OPT2020: 12th Annual Workshop on Optimization for Machine Learning

Reduced-Memory Kalman Based Stochastic Gradient Descent

Jinyi Wang JWANG2242@WISC.EDU
Vivak Patel VIVAK.PATEL@WISC.EDU
University of Wisconsin, Madison, WI, United States

Abstract

We develop a limited-memory method for online linear regression problems named the reduced Kalman-based Stochastic Gradient Descent. Limited-memory methods are intriguing since they address the computational challenges of second order methods and show more robustness to different parameter choices than first order methods. However, current limited-memory methods, such as adaption of L-BFGS to the stochastic case, suffer from one or more of the following drawbacks: (1) assumptions made on objective functions that rule out least squares problems; (2) samples with size larger than the dimensional of problem required when estimating Hessian; (3) less robustness due to new hyper-parameters introduced. Moreover, all the methods do not incorporate the latest observation in the Hessian estimate for the next iterate due to the lack of a relevant proof technique. The standard approach used in current methods requires a conditional independence between the Hessian estimate and the gradient estimate. We tackle those problems by introducing a new limited-memory method that avoids those drawbacks by construction. We give the theoretical guarantee and experimentally demonstrate our method. Importantly, in our method, we fully exploit the up-to-the-moment information in the Hessian estimate and the gradient estimate. We develop a new analysis strategy that allows us to study the convergence when the Hessian and gradient are dependent. Furthermore, this strategy can be adapted for a series of procedures that include dependent Hessian and gradient estimate since no specialties of least squares problems are utilized in our analysis scheme.

1. Introduction

Least squares problems continue to be an intensive area of research especially as the number of equations and dimension of the problem grow with advances in sensor technology [e.g., 13] and developments in higher fidelity models by domain experts [e.g., see Ch. 2 of 2]. In addition, the data for least square problems are generated continuously in many cases, which results in streams of data [1, 13, 14]. To better handle the increasing data size and possible demand of adaptation to streaming model, we reformulate the least squares problem as a statistical estimation problem,

$$\min_{\beta} \mathbb{E}\left[(Y - X'\beta)^2\right], \tag{1}$$

where Y is the resulting random variable corresponding to the elements of the dependent vector of the least squares problem; X is the resulting random vector of dimension p corresponding to the rows of coefficient matrix of the least squares problem and X' is the transpose of X; β is the unknown parameter of dimension p; and $\mathbb{E}$ is the expectation operator over the distribution placed on X and Y. Such a reformulation allow us to employ stochastic approximation methods for solving least square problems.

Among the current stochastic approximation methods, first order methods have very low computational complexity and memory requirements, but their convergence behavior is highly sensitive to the curvature of the objective near the solution [4]. Second order methods, in comparison to first order methods, have been empirically shown to be highly robust to step-size selections (usually a step size of one) [16], while incur a much higher computational cost per iteration and overall storage complexity, which becomes prohibitive as the dimension of the parameter, p, increases.

To balance between computational efficiency and robustness to parameters, limited-memory stochastic approximation methods are an intensive area of research [3, 5, 8, 9, 12, 15, 17, 18]. Current limited-memory stochastic approximation methods are based on deterministic L-BFGS [3, 5, 8, 9, 12, 15, 17, 18]. The adaptations of L-BFGS to stochastic context result in a number of benefits or drawbacks, especially for the linear regression problem: (1) assumptions that preclude the linear regression problem [3, 15, 18]; (2) requirements for a sample size larger than the dimension of the problem for each Hessian update [5, 8, 9, 12]; (3) introducing new hyper-parameters that reduce the robustness of the method [17]. Moreover, all stochastic analogues to L-BFGS do not fully exploit the most recent information in generating the next iterate since they require the conditional independence between the Hessian-estimate and gradient for analysis in their setting. Note, this idiosyncrasy appears to be driven by analysis purposes: the aforementioned methods directly adapt the standard approach for analyzing stochastic approximation methods (e.g., [6]), which depends on the conditional independence between the Hessian-estimate and the gradient-estimate.

For the linear regression problem, rather than finding a general purpose, stochastic extension of L-BFGS, an alterantive, tailored approach is to extend Gauss-Newton (see Ch. 10 of [10]). One of the stochastic extension of Gauss-Newton is named Kalman-based Stochastic Gradient Descent, the convergence of which is robust to the choice of tuning parameters [11]. In this work, we will extend this related Gauss-Newton method to the limited memory case for the online linear regression problem under the name reduced Kalman-based Stochastic Gradient Descent. Importantly, our resulting method makes use of the most recently available information in the Hessian estimate and in the gradient estimate. We develop an analysis strategy that allows us to deal with the dependence between the Hessian estimate and gradient estimate. Fortunately, our analysis strategy does not depend on the properties of the linear regression problem, and can be readily adapted to analyzing procedures that induce dependence between the Hessian estimate and gradient estimate (e.g., AdaGrad, ADAM, etc.). To summarize,

1. We develop a novel limited-memory method that is based on Gauss-Newton and that avoids the difficulties associated with stochastic L-BFGS methods;
2. We develop a novel analysis strategy that can be leveraged to analyze a variety of methods that induce dependence between the Hessian estimate and the gradient estimate.

2. Method & Algorithms

In this section, we introduce how we derive a reduced memory version of Kalman-based Stochastic Gradient Descent (k-SGD). The full k-SGD is derived by choosing optimal step-size and estimated inverse hessian to minimize the mean square error conditioned on $\mathcal{F}_{k+1} = \sigma(X_1, X_2, \ldots, X_{k+1})$ [11].A pseudo-code of k-SGD is summarized in Algorithm 1. Actually, the recursive formula for k-SGD is equivalent to (2), and therefore and be viewed as an adaption of Gauss-Newton method to the stochastic case. More details can be found in [11].

$$\beta_{k+1} = \beta_k + \left(I_p + \frac{1}{\gamma} \sum_{i=1}^{k+1} X_i' X_i \right)^{-1} X_{k+1} \left(Y_{k+1} - \beta_k' X_{k+1} \right). \tag{2}$$

The advantages of k-SGD include the insensitivity to conditioning of Hessian and robustness to the choice of tuning parameters. However, for large dimension problem, it is expensive to store the whole matrix N_k. What's more, the matrix-vector product in Line 6 shows to be time-consuming in the numerical experiments when dimension p is large. Therefore, we construct a reduced memory version of k-SGD for large scale problems. Similar to the L-BFGS procedure, we construct an approximation of the matrix N_k using only a user-defined number of vectors, m. We first obtain a recursive formula to calculate d_k with $\{d_{k-m+1}, \ldots, d_{k-1}\}$ and N_{k-m}, and then introduce a diagonal approximation of N_{k-m}. Our method initializes C_0 as the identity matrix and β_0 as an arbitrary vector, and determines $\{C_k\}$ and $\{\beta_k\}$ by

$$v_{k-(m\wedge k)+j}^{(k+1)} = \left[\overset{\curvearrowleft}{\prod_{s=1}^{j-1}} \left(I_p - \frac{v_{k-(m\wedge k)+s}^{(k+1)} X_{k-m+s}'}{\gamma + \left(v_{k-(m\wedge k)+s}^{(k+1)} \right)' X_{k-m+s}} \right) \right] C_k X_{k-m+j}, \tag{3}$$

$$\beta_{k+1} = \beta_k + \frac{\eta}{\gamma + X_{k+1}' v_{k+1}^{(k+1)}} v_{k+1}^{(k+1)} \left(Y_{k+1} - \beta_k' X_{k+1} \right), \tag{4}$$

$$C_{k+1} = \begin{cases} I_p & k < m \\ \left(C_k^{-1} + \operatorname{diag}\left\{ X_{k-m+1} X_{k-m+1}' \right\} \right)^{-1} & k \geq m \end{cases}, \tag{5}$$

for $j = 1, \ldots, (m \wedge k) + 1$, where $\overset{\curvearrowleft}{\prod}$ represent left multiplication. Here, C_k is the diagonal approximation of N_{k-m}; $\{v_{k-m+1}^{(k+1)}, \ldots, v_{k+1}^{(k+1)}\}$ are the approximated searching directions generated at step k; $\eta > 0$ is a tuning parameter that accounts for the reduced memory approximation to the Hessian estimate.

Algorithm 1: Full k-SGD

Input: Parameter β_0, Hyper-parameter γ
Output: Parameter β

1 $\beta \leftarrow \beta_0$;
2 $N_0 \leftarrow p \times p$ identity matrix ;
3 $k \leftarrow 0$;
4 **while** *true* **do**
5 Read new observation (X_{k+1}, Y_{k+1}) ;
6 $d_{k+1} \leftarrow N_k X_{k+1}$;
7 $s \leftarrow \gamma + d_{k+1}' X_{k+1}$;
8 $\beta \leftarrow \beta + d_{k+1}(Y_{k+1} - \beta' X_{k+1})/s$;
9 $N_{k+1} \leftarrow N_k - d_{k+1} d_{k+1}'/s$;
10 $k \leftarrow k + 1$;
11 **end**

Algorithm 2: Reduced k-SGD

Input: Parameter β_0, Hyper-parameter γ
Output: Parameter β

1 $\beta \leftarrow \beta_0$;
2 $c \leftarrow p \times 1$ array of 1's ;
3 $k \leftarrow 0$;
4 **while** *true* **do**
5 Read new observation (X_{k+1}, Y_{k+1}) ;
6 Compute v_{k+1} by (3) using c and $\{X_{(k-m+1)\vee 1}, \ldots, X_{k+1}\}$;
7 $s \leftarrow \gamma + v_{k+1}' X_{k+1}$;
8 $\beta \leftarrow \beta + \eta v_{k+1}(Y_{k+1} - \beta' X_{k+1})/s$;
9 **if** $k \geq m$ **then**
10 $c \leftarrow \left(c^{-1} + X_{k-m+1}^2 \right)^{-1}$
11 **end**
12 $k \leftarrow k + 1$;
13 **end**

3. Analysis of Convergence

In this section, we prove the convergence rate of our reduced k-SGD method (see Theorem 2). As mentioned in §1, the standard convergence analysis (e.g., [6]) failed since we have dependent Hessian estimated and gradient estimated.[1] The alternative approach in [11] is also not applicable for our method since it requires certain relationships to persist between the initial Hessian estimate through the terminal Hessian estimate.

Therefore, we innovate an approach that can readily handle arbitrary dependencies between the Hessian estimate and gradient estimate. Our new approach has two steps. In the first step, we introduce a "ghost" estimator which replaces the random Hessian estimate with a carefully constructed deterministic quantity. Then, we apply the standard analysis to get the convergence of this "ghost" estimator. In the second step, we analyze the iterates generated by reduced k-SGD against the iterates generated by the ghost estimator, which effectively requires us to compare the random Hessian estimate to the aforementioned carefully constructed deterministic quantity.

We first introduce the assumptions we make for analysis. Assumption 1 formulates the least squares problem. Assumption 2 ensures that (1) is well-defined and has a unique minimizer. Assumption 3 controls the density of $\{X_k\}$.

Assumption 1 *Suppose we have independent identical distributed pairs* $(X_1, Y_1), (X_2, Y_2), \ldots \in \mathbb{R}^p \times \mathbb{R}$*, which have the linear relationship* $Y_k = X_k'\beta_* + \varepsilon_k$*, where* $\beta_* \in \mathbb{R}^p$ *is a fixed vector,* ε_k*'s satisfy* $\mathbb{E}\left[\varepsilon_k | X_k\right] = 0$ *and* $\mathbb{V}\left[\epsilon_k | X_k\right] = \sigma^2 > 0$*.*

Assumption 2 *Assume that* $Q_* = \mathbb{E}[X_1 X_1']$ *exists and* $0 \prec Q_* \prec \infty$*.*

Assumption 3 *Suppose for any* $\{j_1, \ldots, j_s\} \subset \{1, \ldots, p\}$*, and* $\{t_1, \ldots, t_s\}$ *such that* $\sum_{i=1}^s t_i \leq 8$ *we have*

$$\mathbb{E}\left[\prod_{i=1}^{s} (X_{1,j_i})^{t_i}\right] < \infty. \tag{6}$$

We define a ghost estimator $\{\theta_k\}$ by first write out the explicit form of recursive formula (4) and then replace the Hessian estimate by its expectation. The update formula for $\{\theta_k\}$ is

$$\theta_{k+1} = \theta_k + \frac{\eta}{\gamma} J_{Q,k+1}^{-1} X_{k+1} \left(Y_{k+1} - X_{k+1}'\theta_k\right), \tag{7}$$

where the hessian estimate $J_{Q,k+1}$ is defined as

$$J_{Q,k+1} = \begin{cases} I_p + \dfrac{1+k}{\gamma} Q_* & k \leq m \\ I_p + \dfrac{1}{\gamma}\left[(k-m)Q_d + (m+1)Q_*\right] & k > m \end{cases}, \tag{8}$$

where $Q_d = \text{diag}\{Q_*\}$. We show that the "ghost" estimator converges with rate $O(1/k)$.

1. Recall that for stochastic L-BFGS methods, the Hessian estimate and gradient estimate are designed to be conditionally independent; see §1.

Theorem 1 *Suppose that Assumption 1 and Assumption 2 hold. If* $\eta > \frac{\lambda_{\max}(Q_d)}{2\lambda_{\min}(Q_*)}$,

$$\mathbb{E}\left[\|\theta_k - \beta_*\|^2_{Q_*}\right] = O\left(\frac{1}{k}\right), \tag{9}$$

where $\|x\|^2_{Q_*} = x^T Q_* x$.

We can then show the difference between two estimators convergences to 0 with rate $O(1/k)$.

Theorem 2 *With assumption 1, 2 and 3, if the step-size* $\eta > \frac{\lambda_{\max}(Q_d)}{2\lambda_{\min}(Q_*)}$, *then for arbitrary* $\delta > 0$, *there exists a sequence of events* $\{D_k\}_{k=K}^{\infty}$ *satisfies* $D_{k+1} \subset D_k$ *and for* $k \geq K$, $\lim_{k\to\infty} P(D_k) \geq 1 - \delta$, *such that*

$$\mathbb{E}\left[\|\theta_k - \beta_k\|^2_{Q_*} 1_{D_{k-1}}\right] = O\left(\frac{1}{k}\right). \tag{10}$$

4. Numerical Experiments

In this section we are going to compare the behavior of reduced k-SGD with SGD[6], full k-SGD, AdaGrad[7], and the stochastic quasi Newton method [5] on a BlogFeedback data set. All the methods are initialized with 0. We record the number of accessed data points (ADP), elapsed time and the the mean of the residuals squared (MRS). The behaviors are measured with two metrics, the ADP and the time needed to get the same MRS.

BlogFeedback Data Set contains the processed features of the selected blog posts. The goal is to predict the the number of comments in the upcoming 24 hours. We apply a Haar waveket model. The resolutions of all features is set to be 2, which resulted in a parameter of dimension $p = 1960$. From Figure 1, our reduced k-SGD has outperformed SGD, AdaGrad and online L-BFGS. The full k-SGD goes into wrong direction in the first few steps, but then returns to the right track and converges faster than all the other methods when the ADP is large. The online L-BFGS stops making progress early due to some zero gradients in this data set. When turns Figure 2, reduced k-SGD with $m = 1$ still shows good performance. However, the running time of k-SGD with $m = 5$ grows a lot. The total runnning time for full k-SGD is more than 1400 seconds and is not shown in this figure.

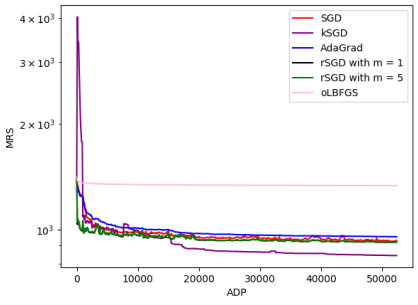

Figure 1: ADP versus Loss

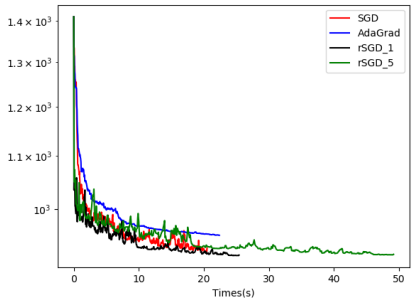

Figure 2: Time versus Loss

5. Conclusion and Future Goals

In this paper, we developed reduced k-SGD method for least squares problems and analyzed the convergence. As a result, our method achieves a convergence rate $O(1/k)$ and avoids the awkwardness of stochastic L-BFGS. That is to say, the assumptions for our method work perfectly for least squares problems; no extra sub-sampling is needed to ensure convergence. Importantly, our method is able to exploit the latest observation in the Hessian estimation with developing a innovative analysis strategy. This strategy is capable of Hessian and gradient estimate that are not conditional independent, in which case the classical analysis strategy has broken down. Furthermore, our method can be readily adapted to a bunch of methods with dependent Hessian estimate and gradient estimate.

There are several things left for the future,

1. show that Theorem 2 convergence with probability 1;
2. construct an efficient implementation of Equation (3);
3. demonstrate the performance of our method on more numerical experiments.

References

[1] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting flu trends using twitter data. In *2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS)*, pages 702–707. IEEE, 2011.

[2] Lorenz Biegler, George Biros, Omar Ghattas, Matthias Heinkenschloss, David Keyes, Bani Mallick, Luis Tenorio, Bart Van Bloemen Waanders, Karen Willcox, and Youssef Marzouk. *Large-scale inverse problems and quantification of uncertainty*, volume 712. Wiley Online Library, 2011.

[3] Raghu Bollapragada, Dheevatsa Mudigere, Jorge Nocedal, Hao-Jun Michael Shi, and Ping Tak Peter Tang. A progressive batching l-bfgs method for machine learning. *arXiv preprint arXiv:1802.05374*, 2018.

[4] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[5] Richard H Byrd, Samantha L Hansen, Jorge Nocedal, and Yoram Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

[6] Kai Lai Chung et al. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3):463–483, 1954.

[7] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul):2121–2159, 2011.

[8] Aryan Mokhtari and Alejandro Ribeiro. Global convergence of online limited memory bfgs. *The Journal of Machine Learning Research*, 16(1):3151–3181, 2015.

[9] Philipp Moritz, Robert Nishihara, and Michael Jordan. A linearly-convergent stochastic l-bfgs algorithm. In *Artificial Intelligence and Statistics*, pages 249–258, 2016.

[10] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[11] Vivak Patel. Kalman-based stochastic gradient method with stop condition and insensitivity to conditioning. *SIAM Journal on Optimization*, 26(4):2620–2648, 2016.

[12] Nicol N Schraudolph, Jin Yu, and Simon Günter. A stochastic quasi-newton method for online convex optimization. In *Artificial intelligence and statistics*, pages 436–443, 2007.

[13] Edmund O Schweitzer, David Whitehead, Armando Guzman, Yanfeng Gong, and Marcos Donolo. Advanced real-time synchrophasor applications. In *proceedings of the 35th Annual Western Protective Relay Conference, Spokane, WA*, 2008.

[14] Joel Swendsen, Dror Ben-Zeev, and Eric Granholm. Real-time electronic ambulatory monitoring of substance use and symptom expression in schizophrenia. *American Journal of Psychiatry*, 168(2):202–209, 2011.

[15] Xiao Wang, Shiqian Ma, Donald Goldfarb, and Wei Liu. Stochastic quasi-newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 27(2):927–956, 2017.

[16] Peng Xu, Fred Roosta, and Michael W Mahoney. Second-order optimization for non-convex machine learning: An empirical study. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 199–207. SIAM, 2020.

[17] Farzad Yousefian, Angelia Nedic, and Uday V Shanbhag. On stochastic and deterministic quasi-newton methods for nonstrongly convex optimization: Asymptotic convergence and rate analysis. *SIAM Journal on Optimization*, 30(2):1144–1172, 2020.

[18] Renbo Zhao, William Benjamin Haskell, and Vincent YF Tan. Stochastic l-bfgs: Improved convergence rates and practical acceleration strategies. *IEEE Transactions on Signal Processing*, 66(5):1155–1169, 2018.

Appendix A. Explict update formula for reduced k-SGD

Lemma 3 *With $C_0 = I_p$ and the same initial point β_0, the updating rule (3), (4), and (5) are equivalent to*

$$J_{C,k+1} = C_k^{-1} + \frac{1}{\gamma} \sum_{j=1}^{(m\wedge k)+1} X_{k-(m\wedge k)+j} X'_{k-(m\wedge k)+j}, \tag{11}$$

$$\beta_{k+1} = \beta_k + \frac{\eta}{\gamma} J_{C,k+1}^{-1} X_{k+1}(Y_{k+1} - X'_{k+1}\beta_k), \tag{12}$$

$$C_{k+1} = \begin{cases} I_p & k < m \\ \left(C_k^{-1} + \text{diag}\{X_{k-m+1}X'_{k-m+1}\}\right)^{-1} & k \geq m \end{cases}. \tag{13}$$

Proof We only prove for the case $k \geq m$. We first work out the explicit form of $v_{k+1}^{(k+1)}$ by substituting previous $v_{k-m+j}^{(k+1)}$'s. When $j = 1$, $v_{k-m+1}^{(k+1)} = C_k X_{k-m+1}$. Then for $j = 2$, we will have

$$\begin{aligned} v_{k-m+2}^{(k+1)} &= \left(I_p - \frac{v_{k-m+1}^{(k+1)} X'_{k-m+1}}{\gamma + \left(v_{k-m+1}^{(k+1)}\right)' X_{k-m+1}} \right) C_k X_{k-m+2} \\ &= \left(C_k - \frac{C_k X_{k-m+1} X'_{k-m+1} C_k}{\gamma + X'_{k-m+1} C_k X_{k-m+1}} \right) X_{k-m+2} \\ &= \left(C_k^{-1} + \frac{1}{\gamma} X_{k-m+1} X'_{k-m+1} \right)^{-1} X_{k-m+2}. \end{aligned} \tag{14}$$

The last equality follows the Sherman-Morrison formula. Actually, if we repeatedly do the same procedure for $j = 2, \ldots, m+1$, we will get an explicit formula for $v_{k-m+j}^{(k+1)}$ with adding more $X_{k-m+j}X_{k-m+j}$'s inside the parentheses. We can prove this by induction. We denote

$$\begin{aligned} M_{k-m+1}^{(k+1)} &= C_k \\ M_{k-m+j+1}^{(k+1)} &= \left[\left(M_{k-m+j}^{(k+1)} \right)^{-1} + X_{k-m+j} X'_{k-m+j} \right]^{-1}, \end{aligned} \tag{15}$$

for $j = 1, \ldots, m$. We claim that

$$v_{k-m+j}^{(k+1)} = M_{k-m+j}^{(k+1)} X_{k-m+j}. \tag{16}$$

for $j = 1, \ldots, m+1$. We have already prove the base case $j = 1$. Now we assume that for $j \leq l$, (16) holds. Then we focus on the case $j = l+1$. Again, we denote

$$q_{k-m+l+1}^{\langle r \rangle} = \left[\prod_{s=1}^{\overset{\curvearrowleft}{r}} \left(I_p - \frac{v_{k-m+s}^{(k+1)} X'_{k-m+s}}{\gamma + \left(v_{k-m+s}^{(k+1)}\right)' X_{k-m+s}} \right) \right] C_k X_{k-m+l+1}, \tag{17}$$

for $r = 0, \ldots, l$. Here we use a second induction with respect to the index r. We claim that

$$q_{k-m+l+1}^{\langle r \rangle} = M_{k-m+r+1}^{(k+1)} X_{k-m+l+1}. \tag{18}$$

The base case $r = 0$ follows the definition of $q_{k-m+l+1}^{\langle 0\rangle}$ and $M_{k-m+1}^{(k+1)}$. We now assume that for $r = t-1$, $1 \le t \le l$,

$$q_{k-m+l+1}^{\langle t-1\rangle} = M_{k-m+t}^{(k+1)} X_{k-m+l+1}. \tag{19}$$

Then for $r = t$,

$$\begin{aligned} q_{k-m+l+1}^{\langle t\rangle} &= \left(I_p - \frac{v_{k-m+t}^{(k+1)} X_{k-m+t}'}{\gamma + \left(v_{k-m+t}^{(k+1)}\right)' X_{k-m+t}} \right) q_{k-m+l+1}^{\langle t-1\rangle} \\ &= \left(I_p - \frac{v_{k-m+t}^{(k+1)} X_{k-m+t}'}{\gamma + \left(v_{k-m+t}^{(k+1)}\right)' X_{k-m+t}} \right) M_{k-m+t}^{(k+1)} X_{k-m+l+1}. \end{aligned} \tag{20}$$

Substituting the first induction hypothesis (16) and using Sherman-Morrison formula,

$$\begin{aligned} q_{k-m+l+1}^{\langle t\rangle} &= \left(I_p - \frac{M_{k-m+t}^{(k+1)} X_{k-m+t} X_{k-m+t}'}{\gamma + X_{k-m+t}' M_{k-m+t}^{(k+1)} X_{k-m+t}} \right) M_{k-m+1}^{(k+1)} X_{k-m+l+1} \\ &= \left(I_p - \frac{M_{k-m+t}^{(k+1)} X_{k-m+t} X_{k-m+t}' M_{k-m+1}^{(k+1)}}{\gamma + X_{k-m+t}' M_{k-m+t}^{(k+1)} X_{k-m+t}} \right) X_{k-m+l+1} \\ &= \left[\left(M_{k-m+t}^{(k+1)}\right)^{-1} + X_{k-m+t} X_{k-m+t}' \right]^{-1} X_{k-m+l+1} \\ &= M_{k-m+t+1}^{(k+1)} X_{k-m+l+1}, \end{aligned} \tag{21}$$

which implies (18) holds when $r = t$. Therefore the second claim holds for all $r = 1, \ldots, l$. When $r = l$,

$$v_{k-m+l+1} = q_{k-m+l+1}^{\langle l\rangle} = M_{k-m+l+1}^{(l+1)} X_{k-m+l+1}. \tag{22}$$

Therefore, (16) holds for the case $j = l+1$, which completes outer induction reasoning. After substituting in the explicit formula of $v_{k+1}^{(k+1)}$, the updating formula for β_{k+1} now becomes

$$\beta_{k+1} = \beta_k + \frac{\eta}{\gamma + X_{k+1}' M_{k+1}^{(k+1)} X_{k+1}} M_{k+1}^{(k+1)} X_{k+1} \left(Y_{k+1} - \beta_k' X_{k+1}\right), \tag{23}$$

We can still simplify this formula by applying Sherman-Morrison formula again,

$$\begin{aligned} \beta_{k+1} &= \beta_k + \frac{\eta}{\gamma} \left[M_{k+1}^{(k+1)} - \frac{M_{k+1}^{(k+1)} X_{k+1} X_{k+1}' M_{k+1}^{(k+1)}}{\gamma + X_{k+1}' M_{k+1}^{(k+1)} X_{k+1}} \right] X_{k+1} \left(Y_{k+1} - X_{k+1}' \beta_k\right) \\ &= \beta_k + \frac{\eta}{\gamma} \left[\left(M_{k+1}^{(k+1)}\right)^{-1} + \frac{1}{\gamma} X_{k+1} X_{k+1}' \right]^{-1} X_{k+1} \left(Y_{k+1} - X_{k+1}' \beta_k\right) \\ &= \beta_k + \frac{\eta}{\gamma} \left[C_k^{-1} + \frac{1}{\gamma} \sum_{j=1}^{m+1} X_{k-m+j} X_{k-m+j}' \right]^{-1} X_{k+1} \left(Y_{k+1} - X_{k+1}' \beta_k\right) \\ &= \beta_k + \frac{\eta}{\gamma} J_{C,k+1}^{-1} X_{k+1} (Y_{k+1} - X_{k+1}' \beta_k). \end{aligned} \tag{24}$$

The last equality follows the definition of $J_{C,k+1}$. ■

Appendix B. The Proof of Theorem 1 and 2

Theorem 1 *Suppose that Assumption 1 and Assumption 2 hold. If* $\eta > \frac{\lambda_{\max}(Q_d)}{2\lambda_{\min}(Q_*)}$,

$$\mathbb{E}\left[\|\theta_k - \beta_*\|_{Q_*}^2\right] = O\left(\frac{1}{k}\right), \tag{9}$$

where $\|x\|_{Q_*}^2 = x^T Q_* x$.

Proof

We can apply the standard approach on $\{\theta_k\}$. Recall that β_* is the true parameter and we want to show that $\{\theta_k\}$ converges to β_*. Importantly, we do not directly analyze the 2-norm of $\theta_k - \beta_*$ to avoid cross term $J_{Q,k+1}^{-1} Q_*$, which is not necessary positive definite even when considering the limit. Instead we consider a matrix norm: specifically, we define $\|x\|_{Q_*}^2 = x^T Q_* x$, for $x \in \mathbb{R}^p$.[2] Denote

$$\begin{aligned} \lambda_{\min}(Q_*) = \xi_1, \quad \lambda_{\max}(Q_*) = \Xi_1, \\ \lambda_{\min}(Q_d) = \xi_2, \quad \lambda_{\max}(Q_d) = \Xi_2. \end{aligned} \tag{25}$$

The following inequalities give the equivalence of $\|\cdot\|_2$ and $\|\cdot\|_{Q_*}$,

$$\begin{aligned} \xi_1 \|x\|_2^2 \le \|x\|_{Q_*}^2 \le \Xi_1 \|x\|_2^2, \\ \frac{1}{\Xi_1}\|x\|_{Q_*}^2 \le \|x\|_2^2 \le \frac{1}{\xi_1}\|x\|_{Q_*}^2. \end{aligned} \tag{26}$$

And we consider the convergence in this norm instead, using (7) and Assumption 1 we have

$$\begin{aligned} \|\theta_k - \beta_*\|_{Q_*}^2 &= \left\|\theta_k - \beta_* + \frac{\eta}{\gamma} J_{Q,k+1}^{-1} X_{k+1}(Y_{k+1} - X_{k+1}'\theta_k)\right\|_{Q_*}^2 \\ &= \left\|\theta_k - \beta_* - \frac{\eta}{\gamma} J_{Q,k+1}^{-1} X_{k+1} X_{k+1}'(\theta_k - \beta_*) + \frac{\eta}{\gamma} J_{Q,k+1}^{-1} X_{k+1}\varepsilon_{k+1}\right\|_{Q_*}^2. \end{aligned} \tag{27}$$

Let $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, $\mathcal{G}_k = \sigma(X_1, Y_1, \ldots, X_k, Y_k)$ and $\mathcal{H}_{k+1} = \sigma(X_1, Y_1, \ldots, X_k, Y_k, X_{k+1})$, where $\sigma(\cdot)$ denotes the induced σ-field. Note $\mathcal{F}_k \subseteq \mathcal{G}_k \subseteq \mathcal{H}_{k+1}$. Then we take the expectation with

2. Analyzing the problem under this norm is equivalent to analyzing the progress of the objective function directly.

respect to $\mathcal{H}_{k+1}$ and use the first and second moment of ε_{k+1},

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{k+1}-\beta_*\|_{Q_*}^2 |\mathcal{H}_{k+1}\right] =& \|\theta_k-\beta_*\|_{Q_*}^2 - \frac{2\eta}{\gamma}(\theta_k-\beta_*)'Q_*J_{Q,k+1}^{-1}X_{k+1}X_{k+1}'(\theta_k-\beta_*) \\
&+ \frac{\eta^2}{\gamma^2}(\theta_k-\beta_*)'X_{k+1}X_{k+1}'J_{Q,k+1}^{-1}Q_* \\
&\times J_{Q,k+1}^{-1}X_{k+1}X_{k+1}'(\theta_k-\beta_*) \\
&+ \frac{\eta^2}{\gamma^2}\sigma^2\left(X_{k+1}'J_{Q,k+1}^{-1}Q_*J_{Q,k+1}^{-1}X_{k+1}\right) \\
\leq& \|\theta_k-\beta_*\|_{Q_*}^2 - \frac{2\eta}{\gamma}(\theta_k-\beta_*)'Q_*J_{Q,k+1}^{-1}X_{k+1}X_{k+1}'(\theta_k-\beta_*) \\
&+ \frac{\eta^2}{\gamma^2}\left\|Q_*^{\frac{1}{2}}J_{Q,k+1}^{-1}\right\|_2^2\|X_{k+1}\|_2^4\|\theta_k-\beta_*\|_2^2 \\
&+ \frac{\eta^2}{\gamma^2}\sigma^2\left\|Q_*^{\frac{1}{2}}J_{Q,k+1}^{-1}\right\|_2^2\|X_{k+1}\|_2^2 .
\end{aligned} \tag{28}$$

The inequality follows the definition of matrix induced 2-norm. Now take expectation with respect to $\mathcal{G}_k$ on both side. Denote $\mathbb{E}\left[\|X_1\|_2^4\right]=\mu_{X,4}$ and $\mathbb{E}\left[\|X_1\|_2^2\right]=\mu_{X,2}$. Then

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{k+1}-\beta_*\|_{Q_*}^2 |\mathcal{G}_k\right] \leq& \|\theta_k-\beta_*\|_{Q_*}^2 - \frac{2\eta}{\gamma}(\theta_k-\beta_*)'Q_*J_{Q,k+1}^{-1}Q_*(\theta_k-\beta_*) \\
&+ \frac{\eta^2}{\gamma^2}\mu_{X,4}\left\|Q_*^{\frac{1}{2}}J_{Q,k+1}^{-1}\right\|_2^2\|\theta_k-\beta_*\|_2^2 + \frac{\eta^2}{\gamma^2}\sigma^2\mu_{X,2}\left\|Q_*^{\frac{1}{2}}J_{Q,k+1}^{-1}\right\|_2^2 \\
\leq& \|\theta_k-\beta_*\|_{Q_*}^2 - \frac{2\eta}{\gamma}\lambda_{\min}\left(Q_*^{\frac{1}{2}}J_{Q,k+1}^{-1}Q_*^{\frac{1}{2}}\right)\|\theta_k-\beta_*\|_{Q_*} \\
&+ \frac{\eta^2}{\gamma^2}\mu_{X,4}\left\|Q_*^{\frac{1}{2}}\right\|_2^2\left\|J_{Q,k+1}^{-1}\right\|_2^2\|\theta_k-\beta_*\|_2^2 \\
&+ \frac{\eta^2}{\gamma^2}\sigma^2\mu_{X,2}\left\|Q_*^{\frac{1}{2}}\right\|_2^2\left\|J_{Q,k+1}^{-1}\right\|_2^2 \\
\leq& \|\theta_k-\beta_*\|_{Q_*}^2 - \frac{2\eta}{\gamma}\xi_1\lambda_{\min}\left(J_{Q,k+1}^{-1}\right)\|\theta_k-\beta_*\|_{Q_*} \\
&+ \frac{\eta^2}{\gamma^2}\mu_{X,4}\Xi_1\lambda_{\max}^2\left(J_{Q,k+1}^{-1}\right)\|\theta_k-\beta_*\|_2^2 \\
&+ \frac{\eta^2}{\gamma^2}\sigma^2\mu_{X,2}\Xi_1\lambda_{\max}^2\left(J_{Q,k+1}^{-1}\right),
\end{aligned} \tag{29}$$

where $\lambda_{\min}(\cdot)$ represents the smallest eigenvalue. The second and third inequality follows the inequalities of eigenvalues. Now since $\lim_{k\to\infty}\frac{1}{k}J_{Q,k+1}=\frac{1}{\gamma}Q_d$, we can bound the right-hand side with eigenvalues of Q_d and get rid of $J_{Q,k+1}$. The accurate bounds is given in Lemma 4.

Using Lemma 4 and (26),

$$\begin{aligned}\mathbb{E}\left[\|\theta_{k+1}-\beta_*\|_{Q_*}^2|\mathcal{G}_k\right] \leq\ & \|\theta_k-\beta_*\|_{Q_*}^2-\frac{1}{k}\frac{2\eta(1-\delta_1)\xi_1}{\Xi_2}\|\theta_k-\beta_*\|_{Q_*}^2 \\ &+\frac{1}{k^2}\frac{\eta^2(1+\delta_2)^2\mu_{X,4}\Xi_1}{\xi_2^2\xi_1}\|\theta_k-\beta_*\|_{Q_*}^2 \\ &+\frac{1}{k^2}\frac{\eta^2\sigma^2\mu_{X,2}(1+\delta_2)^2\Xi_1}{\xi_2^2},\end{aligned} \tag{30}$$

Now the right-hand side is in order $\|\theta_k-\beta_*\|_{Q_*}^2\,[1-O(1/k)]+O(1/k^2)$, we can therefore conclude that $\mathbb{E}\left[\theta_{k+1}-\beta_*\|_{Q_*}^2\right]$ converges to 0 with rate $O(1/k)$ by Lemma 1 in [6].

In detail, for arbitrary δ_3, let $K_2=\frac{\eta(1+\delta_2)^2\Xi_1\Xi_2\mu_{X,4}}{2\delta_3(1-\delta_1)\xi_1^2\xi_2^2}$. When $k>\max\{K_1,K_2\}$, we can control the third term,

$$\begin{aligned}\mathbb{E}\left[\|\theta_{k+1}-\beta_*\|_{Q_*}^2|\mathcal{G}_k\right] \leq\ & \left(1-\frac{1}{k}\frac{2\eta(1-\delta_1)(1-\delta_3)\xi_1}{\Xi_2}\right)\|\theta_k-\beta_*\|_{Q_*}^2 \\ &+\frac{1}{k^2}\frac{\eta^2\sigma^2\mu_{X,2}(1+\delta_2)^2\Xi_1}{\xi_2^2}.\end{aligned} \tag{31}$$

Let $B_1=\frac{2\eta(1-\delta_1)(1-\delta_3)\xi_1}{\Xi_2}$, $B_2=\frac{\eta^2\sigma^2\mu_{X,2}(1+\delta_2)^2\Xi_1}{\xi_2^2}$ and take expectation of both side of (31),

$$\mathbb{E}\left[\|\theta_{k+1}-\beta_*\|_{Q_*}^2\right]\leq\left(1-\frac{B_1}{k}\right)\mathbb{E}\left[\|\theta_k-\beta_*\|_{Q_*}^2\right]+\frac{B_2}{k^2}. \tag{32}$$

If $\eta>\frac{\Xi_2}{2\xi_1}$, we can choose appropriate δ_1,δ_3 such that $B_1>1$. Then we could prove by induction that for $k\geq K=\max\{K_1,K_2\}$,

$$\mathbb{E}\big[\|\theta_k-\beta_*\|_{Q_*}^2\big]\leq\frac{B_3}{k}, \tag{33}$$

where $B_3=\max\left\{K\mathbb{E}\left[\|\theta_K-\beta_*\|_{Q_*}^2\right],\frac{B_2}{B_1-1}\right\}$. We conclude the following result.

We can prove this by induction. For the base case $k=K_3$,

$$\mathbb{E}\big[\|\theta_{K_3}-\beta_*\|_{Q_*}^2\big]=\frac{K_3\mathbb{E}\big[\|\theta_{K_3}-\beta_*\|_{Q_*}^2\big]}{K_3}\leq\frac{B_3}{K_3}. \tag{34}$$

Now assume that for $k=l\geq K_3$, $\mathbb{E}\big[\|\theta_l-\beta_*\|_{Q_*}^2\big]\leq\frac{B_3}{l}$. For $k=l+1$, using (32) and the hypothesis assumption,

$$\begin{aligned}\mathbb{E}\big[\|\theta_{l+1}-\beta_*\|_{Q_*}^2\big] &\leq\left(1-\frac{B_1}{l}\right)\mathbb{E}\left[\|\theta_l-\beta_*\|_{Q_*}^2\right]+\frac{B_2}{l^2} \\ &\leq\left(1-\frac{B_1}{l}\right)\frac{B_3}{l}+\frac{B_2}{l^2}-\frac{B_3}{l+1}+\frac{B_3}{l+1} \\ &=\left(\frac{1}{l(l+1)}-\frac{B_1}{l^2}\right)B_3+\frac{B_1}{l^2}+\frac{B_3}{l+1} \\ &\leq\frac{(1-B_1)B_3+B_2}{l^2}+\frac{B_3}{l+1} \\ &\leq\frac{B_3}{l+1},\end{aligned} \tag{35}$$

which completes the proof. ■

Theorem 2 *With assumption 1, 2 and 3, if the step-size* $\eta > \frac{\lambda_{\max}(Q_d)}{2\lambda_{\min}(Q_*)}$, *then for arbitrary* $\delta > 0$, *there exists a sequence of events* $\{D_k\}_{k=K}^{\infty}$ *satisfies* $D_{k+1} \subset D_k$ *and for* $k \geq K$, $\lim_{k\to\infty} P(D_k) \geq 1 - \delta$, *such that*

$$\mathbb{E}\left[\|\theta_k - \beta_k\|_{Q_*}^2 1_{D_{k-1}}\right] = O\left(\frac{1}{k}\right). \tag{10}$$

Proof Since we have already proved the convergence of the alternative estimator, we can instead analyze the difference between our estimator and the "ghost" estimator.

Let $e_k = \|\beta_k - \theta_k\|_{Q_*}^2$. To analyze the convergence, we again write $\{\mathbb{E}[e_k]\}$ as a recursion. First take the difference of (12) and (7),

$$\begin{aligned}\beta_{k+1} - \theta_{k+1} = \beta_k - \theta_k \\ + \frac{\eta}{\gamma}\left(J_{C,k-1}^{-1} X_{k+1}(Y_{k+1} - X_{k+1}'\beta_k) - J_{Q,k-1}^{-1} X_{k+1}(Y_{k+1} - X_{k+1}'\theta_k)\right)\end{aligned} \tag{36}$$

To analyze the second term, we need to add and subtract several intermediate terms. There are two goals of doing this. First, we want to make use of the convergence of $\theta_k - \beta_*$. Second, we need to get rid of the product of $J_{C,k+1}^{-1}$ and X_{k+1} since it is difficult to control this term as mentioned before. We do some additions and subtractions to make the difference easier to analysis. We first add and subtract a $\frac{\eta}{\gamma} J_{Q,k-1}^{-1} X_{k+1}(Y_{k+1} - X_{k+1}'\beta_k)$,

$$\begin{aligned}\beta_{k+1} - \theta_{k+1} &= \beta_k - \theta_k + \frac{\eta}{\gamma} J_{Q,k-1}^{-1} X_{k+1}(Y_{k+1} - X_{k+1}'\beta_k) \\ &\quad - J_{Q,k-1}^{-1} X_{k+1}(Y_{k+1} - X_{k+1}'\theta_k) + \frac{\eta}{\gamma} J_{C,k-1}^{-1} X_{k+1}(Y_{k+1} - X_{k+1}'\beta_k) \\ &\quad - \frac{\eta}{\gamma} J_{Q,k-1}^{-1} X_{k+1}(Y_{k+1} - X_{k+1}'\beta_k) \\ &= \beta_k - \theta_k - \frac{\eta}{\gamma} J_{Q,k+1}^{-1} X_{k+1} X_{k+1}'(\beta_k - \theta_k) \\ &\quad + \frac{\eta}{\gamma}\left(J_{C,k+1}^{-1} - J_{Q,k+1}^{-1}\right) X_{k+1}(Y_{k+1} - X_{k+1}'\beta_k).\end{aligned} \tag{37}$$

Since we have no idea about $Y_{k+1} - X_{k+1}'\beta_k$, we do another pair of addition and subtraction,

$$\begin{aligned}\beta_{k+1} - \theta_{k+1} &= \beta_k - \theta_k - \frac{\eta}{\gamma} J_{Q,k+1}^{-1} X_{k+1} X_{k+1}'(\beta_k - \theta_k) \\ &\quad + \frac{\eta}{\gamma}\left(J_{C,k+1}^{-1} - J_{Q,k+1}^{-1}\right) X_{k+1}(Y_{k+1} - X_{k+1}'\beta_k) \\ &\quad - \frac{\eta}{\gamma}\left(J_{C,k+1}^{-1} - J_{Q,k+1}^{-1}\right) X_{k+1}(Y_{k+1} - X_{k+1}'\theta_k) \\ &\quad + \frac{\eta}{\gamma}\left(J_{C,k+1}^{-1} - J_{Q,k+1}^{-1}\right) X_{k+1}(Y_{k+1} - X_{k+1}'\beta_k) \\ &= \beta_k - \theta_k - \frac{\eta}{\gamma} J_{Q,k+1}^{-1} X_{k+1} X_{k+1}'(\beta_k - \theta_k) \\ &\quad - \frac{\eta}{\gamma}\left(J_{C,k+1}^{-1} - J_{Q,k+1}^{-1}\right) X_{k+1} X_{k+1}'(\beta_k - \theta_k) \\ &\quad + \frac{\eta}{\gamma}\left(J_{C,k+1}^{-1} - J_{Q,k+1}^{-1}\right) X_{k+1}(Y_{k+1} - X_{k+1}'\theta_k).\end{aligned} \tag{38}$$

We denote $\left(J_{C,k+1}^{-1}-J_{Q,k+1}^{-1}\right)$ by E_{k+1}. Now we can have a recursive formula of $\mathbb{E}[e_k]$. We first calculate the products out, take expectations, control the matrix-vector product with eigenvalues and combine like terms. The result is proved in Lemma 5 and is concluded here.

$$\begin{aligned}\mathbb{E}\left[e_{k+1}|\mathcal{G}_k\right] \leq e_k\left(1-\frac{2\eta}{\gamma}\xi_1\lambda_{\min}\left(J_{Q,k+1}^{-1}\right)+\frac{\eta^2\Xi_1}{\xi_1\gamma^2}\delta_{k,1}\right)\\ +\sqrt{e_k}\|\theta_k-\beta_*\|_{Q_*}\frac{2\eta^2\Xi_1}{\gamma^2\xi_1}\delta_{k,2}+\frac{\eta^2\Xi_1}{\gamma^2\xi_1}\delta_{k,3},\end{aligned} \tag{39}$$

where

$$\begin{aligned}\delta_{k,1} =& \left\|J_{Q,k+1}^{-1}\right\|_2^2\mu_{X,4}+\sqrt{\mu_{X,8}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^4\,|\mathcal{G}_k\right]}\\ &+2\left\|J_{Q,k+1}^{-1}\right\|_2\sqrt{\mu_{X,8}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^2\,|\mathcal{G}_k\right]}+\frac{2\gamma}{\eta}\sqrt{\mu_{X_1,4}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^2\,|\mathcal{G}_k\right]},\\ \delta_{k,2} =& \frac{\gamma}{\eta}\sqrt{\mu_{X_1,4}}\sqrt{\mathbb{E}[\|E_{k+1}\|_F^2|\mathcal{G}_k]}+\sqrt{\mu_{X,8}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^4\,|\mathcal{G}_k\right]}\\ &+\left\|J_{Q,k+1}^{-1}\right\|_2\sqrt{\mu_{X,8}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^2\,|\mathcal{G}_k\right]},\\ \delta_{k,3} =& \|\theta_k-\beta_*\|_{Q_*}^2\sqrt{\mu_{X,8}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^4\,|\mathcal{G}_k\right]}+\xi_1\sigma^2\sqrt{\mu_{X,4}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^4\,|\mathcal{G}_k\right]}.\end{aligned} \tag{40}$$

With Lemma 6, if we can control $\delta_{k,1}$, $\delta_{k,2}$ and $\delta_{k,3}$ appropriately, we will be able to show that $\{\mathbb{E}[e_k]\}$ convergences to 0 with rate $O(1/k)$. To show this, we first notice that in these three terms, only $\mathbb{E}\left[\|E_{k+1}\|_F^j\,|\mathcal{G}_k\right]$ are left to analyze, $j=2,4$. What's more, by Jensen's inequality,

$$\mathbb{E}\left[\|E_{k+1}\|_F^2\,|\mathcal{G}_k\right]\leq\left(\mathbb{E}\left[\|E_{k+1}\|_F^4\,|\mathcal{G}_k\right]\right)^{\frac{1}{2}}. \tag{41}$$

Therefore, we only need to bound $\mathbb{E}\left[\|E_{k+1}\|_F^4\,|\mathcal{G}_k\right]$. Actually, we are able to prove that this term is in $O(1/k^{9/2})$ with high probability in the following paragraphs. With this rate, we can show that the recursive formula of $\mathbb{E}[e_k]$ satisfies the condition in Lemma 7 and therefore $\{\mathbb{E}[e_k]\}$ converges with high probability. This is discussed in appendix D. In conclusion, for arbitrary δ, there exist K_5 such that

$$P\left(\bigcup_{k=K_5}^{\infty}\left\{\mathbb{E}\left[\|E_{k+1}\|_F^4\,|\mathcal{G}_k\right]>\frac{B_4}{k^{9/2}}\right\}\right)<\delta, \tag{42}$$

the exact definition of K_5 is given in the end of appendix D

Then we will complete the proof of Theorem 2. For arbitrary δ, $K_5=K_5(\delta)$ is defined in Lemma 8. Let

$$D_k=\bigcap_{j=K_5}^{k}\left\{\mathbb{E}\left[\|E_{k+1}\|_F^4\,|\mathcal{G}_k\right]\leq\frac{B_4}{k^{9/2}}\right\}, \tag{43}$$

for $k\geq K_5$. Then by (42), $\lim_{k\to\infty}P(D_k)\geq 1-\delta$. Since $D_k\in\mathcal{G}_k$, we have

$$\mathbb{E}\left[e_{k+1}1_{D_k}|\mathcal{G}_k\right]=\mathbb{E}\left[e_{k+1}|\mathcal{G}_k\right]1_{D_k} \tag{44}$$

Using (39),

$$\begin{aligned}\mathbb{E}\left[e_{k+1}1_{D_k}|\mathcal{G}_k\right] \leq e_k 1_{D_k}\left(1-\frac{2\eta}{\gamma}\xi_1\lambda_{\min}\left(J_{Q,k+1}^{-1}\right)+\frac{\eta^2\Xi_1}{\xi_1\gamma^2}\delta_{k,1}1_{D_k}\right)\\ +\sqrt{e_k 1_{D_k}}\|\theta_k-\beta_*\|_{Q_*}\frac{2\eta^2\Xi_1}{\gamma^2\xi_1}\delta_{k,2}1_{D_k}+\frac{\eta^2\Xi_1}{\gamma^2\xi_1}\delta_{k,3}1_{D_k}.\end{aligned} \tag{45}$$

By the definition of $\delta_{k,1}$, $\delta_{k,2}$, $\delta_{k,3}$ and D_k, we have

$$\begin{aligned}\delta_{k,1}1_{D_k} &\leq \left\|J_{Q,k+1}^{-1}\right\|_2^2\mu_{X,4}+\sqrt{\mu_{X,8}}\frac{\sqrt{B_4}}{k^{\frac{9}{4}}}+2\left\|J_{Q,k+1}^{-1}\right\|_2\sqrt{\mu_{X,8}}\frac{B_4^{\frac{1}{4}}}{k^{\frac{9}{8}}}+\frac{2\gamma}{\eta}\sqrt{\mu_{X_1,4}}\frac{B_4^{\frac{1}{4}}}{k^{\frac{9}{8}}},\\ \delta_{k,2}1_{D_k} &\leq \frac{\gamma}{\eta}\sqrt{\mu_{X_1,4}}\frac{B_4^{\frac{1}{4}}}{k^{\frac{9}{8}}}+\sqrt{\mu_{X,8}}\frac{B_4^{\frac{1}{4}}}{k^{\frac{9}{8}}}+\left\|J_{Q,k+1}^{-1}\right\|_2\sqrt{\mu_{X,8}}\frac{B_4^{\frac{1}{4}}}{k^{\frac{9}{8}}},\\ \delta_{k,3}1_{D_k} &\leq \|\theta_k-\beta_*\|_{Q_*}^2\sqrt{\mu_{X,8}}\frac{\sqrt{B_4}}{k^{\frac{9}{4}}}+\xi_1\sigma^2\sqrt{\mu_{X,4}}\frac{\sqrt{B_4}}{k^{\frac{9}{4}}}.\end{aligned} \tag{46}$$

We denote the right-hand side by $\tilde{\delta}_{k,1}$, $\tilde{\delta}_{k,2}$, $\tilde{\delta}_{k,3}$ respectively. Note that $\tilde{\delta}_{k,1}$ and $\tilde{\delta}_{k,2}$ are deterministic. Substituting (46) into (45) and taking expectation on both side,

$$\begin{aligned}\mathbb{E}\left[e_{k+1}1_{D_k}\right] \leq \mathbb{E}\left[e_k 1_{D_k}\right]\left(1-\frac{2\eta}{\gamma}\xi_1\lambda_{\min}\left(J_{Q,k+1}^{-1}\right)+\frac{\eta^2\Xi_1}{\xi_1\gamma^2}\tilde{\delta}_{k,1}\right)\\ +\sqrt{\mathbb{E}\left[e_k 1_{D_k}\right]}\sqrt{\mathbb{E}\left[\|\theta_k-\beta_*\|_{Q_*}^2\right]}\frac{2\eta^2\Xi_1}{\gamma^2\xi_1}\tilde{\delta}_{k,2}+\frac{\eta^2\Xi_1}{\gamma^2\xi_1}\mathbb{E}\left[\tilde{\delta}_{k,3}\right].\end{aligned} \tag{47}$$

Choose appropriate δ_1 and δ_3 such that $B_1=\frac{2\eta(1-\delta_1)(1-\delta_3)\xi_1}{\Xi_2}>1$ and let $\delta_2=1$. Then using Lemma 4 we have when $k>K_1=\max\{m,K_{1,1}(\delta_1),K_{1,2}(1)\}$,

$$\begin{aligned}\lambda_{\min}\left(J_{Q,k+1}^{-1}\right) &\geq \frac{1}{k}\frac{(1-\delta_1)\gamma}{\Xi_2},\\ \lambda_{\max}\left(J_{Q,k+1}^{-1}\right) &\leq \frac{1}{k}\frac{2\gamma}{\xi_2}.\end{aligned} \tag{48}$$

Substituting (48) into (46) we have

$$\tilde{\delta}_{k,1}\leq\frac{B_7}{k^{\frac{9}{8}}}, \tag{49}$$

where

$$B_7=\frac{\gamma^2}{\xi_2^2}\mu_{X,4}+\sqrt{\mu_{X,8}B_4}+\sqrt{\mu_{X,8}}\frac{4\gamma B_4^{\frac{1}{4}}}{\xi_2}+\sqrt{\mu_{X_1,4}}\frac{2\gamma B_4^{\frac{1}{4}}}{\eta}. \tag{50}$$

Then similar as what we did in Theorem 1, for $k\geq K_6=\left(\frac{\eta\Xi_1B_7\Xi_2}{2\xi_1^2(1-\delta_1)\delta_3\gamma^2}\right)^8$, we can dominate $\tilde{\delta}_{k,1}$ with the previous negative term in the parentheses. Now we want to bound the term $\mathbb{E}\left[\|\theta_k-\beta_*\|_{Q_*}^2\right]$. Using Theorem 1, for $k\geq K_3=\max\{K_1,K_2\}$,

$$\mathbb{E}\left[\|\theta_k-\beta_*\|_{Q_*}^2\right]\leq\frac{B_3}{k}. \tag{51}$$

Substituting (48) , (49) and (51) into (47) we have

$$\begin{aligned}\mathbb{E}\left[e_{k+1}1_{D_k}\right] &\leq \mathbb{E}\left[e_k 1_{D_k}\right]\left(1-\frac{B_1}{k}\right)+\sqrt{\mathbb{E}\left[e_k 1_{D_k}\right]}\frac{B_8}{k^{\frac{3}{2}}}+\frac{B_9}{k^2} \\ &\leq \mathbb{E}\left[e_k 1_{D_{k-1}}\right]\left(1-\frac{B_1}{k}\right)+\sqrt{\mathbb{E}\left[e_k 1_{D_{k-1}}\right]}\frac{B_8}{k^{\frac{3}{2}}}+\frac{B_9}{k^2},\end{aligned} \tag{52}$$

where

$$\begin{aligned}B_8 &= B_3^{\frac{1}{2}}B_4^{\frac{1}{4}}\frac{2\eta^2\Xi_1}{\gamma^2\xi_1}\left(\frac{\gamma}{\eta}\sqrt{\mu_{X_1,4}}+\sqrt{\mu_{X,8}}+\sqrt{\mu_{X,8}\frac{\gamma}{\xi_2}}\right) \\ B_9 &= \frac{\eta^2\Xi_1}{\gamma^2\xi_1}\sqrt{B_4}\left(B_3\sqrt{\mu_{X,8}}+\xi_1\sigma^2\sqrt{\mu_{X,4}}\right).\end{aligned} \tag{53}$$

Now we can apply Lemma 6 and the conclusion follows. ■

Appendix C. Technique Lemmas

Lemma 4 *For arbitrary small $\delta_1 > 0$ and $\delta_2 > 0$, there exist $K_1 = \max\{m, K_{1,1}(\delta_1), K_{1,2}(\delta_2)\}$, such that when $k > K_1$,*

$$\lambda_{\min}\left(J_{Q,k+1}^{-1}\right) \geq \frac{1}{k}\frac{(1-\delta_1)\gamma}{\Xi_2}, \tag{54}$$

$$\lambda_{\max}\left(J_{Q,k+1}^{-1}\right) \leq \frac{1}{k}\frac{(1+\delta_2)\gamma}{\xi_2}, \tag{55}$$

where

$$\begin{aligned}K_{1,1}(\delta_1) &= \frac{(1-\delta_1)\left[\gamma+(m+1)\Xi_1-m\Xi_2\right]}{\delta_1\Xi_2}, \\ K_{1,2}(\delta_2) &= \frac{(1+\delta_2)\left[m\xi_2-(m+1)\xi_1-\gamma\right]}{\delta_2\xi_2}.\end{aligned} \tag{56}$$

Proof Using (7), for $k > m$,

$$\begin{aligned}\lambda_{\max}\left(J_{Q,k+1}\right) &\leq 1+\frac{1}{\gamma}\left[\left(k-m\right)\Xi_2+\left(m+1\right)\Xi_1\right], \\ \lambda_{\min}\left(J_{Q,k+1}\right) &\geq 1+\frac{1}{\gamma}\left[\left(k-m\right)\xi_2+\left(m+1\right)\xi_1\right].\end{aligned} \tag{57}$$

Since the reciprocal of eigenvalues of a matrix are the eigenvalues of the inverse matrix, we will have

$$\lambda_{\min}\left(J_{Q,k+1}^{-1}\right) = \lambda_{\max}\left(J_{Q,k+1}\right)^{-1} \geq \left(1+\frac{1}{\gamma}\left[\left(k-m\right)\Xi_2+\left(m+1\right)\Xi_1\right]\right)^{-1}. \tag{58}$$

Then we can get a lower bound of k which can guarantee that (54) holds,

$$k \geq \frac{(1-\delta_1)\left[\gamma+(m+1)\Xi_1-m\Xi_2\right]}{\delta_1\Xi_2}. \tag{59}$$

Similarly,

$$\begin{aligned}\lambda_{\max}\left(J_{Q,k+1}^{-1}\right) &= \lambda_{\min}\left(J_{Q,k+1}\right)^{-1} \\ &\leq \left(1+\frac{1}{\gamma}\left[(k-m)\xi_2+(m+1)\xi_1\right]\right)^{-1}.\end{aligned} \tag{60}$$

We can also get a lower bound of k such that (55) holds,

$$k \geq \frac{(1+\delta_2)\left[m\xi_2-(m+1)\xi_1-\gamma\right]}{\delta_2\xi_2}. \tag{61}$$

■

Lemma 5 *With Assumption 1, 2 and 3,*

$$\begin{aligned}\mathbb{E}\left[e_{k+1}|\mathcal{G}_k\right] \leq e_k\left(1-\frac{2\eta}{\gamma}\xi_1\lambda_{\min}\left(J_{Q,k+1}^{-1}\right)+\frac{\eta^2\Xi_1}{\xi_1\gamma^2}\delta_{k,1}\right) \\ +\sqrt{e_k}\|\theta_k-\beta_*\|_{Q_*}\frac{2\eta^2\Xi_1}{\gamma^2\xi_1}\delta_{k,2}+\frac{\eta^2\Xi_1}{\gamma^2\xi_1}\delta_{k,3},\end{aligned} \tag{62}$$

where

$$\begin{aligned}\delta_{k,1} =& \left\|J_{Q,k+1}^{-1}\right\|_2^2\mu_{X,4}+\sqrt{\mu_{X,8}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^4\,|\mathcal{G}_k\right]} \\ &+2\left\|J_{Q,k+1}^{-1}\right\|_2\sqrt{\mu_{X,8}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^2\,|\mathcal{G}_k\right]}+\frac{2\gamma}{\eta}\sqrt{\mu_{X_1,4}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^2\,|\mathcal{G}_k\right]}, \\ \delta_{k,2} =& \frac{\gamma}{\eta}\sqrt{\mu_{X_1,4}}\sqrt{\mathbb{E}[\|E_{k+1}\|_F^2|\mathcal{G}_k]}+\sqrt{\mu_{X,8}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^4\,|\mathcal{G}_k\right]} \\ &+\left\|J_{Q,k+1}^{-1}\right\|_2\sqrt{\mu_{X,8}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^2\,|\mathcal{G}_k\right]}, \\ \delta_{k,3} =& \|\theta_k-\beta_*\|_{Q_*}^2\sqrt{\mu_{X,8}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^4\,|\mathcal{G}_k\right]}+\xi_1\sigma^2\sqrt{\mu_{X,4}}\sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^4\,|\mathcal{G}_k\right]}.\end{aligned} \tag{63}$$

Proof We first substitute (38) into the definition of e_{k+1},

$$\begin{aligned}
e_{k+1} = e_k &+ \frac{\eta^2}{\gamma^2}\left\|Q_*^{\frac{1}{2}} J_{Q,k+1}^{-1} X_{k+1}\right\|_2^2 \left[X_{k+1}'\left(\beta_k - \theta_k\right)\right]^2 \\
&+ \frac{\eta^2}{\gamma^2}\left\|Q_*^{\frac{1}{2}} E_{k+1} X_{k+1}\right\|_2^2 \left[X_{k+1}'\left(\beta_k - \theta_k\right)\right]^2 \\
&+ \frac{\eta^2}{\gamma^2}\left\|Q_*^{\frac{1}{2}} E_{k+1} X_{k+1}\right\|_2^2 \left(Y_{k+1} - X_{k+1}'\theta_k\right)^2 \\
&- \frac{2\eta}{\gamma}\left(\beta_k - \theta_k\right)' Q_* J_{Q,k+1}^{-1} X_{k+1} X_{k+1}'(\beta_k - \theta_k) \\
&- \frac{2\eta}{\gamma}(\beta_k - \theta_k)' Q_* E_{k+1} X_{k+1} X_{k+1}'(\beta_k - \theta_k) \\
&+ \frac{2\eta}{\gamma}(\beta_k - \theta_k)' Q_* E_{k+1} X_{k+1}\left(Y_{k+1} - X_{k+1}'\theta_k\right) \\
&+ \frac{2\eta^2}{\gamma^2}(\beta_k - \theta_k)' X_{k+1} X_{k+1}' J_{Q,k+1}^{-1} Q_* E_{k+1} X_{k+1} X_{k+1}'(\beta_k - \theta_k) \\
&- \frac{2\eta^2}{\gamma^2}(\beta_k - \theta_k)' X_{k+1} X_{k+1}' J_{Q,k+1}^{-1} Q_* E_{k+1} X_{k+1}\left(Y_{k+1} - X_{k+1}'\theta_k\right) \\
&- \frac{2\eta^2}{\gamma^2}(\beta_k - \theta_k)' X_{k+1} X_{k+1}' E_{k+1} Q_* E_{k+1} X_{k+1}\left(Y_{k+1} - X_{k+1}'\theta_k\right).
\end{aligned} \tag{64}$$

Recall that $\mathcal{H}_{k+1} = \sigma(X_1, Y_1, \ldots, X_k, Y_k, X_{k+1})$ and $\mathcal{G}_k = \sigma(X_1, Y_1, \ldots, X_k, Y_k)$. Compute conditional expectation first with respective to $\mathcal{H}_{k+1}$ using Assumption 1 and the first and second moments of ε_{k+1},

$$\begin{aligned}
\mathbb{E}\left[e_{k+1}|\mathcal{H}_{k+1}\right] = e_k &+ \frac{\eta^2}{\gamma^2}\left\|Q_*^{\frac{1}{2}} J_{Q,k+1}^{-1} X_{k+1}\right\|_2^2 \left[X_{k+1}'(\beta_k - \theta_k)\right]^2 \\
&+ \frac{\eta^2}{\gamma^2}\left\|Q_*^{\frac{1}{2}} E_{k+1} X_{k+1}\right\|_2^2 \left[X_{k+1}'\left(\beta_k - \theta_k\right)\right]^2 \\
&+ \frac{\eta^2}{\gamma^2}\left\|Q_*^{\frac{1}{2}} E_{k+1} X_{k+1}\right\|_2^2 \left[X_{k+1}'\left(\beta_* - \theta_k\right)\right]^2 + \frac{\eta^2}{\gamma^2}\left\|Q_*^{\frac{1}{2}} E_{k+1} X_{k+1}\right\|_2^2 \sigma^2 \\
&- \frac{2\eta}{\gamma}(\beta_k - \theta_k)' Q_* J_{Q,k+1}^{-1} X_{k+1} X_{k+1}'(\beta_k - \theta_k) \\
&- \frac{2\eta}{\gamma}(\beta_k - \theta_k)' Q_* E_{k+1} X_{k+1} X_{k+1}'(\beta_k - \theta_k) \\
&+ \frac{2\eta}{\gamma}(\beta_k - \theta_k)' Q_* E_{k+1} X_{k+1} X_{k+1}'(\beta_* - \theta_k) \\
&+ \frac{2\eta^2}{\gamma^2}(\beta_k - \theta_k)' X_{k+1} X_{k+1}' J_{Q,k+1}^{-1} Q_* E_{k+1} X_{k+1} X_{k+1}'(\beta_k - \theta_k) \\
&- \frac{2\eta^2}{\gamma^2}(\beta_k - \theta_k)' X_{k+1} X_{k+1}' J_{Q,k+1}^{-1} Q_* E_{k+1} X_{k+1} X_{k+1}'(\beta_* - \theta_k) \\
&- \frac{2\eta^2}{\gamma^2}(\beta_k - \theta_k)' X_{k+1} X_{k+1}' E_{k+1} Q_* E_{k+1} X_{k+1} X_{k+1}'(\beta_* - \theta_k).
\end{aligned} \tag{65}$$

We then compute $\mathbb{E}\left[e_{k+1}|\mathcal{G}_k\right]$ using (26). The Cauchy–Schwarz inequality are applied to decouple correlated products. Also we control the matrix-vector product with eigenvalues and Frobenius norm. Here, we use the Frobenius norm rather than l_2 norm since the Frobenius norm is easier to control.

$$
\begin{aligned}
\mathbb{E}[e_{k+1}|\mathcal{G}_k] \leq\ & e_k + e_k \left\| J_{Q,k+1}^{-1} \right\|_2^2 \frac{\eta^2 \Xi_1}{\gamma^2 \xi_1} \mu_{X,4} + e_k \frac{\eta^2 \Xi_1}{\gamma^2 \xi_1} \sqrt{\mu_{X,8}} \sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k \right]} \\
& + \|\theta_k - \beta_*\|_{Q_*}^2 \frac{\eta^2 \Xi_1}{\gamma^2 \xi_1} \sqrt{\mu_{X,8}} \sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k \right]} \\
& + \frac{\eta^2 \Xi_1}{\gamma^2} \sigma^2 \sqrt{\mu_{X,4}} \sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k \right]} \\
& - \frac{2\eta}{\gamma} (\beta_k - \theta_k)' Q_* J_{Q,k+1}^{-1} Q_* (\beta_k - \theta_k) \\
& + e_k \frac{2\eta \Xi_1}{\gamma \xi_1} \sqrt{\mu_{X_1,4}} \sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^2 \,|\mathcal{G}_k \right]} \\
& + \sqrt{e_k} \|\beta_* - \theta_k\|_{Q_*} \frac{2\eta \Xi_1}{\gamma \xi_1} \sqrt{\mu_{X,4}} \sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^2 \,|\mathcal{G}_k \right]} \\
& + e_k \left\| J_{Q,k+1}^{-1} \right\|_2 \frac{2\eta^2 \Xi_1}{\gamma^2 \xi_1} \sqrt{\mu_{X,8}} \sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^2 \,|\mathcal{G}_k \right]} \\
& + \sqrt{e_k} \|\beta_* - \theta_k\|_{Q_*} \left\| J_{Q,k+1}^{-1} \right\|_2 \frac{2\eta^2 \Xi_1}{\gamma^2 \xi_1} \sqrt{\mu_{X,8}} \sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^2 \,|\mathcal{G}_k \right]} \\
& + \sqrt{e_k} \|\beta_* - \theta_k\|_{Q_*} \frac{2\eta^2 \Xi_1}{\gamma^2 \xi_1} \sqrt{\mu_{X,8}} \sqrt{\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k \right]} \\
\leq\ & e_k \left(1 - \frac{2\eta}{\gamma} \xi_1 \lambda_{\min}\left(J_{Q,k+1}^{-1} \right) + \frac{\eta^2 \Xi_1}{\xi_1 \gamma^2} \delta_{k,1} \right) \\
& + \sqrt{e_k} \|\theta_k - \beta_*\|_{Q_*} \frac{2\eta^2 \Xi_1}{\gamma^2 \xi_1} \delta_{k,2} + \frac{\eta^2 \Xi_1}{\gamma^2 \xi_1} \delta_{k,3}.
\end{aligned}
\tag{66}
$$

The last line follows the definition of $\delta_{k,1}$, $\delta_{k,2}$ and $\delta_{k,3}$. ■

Lemma 6 *If a positive sequence $\{b_k\}$ satisfies*

$$
b_{k+1} \leq \left(1 - \frac{\Delta_1}{k} \right) b_k + \frac{\Delta_2}{k^{3/2}} \sqrt{b_k} + \frac{\Delta_3}{k^2} \tag{67}
$$

for all $k > K$ with $\Delta_1 > 1$, $\Delta_2, \Delta_3 > 0$. Then for all $k > K$,

$$
b_k \leq \frac{\Delta_4}{k}, \tag{68}
$$

where $\Delta_4 = \max\left\{ K b_K, \left(\frac{\Delta_2^2 + \sqrt{\Delta_2^2 + 4(1-\Delta_1)\Delta_3}}{2(\Delta_1 - 1)} \right)^2 \right\}$.

Proof We prove this by induction. For $k = K$,

$$b_K = \frac{Kb_K}{K} \leq \frac{\Delta_4}{K}. \tag{69}$$

Now we assume that for $k = l > K$, $b_l < \frac{\Delta_4}{l}$, then for $k = l + 1$,

$$\begin{aligned} b_{l+1} &\leq \left(1 - \frac{\Delta_1}{l}\right)\frac{\Delta_4}{l} + \frac{\Delta_2\sqrt{\Delta_4}}{l^2} + \frac{\Delta_3}{l^2} \\ &\leq \left(1 - \frac{\Delta_1}{l}\right)\frac{\Delta_4}{l} + \frac{\Delta_2\sqrt{\Delta_4}}{l^2} + \frac{\Delta_3}{l^2} - \frac{\Delta_4}{l+1} + \frac{\Delta_4}{l+1} \\ &\leq \frac{\Delta_4}{l(l+1)} + \frac{-\Delta_1\Delta_4 + \Delta_2\sqrt{\Delta_4} + \Delta_3}{l^2} + \frac{\Delta_4}{l+1} \\ &\leq \frac{(1-\Delta_1)\Delta_4 + \Delta_2\sqrt{\Delta_4} + \Delta_3}{l^2} + \frac{\Delta_4}{l+1}. \end{aligned} \tag{70}$$

Since $1 - \Delta_1 < 0$, using the property of quadratic function we have

$$(1-\Delta_1)\Delta_4 + \Delta_2\sqrt{\Delta_4} + \Delta_3 \leq 0, \tag{71}$$

when

$$\sqrt{\Delta_4} \geq \frac{\Delta_2 + \sqrt{\Delta_2^2 + 4(\Delta_1 - 1)\Delta_3}}{2(\Delta_1 - 1)}. \tag{72}$$

Therefore $b_{l+1} \leq \frac{\Delta_4}{l+1}$, which completes the induction reasoning. ■

Appendix D. Bound $\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k\right]$ with high probability

To bound $\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k\right]$, we first break E_{k+1} in several parts so that we can directly analyze the elements in each part. Then we apply Chebyshev's inequality to control the probability of those elements deviating from theirs expectation for a certain distance. Then with a summable series of probabilities, we will have that $\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k\right]$ is controlled for all k large enough with high probability.

Since $J_{Q,k+1}$ and $J_{C,k+1}$ are defined in additive forms, it is easier to analyze elements in the difference of themselves rather than directly analyze E_{k+1}, which is the difference of inverses. Therefore we first rewrite E_{k+1}. Using the definition of E_{k+1} and the properties of Frobenius norm,

$$\begin{aligned} \|E_{k+1}\|_F^4 &= \left\|J_{Q,k+1}^{-1} - J_{C,k+1}^{-1}\right\|_F^4 \\ &= \left\|J_{C,k+1}^{-1}\left(J_{C,k+1} - J_{Q,k+1}\right)J_{Q,k+1}^{-1}\right\|_F^4 \\ &\leq \left\|J_{C,k+1}^{-1}\right\|_F^4 \|(J_{C,k+1} - J_{Q,k+1})\|_F^4 \left\|J_{Q,k+1}^{-1}\right\|_F^4 \end{aligned} \tag{73}$$

For the third term $\left\|J_{Q,k+1}^{-1}\right\|_F^4$, we recall that $J_{Q,k+1}^{-1}$ is deterministic and we give a bound of its eigenvalue in Lemma 4. Using Lemma 4 and taking $\delta_2 = 1$ we have for $k > \max\{m, K_{1,2}(1)\}$

$$\left\|J_{Q,k+1}^{-1}\right\|_F^4 \leq p^2\lambda_{\max}^4\left(J_{Q,k+1}^{-1}\right) \leq \frac{p^2 2^4\gamma^4}{k^4\xi_2^4} \tag{74}$$

Nevertheless, for the first term $\left\|J_{C,k+1}^{-1}\right\|_F^4$, the method we used in Lemma 4 is not applicable. And it is still hard to directly analyze the elements in $J_{C,k+1}^{-1}$. However, due to the special structure of $J_{C,k+1}$, we are able to bound this term by $\|C_k\|_F^4$, where C_k is a diagonal matrix and we can then directly deal with its elements. The bound is obtained by keep using Sherman-Morrison formula and is concluded in the following lemma.

Lemma 7 *If $J_{C,k+1}$ is defined as (11), then for $k \geq m$,*

$$\left\|J_{C,k+1}^{-1}\right\|_F^2 \leq \|C_k\|_F^2 . \tag{75}$$

Proof By (11), $J_{C,k+1} \succeq C_j^{-1}$. Therefore,

$$J_{C,k+1}^{-1} \preceq C_k. \tag{76}$$

Let $Q\Lambda Q'$ be the Schur decomposition of $J_{C,k+1}^{-1}$. Q is positive definite and the columns are given by q_i, $i = 1, \ldots, p$. Recall that p is the dimension of $J_{C,k+1}$. Σ is a diagonal matrix and the diagonal elements of Σ is denoted by $\sigma(J_{C,k+1}^{-1})$. Then,

$$\begin{aligned} \|J_{C,k+1}^{-1}\|_F^2 &= \sum_{i=1}^{p} \sigma_i^2(J_{C,k+1}^{-1}) = \sum_{i=1}^{p} (q_i' J_{C,k+1}^{-1} q_i)^2 \\ &\leq \sum_{i=1}^{p} (q_i' C_k q_i) \leq \|Q' C_k Q\|_F^2 = \|C_k\|_F^2. \end{aligned} \tag{77}$$

■

Now return to $\|E_{k+1}\|_F^4$. We have

$$\|E_{k+1}\|_F^4 \leq \|C_k\|_F^4 \, \|(J_{C,k+1} - J_{Q,k+1})\|_F^4 \left\|J_{Q,k+1}^{-1}\right\|_F^4 . \tag{78}$$

Remember that we need to take expectation conditioning on $\mathcal{G}_k$. The terms $\|C_k\|_F^4$ and $\left\|J_{Q,k+1}^{-1}\right\|_F^4$ on the right-hand side are in $\mathcal{G}_k$ and we only need to consider the conditional expectation of the term $\|J_{C,k+1} - J_{Q,k+1}\|_F^4$. If we try to directly calculate it, the fourth power will generate many cross terms and will entangle the calculations. However, since $J_{C,k+1}$ is defined in an additive form and all the terms expect $X_{k+1}X_{k+1}'$ of $J_{C,k+1}$ are in $\mathcal{G}_k$, we can break $\|J_{C,k+1} - J_{Q,k+1}\|_F^4$ into two parts to decouple $X_{k+1}X_{k+1}'$ with other terms. We remove the last term of $J_{C,k+1}$ and $J_{Q,k+1}$ separately and use $\tilde{J}_{C,k}$ and $\tilde{J}_{Q,k}$ to represent the new matrices:

$$\begin{aligned} \tilde{J}_{C,k} &= C_k^{-1} + \frac{1}{\gamma} \sum_{j=1}^{m \wedge k} X_{k+1-j} X_{k+1-j}', \\ \tilde{J}_{Q,k} &= Q_k^{-1} + \frac{m \wedge k}{\gamma} Q_* \end{aligned} \tag{79}$$

Then by definition,

$$\begin{aligned}\|J_{C,k+1} - J_{Q,k+1}\|_F^2 &= \|\tilde{J}_{C,k} - \tilde{J}_{Q,k} + X_{k+1}X'_{k+1} - Q_*\|_F^2 \\ &\leq 2\left(\|\tilde{J}_{C,k} - \tilde{J}_{Q,k}\|_F^2 + \|X_{k+1}X'_{k+1} - Q_*\|_F^2\right).\end{aligned} \tag{80}$$

The second line follows the inequality of mean. Substitute this into (78),

$$\begin{aligned}\|E_{k+1}\|_F^4 &\leq 4\|C_k\|_F^4 \left(\left\|\tilde{J}_{C,k} - \tilde{J}_{Q,k}\right\|_F^2 + \left\|X_{k+1}X'_{k+1} - Q_*\right\|_F^2\right)^2 \left\|J_{Q,k+1}^{-1}\right\|_F^4 \\ &\leq 8\|C_k\|_F^4 \left(\left\|\tilde{J}_{C,k} - \tilde{J}_{Q,k}\right\|_F^4 + \left\|X_{k+1}X'_{k+1} - Q_*\right\|_F^4\right) \left\|J_{Q,k+1}^{-1}\right\|_F^4\end{aligned} \tag{81}$$

Then we take the conditional expectation on both side,

$$\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k\right] \leq 8\|C_k\|_F^4 \left(\left\|\tilde{J}_{C,k} - \tilde{J}_{Q,k}\right\|_F^4 + L\right) \left\|J_{Q,k+1}^{-1}\right\|_F^4, \tag{82}$$

where $L = \mathbb{E}\left[\left\|X_{k+1}X'_{k+1} - Q_*\right\|_F^4\right]$ is a constant.

Now we can deal with the elements in C_k and $\tilde{J}_{C,k} - \tilde{J}_{Q,k}$. We are going to using Chebyshev's inequality to bound the probability of each element being large. Then we will be able to control $\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k\right]$ for all k large enough with high probability. The following lemma gives the formal statement.

Lemma 8 *Assume that Assumption 2 and 3 hold. For arbitrary small δ, there exist $K_5 = K_5(\delta)$ and a constant B_4 such that the probability that $\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k\right] < \frac{B_4}{k^{9/4}}$ for all $k > K$ is greater than $1 - \delta$.*

Proof We first look at the term $\|C_k\|_F^4$. As we mentioned before, we can control the elements of C_k separately. Recall that C_k is a diagonal matrix and for $k \geq m$,

$$C_k[j,j] = \frac{\gamma}{\sum_{i=1}^{k-m} X_{i,j}^2}. \tag{83}$$

We are going to applying Chebyshev's inequality, which enables us to control the probability of a random deviating from its expectation. Instead of directly applying that, we make the following transformation since we have no idea about the expectation of this reciprocal. Let $\mathbb{E}[X_{1,j}^2] = \mu_{X_{1,j},2}$. For $k > 2m$,

$$\begin{aligned}P\left(C_k[j,j] > \frac{4\gamma}{k\mu_{X_{1,j},2}}\right) &\leq P\left(C_k[j,j] > \frac{2\gamma}{(k-m)\mu_{X_{1,j},2}}\right) \\ &= P\left(\frac{\gamma}{\sum_{i=1}^{k-m} X_{i,j}^2} > \frac{2\gamma}{(k-m)\mu_{X_{1,j},2}}\right) \\ &= P\left(\sum_{i=1}^{k-m}\left(X_{i,j}^2 - \mu_{X_{1,j},2}\right) < -\frac{1}{2}(k-m)\mu_{X_{1,j},2}\right) \\ &\leq P\left(\left|\frac{1}{k-m}\sum_{i=1}^{k-m}(X_{i,j}^2 - \mu_{X_{1,j},2})\right| > \frac{1}{2}\mu_{X_{1,j},2}\right),\end{aligned} \tag{84}$$

Let $U_{i,j} = X_{i,j}^2 - \mu_{X_{1,j},2}$. By Assumption 1 and 3, for fixed j, $U_{i,j}$'s are of independent identical distribution and $\mathbb{E}[U_{i,j}] = 0$, $\mathbb{E}\left[U_{i,j}^2\right] = \mu_{U_{1,j},2} < \infty$, $\mathbb{E}\left[U_{i,j}^4\right] = \mu_{U_{1,j},4} < \infty$. Then we apply Chebyshev's inequality.

$$\begin{aligned} P\left(C_k[j,j] > \frac{4\gamma}{k\mu_{X_{1,j},2}}\right) &\leq \frac{\mathbb{E}\left[\left(\frac{1}{k-m}\sum_{i=1}^{k-m} U_{i,j}\right)^4\right]}{\left(\frac{1}{2}\mu_{X_{1,j},2}\right)^4} \\ &= \frac{\sum_{i=1}^{k-m}\mathbb{E}U_{i,j}^4 + 6\sum_{1\leq i<s\leq k-m}\mathbb{E}\left[U_{i,j}^2 U_{s,j}^2\right]}{(k-m)^4(\frac{1}{2}\mu_{X_{1,j},2})^4} \\ &\leq \frac{\mu_{U_{1,j},4}}{(k-m)^3\left(\frac{1}{2}\mu_{X_{1,j},2}\right)^4} + \frac{3\mu_{U_{1,j},2}^2}{(k-m)^2\left(\frac{1}{2}\mu_{X_{1,j},2}\right)^4}. \end{aligned} \tag{85}$$

We take the fourth moment since we want to construct a summable series of probabilities. Next we sum up those elements.

$$\begin{aligned} P\left(\|C_k\|_F^2 > \frac{2^4\gamma^2}{k^2}\sum_{j=1}^{p}\frac{1}{\mu_{X_{1,j},2}^2}\right) &\leq P\left(\bigcup_{j=1}^{p}\left\{C_k[j,j] > \frac{4\gamma}{k\mu_{X_{1,j},2}}\right\}\right) \\ &\leq \frac{1}{(k-m)^3}\sum_{j=1}^{p}\frac{\mu_{U_{1,j},4}}{\left(\frac{1}{2}\mu_{X_{1,j},2}\right)^4} + \frac{1}{(k-m)^2}\sum_{j=1}^{p}\frac{3\mu_{U_{1,j},2}^2}{\left(\frac{1}{2}\mu_{X_{1,j},2}\right)^4} \\ &\leq \frac{1}{(k-m)^2}\left[\sum_{j=1}^{p}\frac{\mu_{U_{1,j},4}}{\left(\frac{1}{2}\mu_{X_{1,j},2}\right)^4} + \sum_{j=1}^{p}\frac{3\mu_{U_{1,j},2}^2}{\left(\frac{1}{2}\mu_{X_{1,j},2}\right)^4}\right]. \end{aligned} \tag{86}$$

Similarly, we do the same procedure on $\left\|\tilde{J}_{C,k} - \tilde{J}_{Q,k}\right\|_F^4$. Let $\tilde{E} = \tilde{J}_{C,k} - \tilde{J}_{Q,k}$. By (79),

$$\tilde{E}_k[j,s] = \begin{cases} \frac{1}{\gamma}\sum_{i=1}^{k} U_{i,j} & j = s \\ \frac{1}{\gamma}\sum_{i=k-m+1}^{k}\left(X_{i,j}X_{i,s} - \mathbb{E}[X_{i,j}, X_{i,s}]\right) & j \neq s \end{cases} \tag{87}$$

For elements on the diagonal, applying Chebyshev's inequality we have

$$\begin{aligned} P\left(\tilde{E}_k[j,j]^2 > \frac{k^{7/4}}{\gamma^2}\right) &= P\left(\frac{1}{\gamma^2}\left(\sum_{i=1}^{k} U_{i,j}\right)^2 > \frac{k^{7/4}}{\gamma^2}\right) \\ &\leq \frac{\mathbb{E}\left[\left(\sum_{i=1}^{k} U_{i,j}\right)^4\right]}{k^{7/2}} \\ &\leq \frac{\mu_{U_{1,j},4}}{k^{5/2}} + \frac{3\mu_{U_{1,j},2}^2}{k^{3/2}}. \end{aligned} \tag{88}$$

The power $\frac{7}{4}$ is chosen to make sure that the series of probabilities is summable and the bound we get for $\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k\right]$ is in order $O(\frac{1}{k^t})$ with $t > 2$. For $j \neq s$, let $V_{i,j,s} = X_{i,j}X_{i,s} - \mathbb{E}\left[X_{i,j}X_{i,s}\right]$, and by Assumption 3 we have $\mathbb{E}V_{1,j,s} = 0$, $\mathbb{E}V_{1,j,s}^2 = \mu_{V_{1,j,s},2} < \infty$. Then,

$$\begin{aligned} P\left(\tilde{E}_k[j,s]^2 > \frac{k^{7/4}}{\gamma^2}\right) &= P\left(\frac{1}{\gamma^2}\left(\sum_{i=k-m+1}^{k}(V_{i,j,s})\right)^2 > \frac{k^{7/4}}{\gamma^2}\right) \\ &\leq \frac{\mathbb{E}\left[\left(\sum_{i=k-m+1}^{k}(V_{i,j,s})\right)^2\right]}{k^{7/4}} \\ &\leq \frac{m\mu_{V_{1,j,s},2}}{k^{7/4}}. \end{aligned} \tag{89}$$

Combine (88) and (89) we can control $\|\tilde{E}_k\|_F^2$,

$$P\left(\|\tilde{E}_k\|_F^2 > \frac{p^2k^{7/4}}{\gamma^2}\right) \leq \frac{1}{k^{3/2}}\left[\sum_{j=1}^{p}\mu_{U_{1,j},4} + 3\sum_{j=1}^{p}\mu_{U_{1,j},2}^2 + 2m\sum_{1\leq j<s\leq p}\mu_{V_{1,j,s},2}\right]. \tag{90}$$

Then combining (82), (74), (86) and (90), for $k \geq K_4 = \max\{2m, K_{1,2}(1), (\gamma^2 L)^{7/4}\}$,

$$P\left(\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k\right] > \frac{B_4}{k^{9/2}}\right) \leq \frac{B_5}{(k-m)^2} + \frac{B_6}{k^{3/2}}, \tag{91}$$

where

$$\begin{aligned} B_4 &= 2^{16}\left(\sum_{j=1}^{p}\frac{1}{\mu_{X_{1,j},2}^2}\right)^2\frac{p^2\gamma^4}{\xi_2^4}, \\ B_5 &= \sum_{j=1}^{p}\frac{\mu_{U_{1,j},4}}{\left(\frac{1}{2}\mu_{X_{1,j},2}\right)^4} + \sum_{j=1}^{p}\frac{3\mu_{U_{1,j},2}^2}{\left(\frac{1}{2}\mu_{X_{1,j},2}\right)^4}, \\ B_6 &= \sum_{j=1}^{p}\mu_{U_{1,j},4} + 3\sum_{j=1}^{p}\mu_{U_{1,j},2}^2 + 2m\sum_{1\leq j<s\leq p}\mu_{V_{1,j,s},2}. \end{aligned} \tag{92}$$

Then using (91) we have for $l \geq K_4$,

$$\begin{aligned} P\left(\bigcup_{k=l}^{\infty}\left\{\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k\right] > \frac{B_4}{k^{9/2}}\right\}\right) &\leq \sum_{k=l}^{\infty}\left(\frac{B_5}{(k-m)^2} + \frac{B_6}{k^{3/2}}\right) \\ &\leq B_5\int_l^{\infty}\frac{1}{(k-m-1)^2}\,\mathrm{d}k \\ &\quad + B_6\int_l^{\infty}\frac{1}{(k-1)^{3/2}}\,\mathrm{d}k \\ &= \frac{B_5}{l-m-1} + \frac{2B_6}{(l-1)^{1/2}}. \end{aligned} \tag{93}$$

Therefore for arbitrary δ, let $K_5 = \left\lceil \max\left\{2m, K_4, \frac{2B_5}{\delta} + m + 1, \left(\frac{4B_6}{\delta}\right)^2 + 1\right\}\right\rceil$. Then

$$P\left(\bigcup_{k=K_5}^{\infty}\left\{\mathbb{E}\left[\|E_{k+1}\|_F^4 \,|\mathcal{G}_k\right] > \frac{B_4}{k^{9/2}}\right\}\right) < \delta. \tag{94}$$

■