{\Lambda}} - \widehat{\mathbf{L}} \quad (33)$$

where  $\widehat{\mathbf{L}} = \sum_{i=1}^n (\mathbf{I}_T \otimes \mathbf{e}_i^\top) \mathcal{D}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma} (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{S} \mathcal{D}(\mathbf{I}_T \otimes \mathbf{e}_i)$ . Here  $\widehat{\mathbf{L}}$  is a strictly lower triangular matrix. Using the upper bound of  $\|\mathbf{\Gamma}\|_{\text{op}}$  in Lemma D.4, we have  $\|\widehat{\mathbf{L}}\|_{\text{op}} \leq nC(T, \gamma, \eta_{\max}, c_0)$  in the event  $\Omega_1$ . Now we rewrite  $\widehat{\mathbf{K}}^{-1}$  as

$$\widehat{\mathbf{K}}^{-1} = (\widehat{\mathbf{K}} \widehat{\mathbf{\Lambda}}^{-1} \widehat{\mathbf{\Lambda}})^{-1} = \widehat{\mathbf{\Lambda}}^{-1} (\widehat{\mathbf{K}} \widehat{\mathbf{\Lambda}}^{-1})^{-1} = \widehat{\mathbf{\Lambda}}^{-1} (\mathbf{I}_T - \widehat{\mathbf{L}} \widehat{\mathbf{\Lambda}}^{-1})^{-1}.$$

Notice that  $\widehat{\mathbf{L}} \widehat{\mathbf{\Lambda}}^{-1} \in \mathbb{R}^{T \times T}$  is a strictly lower triangular matrix, thus

$$(\mathbf{I}_T - \widehat{\mathbf{L}} \widehat{\mathbf{\Lambda}}^{-1})^{-1} = \sum_{k=0}^{\infty} (\widehat{\mathbf{L}} \widehat{\mathbf{\Lambda}}^{-1})^k = \sum_{k=0}^{T-1} (\widehat{\mathbf{L}} \widehat{\mathbf{\Lambda}}^{-1})^k.$$

By the triangle inequality,

$$\|(\mathbf{I}_T - \widehat{\mathbf{L}} \widehat{\mathbf{\Lambda}}^{-1})^{-1}\|_{\text{op}} \leq \sum_{k=0}^{T-1} \|\widehat{\mathbf{L}} \widehat{\mathbf{\Lambda}}^{-1}\|_{\text{op}}^k \leq C(T, \gamma, \eta_{\max}, c_0, \delta, \|\mathbf{b}^*\|).$$

Therefore, in the event  $\Omega_1 \cap \Omega_2$  which has probability  $\mathbb{P}(\Omega_1 \cap \Omega_2) \geq 1 - e^{-n/18} - e^{-n/2}$ ,

$$\|\widehat{\mathbf{K}}^{-1}\|_{\text{op}} \leq \|\widehat{\mathbf{\Lambda}}^{-1}\|_{\text{op}} \|(\mathbf{I}_T - \widehat{\mathbf{L}} \widehat{\mathbf{\Lambda}}^{-1})^{-1}\|_{\text{op}} \leq n^{-1} C(T, \gamma, \eta_{\max}, c_0, \delta, \|\mathbf{b}^*\|).$$

□

**Lemma C.2.** Under the same conditions of Theorem 3.7 with  $\boldsymbol{\Sigma} = \mathbf{I}_p$ , we have

$$\|\widehat{\mathbf{K}}\|_{\text{op}} \leq n(1 + \|\mathbf{X}\|_{\text{op}}^2 \|\mathbf{\Gamma}\|_{\text{op}}), \quad \|\widehat{\mathbf{A}}\|_{\text{op}} \leq n\|\mathbf{X}\|_{\text{op}}^2 \|\mathbf{\Gamma}\|_{\text{op}}, \quad \|\mathbf{W}\|_{\text{op}} \leq n\|\mathbf{\Gamma}\|_{\text{op}}.$$

*Proof of Lemma C.2.* By the definition of  $\widehat{\mathbf{K}}$  in (31), using  $\|\mathcal{D}\|_{\text{op}} \leq 1$  and  $\|\mathcal{S}\|_{\text{op}} \leq 1$ , we have

$$\|\widehat{\mathbf{K}}\|_{\text{op}} \leq \|\widehat{\mathbf{\Lambda}}\|_{\text{op}} + \|\widehat{\mathbf{L}}\|_{\text{op}} \leq n(1 + \|\mathbf{X}\|_{\text{op}}^2 \|\mathbf{\Gamma}\|_{\text{op}}).$$

Similarly, by the definition of  $\widehat{\mathbf{A}}$  in (24), we have

$$\|\widehat{\mathbf{A}}\|_{\text{op}} \leq n\|\mathbf{X}\|_{\text{op}}^2 \|\mathbf{\Gamma}\|_{\text{op}}.$$

Last, by the definition of  $\mathbf{W}$  in (25), we have  $\|\mathbf{W}\|_{\text{op}} \leq n\|\mathbf{\Gamma}\|_{\text{op}}$ .

□

**Lemma C.3.** Under the same conditions of Theorem 3.6 with  $\Sigma = I_p$ , we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{F}^\top \mathbf{F}(\hat{\mathbf{A}} - \hat{\mathbf{K}}\mathbf{W})^\top\|_F | \varepsilon] &\leq n^{3/2}C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2), \\ \mathbb{E}[\|\mathbf{R}^\top \mathbf{F}(\hat{\mathbf{A}} - \hat{\mathbf{K}}\mathbf{W})^\top\|_F | \varepsilon] &\leq n^2(\frac{\|\varepsilon\|}{n} + \frac{1}{\sqrt{n}})C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2).\end{aligned}$$

*Proof of Lemma C.3.* First, using the definitions of  $\Theta_1, \Theta_2, \Theta_4$  in (27), we have

$$n^{-1}\mathbf{F}^\top \mathbf{F}(\hat{\mathbf{A}} - \hat{\mathbf{K}}\mathbf{W})^\top = n^{-1}\Theta_1\hat{\mathbf{K}}^\top + \Theta_2 + \Theta_4. \quad (34)$$

Hence,

$$\begin{aligned}&\mathbb{E}[\|\mathbf{F}^\top \mathbf{F}(\hat{\mathbf{A}} - \hat{\mathbf{K}}\mathbf{W})^\top\|_F | \varepsilon] \\&= \mathbb{E}[\|(\Theta_1\hat{\mathbf{K}}^\top + n\Theta_2 + n\Theta_4)\|_F | \varepsilon] \quad \text{by (34)} \\&\leq \mathbb{E}[\|\Theta_1\|_F^2 | \varepsilon]^{1/2}\mathbb{E}[\|\hat{\mathbf{K}}\|_{\text{op}}^2 | \varepsilon]^{1/2} + n\mathbb{E}[\|\Theta_2\|_F + \|\Theta_4\|_F | \varepsilon] \quad \text{by the Cauchy-Schwarz inequality} \\&\leq n^{3/2}C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2).\end{aligned}$$

Here the last line uses the upper bounds of  $\mathbb{E}[\|\Theta_k\|_F | \varepsilon]$  from Lemma B.6, and the bound of  $\mathbb{E}[\|\hat{\mathbf{K}}\|_{\text{op}} | \varepsilon]$  follows from Lemma C.2 and the bound of  $\|\Gamma\|_{\text{op}}$  from Lemma D.4. This proves the first inequality.

For the second inequality, we define

$$\begin{aligned}\check{\Theta}_1 &= \frac{\mathbf{R}^\top \mathbf{X}\mathbf{H} + \check{\mathbf{K}}\mathbf{H}^\top \mathbf{H} - \mathbf{R}^\top \mathbf{F}\mathbf{W}^\top}{\|\mathbf{E}\|_F/\sqrt{n} + 1}, \\ \check{\Theta}_2 &= \frac{\mathbf{R}^\top \mathbf{X}\mathbf{X}^\top \tilde{\mathbf{F}} + \check{\mathbf{K}}\mathbf{H}^\top \mathbf{X}^\top \tilde{\mathbf{F}} - p\mathbf{R}^\top \tilde{\mathbf{F}} + \mathbf{R}^\top \mathbf{F}\hat{\mathbf{A}}^\top}{n(\|\mathbf{E}\|_F/\sqrt{n} + 1)}, \\ \check{\Theta}_4 &= \frac{p\mathbf{R}^\top \tilde{\mathbf{F}} - (\check{\mathbf{K}}\mathbf{H}^\top + \mathbf{R}^\top \mathbf{X})(\hat{\mathbf{K}}\mathbf{H}^\top + \tilde{\mathbf{F}}^\top \mathbf{X})^\top}{n(\|\mathbf{E}\|_F/\sqrt{n} + 1)},\end{aligned}$$

where  $\check{\mathbf{K}} = n\mathbf{I}_T - \sum_{i=1}^n (\mathbf{I}_T \otimes (\mathbf{e}_i^\top \mathbf{X}))\Gamma(\mathbf{I}_T \otimes \mathbf{X}^\top)\mathcal{SD}(\mathbf{I}_T \otimes (\mathbf{X}^\top \mathbf{e}_i))$ . Using similar argument that proves Lemma B.6, we can obtain the following bound of  $\check{\Theta}_1, \check{\Theta}_2, \check{\Theta}_4$ .

$$\max_{k \in \{1,2\}} \mathbb{E}[\|\check{\Theta}_k\|_F^2 | \varepsilon] \leq nC(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2), \quad (35)$$

$$\mathbb{E}[\|\check{\Theta}_4\|_F | \varepsilon] \leq n^{1/2}C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2). \quad (36)$$

By the definitions of  $\check{\Theta}_1, \check{\Theta}_2, \check{\Theta}_4$ , we have

$$(\|\mathbf{E}\|_F/\sqrt{n} + 1)[n^{-1}\check{\Theta}_1\hat{\mathbf{K}}^\top + \check{\Theta}_2 + \check{\Theta}_4] = n^{-1}\mathbf{R}^\top \mathbf{F}(\hat{\mathbf{A}} - \hat{\mathbf{K}}\mathbf{W})^\top.$$

Therefore, conditional on  $\varepsilon$ , we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{R}^\top \mathbf{F}(\hat{\mathbf{A}} - \hat{\mathbf{K}}\mathbf{W})^\top\|_F | \varepsilon] &= (\|\mathbf{E}\|_F/\sqrt{n} + 1)\mathbb{E}[\|\check{\Theta}_1\hat{\mathbf{K}}^\top + n\check{\Theta}_2 + n\check{\Theta}_4\|_F | \varepsilon] \\&\leq n^{3/2}(\|\mathbf{E}\|_F/\sqrt{n} + 1)C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2).\end{aligned}$$

This finishes the proof of Lemma C.3.  $\square$

### C.3 Proof of Theorem 3.7

In the event  $\Omega_* = \Omega_1 \cap \Omega_2$ , we know from Lemma C.1 that  $\hat{\mathbf{K}}$  is invertible and  $\|\hat{\mathbf{K}}^{-1}\|_{\text{op}} \leq n^{-1}C$ . Define  $\tilde{\mathbf{W}} = \hat{\mathbf{K}}^{-1}\hat{\mathbf{A}}$ . Using  $\mathbf{R} + \mathbf{F}\tilde{\mathbf{W}}^\top = \mathbf{R} + \mathbf{F}\mathbf{W}^\top + \mathbf{F}(\tilde{\mathbf{W}} - \mathbf{W})^\top$ , we have

$$\begin{aligned}&(\mathbf{R} + \mathbf{F}\tilde{\mathbf{W}}^\top)^\top(\mathbf{R} + \mathbf{F}\tilde{\mathbf{W}}^\top) - (\mathbf{R} + \mathbf{F}\mathbf{W}^\top)^\top(\mathbf{R} + \mathbf{F}\mathbf{W}^\top) \\&= (\mathbf{R} + \mathbf{F}\mathbf{W}^\top)^\top \mathbf{F}(\tilde{\mathbf{W}} - \mathbf{W})^\top + (\tilde{\mathbf{W}} - \mathbf{W})\mathbf{F}^\top(\mathbf{R} + \mathbf{F}\mathbf{W}^\top) + (\tilde{\mathbf{W}} - \mathbf{W})\mathbf{F}^\top \mathbf{F}(\tilde{\mathbf{W}} - \mathbf{W})^\top.\end{aligned}$$

We have by the triangle inequality

$$\begin{aligned}&\|(\mathbf{R} + \mathbf{F}\tilde{\mathbf{W}}^\top)^\top(\mathbf{R} + \mathbf{F}\tilde{\mathbf{W}}^\top) - (\mathbf{R} + \mathbf{F}\mathbf{W}^\top)^\top(\mathbf{R} + \mathbf{F}\mathbf{W}^\top)\|_F \\&\leq 2\|(\mathbf{R} + \mathbf{F}\mathbf{W}^\top)^\top \mathbf{F}(\tilde{\mathbf{W}} - \mathbf{W})^\top\|_F + \|(\tilde{\mathbf{W}} - \mathbf{W})\mathbf{F}^\top \mathbf{F}(\tilde{\mathbf{W}} - \mathbf{W})^\top\|_F \\&\lesssim \|\mathbf{R}^\top \mathbf{F}(\tilde{\mathbf{W}} - \mathbf{W})^\top\|_F + \|\mathbf{W}\mathbf{F}^\top \mathbf{F}(\tilde{\mathbf{W}} - \mathbf{W})^\top\|_F + \|(\tilde{\mathbf{W}} - \mathbf{W})\mathbf{F}^\top \mathbf{F}(\tilde{\mathbf{W}} - \mathbf{W})^\top\|_F.\end{aligned}$$

Recall that in the event  $\Omega_*$ , we have  $\|\widehat{\mathbf{K}}^{-1}\|_{\text{op}} \leq n^{-1}C$ . Using  $\widehat{\mathbf{A}} - \widehat{\mathbf{K}}\mathbf{W} = \widehat{\mathbf{K}}(\widetilde{\mathbf{W}} - \mathbf{W})$  and Lemma C.3, we have

$$\begin{aligned} & \mathbb{E}\left[I(\Omega_*)\|(\mathbf{R} + \mathbf{F}\widetilde{\mathbf{W}}^\top)^\top(\mathbf{R} + \mathbf{F}\widetilde{\mathbf{W}}^\top) - (\mathbf{R} + \mathbf{F}\mathbf{W}^\top)^\top(\mathbf{R} + \mathbf{F}\mathbf{W}^\top)\|_{\text{F}} \mid \varepsilon\right] \\ & \lesssim \mathbb{E}\left[I(\Omega_*)\|\mathbf{R}^\top \mathbf{F}(\widetilde{\mathbf{W}} - \mathbf{W})^\top\|_{\text{F}} \mid \varepsilon\right] \\ & \quad + \mathbb{E}\left[I(\Omega_*)\|\mathbf{W}\mathbf{F}^\top \mathbf{F}(\widetilde{\mathbf{W}} - \mathbf{W})^\top\|_{\text{F}} \mid \varepsilon\right] \\ & \quad + \mathbb{E}\left[I(\Omega_*)\|(\widetilde{\mathbf{W}} - \mathbf{W})\mathbf{F}^\top \mathbf{F}(\widetilde{\mathbf{W}} - \mathbf{W})^\top\|_{\text{F}} \mid \varepsilon\right] \\ & = \mathbb{E}\left[I(\Omega_*)\|\mathbf{R}^\top \mathbf{F}(\widehat{\mathbf{A}} - \widehat{\mathbf{K}}\mathbf{W})^\top (\widehat{\mathbf{K}}^\top)^{-1}\|_{\text{F}} \mid \varepsilon\right] \\ & \quad + \mathbb{E}\left[I(\Omega_*)\|\mathbf{W}\mathbf{F}^\top \mathbf{F}(\widehat{\mathbf{A}} - \widehat{\mathbf{K}}\mathbf{W})^\top (\widehat{\mathbf{K}}^\top)^{-1}\|_{\text{F}} \mid \varepsilon\right] \\ & \quad + \mathbb{E}\left[I(\Omega_*)\|\widehat{\mathbf{K}}^{-1}(\widehat{\mathbf{A}} - \widehat{\mathbf{K}}\mathbf{W})\mathbf{F}^\top \mathbf{F}(\widehat{\mathbf{A}} - \widehat{\mathbf{K}}\mathbf{W})^\top (\widehat{\mathbf{K}}^\top)^{-1}\|_{\text{F}} \mid \varepsilon\right]. \end{aligned}$$

According to Lemma C.3 and the bound of  $\|\widehat{\mathbf{K}}^{-1}\|_{\text{op}}$  in Lemma C.1, the first conditional expectation is bounded from above by

$$n\left(\frac{\|\varepsilon\|}{n} + \frac{1}{\sqrt{n}}\right)C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2).$$

Using the bound of  $\|\mathbf{K}^{-1}\|_{\text{op}}$  in Lemma C.1, the bound of  $\|\mathbf{W}\|_{\text{op}}$  in Lemma C.2, and the bound of  $\mathbb{E}[\|\mathbf{F}^\top \mathbf{F}(\widehat{\mathbf{A}} - \widehat{\mathbf{K}}\mathbf{W})^\top\|_{\text{F}} \mid \varepsilon]$  in Lemma C.3, the second conditional expectation is bounded from above by

$$n^{1/2}C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2).$$

Similarly, the third conditional expectation is bounded from above by

$$n^{1/2}C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2).$$

In summary, we have

$$\begin{aligned} & n^{-1}\mathbb{E}[I(\Omega_*)\|(\mathbf{R} + \mathbf{F}\widetilde{\mathbf{W}}^\top)^\top(\mathbf{R} + \mathbf{F}\widetilde{\mathbf{W}}^\top) - (\mathbf{R} + \mathbf{F}\mathbf{W}^\top)^\top(\mathbf{R} + \mathbf{F}\mathbf{W}^\top)\|_{\text{F}} \mid \varepsilon] \\ & \leq \frac{1}{\sqrt{n}}\left(\frac{\|\varepsilon\|}{\sqrt{n}} + 1\right)C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2). \end{aligned}$$

Since  $\tilde{r}_t$  and  $\hat{r}_t$  are the  $t$ -th diagonal entries of  $(\mathbf{R} + \mathbf{F}\widetilde{\mathbf{W}}^\top)^\top(\mathbf{R} + \mathbf{F}\widetilde{\mathbf{W}}^\top)/n$  and  $(\mathbf{R} + \mathbf{F}\mathbf{W}^\top)^\top(\mathbf{R} + \mathbf{F}\mathbf{W}^\top)/n$ , respectively, we have

$$\begin{aligned} & \mathbb{E}[I(\Omega_*)|\tilde{r}_t - \hat{r}_t| \mid \varepsilon] \\ & \leq \mathbb{E}\left[I(\Omega_*)\|(\mathbf{R} + \mathbf{F}\widetilde{\mathbf{W}}^\top)^\top(\mathbf{R} + \mathbf{F}\widetilde{\mathbf{W}}^\top) - (\mathbf{R} + \mathbf{F}\mathbf{W}^\top)^\top(\mathbf{R} + \mathbf{F}\mathbf{W}^\top)\|_{\text{F}} \mid \varepsilon\right] \\ & \leq \frac{1}{\sqrt{n}}\left(\frac{\|\varepsilon\|}{\sqrt{n}} + 1\right)C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2). \end{aligned}$$

Using the same argument in the proof of Theorem 3.6, we have for any  $\epsilon > 0$ ,

$$\begin{aligned} \mathbb{P}\left(I(\Omega_*)|\tilde{r}_t - \hat{r}_t| > \epsilon \mid \varepsilon\right) & \leq \min\left(1, \frac{C(T, \gamma, \eta_{\max}, c_0, \delta)}{\epsilon}\left(\frac{1}{\sqrt{n}} + \frac{\|\varepsilon\|}{n}\right)\right) \\ & \leq \max\left\{1, \frac{C(T, \gamma, \eta_{\max}, c_0, \delta)}{\epsilon}\right\} \min\left(1, \frac{1}{\sqrt{n}} + \frac{\|\varepsilon\|}{n}\right). \end{aligned}$$

Taking expectation with respect to  $\varepsilon$ , we have

$$\mathbb{P}\left(I(\Omega_*)|\tilde{r}_t - \hat{r}_t| > \epsilon\right) \leq \max\left\{1, \frac{C(T, \gamma, \eta_{\max}, c_0, \delta)}{\epsilon}\right\} \mathbb{E}\left[\min\left(1, \frac{1}{\sqrt{n}} + \frac{\|\varepsilon\|}{n}\right)\right].$$

Using the union bound and  $\mathbb{P}(\Omega_*) \geq 1 - e^{-n/18} - e^{-n/2} \geq 1 - 2e^{-n/18}$ , we obtain

$$\mathbb{P}\left(|\tilde{r}_t - \hat{r}_t| > \epsilon\right) \leq 2e^{-n/18} + \max\left\{1, \frac{C(T, \gamma, \eta_{\max}, c_0, \delta)}{\epsilon}\right\} \left[\frac{1}{\sqrt{n}} + \mathbb{E}[\min(1, \frac{\|\varepsilon\|}{n})]\right].$$

Using the triangle inequality and the tail probability of  $|\hat{r}_t - r_t|$  in Theorem 3.6, we have

$$\mathbb{P}\left(|\tilde{r}_t - r_t| > \epsilon\right) \leq 2e^{-n/18} + \max\left\{1, \frac{C(T, \gamma, \eta_{\max}, c_0, \delta)}{\epsilon}\right\} \left[\frac{1}{\sqrt{n}} + \mathbb{E}[\min(1, \frac{\|\varepsilon\|}{n})]\right].$$

## D Proof of results in Appendix B.2

### D.1 Proof of Lemma B.1

By assumption, we know  $\hat{\mathbf{b}}^1$  is independent of  $\mathbf{X}$  and  $\hat{\mathbf{b}}^{t+1} = \phi_t(\hat{\mathbf{b}}^t + \frac{\eta_t}{|I_t|} \mathbf{X}^\top \mathbf{S}_t \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^t))$  from (4). Recall that  $\mathbf{D}_t = \frac{\partial \psi(\mathbf{u})}{\partial \mathbf{u}} \Big|_{\mathbf{u}=\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^t}$  and  $\tilde{\mathbf{D}}_t = \frac{\partial \phi_t(\mathbf{v})}{\partial \mathbf{v}} \Big|_{\mathbf{v}=\hat{\mathbf{b}}^t + \frac{\eta_t}{|I_t|} \mathbf{X}^\top \mathbf{S}_t \psi(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^t)}$ . Let a dot denote the derivative with respect to  $x_{ij}$ . By product rule and chain rule and using  $\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^t = \boldsymbol{\varepsilon} - \mathbf{X}(\hat{\mathbf{b}}^t - \mathbf{b}^*)$ , we have

$$\begin{aligned} \dot{\mathbf{b}}^{t+1} &= \tilde{\mathbf{D}}_t \left[ \dot{\mathbf{b}}^t + \frac{\eta_t}{|I_t|} \left( \dot{\mathbf{X}}^\top \mathbf{S}_t \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^t) - \mathbf{X}^\top \mathbf{S}_t \mathbf{D}_t (\dot{\mathbf{X}}(\hat{\mathbf{b}}^t - \mathbf{b}^*) + \mathbf{X}\dot{\mathbf{b}}^t) \right) \right] \\ &= \tilde{\mathbf{D}}_t \left[ \dot{\mathbf{b}}^t + \frac{\eta_t}{|I_t|} \left( \dot{\mathbf{X}}^\top \mathbf{F}_t - \mathbf{X}^\top \mathbf{S}_t \mathbf{D}_t (\dot{\mathbf{X}} \mathbf{H}_t + \mathbf{X}\dot{\mathbf{b}}^t) \right) \right], \end{aligned}$$

where the last line uses  $\mathbf{F}_t = \mathbf{S}_t \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^t)$  and  $\mathbf{H}_t = \hat{\mathbf{b}}^t - \mathbf{b}^*$ . Arranging terms gives

$$-\tilde{\mathbf{D}}_t \left( \mathbf{I}_p - \frac{\eta_t}{|I_t|} \mathbf{X}^\top \mathbf{S}_t \mathbf{D}_t \mathbf{X} \right) \dot{\mathbf{b}}^t + \dot{\mathbf{b}}^{t+1} = \frac{\eta_t}{|I_t|} \tilde{\mathbf{D}}_t (\dot{\mathbf{X}}^\top \mathbf{F}_t - \mathbf{X}^\top \mathbf{S}_t \mathbf{D}_t \dot{\mathbf{X}} \mathbf{H}_t).$$

Let  $\mathbf{P}_t = \tilde{\mathbf{D}}_t (\mathbf{I}_p - \frac{\eta_t}{|I_t|} \mathbf{X}^\top \mathbf{S}_t \mathbf{D}_t \mathbf{X})$  and  $\mathbf{a}_t = \frac{\eta_t}{|I_t|} \tilde{\mathbf{D}}_t (\dot{\mathbf{X}}^\top \mathbf{F}_t - \mathbf{X}^\top \mathbf{S}_t \mathbf{D}_t \dot{\mathbf{X}} \mathbf{H}_t)$ , we can rewrite the above recursion of  $\dot{\mathbf{b}}^t$  as a linear system:

$$\underbrace{\begin{bmatrix} \mathbf{I}_p & & & \\ -\mathbf{P}_1 & \mathbf{I}_p & & \\ & \ddots & \ddots & \\ & & -\mathbf{P}_{T-1} & \mathbf{I}_p \end{bmatrix}}_{\mathcal{M}} \begin{bmatrix} \dot{\mathbf{b}}^1 \\ \dot{\mathbf{b}}^2 \\ \vdots \\ \dot{\mathbf{b}}^T \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{0} & \mathbf{0} & & \\ \mathbf{I}_p & \mathbf{0} & & \\ & \ddots & \ddots & \\ & & \mathbf{I}_p & \mathbf{0} \end{bmatrix}}_{\mathbf{L}} \underbrace{\begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_T \end{bmatrix}}_{\mathbf{a}}.$$

Solving the above system, we have  $\dot{\mathbf{b}}^t = (\mathbf{e}_t^\top \otimes \mathbf{I}_p) \mathcal{M}^{-1} \mathbf{L} \mathbf{a}$ . Since  $\dot{\mathbf{X}} = \frac{\partial \mathbf{X}}{\partial x_{ij}} = \mathbf{e}_i \mathbf{e}_j^\top$ ,  $\mathbf{a}_t$  can be further simplified as

$$\mathbf{a}_t = \frac{\eta_t}{|I_t|} \tilde{\mathbf{D}}_t (\mathbf{e}_j \mathbf{e}_i^\top \mathbf{F}_t - \mathbf{X}^\top \mathbf{S}_t \mathbf{D}_t \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}_t).$$

Using  $\mathcal{D} = \sum_{t=1}^T ((\mathbf{e}_t \mathbf{e}_t^\top) \otimes \mathbf{D}_t)$ ,  $\tilde{\mathcal{D}} = \sum_{t=1}^T ((\mathbf{e}_t \mathbf{e}_t^\top) \otimes \tilde{\mathbf{D}}_t)$ ,  $\mathcal{S} = \sum_{t=1}^T ((\mathbf{e}_t \mathbf{e}_t^\top) \otimes \mathbf{S}_t)$ , and  $\boldsymbol{\Lambda} = \sum_{t=1}^T \frac{\eta_t}{|I_t|} \mathbf{e}_t \mathbf{e}_t^\top$ , we have

$$\begin{aligned} \mathbf{a} &= (\boldsymbol{\Lambda} \otimes \mathbf{I}_p) \tilde{\mathcal{D}} [\text{vec}(\mathbf{e}_j \mathbf{e}_i^\top \mathbf{F}) - (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{S} \mathcal{D} \text{vec}(\mathbf{e}_i \mathbf{e}_j^\top \mathbf{H})] \\ &= (\boldsymbol{\Lambda} \otimes \mathbf{I}_p) \tilde{\mathcal{D}} [((\mathbf{F}^\top \mathbf{e}_i) \otimes \mathbf{e}_j) - (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{S} \mathcal{D} ((\mathbf{H}^\top \mathbf{e}_j) \otimes \mathbf{e}_i)]. \end{aligned}$$

Plugging this expression for  $\mathbf{a}$  into  $\dot{\mathbf{b}}^t = (\mathbf{e}_t^\top \otimes \mathbf{I}_p) \mathcal{M}^{-1} \mathbf{L} \mathbf{a}$  gives

$$\frac{\partial \hat{\mathbf{b}}^t}{\partial x_{ij}} = (\mathbf{e}_t^\top \otimes \mathbf{I}_p) \mathcal{M}^{-1} \mathbf{L} (\boldsymbol{\Lambda} \otimes \mathbf{I}_p) \tilde{\mathcal{D}} [((\mathbf{F}^\top \mathbf{e}_i) \otimes \mathbf{e}_j) - (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{S} \mathcal{D} ((\mathbf{H}^\top \mathbf{e}_j) \otimes \mathbf{e}_i)].$$

This finishes the proof of Lemma B.1.

### D.2 Proof of Lemma B.2

By definition,  $\mathbf{F}_{lt} = \mathbf{e}_l^\top \mathbf{S}_t \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^t)$ . By the chain rule of differentiation, we have

$$\frac{\partial \mathbf{F}_{lt}}{\partial x_{ij}} = \mathbf{e}_l^\top \frac{\partial \mathbf{S}_t \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^t)}{\partial x_{ij}} = -\mathbf{e}_l^\top \mathbf{S}_t \mathbf{D}_t (\mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} \mathbf{e}_t + \mathbf{X} \frac{\partial \hat{\mathbf{b}}^t}{\partial x_{ij}}).$$

Notice that  $(\mathbf{e}_t \otimes (\mathbf{D}_t \mathbf{S}_t)) = \mathcal{D} \mathcal{S} (\mathbf{e}_t \otimes \mathbf{I}_n)$ . The desired formula then follows by plugging in the expression of  $\frac{\partial \hat{\mathbf{b}}^t}{\partial x_{ij}}$  in Lemma B.1.

### D.3 Proof of Lemma B.3

The desired identity follows by

$$\frac{\partial \tilde{\mathbf{F}}_{lt}}{\partial x_{ij}} = \mathbf{e}_l^\top \frac{\partial \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^t)}{\partial x_{ij}} = -\mathbf{e}_l^\top \mathbf{D}_t (\mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} \mathbf{e}_t + \mathbf{X} \frac{\partial \hat{\mathbf{b}}^t}{\partial x_{ij}})$$

and the expression of  $\frac{\partial \hat{\mathbf{b}}^t}{\partial x_{ij}}$  in Lemma B.1.

#### D.4 Alternative expressions for the matrices defined in Definition B.4

This section derives alternative expressions for the matrices  $\Upsilon_1, \dots, \Upsilon_5$  defined in Definition B.4.

We first study  $\Upsilon_1$  in (18). Using  $\mathbf{F} = \sum_{t=1}^T \mathbf{F} \mathbf{e}_t \mathbf{e}_t^\top$ , we have by Lemma B.2

$$\sum_{j=1}^n \frac{\partial \mathbf{e}_j^\top \mathbf{F}}{\partial x_{ij}} = \sum_{i=1}^n \sum_{t=1}^T \frac{\partial F_{it}}{\partial x_{ij}} \mathbf{e}_t^\top = \sum_{i=1}^n \sum_{t=1}^T D_{ij}^{it} \mathbf{e}_t^\top + \sum_{i=1}^n \sum_{t=1}^T \Delta_{ij}^{it} \mathbf{e}_t^\top. \quad (37)$$

Now we compute the two terms in the right-hand side of (37). For the first term, using the expression of  $D_{ij}^{it}$  in Lemma B.2,

$$\begin{aligned} & \sum_{i,t} D_{ij}^{it} \mathbf{e}_t^\top \\ &= -\mathbf{e}_j^\top \mathbf{H} \sum_{t=1}^T \text{Tr}(\mathbf{S}_t \mathbf{D}_t) \mathbf{e}_t \mathbf{e}_t^\top + \sum_{i,t} ((\mathbf{e}_j^\top \mathbf{H}) \otimes \mathbf{e}_i^\top) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma}^\top (\mathbf{e}_t \otimes (\mathbf{X}^\top \mathbf{D}_t \mathbf{S}_t \mathbf{e}_i)) \mathbf{e}_t^\top \\ &= -\mathbf{e}_j^\top \mathbf{H} \sum_{t=1}^T \text{Tr}(\mathbf{S}_t \mathbf{D}_t) \mathbf{e}_t \mathbf{e}_t^\top + \mathbf{e}_j^\top \mathbf{H} \sum_{i=1}^n (\mathbf{I}_T \otimes \mathbf{e}_i^\top) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{e}_i) \\ &= -\mathbf{e}_j^\top \mathbf{H} \widetilde{\mathbf{K}}^\top \end{aligned} \quad \text{by (26).}$$

Next, we compute the second term in the right-hand side of (37). Using the expression of  $\Delta_{ij}^{it}$  in Lemma B.2,

$$\begin{aligned} \sum_{i,t} \Delta_{ij}^{it} \mathbf{e}_t^\top &= -\sum_{i,t} ((\mathbf{e}_i^\top \mathbf{F}) \otimes \mathbf{e}_j^\top) \mathbf{\Gamma}^\top (\mathbf{e}_t \otimes (\mathbf{X}^\top \mathbf{D}_t \mathbf{S}_t \mathbf{e}_i)) \mathbf{e}_t^\top \\ &= -\mathbf{e}_j^\top \underbrace{\sum_i ((\mathbf{e}_i^\top \mathbf{F}) \otimes \mathbf{I}_p) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{e}_i)}_{\Upsilon_1}. \end{aligned}$$

The identity (18) then follows by substituting the above two expressions into (37).

To study  $\Upsilon_2$  in (19), we use a similar procedure. Using the mixed property of Kronecker product and the fact that the transpose of a scalar remains the same, we have

$$\begin{aligned} \sum_{j=1}^p \frac{\partial \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} &= \sum_{j=1}^p \sum_{t=1}^T \frac{\partial \mathbf{e}_j^\top \widehat{\mathbf{b}}^t}{\partial x_{ij}} \mathbf{e}_t^\top = \sum_{j,t} \mathbf{e}_j^\top (\mathbf{e}_t^\top \otimes \mathbf{I}_p) \mathbf{\Gamma}^\top ((\mathbf{F}^\top \mathbf{e}_i) \otimes \mathbf{e}_j) \mathbf{e}_t^\top \\ &\quad - \sum_{j,t} \mathbf{e}_j^\top (\mathbf{e}_t^\top \otimes \mathbf{I}_p) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{SD}((\mathbf{H}^\top \mathbf{e}_j) \otimes \mathbf{e}_i) \mathbf{e}_t^\top \quad \text{by (15)} \\ &= \mathbf{e}_i^\top \mathbf{F} \sum_j (\mathbf{I}_T \otimes \mathbf{e}_j^\top) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{e}_j) \\ &\quad - \sum_{j,t} ((\mathbf{e}_j^\top \mathbf{H}) \otimes \mathbf{e}_i^\top) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma}^\top (\mathbf{e}_t \otimes \mathbf{e}_j) \mathbf{e}_t^\top \\ &= \mathbf{e}_i^\top \mathbf{F} \sum_j (\mathbf{I}_T \otimes \mathbf{e}_j^\top) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{e}_j) \\ &\quad - \mathbf{e}_i^\top \sum_j ((\mathbf{e}_j^\top \mathbf{H}) \otimes \mathbf{I}_n) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{e}_j) \\ &= \mathbf{e}_i^\top \mathbf{F} \mathbf{W}^\top - \mathbf{e}_i^\top \Upsilon_2, \end{aligned}$$

where  $\mathbf{W} = \sum_j (\mathbf{I}_T \otimes \mathbf{e}_j^\top) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{e}_j)$  and  $\Upsilon_2 = \sum_j ((\mathbf{e}_j^\top \mathbf{H}) \otimes \mathbf{I}_n) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{e}_j)$ .

To study  $\Upsilon_3$  in (20), we use the product rule of differentiation and (18)-(19):

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^p \frac{\partial \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} &= \sum_{i=1}^n \sum_{j=1}^p \frac{\partial \mathbf{F}^\top}{\partial x_{ij}} \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} + \mathbf{F}^\top \sum_{i=1}^n \sum_{j=1}^p \mathbf{e}_i \frac{\partial \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} \\ &= - \sum_j (\widetilde{\mathbf{K}} \mathbf{H}^\top \mathbf{e}_j + \Upsilon_1^\top \mathbf{e}_j) \mathbf{e}_j^\top \mathbf{H} + \mathbf{F}^\top \sum_i \mathbf{e}_i (\mathbf{e}_i^\top \mathbf{F} \mathbf{W}^\top - \mathbf{e}_i^\top \Upsilon_2) \\ &= - \widetilde{\mathbf{K}} \mathbf{H}^\top \mathbf{H} + \mathbf{F}^\top \mathbf{F} \mathbf{W}^\top - \underbrace{(\Upsilon_1^\top \mathbf{H} + \mathbf{F}^\top \Upsilon_2)}_{\Upsilon_3}.\end{aligned}$$

For  $\Upsilon_4$  in (21), by the product rule,

$$\begin{aligned}&\sum_{i=1}^n \sum_{j=1}^p \frac{\partial \mathbf{H}^\top \mathbf{X}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} \\ &= \sum_{i,j} \frac{\partial \mathbf{H}^\top}{\partial x_{ij}} \mathbf{X}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} + \mathbf{H}^\top \sum_{i,j} \mathbf{e}_j \mathbf{e}_i^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} + \mathbf{H}^\top \mathbf{X}^\top \sum_{i,j} \mathbf{e}_i \mathbf{e}_j^\top \frac{\partial \mathbf{H}}{\partial x_{ij}} \\ &= \sum_{i,j} \frac{\partial \mathbf{H}^\top}{\partial x_{ij}} \mathbf{X}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} + n \mathbf{H}^\top \mathbf{H} + \mathbf{H}^\top \mathbf{X}^\top (\mathbf{F} \mathbf{W}^\top - \Upsilon_2) \quad \text{by (19)}.\end{aligned}$$

We then compute the first term of the last line as follows

$$\begin{aligned}&\sum_{i,j} \frac{\partial \mathbf{H}^\top}{\partial x_{ij}} \mathbf{X}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} \\ &= \sum_{i,j,t} \mathbf{e}_t \mathbf{e}_t^\top \frac{\partial \mathbf{H}^\top}{\partial x_{ij}} \mathbf{X}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} \\ &= \sum_{i,j,t} \mathbf{e}_t \left( \frac{\partial \widehat{\mathbf{b}}^t}{\partial x_{ij}} \right)^\top \mathbf{X}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} \\ &= \sum_{i,j,t} \mathbf{e}_t \mathbf{e}_i^\top \mathbf{X} \frac{\partial \widehat{\mathbf{b}}^t}{\partial x_{ij}} \mathbf{e}_j^\top \mathbf{H} \\ &= - \sum_{i,j,t} \mathbf{e}_t \mathbf{e}_i^\top \mathbf{X} (\mathbf{e}_t^\top \otimes \mathbf{I}_p) \Gamma(\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{SD}((\mathbf{H}^\top \mathbf{e}_j) \otimes \mathbf{e}_i) \mathbf{e}_j^\top \mathbf{H} + \widetilde{\Upsilon}_1 \quad \text{by (15)} \\ &= - \sum_i (\mathbf{I}_T \otimes (\mathbf{e}_i^\top \mathbf{X})) \Gamma(\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{SD}(\mathbf{I}_T \otimes \mathbf{e}_i) \mathbf{H}^\top \mathbf{H} + \widetilde{\Upsilon}_1 \\ &= - \widetilde{\mathbf{A}} \mathbf{H}^\top \mathbf{H} + \widetilde{\Upsilon}_1 \quad \text{by (23)},\end{aligned}$$

where

$$\widetilde{\Upsilon}_1 = \sum_{i,j,t} \mathbf{e}_t \mathbf{e}_i^\top \mathbf{X} (\mathbf{e}_t^\top \otimes \mathbf{I}_p) \Gamma((\mathbf{F}^\top \mathbf{e}_i) \otimes \mathbf{e}_j) \mathbf{e}_j^\top \mathbf{H} = \sum_i (\mathbf{I}_T \otimes \mathbf{e}_i^\top \mathbf{X}) \Gamma((\mathbf{F}^\top \mathbf{e}_i) \otimes \mathbf{I}_p) \mathbf{H}.$$

Combining the above pieces, we have

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^p \frac{\partial \mathbf{H}^\top \mathbf{X}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} &= - \widetilde{\mathbf{A}} \mathbf{H}^\top \mathbf{H} + \widetilde{\Upsilon}_1 + n \mathbf{H}^\top \mathbf{H} + \mathbf{H}^\top \mathbf{X}^\top (\mathbf{F} \mathbf{W}^\top - \Upsilon_2) \\ &= (n \mathbf{I}_T - \widetilde{\mathbf{A}}) \mathbf{H}^\top \mathbf{H} + \mathbf{H}^\top \mathbf{X}^\top \mathbf{F} \mathbf{W}^\top - \underbrace{(\mathbf{H}^\top \mathbf{X}^\top \Upsilon_2 - \widetilde{\Upsilon}_1)}_{\Upsilon_4}.\end{aligned}$$

This provides an alternative expression for  $\Upsilon_4$  in (21).

Last, we study  $\Upsilon_5$  in (22). We have

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^p \frac{\partial \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \tilde{\mathbf{F}}}{\partial x_{ij}} &= \sum_{i,j} \left[ \frac{\partial \mathbf{F}^\top \mathbf{e}_i}{\partial x_{ij}} \mathbf{e}_j^\top \mathbf{X}^\top \tilde{\mathbf{F}} + \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{e}_j \mathbf{e}_i^\top \tilde{\mathbf{F}} + \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \frac{\partial \tilde{\mathbf{F}}}{\partial x_{ij}} \right] \\ &= -(\tilde{\mathbf{K}} \mathbf{H}^\top + \Upsilon_1^\top) \mathbf{X}^\top \tilde{\mathbf{F}} + p \mathbf{F}^\top \tilde{\mathbf{F}} + \sum_{i,j} \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \frac{\partial \tilde{\mathbf{F}}}{\partial x_{ij}}. \end{aligned}$$

It remains to compute the third term in the last display. Using the fact that  $\tilde{\mathbf{F}} = \sum_{l=1}^n \sum_{t=1}^T \mathbf{e}_l \mathbf{e}_l^\top \tilde{\mathbf{F}} \mathbf{e}_t \mathbf{e}_t^\top$ , we have by Lemma B.3 that

$$\frac{\partial \tilde{\mathbf{F}}}{\partial x_{ij}} = \sum_{l,t} \mathbf{e}_l \frac{\partial \mathbf{e}_l^\top \tilde{\mathbf{F}} \mathbf{e}_t}{\partial x_{ij}} \mathbf{e}_t^\top = \sum_{l,t} \mathbf{e}_l (\tilde{D}_{ij}^{lt} + \tilde{\Delta}_{ij}^{lt}) \mathbf{e}_t^\top.$$

Using

$$\begin{aligned} \tilde{D}_{ij}^{lt} &= -\mathbf{e}_l^\top \mathbf{D}_t \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} \mathbf{e}_t + ((\mathbf{e}_j^\top \mathbf{H}) \otimes \mathbf{e}_i^\top) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \Gamma^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{D}(\mathbf{e}_t \otimes \mathbf{e}_l), \\ \tilde{\Delta}_{ij}^{lt} &= -((\mathbf{e}_i^\top \mathbf{F}) \otimes \mathbf{e}_j^\top) \Gamma^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{D}(\mathbf{e}_t \otimes \mathbf{e}_l), \end{aligned}$$

we find

$$\begin{aligned} \sum_{i,j} \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \frac{\partial \tilde{\mathbf{F}}}{\partial x_{ij}} &= \sum_{i,j,l,t} \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \mathbf{e}_l \frac{\partial \mathbf{e}_l^\top \tilde{\mathbf{F}} \mathbf{e}_t}{\partial x_{ij}} \mathbf{e}_t^\top \\ &= \sum_{i,j,l,t} \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \mathbf{e}_l \tilde{\Delta}_{ij}^{lt} \mathbf{e}_t^\top + \sum_{i,j,l,t} \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \mathbf{e}_l \tilde{D}_{ij}^{lt} \mathbf{e}_t^\top. \end{aligned}$$

We now compute the two terms in the above display. For the first term, we have

$$\begin{aligned} &\sum_{i,j,l,t} \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \mathbf{e}_l \tilde{\Delta}_{ij}^{lt} \mathbf{e}_t^\top \\ &= \sum_{i,j,l,t} \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \mathbf{e}_l [ -((\mathbf{e}_i^\top \mathbf{F}) \otimes \mathbf{e}_j^\top) \Gamma^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{D}(\mathbf{e}_t \otimes \mathbf{e}_l) ] \mathbf{e}_t^\top \\ &= -\mathbf{F}^\top \mathbf{F} \sum_{l=1}^n (\mathbf{I}_T \otimes (\mathbf{e}_l^\top \mathbf{X})) \Gamma^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{D}(\mathbf{I}_T \otimes \mathbf{e}_l) \\ &= -\mathbf{F}^\top \mathbf{F} \hat{\mathbf{A}}^\top \end{aligned} \quad \text{by (24).}$$

For the second term, we have

$$\begin{aligned} &\sum_{i,j,l,t} \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \mathbf{e}_l \tilde{D}_{ij}^{lt} \mathbf{e}_t^\top \\ &= \sum_{i,j,l,t} \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \mathbf{e}_l [ -\mathbf{e}_l^\top \mathbf{D}_t \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} \mathbf{e}_t + ((\mathbf{e}_j^\top \mathbf{H}) \otimes \mathbf{e}_i^\top) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \Gamma^\top (\mathbf{e}_t \otimes (\mathbf{X}^\top \mathbf{D}_t \mathbf{e}_l)) ] \mathbf{e}_t^\top \\ &= - \underbrace{\sum_{t=1}^T \mathbf{F}^\top \mathbf{D}_t \mathbf{X} \mathbf{H} \mathbf{e}_t + \sum_{l=1}^n (\mathbf{e}_l^\top \mathbf{X} \mathbf{H} \otimes \mathbf{F}^\top) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \Gamma^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{D}(\mathbf{I}_T \otimes \mathbf{e}_l)}_{\tilde{\Upsilon}_2}. \end{aligned}$$

Thus, we have established

$$\sum_{i=1}^n \sum_{j=1}^p \frac{\partial \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \tilde{\mathbf{F}}}{\partial x_{ij}} = -\tilde{\mathbf{K}} \mathbf{H}^\top \mathbf{X}^\top \tilde{\mathbf{F}} + p \mathbf{F}^\top \tilde{\mathbf{F}} - \mathbf{F}^\top \mathbf{F} \hat{\mathbf{A}}^\top - \underbrace{(\Upsilon_1^\top \mathbf{X}^\top \tilde{\mathbf{F}} - \tilde{\Upsilon}_2)}_{\Upsilon_5}.$$

This provides an alternative expression for  $\Upsilon_5$  in (22).



## D.5 Preparation results for proving Lemmas B.5 and B.6

**Lemma D.1** (Moment bounds for  $\mathbf{H}, \mathbf{F}, \tilde{\mathbf{F}}$ ). *Under Assumptions 3.1, 3.3 and 3.5 with  $\Sigma = \mathbf{I}_p$ . Let  $\mathbf{H}, \mathbf{F}$  be defined in (13), we have for any finite integer  $k$ ,*

$$\begin{aligned}\mathbb{E}[\|\mathbf{X}/\sqrt{n}\|_{\text{op}}^{2k}] &\leq C(\gamma, k), \\ \mathbb{E}[\|\mathbf{H}\|_{\text{F}}^{2k} \mid \varepsilon] &\leq C(T, \gamma, c_0, \eta_{\max}, k)(\delta^2 + \|\mathbf{b}^*\|)^{2k}, \\ \mathbb{E}[\|\mathbf{F}/\sqrt{n}\|_{\text{F}}^{2k} \mid \varepsilon] &\leq \mathbb{E}[\|\tilde{\mathbf{F}}/\sqrt{n}\|_{\text{F}}^{2k} \mid \varepsilon] \leq C(T, k)\delta^{2k}.\end{aligned}$$

*Proof of Lemma D.1.* For the first inequality, according to Assumption 3.1 and  $\Sigma = \mathbf{I}_p$ ,  $\mathbf{X}$  has i.i.d. standard normal entries. By [11, Theorem II.13], there exists a random variable  $z \sim N(0, 1)$  such that  $\|\mathbf{X}\|_{\text{op}} \leq \sqrt{n} + \sqrt{p} + z$  almost surely. Under Assumption 3.5 that  $p/n \leq \gamma$ , we have  $\mathbb{E}[\|\mathbf{X}/\sqrt{n}\|_{\text{op}}^k] \leq C(\gamma, k)$  for any finite integer  $k$ .

For the second inequality, since  $\|\mathbf{H}\|_{\text{F}}^2 = \sum_{t=1}^T \|\hat{\mathbf{b}}^t - \mathbf{b}^*\|^2$ , it suffices to bound  $\|\hat{\mathbf{b}}^t - \mathbf{b}^*\|^2$  for each  $t \in [T]$ . Define the sequence of scalars  $a_t \stackrel{\text{def}}{=} \max\{\|\hat{\mathbf{b}}^t\|, \delta\}$ . Since  $\hat{\mathbf{b}}^t = \phi_{t-1}(\hat{\mathbf{b}}^{t-1} + \frac{\eta_t}{n_t} \mathbf{X}^\top \mathbf{S}_t \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^{t-1}))$  where  $n_t := |I_t|$ . Note that  $\|\psi(\mathbf{y}_{I_t} - \mathbf{X}_{I_t} \mathbf{b})\| \leq \sqrt{|I_t|}\delta$  by Assumption 3.3, we have

$$\begin{aligned}\|\hat{\mathbf{b}}^t - \mathbf{0}\| &= \|\hat{\mathbf{b}}^t - \phi_{t-1}(\mathbf{0})\| && \text{since } \phi_{t-1} = \mathbf{0} \\ &\leq \|\hat{\mathbf{b}}^{t-1} + \frac{\eta_t}{|I_t|} \mathbf{X}^\top \mathbf{S}_t \psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^{t-1})\| && \text{since } \phi_t \text{ is 1-Lipschitz} \\ &= \|\hat{\mathbf{b}}^{t-1} + \frac{\eta_t}{|I_t|} \mathbf{X}_{I_t}^\top \psi(\mathbf{y}_{I_t} - \mathbf{X}_{I_t} \hat{\mathbf{b}}^{t-1})\| \\ &\leq \|\hat{\mathbf{b}}^{t-1}\| + \frac{\eta_t}{\sqrt{|I_t|}} \|\mathbf{X}_{I_t}\|_{\text{op}} \delta && \text{by the triangle inequality} \\ &\leq \|\hat{\mathbf{b}}^{t-1}\| + \frac{\eta_{\max}}{\sqrt{c_0 n}} \|\mathbf{X}\|_{\text{op}} \delta && \text{by } |I_t| \geq c_0 n \\ &\leq a_{t-1} + \frac{\eta_{\max}}{\sqrt{c_0}} \|\mathbf{X}/\sqrt{n}\|_{\text{op}} a_{t-1} \\ &= (1 + \frac{\eta_{\max}}{\sqrt{c_0}} \|\mathbf{X}/\sqrt{n}\|_{\text{op}}) a_{t-1}.\end{aligned}$$

Since  $a_t = \max\{\|\hat{\mathbf{b}}^t\|, \delta\}$  and  $\delta \leq a_{t-1}$ , we have

$$a_t \leq (1 + \frac{\eta_{\max}}{\sqrt{c_0}} \|\mathbf{X}/\sqrt{n}\|_{\text{op}}) a_{t-1}.$$

Notice  $a_1 = \delta$  since  $\hat{\mathbf{b}}^1 = \mathbf{0}_p$ , we obtain

$$a_t \leq (1 + \frac{\eta_{\max}}{\sqrt{c_0}} \|\mathbf{X}/\sqrt{n}\|_{\text{op}})^{t-1} \delta.$$

Hence, using the inequality  $\|\hat{\mathbf{b}}^t - \mathbf{b}^*\|^2 \leq 2\|\hat{\mathbf{b}}^t\|^2 + 2\|\mathbf{b}^*\|^2$ , we have

$$\begin{aligned}\|\mathbf{H}\|_{\text{F}}^2 &\lesssim \sum_{t=1}^T (\|\hat{\mathbf{b}}^t\|^2 + \|\mathbf{b}^*\|^2) \\ &\leq \sum_{t=1}^T [(1 + \frac{\eta_{\max}}{\sqrt{c_0}} \|\mathbf{X}/\sqrt{n}\|_{\text{op}})^{2t-2} \delta^2 + \|\mathbf{b}^*\|^2] \\ &\leq T(\delta^2 + \|\mathbf{b}^*\|^2)(1 + \frac{\eta_{\max}}{\sqrt{c_0}} \|\mathbf{X}/\sqrt{n}\|_{\text{op}})^{2T}.\end{aligned}\tag{38}$$

Taking conditional expectation on both sides given  $\varepsilon$ , the desired moment bound for  $\mathbf{H}$  follows from the moment bound for  $\|\mathbf{X}/\sqrt{n}\|_{\text{op}}$ .

For the third inequality, since  $|\psi(x)| \leq \delta$  from Assumption 3.3, we have  $\|\psi(\mathbf{u})\| \leq n\delta^2$  for any  $\mathbf{u} \in \mathbb{R}^n$ . By the definitions of  $\mathbf{F}$  and  $\tilde{\mathbf{F}}$ , we have

$$\|\mathbf{F}\|_{\text{F}}^2 \leq \|\tilde{\mathbf{F}}\|_{\text{F}}^2 = \sum_{t=1}^T \|\psi(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}^t)\|^2 \leq Tn\delta^2.$$

Since the above display holds for any  $\varepsilon$ , it implies the desired conditional moment bounds for  $\mathbf{F}, \tilde{\mathbf{F}}$ .  $\square$

**Lemma D.2** (Frobenius norm bound for  $\mathbf{X}\mathbf{H}$ ). *Under Assumptions 3.1, 3.3 and 3.5, we have*

$$\|\mathbf{X}\mathbf{H}\|_{\text{F}}^2 \leq C(T, \gamma, \eta_{\max}, c_0)n(\delta^2 + \|\mathbf{b}^*\|^2)$$

with probability at least  $1 - \exp(-n/2)$ .

*Proof of Lemma D.2.* By [11, Theorem II.13], under Assumption 3.1 with  $\Sigma = \mathbf{I}_p$ , we have

$$\mathbb{P}(\|\mathbf{X}/\sqrt{n}\|_{\text{op}} \leq 2 + \sqrt{\gamma}) \geq 1 - \exp(-n/2).$$

Using  $\|\mathbf{X}\mathbf{H}\|_{\text{F}}^2 \leq \|\mathbf{X}\|_{\text{op}}^2 \|\mathbf{H}\|_{\text{F}}^2$  and the bound (38), we have

$$\|\mathbf{X}\mathbf{H}\|_{\text{F}}^2 \leq nC(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2)$$

holds with probability at least  $1 - \exp(-n/2)$ .  $\square$

**Lemma D.3** (Operator norm bound for  $\mathcal{M}$ ). *Under Assumptions 3.1, 3.3 and 3.5, we have*

$$\|\mathcal{M}^{-1}\|_{\text{op}} \leq C(T)(1 + \xi)^T,$$

where  $\xi = \frac{\eta_{\max}}{c_0 n} \|\mathbf{X}\|_{\text{op}}^2$ .

*Proof of Lemma D.3.* By the definition of  $\mathcal{M}$  in Lemma B.1, we have

$$\mathcal{M} = \begin{bmatrix} \mathbf{I}_p & & & & \\ -\mathbf{P}_1 & \mathbf{I}_p & & & \\ & \ddots & \ddots & & \\ & & -\mathbf{P}_{T-1} & \mathbf{I}_p & \end{bmatrix} \quad \text{where } \mathbf{P}_t = \tilde{\mathbf{D}}_t(\mathbf{I}_p - \frac{\eta_t}{|I_t|} \mathbf{X}^\top \mathbf{S}_t \mathbf{D}_t \mathbf{X}).$$

Hence, we can write  $\mathcal{M} = \mathbf{I}_{pT} - \mathbf{A}$ , where  $\mathbf{A}$  is the lower triangular matrix with off-diagonal blocks  $\mathbf{P}_1, \dots, \mathbf{P}_{T-1}$ . Using the matrix identity  $(\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k$  and noticing  $\mathbf{A}^k = \mathbf{0}$  for  $k \geq T$ , we have

$$\mathcal{M}^{-1} = \sum_{k=0}^{T-1} \begin{bmatrix} \mathbf{0} & & & & \\ \mathbf{P}_1 & \mathbf{0} & & & \\ & \ddots & \ddots & & \\ & & \mathbf{P}_{T-1} & \mathbf{0} & \end{bmatrix}^k.$$

Taking operator norm on both sides, we obtain

$$\|\mathcal{M}^{-1}\|_{\text{op}} \leq \sum_{k=0}^{T-1} \left( \sum_{t=1}^{T-1} \|\mathbf{P}_t\|_{\text{op}} \right)^k. \quad (39)$$

Since  $\mathbf{D}_t = \frac{\partial \psi(\mathbf{u})}{\partial \mathbf{u}} \big|_{\mathbf{u}=\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^t}$  and  $\psi$  is 1-Lipschitz, we have  $\|\mathbf{D}_t\|_{\text{op}} \leq 1$ . By the definition that  $\mathbf{S}_t = \sum_{i \in I_t} \mathbf{e}_i \mathbf{e}_i^\top$ , we know  $\|\mathbf{S}_t\|_{\text{op}} \leq 1$ . Since  $|I_t| \geq c_0 n$  and  $\eta_t \leq \eta_{\max}$  for any  $t \in [T]$ , we have

$$\|\mathbf{P}_t\|_{\text{op}} \leq 1 + \frac{\eta_t}{|I_t|} \|\mathbf{X}_{I_t}\|_{\text{op}}^2 \leq 1 + \frac{\eta_{\max}}{c_0 n} \|\mathbf{X}\|_{\text{op}}^2 \stackrel{\text{def}}{=} 1 + \xi.$$

Plugging this inequality into (39) gives

$$\|\mathcal{M}^{-1}\|_{\text{op}} \leq \sum_{k=0}^{T-1} (T(1 + \xi))^k \leq C(T)(1 + \xi)^T.$$

$\square$

**Lemma D.4.** *Under the same conditions as Lemma D.3, we have*

$$\|\mathbf{\Gamma}\|_{\text{op}} \leq n^{-1}C(T, \eta_{\max}, c_0)(1 + \xi)^T,$$

where  $\xi = \frac{\eta_{\max}}{c_0 n} \|\mathbf{X}\|_{\text{op}}^2$ .

*Proof of Lemma D.4.* By the definition of  $\mathbf{\Gamma}$  in Lemma B.1, we have  $\mathbf{\Gamma} = \mathcal{M}^{-1} \mathbf{L}(\mathbf{\Lambda} \otimes \mathbf{I}_p) \tilde{\mathcal{D}}$ . Notice that  $\mathbf{\Lambda} = \sum_{t=1}^T \frac{\eta_t}{|I_t|} \mathbf{e}_t \mathbf{e}_t^\top$ , we have  $\|\mathbf{\Lambda}\|_{\text{op}} = \max_{t \in [T]} \frac{\eta_t}{|I_t|} \leq n^{-1} \frac{\eta_{\max}}{c_0}$  using  $|I_t| \geq c_0 n$  and  $\eta_t \leq \eta_{\max}$ . Since  $\phi$  is 1-Lipschitz, we have  $\|\tilde{\mathcal{D}}\|_{\text{op}} \leq 1$ . By definition of  $\mathbf{L}$  in Lemma B.1, we have  $\|\mathbf{L}\|_{\text{op}} = 1$ . Using these upper bounds of  $\|\mathbf{L}\|_{\text{op}}$ ,  $\|\mathbf{\Lambda}\|_{\text{op}}$ ,  $\|\tilde{\mathcal{D}}\|_{\text{op}}$  and the upper bound of  $\|\mathcal{M}^{-1}\|_{\text{op}}$  in Lemma D.3, we obtain

$$\begin{aligned} \|\mathbf{\Gamma}\|_{\text{op}} &\leq \|\mathcal{M}^{-1}\|_{\text{op}} \|\mathbf{L}\|_{\text{op}} \|\mathbf{\Lambda} \otimes \mathbf{I}_p\|_{\text{op}} \|\tilde{\mathcal{D}}\|_{\text{op}} \\ &\leq n^{-1} C(T, \eta_{\max}, c_0) (1 + \xi)^T. \end{aligned}$$

□

**Lemma D.5** (Moment bounds for derivative of  $\mathbf{H}, \mathbf{F}, \tilde{\mathbf{F}}$ ). *Under Assumptions 3.1, 3.3 and 3.5 and  $\mathbf{\Sigma} = \mathbf{I}_p$ , we have for any finite integer  $k$ ,*

$$\begin{aligned} \mathbb{E} \left[ \left( \sum_{i=1}^n \sum_{j=1}^p \left\| \frac{\partial \mathbf{H}}{\partial x_{ij}} \right\|_{\text{F}}^2 \right)^k \middle| \epsilon \right] &\leq C(T, \gamma, \eta_{\max}, c_0) (\delta^2 + \|\mathbf{b}^*\|^2)^{2k}, \\ \mathbb{E} \left[ \left( \sum_{i=1}^n \sum_{j=1}^p \left\| \frac{\partial \mathbf{F} / \sqrt{n}}{\partial x_{ij}} \right\|_{\text{F}}^2 \right)^k \middle| \epsilon \right] &\leq \mathbb{E} \left[ \left( \sum_{i=1}^n \sum_{j=1}^p \left\| \frac{\partial \tilde{\mathbf{F}} / \sqrt{n}}{\partial x_{ij}} \right\|_{\text{F}}^2 \right)^k \middle| \epsilon \right] \leq C(T, \gamma, \eta_{\max}, c_0) (\delta^2 + \|\mathbf{b}^*\|^2)^{2k}. \end{aligned}$$

*Proof of Lemma D.5.* We first prove the first bound. By Lemma B.1, we have

$$\frac{\partial \hat{\mathbf{b}}^t}{\partial x_{ij}} = (\mathbf{e}_t^\top \otimes \mathbf{I}_p) \mathbf{\Gamma} [((\mathbf{F}^\top \mathbf{e}_i) \otimes \mathbf{e}_j) - (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{SD}((\mathbf{H}^\top \mathbf{e}_j) \otimes \mathbf{e}_i)].$$

Hence, using  $\frac{\partial \mathbf{e}_k^\top \mathbf{H} \mathbf{e}_t}{\partial x_{ij}} = \frac{\partial \mathbf{e}_k^\top \hat{\mathbf{b}}^t}{\partial x_{ij}}$ , we have

$$\begin{aligned} \frac{\partial \mathbf{e}_k^\top \mathbf{H} \mathbf{e}_t}{\partial x_{ij}} &= (\mathbf{e}_t^\top \otimes \mathbf{e}_k^\top) \mathbf{\Gamma} [((\mathbf{F}^\top \mathbf{e}_i) \otimes \mathbf{e}_j) - (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{SD}((\mathbf{H}^\top \mathbf{e}_j) \otimes \mathbf{e}_i)] \\ &= (\mathbf{e}_t^\top \otimes \mathbf{e}_k^\top) \mathbf{\Gamma} [(\mathbf{F}^\top \otimes \mathbf{I}_p)(\mathbf{e}_i \otimes \mathbf{e}_j) - (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{SD}(\mathbf{H}^\top \otimes \mathbf{I}_n)(\mathbf{e}_j \otimes \mathbf{e}_i)]. \end{aligned}$$

Using the above equality,  $\sum_{i,j,t,k} [(\mathbf{e}_t^\top \otimes \mathbf{e}_k^\top) \mathbf{A}(\mathbf{e}_j \otimes \mathbf{e}_i)]^2 = \|\mathbf{A}\|_{\text{F}}^2$  for  $\mathbf{A} \in \mathbb{R}^{pT \times np}$ , and the triangle inequality, we have

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^p \left\| \frac{\partial \mathbf{H}}{\partial x_{ij}} \right\|_{\text{F}}^2 &= \sum_{i=1}^n \sum_{j=1}^p \sum_{k=1}^p \sum_{t=1}^T \left( \frac{\partial \mathbf{e}_k^\top \mathbf{H} \mathbf{e}_t}{\partial x_{ij}} \right)^2 \\ &\lesssim \|\mathbf{\Gamma}(\mathbf{F}^\top \otimes \mathbf{I}_p)\|_{\text{F}}^2 + \|\mathbf{\Gamma}(\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{SD}(\mathbf{H}^\top \otimes \mathbf{I}_n)\|_{\text{F}}^2 \\ &\leq p \|\mathbf{\Gamma}\|_{\text{op}}^2 \|\mathbf{F}\|_{\text{F}}^2 + n \|\mathbf{\Gamma}\|_{\text{op}}^2 \|\mathbf{X}\|_{\text{op}}^2 \|\mathcal{S}\|_{\text{op}}^2 \|\mathcal{D}\|_{\text{op}}^2 \|\mathbf{H}\|_{\text{F}}^2 \\ &\leq p \|\mathbf{\Gamma}\|_{\text{op}}^2 \|\mathbf{F}\|_{\text{F}}^2 + n \|\mathbf{\Gamma}\|_{\text{op}}^2 \|\mathbf{X}\|_{\text{op}}^2 \|\mathbf{H}\|_{\text{F}}^2 \\ &\leq C(T, \gamma, \eta_{\max}, c_0) (1 + \xi)^{2T} \|\mathbf{F}\|_{\text{F}}^2 / n + C(T, \gamma, \eta_{\max}, c_0) (1 + \xi)^{2T} \|\mathbf{X}\|_{\text{op}}^2 / n \|\mathbf{H}\|_{\text{F}}^2, \end{aligned}$$

where the last inequality uses Lemma D.4. Taking the conditional expectation on both sides given  $\epsilon$ , the desired moment bound follows from the moment bounds of  $\mathbf{X}, \mathbf{H}, \mathbf{F}$  in Lemma D.1.

Now we prove the second bound. By definition, the  $t$ -th column of  $\mathbf{F}$  is  $F_t = \mathbf{S}_t \psi(\mathbf{y} - \mathbf{X} \hat{\mathbf{b}}^t)$ , it can be written using the  $t$ -th column of  $\tilde{\mathbf{F}}$  as  $F_t = \mathbf{S}_t \tilde{F}_t$ . Since  $\|\mathbf{S}_t\|_{\text{op}} \leq 1$ , we have

$$\left\| \frac{\partial \mathbf{F}}{\partial x_{ij}} \right\|_{\text{F}}^2 = \sum_t \left\| \frac{\partial F_t}{\partial x_{ij}} \right\|^2 \leq \sum_t \left\| \frac{\partial \tilde{F}_t}{\partial x_{ij}} \right\|^2 = \left\| \frac{\partial \tilde{\mathbf{F}}}{\partial x_{ij}} \right\|_{\text{F}}^2.$$

By Lemma B.3, we have  $\frac{\partial \mathbf{e}_i^\top \tilde{\mathbf{F}} \mathbf{e}_t}{\partial x_{ij}} = \tilde{D}_{ij}^{tt} + \tilde{\Delta}_{ij}^{lt}$  where

$$\begin{aligned} \tilde{D}_{ij}^{tt} &= -\mathbf{e}_i^\top \mathbf{D}_t \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H} \mathbf{e}_t + ((\mathbf{e}_j^\top \mathbf{H}) \otimes \mathbf{e}_i^\top) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{D}(\mathbf{e}_t \otimes \mathbf{e}_l), \\ &= ((\mathbf{e}_j^\top \mathbf{H}) \otimes \mathbf{e}_i^\top) [-\mathbf{I}_{pT} + \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{X}^\top)] \mathcal{D}(\mathbf{e}_t \otimes \mathbf{e}_l), \\ \tilde{\Delta}_{ij}^{lt} &= -((\mathbf{e}_i^\top \mathbf{F}) \otimes \mathbf{e}_j^\top) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{D}(\mathbf{e}_t \otimes \mathbf{e}_l). \end{aligned}$$

Using  $\sum_{i,j,t,k} [(e_j^\top \otimes e_i^\top) \mathbf{A} (e_t \otimes e_l)]^2 = \|\mathbf{A}\|_F^2$  for  $\mathbf{A} \in \mathbb{R}^{np \times nT}$  and the triangle inequality, we have

$$\begin{aligned} & \sum_{i=1}^n \sum_{j=1}^p \left\| \frac{\partial \tilde{\mathbf{F}}}{\partial x_{ij}} \right\|_F^2 \\ &= \sum_{i=1}^n \sum_{j=1}^p \sum_{l=1}^n \sum_{t=1}^T (\tilde{D}_{ij}^{lt} + \tilde{\Delta}_{ij}^{lt})^2 \\ &\lesssim \|(\mathbf{H} \otimes \mathbf{I}_n)[- \mathbf{I}_{pT} + \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X})\mathbf{\Gamma}^\top(\mathbf{I}_T \otimes \mathbf{X}^\top)]\mathcal{D}\|_F^2 + \|(\mathbf{F} \otimes \mathbf{I}_p)\mathbf{\Gamma}^\top(\mathbf{I}_T \otimes \mathbf{X}^\top)\mathcal{D}\|_F^2 \\ &\lesssim n\|\mathbf{H}\|_F^2(1 + \|\mathbf{X}\|_{\text{op}}^4\|\mathbf{\Gamma}\|_{\text{op}}^2) + \|\mathbf{F}\|_F^2\|\mathbf{\Gamma}\|_{\text{op}}^2\|\mathbf{X}\|_{\text{op}}^2, \end{aligned}$$

where the last inequality uses  $\|\mathcal{D}\|_{\text{op}} \leq 1$  and  $\|\mathcal{S}\|_{\text{op}} \leq 1$ . Taking the conditional expectation on both sides given  $\varepsilon$ , the desired moment bound follows from the bound of  $\mathbf{\Gamma}$  in Lemma D.4 and the moment bounds of  $\mathbf{X}$ ,  $\mathbf{H}$ ,  $\mathbf{F}$  in Lemma D.1.  $\square$

## D.6 Proof of Lemma B.5

**Bound of  $\Upsilon_1$ .** By the expression for the matrix  $\Upsilon_1 \in \mathbb{R}^{p \times T}$  obtained in Appendix D.4,

$$\begin{aligned} \Upsilon_1 &= \sum_{i=1}^n ((e_i^\top \mathbf{F}) \otimes \mathbf{I}_p) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{DS}(\mathbf{I}_T \otimes e_i) \\ &= \sum_{i=1}^n \sum_{t=1}^T ((e_i^\top \mathbf{F} e_t e_t^\top) \otimes \mathbf{I}_p) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{DS}(\mathbf{I}_T \otimes e_i) \\ &= \sum_{i=1}^n \sum_{t=1}^T (e_t^\top \otimes \mathbf{I}_p) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{DS}(\mathbf{I}_T \otimes e_i) e_i^\top \mathbf{F} e_t \\ &= \sum_{t=1}^T (e_t^\top \otimes \mathbf{I}_p) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{DS}(\mathbf{I}_T \otimes (\mathbf{F} e_t)). \end{aligned}$$

By the triangle inequality and  $\|\mathcal{D}\|_{\text{op}} \vee \|\mathcal{S}\|_{\text{op}} \leq 1$ ,  $\|\Upsilon_1\|_{\text{op}} \leq T\|\mathbf{\Gamma}\|_{\text{op}}\|\mathbf{X}\|_{\text{op}}\|\mathbf{F}\|_F$ . Using the bound of  $\|\mathbf{\Gamma}\|_{\text{op}}$  in Lemma D.4, the moment bound of  $\|\mathbf{X}\|_{\text{op}}$ ,  $\|\mathbf{F}\|_F$  in Lemma D.1 gives

$$\mathbb{E}[\|\Upsilon_1\|_{\text{op}}^2 \mid \varepsilon] \leq C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2).$$

**Bound of  $\Upsilon_2$ .** By the expression for the matrix  $\Upsilon_2 \in \mathbb{R}^{n \times T}$  obtained in Appendix D.4, we have

$$\begin{aligned} \Upsilon_2 &= \sum_{j=1}^p ((e_j^\top \mathbf{H}) \otimes \mathbf{I}_n) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes e_j) \\ &= \sum_{j=1}^p \sum_{t=1}^T ((e_j^\top \mathbf{H} e_t e_t^\top) \otimes \mathbf{I}_n) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes e_j) \\ &= \sum_{j=1}^p \sum_{t=1}^T (e_t^\top \otimes \mathbf{I}_n) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes e_j) e_j^\top \mathbf{H} e_t \\ &= \sum_{t=1}^T (e_t^\top \otimes \mathbf{I}_n) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes (\mathbf{H} e_t)). \end{aligned}$$

By the triangle inequality and  $\|\mathcal{D}\|_{\text{op}} \vee \|\mathcal{S}\|_{\text{op}} \leq 1$ ,  $\|\Upsilon_2\|_{\text{op}} \leq T\|\mathbf{\Gamma}\|_{\text{op}}\|\mathbf{X}\|_{\text{op}}\|\mathbf{H}\|_F$ . Similar to the moment bound of  $\|\Upsilon_1\|_{\text{op}}$ , we obtain

$$\mathbb{E}[\|\Upsilon_2\|_{\text{op}}^2 \mid \varepsilon] \leq C(T, \gamma, \eta_{\max}, c_0)n^{-1}(\delta^2 + \|\mathbf{b}^*\|^2).$$

**Bound of  $\Upsilon_3$ .** By the expression for the matrix  $\Upsilon_3 \in \mathbb{R}^{T \times T}$  obtained in Appendix D.4, we have  $\Upsilon_3 = (\Upsilon_1^\top \mathbf{H} + \mathbf{F}^\top \Upsilon_2)$ . It directly follows that

$$\|\Upsilon_3\|_{\text{op}} \leq \|\Upsilon_1\|_{\text{op}} \|\mathbf{H}\|_{\text{F}} + \|\mathbf{F}\|_{\text{F}} \|\Upsilon_2\|_{\text{op}}.$$

Using the triangle inequality and the moment bounds of  $\|\mathbf{H}\|_{\text{F}}$ ,  $\|\mathbf{F}\|_{\text{F}}$  in Lemma D.1 and the moment bounds of  $\|\Upsilon_1\|_{\text{op}}$ ,  $\|\Upsilon_2\|_{\text{op}}$  we just obtained, we have

$$\mathbb{E}[\|\Upsilon_3\|_{\text{op}}^2 \mid \varepsilon] \leq C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2).$$

**Bound of  $\Upsilon_4$ .** By the expression for the matrix  $\Upsilon_4 \in \mathbb{R}^{T \times T}$  obtained in Appendix D.4, we have

$$\Upsilon_4 = \mathbf{H}^\top \mathbf{X}^\top \Upsilon_2 - \tilde{\Upsilon}_1,$$

where  $\tilde{\Upsilon}_1 = \sum_i (\mathbf{I}_T \otimes \mathbf{e}_i^\top \mathbf{X}) \Gamma((\mathbf{F}^\top \mathbf{e}_i) \otimes \mathbf{I}_p) \mathbf{H}$ . We can rewrite  $\tilde{\Upsilon}_1$  as

$$\begin{aligned} \tilde{\Upsilon}_1 &= \sum_{i=1}^n (\mathbf{I}_T \otimes \mathbf{e}_i^\top \mathbf{X}) \Gamma((\mathbf{F}^\top \mathbf{e}_i) \otimes \mathbf{I}_p) \mathbf{H} \\ &= \sum_{i=1}^n \sum_{t=1}^T (\mathbf{I}_T \otimes \mathbf{e}_i^\top \mathbf{X}) \Gamma((\mathbf{e}_t \mathbf{e}_t^\top \mathbf{F}^\top \mathbf{e}_i) \otimes \mathbf{I}_p) \mathbf{H} \\ &= \sum_{i=1}^n \sum_{t=1}^T (\mathbf{I}_T \otimes (\mathbf{e}_t^\top \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_i^\top \mathbf{X})) \Gamma(\mathbf{e}_t \otimes \mathbf{I}_p) \mathbf{H} \\ &= \sum_{t=1}^T (\mathbf{I}_T \otimes (\mathbf{e}_t^\top \mathbf{F}^\top \mathbf{X})) \Gamma(\mathbf{e}_t \otimes \mathbf{I}_p) \mathbf{H}. \end{aligned}$$

We have by the triangle inequality,

$$\|\tilde{\Upsilon}_1\|_{\text{op}} \leq T \|\Gamma\|_{\text{op}} \|\mathbf{X}\|_{\text{op}} \|\mathbf{F}\|_{\text{F}} \|\mathbf{H}\|_{\text{F}}.$$

By the triangle inequality and the upper bound of  $\|\Upsilon_2\|_{\text{op}}$ , we have

$$\|\Upsilon_4\|_{\text{op}} \leq \|\Upsilon_2\|_{\text{op}} \|\mathbf{X}\|_{\text{op}} \|\mathbf{H}\|_{\text{F}} + \|\tilde{\Upsilon}_1\|_{\text{op}} \leq T \|\Gamma\|_{\text{op}} \|\mathbf{X}\|_{\text{op}}^2 \|\mathbf{H}\|_{\text{F}}^2 + T \|\Gamma\|_{\text{op}} \|\mathbf{X}\|_{\text{op}} \|\mathbf{F}\|_{\text{F}} \|\mathbf{H}\|_{\text{F}}.$$

Squaring both sides and taking conditional expectation, we have

$$\mathbb{E}[\|\Upsilon_4\|_{\text{op}}^2 \mid \varepsilon] \leq C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2),$$

thanks to the upper bound of  $\|\Gamma\|_{\text{op}}$  in Lemma D.4 and the moment bounds of  $\|\mathbf{X}\|_{\text{op}}$ ,  $\|\mathbf{F}\|_{\text{F}}$ ,  $\|\mathbf{H}\|_{\text{F}}$  in Lemma D.1.

**Bound of  $\Upsilon_5$ .** By the expression for the matrix  $\Upsilon_5 \in \mathbb{R}^{T \times T}$  obtained in Appendix D.4, we have  $\Upsilon_5 = \Upsilon_1^\top \mathbf{X}^\top \tilde{\mathbf{F}} - \tilde{\Upsilon}_2$ , where

$$\tilde{\Upsilon}_2 = - \sum_{t=1}^T \mathbf{F}^\top \mathbf{D}_t \mathbf{X} \mathbf{H} \mathbf{e}_t + \sum_{l=1}^n (\mathbf{e}_l^\top \mathbf{X} \mathbf{H} \otimes \mathbf{F}^\top) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \Gamma^\top (\mathbf{I}_T \otimes \mathbf{X}^\top) \mathcal{D}(\mathbf{I}_T \otimes \mathbf{e}_l).$$

By the triangle inequality,

$$\|\tilde{\Upsilon}_2\|_{\text{op}} \leq T \|\mathbf{X}\|_{\text{op}} \|\mathbf{F}\|_{\text{F}} \|\mathbf{H}\|_{\text{F}} + \|\mathbf{X}\|_{\text{op}}^3 \|\Gamma\|_{\text{op}} \|\mathbf{F}\|_{\text{F}} \|\mathbf{H}\|_{\text{F}}.$$

Using the moment bounds of  $\|\Upsilon_1\|_{\text{op}}$ ,  $\|\mathbf{X}\|_{\text{op}}$ ,  $\|\mathbf{F}\|_{\text{F}}$ ,  $\|\tilde{\mathbf{F}}\|_{\text{F}}$ ,  $\|\mathbf{H}\|_{\text{F}}$ , we have

$$\mathbb{E}[\|\Upsilon_5\|_{\text{op}}^2 \mid \varepsilon] \leq n^2 C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2).$$

## D.7 Proof of Lemma B.6

We first state three useful lemmas.

**Lemma D.6** (Adopted from Lemma E.10 of [26]). Let  $\mathbf{U}, \mathbf{V} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^{n \times T}$  be two locally Lipschitz functions of  $\mathbf{Z}$  with i.i.d.  $\mathcal{N}(0, 1)$  entries, then

$$\begin{aligned} & \mathbb{E} \left[ \left\| \mathbf{U}^\top \mathbf{Z} \mathbf{V} - \sum_{j=1}^p \sum_{i=1}^n \frac{\partial}{\partial z_{ij}} \left( \mathbf{U}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{V} \right) \right\|_{\text{F}}^2 \right] \\ & \leq \mathbb{E} \|\mathbf{U}\|_{\text{F}}^2 \|\mathbf{V}\|_{\text{F}}^2 + \mathbb{E} \sum_{ij} \left[ 2 \|\mathbf{V}\|_{\text{F}}^2 \left\| \frac{\partial \mathbf{U}}{\partial z_{ij}} \right\|_{\text{F}}^2 + 2 \|\mathbf{U}\|_{\text{F}}^2 \left\| \frac{\partial \mathbf{V}}{\partial z_{ij}} \right\|_{\text{F}}^2 \right]. \end{aligned}$$

**Lemma D.7** (Adopted from Lemma F.5 of [5]). Let  $\mathbf{U}, \mathbf{V} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^{n \times T}$  be two locally Lipschitz functions of  $\mathbf{Z}$  with i.i.d.  $\mathcal{N}(0, 1)$  entries. Provided the following expectations are finite, we have

$$\begin{aligned} & \mathbb{E} \left[ \left\| p \mathbf{U}^\top \mathbf{V} - \sum_{j=1}^p \left( \sum_{i=1}^n \partial_{ij} \mathbf{U}^\top \mathbf{e}_i - \mathbf{U}^\top \mathbf{Z} \mathbf{e}_j \right) \left( \sum_{i=1}^n \partial_{ij} \mathbf{e}_i^\top \mathbf{V} - \mathbf{e}_j^\top \mathbf{Z}^\top \mathbf{V} \right) \right\|_{\text{F}}^2 \right] \\ & \leq (1 + 2\sqrt{p}) (\mathbb{E} [\|\mathbf{U}\|_{\text{F}}^4]^{1/2} + \mathbb{E} [\|\mathbf{V}\|_{\text{F}}^4]^{1/2} + \mathbb{E} [\|\mathbf{U}\|_{\partial}^4]^{1/2} + \mathbb{E} [\|\mathbf{V}\|_{\partial}^4]^{1/2}), \end{aligned}$$

where  $\partial_{ij} \mathbf{U} = \partial \mathbf{U} / \partial z_{ij}$  and  $\|\mathbf{U}\|_{\partial} = (\sum_{i=1}^n \sum_{j=1}^p \|\partial_{ij} \mathbf{U}\|_{\text{F}}^2)^{1/2}$ .

We will use the above two lemmas, conditionally on  $\epsilon$ , to bound the conditional moments of  $\Theta_1, \Theta_2, \Theta_3, \Theta_4$  and  $\Theta_5$  given  $\epsilon$ .

**Bound of  $\Theta_1$ .** By the definition of  $\Theta_1$ , we have

$$\begin{aligned} \mathbf{F}^\top \mathbf{X} \mathbf{H} - \sum_{i,j} \frac{\partial \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} &= \mathbf{F}^\top \mathbf{X} \mathbf{H} + \widetilde{\mathbf{K}} \mathbf{H}^\top \mathbf{H} - \mathbf{F}^\top \mathbf{F} \mathbf{W}^\top + \Upsilon_3 \quad \text{by (20)} \\ &= \Theta_1 + \Upsilon_3. \end{aligned}$$

Applying Lemma D.6 conditionally on  $\epsilon$  to  $(\mathbf{Z}, \mathbf{U}, \mathbf{V}) = (\mathbf{X}, \mathbf{F}, \mathbf{H})$  gives

$$\begin{aligned} \mathbb{E} [\|\Theta_1\|_{\text{F}}^2 \mid \epsilon] &\lesssim \mathbb{E} \left[ \left\| \mathbf{F}^\top \mathbf{X} \mathbf{H} - \sum_{i,j} \frac{\partial \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} \right\|_{\text{F}}^2 \mid \epsilon \right] + \mathbb{E} [\|\Upsilon_3\|_{\text{F}}^2 \mid \epsilon] \\ &\lesssim \mathbb{E} [\|\mathbf{F}\|_{\text{F}}^2 \|\mathbf{H}\|_{\text{F}}^2 \mid \epsilon] + \mathbb{E} \sum_{ij} \left[ \|\mathbf{H}\|_{\text{F}}^2 \left\| \frac{\partial \mathbf{F}}{\partial x_{ij}} \right\|_{\text{F}}^2 + \|\mathbf{F}\|_{\text{F}}^2 \left\| \frac{\partial \mathbf{H}}{\partial x_{ij}} \right\|_{\text{F}}^2 \mid \epsilon \right] + \mathbb{E} [\|\Upsilon_3\|_{\text{F}}^2 \mid \epsilon] \\ &\leq nC(T, \gamma, \eta_{\max}, c_0) (\delta^2 + \|\mathbf{b}^*\|^2), \end{aligned}$$

where the last line uses the moment bounds of  $\|\mathbf{F}\|_{\text{F}}$ ,  $\|\mathbf{H}\|_{\text{F}}$  in Lemma D.1, the moment bounds of  $\|\frac{\partial \mathbf{H}}{\partial x_{ij}}\|_{\text{F}}$ ,  $\|\frac{\partial \mathbf{F}}{\partial x_{ij}}\|_{\text{F}}$  in Lemma D.5, the moment bound of  $\|\Upsilon_3\|_{\text{op}}$  in Lemma B.5.

**Bound of  $\Theta_2$ .** By the definition of  $\Theta_2$ , we have by (21)

$$\begin{aligned} \mathbf{F}^\top \mathbf{X} \mathbf{X}^\top \widetilde{\mathbf{F}} - \sum_{i,j} \frac{\partial \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \widetilde{\mathbf{F}}}{\partial x_{ij}} &= \mathbf{F}^\top \mathbf{X} \mathbf{X}^\top \widetilde{\mathbf{F}} + \widetilde{\mathbf{K}} \mathbf{H}^\top \mathbf{X}^\top \widetilde{\mathbf{F}} - p \mathbf{F}^\top \widetilde{\mathbf{F}} + \mathbf{F}^\top \mathbf{F} \widehat{\mathbf{A}}^\top + \Upsilon_5 \\ &= n\Theta_2 + \Upsilon_5. \end{aligned}$$

Applying Lemma D.6 conditionally on  $\epsilon$  to  $(\mathbf{Z}, \mathbf{U}, \mathbf{V}) = (\mathbf{X}, \mathbf{F}, \mathbf{X}^\top \widetilde{\mathbf{F}})$  gives

$$\begin{aligned} n^2 \mathbb{E} [\|\Theta_2\|_{\text{F}}^2 \mid \epsilon] &\lesssim \mathbb{E} \left[ \left\| \mathbf{F}^\top \mathbf{X} \mathbf{X}^\top \widetilde{\mathbf{F}} - \sum_{i,j} \frac{\partial \mathbf{F}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{X}^\top \widetilde{\mathbf{F}}}{\partial x_{ij}} \right\|_{\text{F}}^2 \mid \epsilon \right] + \mathbb{E} [\|\Upsilon_5\|_{\text{F}}^2 \mid \epsilon] \\ &\lesssim \mathbb{E} [\|\mathbf{F}\|_{\text{F}}^2 \|\mathbf{X}^\top \widetilde{\mathbf{F}}\|_{\text{F}}^2 \mid \epsilon] + \mathbb{E} \sum_{ij} \left[ \|\mathbf{X}^\top \widetilde{\mathbf{F}}\|_{\text{F}}^2 \left\| \frac{\partial \mathbf{F}}{\partial x_{ij}} \right\|_{\text{F}}^2 + \|\mathbf{F}\|_{\text{F}}^2 \left\| \frac{\partial \mathbf{X}^\top \widetilde{\mathbf{F}}}{\partial x_{ij}} \right\|_{\text{F}}^2 \mid \epsilon \right] + \mathbb{E} [\|\Upsilon_5\|_{\text{F}}^2 \mid \epsilon] \\ &\lesssim \mathbb{E} [\|\widetilde{\mathbf{F}}\|_{\text{F}}^4 \|\mathbf{X}\|_{\text{op}}^2 \mid \epsilon] + \mathbb{E} \left[ (1 + \|\mathbf{X}\|_{\text{op}}^2) \|\widetilde{\mathbf{F}}\|_{\text{F}}^2 \sum_{ij} \left\| \frac{\partial \widetilde{\mathbf{F}}}{\partial x_{ij}} \right\|_{\text{F}}^2 \mid \epsilon \right] + T \mathbb{E} [\|\Upsilon_5\|_{\text{op}}^2 \mid \epsilon] \\ &\leq n^3 C(T, \gamma, \eta_{\max}, c_0) (\delta^2 + \|\mathbf{b}^*\|^2). \end{aligned}$$

Here, the penultimate line uses  $\|\mathbf{F}\|_F \leq \|\tilde{\mathbf{F}}\|_F$ ,  $\|\frac{\partial \mathbf{F}}{\partial x_{ij}}\|_F \leq \|\frac{\partial \tilde{\mathbf{F}}}{\partial x_{ij}}\|_F$ , and  $\|\frac{\partial \mathbf{X}^\top \tilde{\mathbf{F}}}{\partial x_{ij}}\|_F = \|\mathbf{e}_j^\top \mathbf{e}_i^\top \frac{\partial \tilde{\mathbf{F}}}{\partial x_{ij}} + \mathbf{X}^\top \frac{\partial \tilde{\mathbf{F}}}{\partial x_{ij}}\|_F \leq \|\frac{\partial \tilde{\mathbf{F}}}{\partial x_{ij}}\|_F + \|\mathbf{X}\|_{\text{op}} \|\frac{\partial \tilde{\mathbf{F}}}{\partial x_{ij}}\|_F$ . The last line uses the moment bounds of  $\|\mathbf{X}\|_{\text{op}}$ ,  $\|\tilde{\mathbf{F}}\|_F$  in Lemma D.1, the moment bound of  $\|\frac{\partial \tilde{\mathbf{F}}}{\partial x_{ij}}\|_F$  in Lemma D.5, the moment bound of  $\|\Upsilon_5\|_{\text{op}}$  in Lemma B.5.

**Bound of  $\Theta_3$ .** By the definition of  $\Theta_3$ , we have

$$\begin{aligned} & \mathbf{H}^\top \mathbf{X}^\top \mathbf{X} \mathbf{H} - \sum_{i,j} \frac{\partial \mathbf{H}^\top \mathbf{X}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} \\ &= \mathbf{H}^\top \mathbf{X}^\top \mathbf{X} \mathbf{H} - (n\mathbf{I}_T - \tilde{\mathbf{A}}) \mathbf{H}^\top \mathbf{H} - \mathbf{H}^\top \mathbf{X}^\top \mathbf{F} \mathbf{W}^\top + \Upsilon_4 \quad \text{by (21)} \\ &= \Theta_3 + \Upsilon_4. \end{aligned}$$

Applying Lemma D.6 conditionally on  $\varepsilon$  to  $(\mathbf{Z}, \mathbf{U}, \mathbf{V}) = (\mathbf{X}, \mathbf{X} \mathbf{H}, \mathbf{H})$  gives

$$\begin{aligned} \mathbb{E}[\|\Theta_3\|_F^2 \mid \varepsilon] &\lesssim \mathbb{E}\left[\left\|\mathbf{H}^\top \mathbf{X}^\top \mathbf{X} \mathbf{H} - \sum_{i,j} \frac{\partial \mathbf{H}^\top \mathbf{X}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}}\right\|^2 \mid \varepsilon\right] + \mathbb{E}[\|\Upsilon_4\|_F^2 \mid \varepsilon] \\ &\lesssim \mathbb{E}[\|\mathbf{X} \mathbf{H}\|_F^2 \|\mathbf{H}\|_F^2 \mid \varepsilon] + \mathbb{E} \sum_{ij} \left[ \|\mathbf{X} \mathbf{H}\|_F^2 \left\|\frac{\partial \mathbf{H}}{\partial x_{ij}}\right\|_F^2 + \|\mathbf{H}\|_F^2 \left\|\frac{\partial \mathbf{X} \mathbf{H}}{\partial x_{ij}}\right\|_F^2 \mid \varepsilon \right] + \mathbb{E}[\|\Upsilon_4\|_F^2 \mid \varepsilon] \\ &\lesssim \mathbb{E}[\|\mathbf{X}\|_{\text{op}}^2 \|\mathbf{H}\|_F^4 \mid \varepsilon] + \mathbb{E}\left[(1 + \|\mathbf{X}\|_{\text{op}}^2) \|\mathbf{H}\|_F^2 \sum_{ij} \left\|\frac{\partial \mathbf{H}}{\partial x_{ij}}\right\|_F^2 \mid \varepsilon\right] + T \mathbb{E}[\|\Upsilon_4\|_{\text{op}}^2 \mid \varepsilon] \\ &\leq nC(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2). \end{aligned}$$

Here, the penultimate line uses  $\|\frac{\partial \mathbf{X} \mathbf{H}}{\partial x_{ij}}\|_F = \|\mathbf{e}_i \mathbf{e}_j^\top \frac{\partial \mathbf{H}}{\partial x_{ij}} + \mathbf{X} \frac{\partial \mathbf{H}}{\partial x_{ij}}\|_F \leq (1 + \|\mathbf{X}\|_{\text{op}}) \|\frac{\partial \mathbf{H}}{\partial x_{ij}}\|_F$ , and the last line uses the moment bounds of  $\|\mathbf{X}\|_{\text{op}}$ ,  $\|\mathbf{H}\|_F$  in Lemma D.1, the moment bounds of  $\|\frac{\partial \mathbf{H}}{\partial x_{ij}}\|_F$  in Lemma D.5, the moment bound of  $\|\Upsilon_4\|_{\text{op}}$  in Lemma B.5.

**Bound of  $\Theta_4$ .** By definition, we have

$$\sum_{i=1}^n \frac{\partial \mathbf{F}^\top \mathbf{e}_i}{\partial x_{ij}} = -(\tilde{\mathbf{K}} \mathbf{H}^\top + \Upsilon_1^\top) \mathbf{e}_j \text{ and } \sum_{i=1}^n \frac{\partial \tilde{\mathbf{F}}^\top \mathbf{e}_i}{\partial x_{ij}} = -(\widehat{\mathbf{K}} \mathbf{H}^\top + \tilde{\Upsilon}_1^\top) \mathbf{e}_j.$$

Using  $\sum_{j=1}^p \mathbf{e}_j \mathbf{e}_j^\top = \mathbf{I}_p$ , we find

$$\begin{aligned} & \sum_{j=1}^p \left( \sum_{i=1}^n \frac{\partial \mathbf{F}^\top \mathbf{e}_i}{\partial x_{ij}} - \mathbf{F}^\top \mathbf{X} \mathbf{e}_j \right) \left( \sum_{i=1}^n \frac{\partial \tilde{\mathbf{F}}^\top \mathbf{e}_i}{\partial x_{ij}} - \tilde{\mathbf{F}}^\top \mathbf{X} \mathbf{e}_j \right)^\top \\ &= (\tilde{\mathbf{K}} \mathbf{H}^\top + \mathbf{F}^\top \mathbf{X} + \Upsilon_1^\top) (\widehat{\mathbf{K}} \mathbf{H}^\top + \tilde{\mathbf{F}}^\top \mathbf{X} + \tilde{\Upsilon}_1^\top). \end{aligned}$$

This further implies

$$\begin{aligned} & p \mathbf{F}^\top \tilde{\mathbf{F}} - \sum_{j=1}^p \left( \sum_{i=1}^n \frac{\partial \mathbf{F}^\top \mathbf{e}_i}{\partial x_{ij}} - \mathbf{F}^\top \mathbf{X} \mathbf{e}_j \right) \left( \sum_{i=1}^n \frac{\partial \tilde{\mathbf{F}}^\top \mathbf{e}_i}{\partial x_{ij}} - \tilde{\mathbf{F}}^\top \mathbf{X} \mathbf{e}_j \right)^\top \\ &= n \Theta_4 + \underbrace{\Upsilon_1^\top (\widehat{\mathbf{K}} \mathbf{H}^\top + \tilde{\mathbf{F}}^\top \mathbf{X}) + (\tilde{\mathbf{K}} \mathbf{H}^\top + \mathbf{F}^\top \mathbf{X}) \tilde{\Upsilon}_1^\top + \Upsilon_1^\top \tilde{\Upsilon}_1^\top}_{\tilde{\Upsilon}_4}. \end{aligned}$$

Applying Lemma D.7 conditionally on  $\varepsilon$  to  $(\mathbf{Z}, \mathbf{U}, \mathbf{V}) = (\mathbf{X}, \mathbf{F}, \tilde{\mathbf{F}})$ ,

$$\begin{aligned} n \mathbb{E}[\|\Theta_4\|_F \mid \varepsilon] &\lesssim \mathbb{E}\left[\left\|p \mathbf{F}^\top \tilde{\mathbf{F}} - \sum_{j=1}^p \left( \sum_{i=1}^n \frac{\partial \mathbf{F}^\top \mathbf{e}_i}{\partial x_{ij}} - \mathbf{F}^\top \mathbf{X} \mathbf{e}_j \right) \left( \sum_{i=1}^n \frac{\partial \tilde{\mathbf{F}}^\top \mathbf{e}_i}{\partial x_{ij}} - \tilde{\mathbf{F}}^\top \mathbf{X} \mathbf{e}_j \right)^\top\right\|_F \mid \varepsilon\right] + \mathbb{E}[\|\tilde{\Upsilon}_4\|_F \mid \varepsilon] \\ &\lesssim (1 + 2\sqrt{p}) (\mathbb{E}[\|\mathbf{F}\|_F^4 \mid \varepsilon]^{1/2} + \mathbb{E}[\|\tilde{\mathbf{F}}\|_F^4 \mid \varepsilon]^{1/2} + \mathbb{E}[\|\mathbf{F}\|_\partial^4 \mid \varepsilon]^{1/2} + \mathbb{E}[\|\tilde{\mathbf{F}}\|_\partial^4 \mid \varepsilon]^{1/2}) + T \mathbb{E}[\|\tilde{\Upsilon}_4\|_{\text{op}}^2 \mid \varepsilon] \\ &\leq n^{3/2} C(T, \gamma, \eta_{\max}, c_0)(\delta^2 + \|\mathbf{b}^*\|^2). \end{aligned}$$

Here, the last inequality uses the moment bounds of  $\|\mathbf{F}\|_F$ ,  $\|\tilde{\mathbf{F}}\|_F$  in Lemma D.1, and the moment bounds of  $\|\mathbf{F}\|_\partial^2$ ,  $\|\tilde{\mathbf{F}}\|_\partial^2$  in Lemma D.5, and the moment bound of  $\|\tilde{\Upsilon}_4\|_{\text{op}}$  in Lemma B.5.

**Bound of  $\Theta_5$ .** By the identity (19), we have

$$\begin{aligned} & \sum_{i=1}^n \left( \sum_{j=1}^p \frac{\partial \mathbf{H}^\top \mathbf{e}_j}{\partial x_{ij}} - \mathbf{H}^\top \mathbf{X}^\top \mathbf{e}_i \right) \left( \sum_{j=1}^p \frac{\partial \mathbf{H}^\top \mathbf{e}_j}{\partial x_{ij}} - \mathbf{H}^\top \mathbf{X}^\top \mathbf{e}_i \right)^\top \\ &= (\mathbf{W} \mathbf{F}^\top - \mathbf{H}^\top \mathbf{X}^\top - \mathbf{\Upsilon}_2^\top) (\mathbf{W} \mathbf{F}^\top - \mathbf{H}^\top \mathbf{X}^\top - \mathbf{\Upsilon}_2^\top)^\top. \end{aligned}$$

By the definition of  $\Theta_5$ , we have

$$\begin{aligned} & n \mathbf{H}^\top \mathbf{H} - \sum_{i=1}^n \left( \sum_{j=1}^p \frac{\partial \mathbf{H}^\top \mathbf{e}_j}{\partial x_{ij}} - \mathbf{H}^\top \mathbf{X}^\top \mathbf{e}_i \right) \left( \sum_{j=1}^p \frac{\partial \mathbf{H}^\top \mathbf{e}_j}{\partial x_{ij}} - \mathbf{H}^\top \mathbf{X}^\top \mathbf{e}_i \right)^\top \\ &= \Theta_5 + \underbrace{\mathbf{\Upsilon}_2^\top (\mathbf{W} \mathbf{F}^\top - \mathbf{H}^\top \mathbf{X}^\top)^\top + (\mathbf{W} \mathbf{F}^\top - \mathbf{H}^\top \mathbf{X}^\top) \mathbf{\Upsilon}_2 - \mathbf{\Upsilon}_2^\top \mathbf{\Upsilon}_2}_{\tilde{\mathbf{\Upsilon}}_5}. \end{aligned}$$

Here  $\mathbf{\Upsilon}_2 = \sum_j ((\mathbf{e}_j^\top \mathbf{H}) \otimes \mathbf{I}_n) \mathcal{DS}(\mathbf{I}_T \otimes \mathbf{X}) \mathbf{\Gamma}^\top (\mathbf{I}_T \otimes \mathbf{e}_j)$ .

Applying Lemma D.7 conditionally on  $\varepsilon$  to  $(\mathbf{Z}, \mathbf{U}, \mathbf{V}) = (\mathbf{X}^\top, \mathbf{H}, \mathbf{H})$  (i.e., consider the mapping from  $\mathbb{R}^{p \times n}$  to  $\mathbb{R}^{p \times T}$ :  $\mathbf{X}^\top \mapsto \mathbf{H}$ ) gives

$$\begin{aligned} & \mathbb{E}[\|\Theta_5\|_F \mid \varepsilon] \\ & \leq \mathbb{E} \left[ \left\| n \mathbf{H}^\top \mathbf{H} - \sum_{i=1}^n \left( \sum_{j=1}^p \frac{\partial \mathbf{H}^\top \mathbf{e}_j}{\partial x_{ij}} - \mathbf{H}^\top \mathbf{X}^\top \mathbf{e}_i \right) \left( \sum_{j=1}^p \frac{\partial \mathbf{H}^\top \mathbf{e}_j}{\partial x_{ij}} - \mathbf{H}^\top \mathbf{X}^\top \mathbf{e}_i \right)^\top \right\|_F \mid \varepsilon \right] + \mathbb{E}[\|\tilde{\mathbf{\Upsilon}}_5\|_F \mid \varepsilon] \\ & \lesssim (1 + 2\sqrt{p}) (\mathbb{E}[\|\mathbf{H}\|_F^4 \mid \varepsilon]^{1/2} + \mathbb{E}[\|\mathbf{H}\|_\partial^4 \mid \varepsilon]^{1/2}) + T \mathbb{E}[\|\tilde{\mathbf{\Upsilon}}_5\|_{\text{op}}^2 \mid \varepsilon] \\ & \leq n^{1/2} C(T, \gamma, \eta_{\max}, c_0) (\delta^2 + \|\mathbf{b}^*\|^2). \end{aligned}$$

Here, the last line use the moment bounds of  $\|\mathbf{H}\|_F$  in Lemma D.1, the moment bound of  $\|\mathbf{H}\|_\partial^2$  in Lemma D.5, and the moment bound of  $\|\tilde{\mathbf{\Upsilon}}_5\|_{\text{op}}$  in Lemma B.5.

**Bound of  $\Theta_6$ .** Using (19), we have

$$\sum_{i=1}^n \sum_{j=1}^p \frac{\partial \mathbf{E}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} = \mathbf{E}^\top (\mathbf{F} \mathbf{W}^\top - \mathbf{\Upsilon}_2).$$

It follows that

$$\begin{aligned} \mathbf{E}^\top \mathbf{X} \mathbf{H} - \sum_{i=1}^n \sum_{j=1}^p \frac{\partial \mathbf{E}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} &= \mathbf{E}^\top \mathbf{X} \mathbf{H} - \mathbf{E}^\top (\mathbf{F} \mathbf{W}^\top - \mathbf{\Upsilon}_2) \\ &= \|\mathbf{E}\|_F \Theta_6 + \mathbf{E}^\top \mathbf{\Upsilon}_2. \end{aligned}$$

Thus, using  $\tilde{\mathbf{E}} = \mathbf{E} / \|\mathbf{E}\|_F$ , we have

$$\Theta_6 = \tilde{\mathbf{E}}^\top \mathbf{X} \mathbf{H} - \sum_{i=1}^n \sum_{j=1}^p \frac{\partial \tilde{\mathbf{E}}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} - \tilde{\mathbf{E}}^\top \mathbf{\Upsilon}_2.$$

Applying Lemma D.6 conditionally on  $\varepsilon$  to  $(\mathbf{Z}, \mathbf{U}, \mathbf{V}) = (\mathbf{X}, \tilde{\mathbf{E}}, \mathbf{H})$  gives

$$\begin{aligned} \mathbb{E}[\|\Theta_6\|_F^2 \mid \varepsilon] & \lesssim \mathbb{E} \left[ \left\| \tilde{\mathbf{E}}^\top \mathbf{X} \mathbf{H} - \sum_{i=1}^n \sum_{j=1}^p \frac{\partial \tilde{\mathbf{E}}^\top \mathbf{e}_i \mathbf{e}_j^\top \mathbf{H}}{\partial x_{ij}} \right\|^2 \mid \varepsilon \right] + \mathbb{E}[\|\tilde{\mathbf{E}}^\top \mathbf{\Upsilon}_2\|_F^2 \mid \varepsilon] \\ & \lesssim \mathbb{E}[\|\tilde{\mathbf{E}}\|_F^2 \|\mathbf{H}\|_F^2 \mid \varepsilon] + \mathbb{E} \left[ \left\| \tilde{\mathbf{E}} \right\|_F^2 \left\| \frac{\partial \mathbf{H}}{\partial x_{ij}} \right\|_F^2 \mid \varepsilon \right] + \mathbb{E}[\|\tilde{\mathbf{E}}^\top \mathbf{\Upsilon}_2\|_F^2 \mid \varepsilon] \\ & \leq \mathbb{E}[\|\mathbf{H}\|_F^2 \mid \varepsilon] + \mathbb{E} \left[ \left\| \frac{\partial \mathbf{H}}{\partial x_{ij}} \right\|_F^2 \mid \varepsilon \right] + \mathbb{E}[\|\mathbf{\Upsilon}_2\|_{\text{op}}^2 \mid \varepsilon] \\ & \leq C(T, \gamma, \eta_{\max}, c_0) (\delta^2 + \|\mathbf{b}^*\|^2). \end{aligned}$$

Here, the last line uses the moment bound of  $\|\mathbf{H}\|_F$  in Lemma D.1, the moment bounds of  $\|\frac{\partial \mathbf{H}}{\partial x_{ij}}\|_F$  in Lemma D.5, and the moment bound of  $\|\mathbf{\Upsilon}_2\|_{\text{op}}$  in Lemma B.5. This finishes the proof of Lemma B.6.



## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction clearly state the claims made in the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Our results only holds under several assumptions discussed in the paper.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: All the proofs are provided in the supplemental material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All the code needed to reproduce the results is in the supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All the code are provided in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The choices of step size for GD and SGD are provided in Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All the figures include the 2-standard error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: It is provided in the README file of the supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Given the theoretical nature of our work, it does not have any negative societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper is of the theoretical nature, it does not invent any new models or datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

#### 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We write all the code by ourselves, and did not use other code or data.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.