Journal of Multivariate Analysis 191 (2022) 105017

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Online statistical inference for parameters estimation with linear-equality constraints

Ruiqi Liu [a,*], Mingao Yuan [b], Zuofeng Shang [c]

[a] Department of Mathematics and Statistics, Texas Tech University, TX 79409, USA
[b] Department of Statistics, North Dakota State University, ND 58108, USA
[c] Department of Mathematical Sciences, New Jersey Institute of Technology, NJ 07102, USA

ARTICLE INFO

Article history:
Received 7 January 2022
Received in revised form 12 April 2022
Accepted 13 April 2022
Available online 23 May 2022

AMS 2020 subject classifications:
primary 62F12
secondary 62L20

Keywords:
Online inference
Constrained optimization
Projected stochastic gradient descent algorithm

ABSTRACT

Stochastic gradient descent (SGD) and projected stochastic gradient descent (PSGD) are scalable algorithms to compute model parameters in unconstrained and constrained optimization problems. In comparison with SGD, PSGD forces its iterative values into the constrained parameter space via projection. From a statistical point of view, this paper studies the limiting distribution of PSGD-based estimate when the true parameters satisfy some linear-equality constraints. Our theoretical findings reveal the role of projection played in the uncertainty of the PSGD-based estimate. As a byproduct, we propose an online hypothesis testing procedure to test the linear-equality constraints. Simulation studies on synthetic data and an application to a real-world dataset confirm our theory.

1. Introduction

With the rapid increase in availability of data in the past two decades or so, many classical optimization methods for statistical problems such as gradient descent, expectation–maximization or Fisher scoring cannot be applied in the presence of large datasets, or when the observations are collected one-by-one in an online fashion [4,19]. To overcome the difficulty in the era of big data, a computationally scalable algorithm called stochastic gradient descent (SGD) proposed in the seminal work [16] has been widely applied and achieved great success [1,5,22]. In comparison with classical optimization methods, one appealing feature of SGD is that the algorithm only requires accessing a single observation during each iteration, which makes it scale well with big data and computationally feasible with streaming data.

Due to the success of SGD, the studies of its theoretical properties have drawn a great deal of attention. The theoretical analysis of SGD can be categorized into two directions based on different research interests. The first direction is about the convergence rate. Existing literature shows that SGD algorithm can achieve a (in terms of regret) $O(1/T)$ convergence rate for strongly convex objective functions (e.g., see [2,8]), and a $O(1/\sqrt{T})$ rate for general convex cases [11], where T is the number of iterations. The second direction focuses on applying SGD to statistical inference. It was proved that the SGD estimate is asymptotic normal (e.g., see [12]) under suitable conditions. However, unlike classical parameter estimates, the SGD estimate may not be root-T consistent, and its convergence rate depends on the learning rate. To improve the convergence rate, [14,18] independently proposed the averaged stochastic gradient descent (ASGD) estimate, which was

* Corresponding author.
E-mail address: ruiqliu@ttu.edu (R. Liu).

https://doi.org/10.1016/j.jmva.2022.105017
0047-259X/

obtained by averaging the updated values in all iterations. They showed that the ASGD estimate is root-T consistent, while its asymptotic normality was proved by [15]. Following [15], there is a vast amount of work related to conducting statistical inference based on ASGD estimates. For example, [19] proposed a hierarchical incremental gradient descent (HIGrad) procedure to construct the confidence interval for the unknown parameters. In comparison with ASGD estimate, the flexible structure makes HiGrad easier to parallelize. In [4], the authors developed an online bootstrap algorithm to construct the confidence interval, which is still applicable when there is no explicit formula for the covariance matrix of the ASGD estimate. Recently, [3] proposed a plug-in estimate and a batch-means estimate for the asymptotic covariance matrix. With strong convexity assumption on the objective function, they proved the convergence rate of the estimates.

When there are constraints imposed on the parameters, the SGD algorithm is often combined with projection, which forces the iterated values into the constrained parameter space. The convergence rate of this projected stochastic gradient descent (PSGD) is also well studied (e.g., see [11]), which is proved to be the same as that of SGD. In the view of statistical inference, [9] studied the asymptotic distribution of PSGD estimate when the model parameters are in the interior of the constrained parameter space. It was proved that the projection operation only happens a finite number of times almost surely. As a consequence, the limiting distribution of PSGD estimate is exactly the same as that of SGD estimate. Recently, [6] studied the limiting distribution of averaged projected stochastic gradient descent (APSGD) estimate, which is the averaged version of PSGD. When the model parameters are in the interior of the constrained parameter space, APSGD and ASGD estimates have the same limiting distribution.

This paper aims to quantify the uncertainty in APSGD estimates when the model parameters satisfy some linear-equality constraints. Compared to the existing literature, a significant difference of our model is that the model parameters are not in the interior of the constrained parameter space. Therefore, the projection operation will take place during every iteration, and the limiting distribution of the APSGD estimate turns out to be a degenerate multivariate normal distribution. The contribution of current work is threefold:

(i) We derive the limiting distribution of the APSGD estimate, which is proved to be at least as efficient as ASGD estimate under mild conditions.
(ii) An online specification test for the linear-equality constraints is proposed based on the difference between APSGD and ASGD estimates.
(iii) Our findings reveal that, when the true parameters are not in the interior of the parameter space, the APSGD and ASGD estimates could have different limiting distributions.

This paper is organized as follows. In Section 2, we mathematically formulate the parameters estimation problem with linear-equality constraints. Section 3 proposes the APSGD estimate and studies its asymptotic properties. An online specification test is proposed in Section 4. All the mathematical proofs are deferred to the appendix. A set of Monte Carlo simulations to investigate the finite sample performance of the proposed methods and an application to a real-world dataset are provided in a supplementary material.

2. Problem formulation

We consider the problem to conduct statistical inference about the model parameter

$$\theta^* = \operatorname*{argmin}_{\theta\in\mathbb{R}^p}\big\{L(\theta) \coloneqq \mathrm{E}[l(\theta, Z)]\big\}, \tag{1}$$

where $l(\theta, Z)$ is the loss function, and Z is a single copy drawn from an unknown distribution F_{θ^*}. Moreover, we assume that additional information about the truth θ^* is available:

$$B\theta^* = b, \tag{2}$$

where B and b are some prespecified matrix and vector with comfortable dimensions. The loss function specified by (1) is quite general and covers many popular statistical models, which are illustrated by the following examples.

Example 1 (*Mean Estimation*)**.** Suppose $Z \in \mathbb{R}^p$ is random vector with mean $\theta^* = \mathrm{E}(Z)$. The loss function becomes $l(\theta, z) = \frac{1}{2}\|z-\theta\|^2$ with $\theta, z \in \mathbb{R}^p$.

Example 2 (*Linear Regression*)**.** Let the random vector be $Z = (Y, X^\top)^\top$ with $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$ satisfying $Y = X^\top\theta^* + \epsilon$. Here $\epsilon \in \mathbb{R}$ is the random noise with zero mean. The loss function can be chosen as $l(\theta, z) = \frac{1}{2}(y - x^\top\theta)^2$ with $y \in \mathbb{R}, x, \theta \in \mathbb{R}^p$, and $z = (y, x^\top)^\top$.

Example 3 (*Logistic Regression*)**.** Suppose that the observation $Z = (Y, X^\top)^\top$ with $Y \in \{-1, 1\}$ and $X \in \mathbb{R}^p$ satisfying $\Pr(Y = y|X = x) = [1 + \exp(-yx^\top\theta^*)]^{-1}$. The loss function is $l(\theta, z) = \log(1 + \exp(-yx^\top\theta))$ with $y \in \{-1, 1\}, x, \theta \in \mathbb{R}^p$, and $z = (y, x^\top)^\top$.

Example 4 (*Maximal Likelihood Estimation*)**.** Let F_{θ^*} be the distribution of Z, and the function form of F_{θ^*} is known except the value of θ^*. The loss function is the negative log likelihood: $l(\theta, z) = -\log(F_\theta(z))$.

In general, the function form of $L(\theta)$ is unknown, as it relies on the distribution F_{θ^*}. Instead, classical statistical methods estimate θ^* based on the sample counterpart of $L(\theta)$ as follows:

$$\tilde{\theta}_T = \underset{\theta\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{T}\sum_{t=1}^{T} l(\theta, Z_t) \quad \text{s.t.} \quad B\theta = b, \tag{3}$$

where $Z_1, \ldots, Z_T$ are the i.i.d. observations generated from distribution F_{θ^*}. However, the computation of $\tilde{\theta}_T$ in (3) involves calculating a summation among T terms, which is not efficient when sample size T is large. Moreover, in many real-world scenarios, the observations are collected sequentially in an online fashion. With the growing number of observations, data storage devices cannot store all the collected observations or there is no enough memory to load the whole dataset. In this case, the classical estimation procedures are not computationally feasible.

Before proceeding, we introduction some notation. Let $\|v\| = \sqrt{v^\top v}$ denote the Euclidean norm of the vector v. For any matrix $A \in \mathbb{R}^{q\times k}$, we define $\|A\| = \sup_{x\in\mathbb{R}^k} \sqrt{x^\top A^\top A x}$ as its operator norm, A^- as its Moore–Penrose inverse, and rank(A) as its rank. For two symmetric matrices $V_1, V_2 \in \mathbb{R}^{k\times k}$, we say $V_1 \succeq V_2$ if $x^\top V_1 x \geq x^\top V_2 x$ for all $x \in \mathbb{R}^k$. We use the notation $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{\mathbb{L}}$ to denote convergence in probability and in distribution, respectively. For $t \geq 1$, we denote $\mathcal{F}_t$ as the sigma algebra generated by $\{Z_1, \ldots, Z_t\}$. We denote $\chi^2(k)$ as the chi-square distribution with degree of freedom k, and $\chi^2(\delta, k)$ as the non-central chi-squared distribution with noncentrality parameter δ and degree of freedom k, for positive integer k and positive constant δ.

3. Projected Polyak–Ruppert averaging

To overcome the drawbacks of the classical methods, we consider the following PSGD algorithm. Choosing an initial value $\theta_0 \in \mathbb{R}^p$, we recursively update the value as follows:

$$\theta_t = \Pi(\theta_{t-1} - \gamma_t \nabla l(\theta_{t-1}, Z_t)), \tag{4}$$

where $\Pi(\cdot)$ is the projection operator onto the affine set $\{\theta \in \mathbb{R}^p : B\theta = b\}$, and $\gamma_t > 0$ is the predetermined learning rate (or step size). The updating equation in (4) can be explicitly written in matrix form as

$$\theta_t = c + P[\theta_{t-1} - \gamma_t \nabla l(\theta_{t-1}, Z_t) - c],$$

where $P \in \mathbb{R}^{p\times p}$ is the orthogonal projection matrix onto Ker(B), and $c \in \mathbb{R}^p$ is any vector satisfying $Bc = b$. Following [15], we define the APSGD estimate as follows:

$$\bar{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \theta_t. \tag{5}$$

By projection operation in (4), the estimate $\bar{\theta}_T$ satisfies (2). It is worth mentioning that, the average in (5) can be updated recursively in an online fashion as

$$\bar{\theta}_t = \frac{t-1}{t}\bar{\theta}_{t-1} + \frac{1}{t}\theta_t,$$

which is also obtainable with a large sample size. To discuss the theoretical properties of $\bar{\theta}_T$, we need the following Assumption.

Assumption A1. There exist constants $K, \epsilon > 0$ such that the following statements hold.

(i) The learning rate satisfies $\gamma_t = \gamma t^{-\rho}$, for some constants $\gamma > 0$ and $\rho \in (1/2, 1)$.
(ii) The objective function $L(\theta)$ is convex and continuously differentiable for all $\theta \in \mathbb{R}^p$. Moreover, it is twice continuously differentiable at $\theta = \theta^*$, where θ^* is the unique minimizer of $L(\theta)$.
(iii) For all $\theta, \tilde{\theta} \in \mathbb{R}^p$, the inequality $\|\nabla L(\theta) - \nabla L(\tilde{\theta})\| \leq K\|\theta - \tilde{\theta}\|$ holds.
(iv) The Hessian matrix $G := \nabla^2 L(\theta^*) \in \mathbb{R}^{p\times p}$ is positive definite. Furthermore, the inequality $\|\nabla^2 L(\theta) - \nabla^2 L(\theta^*)\| \leq K\|\theta - \theta^*\|$ holds for all θ with $\|\theta - \theta^*\| \leq \epsilon$.
(v) For all $\theta \in \mathbb{R}^p$, it holds that $E(\|\nabla l(\theta, Z)\|^2) \leq K(1 + \|\theta\|^2)$, and the matrix $S := \mathrm{E}(\nabla l(\theta^*, Z)\nabla l^\top(\theta^*, Z)) \in \mathbb{R}^{p\times p}$ is positive definite.
(vi) For all θ with $\|\theta - \theta^*\| \leq \epsilon$, it holds that $\mathrm{E}(\|\nabla l(\theta, Z) - \nabla l(\theta^*, Z)\|^2) \leq \delta(\|\theta - \theta^*\|)$, where $\delta(\cdot)$ is a function such that $\delta(v) \to$ as $v \to 0$.
(vii) For each $\theta \in \mathbb{R}^p$, there exist a constant $\epsilon_\theta > 0$ and a measurable function $M_\theta(z)$ with $\mathrm{E}(M_\theta(Z)) < \infty$ such that

$$\sup_{\tilde{\theta}:\|\tilde{\theta}-\theta\|\leq\epsilon_\theta} \|\nabla l(\tilde{\theta}, Z)\| \leq M_\theta(Z) \quad \text{almost surely.}$$

(viii) The projection matrix P satisfies $P^2 = P^\top = P$ and rank$(P) = d$ for some integer $d \in \{0, \ldots, p\}$.

Remark 1. Assumption A1(i) specifies the learning rate for tth iteration. The learning rate satisfies $\sum_{t=1}^{\infty}\gamma_t = \infty$ and $\sum_{t=1}^{\infty}\gamma_t^2 < \infty$, which is widely used in literature [4,15,19]. Assumptions A1(ii)–A1(vii) are regularity conditions about the objective function $L(\theta)$ and the lose function $l(\theta, z)$, which are standard and also adopted in [4]. Assumption A1(viii) is to characterize the linear-equality constraint $B\theta^* = b$. In particular, when $P = I$ and $d = p$, the APSGD estimate $\overline{\theta}_T$ in (5) becomes the ASGD estimate without projection in [15].

Theorem 1. *Under Assumption A1, it follows that*

$$\overline{\theta}_T = \theta^* - \frac{1}{T}\sum_{t=1}^{T}(PGP)^{-}\zeta_t + o_p(T^{-1/2}),$$

where $\zeta_t = \nabla l(\theta_{t-1}, Z_t) - \nabla L(\theta_{t-1})$. *Moreover, the following statement holds:*

$$\sqrt{T}(\overline{\theta}_T - \theta^*) \xrightarrow{\mathbb{L}} N(0, (PGP)^{-}S(PGP)^{-}).$$

Theorem 1 provides the asymptotic expansion and limiting distribution of the APSGD estimate $\overline{\theta}_T$. Notice that $\theta_{t-1} \in \mathcal{F}_{t-1}$, and Z_t is independent from $\mathcal{F}_{t-1}$, so $\mathrm{E}(\zeta_t|\mathcal{F}_{t-1}) = 0$, which implies that $\zeta_1, \ldots, \zeta_T$ is a martingale-difference process. Under Assumption A1, we can apply the martingale central limit theorem (e.g., see [13]) to derive the limiting distribution. It is worth mentioning the differences and connections between Theorem 1 and the existing results. First, [15] considered an unconstrained parameter space and showed that the ASGD estimate is asymptotically distributed as $N(0, G^{-1}SG^{-1})$. Theorem 1 can be viewed as an extension of [15] from $P = I$ to a general projection matrix P. Second, [6] studied the APSGD estimate when the model parameters are in the interior of the constrained parameter space, and they showed that APSGD have the same limiting distribution as PSGD. However, Theorem 1 reveals the different limiting distributions of APSGD and PSGD in our model. The reason behind this difference is that our model parameter θ^* is not in the interior of the constrained parameter space $\{\theta \in \mathbb{R}^p : B\theta = b\}$.

Let us revisit examples in previous section and investigate the limiting distributions of the corresponding APSGD estimates.

Example 1 (*Continued*)**.** Suppose the covariance of Z is Σ. We can verify $\nabla l(\theta, z) = -(z - \theta)$, $\nabla^2 l(\theta, z) = I$, $G = I$, and $S = \Sigma$. So the asymptotic covariance of the APSGD estimate is $P\Sigma P$.

Example 2 (*Continued*)**.** Suppose ϵ is independent from X with $\mathrm{E}(\epsilon) = 0$, $\mathrm{E}(\epsilon^2) = \sigma^2$. It can be verified that $\nabla l(\theta, z) = -(y - x^\top\theta)x$, $\nabla^2 l(\theta, z) = xx^\top$, $G = \mathrm{E}(XX^\top)$, and $S = \sigma^2\mathrm{E}(XX^\top) = \sigma^2 G$. Hence, the APSGD estimate is asymptotically with covariance matrix $\sigma^2(PGP)^{-}$.

Example 3 (*Continued*)**.** Suppose ϵ is independent from X with $\mathrm{E}(\epsilon) = 0$, $\mathrm{E}(\epsilon^2) = \sigma^2$, and $V = \mathrm{E}(XX^\top)$. It is not difficult to verify that

$$\nabla l(\theta, z) = \frac{-yx}{1 + \exp(yx^\top\theta)}, \quad \nabla^2 l(\theta, z) = \frac{\exp(yx^\top\theta)}{[1 + \exp(yx^\top\theta)]^2}xx^\top, \quad G = S = \mathrm{E}\left(\frac{\exp(X^\top\theta^*)}{[1 + \exp(X^\top\theta^*)]^2}XX^\top\right).$$

As a consequence, the APSGD estimate is asymptotically normal with covariance matrix $(PGP)^{-}$.

Example 4 (*Continued*)**.** Assume almost surely for all Z, the map $\theta \to F_\theta(Z)$ is twice continuously differentiable. Due to the properties of log likelihood function, the Fisher information matrix satisfies $I_{\theta^*} := \mathrm{E}[\nabla^2 l(\theta^*, Z)] = \mathrm{E}[\nabla l(\theta^*, Z)\nabla l(\theta^*, Z)] = G = S$. Therefore, we show that the covariance matrix is $(PI_{\theta^*}P)^{-}$.

It is worth discussing the role of the constraint (2) played in the estimation. For this purpose, let us denote $\overline{\theta}_{T,I}$ and $\overline{\theta}_{T,P}$ as the APSGD estimates using projection matrices I and P, respectively. By Theorem 1, their asymptotic covariance matrices are $V_I := G^{-1}SG^{-1}$ and $V_P := (PGP)^{-}S(PGP)^{-}$. For a general loss function $l(\theta, z)$, the performance $\overline{\theta}_{T,P}$ is not necessarily better than $\overline{\theta}_{T,I}$. To see this, let us consider a special case of Example 1.

Example 1 (*Continued*)**.** Suppose $\theta^* = (\theta_1^*, \theta_2^*)^\top \in \mathbb{R}^2$, $B = (1, -1)$ and $b = (0, 0)^\top$. The linear-equality constraint in (2) becomes $\theta_1^* = \theta_2^*$. Moreover, we assume $\Sigma = \mathrm{Diag}(\sigma^2, 3\sigma^2)$. We can verify that

$$V_P = \begin{pmatrix}\sigma^2 & \sigma^2\\ \sigma^2 & \sigma^2\end{pmatrix}, \quad V_I = \begin{pmatrix}\sigma^2 & 0\\ 0 & 3\sigma^2\end{pmatrix}.$$

As a consequence, neither $V_P \succeq V_I$ nor $V_I \succeq V_P$ holds.

However, for a board class of loss functions, the following Lemma suggests $\overline{\theta}_{T,P}$ is at least as efficient as $\overline{\theta}_{T,I}$.

Lemma 1. *Under Assumption A1, if* $S = cG$ *for some constant* $c > 0$, *then* $V_I = cG^{-1}$ *and* $V_P = c(PGP)^{-}$. *Moreover, it follows that* $V_I \succeq V_P$, *and the equality* $V_I = V_P$ *holds if and only if* $P = I$.

Lemma 1 indicates that, under an additional condition, the estimation performance of $\overline{\theta}_{T,P}$ is improved by utilizing the additional information in (2). The additional condition $S = cG$ holds for many popular models, including Examples 2–4. In particular, for the negative log likelihood loss function in Example 4, the asymptotic covariance matrix $(PI_{\theta^*}P)^-$ coincides the Cramér–Rao lower bound for constrained maximal likelihood model (e.g., see [7,10].

To apply Theorem 1, the unknown covariance matrix needs to be estimated. For this purpose, the following regularity conditions on $l(\theta, z)$ are imposed.

Assumption A2. There exists a constant $\epsilon > 0$ such that, for each θ with $\|\theta - \theta^*\| \le \epsilon$, the function $\theta \to l(\theta, Z)$ has a continuous Hessian matrix $\nabla^2 l(\theta, Z)$ almost surely. Moreover, there exists a measurable function $M(z)$ with $\mathrm{E}(M(Z)) < \infty$ satisfying $\|\nabla^2 l(\theta, Z)\| \le M(Z)$ for all θ with $\|\theta - \theta^*\| \le \epsilon$ almost surely.

The existence of the second-order derivatives of $\theta \to l(\theta, z)$ in Assumption A2 is to estimate $G = \nabla^2 L(\theta^*)$ based on its sample counterpart, while the dominating function $M(Z)$ is required to allow changing the order of the gradient operator and expectation, namely, $\nabla^2 \mathrm{E}[l(\theta^*, Z)] = \mathrm{E}[\nabla^2 l(\theta^*, Z)]$. To estimate the covariance matrix, let us define

$$\hat{G}_T = \frac{1}{T}\sum_{t=1}^{T} \nabla^2 l(\overline{\theta}_t, Z_t), \quad \hat{S}_T = \frac{1}{T}\sum_{t=1}^{T} \nabla l(\overline{\theta}_t, Z_t)\nabla l^\top(\overline{\theta}_t, Z_t), \tag{6}$$

which both can be recursively calculate by

$$\hat{G}_t = \frac{t-1}{t}\hat{G}_{t-1} + \frac{1}{t}\nabla^2 l(\overline{\theta}_t, Z_t), \quad \hat{S}_t = \frac{t-1}{t}\hat{S}_{t-1} + \frac{1}{t}\nabla l(\overline{\theta}_t, Z_t)\nabla l^\top(\overline{\theta}_t, Z_t).$$

The following lemma provides a consistent estimate for the covariance matrix.

Lemma 2. *Under Assumptions A1 and A2, it follows that* $(P\hat{G}_T P)^- \hat{S}_T (P\hat{G}_T P)^- = (PGP)^- S (PGP)^- + o_p(1)$.

Combining Theorem 1 with Lemma 2, we can construct an $(1-\alpha)\times 100\%$ confidence interval for the function $g(\theta^*)$ as

$$g(\overline{\theta}_T) \pm z_{\alpha/2}\sqrt{\frac{\nabla g^\top(\overline{\theta}_T)(P\hat{G}_T P)^- \hat{S}_T (P\hat{G}_T P)^- \nabla g(\overline{\theta}_T)}{T}}, \tag{7}$$

where $z_{\alpha/2}$ is the $\alpha/2\times 100\%$ upper quartile of standard normal. Since $\overline{\theta}_T$, $\hat{G}_T$ and $\hat{S}_T$ can be computed in an online fashion, so is the confidence interval in (7).

4. Specification test

As a byproduct of Theorem 1, we propose a specification test for the constraint in (2). Specifically, we aim to test the following hypotheses:

$$H_0 : B\theta^* = b \quad \text{vs.} \quad H_1 : B\theta^* = b + \beta \text{ for some } \beta \neq 0.$$

For this purpose, we define the test statistic

$$\kappa_T = T(\overline{\theta}_{T,P} - \overline{\theta}_{T,I})^\top \hat{W}^- (\overline{\theta}_{T,P} - \overline{\theta}_{T,I}). \tag{8}$$

Here $\hat{W} = (I-P)\hat{G}_{T,I}^{-1}\hat{S}_{T,I}\hat{G}_{T,I}^{-1}(I-P)$ is a weight matrix with $\hat{G}_{T,I}$ and $\hat{S}_{T,I}$ being the matrices in (6) calculated using projection matrix I. Essentially, $\hat{W}$ estimates the weight matrix $W = (I-P)G^{-1}SG^{-1}(I-P)$. The idea of the proposed test statistic in (8) is simple and straightforward. Under H_0, both $\overline{\theta}_{T,P}$ and $\overline{\theta}_{T,I}$ consistently estimate θ^*. Hence, their difference, as well as κ_T, should be around zero. However, under H_1, due to model misspecification, $\overline{\theta}_{T,P}$ is inconsistent, and the difference $\overline{\theta}_{T,P} - \overline{\theta}_{T,I}$ does not vanish. Based on (8), we propose the following asymptotic size α testing procedure:

$$\text{reject } H_0 \quad \text{if } \kappa_T > \chi^2_\alpha(p-d), \tag{9}$$

where $\chi^2_\alpha(p-d)$ is the $\alpha\times 100\%$ upper quartile of χ^2 distribution with degree $p-d$. The following theorem reveals the limiting behavior of the statistic κ_T and the validity of the proposed testing procedure.

Theorem 2. *Suppose Assumptions A1 and A2 are satisfied. Then the following statements are true:*

(i) *Under* $H_0 : B\theta^* = b$, *the convergence* $\kappa_T \xrightarrow{\mathbb{L}} \chi^2(p-d)$ *holds.*

(ii) *Under* $H_1 : B\theta^* = b + \beta$ *for some* $\beta \neq 0$, *it follows that* $\kappa_T \to \infty$ *in probability.*

(iii) *Under* $H_a : B\theta^* = b + \frac{\beta}{\sqrt{T}}$ *for some* $\beta \neq 0$, *it holds that* $\kappa_T \xrightarrow{\mathbb{L}} \chi^2(\mu^\top W^- \mu, p-d)$, *where* $\mu \in \mathbb{R}^p$ *is any vector satisfying* $B\mu = \beta$.

As a consequence, for any $\alpha \in (0, 1)$, it follows that

$$\lim_{T\to\infty} \Pr(\kappa_T > \chi^2_\alpha(p-d)|H_0) = \alpha, \quad \lim_{T\to\infty} \Pr(\kappa_T > \chi^2_\alpha(p-d)|H_1) = 1.$$

Theorem 2 provides the asymptotic distributions of κ_T under null hypothesis H_0 and local alternative hypothesis H_a, which are chi-square and noncentral chi-squared, respectively. Moreover, it shows that κ_T will diverge under alternative hypothesis H_1. Consequently, it verifies that testing procedure in (9) is consistent and has an asymptotic size α.

CRediT authorship contribution statement

Ruiqi Liu: Conceptualization, Methodology, Writing – review & editing. **Mingao Yuan:** Software, Validation. **Zuofeng Shang:** Supervision.

Acknowledgments

The authors gratefully acknowledge the constructive comments and suggestions from the Editor-in-Chief Dr. Dietrich von Rosen, an associate editor, and two anonymous referees. Zuofeng Shang acknowledges supports by National Science Foundation DMS-1764280 and DMS-1821157.

Appendix

In the appendix, we collect all the mathematical proofs of the main theorems and related lemmas.

Lemma A.1. *Let $H \in \mathbb{R}^{p\times p}$ be a positive definite matrix and suppose that $\gamma_t = \gamma t^{-\rho}$ for some constants $\gamma > 0$, $\rho \in (1/2, 1)$. Let us define squared matrices*

$$W_j^j = I, \quad W_j^t = (I - \gamma_{t-1}H)W_j^{t-1} = \cdots = \prod_{k=j}^{t-1}(I - \gamma_k H) \quad \text{for } t \geq j,$$

$$\overline{W}_j^t = \gamma_j \sum_{i=j}^{t-1} W_j^i = \gamma_j \sum_{i=j}^{t-1}\prod_{k=j}^{i-1}(I - \gamma_k H).$$

Then the following statements hold:

(i) *There are constants $K > 0$ such that $\|\overline{W}_j^t\| \leq K$ for all j and all $t \geq j$.*
(ii) $\frac{1}{t}\sum_{j=0}^{t-1} \|\overline{W}_j^t - H^{-1}\| \to 0$ *as* $t \to \infty$.

Proof. This is Lemma 1 of [15]. □

Lemma A.2. *Let $A \in \mathbb{R}^{p\times p}$ be a positive definite matrix and P be a projection matrix such that $P = P^\top$ and rank$(P) = d$. Then there exists an orthonormal matrix $U \in \mathbb{R}^{p\times p}$ such that*

$$U^\top PU = \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix}, \quad U^\top PAPU = \begin{pmatrix} \Omega_d & 0 \\ 0 & 0 \end{pmatrix}, \quad U^\top (PAP)^- U = \begin{pmatrix} \Omega_d^{-1} & 0 \\ 0 & 0 \end{pmatrix},$$

where $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix, and Ω_d is a diagonal matrix with diagonal elements $\rho_1, \ldots, \rho_d > 0$. Moreover, it follows that $(PAP)^- P = (PAP)^-$, $P(PAP)^- = (PAP)^-$ and $(PAP)^-(PAP)x = x$ for all x satisfying $Px = x$.

Proof. For any $x \in \mathbb{R}^p$ with $PAPx = 0$, it holds that $x^\top PAPx = 0$ and $Px = 0$ by the positive definiteness of A. Clearly, $Px = 0$ implies $PAPx = 0$. Therefore, we conclude that Ker(PAP) = Ker(P) and rank(PAP) = rank$(P) = d$.

For simplicity, we denote $S = PAP$. By direct examination, S and P are diagonalizable, and they commute. By simple linear algebra, there exist eigenvectors $u_1, u_2, \ldots, u_p$ that simultaneously diagonalize P and S. W.L.O.G, we assume $Pu_i = u_i$ for $i \in \{1, \ldots, d\}$ and $Pu_i = 0$ for $i \in \{d+1, \ldots, p\}$. We further assume $\rho_1, \rho_2, \ldots, \rho_p$ to be the eigenvalues of S corresponding to the eigenvectors $u_1, u_2, \ldots, u_p$. By the above notation, it shows that

$$\rho_i u_i = Su_i = PAPu_i = 0 \quad \text{for } i \in \{d+1, \ldots, p\}.$$

Since rank$(S) = d$, we conclude that $\rho_i > 0$ for $i \in \{1, \ldots, d\}$. As a consequence, $U = (u_1, \ldots, u_p)$ and $\Omega_d = \text{Diag}(\rho_1, \ldots, \rho_p)$ will be the desired choices. Moreover, it is not difficult to verify that

$$(PAP)^- P = U\begin{pmatrix} \Omega_d^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^\top U \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix} U^\top = U \begin{pmatrix} \Omega_d^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^\top = (PAP)^-.$$

Similarly, we can prove that $P(PAP)^- = (PAP)^-$. Suppose that x satisfies $Px = x$, then $x = \sum_{i=1}^d c_i u_i$ for some $c_1, \ldots, c_d \in \mathbb{R}$. As a consequence, it follows that $PAPx = PAP\sum_{i=1}^d c_i u_i = \sum_{i=1}^d c_i \rho_i u_i$. Notice that $PAP = \sum_{i=1}^d \rho_i u_i u_i^\top$ and $(PAP)^- = \sum_{i=1}^d \rho_i^{-1} u_i u_i^\top$, we have $(PAP)^-(PAP)x = \sum_{i=1}^d \rho_i^{-1} u_i u_i^\top \sum_{i=1}^d c_i \rho_i u_i = \sum_{i=1}^d c_i u_i = x$. □

Lemma A.3. *Under Assumption A1, it follows that*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{t-1} \|\gamma_j \sum_{k=j}^{t-1} \prod_{i=j+1}^{k} P(I - \gamma_i G)P - (PGP)^-\| = 0.$$

Moreover, there is a constant $K > 0$ such that $\|\gamma_j \sum_{k=j}^{t-1} \prod_{i=j+1}^{k} P(I - \gamma_i G)P\| \le K$ for all j and all $t \ge j$.

Proof. Since G is positive definite by Assumption A1(iv), it follows from Lemma A.2 that

$$U^\top P(I - \gamma_i G)PU = U^\top PU - \gamma_i U^\top PGPU = \begin{pmatrix} I_d - \gamma_i \Omega_d & 0 \\ 0 & 0 \end{pmatrix},$$

where U is an orthonormal matrix, $I_d \in \mathbb{R}^{d\times d}$ is the identity matrix, and $\Omega_d \in \mathbb{R}^{d\times d}$ is a diagonal and positive definite matrix. As a consequence, we have

$$\prod_{i=j+1}^{k} P(I - \gamma_i G)P = U \begin{pmatrix} \prod_{i=j+1}^{k}(I_d - \gamma_i \Omega_d) & 0 \\ 0 & 0 \end{pmatrix} U^\top.$$

By Lemma A.1, we have

$$\lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{t-1} \|\gamma_j \sum_{k=j}^{t-1} \prod_{i=j+1}^{k} (I_d - \gamma_i \Omega_d) - \Omega_d^{-1}\| = 0,$$

which further leads to the first statement according to Lemma A.2. Applying Lemma A.2 again, we obtain the second conclusion. □

Lemma A.4. *Let c_1 and c_2 be arbitrary positive constants. Support that $\gamma_t = \gamma t^{-\rho}$ for some constants $\gamma > 0$ and $\rho \in (1/2, 1)$. Moreover, assume a sequence $\{B_t\}_{t=1}^\infty$ satisfies*

$$B_t \le \frac{\gamma_{t-1}(1 - c_1\gamma_t)}{\gamma_t} B_{t-1} + c_2\gamma_t.$$

Then $\sup_{1\le t<\infty} B_t < \infty$.

Proof. This Lemma A.10 in [19]. □

Lemma A.5. *Let $F(x)$ be a differentiable convex function defined on $\mathbb{R}^p$ with a unique minimizer x^*. Suppose there exist constants $\rho, r > 0$ such that $x \to F(x) - \frac{\rho}{2}\|x\|^2$ is convex for all x with $\|x - x^*\| \le r$. Then for all $x \in \mathbb{R}^p$, it holds that $(x - x^*)^\top \nabla F(x) \ge \rho\|x - x^*\| \min\{\|x - x^*\|, r\}$.*

Proof. This is Lemma B.1 in [19]. □

Proof of Theorem 1. We sketch the proof of Theorem 1. By iteration formula in (4), we have

$$\theta_t = c + P(\theta_{t-1} - \gamma_t y_t - c), \quad y_t = \nabla l(\theta_{t-1}, Z_t) = \nabla L(\theta_{t-1}) + [\nabla l(\theta_{t-1}, Z_t) - \nabla L(\theta_{t-1})] := R(\theta_{t-1}) + \zeta_t, \tag{10}$$

where $c \in \mathbb{R}^p$ is any vector satisfying $Bc = b$. Let $\Delta_t = \theta_t - \theta^*$. Since $P(\theta^* - c) = \theta^* - c$, it follows that

$$\begin{aligned}
\Delta_t = \theta_t - \theta^* &= c + P(\theta_{t-1} - \gamma_t y_t - c) - \theta^* = c + P(\Delta_{t-1} - \gamma_t y_t + \theta^* - c) - \theta^* = P\Delta_{t-1} - \gamma_t P y_t \\
&= P\Delta_{t-1} - \gamma_t PR(\theta_{t-1}) - \gamma_t P\zeta_t = P\Delta_{t-1} - \gamma_t PG\Delta_{t-1} - \gamma_t P\zeta_t - \gamma_t P(R(\theta_{t-1}) - G\Delta_{t-1}) \\
&= P(I - \gamma_t G)\Delta_{t-1} - \gamma_t P\zeta_t - \gamma_t P(R(\theta_{t-1}) - G\Delta_{t-1}) \\
&= \left[\prod_{j=1}^{t} P(I - \gamma_j G)\right]\Delta_0 + \sum_{j=1}^{t}\left[\prod_{i=j+1}^{t} P(I - \gamma_i G)P\right]\gamma_j P\zeta_j + \sum_{j=1}^{t}\left[\prod_{i=j+1}^{t} P(I - \gamma_i G)P\right]\gamma_j P(R(x_{j-1}) - G\Delta_{j-1}).
\end{aligned}$$

Taking average, we show that

$$\frac{1}{T}\sum_{t=1}^{T}\Delta_t = \frac{1}{T}\sum_{t=1}^{T}\Big[\prod_{j=1}^{t}P(I-\gamma_j G)\Big]\Delta_0 + \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{t}\Big[\prod_{i=j+1}^{t}P(I-\gamma_i G)P\Big]\gamma_j P\zeta_j + \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{t}\Big[\prod_{i=j+1}^{t}P(I-\gamma_i G)P\Big]\gamma_j P(R(x_{j-1})-G\Delta_{j-1}) := S_1+S_2+S_3. \tag{11}$$

In Lemmas A.10 and A.11, we will show that

$$S_1+S_2=\frac{1}{T}(PGP)^{-}\sum_{t=1}^{T}\zeta_t+o_p(T^{-1/2}),\quad S_3=o_p(T^{-1/2}).$$

Finally, we prove the asymptotic normality based on martingale C.L.T. in Lemma A.12. □

Lemma A.6. *Under Assumption A1, the following statements hold for some constants $\epsilon, K>0$.*

(i) $(\theta-\theta^*)^\top R(\theta)\ge\epsilon\|\theta-\theta^*\|\min\{\|\theta-\theta^*\|,\epsilon\}$ *for all* $\theta\in\mathbb{R}^p$.
(ii) $E(\zeta_t|\mathcal{F}_{t-1})=0$.
(iii) $E(\|\zeta_t\|^2|\mathcal{F}_{t-1})\le K(1+\|\theta_{t-1}\|^2)$ *almost surely.*
(iv) $\|R(\theta)\|^2\le K(1+\|\theta\|^2)$.
(v) $\|R(\theta)-G(\theta-\theta^*)\|\le K\|\theta-\theta^*\|^2$ *for all* θ *with* $\|\theta-\theta^*\|\le\epsilon$.

Proof. For statement (i), by Assumption A1(iv), we know $L(\theta)$ satisfies the conditions in Lemma A.5 with some $\rho, r>0$. Therefore, it follows that

$$(\theta-\theta^*)^\top R(\theta)=(\theta-\theta^*)^\top\nabla L(\theta)\ge\epsilon\|\theta-\theta^*\|\min\{\|\theta-\theta^*\|,\epsilon\},$$

where $\epsilon=\min\{\rho,r\}$.

For statement (ii), since $\theta_{t-1}\in\mathcal{F}_{t-1}$ and Z_t is independent from $\mathcal{F}_{t-1}$, we have $\mathrm{E}(\nabla l(\theta_{t-1},Z_t)|\mathcal{F}_{t-1})=\nabla L(\theta_{t-1})$.

Similarly, by Assumption A1(v), the statement (iii) follows from the inequality below:

$$\mathrm{E}(\|\zeta_t\|^2|\mathcal{F}_{t-1})=\mathrm{E}[\|\nabla l(\theta_{t-1},Z_t)-\nabla L(\theta_{t-1})\|^2|\mathcal{F}_{t-1}]\le\mathrm{E}(\|\nabla l(\theta_{t-1},Z)\|^2|\mathcal{F}_{t-1})\le K(1+\|\theta_{t-1}\|^2).$$

Statement (iv) follows, as $\|R(\theta)\|^2=\|\nabla L(\theta)\|^2=\|\nabla L(\theta)-\nabla L(\theta^*)\|^2\le K^2\|\theta-\theta^*\|^2\le 2K^2(\|\theta^*\|^2+\|\theta\|^2)$ by Assumption A1(iii).

To prove statement (v), by Assumption A1(iv) and Taylor expansion, we have

$$\begin{aligned}\|R(\theta)-G(\theta-\theta^*)\|&=\|R(\theta)-R(\theta^*)-G(\theta-\theta^*)\|=\|R(\theta)-R(\theta^*)-\nabla^2L(\theta^*)(\theta-\theta^*)\|\\&=\|\nabla^2L(\tilde\theta)(\theta-\theta^*)-\nabla^2L(\theta^*)(\theta-\theta^*)\|\le K\|\theta-\theta^*\|^2\ \text{ for all }\theta\text{ with }\|\theta-\theta^*\|\le\epsilon,\end{aligned}$$

where $\tilde\theta$ is a vector between θ and θ^*. □

Lemma A.7. *Suppose Assumption A1 holds. Then there exists a constant $K>0$ such that*

$$\|R(\theta)-G(\theta-\theta^*)\|\le K\|\theta-\theta^*\|^2\ \text{ for all }\theta\in\mathbb{R}^p.$$

Proof. By Assumptions A1(iii) and A1(iv), we have

$$\begin{aligned}\|R(\theta)-G(\theta-\theta^*)\|&=\|R(\theta)-R(\theta^*)-G(\theta-\theta^*)\|\le\|R(\theta)-R(\theta^*)\|+\|G(\theta-\theta^*)\|\le(K+\|G\|)\|\theta-\theta^*\|\\&\le(K+\|G\|)\|\theta-\theta^*\|^2/\epsilon\ \text{ for all }\theta\text{ with }\|\theta-\theta^*\|\ge\epsilon.\end{aligned}$$

Combining with statement (v) in Lemma A.6, we complete the proof. □

Lemma A.8. *Under Assumption A1, it holds that $\lim_{t\to\infty}\theta_t=\theta^*$ almost surely.*

Proof. Notice that $\theta_t-c\in\mathrm{Ker}(B)$ for all $t\ge1$, so it follows that

$$\theta_t-\theta^*=c+P[\theta_{t-1}-\gamma_t\nabla l(\theta_{t-1},Z_t)-c]-\theta^*=\theta_{t-1}-\theta^*-\gamma_tP[R(\theta_{t-1})+\zeta_t].$$

Moreover $P\Delta_t=\Delta_t$ for $t\ge1$, we have

$$\begin{aligned}\|\Delta_t\|^2&=\|\Delta_{t-1}-\gamma_tPR(\theta_{t-1})-\gamma_tP\zeta_t\|^2=\|\Delta_{t-1}-\gamma_tPR(\theta_{t-1})\|^2-2\gamma_t\Delta_{t-1}^\top P\zeta_t+2\gamma_t^2R^\top(\theta_{t-1})P\zeta_t+\gamma_t^2\|P\zeta_t\|^2\\&=\|\Delta_{t-1}\|^2+\gamma_t^2\|PR(\theta_{t-1})\|^2-2\gamma_t\Delta_{t-1}^\top R(\theta_{t-1})-2\gamma_t\Delta_{t-1}^\top\zeta_t+2\gamma_t^2R^\top(\theta_{t-1})P\zeta_t+\gamma_t^2\|P\zeta_t\|^2\\&\le\|\Delta_{t-1}\|^2+\gamma_t^2\|R(\theta_{t-1})\|^2-2\gamma_t\Delta_{t-1}^\top R(\theta_{t-1})-2\gamma_t\Delta_{t-1}^\top\zeta_t+2\gamma_t^2R^\top(\theta_{t-1})P\zeta_t+\gamma_t^2\|\zeta_t\|^2\ \text{ for all }t\ge2.\end{aligned}\tag{12}$$

Taking conditional expectation on both sides of (12) and by Lemma A.6, we show that there exist constants $K, \epsilon > 0$ such that

$$\begin{aligned}
\mathrm{E}(\|\Delta_t\|^2|\mathcal{F}_{t-1}) &\le \|\Delta_{t-1}\|^2 - 2\gamma_t\Delta_{t-1}^\top R(\theta_{t-1}) + \gamma_t^2\|R(\theta_{t-1})\|^2 + \gamma_t^2\mathrm{E}(\|\zeta_t\|^2|\mathcal{F}_{t-1}) \\
&\le \|\Delta_{t-1}\|^2 - 2\gamma_t\Delta_{t-1}^\top R(\theta_{t-1}) + \gamma_t^2K(1+\|\theta_{t-1}\|^2) + \gamma_t^2K(1+\|\theta_{t-1}\|^2) \\
&= \|\Delta_{t-1}\|^2 - 2\gamma_t\Delta_{t-1}^\top R(\theta_{t-1}) + 2\gamma_t^2K(1+\|\theta_{t-1}\|^2) \\
&\le \|\Delta_{t-1}\|^2 + 2\gamma_t^2K(1+2\|\theta^*\|^2+2\|\Delta_{t-1}\|^2) - 2\gamma_t\Delta_{t-1}^\top R(\theta_{t-1}) \\
&= (1+4\gamma_t^2K)\|\Delta_{t-1}\|^2 + 2\gamma_t^2K(1+2\|\theta^*\|^2) - 2\gamma_t\Delta_{t-1}^\top R(\theta_{t-1}) \\
&\le (1+4\gamma_t^2K)\|\Delta_{t-1}\|^2 + 2\gamma_t^2K(1+2\|\theta^*\|^2) - 2\gamma_t\epsilon\|\Delta_{t-1}\|\min\{\|\Delta_{t-1}\|,\epsilon\}. \qquad (13)
\end{aligned}$$

Since $\sum_{t=1}^\infty \gamma_t = \infty$ and $\sum_{t=1}^\infty \gamma_t^2 < \infty$, applying Robbins–Siegmund Theorem (e.g., see [17]), we have $\|\Delta_t\|^2 \to V$ almost surely for some random variable V, and

$$\sum_{t=1}^\infty 2\gamma_t\epsilon\|\Delta_{t-1}\|\min\{\|\Delta_{t-1}\|,\epsilon\} < \infty \ \text{ almost surely.}$$

As a consequence, it follows that $\lim_{t\to\infty}\|\Delta_{t-1}\| \to 0$ almost surely. □

Lemma A.9. *Suppose Assumption A1 holds. Then for any $M > 0$, there exists a constant $K_M > 0$ such that*

$$E[\|\theta_t - \theta^*\|^2 I(\tau_M > t)] \le K_M\gamma_t \ \text{ for all } t \ge 0,$$

where $\tau_M = \inf\{i \ge 1 : \|\theta_i - \theta^*\| > M\}$ is a stopping time.

Proof. By Lemma A.8, for any $\delta > 0$, there exists a $M > 0$ such that $\Pr(\sup_{1\le t<\infty}\|\theta_t - \theta^*\| \le M) \ge 1-\delta$. Notice $\{\tau_M > t\} \in \mathcal{F}_t$ and on event $\{\tau_M > t\}$, $\|\theta_1 - \theta^*\|, \ldots, \|\theta_t - \theta^*\|$ are bounded by M, using (12), Lemmas A.6 and A.7, we have

$$\begin{aligned}
\|\Delta_t\|^2 I(\tau_M > t) &\le \|\Delta_t\|^2 I(\tau_M > t-1) \\
&\le \left(\|\Delta_{t-1}\|^2 + \gamma_t^2\|R(\theta_{t-1})\|^2 - 2\gamma_t\Delta_{t-1}^\top R(\theta_{t-1}) - 2\gamma_t\Delta_{t-1}^\top\zeta_t + 2\gamma_t^2R^\top(\theta_{t-1})P\zeta_t + \gamma_t^2\|\zeta_t\|^2\right) \\
&\quad \times I(\tau_M > t-1).
\end{aligned}$$

By similar calculation in (13), we show that there exist constants $K, \epsilon > 0$ such that

$$\mathrm{E}(\|\Delta_t\|^2 I(\tau_M > t)|\mathcal{F}_{t-1}) \le \left(1 + 4\gamma_t^2K\|\Delta_{t-1}\|^2 + 2\gamma_t^2K(1+2\|\theta^*\|^2) - 2\gamma_t\epsilon\|\Delta_{t-1}\|\min\{\|\Delta_{t-1}\|,\epsilon\}\right) I(\tau_M > t-1).$$

Notice that $\|\Delta_{t-1}\|\min\{\|\Delta_{t-1}\|,\epsilon\} = \|\Delta_{t-1}\|^2$ if $\|\Delta_{t-1}\| \le \epsilon$, and $\|\Delta_{t-1}\|\min\{\|\Delta_{t-1}\|,\epsilon\} = \|\Delta_{t-1}\|\epsilon \ge \|\Delta_{t-1}\|^2\epsilon/M$ if $\epsilon < \|\Delta_{t-1}\| \le M$, we conclude that

$$\begin{aligned}
\mathrm{E}(\|\Delta_t\|^2 I(\tau_M > t)|\mathcal{F}_{t-1}) &\le \left((1+4\gamma_t^2K)\|\Delta_{t-1}\|^2 + 2\gamma_t^2K(1+2\|\theta^*\|^2) - 2\gamma_t\epsilon^2M^{-1}\|\Delta_{t-1}\|^2\right) I(\tau_M > t-1) \\
&\le (1 - 2\gamma_t\epsilon^2M^{-1} + 4\gamma_t^2K)\|\Delta_{t-1}\|^2 I(\tau_M > t-1) + 2\gamma_t^2K(1+2\|\theta^*\|^2).
\end{aligned}$$

where we use the fact that $\|\Delta_{t-1}\| \le M$ on event $\{\tau_M > t-1\}$. Taking expectation again, if $\gamma_t \le \epsilon^2/(4MK)$, then it follows that

$$\begin{aligned}
\mathrm{E}[\|\Delta_t\|^2 I(\tau_M > t)] &\le (1 - 2\gamma_t\epsilon^2M^{-1} + 4\gamma_t^2K)\mathrm{E}[\|\Delta_{t-1}\|^2 I(\tau_M > t-1)] + 2\gamma_t^2K(1+2\|\theta^*\|^2) \\
&\le (1 - \gamma_t\epsilon^2M^{-1})\mathrm{E}[\|\Delta_{t-1}\|^2 I(\tau_M > t-1)] + 2\gamma_t^2K(1+2\|\theta^*\|^2).
\end{aligned}$$

Applying Lemma A.4, we conclude that, there exists a constant $K_M > 0$ such that $\mathrm{E}[\|\Delta_t\|^2 I(\tau_M > t)] \le K_M\gamma_t$ for all $t \ge 0$. □

Lemma A.10. *Under Assumption A1, it follows that*

$$S_1 + S_2 = \frac{1}{T}(PGP)^- \sum_{t=2}^T \zeta_t + o_p(T^{-1/2}),$$

where S_1 and S_2 are defined in (11).

Proof. Let $\hat\theta_0 = \theta_0 \in \mathbb{R}^p$ be the initial value for iteration. We define sequence

$$\hat\theta_t = c + P(\hat\theta_{t-1} - \gamma_t h_t - c) \text{ with } h_t = G\hat\theta_{t-1} - G\theta^* + \zeta_t, \ \text{ for } t \ge 1,$$

where $G \in \mathbb{R}^{p\times p}$ is the positive definite matrix defined in Assumption A1(iv), ζ_t is the process defined in (10), and $c \in \mathbb{R}^p$ satisfies $Bc = b$. The proof is divided into four steps.

Step 1: This step is to show that $\lim_{t\to\infty}\hat{\theta}_t = \theta^*$ almost surely. Let us define $\hat{\Delta}_t = \hat{\theta}_t - \theta^*$, which is different from $\Delta_t = \theta_t - \theta^*$. By the fact that $P(\theta^* - c) = \theta^* - c$, we have

$$\begin{aligned}\hat{\Delta}_t &= c + P(\hat{\theta}_{t-1} - \gamma_t h_t - c) - \theta^* = c + P(\hat{\Delta}_{t-1} - \gamma_t h_t + \theta^* - c) - \theta^* = P\hat{\Delta}_{t-1} - \gamma_t P h_t \\ &= P\hat{\Delta}_{t-1} - \gamma_t P(G\hat{\theta}_{t-1} - G\theta^* + \zeta_t) = P\hat{\Delta}_{t-1} - \gamma_t PG\hat{\Delta}_{t-1} - \gamma_t P\zeta_t = P(I - \gamma_t G)\hat{\Delta}_{t-1} - \gamma_t P\zeta_t. \end{aligned} \tag{14}$$

As a consequence, it follows from (14) that

$$\begin{aligned}\|\hat{\Delta}_t\|^2 &= \|P(I - \gamma_t G)\hat{\Delta}_{t-1}\|^2 - 2\gamma_t \zeta_t^\top P(I - \gamma_t G)\hat{\Delta}_{t-1} + \gamma_t^2\|P\zeta_t\|^2 \\ &\le (1 - \gamma_t\lambda)^2\|\hat{\Delta}_{t-1}\|^2 - 2\gamma_t\zeta_t^\top P(I - \gamma_t G)\hat{\Delta}_{t-1} + \gamma_t^2\|\zeta_t\|^2, \end{aligned} \tag{15}$$

where $\lambda > 0$ is the smallest eigenvalue of G. Taking conditional expectation, it follows that

$$\begin{aligned}\mathrm{E}(\|\hat{\Delta}_t\|^2|\mathcal{F}_{t-1}) &\le (1 - 2\gamma_t\lambda + \gamma_t^2\lambda^2)\|\hat{\Delta}_{t-1}\|^2 + \gamma_t^2\mathrm{E}(\|\zeta_t\|^2|\mathcal{F}_{t-1}) = (1 + \gamma_t^2\lambda^2)\|\hat{\Delta}_{t-1}\|^2 + \gamma_t^2\mathrm{E}(\|\zeta_t\|^2|\mathcal{F}_{t-1}) \\ &\quad - 2\gamma_t\lambda\|\hat{\Delta}_{t-1}\|^2 \\ &\le (1 + \gamma_t^2\lambda^2)\|\hat{\Delta}_{t-1}\|^2 + \gamma_t^2 K(1 + \|\theta_t\|^2) - 2\gamma_t\lambda\|\hat{\Delta}_{t-1}\|^2 \\ &\le (1 + \gamma_t^2\lambda^2)\|\hat{\Delta}_{t-1}\|^2 + \gamma_t^2 K(1 + 2\|\Delta_t\|^2 + 2\|\theta^*\|^2) - 2\gamma_t\lambda\|\hat{\Delta}_{t-1}\|^2, \end{aligned}$$

where Lemma A.6(iii) is used. Since $\lim_{t\to\infty}\|\Delta_t\| = 0$ almost surely by Lemma A.8 and $\sum_{t=1}^{\infty}\gamma_t^2 < \infty$ by Assumption A1(i), it follows that $\sum_{t=1}^{\infty}\gamma_t^2 K(1+2\|\Delta_t\|^2+2\|\theta^*\|^2) < \infty$ almost surely. Hence, Robbins–Siegmund Theorem (e.g., see [17]) implies that

$$\lim_{t\to\infty}\|\hat{\Delta}_t\|^2 \to \hat{V}, \quad \sum_{t=1}^{\infty}\gamma_t\|\hat{\Delta}_t\|^2 < \infty \ \text{ almost surely},$$

for some random variable $\hat{V}$. Since $\sum_{t=1}^{\infty}\gamma_t = \infty$, we conclude that $\lim_{t\to\infty}\|\hat{\Delta}_t\|^2 = 0$ almost surely.
Step 2: Let us define stopping times $\hat{\tau}_M = \inf\{j \ge 1 : \|\hat{\Delta}_j\| > M\}$ and $\tau_M = \inf\{j \ge 1 : \|\Delta_j\| > M\}$ for $M > 0$. This step is to prove that for any $M > 0$, there exists a constant $K_M > 0$ such that

$$\mathrm{E}[\|\hat{\Delta}_t\|^2 I(\hat{\tau}_M > t, \tau_M > t)] \le K_M\gamma_t \ \text{ for all } t \ge 1. \tag{16}$$

Using (15) again, we have

$$\begin{aligned}\|\hat{\Delta}_t\|^2 I(\hat{\tau}_M > t, \tau_M > t) &\le \|\hat{\Delta}_t\|^2 I(\hat{\tau}_M > t-1, \tau_M > t-1) \\ &\le (1 - \gamma_t\lambda)^2\|\hat{\Delta}_{t-1}\|^2 I(\hat{\tau}_M > t-1, \tau_M > t-1) + \gamma_t^2\|\zeta_t\|^2 I(\hat{\tau}_M > t-1, \tau_M > t-1) \\ &\quad - 2\gamma_t\zeta_t^\top P(I - \gamma_t G)\hat{\Delta}_{t-1} I(\hat{\tau}_M > t-1, \tau_M > t-1). \end{aligned}$$

Taking conditional expectation and noticing that $\{\hat{\tau}_M > t-1, \tau_M > t-1\} \in \mathcal{F}_{t-1}$, Lemma A.6(iii) further leads to

$$\begin{aligned}\mathrm{E}[\|\hat{\Delta}_t\|^2 I(\hat{\tau}_M > t, \tau_M > t)|\mathcal{F}_{t-1}] &\le \Big((1 - \gamma_t\lambda)^2\|\hat{\Delta}_{t-1}\|^2 + \gamma_t^2\mathrm{E}(\|\zeta_t\|^2|\mathcal{F}_{t-1})\Big) I(\hat{\tau}_M > t-1, \tau_M > t-1) \\ &\le \Big((1 - \gamma_t\lambda)^2\|\hat{\Delta}_{t-1}\|^2 + \gamma_t^2 K(1 + \|\theta_{t-1}\|^2)\Big) I(\hat{\tau}_M > t-1, \tau_M > t-1) \\ &\le \Big((1 - \gamma_t\lambda)^2\|\hat{\Delta}_{t-1}\|^2 + \gamma_t^2 K(1 + 2\|\Delta_{t-1}\|^2 + 2\|\theta^*\|^2)\Big) I(\hat{\tau}_M > t-1, \tau_M > t-1) \\ &\le (1 - 2\lambda\gamma_t + \lambda^2\gamma_t^2)\|\hat{\Delta}_{t-1}\|^2 I(\hat{\tau}_M > t-1, \tau_M > t-1) + 2\gamma_t^2 K(1 + M^2 + \|\theta^*\|^2), \end{aligned}$$

where we use the fact that $\|\Delta_{t-1}\| \le M$ when $\tau_M > t-1$. Taking expectation again, we have

$$\begin{aligned}\mathrm{E}[\|\hat{\Delta}_t\|^2 I(\hat{\tau}_M > t, \tau_M > t)] &\le (1 - 2\lambda\gamma_t + \lambda^2\gamma_t^2)\mathrm{E}[\|\hat{\Delta}_{t-1}\|^2 I(\hat{\tau}_M > t-1, \tau_M > t-1)] + 2\gamma_t^2 K(1 + M^2 + \|\theta^*\|^2) \\ &\le (1 - \lambda\gamma_t)\mathrm{E}[\|\hat{\Delta}_{t-1}\|^2 I(\hat{\tau}_M > t-1, \tau_M > t-1)] + 2\gamma_t^2 K(1 + M^2 + \|\theta^*\|^2), \end{aligned}$$

where we use the fact that $\gamma_t \le 1/\lambda$ for large t. The above inequality further implies that

$$\frac{\mathrm{E}[\|\hat{\Delta}_t\|^2 I(\hat{\tau}_M > t, \tau_M > t)]}{\gamma_t} \le \frac{\gamma_{t-1}(1 - \lambda\gamma_t)}{\gamma_t}\frac{\mathrm{E}[\|\Delta_{t-1}\|^2 I(\hat{\tau}_M > t-1, \tau_M > t-1)]}{\gamma_{t-1}} + 2\gamma_t K(1 + M^2 + \|\theta^*\|^2).$$

Now applying Lemma A.4, we conclude that $\sup_{1\le t<\infty}\mathrm{E}[\|\hat{\Delta}_t\|^2 I(\hat{\tau}_M > t, \tau_M > t)]/\gamma_t < \infty$, which further implies (16).
Step 3: This step is to show

$$\frac{1}{\sqrt{T}}\sum_{t=2}^{T}\frac{\hat{\theta}_{t-1} - \hat{\theta}_t}{\gamma_t} = o_p(1). \tag{17}$$

Since both $\hat{\theta}_t$ and θ_t are strongly consistent by Step 1 and Lemma A.8, for any $\epsilon > 0$, there exists a constant $M > 0$ such that

$$\Pr(\sup_{1\le t<\infty} \|\hat{\Delta}_t\| \le M) \ge 1-\epsilon, \quad \Pr(\sup_{1\le t<\infty} \|\Delta_t\| \le M) \ge 1-\epsilon. \tag{18}$$

By direction examination, it follows that

$$\begin{aligned}\frac{1}{\sqrt{T}}\sum_{t=2}^{T}\frac{\hat{\theta}_{t-1}-\hat{\theta}_t}{\gamma_t} &= \frac{1}{\sqrt{T}}\sum_{t=2}^{T}\frac{\hat{\theta}_{t-1}-\theta^*+\theta^*-\hat{\theta}_t}{\gamma_t} = \frac{1}{\sqrt{T}}\sum_{t=2}^{T}\frac{\hat{\theta}_{t-1}-\theta^*}{\gamma_t} - \frac{1}{\sqrt{T}}\sum_{t=2}^{T}\frac{\hat{\theta}_t-\theta^*}{\gamma_t}\\ &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T-1}\frac{\hat{\theta}_t-\theta^*}{\gamma_{t+1}} - \frac{1}{\sqrt{T}}\sum_{t=2}^{T}\frac{\hat{\theta}_t-\theta^*}{\gamma_t} := D_1 - D_2 + D_3,\end{aligned}$$

where

$$D_1 = \frac{1}{\sqrt{T}}\frac{\hat{\theta}_1-\theta^*}{\gamma_2}, \quad D_2 = \frac{1}{\sqrt{T}}\frac{\hat{\theta}_T-\theta^*}{\gamma_T}, \quad D_3 = \frac{1}{\sqrt{T}}\sum_{t=2}^{T-1}(\hat{\theta}_t-\theta^*)(\gamma_{t+1}^{-1}-\gamma_t^{-1}).$$

It suffices to bound the above three terms. Clearly $D_1 = o_p(1)$. For D_2, we have the following bound

$$\begin{aligned}\|D_2\| &= \frac{1}{\sqrt{T}\gamma_T}\|\hat{\theta}_T-\theta^*\| = \frac{1}{\sqrt{T}\gamma_T}\|\hat{\Delta}_T\|\\ &\le \frac{1}{\sqrt{T}\gamma_T}\|\hat{\Delta}_T\|I(\hat{\tau}_M > T, \tau_M > T) + \frac{1}{\sqrt{T}\gamma_T}\|\hat{\Delta}_T\|I(\hat{\tau}_M \le T) + \frac{1}{\sqrt{T}\gamma_T}\|\hat{\Delta}_T\|I(\tau_M \le T) := D_{21}+D_{22}+D_{23}.\end{aligned}$$

By (16) and Assumption A1(i), we have

$$\mathrm{E}(D_{21}) \le \frac{1}{\sqrt{T}\gamma_T}\sqrt{\mathrm{E}[\|\Delta_T\|^2 I(\hat{\tau_M} > T, \tau_M > T)]} \le \sqrt{\frac{K_M}{T\gamma_T}} = \sqrt{\frac{K_M}{\gamma T^{1-\rho}}} \to 0.$$

The definitions of $\hat{\tau}_M$ and τ_M indicate that $\{\sup_{1\le t<\infty}\|\hat{\Delta}_t\| \le M\} \subset \{\hat{\tau}_M > T\}$ and $\{\sup_{1\le t<\infty}\|\Delta_t\| \le M\} \subset \{\tau_M > T\}$. By (18), we see that

$$\Pr(\hat{\tau}_M > T) \ge \Pr(\sup_{1\le t<\infty}\|\hat{\Delta}_t\| \le M) \ge 1-\epsilon, \quad \Pr(\tau_M > T) \ge \Pr(\sup_{1\le t<\infty}\|\Delta_t\| \le M) \ge 1-\epsilon. \tag{19}$$

Since for any $\delta > 0$, it follows that

$$D_{22} = \begin{cases}\frac{1}{\sqrt{T}\gamma_T}\|\hat{\Delta}_T\| & \text{if } \hat{\tau}_M \le T;\\ 0 & \text{if } \hat{\tau}_M > T;\end{cases} \quad D_{23} = \begin{cases}\frac{1}{\sqrt{T}\gamma_T}\|\hat{\Delta}_T\| & \text{if } \tau_M \le T;\\ 0 & \text{if } \tau_M > T,\end{cases}$$

we see that

$$\{D_{22} > \delta/3\} \subset \{\hat{\tau}_M \le T\}, \quad \{D_{23} > \delta/3\} \subset \{\tau_M \le T\} \tag{20}$$

Combining the above inequalities, for any $\delta > 0$, we deduce that

$$\begin{aligned}\Pr(\|D_2\| > \delta) &\le \Pr(D_{21} > \delta/3) + \Pr(D_{22} > \delta/3) + \Pr(D_{23} > \delta/3)\\ &\le \frac{3}{\delta}\sqrt{\frac{K_M}{\gamma T^{1-\rho}}} + \Pr(\hat{\tau}_M \le T) + \Pr(\tau_M \le T) \le \frac{3}{\delta}\sqrt{\frac{K_M}{\gamma T^{1-\rho}}} + 2\epsilon,\end{aligned}$$

which further implies that $\lim_{T\to\infty}\Pr(\|D_2\| > \delta) \le 2\epsilon$. Since $\epsilon > 0$ can be arbitrarily chosen, we show that $D_2 = o_p(1)$. To handle D_3, we use the following decomposition

$$\begin{aligned}\|D_3\| &\le \frac{1}{\sqrt{T}}\sum_{t=2}^{T-1}\|\hat{\theta}_t-\theta^*\|\,|\gamma_{t+1}^{-1}-\gamma_t^{-1}|I(\hat{\tau}_M > t, \tau_M > t) + \frac{1}{\sqrt{T}}\sum_{t=2}^{T-1}\|\hat{\theta}_t-\theta^*\|\,|\gamma_{t+1}^{-1}-\gamma_t^{-1}|I(\hat{\tau}_M \le t)\\ &\quad + \frac{1}{\sqrt{T}}\sum_{t=2}^{T-1}\|\hat{\theta}_t-\theta^*\|\,|\gamma_{t+1}^{-1}-\gamma_t^{-1}|I(\tau_M \le t)\\ &= \frac{1}{\sqrt{T}}\sum_{t=2}^{T-1}\|\hat{\Delta}_t\|\,|\gamma_{t+1}^{-1}-\gamma_t^{-1}|I(\hat{\tau}_M > t, \tau_M > t) + \frac{1}{\sqrt{T}}\sum_{t=2}^{T-1}\|\hat{\Delta}_t\|\,|\gamma_{t+1}^{-1}-\gamma_t^{-1}|I(\hat{\tau}_M \le T)\\ &\quad + \frac{1}{\sqrt{T}}\sum_{t=2}^{T-1}\|\hat{\Delta}_t\|\,|\gamma_{t+1}^{-1}-\gamma_t^{-1}|I(\tau_M \le T) := D_{31}+D_{32}+D_{33}.\end{aligned}$$

We obtain from (16) that

$$\begin{aligned}
&\mathrm{E}\left(\sum_{t=2}^{\infty}\frac{1}{\sqrt{t}}\|\hat{\Delta}_t\|\,|\gamma_{t+1}^{-1}-\gamma_t^{-1}|I(\hat{\tau}_M>t,\tau_M>t)\right)\le\sum_{t=2}^{\infty}\frac{|\gamma_{t+1}^{-1}-\gamma_t^{-1}|}{\sqrt{t}}\mathrm{E}[\|\hat{\Delta}_t\|I(\hat{\tau}_M>t,\tau_M>t)]\\
&\le\sum_{t=2}^{\infty}\frac{|\gamma_{t+1}^{-1}-\gamma_t^{-1}|}{\sqrt{t}}\sqrt{\mathrm{E}[\|\hat{\Delta}_t\|^2I(\hat{\tau}_M>t,\tau_M>t)]}\le\sum_{t=2}^{\infty}\frac{|\gamma_{t+1}^{-1}-\gamma_t^{-1}|}{\sqrt{t}}\sqrt{K_M\gamma_t}\\
&=\sqrt{K_M}\sum_{t=2}^{\infty}\frac{1}{\sqrt{t}}\frac{1}{\gamma}|(t+1)^{\rho}-t^{\rho}|\sqrt{\gamma t^{-\rho}}=\sqrt{\gamma K_M}\sum_{t=2}^{\infty}\frac{1}{\sqrt{t}}t^{\rho}\left[\left(\frac{t+1}{t}\right)^{\rho}-1\right]\sqrt{t^{-\rho}}\\
&\le\sqrt{\gamma K_M}\sum_{t=2}^{\infty}\frac{1}{\sqrt{t}}t^{\rho}\left[\left(\frac{t+1}{t}\right)-1\right]\sqrt{t^{-\rho}}=\sqrt{\gamma K_M}\sum_{t=2}^{\infty}\frac{1}{t^{3/2-\rho/2}}<\infty,
\end{aligned}$$

where we use Assumption A1(i) that $\gamma_t=\gamma t^{\rho}$ for some $\rho\in(1/2,1)$. The above inequality also implies that

$$\sum_{t=2}^{\infty}\frac{1}{\sqrt{t}}\|\hat{\Delta}_t\|\,|\gamma_{t+1}^{-1}-\gamma_t^{-1}|I(\hat{\tau}_M>t,\tau_M>t)<\infty\ \text{ almost surely.}$$

As a consequence of Kronecker's lemma, we show that $D_{31}=o_p(1)$. Using (19) and similar arguments as (20), for any $\delta>0$, we have

$$\begin{aligned}
\Pr(\|D_3\|>\delta)&\le\Pr(D_{31}>\delta/3)+\Pr(D_{32}>\delta/3)+\Pr(D_{33}>\delta/3)\\
&\le\Pr(D_{31}>\delta/3)+\Pr(\hat{\tau}_M\le T)+\Pr(\tau_M\le T)\le\Pr(D_{31}>\delta/3)+2\epsilon.
\end{aligned}$$

Taking limit, it holds that $\lim_{T\to\infty}\Pr(\|D_3\|>\delta)\le 2\epsilon$. Since $\epsilon>0$ can be arbitrarily chosen, we show that $D_3=o_p(1)$. Combining the rates of D_1,D_2,D_3, we verify (17)

Step 4: Using (14), we have

$$\hat{\Delta}_t=P(I-\gamma_tG)\hat{\Delta}_{t-1}-\gamma_tP\zeta_t=P\hat{\Delta}_{t-1}-\gamma_tPG\hat{\Delta}_{t-1}-\gamma_tP\zeta_t.$$

Since $P\hat{\Delta}_t=\hat{\Delta}_t$ for $t\ge1$, it further leads to

$$\gamma_tPGP\hat{\Delta}_{t-1}=\gamma_tPG\hat{\Delta}_{t-1}=-P(\hat{\Delta}_t-\hat{\Delta}_{t-1})-\gamma_tP\zeta_t=-P(\hat{\theta}_t-\hat{\theta}_{t-1})-\gamma_tP\zeta_t\quad\text{for all } t\ge2.$$

Taking summation, we show that

$$\sum_{t=2}^{T+1}PGP\hat{\Delta}_{t-1}=PGP\hat{\Delta}_T+\sum_{t=2}^{T}PGP\hat{\Delta}_{t-1}=PGP\hat{\Delta}_T-P\sum_{t=2}^{T}\frac{\theta_t-\theta_{t-1}}{\gamma_t}-P\sum_{t=2}^{T}\zeta_t,$$

which further implies that

$$\frac{1}{\sqrt{T}}PGP\sum_{t=1}^{T}\hat{\Delta}_t=\frac{1}{\sqrt{T}}PGP\hat{\Delta}_T-\frac{1}{\sqrt{T}}P\sum_{t=2}^{T}\frac{\theta_t-\theta_{t-1}}{\gamma_t}-\frac{1}{\sqrt{T}}P\sum_{t=2}^{T}\zeta_t:=E_1-E_2-E_3.$$

Using the strong consistency of $\hat{\theta}_t$ in Step 1 and (17) in Step 3, we show that $E_1=o_p(1)$ and $E_2=o_p(1)$. Moreover, by iterative substitution and the fact that $\hat{\theta}_0=\theta_0$, (14) leads to

$$\begin{aligned}
\hat{\Delta}_t&=P(I-\gamma_tG)\hat{\Delta}_{t-1}-\gamma_tP\zeta_t=\left[\prod_{j=1}^{t}P(I-\gamma_jG)\right]\hat{\Delta}_0+\sum_{j=1}^{t}\left[\prod_{i=j+1}^{t}P(I-\gamma_iG)\right]\gamma_jP\zeta_j\\
&=\left[\prod_{j=1}^{t}P(I-\gamma_jG)\right]\Delta_0+\sum_{j=1}^{t}\left[\prod_{i=j+1}^{t}P(I-\gamma_iG)P\right]\gamma_jP\zeta_j,
\end{aligned}$$

which, by averaging, further implies that

$$\frac{1}{T}\sum_{t=1}^{T}\hat{\Delta}_t=\frac{1}{T}\sum_{t=1}^{T}\left[\prod_{j=1}^{t}P(I-\gamma_jA)\right]\Delta_0+\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{t}\left[\prod_{i=j+1}^{t}P(I-\gamma_iG)P\right]\gamma_jP\zeta_j=S_1+S_2.$$

Notice that $(PGP)^-P=(PGP)^-$ by Lemma A.2, we complete the proof. □

Lemma A.11. *Under Assumption A1, it follows that $S_3=o_p(T^{-1/2})$, where S_3 is defined in (11).*

Proof. Changing the order of summation leads to

$$S_3 = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{t}\Big[\prod_{i=j+1}^{t} P(I-\gamma_i G)P\Big]\gamma_j P(R(\theta_{j-1}) - G\Delta_{j-1}) = \frac{1}{T}\sum_{j=1}^{T}\sum_{t=j}^{T}\Big[\prod_{i=j+1}^{t} P(I-\gamma_i G)P\Big]\gamma_j P(R(\theta_{j-1}) - G\Delta_{j-1}).$$

By Lemma A.8, for any $\epsilon > 0$, there exists a constant $M > 0$ such that

$$\Pr(\tau_M > T) \geq \Pr(\sup_{1\leq t<\infty} \|\Delta_t\| \leq M) \geq 1-\epsilon, \tag{21}$$

where $\tau_M = \inf\{j \geq 1 : \|\Delta_j\| > M\}$ is the stopping time defined in Lemma A.9. Let us define constant $\alpha_j^T = \gamma_j \sum_{t=j}^{T}\big(\prod_{i=j+1}^{t} P(I-\gamma_i G)P\big)$. Lemmas A.3 and A.7 lead to

$$\begin{aligned}
\|\sqrt{T}S_3\| &\leq \Big\|\frac{1}{\sqrt{T}}\sum_{j=1}^{T}\alpha_j^T P(R(\theta_{j-1}) - G\Delta_{j-1})\Big\| \leq \frac{1}{\sqrt{T}}\sum_{j=1}^{T}\|\alpha_j^T\|\|P(R(\theta_{j-1}) - G\Delta_{j-1})\| \leq \frac{K}{\sqrt{T}}\sum_{j=1}^{T}\|R(\theta_{j-1}) - G\Delta_{j-1}\| \\
&\leq \frac{K^2}{\sqrt{T}}\sum_{j=1}^{T}\|\Delta_{j-1}\|^2 \leq \frac{K^2}{\sqrt{T}}\sum_{j=1}^{T}\|\Delta_{j-1}\|^2 I(\tau_M \leq j-1) + \frac{K^2}{\sqrt{T}}\sum_{j=1}^{T}\|\Delta_{j-1}\|^2 I(\tau_M > j-1) \\
&\leq \frac{K^2}{\sqrt{T}}\sum_{j=1}^{T}\|\Delta_{j-1}\|^2 I(\sup_{1\leq t<\infty}\|\Delta_t\| > M) + \frac{K^2}{\sqrt{T}}\sum_{j=1}^{T}\|\Delta_{j-1}\|^2 I(\tau_M > j-1) := S_{31} + S_{32}.
\end{aligned}$$

For the first term, using (21) and similar arguments as (20), for any $\delta > 0$, we have

$$\Pr(S_{31} > \delta/2) \leq \Pr(\sup_{1\leq t<\infty}\|\Delta_t\| > M) \leq \epsilon.$$

For the second term, Lemma A.9 implies that

$$\mathrm{E}\Big(\sum_{j=1}^{\infty}\frac{\|\Delta_j\|^2 I(\tau_M > j)}{j^{1/2}}\Big) \leq K_M\sum_{j=1}^{\infty}\frac{\gamma_j}{j^{1/2}} = K_M\sum_{j=1}^{\infty}\frac{\gamma}{j^{1/2+\rho}} < \infty,$$

where we use Assumption A1(i) that $\gamma_j = \gamma j^{-\rho}$ for some $\rho \in (1/2, 1)$. The above inequality also implies that $\Pr(\sum_{j=1}^{\infty} j^{-1/2}\|\Delta_j\|^2 I(\tau_M > j) < \infty) = 1$. Applying Kronecker's lemma, we show that $S_{32} \to 0$ almost surely as $T \to \infty$. Combining the bounds of S_{31} and S_{32}, we conclude that

$$\lim_{T\to\infty}\Pr(\|\sqrt{T}S_3\| > \delta) \leq \lim_{T\to\infty}\Pr(S_{31} > \delta/2) + \lim_{T\to\infty}\Pr(S_{32} > \delta/2) \leq \epsilon.$$

Since $\epsilon > 0$ can be arbitrarily chosen, we show that $\sqrt{T}S_3 = o_p(1)$. □

Lemma A.12. *Under Assumption A1, it follows that* $T^{-1/2}\sum_{t=1}^{T}\zeta_t \xrightarrow{\mathbb{L}} N(0, S)$.

Proof. We decompose the process ζ_t as follows:

$$\zeta_t = \nabla l(\theta_{t-1}, Z_t) - \nabla L(\theta_{t-1}) = \nabla l(\theta^*, Z_t) + [\nabla l(\theta_{t-1}, Z_t) - \nabla l(\theta^*, Z_t) - \nabla L(\theta_{t-1}) + \nabla L(\theta^*)] := \eta_t + \xi_t.$$

Assumption A1(vi) and Lemma A.8 imply that

$$\mathrm{E}(\|\xi_t\|^2|\mathcal{F}_{t-1}) \leq \mathrm{E}(\|\nabla l(\theta_{t-1}, Z_t) - \nabla l(\theta^*, Z_t)\|^2|\mathcal{F}_{t-1}) \leq \delta(\|\theta_{t-1} - \theta^*\|) \to 0 \ \text{ almost surely.}$$

Moreover, by Cauchy–Schwarz inequality, it follows that

$$\mathrm{E}(\|\eta_t\xi_t^\top\||\mathcal{F}_{t-1}) \leq \mathrm{E}(\|\eta_t\|\|\xi_t\||\mathcal{F}_{t-1}) \leq \sqrt{\mathrm{E}(\|\eta_t\|^2)}\sqrt{\mathrm{E}(\|\xi_t\|^2|\mathcal{F}_{t-1})} \to 0 \ \text{ almost surely.}$$

As a consequence of the above two inequalities, we show that

$$\mathrm{E}(\zeta_t\zeta_t^\top|\mathcal{F}_{t-1}) = \mathrm{E}(\eta_t\eta_t^\top) + 2\mathrm{E}(\eta_t\xi_t^\top|\mathcal{F}_{t-1}) + \mathrm{E}(\xi_t\xi_t^\top|\mathcal{F}_{t-1}) \to S \ \text{ almost surely,}$$

where $S = \mathrm{E}[\nabla l(\theta^*, Z)\nabla l^\top(\theta^*, Z)] \in \mathbb{R}^{p\times p}$ is a positive definite matrix defined in Assumption A1(v). For any $\epsilon > 0$, direct calculation leads to

$$\begin{aligned}
&\mathrm{E}(\|\zeta_t\|^2 I(\|\zeta_t\| > \epsilon\sqrt{T})|\mathcal{F}_{t-1}) \leq \mathrm{E}[2(\|\eta_t\|^2 + \|\xi_t\|^2)I(\|\eta_t\| + \|\xi_t\| > \epsilon\sqrt{T})|\mathcal{F}_{t-1}] \\
&\leq \mathrm{E}[2(\|\eta_t\|^2 + \|\xi_t\|^2)I(2\|\eta_t\| > \epsilon\sqrt{T}, \|\eta_t\| \geq \|\xi_t\|)|\mathcal{F}_{t-1}] + \mathrm{E}[2(\|\eta_t\|^2 + \|\xi_t\|^2)I(2\|\xi_t\| > \epsilon\sqrt{T}, \|\eta_t\| < \|\xi_t\|)|\mathcal{F}_{t-1}] \\
&\leq 4\mathrm{E}[\|\eta_t\|^2 I(\|\eta_t\| \geq \epsilon\sqrt{T}/2)|\mathcal{F}_{t-1}] + 4\mathrm{E}[\|\xi_t\|^2 I(\|\xi_t\| \geq \epsilon\sqrt{T}/2)|\mathcal{F}_{t-1}] \leq 4\mathrm{E}[\|\eta_t\|^2 I(\|\eta_t\| \geq \epsilon\sqrt{T}/2)] \\
&+ 4\delta(\|\theta_{t-1} - \theta^*\|).
\end{aligned}$$

Since $\theta_j = \nabla l(\theta^*, Z_j)$ are i.i.d., and $\lim_{t\to\infty} \delta(\|\theta_{t-1} - \theta^*\|) = 0$ almost surely, we conclude that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathrm{E}(\|\zeta_t\|^2 I(\|\zeta_t\| > \epsilon\sqrt{T})|\mathcal{F}_{t-1}) = 0 \quad \text{almost surely.}$$

By the C.L.T. for martingale-difference arrays (e.g., see [13]), we prove the asymptotic normality. □

Proof of Lemma 1. It suffices to show that $G^{-1} - (PGP)^-$ is positive semidefinite and has rank $p - d$. Since $\mathrm{rank}(P) = d$ by Assumption A1(viii), and P is diagonalizable, there exists an orthogonal matrix $U \in \mathbb{R}^{p\times p}$ such that

$$P = U \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix} U^\top, \quad U^\top G U = \begin{pmatrix} X & Y \\ Y^\top & Z \end{pmatrix},$$

for some matrices X, Y, Z with comfortable dimensions. As a consequence, it follows that

$$PGP = U \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix} U^\top G U \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix} U^\top = U \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} X & Y \\ Y^\top & Z \end{pmatrix} \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix} U^\top = U \begin{pmatrix} X & 0 \\ 0 & 0 \end{pmatrix} U^\top.$$

Let $S = Z - Y^\top X^{-1} Y \in \mathbb{R}^{(p-d)\times(p-d)}$ be the Schur complement of X. Since G is positive definite by Assumption A1(iv), so is S. The matrix block inversion formula implies that

$$\begin{aligned} G^{-1} &= U \begin{pmatrix} X & Y \\ Y^\top & Z \end{pmatrix}^{-1} U^\top = U \begin{pmatrix} X^{-1} + X^{-1}YS^{-1}Y^\top X^{-1} & -X^{-1}YS^{-1} \\ -S^{-1}Y^\top X^{-1} & S^{-1} \end{pmatrix} U^\top \\ &= U \begin{pmatrix} X^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^\top + U \begin{pmatrix} X^{-1}YS^{-1}Y^\top X^{-1} & -X^{-1}YS^{-1} \\ -S^{-1}Y^\top X^{-1} & S^{-1} \end{pmatrix} \\ U^\top &= (PGP)^- + U \begin{pmatrix} X^{-1}YS^{-1/2} \\ -S^{-1/2} \end{pmatrix} \left(S^{-1/2}Y^\top X^{-1}, -S^{-1/2}\right) U^\top, \end{aligned}$$

which proves the positive semidefiniteness. Because $\mathrm{rank}(S^{-1/2}) = p - d$ and $(S^{-1/2}Y^\top X^{-1}, -S^{-1/2}) \in \mathbb{R}^{(p-d)\times p}$, we verify that $G^{-1} - (PGP)^-$ has rank $p - d$. □

Lemma A.13. *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric with eigenvalues $\lambda_1 \geq \cdots \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \hat{\lambda}_p$. Fixing $1 \leq r \leq s \leq p$, let us define $d = s - r + 1$, and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{p\times d}$, $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $\Sigma v_j = \lambda_j v_j$ and $\hat{\Sigma}\hat{v}_j = \hat{\lambda}_j v_j$ for $j \in \{r, r+1, \ldots, s\}$. If $e := \inf\{|\hat{\lambda} - \lambda| : \lambda \in [\lambda_s, \lambda_r], \hat{\lambda} \in (-\infty, \hat{\lambda}_{s-1}] \cup [\hat{\lambda}_{r+1}, \infty)\} > 0$, where $\hat{\lambda}_0 := -\infty$ and $\hat{\lambda}_{p+1} := \infty$, then it follows that $\|VV^\top - \hat{V}\hat{V}^\top\| \leq 2\|\hat{\Sigma} - \Sigma\|/e$. Moreover, the eigenvalues satisfy $|\hat{\lambda}_i - \lambda_i| \leq \|\hat{\Sigma} - \Sigma\|$.*

Proof. It follows from Davis–Kahan Theorem (e.g., see [21]) and Weyl's inequality [20]. □

Lemma A.14. *Let $\Sigma, \hat{\Sigma}_n \in \mathbb{R}^{p\times p}$ be positive semidefinite matrices such that $\mathrm{rank}(\hat{\Sigma}_n) = \mathrm{rank}(\Sigma)$ and $\hat{\Sigma}_n \to \Sigma$ as $n \to \infty$. Then $\lim_{n\to\infty} \hat{\Sigma}_n^- = \Sigma^-$.*

Proof. Let distinct eigenvalues of Σ be $\rho_1 > \rho_2 > \cdots > \rho_d = 0$, and suppose that there are $k_j \geq 1$ eigenvalues $\lambda_{j,1} = \lambda_{j,2} = \cdots = \lambda_{j,k_j}$ equal to ρ_j, for $j \in \{1, \ldots, d\}$. We denote $v_{j,s}$ as the eigenvector corresponding to eigenvalue $\lambda_{j,s}$. Similarly, we define $(\hat{\lambda}_{j,s}, \hat{v}_{j,s})$ as the eigenpair of $\hat{\Sigma}_n$ for $j \in \{1, \ldots, d\}$ and $s \in \{1, \ldots, k_j\}$. However, in general, we do not have $\hat{\lambda}_{j,1} = \hat{\lambda}_{j,2} = \cdots = \hat{\lambda}_{j,k_j}$. Moreover, the eigenvalues can be chosen to be in an increasing order such that

$$\begin{aligned} \hat{\lambda}_{j,1} \geq \hat{\lambda}_{j,1} \geq \cdots \geq \hat{\lambda}_{j,k_j} \quad & \text{for all } j \in \{1, \ldots, d\}, \\ \hat{\lambda}_{1,s_1} \geq \hat{\lambda}_{2,s_2} \geq \cdots \geq \hat{\lambda}_{d,s_d} \quad & \text{for all } s_j \in \{1, \ldots, k_j\} \text{ and } j \in \{1, \ldots, d\}. \end{aligned}$$

By Lemma A.13, we see that $\hat{\lambda}_{j,s} \to \lambda_{j,s} = \rho_j$ for all $j \in \{1, \ldots, d\}$. As a consequence, when n is sufficiently large, there exists a constant $\epsilon > 0$ such that

$$\rho_{j+1} < \rho_j - \epsilon \leq \hat{\lambda}_{j,s} \leq \rho_j + \epsilon < \rho_{j-1} \quad \text{for all } s \in \{1, \ldots, k_j\} \text{ and } j \in \{1, \ldots, d-1\}.$$

Since $\mathrm{rank}(\Sigma_n) = \mathrm{rank}(\Sigma)$, it holds that $\hat{\lambda}_{d,s} = \lambda_{d,s} = \rho_d = 0$. For each $j \in \{1, \ldots, d-1\}$, applying Lemma A.13 to eigenpairs $(\lambda_{j,s}, v_{j,s})$ and $(\hat{\lambda}_{j,s}, \hat{v}_{j,s})$ with $s \in \{1, \ldots, k_j\}$, we have $e \geq \epsilon$ and

$$\left\| \sum_{s=1}^{k_j} \hat{v}_{j,s}\hat{v}_{j,s}^\top - \sum_{s=1}^{k_j} v_{j,s}v_{j,s}^\top \right\| \leq 2\|\hat{\Sigma}_n - \Sigma\|/\epsilon = o_p(1),$$

which further implies that

$$\begin{aligned}
&\left\| \sum_{s=1}^{k_j} \hat{\lambda}_{j,s}^{-1} \hat{v}_{j,s} \hat{v}_{j,s}^\top - \sum_{s=1}^{k_j} \lambda_{j,s}^{-1} v_{j,s} v_{j,s}^\top \right\| = \left\| \sum_{s=1}^{k_j} \hat{\lambda}_{j,s}^{-1} \hat{v}_{j,s} \hat{v}_{j,s}^\top - \sum_{s=1}^{k_j} \rho_j^{-1} v_{j,s} v_{j,s}^\top \right\| \\
&\le \left\| \sum_{s=1}^{k_j} \hat{\lambda}_{j,s}^{-1} \hat{v}_{j,s} \hat{v}_{j,s}^\top - \sum_{s=1}^{k_j} \rho_j^{-1} \hat{v}_{j,s} \hat{v}_{j,s}^\top \right\| + \left\| \sum_{s=1}^{k_j} \rho_j^{-1} \hat{v}_{j,s} \hat{v}_{j,s}^\top - \sum_{s=1}^{k_j} \rho_j^{-1} v_{j,s} v_{j,s}^\top \right\| \\
&\le \sum_{s=1}^{k_j} |\hat{\lambda}_{j,s}^{-1} - \rho_j^{-1}| \|\hat{v}_{j,s} \hat{v}_{j,s}^\top\| + \rho_j^{-1} \left\| \sum_{s=1}^{k_j} \hat{v}_{j,s} \hat{v}_{j,s}^\top - \sum_{s=1}^{k_j} v_{j,s} v_{j,s}^\top \right\| = \sum_{s=1}^{k_j} |\hat{\lambda}_{j,s}^{-1} - \rho_j^{-1}| + o_p(1),
\end{aligned}$$

where we used the fact that $\rho_j > 0$ for $j \in \{1, \ldots, d-1\}$. Finally, notice that

$$\sum_{s=1}^{k_j} |\hat{\lambda}_{j,s}^{-1} - \rho_j^{-1}| = o_p(1), \quad \Sigma^- = \sum_{j=1}^{d-1} \sum_{s=1}^{k_j} \lambda_{j,s}^{-1} v_{j,s} v_{j,s}^\top, \quad \hat{\Sigma}^- = \sum_{j=1}^{d-1} \sum_{s=1}^{k_j} \hat{\lambda}_{j,s}^{-1} \hat{v}_{j,s} \hat{v}_{j,s}^\top,$$

we complete the proof. □

Lemma A.15. *Suppose a sequence of matrices $\{A_n\}_{n=1}^\infty \in \mathbb{R}^{p\times p}$ satisfies $\lim_{n\to\infty} A_n = A$ where $A \in \mathbb{R}^{p\times p}$ is positive definite. Let $P \in \mathbb{R}^{p\times p}$ be a projection matrix such that $P^2 = P$ and $P^\top = P$. Then $\lim_{n\to\infty}(PA_nP)^- = (PAP)^-$.*

Proof. Since A is positive definite, so is A_n when n is sufficiently large. Hence PA_nP and PAP both have the same rank as P. The desired result follows from Lemma A.14. □

Lemma A.16. *Under Assumptions A1 and A2, it follows that $\hat{G}_T = G + o_p(1)$, $(P\hat{G}_TP)^- = (PGP)^- + o_p(1)$, and $\hat{S}_T = S + o_p(1)$.*

Proof. Since $\overline{\theta}_T \to \theta^*$ almost surely as $T \to \infty$ by Lemma A.8, it follows from the continuity of $\theta \to \nabla^2 l(\theta, Z)$ at θ^* in Assumption A2 that $\lim_{T\to\infty} \|\nabla^2 l(\overline{\theta}_T, Z_T) - \nabla^2 l(\theta^*, Z_T)\| = 0$ almost surely. As a consequence, when $T \to \infty$, we have

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \left(\nabla^2 l(\overline{\theta}_t, Z_t) - \nabla^2 l(\theta^*, Z_t) \right) \right\| \le \frac{1}{T} \sum_{t=1}^{T} \left\| \nabla^2 l(\overline{\theta}_t, Z_t) - \nabla^2 l(\theta^*, Z_t) \right\| \to 0 \text{ almost surely.}$$

By Assumption A2, Lebesgue's Dominated Convergence Theorem, and L.L.N., we can see

$$\frac{1}{T} \sum_{t=1}^{T} \nabla^2 l(\theta^*, Z_t) = \mathrm{E}[\nabla^2 l(\theta^*, Z_t)] + o_p(1) = \nabla^2 L(\theta^*) + o_p(1).$$

Combining the above, we show that $\hat{G}_T = G + o_p(1)$.

Similarly, we have $\lim_{T\to\infty} \|\nabla l(\overline{\theta}_T, Z_T) \nabla l^\top(\overline{\theta}_T, Z_T) - \nabla l(\theta^*, Z_T) \nabla l^\top(\theta^*, Z_T)\| = 0$ almost surely by the differentiability of $\theta \to \nabla l(\theta, Z)$ in Assumption A2. Moreover, by L.L.N., we can derive $\hat{S}_T = S + o_p(1)$. Finally, applying Lemma A.15, we complete the proof. □

Proof of Lemma 2. Lemma 2 is a direct consequence of Lemma A.16. □

Proof of Theorem 2. Under H_0, by Theorem 1, it follows that

$$\overline{\theta}_{T,P} - \theta^* = -\frac{1}{T} \sum_{t=1}^{T} (PGP)^- \zeta_t + o_p(T^{-1/2}), \quad \overline{\theta}_{T,I} - \theta^* = -\frac{1}{T} \sum_{t=1}^{T} G^{-1} \zeta_t + o_p(T^{-1/2}).$$

Since $P(PGP)^- = (PGP)^-$ by Lemma A.2, we have

$$(I-P)(\overline{\theta}_{T,P} - \overline{\theta}_{T,I}) = \frac{1}{T} \sum_{t=1}^{T} (I-P)[G^{-1} - (PGP)^-]\zeta_t + o_p(T^{-1/2}) = \frac{1}{T} \sum_{t=1}^{T} (I-P) G^{-1} \zeta_t + o_p(T^{-1/2}).$$

By Lemma A.12, we show that $\sqrt{T}(\overline{\theta}_{T,P} - \overline{\theta}_{T,I}) \xrightarrow{\mathbb{L}} N(0, W)$, where $W = (I-P)G^{-1}SG^{-1}(I-P)$. By delta method, we have $\sqrt{T}[W^-]^{1/2}(\overline{\theta}_{T,P} - \overline{\theta}_{T,I}) \xrightarrow{\mathbb{L}} N(0, V)$, where $V = \mathrm{Diag}(1, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^{p\times p}$ is a squared matrix with rank $p-d$. The above convergence further leads to $T(\overline{\theta}_{T,P} - \overline{\theta}_{T,I})^\top W^- (\overline{\theta}_{T,P} - \overline{\theta}_{T,I}) \xrightarrow{\mathbb{L}} \chi^2(p-d)$. By Lemma A.16, it follows that $\hat{G}_{T,I} = G + o_p(1)$ and $\hat{S}_{T,I} = S + o_p(1)$. Moreover, both W and $\hat{W}$ are of rank $p-d$. As a consequence of Lemma A.14, it follows $\hat{W} = W + o_p(1)$. Applying Slutsky's Theorem, we compete the proof of the result under H_0.

Under H_1, since $B\theta^* = b + \beta$, for some $\beta \neq 0$. Consider the following decomposition $\theta^* = \tilde{\theta}^* + \mu$ with $B\tilde{\theta}^* = b$ and $B\mu = \beta$. Clearly, $(I-P)\mu \neq 0$, as $(I-P)\mu = 0$ implies $P\mu = \mu$ and $\mu \in \mathrm{Ker}(B)$, which is impossible. Since $B\overline{\theta}_{T,P} = B\tilde{\theta}^* = b$, we have

$$\begin{aligned}(I-P)(\overline{\theta}_{T,P} - \overline{\theta}_{T,I}) &= (I-P)(\overline{\theta}_{T,P} - \theta^* + \theta^* - \overline{\theta}_{T,I}) = (I-P)(\overline{\theta}_{T,P} - \tilde{\theta}^* - \mu + \theta^* - \overline{\theta}_{T,I}) \\ &= -(I-P)\mu - (I-P)(\overline{\theta}_{T,I} - \theta^*). \end{aligned} \tag{22}$$

Moreover, by Lemma A.2, we have $\hat{W}^-(I-P) = (I-P)\hat{W}^- = W^-$. Following (22), we have

$$T(\overline{\theta}_{T,P} - \overline{\theta}_{T,I})^\top \hat{W}^-(\overline{\theta}_{T,P} - \overline{\theta}_{T,I}) = T\mu^\top \hat{W}^-\mu + T(\overline{\theta}_{T,I} - \theta^*)^\top \hat{W}^-(\overline{\theta}_{T,I} - \theta^*) + 2T(\overline{\theta}_{T,I} - \theta^*)^\top \hat{W}^-\mu := J_1 + J_2 + J_3.$$

For S_1, let $\hat{\lambda}_1, \hat{\lambda}_{p-d}$ and λ_1, λ_{p-d} be the largest and smallest non-zero eigenvalues of $\hat{W}$ and W respectively. By Lemma A.14, we know $\hat{\lambda}_1 \leq 2\lambda_1$ and $\hat{\lambda}_{p-d} \geq \lambda_{p-d}/2$ with probability approaching 1. Then by Lemma A.2, we conclude that

$$J_1 \geq \frac{T}{\hat{\lambda}}\|(I-P)\mu\|^2 \geq \frac{T}{2\lambda}\|(I-P)\mu\|^2 \quad \text{with probability approaching 1.}$$

Since Theorem 1 implies that $\overline{\theta}_{T,I} - \theta^* = O_p(T^{-1/2})$, it follows that

$$J_2 \leq T\|W^-\|\|\overline{\theta}_{T,I} - \theta^*\|^2 \leq \frac{T}{\hat{\lambda}_{p-d}}\|\overline{\theta}_{T,I} - \theta^*\|^2 \leq \frac{2T}{\lambda_{p-d}}\|\overline{\theta}_{T,I} - \theta^*\|^2 = O_p(1).$$

Similarly, by Cauchy–Schwarz inequality, we can show

$$|J_3| \leq 2T\|\hat{W}^-\|\|\overline{\theta}_{T,I} - \theta^*\|\|\mu\| \leq \frac{2T}{\hat{\lambda}_{p-d}}\|\overline{\theta}_{T,I} - \theta^*\|\|\mu\| \leq \frac{4T}{\lambda_{p-d}}\|\overline{\theta}_{T,I} - \theta^*\|\|\mu\| = O_p(T^{1/2}).$$

Combining the three bounds, we prove that $T(\overline{\theta}_{T,P} - \overline{\theta}_{T,I})^\top \hat{W}^-(\overline{\theta}_{T,P} - \overline{\theta}_{T,I}) \to \infty$ with probability approaching 1.

Suppose the local alternative $H_a : B\theta^* = b + \beta/\sqrt{T}$ holds. Consider the following decomposition $\theta^* = \tilde{\theta}^* + \mu/\sqrt{T}$ with $B\tilde{\theta}^* = b$ and $B\mu = \beta$. By Lemma A.2, we have $(\hat{W}^-)^{1/2}(I-P) = (I-P)(\hat{W}^-)^{1/2} = (\hat{W}^-)^{1/2}$. By similar proof to (22), we have

$$(I-P)(\overline{\theta}_{T,P} - \overline{\theta}_{T,I}) = -(I-P)\mu/\sqrt{T} - (I-P)(\overline{\theta}_{T,I} - \theta^*),$$

which further leads to

$$(\hat{W}^-)^{1/2}(\overline{\theta}_{T,P} - \overline{\theta}_{T,I}) = (\hat{W}^-)^{1/2}(I-P)(\overline{\theta}_{T,P} - \overline{\theta}_{T,I}) = -(\hat{W}^-)^{1/2}(I-P)\mu/\sqrt{T} - (\hat{W}^-)^{1/2}(\overline{\theta}_{T,I} - \theta^*) := R_1 - R_2.$$

Since $\hat{W}^- = W + o_p(1)$, it follows that $\sqrt{T}R_1 = -(W^-)^{1/2}(I-P)\mu + o_p(1) = -(W^-)^{1/2}\mu + o_p(1)$. Moreover, Theorem 1 implies that

$$\sqrt{T}R_2 = (W^-)^{1/2}(\overline{\theta}_{T,I} - \theta^*) + o_p(1) \xrightarrow{\mathbb{L}} N(0, (W^-)^{1/2}G^{-1}SG^{-1}(W^-)^{1/2}).$$

By direct calculation, it can be verified that

$$(W^-)^{1/2}G^{-1}SG^{-1}(W^-)^{1/2} = (W^-)^{1/2}(I-P)G^{-1}SG^{-1}(I-P)(W^-)^{1/2} = (W^-)^{1/2}W(W^-)^{1/2} = V,$$

where $V = \mathrm{Diag}(1, \ldots, 1, 0, \ldots, 0) \in \mathbb{R}^{p\times p}$ is a squared matrix with rank $p-d$. As a consequence, we show that $T(\overline{\theta}_{T,P} - \overline{\theta}_{T,I})^\top \hat{W}^-(\overline{\theta}_{T,P} - \overline{\theta}_{T,I}) \xrightarrow{\mathbb{L}} \chi^2(\mu^\top W^-\mu, p-d)$. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2022.105017.

References

[1] L. Bottou, Stochastic gradient learning in neural networks, in: Proceedings of Neuro-Nîmes 91, EC2, Nimes, France, 1991.
[2] L. Bottou, F.E. Curtis, J. Nocedal, Optimization methods for large-scale machine learning, SIAM Rev. 60 (2) (2018) 223–311.
[3] X. Chen, J.D. Lee, X.T. Tong, Y. Zhang, Statistical inference for model parameters in stochastic gradient descent, Ann. Statist. 48 (1) (2020) 251–273.
[4] Y. Fang, J. Xu, L. Yang, Online bootstrap confidence intervals for the stochastic gradient descent estimator, J. Mach. Learn. Res. 19 (1) (2018) 3053–3073.
[5] R. Gemulla, E. Nijkamp, P.J. Haas, Y. Sismanis, Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent, in: KDD '11, Association for Computing Machinery, New York, 2011, pp. 69–77.
[6] A. Godichon-Baggioni, B. Portier, An averaged projected Robbins-Monro algorithm for estimating the parameters of a truncated spherical distribution, Electron. J. Stat. 11 (1) (2017) 1890–1927.
[7] J.D. Gorman, A.O. Hero, Lower bounds for parametric estimation with constraints, IEEE Trans. Inform. Theory 36 (6) (1990) 1285–1301.
[8] R.M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, P. Richtárik, in: K. Chaudhuri, R. Salakhutdinov (Eds.), SGD: General Analysis and Improved Rates, in: Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 5200–5209.

[9] L. Jérôme, A central limit theorem for robbins monro algorithms with projections, 2005, preprint on webpage at https://cermics.enpc.fr/cermics-rapports-recherche/2005/CERMICS-2005/CERMICS-2005-285.pdf. Accessed on 03.20.2022.
[10] T.J. Moore, B.M. Sadler, R.J. Kozick, Maximum-likelihood estimation, the Cramér–Rao bound, and the method of scoring with parameter constraints, IEEE Trans. Signal Process. 56 (3) (2008) 895–908.
[11] A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM J. Optim. 19 (4) (2009) 1574–1609.
[12] M. Pelletier, Asymptotic almost sure efficiency of averaged stochastic algorithms, SIAM J. Control Optim. 39 (1) (2000) 49–72.
[13] D. Pollard, Convergence of Stochastic Processes, Springer-Verlag, Berlin, Heidelberg, 1984.
[14] B.T. Polyak, New method of stochastic approximation type, Autom. Remote Control 51 (7 pt 2) (1990) 937–946.
[15] B.T. Polyak, A.B. Juditsky, Acceleration of stochastic approximation by averaging, SIAM J. Control Optim. 30 (4) (1992) 838–855.
[16] H. Robbins, S. Monro, A stochastic approximation method, Ann. Math. Stat. 22 (3) (1951) 400–407.
[17] H. Robbins, D. Siegmund, A convergence theorem for non negative almost supermartingales and some applications, in: J.S. Rustagi (Ed.), Optimizing Methods in Statistics, Academic Press, 1971, pp. 233–257.
[18] D. Ruppert, Stochastic approximation, in: B.K. Ghosh, P.K. Sen (Eds.), Handbook of Sequential Analysis, Marcel Dekker, New York, 1991, pp. 503–529.
[19] W.J. Su, Y. Zhu, Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent, 2018, arXiv preprint arXiv:1802.04876.
[20] H. Weyl, Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung), Math. Ann. 71 (4) (1912) 441–479.
[21] Y. Yu, T. Wang, R.J. Samworth, A useful variant of the Davis–Kahan theorem for statisticians, Biometrika 102 (2) (2015) 315–323.
[22] T. Zhang, Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms, in: Proceedings of International Conference on Machine Learning, Association for Computing Machinery, New York, 2004, pp. 919–926.