Supplementary Material for Online Statistical Inference for Stochastic Optimization via Kiefer-Wolfowitz Methods

The supplementary material is organized as follows:

1. Section A is a supplementary material to Section 3. Proofs for the two-query approximation are presented in Section A.1. We provide relaxed conditions and the corresponding technical details for the logistic regression and quantile regression in Section A.2. In Section A.3, we further provide illustrations of the choices of directions $\mathcal{P}_{\boldsymbol{v}}$ introduced in Section 3.1 of the main text. Proofs for the multi-query extension (Section 3.2) are given in Section A.4.

2. Section B provides technical details and additional discussions for Section 4. Section B.1 includes the proof of theoretical results of online statistical inference procedures. Section B.2 provides results for the `(KW)` version of stochastic Newton's method as a bi-product of the results in Section 4.

3. In Section C, we present additional results of numerical experiments.

Throughout the supplementary material, we will assume, without loss of generality, $F(\cdot)$ achieves its minimum at $\boldsymbol{\theta}^\star = \mathbf{0}$ and $F(\mathbf{0}) = 0$. We now introduce some notations as follows,

$$\begin{aligned}
\boldsymbol{\xi}_n &= \nabla F(\boldsymbol{\theta}_{n-1}) - \mathbb{E}_{n-1}(\frac{1}{h_n}[F(\boldsymbol{\theta}_{n-1} + h_n\boldsymbol{v}_n) - F(\boldsymbol{\theta}_{n-1})]\boldsymbol{v}_n),\\
\boldsymbol{\gamma}_n &= \mathbb{E}_{n-1}\frac{1}{h_n}[F(\boldsymbol{\theta}_{n-1} + h_n\boldsymbol{v}_n) - F(\boldsymbol{\theta}_{n-1})]\boldsymbol{v}_n - \frac{1}{h_n}[F(\boldsymbol{\theta}_{n-1} + h_n\boldsymbol{v}_n) - F(\boldsymbol{\theta}_{n-1})]\boldsymbol{v}_n,\\
\boldsymbol{\varepsilon}_n &= \frac{1}{h_n}[F(\boldsymbol{\theta}_{n-1} + h_n\boldsymbol{v}_n) - F(\boldsymbol{\theta}_{n-1})]\boldsymbol{v}_n - \frac{1}{h_n}[f(\boldsymbol{\theta}_{n-1} + h_n\boldsymbol{v}_n;\boldsymbol{\zeta}_n) - f(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n)]\boldsymbol{v}_n.
\end{aligned}$$

A Supplementary Material for Section 3

A.1 Two-query approximation

Proof of Lemma 2.4

Proof. By definition, $\mathbb{E}_{\boldsymbol{\zeta}}\, \widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta}) = \frac{1}{h}\Delta_{h,\boldsymbol{v}}F(\boldsymbol{\theta})\boldsymbol{v} = \frac{1}{h}\left[F(\boldsymbol{\theta}+h\boldsymbol{v}) - F(\boldsymbol{\theta})\right]\boldsymbol{v}$. For the first inequality, we have

$$\begin{aligned}
\|\mathbb{E}\, \widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta})\| &= \left\|\mathbb{E}\, \frac{1}{h}\left[F(\boldsymbol{\theta}+h\boldsymbol{v}) - F(\boldsymbol{\theta})\right]\boldsymbol{v} - \nabla F(\boldsymbol{\theta})\right\| \\
&= \left\|\mathbb{E}\, \boldsymbol{v}\boldsymbol{v}^\top \nabla F(\boldsymbol{\theta}) + \frac{1}{2}h\mathbb{E}\, \boldsymbol{v}\boldsymbol{v}^\top \nabla^2 F(\boldsymbol{\theta}_{h,\boldsymbol{v}})\boldsymbol{v} - \nabla F(\boldsymbol{\theta})\right\| \\
&= \frac{1}{2}h\left\|\mathbb{E}\, \boldsymbol{v}\boldsymbol{v}^\top \nabla^2 F(\boldsymbol{\theta}_{h,\boldsymbol{v}})\boldsymbol{v}\right\| \\
&\leq \frac{1}{2}hL_f\mathbb{E}\|\boldsymbol{v}\|^3, \qquad \text{(A.1)}
\end{aligned}$$

where in the third equality we use the Taylor expansion of $F(\boldsymbol{\theta})$, and $\boldsymbol{\theta}_{h,\boldsymbol{v}}$ comes from the remainder term of the Taylor expansion. □

Proof of Proposition 3.1

Proposition 3.1. *Assume Assumptions 1, 2, and 4 hold. Set the step size as $\eta_n = \eta_0 n^{-\alpha}$ for some constant $\eta_0 > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$ and the spacing parameter as $h_n = h_0 n^{-\gamma}$ for some constant $h_0 > 0$, and $\gamma \in \left(\frac{1}{2}, 1\right)$. The* `(KW)` *iterate $\boldsymbol{\theta}_n$ converges to $\boldsymbol{\theta}^\star$ almost surely. Furthermore, for sufficiently large n, we have for $0 < \delta \leq 2$,*

$$\mathbb{E}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star\|^{2+\delta} \leq Cn^{-\alpha(2+\delta)/2}.$$

where the constant C depends on $d, \lambda, L_f, \alpha, \gamma, \eta_0, h_0$.

Remark A.1. *The parameter dependency in Proposition 3.1 could be given explicitly as follows,*

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\theta}_n - \boldsymbol{\theta}^*\|^2 \leq\ & \exp\left(CM_1\eta_0/(2\alpha-1) + CM_2/(2\beta-1) - C\lambda\eta_0 n^{1-\alpha}/(1-\alpha)\right)\|\boldsymbol{\theta}_0\|^2 \\
&+ M_3\left(\exp\left(-C\lambda\eta_0 n^{1-\alpha}/(1-\alpha)\right) + \frac{\eta_0 n^{-\alpha}}{\lambda}\right) \\
&+ \frac{M_3}{M_1}\exp\left(CM_1\eta_0/(2\alpha-1) + CM_2/(2\beta-1) - C\lambda\eta_0 n^{1-\alpha}/(1-\alpha)\right),
\end{aligned}$$

where the constant C above is a universal constant that does not depend on any constant/parameters in the assumptions. The other terms M_1, M_2, M_3 above are given below,

$$M_1 = C\left(L_f^2\mathbb{E}\|\boldsymbol{v}\|^4 + M^{\frac{2}{2+\delta}}\mathbb{E}\|\boldsymbol{v}\|^4 + L_f^2\right),$$
$$M_2 = CL_f^2\mathbb{E}\|\boldsymbol{v}\|^3,$$
$$M_3 = C\left(\mathbb{E}\|\boldsymbol{v}\|^3 + \left(h_n^2L_f^2\mathbb{E}\|\boldsymbol{v}\|^6 + M^{\frac{2}{2+\delta}}\mathbb{E}\|\boldsymbol{v}\|^4(h_n^2\|\boldsymbol{v}\|^2+1)\right)\right).$$

We will prove both Proposition 3.1 and Remark A.1 below.

Proof. We first give some bounds on $\boldsymbol{\xi}_n, \boldsymbol{\gamma}_n, \boldsymbol{\varepsilon}_n$. By definition, $\mathbb{E}_{n-1}\boldsymbol{\gamma}_n = \mathbb{E}_{n-1}\boldsymbol{\varepsilon}_n = 0$. From (A.1),

$$\|\boldsymbol{\xi}_n\| \le \frac{1}{2}h_nL_f\mathbb{E}\|\boldsymbol{v}\|^3. \tag{A.2}$$

We can bound $\boldsymbol{\gamma}_n$ by the following

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\gamma}_n\|^2 &\le \mathbb{E}\left\|\frac{1}{h_n}[F(\boldsymbol{\theta}_{n-1}+h_n\boldsymbol{v}_n) - F(\boldsymbol{\theta}_{n-1})]\boldsymbol{v}_n\right\|^2\\ &\le \mathbb{E}\|\langle\nabla F(\boldsymbol{\theta}_{n-1}), \boldsymbol{v}_n\rangle\boldsymbol{v}_n\|^2 + \frac{1}{4}h_n^2L_f^2\mathbb{E}\|\boldsymbol{v}\|^6\\ &\le L_f^2\mathbb{E}\|\boldsymbol{v}\|^4\mathbb{E}\|\boldsymbol{\theta}_{n-1}\|^2 + \frac{1}{4}h_n^2L_f^2\mathbb{E}\|\boldsymbol{v}\|^6.\end{aligned} \tag{A.3}$$

We also have the following fact for $\boldsymbol{\varepsilon}$.

$$\begin{aligned}&\mathbb{E}_{n-1}\left[\|\boldsymbol{\varepsilon}_n\|^2|\boldsymbol{v}_n\right]\\ =\ &\mathbb{E}_{n-1}\left[\left\|\frac{1}{h_n}\int_0^{h_n}\langle\nabla F(\boldsymbol{\theta}_{n-1}+s\boldsymbol{v}_n) - \nabla f(\boldsymbol{\theta}_{n-1}+s\boldsymbol{v}_n;\boldsymbol{\zeta}_n), \boldsymbol{v}_n\rangle\boldsymbol{v}_n\,\mathrm{d}s\right\|^2\Big|\boldsymbol{v}_n\right]\\ \le\ &\|\boldsymbol{v}_n\|^4\mathbb{E}_{n-1}\left[\frac{1}{h_n}\int_0^{h_n}\|\nabla F(\boldsymbol{\theta}_{n-1}+s\boldsymbol{v}_n) - \nabla f(\boldsymbol{\theta}_{n-1}+s\boldsymbol{v}_n;\boldsymbol{\zeta}_n)\|^2\,\mathrm{d}s\Big|\boldsymbol{v}_n\right]\\ \le\ &M^{\frac{2}{2+\delta}}\|\boldsymbol{v}_n\|^4\frac{1}{h_n}\int_0^{h_n}(\|\boldsymbol{\theta}_{n-1}+s\boldsymbol{v}_n\|^2+1)\,\mathrm{d}s\\ \le\ &M^{\frac{2}{2+\delta}}\|\boldsymbol{v}_n\|^4(\|\boldsymbol{\theta}_{n-1}\|^2 + h_n^2\|\boldsymbol{v}_n\|^2+1),\end{aligned} \tag{A.4}$$

where in the second inequality, we use Assumption 2.

Now decompose the update step as follows,

$$\begin{aligned}\boldsymbol{\theta}_n &= \boldsymbol{\theta}_{n-1} - \eta_n\frac{1}{h_n}[f(\boldsymbol{\theta}_{n-1}+h_n\boldsymbol{u}_n;\boldsymbol{\zeta}_n) - f(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n)]\\ &= \boldsymbol{\theta}_{n-1} - \eta_n\nabla F(\boldsymbol{\theta}_{n-1}) + \eta_n(\boldsymbol{\xi}_n+\boldsymbol{\gamma}_n+\boldsymbol{\varepsilon}_n).\end{aligned}$$

Therefore, we can derive that,

$$\begin{aligned}\|\boldsymbol{\theta}_n\|^2 \leq & \|\boldsymbol{\theta}_{n-1}\|^2 - 2\eta_n\langle\nabla F(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_{n-1}\rangle + 2\eta_n\langle\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n, \boldsymbol{\theta}_{n-1}\rangle \\ & + \eta_n^2\|\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n - \nabla F(\boldsymbol{\theta}_{n-1})\|^2. \end{aligned} \tag{A.5}$$

For the first part in the RHS of the above inequality, we have,

$$\langle\nabla F(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_{n-1}\rangle \geq F(\boldsymbol{\theta}_{n-1}) + \frac{\lambda}{2}\|\boldsymbol{\theta}_{n-1}\|^2 \geq \lambda\|\boldsymbol{\theta}_{n-1}\|^2,$$

with strong convexity property. Moreover,

$$\begin{aligned} |\eta_n\mathbb{E}_{n-1}\langle\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n, \boldsymbol{\theta}_{n-1}\rangle| &= \eta_n\left|\mathbb{E}_{n-1}\langle\boldsymbol{\xi}_n, \boldsymbol{\theta}_{n-1}\rangle\right| \\ &\leq \frac{1}{2}\eta_n h_n L_f\|\boldsymbol{\theta}_{n-1}\|\mathbb{E}\|\boldsymbol{v}\|^3 \\ &\leq CL_f^2\mathbb{E}\|\boldsymbol{v}\|^3 h_n^2\|\boldsymbol{\theta}_{n-1}\|^2 + C\mathbb{E}\|\boldsymbol{v}\|^3\eta_n^2, \end{aligned} \tag{A.6}$$

$$\begin{aligned} \mathbb{E}_{n-1}\|\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n - \nabla F(\boldsymbol{\theta}_{n-1})\|^2 &\leq 4\|\boldsymbol{\xi}_n\|^2 + 4\|\boldsymbol{\gamma}_n\|^2 + 4\|\boldsymbol{\varepsilon}_n\|^2 + 4\|\nabla F(\boldsymbol{\theta}_{n-1})\|^2 \\ &\leq h_n^2 L_f^2\mathbb{E}(\|\boldsymbol{v}\|^3)^2 + 4L_f^2\mathbb{E}\|\boldsymbol{v}\|^4\|\boldsymbol{\theta}_{n-1}\|^2 + h_n^2 L_f^2\mathbb{E}\|\boldsymbol{v}\|^6 \\ &\quad + 4M^{\frac{2}{2+\delta}}\mathbb{E}\|\boldsymbol{v}_n\|^4(\|\boldsymbol{\theta}_{n-1}\|^2 + h_n^2\|\boldsymbol{v}_n\|^2 + 1) + 4L_f^2\|\boldsymbol{\theta}_{n-1}\|^2 \\ &:= M_1\|\boldsymbol{\theta}_{n-1}\|^2 + M_2 \end{aligned} \tag{A.7}$$

where we use Cauchy-Schwarz inequality in (A.6), (A.7) and $M_1 = C\big(L_f^2\mathbb{E}\|\boldsymbol{v}\|^4 + M^{\frac{2}{2+\delta}}\mathbb{E}\|\boldsymbol{v}\|^4 + L_f^2\big), M_2 = C\big(h_n^2 L_f^2\mathbb{E}\|\boldsymbol{v}\|^6 + M^{\frac{2}{2+\delta}}\mathbb{E}\|\boldsymbol{v}\|^4(h_n^2\|\boldsymbol{v}\|^2 + 1)\big)$. So combining all inequalities, we have

$$\mathbb{E}_{n-1}\|\boldsymbol{\theta}_n\|^2 \leq \left[1 - 2\lambda\eta_n + M_1\eta_n^2 + M_3 h_n^2\right]\|\boldsymbol{\theta}_{n-1}\|^2 + M_4\eta_n^2, \tag{A.8}$$

where M_3, M_4 is defined by $M_3 = CL_f^2\mathbb{E}\|\boldsymbol{v}\|^3$, $M_4 = C(\mathbb{E}\|\boldsymbol{v}\|^3 + M_2)$. Following the proof of Theorem 1 of Moulines and Bach (2011), we can apply the recursion and get

$$\mathbb{E}\|\boldsymbol{\theta}_n\|^2 \leq \prod_{k=1}^{n}\left[1 - 2\lambda\eta_k + M_1\eta_k^2 + CM_3 h_k^2\right]\|\boldsymbol{\theta}_0\|^2 + M_4\sum_{k=1}^{n}\prod_{i=k+1}^{n}\left[1 - 2\lambda\eta_i + M_1\eta_k^2 + M_3 h_k^2\right]\eta_k^2.$$

We can then bound the first term on the RHS,

$$\prod_{k=1}^{n}\left[1 - 2\lambda\eta_k + M_1\eta_k^2 + M_3 h_k^2\right] \leq \exp\left(-2\lambda\sum_{k=1}^{n}\eta_k\right)\exp\left(M_1\sum_{k=1}^{n}\eta_k^2\right)\exp\left(M_3\sum_{k=1}^{n}h_k^2\right),$$

as well as the second term on the RHS

$$\begin{aligned}
&\sum_{k=1}^{n}\prod_{i=k+1}^{n}\left[1-2\lambda\eta_i+M_1\eta_k^2+M_3h_k^2\right]\eta_k^2\\
\leq\ &\exp\left(-\lambda\sum_{k=m+1}^{n}\eta_k\right)\sum_{k=1}^{n}\eta_k^2+\frac{\eta_m}{\lambda}+\frac{1}{M_1}\exp\left(M_1\sum_{k=1}^{n_0}\eta_k^2\right)\exp\left(M_3\sum_{k=1}^{n_0}h_k^2\right)\exp\left(-\lambda\sum_{k=1}^{n}\eta_k\right),
\end{aligned}$$

where we denote by $n_0=\inf\{k\in\mathbb{N}, 1-2\lambda\eta_k+M_1\eta_k^2+M_3h_k^2\leq 1-\lambda\eta_k\}$ and m is any integer in $\{1,\ldots,n\}$. Choose $m=n/2$ and bound n_0 by n. Notice that $\sum_{k=1}^{n}\eta_k^2$ converge. So we can get

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\theta}_n\|^2\leq\ &\exp\left(CM_1\eta_0/(2\alpha-1)+CM_3/(2\beta-1)-C\lambda\eta_0n^{1-\alpha}/(1-\alpha)\right)\|\boldsymbol{\theta}_0\|^2\\
&+M_4\left(\exp\left(-C\lambda\eta_0n^{1-\alpha}/(1-\alpha)\right)+\frac{\eta_0n^{-\alpha}}{\lambda}\right)\\
&+\frac{M_4}{M_1}\exp\left(CM_1\eta_0/(2\alpha-1)+CM_3/(2\beta-1)-C\lambda\eta_0n^{1-\alpha}/(1-\alpha)\right).
\end{aligned}$$

Only the term $M_4\eta_0n^{-\alpha}/\lambda$ decreases at the order of $O(n^{-\alpha})$ while all the other terms decrease much faster.

Notice that up to this point, all C's are universal constants which do not depend on any parameters in the assumptions. From now on, we will absorb all parameters (other than n) into C to make the asymptotic analysis more clear.

By martingale convergence theorem, $\|\boldsymbol{\theta}_n\|$ converges almost surely. Because its second moment converges to $\mathbf{0}$, it must converge to $\mathbf{0}$ almost surely.

We now show that,

$$\mathbb{E}\|\boldsymbol{\theta}_n-\boldsymbol{\theta}^\star\|^{2+\delta}\leq Cn^{-\alpha(2+\delta)/2}.$$

By same arguments as in (A.2), (A.3), (A.4), we can get $\|\boldsymbol{\xi}_n\|^{2+\delta}\leq Ch_n^{2+\delta}$, $\mathbb{E}_{n-1}\|\boldsymbol{\gamma}_n\|^{2+\delta}\leq\|\boldsymbol{\theta}_{n-1}\|^{2+\delta}+Ch_n^{2+\delta}$, $\mathbb{E}_{n-1}\left[\|\boldsymbol{\varepsilon}_n\|^{2+\delta}\right]\leq C(\|\boldsymbol{\theta}_{n-1}\|^{2+\delta}+1)$.

By similar arguments as in Lemma B.3, there exists constants C such that for any $\boldsymbol{a},\boldsymbol{b}$,

$$\|\boldsymbol{a}+\boldsymbol{b}\|^{2+\delta}\leq\|\boldsymbol{a}\|^{2+\delta}+(2+\delta)\langle\boldsymbol{a},\boldsymbol{b}\rangle\|\boldsymbol{a}\|^{\delta}+C\|\boldsymbol{a}\|^{\delta}\|\boldsymbol{b}\|^2+C\|\boldsymbol{b}\|^{2+\delta}.$$

So we have the bound

$$\begin{aligned}\mathbb{E}_{n-1}\|\boldsymbol{\theta}_n\|^{2+\delta} \le \|\boldsymbol{\theta}_{n-1}\|^{2+\delta} &+ \eta_n(2+\delta)\mathbb{E}_{n-1}\langle\boldsymbol{\theta}_{n-1}, -\nabla F(\boldsymbol{\theta}_{n-1}) + \boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n\rangle\|\boldsymbol{\theta}_{n-1}\|^{\delta} \\ &+ C\eta_n^2\|\boldsymbol{\theta}_{n-1}\|^{\delta}\mathbb{E}_{n-1}\| - \nabla F(\boldsymbol{\theta}_{n-1}) + \boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n\|^2 \\ &+ C\eta_n^{2+\delta}\mathbb{E}_{n-1}\| - \nabla F(\boldsymbol{\theta}_{n-1}) + \boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n\|^{2+\delta} \\ \le (1-(2+\delta)\lambda\eta_n)\|\boldsymbol{\theta}_{n-1}\|^{2+\delta} &+ C\eta_n h_n\|\boldsymbol{\theta}_{n-1}\|^{1+\delta} \\ &+ C\eta_n^2(\|\boldsymbol{\theta}_{n-1}\|^2+1)\|\boldsymbol{\theta}_{n-1}\|^{\delta} + C\eta_n^{2+\delta}(\|\boldsymbol{\theta}_{n-1}\|^{2+\delta}+1).\end{aligned}$$

If $0 < \delta \le 1$, by previous bound $\mathbb{E}\|\boldsymbol{\theta}_n\|^2 \le Cn^{-\alpha}$, we can get $\mathbb{E}\|\boldsymbol{\theta}_n\|^{1+\delta} \le Cn^{-\alpha(1+\delta)/2}$ and $\mathbb{E}\|\boldsymbol{\theta}_n\|^{\delta} \le Cn^{-\alpha\delta/2}$ by Hölder's inequality. So we can further get

$$\mathbb{E}\|\boldsymbol{\theta}_n\|^{2+\delta} \le (1 - Cn^{-\alpha} + Cn^{-2\alpha})\mathbb{E}\|\boldsymbol{\theta}_{n-1}\|^{2+\delta} + Cn^{-(2+\delta)\alpha/2},$$

which implies $\mathbb{E}\|\boldsymbol{\theta}_n\|^{2+\delta} \le Cn^{-(2+\delta)\alpha/2}$ as in the above proof after (A.8).

Now the case for $0 < \delta \le 1$ is proved. We can then use induction. If $\mathbb{E}\|\boldsymbol{\theta}_n\|^{2+\delta} \le Cn^{-(2+\delta)\alpha/2}$ for all $\delta \le n$, then we can use the same method to prove the same inequality holds for $\delta \in (n, n+1]$. Thus the inequality holds for all δ. □

Proof of Lemma 3.2

Proof. By Assumption 2, we know that

$$\mathbb{E}\|\nabla f(\boldsymbol{\theta};\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta})\|^{2+\delta} \le M(\|\boldsymbol{\theta}\|^{2+\delta} + d^{2+\delta}).$$

Therefore, the following holds some constant $C > 0$,

$$\mathbb{E}\|\nabla f(\boldsymbol{\theta};\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta})\|^2 \le C(\|\boldsymbol{\theta}\|^2 + d^2). \tag{A.9}$$

In particular,

$$\mathbb{E}\|\nabla f(\mathbf{0};\boldsymbol{\zeta}) - \nabla F(\mathbf{0})\|^2 \le C. \tag{A.10}$$

From Assumption 3, we can get the following estimate for the Hessian matrix $\nabla^2 f(\boldsymbol{\theta};\boldsymbol{\zeta})$,

$$\begin{aligned}\mathbb{E}\|\nabla^2 f(\boldsymbol{\theta};\boldsymbol{\zeta})\|^2 &\le 2\mathbb{E}\|\nabla^2 f(\mathbf{0};\boldsymbol{\zeta})\|^2 + 2\mathbb{E}\left\|\nabla^2 f(\boldsymbol{\theta};\boldsymbol{\zeta}) - \nabla^2 f(\mathbf{0};\boldsymbol{\zeta})\right\|^2 \\ &\le C(1+\|\boldsymbol{\theta}\|^2).\end{aligned}$$

Using the above observation, we find that

$$\begin{aligned}
&\mathbb{E}\|\nabla f(\boldsymbol{\theta};\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta}) - \nabla f(\mathbf{0};\boldsymbol{\zeta}) + \nabla F(\mathbf{0})\|^2 \\
&\le C\|\boldsymbol{\theta}\|^2 + 2\mathbb{E}\|\nabla f(\boldsymbol{\theta};\boldsymbol{\zeta}) - \nabla f(\mathbf{0};\boldsymbol{\zeta}))\|^2 \\
&= C\|\boldsymbol{\theta}\|^2 + 2\mathbb{E}\left\|\int_0^1 \nabla^2 f(s\boldsymbol{\theta};\boldsymbol{\zeta})\boldsymbol{\theta}\,\mathrm{d}s\right\|^2 \\
&\le C\|\boldsymbol{\theta}\|^2 + 2\mathbb{E}\int_0^1 \|\nabla^2 f(s\boldsymbol{\theta};\boldsymbol{\zeta})\boldsymbol{\theta}\|^2\,\mathrm{d}s \\
&\le C\|\boldsymbol{\theta}\|^2(1 + \int_0^1 \mathbb{E}\|\nabla^2 f(s\boldsymbol{\theta};\boldsymbol{\zeta})\|^2\,\mathrm{d}s) \\
&\le C\|\boldsymbol{\theta}\|^2(1 + \|\boldsymbol{\theta}\|^2). \qquad \text{(A.11)}
\end{aligned}$$

Define the function $\Sigma(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ by

$$\Sigma(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \mathbb{E}(\nabla f(\boldsymbol{\theta}_1;\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta}_1))(\nabla f(\boldsymbol{\theta}_2;\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta}_2))^\top.$$

Then combining inequalities (A.9), (A.10), (A.11), we have

$$\begin{aligned}
\|\Sigma(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) - S\| &\le \mathbb{E}\|(\nabla f(\boldsymbol{\theta}_1;\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta}_1))(\nabla f(\boldsymbol{\theta}_2;\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta}_2))^\top \\
&\quad - (\nabla f(\mathbf{0};\boldsymbol{\zeta}) - \nabla F(\mathbf{0}))(\nabla f(\mathbf{0};\boldsymbol{\zeta}) - \nabla F(\mathbf{0}))^\top\| \\
&\le \mathbb{E}\|\nabla f(\boldsymbol{\theta}_1;\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta}_1)\|\|\nabla f(\boldsymbol{\theta}_2;\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta}_2) - \nabla f(\mathbf{0};\boldsymbol{\zeta}) + \nabla F(\mathbf{0})\| \\
&\quad + \mathbb{E}\|\nabla f(\boldsymbol{\theta}_1;\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta}_2) - \nabla f(\mathbf{0};\boldsymbol{\zeta}) + \nabla F(\mathbf{0})\|\|\nabla f(\mathbf{0};\boldsymbol{\zeta}) - \nabla F(\mathbf{0})\| \\
&\le C(d + \|\boldsymbol{\theta}_1\|)\|\boldsymbol{\theta}_2\|(1 + \|\boldsymbol{\theta}_2\|) + C\|\boldsymbol{\theta}_1\|(1 + \|\boldsymbol{\theta}_1\|). \qquad \text{(A.12)}
\end{aligned}$$

Notice that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\zeta}}\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta})\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta})^\top - (\frac{1}{h}\Delta_{h,\boldsymbol{v}}F(\boldsymbol{\theta})\boldsymbol{v})(\frac{1}{h}\Delta_{h,\boldsymbol{v}}F(\boldsymbol{\theta})\boldsymbol{v})^\top \\
&= \mathbb{E}_{\boldsymbol{\zeta}}(\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta}) - \frac{1}{h}\Delta_{h,\boldsymbol{v}}F(\boldsymbol{\theta})\boldsymbol{v})(\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta}) - \frac{1}{h}\Delta_{h,\boldsymbol{v}}F(\boldsymbol{\theta})\boldsymbol{v})^\top \\
&= \frac{1}{h^2}\mathbb{E}_{\boldsymbol{\zeta}}\boldsymbol{v}(f(\boldsymbol{\theta} + h\boldsymbol{v};\boldsymbol{\zeta}) - f(\boldsymbol{\theta};\boldsymbol{\zeta}) - F(\boldsymbol{\theta} + h\boldsymbol{v}) + F(\boldsymbol{\theta}))^2\boldsymbol{v}^\top \\
&= \frac{1}{h^2}\mathbb{E}_{\boldsymbol{\zeta}}\boldsymbol{v}\boldsymbol{v}^\top\Big[\int_0^h\int_0^h (\nabla F(\boldsymbol{\theta} + s_1\boldsymbol{v}) - \nabla f(\boldsymbol{\theta} + s_1\boldsymbol{v};\boldsymbol{\zeta})) \\
&\quad (\nabla F(\boldsymbol{\theta} + s_2\boldsymbol{v}) - \nabla f(\boldsymbol{\theta} + s_2\boldsymbol{v};\boldsymbol{\zeta}))^\top\,\mathrm{d}s_1\,\mathrm{d}s_2\Big]\boldsymbol{v}\boldsymbol{v}^\top \\
&= \frac{1}{h^2}\mathbb{E}_{\boldsymbol{\zeta}}\boldsymbol{v}\boldsymbol{v}^\top\int_0^h\int_0^h \Sigma(\boldsymbol{\theta} + s_1\boldsymbol{v}, \boldsymbol{\theta} + s_2\boldsymbol{v})\,\mathrm{d}s_1\,\mathrm{d}s_2\boldsymbol{v}\boldsymbol{v}^\top.
\end{aligned}$$

We can use (A.12) and derive that

$$\begin{aligned}&\|\mathbb{E}_{\boldsymbol{\zeta}}\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta})\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta})^\top - (\frac{1}{h}\Delta_{h,\boldsymbol{v}}F(\boldsymbol{\theta})\boldsymbol{v})(\frac{1}{h}\Delta_{h,\boldsymbol{v}}F(\boldsymbol{\theta})\boldsymbol{v})^\top - \boldsymbol{v}\boldsymbol{v}^\top S\boldsymbol{v}\boldsymbol{v}^\top\| \\ \leq\ & C\|\boldsymbol{v}\|^4(\|\boldsymbol{\theta}\| + h\|\boldsymbol{v}\|)(1 + \|\boldsymbol{\theta}\| + h\|\boldsymbol{v}\|)(d + \|\boldsymbol{\theta}\| + h\|\boldsymbol{v}\|).\end{aligned}$$

Now we have

$$\begin{aligned}&\|\mathbb{E}\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta})\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta})^\top - \mathbb{E}(\frac{1}{h}\Delta_{h,\boldsymbol{v}}F(\boldsymbol{\theta})\boldsymbol{v})(\frac{1}{h}\Delta_{h,\boldsymbol{v}}F(\boldsymbol{\theta})\boldsymbol{v})^\top - \mathbb{E}\boldsymbol{v}\boldsymbol{v}^\top S\boldsymbol{v}\boldsymbol{v}^\top\| \\ \leq\ & C\mathbb{E}\|\boldsymbol{v}\|^4(\|\boldsymbol{\theta}\| + h\|\boldsymbol{v}\|)(1 + \|\boldsymbol{\theta}\| + h\|\boldsymbol{v}\|)(d + \|\boldsymbol{\theta}\| + h\|\boldsymbol{v}\|).\end{aligned} \tag{A.13}$$

By the same argument,

$$\begin{aligned}&\|\mathbb{E}(\frac{1}{h}\Delta_{h,\boldsymbol{v}}F(\boldsymbol{\theta})\boldsymbol{v})(\frac{1}{h}\Delta_{h,\boldsymbol{v}}F(\boldsymbol{\theta})\boldsymbol{v})^\top\| \\ \leq\ & \frac{1}{h^2}\mathbb{E}\left\|\boldsymbol{v}\boldsymbol{v}^\top\left[\int_0^h\int_0^h (\nabla F(\boldsymbol{\theta}+s_1\boldsymbol{v}))\,(\nabla F(\boldsymbol{\theta}+s_2\boldsymbol{v}))^\top\,\mathrm{d}s_1\,\mathrm{d}s_2\right]\boldsymbol{v}\boldsymbol{v}^\top\right\| \\ \leq\ & C\mathbb{E}\|\boldsymbol{v}\|^4(\|\boldsymbol{\theta}\|^2 + h^2\|\boldsymbol{v}\|^2).\end{aligned}$$

So we finally get

$$\|\mathbb{E}\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta})\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta})^\top - \mathbb{E}\boldsymbol{v}\boldsymbol{v}^\top S\boldsymbol{v}\boldsymbol{v}^\top\| \leq C\mathbb{E}\|\boldsymbol{v}\|^4(\|\boldsymbol{\theta}\| + h\|\boldsymbol{v}\|)(1 + \|\boldsymbol{\theta}\| + h\|\boldsymbol{v}\|)(d + \|\boldsymbol{\theta}\| + h\|\boldsymbol{v}\|).$$

for some constant $C > 0$. □

Proof of Theorem 3.3

Proof. We follow the proof in Polyak and Juditsky (1992). The update step is

$$\begin{aligned}\boldsymbol{\theta}_n &= \boldsymbol{\theta}_{n-1} - \eta_n\nabla F(\boldsymbol{\theta}_{n-1}) + \eta_n(\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n) \\ &= (I_d - \eta_n H)\boldsymbol{\theta}_{n-1} + \eta_n(H\boldsymbol{\theta}_{n-1} - \nabla F(\boldsymbol{\theta}_{n-1}) + \boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n).\end{aligned}$$

By the argument in Polyak and Juditsky (1992), we only need to prove the following three conditions. First,

$$\sum_{i=1}^{\infty}\frac{1}{\sqrt{i}}\mathbb{E}\|H\boldsymbol{\theta}_{i-1} - \nabla F(\boldsymbol{\theta}_{i-1}) + \boldsymbol{\xi}_i\|, \tag{A.14}$$

is bounded almost surely. Furthermore, we have

$$\mathbb{E}\|\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i\|^2, \tag{A.15}$$

is bounded almost surely. Moreover, when $t \to \infty$, we have the following convergence in probability,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i) \Longrightarrow \mathcal{N}(\mathbf{0}, Q). \tag{A.16}$$

By Assumption 3, we know that

$$\|\nabla^2 F(\boldsymbol{\theta}) - \nabla^2 F(\boldsymbol{\theta'})\|^2 \leq L_g \|\boldsymbol{\theta} - \boldsymbol{\theta'}\|^2.$$

By Taylor expansion,

$$\|H\boldsymbol{\theta}_{i-1} - \nabla F(\boldsymbol{\theta}_{i-1})\| \leq C\|\boldsymbol{\theta}_{i-1}\|^2.$$

By facts (A.2) to (A.4), we know that

$$\mathbb{E}\|H\boldsymbol{\theta}_{i-1} - \nabla F(\boldsymbol{\theta}_{i-1}) + \boldsymbol{\xi}_i\| \leq Ci^{-\alpha},$$

which indicates that condition (A.14) holds.

Because $\boldsymbol{\gamma}_i$ converges to $\mathbf{0}$ almost surely and $\boldsymbol{\varepsilon}_i$ has bounded variance. So condition (A.15) holds. To prove condition (A.16), it suffices to verify that,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i \Longrightarrow \mathcal{N}(\mathbf{0}, Q).$$

By martingale central limit theorem (Durrett, 2019, Theorem 8.2.8), we only need to verify two conditions,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{i-1}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^\top] \to Q, \tag{A.17}$$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\|\boldsymbol{\varepsilon}_i\|^2 \mathbf{1}_{\|\boldsymbol{\varepsilon}_i\| > a\sqrt{n}}\right] \to 0, \tag{A.18}$$

in probability for all $a > 0$.

Notice that (A.13) is equivalent to the following inequality,

$$\|\mathbb{E}_{n-1}\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^\top - \mathbb{E}\boldsymbol{v}\boldsymbol{v}^\top S\boldsymbol{v}\boldsymbol{v}^\top\| \leq C(\|\boldsymbol{\theta}_{n-1}\| + h_n)(1 + \|\boldsymbol{\theta}_{n-1}\|^3 + h_n^3). \tag{A.19}$$

Thus $\mathbb{E}_{n-1}[\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^\top]$ converges almost surely to Q and condition (A.17) holds.

Now consider the quantity in (A.18), by Proposition 3.1,

$$\mathbb{E}_{i-1}\left[\|\boldsymbol{\varepsilon}_i\|^2\mathbf{1}_{\|\boldsymbol{\varepsilon}_i\|>a\sqrt{n}}\right] \le \left[\mathbb{E}_{i-1}\left[\|\boldsymbol{\varepsilon}_i\|^{2+\delta}\right]\right]^{\frac{2}{2+\delta}}\left[\mathbb{E}_{i-1}\left[\mathbf{1}_{\|\boldsymbol{\varepsilon}_i\|>a\sqrt{n}}\right]\right]^{\frac{\delta}{2+\delta}}.$$

Note that

$$\mathbb{E}_{i-1}\left[\mathbf{1}_{\|\boldsymbol{\varepsilon}_i\|>a\sqrt{n}}\right] = \mathbb{P}_{i-1}\left(\|\boldsymbol{\varepsilon}_i\| > a\sqrt{n}|\boldsymbol{\theta}_{i-1}\right) \le \frac{1}{a\sqrt{n}}\mathbb{E}_{i-1}\|\boldsymbol{\varepsilon}_i\|.$$

Therefore, it can be bounded by

$$\mathbb{E}_{i-1}\left[\|\boldsymbol{\varepsilon}_i\|^2\mathbf{1}_{\|\boldsymbol{\varepsilon}_i\|>a\sqrt{n}}\right] \le C\left(\frac{1}{a\sqrt{n}}\right)^{\frac{\delta}{2+\delta}}\left(1+\|\boldsymbol{\theta}_{i-1}\|^{2+\delta}\right)^{\frac{2}{2+\delta}}\left(1+\|\boldsymbol{\theta}_{i-1}\|\right)^{\frac{\delta}{2+\delta}},$$

from which we can obtain that

$$\mathbb{E}[\|\boldsymbol{\varepsilon}_i\|^2\mathbf{1}_{\|\boldsymbol{\varepsilon}_i\|>a\sqrt{n}}] \le C\left(\frac{1}{a\sqrt{n}}\right)^{\frac{\delta}{2+\delta}}. \tag{A.20}$$

We find that condition (A.18) holds when n goes to infinity:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\boldsymbol{\varepsilon}_i\|^2\mathbf{1}_{\|\boldsymbol{\varepsilon}_i\|>a\sqrt{n}}\right] \le C\left(\frac{1}{a\sqrt{n}}\right)^{\frac{\delta}{2+\delta}} \to 0.$$

Therefore, we conclude the result. □

Proof of Proposition 3.5

Proof. For $Q^{(\mathtt{G})}$, let $\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},I_d)$, and we now calculate $\mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top S\boldsymbol{z}\boldsymbol{z}^\top$. The (i,i)-th entry is

$$\mathbb{E}\sum_{j,k} z_iz_jS_{jk}z_kz_i = \sum_{j\neq i}S_{jj}+3S_{ii} = 2S_{ii}+\mathrm{tr}(S).$$

For $i\neq j$, the (i,j)th entry is

$$\mathbb{E}\sum_{k,l} z_iz_kS_{kl}z_lz_j = 2S_{ij}.$$

So $\mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top S\boldsymbol{z}\boldsymbol{z}^\top = 2S+\mathrm{tr}(S)I_d$.

For $Q^{(\mathtt{S})}$, let $\boldsymbol{v}$ be sampled from the uniform distribution on the sphere $\|\boldsymbol{v}\|=d$. The Gaussian vector $\boldsymbol{z}$ can be decomposed into independent radius part and spherical part,

$$\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \mathbb{E}\left[\|\boldsymbol{z}\|^2\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}\frac{\boldsymbol{z}^\top}{\|\boldsymbol{z}\|}\right] = \mathbb{E}\boldsymbol{v}\boldsymbol{v}^\top,$$
$$\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top S\boldsymbol{z}\boldsymbol{z}^\top] = \mathbb{E}\left[\|\boldsymbol{z}\|^4\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}\frac{\boldsymbol{z}^\top}{\|\boldsymbol{z}\|}S\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}\frac{\boldsymbol{z}^\top}{\|\boldsymbol{z}\|}\right] = \frac{d+2}{d}\mathbb{E}\boldsymbol{v}\boldsymbol{v}^\top S\boldsymbol{v}\boldsymbol{v}^\top.$$

Now we have

$$\mathbb{E}\boldsymbol{v}\boldsymbol{v}^\top = I_d, \;\; \mathbb{E}\boldsymbol{v}\boldsymbol{v}^\top S\boldsymbol{v}\boldsymbol{v}^\top = \frac{d}{d+2}(2S + \mathrm{tr}(S)I_d).$$

For $Q^{(\mathtt{U})}$, let $\boldsymbol{u}$ obey the uniform distribution on $\{\sqrt{d}e_1, \ldots, \sqrt{d}e_d\}$. By direct calculation, we have

$$\mathbb{E}\boldsymbol{u}\boldsymbol{u}^\top S\boldsymbol{u}\boldsymbol{u}^\top = \sum_{j=1}^{d} \frac{1}{d} \cdot d^2 S_{jj} = d \ \mathrm{diag}(S).$$

The final two cases for $Q^{(\mathtt{U})}, Q^{(\mathtt{P})}$ can also be verified by direct calculation. □

A.2 Extensions to local strong convexity and nonsmoothness

Asymptotic behavior for locally strongly convex loss function

To comply with the settings of the logistic regression, we need to consider a relaxed version of Assumption 1 as follows,

Assumption 1′. *The population loss function $F(\boldsymbol{\theta})$ is twice continuously differentiable, convex and L_f-smooth. In addition, there exists $\delta_1 > 0$ such that for all $\boldsymbol{\theta}$ in the δ_1-ball centered at $\boldsymbol{\theta}^\star$, the Hessian matrix $\nabla^2 F(\boldsymbol{\theta})$ is positive-definite.*

Assumption 1′ considers local strong convexity of the population objective $F(\cdot)$ at the minimizer $\boldsymbol{\theta}^\star$. Intuitively, after a number of steps in the (KW) SGD update, the estimated parameter $\boldsymbol{\theta}_n$ would be sufficiently close to $\boldsymbol{\theta}^\star$ and we have the strong convexity from there. This assumption naturally suites the settings of the logistic regression.

Theorem A.2. *Let Assumption 1′, and 2 to 4 hold. Set the step size as $\eta_n = \eta_0 n^{-\alpha}$ for some constant $\eta_0 > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$, and the spacing parameter as $h_n = h_0 n^{-\gamma}$ for some constant $h_0 > 0$, and $\gamma \in \left(\frac{1}{2}, 1\right)$. The averaged estimator $\overline{\boldsymbol{\theta}}_n$ satisfies,*

$$\sqrt{n}\ \left(\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\right) \Longrightarrow \mathcal{N}\left(\mathbf{0}, H^{-1}QH^{-1}\right), \qquad as \quad n \to \infty.$$

Proof. Under assumption 1′, using Lemma B.1 in Su and Zhu (2018), for all $\boldsymbol{\theta}$ in the δ_1-ball centered at $\mathbf{0}$, we have

$$\langle \boldsymbol{\theta}, \nabla F(\boldsymbol{\theta}) \rangle \geq \rho \|\boldsymbol{\theta}\| \min\left\{\|\boldsymbol{\theta}\|, \delta_1\right\}. \tag{A.21}$$

for some $\rho > 0$. For the first part in the RHS of the previous inequality (A.5), using inequality (A.21), we have

$$\langle \nabla F(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta}_{n-1} \rangle \geq \rho \|\boldsymbol{\theta}_{n-1}\| \min\left\{\|\boldsymbol{\theta}_{n-1}\|, \delta_1\right\},$$

Furthermore, by (A.6) and (A.7), we have

$$\mathbb{E}_{n-1}\|\boldsymbol{\theta}_n\|^2 \leq \left(1 + C_2^2\eta_n^2\right)\|\boldsymbol{\theta}_{n-1}\|^2 - 2\eta_n\rho\|\boldsymbol{\theta}_{n-1}\| \min\left\{\|\boldsymbol{\theta}_{n-1}\|, \delta_1\right\} + C_2\eta_n^2.$$

By Robbins-Siegmund theorem, $\|\boldsymbol{\theta}_n\|^2$ converges to some random variable almost surely and

$$\sum_{n=1}^{\infty} 2\eta_n\rho\|\boldsymbol{\theta}_n\| \min\left\{\|\boldsymbol{\theta}_n\|, \delta_1\right\} < \infty.$$

Combining the fact that $\sum_{n=1}^{\infty}\eta_n = \infty$ we can yield that $\boldsymbol{\theta}_n$ converges almost surely to $\mathbf{0}$. The rest part follows from the proof of Theorem 3.3. □

Asymptotic behavior of (AKW) estimator for nonsmooth loss functions

We present Theorem 3.3 in the main paper with strengthened Assumptions 2 and 3, where we assume the existence of the gradient of the inaccessible (RM) stochastic gradient $g(\boldsymbol{\theta};\boldsymbol{\zeta})$ and its Gram matrix $S = \mathbb{E}[g(\boldsymbol{\theta}^\star;\boldsymbol{\zeta})g(\boldsymbol{\theta}^\star;\boldsymbol{\zeta})^\top]$. The theoretical analysis of the asymptotic distribution of the (AKW) estimator remains working with a weakened assumption, which is a natural fit to some nonsmooth loss functions $F(\boldsymbol{\theta})$ including the quantile regression in Example 2.3.

Assumption A.3. *Assume there exists $C_1 > 0$ such that $\left\|\mathbb{E}\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta}) - \nabla F(\boldsymbol{\theta})\right\| \leq C_1 h$ for any $h > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^d$. Further assume there exists $C_2 > 0$ such that $\mathbb{E}\left[\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta}^\star;\boldsymbol{\zeta})\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta}^\star;\boldsymbol{\zeta})^\top\right] = Q + \Delta_h$ for some matrix $Q \in \mathbb{R}^{d\times d}$ and $\|\Delta_h\| \leq C_2 h^\iota$, for some $\iota > 0$. Moreover, for some $0 < \delta \leq 2$, there exists $M > 0$,*

$$\mathbb{E}\|\widehat{g}_{h,\boldsymbol{v}}(\boldsymbol{\theta};\boldsymbol{\zeta}_n) - \nabla F(\boldsymbol{\theta})\|^{2+\delta} \leq M(\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|^{2+\delta} + h^{2+\delta}).$$

Theorem A.4. *Let Assumption 1 and A.3 hold. Under the step size and spacing parameter conditions specified in Theorem 3.3, the averaged estimator $\overline{\boldsymbol{\theta}}_n$ satisfies,*

$$\sqrt{n}\ \left(\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\right) \Longrightarrow \mathcal{N}\left(\mathbf{0}, H^{-1}QH^{-1}\right), \qquad as \quad n \to \infty. \tag{A.22}$$

Proof. Under Assumption 1 and A.3, the conclusions in Lemma 2.4 and Lemma 3.2 naturally hold. The rest of the proof follows from the proof in Proposition 3.1 and Theorem 3.3. □

A.3 Illustration of choices of directions $\mathcal{P}_{\boldsymbol{v}}$

We first note that $Q^{(\mathtt{G})} \succ Q^{(\mathtt{S})}$ regardless of the dimension d and Gram matrix S. Intuitively, when the direction $\boldsymbol{v}$ is generated by Gaussian (G), it can be decomposed into two independent random variables: the radical part $\|\boldsymbol{v}\|$ and the spherical part $\boldsymbol{v}/\|\boldsymbol{v}\|$. The spherical part $\boldsymbol{v}/\|\boldsymbol{v}\|$ follows the same distribution as the uniform distribution on the sphere with radius d (which is identical to (S)). The extra randomness in the radical part $\|\boldsymbol{v}\|^2 \sim \chi^2(d)$ leads to a larger magnitude of Q compared to that of (S). Therefore the (AKW) estimator with Gaussian directions (G) is always inferior to that with spherical directions (S), asymptotically. However, for the other candidates, they are not directly comparable, and the optimal choice of $\mathcal{P}_{\boldsymbol{v}}$ depends on the optimality criterion, and Gram matrix S.

As a simple illustration, we consider $S = \mathrm{diag}(1, r_0)$ for some $r_0 > 0$. We have

(S) Spherical: $Q^{(\mathtt{S})} = \mathrm{diag}\left(\frac{r_0+3}{2}, \frac{3r_0+1}{2}\right)$.

(I) Uniform in a natural coordinate basis: $Q^{(\mathtt{I})} = \mathrm{diag}(2, 2r_0)$.

(U) Uniform in an arbitrary orthonormal basis U: when $U = (\cos\omega, \sin\omega; -\sin\omega, \cos\omega)$ and $\omega = 0$, we have $Q^{(\mathtt{U})} = Q^{(\mathtt{I})} = \mathrm{diag}(2, 2r_0)$; when $\omega = \pi/4$, we have $Q^{(\mathtt{U})} = \mathrm{diag}(1+r_0, 1+r_0)$.

(P) Non-uniform in a natural coordinate basis: $\mathrm{diag}\left(\frac{1}{p_1}, \frac{r_0}{1-p_1}\right), p_1 \in (0,1)$.

From the above we can see that, the choices of the distribution of direction vectors $\mathcal{P}_{\boldsymbol{v}}$ depends on the optimality-criteria on comparing the covariance matrices. Specifically in the above example, if one seeks to minimize

- the trace of covariance matrix, we have

$$\mathrm{tr}(Q^{(\mathtt{S})}) = \mathrm{tr}(Q^{(\mathtt{I})}) = \mathrm{tr}(Q^{(\mathtt{U})}) = 2 + 2r_0, \quad \mathrm{tr}(Q^{(\mathtt{P})}) = \frac{1}{p_1} + \frac{r_0}{1-p_1},$$

  and the optimal distribution that minimizes the trace depends on the value of p_1.

- the determinant of covariance matrix, we have

$$\det(Q^{(\mathtt{S})}) = \frac{3r_0^2 + 10r_0 + 3}{4}, \qquad \det(Q^{(\mathtt{I})}) = 4r_0,$$
$$\det(Q^{(\mathtt{U})}) = \frac{-\cos(4\omega)(r_0-1)^2 + r_0^2 + 6r_0 + 1}{2}, \quad \det(Q^{(\mathtt{P})}) = \frac{r_0}{p_1(1-p_1)}.$$

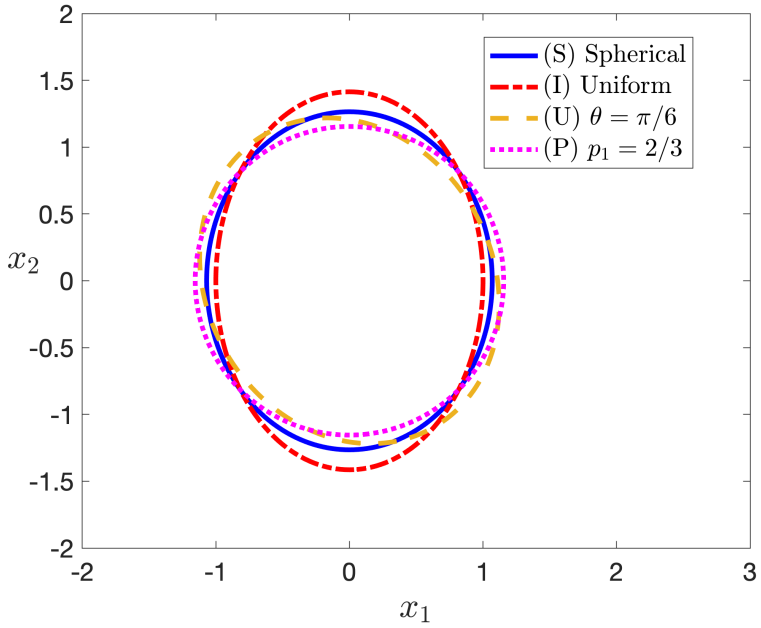


Figure A.1: Comparison of Q matrices under different direction distributions $\mathcal{P}_{\boldsymbol{v}}$ when $S = \mathrm{diag}(1, 1/2)$.

By a simple derivation, we have $\det(Q^{(\mathtt{S})}) \geq \det(Q^{(\mathtt{U})}) \geq \det(Q^{(\mathtt{I})})$ and $\det(Q^{(\mathtt{P})}) \geq \det(Q^{(\mathtt{I})})$.

- the operator norm of covariance matrix, i.e., the largest eigenvalue, we have

$$\lambda_{\max}(Q^{(\mathtt{S})}) = \frac{r_0 + 3}{2}, \qquad \lambda_{\max}(Q^{(\mathtt{I})}) = 2,$$
$$\lambda_{\max}(Q^{(\mathtt{P})}) = \max\left\{\frac{1}{p_1}, \frac{r_0}{1 - p_1}\right\}, \quad \lambda_{\max}(Q^{(\mathtt{U})}) = r_0 + 1 + (1 - r_0)\left|\cos(2\omega)\right|.$$

The smallest operator norm for $Q^{(\mathtt{P})}$ is given by $p_1 = \frac{1}{1+r_0}$. When $r_0 \leq 1$, and $0 \leq \omega \leq \pi/6$, we have $\lambda_{\max}(Q^{(\mathtt{I})}) \geq \lambda_{\max}(Q^{(\mathtt{U})}) \geq \lambda_{\max}(Q^{(\mathtt{S})}) \geq \lambda_{\max}(Q^{(\mathtt{P})})$. When $r_0 \geq 1$, and $0 \leq \omega \leq \pi/6$, we have $\lambda_{\max}(Q^{(\mathtt{P})}) \geq \lambda_{\max}(Q^{(\mathtt{S})}) \geq \lambda_{\max}(Q^{(\mathtt{U})}) \geq \lambda_{\max}(Q^{(\mathtt{I})})$. For other choices of ω, we can obtain a comparison analogously.

In general, it is natural to use Loewner order to compare two positive semi-definite matrix $A, B \in \mathbb{R}^{d\times d}$, i.e., $A \succeq B$ if $\boldsymbol{x}^\top A\boldsymbol{x} \geq \boldsymbol{x}^\top B\boldsymbol{x}$ for any $\boldsymbol{x} \in \mathbb{R}^d$. It is equivalent to say, for any positive constant $c > 0$, the ellipsoid $\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x}^\top A\boldsymbol{x} \leq c\}$ contains the ellipsoid $\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x}^\top B\boldsymbol{x} \leq c\}$. To better illustrate the result, we consider the 2-dimensional case where $S = \mathrm{diag}(1, 1/2)$ and plot the ellipse $\{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{x}^\top Q^{(\cdot)}\boldsymbol{x} = 2\}$. In Figure A.1, we compare $Q^{(\mathtt{S})}$, $Q^{(\mathtt{I})}$ (as a special case of $Q^{(\mathtt{U})}$

with $\theta = 0$), $Q^{(\mathtt{U})}$ with $\theta = \frac{\pi}{6}$, and $Q^{(\mathtt{P})}$ with $p_1 = \frac{1}{1+r_0} = \frac{2}{3}$. As can be inferred from the plot, none of the ellipsoids contain any other ellipsoids.

As shown in this illustrative example, there is no unique optimal direction distribution, and a practitioner might choose a search direction based on her favorable optimality criterion.

Lastly, in the following Remark A.5, we show that, if the optimality criterion degenerates to one dimension, one may utilize the non-uniform distribution (P) to obtain a smaller limiting variance. In particular, consider the application where we are only interested in the first coordinate of $\boldsymbol{\theta}^\star$, in which cases the optimality criterion of the limiting variance is on $\theta_1^\star$. We will show that the (AKW) estimator with the non-uniform distribution (P) achieves the Cramér-Rao lower bound.

Remark A.5. *Assume the population loss function $F(\cdot)$ has Hessian $H = I_d$. Considering a non-uniform sampling* (P) *from $\{\boldsymbol{e}_k\}_{k=1}^d$ for the direction distribution $\mathcal{P}_{\boldsymbol{v}}$. We choose $\boldsymbol{v} = \boldsymbol{e}_k$ with probability p_k for $k = 1, 2, \ldots, d$, where $p_1 = 1 - p$ for some constant $p \in (0, 1]$ and $p_k = p/(d-1)$ for $k \neq 1$. Define i.i.d. random variables k_n where $k_n = 1$ with probability $1 - p$ and $k_n = 2, \ldots, d$ uniformly with probability $p/(d-1)$. The gradient estimator is defined by,*

$$\widehat{g}(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n) = \frac{f(\boldsymbol{\theta}_{n-1} + h_n \boldsymbol{e}_{k_n}; \boldsymbol{\zeta}_n) - f(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n)}{h_n p_n} \boldsymbol{e}_{k_n},$$

where $p_n = 1 - p$ if $k_n = 1$, $p_n = p/(d-1)$ for $k_n > 1$. By the same argument as Proposition 3.5, the variance for $\overline{\boldsymbol{\theta}}_n$ in the direction $\boldsymbol{e}_1$ is,

$$n\mathrm{Var}\left(\boldsymbol{e}_1^\top (\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star)\right) = \frac{S_{11}}{1-p}.$$

As $p \to 0$, we approximately obtain the optimal variance given by Cramér-Rao lower bound in the direction $\boldsymbol{e}_1$. However, in order to approach the optimal variance in the direction $\boldsymbol{e}_1$, we increase the magnitude of variance in all other directions, where the variance in other directions is given by $n\mathrm{Var}\left(\boldsymbol{e}_k^\top (\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star)\right) = (d-1)S_{kk}/p$ for $k = 2, \ldots, d$.

A.4 Multi-query approximation

Proof of Theorem 3.6

Proof. The convergence result can be obtained as in the two function evaluation case. The only difference is the following calculation:

$$\mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{v}_i\boldsymbol{v}_i^\top\right)S\left(\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{v}_i\boldsymbol{v}_i^\top\right)=\frac{1}{m}\mathbb{E}\boldsymbol{v}\boldsymbol{v}^\top S\boldsymbol{v}\boldsymbol{v}^\top+\frac{m-1}{m}S,$$

which implies the desired result. □

Proof of Theorem 3.7

Proof. It is clear that $Q_m = S$ for $m = d$. We need to compute the quantity

$$Q_m=\frac{d^2}{m^2}\mathbb{E}\Big(\sum_{i=1}^{m}\boldsymbol{v}_i\boldsymbol{v}_i^\top\Big)S\Big(\sum_{i=1}^{m}\boldsymbol{v}_i\boldsymbol{v}_i^\top\Big),$$

which can be simplifies to

$$Q_m=\frac{d^2}{m^2}\mathbb{E}\Big(\sum_{i=1}^{m}\boldsymbol{v}_i\boldsymbol{v}_i^\top S\boldsymbol{v}_i\boldsymbol{v}_i^\top\Big)+\frac{d^2}{m^2}\mathbb{E}\Big(\sum_{i\neq j}\boldsymbol{v}_i\boldsymbol{v}_i^\top S\boldsymbol{v}_j\boldsymbol{v}_j^\top\Big).$$

By symmetry, it equals to

$$Q_m=\frac{d^2}{m}\mathbb{E}\boldsymbol{v}_1\boldsymbol{v}_1^\top S\boldsymbol{v}_1\boldsymbol{v}_1^\top+\frac{d^2(m-1)}{m}\mathbb{E}\boldsymbol{v}_1\boldsymbol{v}_1^\top S\boldsymbol{v}_2\boldsymbol{v}_2^\top.$$

We know $\mathbb{E}\boldsymbol{v}_1\boldsymbol{v}_1^\top S\boldsymbol{v}_1\boldsymbol{v}_1^\top=\frac{1}{d^2}Q$ and $Q_d = S$. So we can solve for $\mathbb{E}\boldsymbol{v}_1\boldsymbol{v}_1^\top S\boldsymbol{v}_2\boldsymbol{v}_2^\top$ and get

$$\mathbb{E}\boldsymbol{v}_1\boldsymbol{v}_1^\top S\boldsymbol{v}_2\boldsymbol{v}_2^\top=\frac{1}{d(d-1)}(\frac{1}{d}Q-\mathrm{diag}(S)).$$

Therefore,

$$\begin{aligned}Q_m&=\frac{1}{m}Q+\frac{d(m-1)}{m(d-1)}(\frac{1}{d}Q-\mathrm{diag}S)\\&=\frac{d-m}{m(d-1)}Q+\frac{d(m-1)}{m(d-1)}S.\end{aligned}$$

□

B Proofs of Results in Section 4

B.1 Proof of Lemma 4.1

Before we come to the proof of the Hessian estimator (17) in Lemma 4.1, we first introduce a naive method to estimate Hessian matrix H which we omit in the main text.

Inspired by the previous gradient estimator, we can estimate the Hessian matrix H by the following

$$\widehat{G}_n = \frac{1}{mh_n^2}\sum_{j=1}^{m}\left[\Delta_{h_n \boldsymbol{v}_n^{(j)}} f(\boldsymbol{\theta}_{n-1} + h_n \boldsymbol{u}_n^{(j)}; \boldsymbol{\zeta}_n) - \Delta_{h_n \boldsymbol{v}_n^{(j)}} f(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n)\right] \boldsymbol{u}_n^{(j)} \boldsymbol{v}_n^{(j)\top},$$

where $\{\boldsymbol{u}_n^{(j)}\}_{j=1}^m$ and $\{\boldsymbol{v}_n^{(j)}\}_{j=1}^m$ are *i.i.d.* random vectors and $m > 0$ is a parameter (which might be different from m in the previous section). Therefore, our naive Hessian estimator is,

$$\widetilde{H}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{G}_i + \widehat{G}_i^\top}{2}. \tag{B.1}$$

where the $(\widehat{G}_i + \widehat{G}_i^\top)/2$ term ensures the symmetry of $\widetilde{H}_n$. The function query complexity is $\mathcal{O}(m)$ per step for this Hessian estimation.

Now we restate our Lemma 4.1 for the both estimators (B.1) and (17).

Lemma B.1. *Under the assumptions in Theorem 3.3, we have the following result for the Hessian estimator* (B.1)*,*

$$\mathbb{E}\|\widetilde{H}_n - H\|^2 \le C_1 n^{-\alpha} + C_2\left(1 + \frac{1}{m}\right) n^{-1}. \tag{B.2}$$

The Hessian estimator (17) *satisfies,*

$$\mathbb{E}\|\widetilde{H}_n - H\|^2 \le C_1 n^{-\alpha} + C_2 p^{-1} n^{-1}. \tag{B.3}$$

Proof. In the case of naive Hessian estimator (B.1), we decompose $\widetilde{H}_n - H$ as follows,

$$\begin{aligned}
\widetilde{H}_n - H &= \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{G}_i + \widehat{G}_n^\top}{2} - H \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\widehat{G}_i + \widehat{G}_i^\top}{2} - \left(\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{u}_i^{(j)} \boldsymbol{u}_i^{(j)\top} \right) \nabla^2 f(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n) \left(\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{v}_i^{(j)} \boldsymbol{v}_i^{(j)\top} \right) \right) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \left(\left(\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{u}_i^{(j)} \boldsymbol{u}_i^{(j)\top} \right) \nabla^2 f(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n) \left(\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{v}_i^{(j)} \boldsymbol{v}_i^{(j)\top} \right) - \nabla^2 f(\boldsymbol{\theta}_{i-1}; \boldsymbol{\zeta}_i) \right) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \left[\nabla^2 f(\boldsymbol{\theta}_{i-1}; \boldsymbol{\zeta}_i) - \nabla^2 f(\mathbf{0}; \boldsymbol{\zeta}_i) \right] + \frac{1}{n}\sum_{i=1}^{n} \left(\nabla^2 f(\mathbf{0}; \boldsymbol{\zeta}_i) - H \right). \qquad \text{(B.4)}
\end{aligned}$$

For the first term in the decomposition (B.4),

$$\begin{aligned}
&\mathbb{E}_{n-1} \Bigg[\Big\| \frac{1}{h_n^2} \big[f(\boldsymbol{\theta}_{n-1} + h_n \boldsymbol{u} + h_n \boldsymbol{v}; \boldsymbol{\zeta}_n) - f(\boldsymbol{\theta}_{n-1} + h_n \boldsymbol{u}; \boldsymbol{\zeta}_n) - f(\boldsymbol{\theta}_{n-1} + h_n \boldsymbol{v}; \boldsymbol{\zeta}_n) \\
&\quad + f(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n) \big] \boldsymbol{u}\boldsymbol{v}^\top - \boldsymbol{u}\boldsymbol{u}^\top \nabla^2 f(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n) \boldsymbol{v}\boldsymbol{v}^\top \Big\|^2 \Big| \boldsymbol{u}, \boldsymbol{v} \Bigg] \\
&\le \mathbb{E}_{n-1} \left[\left\| \frac{1}{h_n^2} \boldsymbol{u}\boldsymbol{u}^\top \int_0^{h_n} \int_0^{h_n} \nabla^2 f(\boldsymbol{\theta}_{n-1} + s_1 \boldsymbol{u} + s_2 \boldsymbol{v}; \boldsymbol{\zeta}_n) - \nabla^2 f(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n) \, \mathrm{d}s_1 \, \mathrm{d}s_2 \boldsymbol{v}\boldsymbol{v}^\top \right\|^2 \Big| \boldsymbol{u}, \boldsymbol{v} \right] \\
&\le \frac{1}{h_n^2} \|\boldsymbol{u}\|^2 \|\boldsymbol{v}\|^2 \int_0^{h_n} \int_0^{h_n} \mathbb{E}_{n-1} \left[\left\| \nabla^2 f(\boldsymbol{\theta}_{n-1} + s_1 \boldsymbol{u} + s_2 \boldsymbol{v}; \boldsymbol{\zeta}_n) - \nabla^2 f(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n) \right\|^2 \Big| \boldsymbol{u}, \boldsymbol{v} \right] \mathrm{d}s_1 \, \mathrm{d}s_2 \\
&\le \frac{C}{h_n^2} \|\boldsymbol{u}\|^2 \|\boldsymbol{v}\|^2 \int_0^{h_n} \int_0^{h_n} \|s_1 \boldsymbol{u} + s_2 \boldsymbol{v}\|^2 \, \mathrm{d}s_1 \, \mathrm{d}s_2 \le C h_n^2 \|\boldsymbol{u}\|^2 \|\boldsymbol{v}\|^2 (\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2).
\end{aligned}$$

The above derivation implies that

$$\mathbb{E} \| \widehat{G}_n - \left(\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{u}_i^{(j)} \boldsymbol{u}_i^{(j)\top} \right) \nabla^2 f(\boldsymbol{\theta}_{n-1}; \boldsymbol{\zeta}_n) \left(\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{v}_i^{(j)} \boldsymbol{v}_i^{(j)\top} \right) \| \le C h_n^2.$$

Therefore, we can show that

$$\begin{aligned}
&\mathbb{E} \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\widehat{G}_i + \widehat{G}_i^\top}{2} - \left(\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{u}_i^{(j)} \boldsymbol{u}_i^{(j)\top} \right) \nabla^2 f(\boldsymbol{\theta}_{i-1}; \boldsymbol{\zeta}_i) \left(\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{v}_i^{(j)} \boldsymbol{v}_i^{(j)\top} \right) \right) \right\|^2 \\
&\le \mathbb{E} \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\widehat{G}_i - \left(\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{u}_i^{(j)} \boldsymbol{u}_i^{(j)\top} \right) \nabla^2 f(\boldsymbol{\theta}_{i-1}; \boldsymbol{\zeta}_i) \left(\frac{1}{m}\sum_{j=1}^{m} \boldsymbol{v}_i^{(j)} \boldsymbol{v}_i^{(j)\top} \right) \right) \right\|^2 \\
&\le C \frac{1}{n}\sum_{i=1}^{n} h_i^2 \le C n^{-2\gamma}, \qquad \text{(B.5)}
\end{aligned}$$

where in the first inequality, we use the fact that, $\widehat{G}_i$ and $\widehat{G}_i^\top$ has the same distribution.

For the second term, notice that

$$\begin{aligned}
&\mathbb{E}_{n-1}\left\|\left(\frac{1}{m}\sum_{j=1}^{m}\boldsymbol{u}_j\boldsymbol{u}_j^\top\right)\nabla^2 f(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n)\left(\frac{1}{m}\sum_{j=1}^{m}\boldsymbol{v}_j\boldsymbol{v}_j^\top\right)-\nabla^2 f(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n)\right\|^2\\
\le\ &\mathbb{E}_{n-1}\left\|\frac{1}{m}\boldsymbol{u}_i\boldsymbol{u}_i^\top-I_d\right\|^2\left\|\nabla^2 f(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n)\right\|^2\left\|\frac{1}{m}\boldsymbol{v}\boldsymbol{v}^\top-I_d\right\|^2\\
&+\mathbb{E}_{n-1}\left\|\frac{1}{m}\boldsymbol{u}_i\boldsymbol{u}_i^\top-I_d\right\|^2\left\|\nabla^2 f(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n)\right\|^2+\mathbb{E}_{n-1}\left\|\nabla^2 f(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n)\right\|^2\left\|\frac{1}{m}\boldsymbol{v}\boldsymbol{v}^\top-I_d\right\|^2\\
\le\ &\frac{C}{m}\left(1+\|\boldsymbol{\theta}_{n-1}\|^2\right).
\end{aligned}$$

Furthermore, the second term is a sum of martingale difference sequence and we have

$$\begin{aligned}
&\mathbb{E}\|\frac{1}{n}\sum_{i=1}^{n}\left(\left(\frac{1}{m}\sum_{j=1}^{m}\boldsymbol{u}_i^{(j)}\boldsymbol{u}_i^{(j)\top}\right)\nabla^2 f(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n)\left(\frac{1}{m}\sum_{j=1}^{m}\boldsymbol{v}_i^{(j)}\boldsymbol{v}_i^{(j)\top}\right)-\nabla^2 f(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)\right)\|^2\\
=\ &\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\left(\left(\frac{1}{m}\sum_{j=1}^{m}\boldsymbol{u}_i^{(j)}\boldsymbol{u}_i^{(j)\top}\right)\nabla^2 f(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n)\left(\frac{1}{m}\sum_{j=1}^{m}\boldsymbol{v}_i^{(j)}\boldsymbol{v}_i^{(j)\top}\right)-\nabla^2 f(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)\right)\|^2\\
\le\ &C\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{m}\left(1+\mathbb{E}\|\boldsymbol{\theta}_{n-1}\|^2\right)\le C\frac{1}{mn}.
\end{aligned}\tag{B.6}$$

For the third term in (B.4), we have

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)-\nabla^2 f(\mathbf{0};\boldsymbol{\zeta}_i)\right\|^2&\le\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla^2 f(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)-\nabla^2 f(\mathbf{0};\boldsymbol{\zeta}_i)\right\|^2\\
&\le\frac{C}{n}\sum_{i=1}^{n}\mathbb{E}\|\boldsymbol{\theta}_i\|^2\le Cn^{-\alpha}.
\end{aligned}\tag{B.7}$$

For the final term, we have

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f(\mathbf{0};\boldsymbol{\zeta}_i)-H\right\|^2&\le\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla^2 f(\mathbf{0};\boldsymbol{\zeta}_i)-H\right\|^2\\
&\le\frac{C}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla^2 f(\mathbf{0};\boldsymbol{\zeta}_i)^2-H^2\right\|\le Cn^{-1},
\end{aligned}\tag{B.8}$$

where the second inequality is due to the fact that it is an equality in Frobenius norm.

Combine the previous estimates (B.5), (B.6), (B.7) and (B.8), our naive Hessian estimator satisfies,

$$\mathbb{E}\left\|\widetilde{H}_n-H\right\|^2\le Cn^{-\alpha}+C(1+\frac{1}{m})n^{-1}.$$

Similarly, for the Hessian estimator (17), we have the following decomposition,

$$\begin{aligned}\widetilde{H}_n - H =& \frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{G}_i + \widetilde{G}_i^\top}{2} - H \\ =& \frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{G}_i + \widetilde{G}_i^\top}{2} - \frac{\widehat{G}_i + \widehat{G}_i^\top}{2} + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\widehat{G}_i + \widehat{G}_i^\top}{2} - \nabla^2 f(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)\right) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\left[\nabla^2 f(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i) - \nabla^2 f(\mathbf{0};\boldsymbol{\zeta}_i)\right] + \frac{1}{n}\sum_{i=1}^{n}\nabla^2 f(\mathbf{0};\boldsymbol{\zeta}_i) - H. \end{aligned} \tag{B.9}$$

Given $\widehat{G}_n$, our Bernoulli sampling Hessian estimator $\widetilde{G}_n$ satisfies,

$$\begin{aligned}\mathbb{E}\left\|\widetilde{G}_n - \widehat{G}_n\right\|_{\text{Fro}}^2 &= \mathbb{E}\left[\sum_{j=1}^{d}\sum_{k=1}^{d}\frac{1}{p}\left(\widehat{G}_n^{(jk)}B_n^{(jk)} - \widehat{G}_n^{(jk)}\right)^2\right] \\ &= \sum_{j=1}^{d}\sum_{k=1}^{d}\mathbb{E}\left(\frac{1}{p}B_n^{(jk)} - 1\right)^2\left(\widehat{G}_n^{(jk)}\right)^2 \\ &= \frac{1-p}{p}\sum_{j=1}^{d}\sum_{k=1}^{d}\mathbb{E}\left(\widehat{G}_n^{(jk)}\right)^2 = \frac{1-p}{p}\|\widehat{G}_n\|_{\text{Fro}}^2,\end{aligned}$$

where the entries of B_n are *i.i.d.* and follow a Bernoulli distribution, i.e., $B_n^{(k\ell)} \sim \text{Bernoulli}(p)$, for some fixed $p \in (0,1)$. Here the second equality uses the fact that $B_i^{(jk)}$ are independent from each other. Therefore,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{G}_i - \widehat{G}_i\right\|^2 \le \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{G}_i - \widehat{G}_i\right\|_{\text{Fro}}^2 \le C\frac{1-p}{p}n^{-2}\sum_{i=1}^{n}\mathbb{E}\left\|\widehat{G}_i\right\|^2.$$

With $1/t\sum_{i=1}^{n}\mathbb{E}\|\widehat{G}_i\|^2 \le C + Cn^{-\alpha}$, the first term in decomposition (B.9) satisfies,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{G}_i - \widehat{G}_i\right\|^2 \le C\frac{1-p}{p}n^{-1}. \tag{B.10}$$

Other terms can be bounded similarly as in the first case:

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{G}_i + \widehat{G}_i^\top}{2} - \nabla^2 f(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i)\right\|^2 \le Cn^{-2\gamma}, \tag{B.11}$$

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f(\boldsymbol{\theta}_{i-1};\boldsymbol{\zeta}_i) - \nabla^2 f(\mathbf{0};\boldsymbol{\zeta}_i)\right\|^2 \le Cn^{-\alpha}, \tag{B.12}$$

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla^2 f(\mathbf{0};\boldsymbol{\zeta}_i) - H\right\|^2 \le Cn^{-1}. \tag{B.13}$$

Combine inequality (B.10), (B.11), (B.12) and (B.13), we obtain the desired result for Hessian estimator (17). □

Proof of Theorem 4.3

To prove Theorem 4.3, we first present the following lemma on the error rate of $\widehat{Q}_n$.

Lemma B.2. *Under conditions in Theorem 4.3, our online Gram matrix estimate $\widehat{Q}_n$ has the following convergence rate,*

$$\mathbb{E}\|\widehat{Q}_n - Q\| \leq Cn^{-\alpha/2}.$$

Proof. Recall the update rule,

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} - \eta_n \nabla F(\boldsymbol{\theta}_{n-1}) + \eta_n(\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n),$$

and our Gram matrix estimate $\widehat{Q}_n$ is,

$$\widehat{Q}_n = \frac{1}{n}\sum_{i=1}^{n}(\nabla F(\boldsymbol{\theta}_{i-1}) - \boldsymbol{\xi}_i - \boldsymbol{\gamma}_i - \boldsymbol{\varepsilon}_i)(\nabla F(\boldsymbol{\theta}_{i-1}) - \boldsymbol{\xi}_i - \boldsymbol{\gamma}_i - \boldsymbol{\varepsilon}_i)^\top.$$

It can be seen that we have the following estimates,

$$\begin{aligned}
\mathbb{E}_{n-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F(\boldsymbol{\theta}_{i-1})\nabla F(\boldsymbol{\theta}_{i-1})^\top\right\| &\leq C\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{n-1}\|\boldsymbol{\theta}_{i-1}\|^2 \leq Cn^{-\alpha},\\
\mathbb{E}_{n-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right\| &\leq C\frac{1}{n}\sum_{i=1}^{n}h_n^2 \leq Cn^{-2\gamma},\\
\mathbb{E}_{n-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^\top\right\| &\leq C\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}_{n-1}\|\boldsymbol{\theta}_{i-1}\|^2 + h_n^2) \leq Cn^{-\alpha},\\
\mathbb{E}_{n-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top\right\| &\leq C\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}_{n-1}\|\boldsymbol{\theta}_{i-1}\|^2 + h_n^2 + 1) \leq C.
\end{aligned}$$

The crossing terms between them can be bounded by Cauchy-Schwarz inequality. Therefore, we can find that all terms in $\widehat{Q}_n$ except $\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top/t$ can be bounded by $Cn^{-\alpha/2}$. So it suffices to prove,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top - Q\right\| \leq Cn^{-\alpha/2}. \tag{B.14}$$

Define a new sequence $z_n := \boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^\top - \mathbb{E}_{n-1}\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^\top$. Then z_n is a martingale difference sequence and we have

$$\begin{aligned}
\left\|\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^\top - Q\right\| &\leq \|z_n\| + \left\|\mathbb{E}_{n-1}\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^\top - Q\right\|\\
&\leq \|z_n\| + C\left(\|\boldsymbol{\theta}_{n-1}\| + \|\boldsymbol{\theta}_{n-1}\|^4 + h_n + h_n^4\right),
\end{aligned}$$

where the last inequality leverages inequality (A.19). Now we have,

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\top - Q\right\| &\le \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} z_i\right\| + C\mathbb{E}\left(\|\boldsymbol{\theta}_{n-1}\| + \|\boldsymbol{\theta}_{n-1}\|^4 + h_n + h_n^4\right) \\ &\le \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} z_i\right\| + Cn^{-\alpha/2}.\end{aligned}$$

Thus we turn the proof of (B.14) into,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} z_i\right\| \le Cn^{-1/2}. \tag{B.15}$$

By Hölder's inequality, it can be derived that,

$$\mathbb{E}_{n-1}\|z_n\|^2 \le \mathbb{E}_{n-1}\|\boldsymbol{\varepsilon}_n\|^4 \le C(\|\boldsymbol{\theta}_{n-1}\|^4 + h_n^4 + 1).$$

Combine Lemma B.3 with Lemma 3.1, we have

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} z_i\right\|^2 \le \frac{1}{n^2}\sum_{i=1}^{n} C\mathbb{E}\left(\|\boldsymbol{\theta}_{i-1}\|^4 + h_i^4 + 1\right) \le Cn^{-1}.$$

Therefore, condition (B.15) is satisfied through Jensen's inequality. □

We now come back to the main proof of Theorem 4.3.

Theorem 4.3. *Assume Assumptions 1 to 4 hold for $\delta = 2$. Set the step size as $\eta_n = \eta_0 n^{-\alpha}$ for some constant $\eta_0 > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$, and the spacing parameter as $h_n = h_0 n^{-\gamma}$ for some constant $h_0 > 0$, and $\gamma \in \left(\frac{1}{2}, 1\right)$. We have*

$$\mathbb{E}\left\|\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1} - H^{-1}QH^{-1}\right\| \le Cn^{-\alpha/2}.$$

Proof. For the thresholding estimator $\widehat{H}_n$, since $\|\widehat{H}_n - \widetilde{H}_n\| \le \|\widetilde{H}_n - H\|$ by construction, it is consistent with the rate below,

$$\mathbb{E}\|\widehat{H}_n - H\|^2 \le 2\mathbb{E}\|\widetilde{H}_n - H\|^2 + 2\mathbb{E}\|\widehat{H}_n - \widetilde{H}_n\|^2 \le 4\mathbb{E}\|\widetilde{H}_n - H\|^2 \le Cn^{-\alpha}, \tag{B.16}$$

where the last inequality from Lemma 4.1.

By Lemma B.4, the inverse matrix error satisfies,

$$
\begin{aligned}
&\mathbb{E}\|\widehat{H}_n^{-1} - H^{-1}\|^2 \\
\leq\ & \mathbb{E}\left[\mathbf{1}_{\|H^{-1}(\widehat{H}_n - H)\| \leq 1/2} 2\|\widehat{H}_n - H\|\|H^{-1}\|^2 + \mathbf{1}_{\|H^{-1}(\widehat{H}_n - H)\| \geq 1/2}\|\widehat{H}_n^{-1} - H^{-1}\|\right]^2 \\
\leq\ & 8\|H^{-1}\|^4\mathbb{E}\|\widehat{H}_n - H\|^2 + 2(\kappa_1^{-1} + \lambda_{\min}^{-1}(H))^2\mathbb{P}\left(\|H^{-1}(\widehat{H}_n - H)\| \geq \frac{1}{2}\right) \\
\leq\ & 8\|H^{-1}\|^4\mathbb{E}\|\widehat{H}_n - H\|^2 + \frac{1}{2\lambda^2}(\kappa_1^{-1} + \lambda_{\min}^{-1}(H))^2\mathbb{E}\|\widehat{H}_n - H\|^2 \\
\leq\ & C\, n^{-\alpha}, \qquad \text{(B.17)}
\end{aligned}
$$

where the third inequality follows from Markov's inequality and the last one from (B.16).

We now consider our target term, with our previous results (B.16), (B.17), and Lemma B.2, we can obtain that,

$$
\begin{aligned}
&\mathbb{E}\left\|\widehat{H}_n^{-1}\widehat{Q}_n\widehat{H}_n^{-1} - H^{-1}QH^{-1}\right\| \\
=\ & \mathbb{E}\left\|\widehat{H}_n^{-1}(\widehat{Q}_n - Q)\widehat{H}_n^{-1} + (H^{-1} + \widehat{H}_n^{-1} - H^{-1})Q(H^{-1} + \widehat{H}_n^{-1} - H^{-1}) - H^{-1}QH^{-1}\right\| \\
\leq\ & \mathbb{E}\left\|\widehat{H}_n^{-1}(\widehat{Q}_n - Q)\widehat{H}_n^{-1}\right\| + \mathbb{E}\left\|H^{-1}Q(\widehat{H}_n^{-1} - H^{-1})\right\| + \mathbb{E}\left\|(\widehat{H}_n^{-1} - H^{-1})QH^{-1}\right\| \\
&+\mathbb{E}\left\|(\widehat{H}_n^{-1} - H^{-1})Q(\widehat{H}_n^{-1} - H^{-1})\right\| \\
\leq\ & \kappa_1^{-2}\mathbb{E}\left\|\widehat{Q}_n - Q\right\| + 2\lambda^{-1}\|Q\|\mathbb{E}\left\|\widehat{H}_n^{-1} - H^{-1}\right\| + \|Q\|\mathbb{E}\left\|\widehat{H}_n^{-1} - H^{-1}\right\|^2 \\
\leq\ & Cn^{-\alpha/2},
\end{aligned}
$$

which completes the proof. □

Proof of Theorem 4.4

Proof. We first show that we can extend our result in Theorem 3.3 to the following form,

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr \rfloor}\boldsymbol{\theta}_i \Longrightarrow \Sigma^{1/2}\boldsymbol{W}_r, \quad r \in [0,1].
$$

where $\Sigma = H^{-1}QH^{-1}$ and $\boldsymbol{W}_r$ is a d-dimensional vector of independent standard Brownian motions on $[0,1]$. For any $r \in [0,1]$, we consider the following partial summation process,

$$
\overline{B}_n(r) = \frac{1}{n}\sum_{i=1}^{\lfloor nr \rfloor}\Delta_i,
$$

where $\Delta_i = \boldsymbol{\theta}_i - \boldsymbol{\theta}^\star = \boldsymbol{\theta}_i$. Now consider the following alternative partial summation process,

$$\overline{B}_n'(r) = \frac{1}{n}\sum_{i=1}^{\lfloor nr \rfloor} \Delta_i',$$

where

$$\Delta_i' = \Delta_{i-1}' - \eta_i H \Delta_{i-1}' + \eta_n(\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n), \quad \Delta_0' = \Delta_0 = \boldsymbol{\theta}_0.$$

From Theorem 2 in Polyak and Juditsky (1992), we know that $\sqrt{n}\sup_r |\overline{B}_n'(r) - \overline{B}_n(r)| = o_p(1)$. Now we consider the weak convergence of $\overline{B}_n'(r)$ instead. Using the decomposition below,

$$\sqrt{n}\overline{B}_n'(r) = \frac{1}{\sqrt{n}\lfloor nr \rfloor \eta_{\lfloor nr \rfloor}}\boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr \rfloor} H^{-1}(\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n) + \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr \rfloor} w_i^{\lfloor nr \rfloor}(\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n),$$

where $1/\sqrt{n}\sum_{i=1}^n \|w_i^n\| \to 0$. Using the result from Lemma 3.1, the first and the third terms on the RHS are $o_p(1)$. Combining Theorem 4.2 from Hall and Heyde (1980) and Equation (A.16), we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr \rfloor} H^{-1}(\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n) \Longrightarrow \Sigma^{1/2}\boldsymbol{W}_r.$$

Therefore, for any $\boldsymbol{w} \in \mathbb{R}^d$, we have

$$C_n(r) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr \rfloor} \boldsymbol{w}^\top \boldsymbol{\theta}_i \Rightarrow \boldsymbol{w}^\top(\boldsymbol{w}^\top \Sigma \boldsymbol{w})^{1/2} W_r, \quad r \in [0,1].$$

Here W_r is the standard one dimensional Brownian motion. In addition,

$$\boldsymbol{w}^\top V_n \boldsymbol{w} = \frac{1}{n}\sum_{i=1}^n \left[C_n\left(\frac{i}{n}\right) - \frac{i}{n}C_n(1)\right]\left[C_n\left(\frac{i}{n}\right) - \frac{i}{n}C_n(1)\right]^\top.$$

Notice that $\boldsymbol{w}^\top(\overline{\boldsymbol{\theta}}_n) = \frac{1}{\sqrt{n}}C_n(1)$, and

$$n\frac{(\boldsymbol{w}^\top \overline{\boldsymbol{\theta}}_n)^2}{\boldsymbol{w}^\top V_n \boldsymbol{w}} \Rightarrow \frac{W_1^2}{\int_0^1 (W_r - rW_1)^2 \,\mathrm{d}r},$$

using the continuous mapping theorem. □

B.1.1 Technical Lemmas

The following lemma is from Assouad (1975). We include the proof here.

Lemma B.3 (Assouad (1975)). *Let $\{\boldsymbol{X}_n\}$ be a martingale difference sequence, i.e. $\mathbb{E}[\boldsymbol{X}_n|\boldsymbol{X}_{n-1}] = 0$. For any $1 \le p \le 2$ and any norm $\|\cdot\|$ on $\mathbb{R}^d$, there exists a constant C such that*

$$\mathbb{E}\left\|\sum_{i=1}^n \boldsymbol{X}_i\right\|^p \le C\sum_{i=1}^n \mathbb{E}\left[\|\boldsymbol{X}_i\|^p|\boldsymbol{X}_{i-1}\right].$$

Proof. We would like to show that there exists a constant C (which depends on d and p) such that for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$,

$$\frac{1}{2}\left(\|\boldsymbol{a}+\boldsymbol{b}\|_2^p + \|\boldsymbol{a}-\boldsymbol{b}\|_2^p\right) \le \|\boldsymbol{a}\|_2^p + C\|\boldsymbol{b}\|_2^p,$$

where $\|\cdot\|_2$ is the 2-norm. To see this, in the one dimensional case, this is equivalent to

$$\frac{1}{2}\left(|1+x|^p + |1-x|^p\right) \le 1 + C|x|^p.$$

At $x = 1$, the left hand side is differentiable and its first derivative is 0, so there exists a constant C such that the inequality holds in a neighborhood of $x = 1$. At $x \to \pm\infty$, the inequality also holds with some constant C. So it is easy to find a constant C such that the inequality holds for all x. The proof for the d-dimensional case is the same.

Using the above inequality, we have

$$\begin{aligned}\mathbb{E}_{n-1}\left\|\sum_{i=1}^n \boldsymbol{X}_i\right\|_2^p &= \mathbb{E}_{n-1}\left\|\sum_{i=1}^{n-1}\boldsymbol{X}_i + \boldsymbol{X}_n\right\|_2^p \\ &\le 2\left\|\sum_{i=1}^{n-1}\boldsymbol{X}_i\right\|_2^p + 2C\mathbb{E}_{n-1}\|\boldsymbol{X}_n\|_2^p - \mathbb{E}_{n-1}\left\|\sum_{i=1}^{n-1}\boldsymbol{X}_i - \boldsymbol{X}_n\right\|_2^p.\end{aligned}$$

On the other hand,

$$\mathbb{E}_{n-1}\left\|\sum_{i=1}^{n-1}\boldsymbol{X}_i - \boldsymbol{X}_n\right\|_2^p \ge \left\|\sum_{i=1}^{n-1}\boldsymbol{X}_i - \mathbb{E}_{n-1}\boldsymbol{X}_n\right\|_2^p = \left\|\sum_{i=1}^{n-1}\boldsymbol{X}_i\right\|_2^p.$$

So

$$\mathbb{E}_{n-1}\left\|\sum_{i=1}^n \boldsymbol{X}_i\right\|_2^p \le \left\|\sum_{i=1}^{n-1}\boldsymbol{X}_i\right\|_2^p + 2C\mathbb{E}_{n-1}\|\boldsymbol{X}_n\|_2^p.$$

By induction, we then have

$$\mathbb{E}\left\|\sum_{i=1}^n \boldsymbol{X}_i\right\|_2^p \le 2C\sum_{i=1}^n \mathbb{E}\left[\|\boldsymbol{X}_i\|_2^p|\boldsymbol{X}_{i-1}\right].$$

For any general norm, there exists a constant C such that

$$\frac{1}{C}\|X\| \leq \|X\|_2 \leq C\|X\|.$$

So the same result holds for any norm. □

We now provide a matrix perturbation inequality from Chen et al. (2021).

Lemma B.4. *If a matrix $B = A + E$ where A and B are invertible, we have,*

$$\left\|B^{-1} - A^{-1}\right\| \leq \|A^{-1}\|^2\|E\|\frac{1}{1 - \|A^{-1}E\|}.$$

Proof. Notice that

$$\begin{aligned} B^{-1} = (A+E)^{-1} &= A^{-1} - A^{-1}\left(A^{-1} + E^{-1}\right)^{-1}A^{-1} \\ &= A^{-1} - A^{-1}E\left(A^{-1}E + I\right)^{-1}A^{-1}. \end{aligned}$$

Therefore, the inversion error is,

$$\begin{aligned} \|B^{-1} - A^{-1}\| &= \left\|A^{-1}E\left(A^{-1}E + I\right)^{-1}A^{-1}\right\| \\ &\leq \|A^{-1}\|^2\|E\|\|(A^{-1}E + I)^{-1}\| \\ &\leq \|A^{-1}\|^2\|E\|\frac{1}{\lambda_{\min}(A^{-1}E + I)} \\ &\leq \|A^{-1}\|^2\|E\|\frac{1}{1 - \|A^{-1}E\|}, \end{aligned}$$

where we use Weyl's inequality in the last inequality. □

B.2 Finite-difference stochastic Newton method

As a by-product and an application, the online finite-difference estimator of Hessian in (19) enables us to develop the `(KW)` version of the stochastic Newton's method. Existing literature that handles the `(RM)` version of the stochastic Newton's method traces back to Ruppert (1985). Given an initial point $\boldsymbol{\theta}_0$, the `(KW)` stochastic Newton's method has the following updating rule,

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} - \frac{1}{n}\widehat{H}_{n-1}^{-1}\widehat{g}_{h_n,\boldsymbol{v}_n}(\boldsymbol{\theta}_{n-1};\boldsymbol{\zeta}_n), \tag{B.18}$$

Here $\widehat{H}_n^{-1}$ a recursive estimator of H^{-1}. We modify the thresholding Hessian estimator $\widehat{H}_n$ in (19) as follows. Let $U\widetilde{\Lambda}_n U^\top$ be the eigenvalue decomposition of $\widetilde{H}_n$ in (17), and define

$$\widehat{H}_n = U\widehat{\Lambda}_n U^\top, \quad \widehat{\Lambda}_{n,kk} = \max\left\{\kappa_1, \min\left\{\kappa_2, \widetilde{\Lambda}_{n,kk}\right\}\right\}, \ k = 1, 2, \ldots, d, \tag{B.19}$$

for some constants $0 < \kappa_1 < \lambda < L_f < \kappa_2$, where λ, L_f are defined in Assumption 1.

Theorem B.5. *Under the assumptions in Theorem 3.3, the Hessian estimator $\widehat{H}_n$ in (B.19) converges in probability to the empirical Hessian matrix H. The stochastic Newton estimator $\boldsymbol{\theta}_n$ in (B.18) converges to $\boldsymbol{\theta}^\star$ almost surely and has the following limiting distribution,*

$$\sqrt{n}\left(\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star\right) \Longrightarrow \mathcal{N}\left(\mathbf{0}, H^{-1}QH^{-1}\right), \tag{B.20}$$

for the same Q as in Theorem 3.3.

Theorem B.5 states that the final iterate of the `(KW)` stochastic Newton method (B.18) entails the same asymptotic distribution as the averaged `(AKW)` estimator (6). In contrast to `(AKW)`, (B.18) leverages additional Hessian information to achieve the asymptotic normality and efficiency. Nevertheless, the numerical implementation of the `(KW)` stochastic Newton's method requires to update a Hessian estimator $\widehat{H}_n$ in all iterations, which demands significant additional computation unless such an estimator is yet computed and maintained along the procedure for other purposes.

Proof of Theorem B.5

Proof. Notice that

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} - \frac{1}{n}H_{n-1}^{-1}\nabla F(\boldsymbol{\theta}_{n-1}) + \frac{1}{n}\overline{H}_{n-1}^{-1}\left(\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n\right),$$

where $\boldsymbol{\xi}_n, \boldsymbol{\gamma}_n, \boldsymbol{\varepsilon}_n$ are defined at the beginning of the supplement. We now show that Lemma 3.1 holds under $\alpha = 1$. Following from the same logic in Lemma 3.1, we can show that there exists some universal constant $n_0 > 0$, such that for all $n > n_0$, and some constants C_1, C_2,

$$\mathbb{E}_{n-1}\|\boldsymbol{\theta}_n\|^2 \leq \left(1 - \frac{C_1}{n}\right)\|\boldsymbol{\theta}_{n-1}\|^2 + C_2 n^{-2}. \tag{B.21}$$

Therefore, $\boldsymbol{\theta}_n \to 0$ almost surely by martingale convergence theorem (Robbins and Monro, 1951). Now we consider the convergence rate of $\boldsymbol{\theta}_n$.

Using the proof in Lemma 3.1, we can show that

$$\mathbb{E}_{n-1}\|\boldsymbol{\theta}_n\|^2 \leq C\left(n^{-C_1/2} + n^{-1}\right). \tag{B.22}$$

Similarly,

$$\mathbb{E}_{n-1}\|\boldsymbol{\theta}_n\|^{2+\delta} \leq C\left(n^{-C_1/2} + n^{-(1+\delta)}\right). \tag{B.23}$$

Now we consider the limiting distribution.

$$\begin{aligned}\boldsymbol{\theta}_n &= \boldsymbol{\theta}_{n-1} - \frac{1}{n}H^{-1}\nabla F(\boldsymbol{\theta}_{n-1}) - \frac{1}{n}\left(H_{n-1}^{-1} - H^{-1}\right)\nabla F(\boldsymbol{\theta}_{n-1}) + \frac{1}{n}H_{n-1}^{-1}\left(\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n\right)\\ &= \left(1 - \frac{1}{n}\right)\boldsymbol{\theta}_{n-1} - \frac{1}{n}H^{-1}\boldsymbol{\delta}_n - \frac{1}{n}\left(H_{n-1}^{-1} - H^{-1}\right)\nabla F(\boldsymbol{\theta}_{n-1}) + \frac{1}{n}H_{n-1}^{-1}\left(\boldsymbol{\xi}_n + \boldsymbol{\gamma}_n + \boldsymbol{\varepsilon}_n\right),\end{aligned}$$

where $\boldsymbol{\delta}_n = \nabla F(\boldsymbol{\theta}_{n-1}) - H\boldsymbol{\theta}_{n-1}$. By induction, we can find that

$$\boldsymbol{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}H_k^{-1}\boldsymbol{\varepsilon}_{k+1} + \frac{1}{n}\sum_{k=0}^{n-1}H_k^{-1}\left(\boldsymbol{\xi}_{k+1} + \boldsymbol{\gamma}_{k+1}\right) - \frac{1}{n}H^{-1}\sum_{k=0}^{n-1}\boldsymbol{\delta}_{k+1} - \frac{1}{n}\sum_{k=0}^{n-1}\left(H_k^{-1} - H^{-1}\right)\nabla F(\boldsymbol{\theta}_k).$$

The last three terms in the RHS above all converge to zero due to Assumption 4. Now we only need to show that $\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}H_k^{-1}\boldsymbol{\varepsilon}_{k+1}$ converges to a normal distribution. Consider

$$\mathbb{E}_k\left[H_k^{-1}\boldsymbol{\varepsilon}_{k+1}\boldsymbol{\varepsilon}_{k+1}^\top H_k^{-1}\right] = H_k^{-1}\mathbb{E}_k\left[\boldsymbol{\varepsilon}_{k+1}\boldsymbol{\varepsilon}_{k+1}^\top\right]H_k^{-1},$$

recall that in (A.19) we have shown that $\mathbb{E}_k\left[\boldsymbol{\varepsilon}_{k+1}\boldsymbol{\varepsilon}_{k+1}^\top\right]$ converges almost surely to Q. Therefore, by Assumption 4, $\mathbb{E}_k\left[H_k^{-1}\boldsymbol{\varepsilon}_{k+1}\boldsymbol{\varepsilon}_{k+1}^\top H_k^{-1}\right]$ converges in probability to $H^{-1}QH^{-1}$.

Obviously, we can get the tail bound similar to (A.20) and by martingale central limit theorem (Duflo, 1997, Theorem 2.1.9),

$$\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1}H_k^{-1}\boldsymbol{\varepsilon}_{k+1} \Longrightarrow \mathcal{N}\left(0, H^{-1}QH^{-1}\right).$$

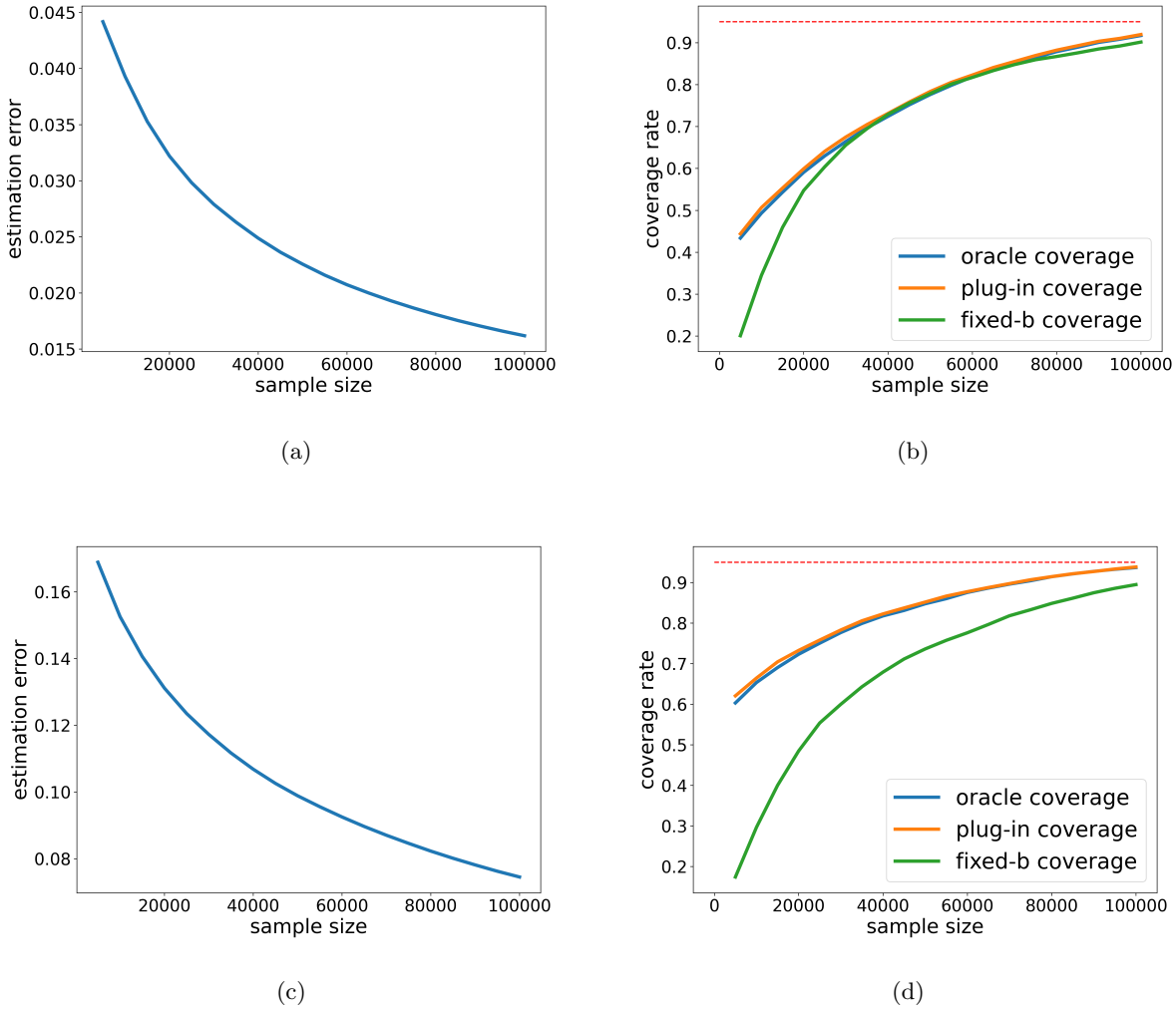
□

C Additional Results of Numerical Experiments

In this section, we present additional results of numerical experiments.

d	$\mathcal{P}_{\boldsymbol{v}}$	Estimation error		Average coverage rate		Average length	
		Parameter	Plug-in Cov	Plug-in	Oracle	Plug-in	Oracle
5	(I)	0.0299	0.0697	0.9440	0.9464	3.3620	3.2580
		(0.0131)	(0.0514)	(0.1196)	(0.1207)	(0.8587)	-
	(S)	0.0321	0.0712	0.9484	0.9468	3.2746	3.2653
		(0.0137)	(0.0507)	(0.1245)	(0.1116)	(0.8135)	-
	(G)	0.0360	0.0813	0.9508	0.9464	3.8779	3.8635
		(0.0149)	(0.0537)	(0.1196)	(0.1103)	(0.9655)	-
20	(I)	0.0799	0.1213	0.9383	0.9369	5.6873	5.6356
		(0.0146)	(0.0359)	(0.0577)	(0.0561)	(0.6775)	-
	(S)	0.0838	0.1281	0.9357	0.9347	5.4677	5.4178
		(0.0153)	(0.0382)	(0.0557)	(0.0580)	(0.6523)	-
	(G)	0.0859	0.1282	0.9379	0.9372	5.7343	5.6822
		(0.0152)	(0.0359)	(0.0548)	(0.0543)	(0.6820)	-
100	(I)	0.2867	0.7685	0.9608	0.9041	12.7375	10.4868
		(0.0253)	(0.2933)	(0.0185)	(0.0314)	(0.8942)	-
	(S)	0.2913	0.7801	0.9615	0.9032	13.1285	10.7976
		(0.0256)	(0.3115)	(0.0215)	(0.0313)	(0.9803)	-
	(G)	0.2925	0.7845	0.9618	0.9043	13.2771	10.9051
		(0.0259)	(0.3146)	(0.0191)	(0.0320)	(0.9883)	-

Table C.2: Comparison among different direction distributions $\mathcal{P}_{\boldsymbol{v}}$ (Detailed specification of `(I),(S),(G)` can be referred to Section 3.1). We consider the logistic regression model with equicorrelation covariance design, and the `(AKW)` estimators are computed based on the case of two function queries ($m = 1$). Corresponding standard errors are reported in the brackets.



(a) (b)

(c) (d)

Figure C.1: Convergence of the parameter estimation error $\|\overline{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\|$ and coverage rates v.s. the sample size n when $d = 20$ and Σ is in the equicorrelation design. Plots (a) to (b) show the cases of linear regression and plot (c) to (d) show the cases of logistic regression. Dashed lines in plots (b) and (d) correspond to the nominal 95% coverage.

d	$\mathcal{P}_{\boldsymbol{v}}$	Estimation error		Average coverage rate		Average length	
		Parameter	Plug-in Cov	Plug-in	Oracle	Plug-in	Oracle
5	`(I)`	0.0031	0.0384	0.9448	0.9436	1.7555	1.7533
		(0.0010)	(0.0106)	(0.1035)	(0.1040)	(0.0082)	-
	`(S)`	0.0030	0.0406	0.9472	0.9456	1.7556	1.7533
		(0.0009)	(0.0088)	(0.0976)	(0.0984)	(0.0075)	-
	`(G)`	0.0036	0.0623	0.9440	0.9432	2.0780	2.0745
		(0.0011)	(0.0151)	(0.1061)	(0.1087)	(0.0166)	-
20	`(I)`	0.0135	0.1126	0.9319	0.9288	3.5337	3.5065
		(0.0023)	(0.0190)	(0.0594)	(0.0616)	(0.0164)	-
	`(S)`	0.0135	0.1103	0.9306	0.9281	3.5348	3.5065
		(0.0021)	(0.0128)	(0.0575)	(0.0614)	(0.0168)	-
	`(G)`	0.0141	0.1273	0.9308	0.9283	3.7100	3.6777
		(0.0022)	(0.0180)	(0.0572)	(0.0571)	(0.0213)	-
100	`(I)`	0.0748	0.5707	0.9309	0.9012	8.6675	7.8397
		(0.0062)	(0.0648)	(0.0261)	(0.0336)	(0.1081)	-
	`(S)`	0.0750	0.5348	0.9310	0.8990	8.6814	7.8398
		(0.0059)	(0.0401)	(0.0243)	(0.0323)	(0.1001)	-
	`(G)`	0.0757	0.5548	0.9312	0.8990	8.7837	7.9178
		(0.0058)	(0.0441)	(0.0238)	(0.0321)	(0.1042)	-

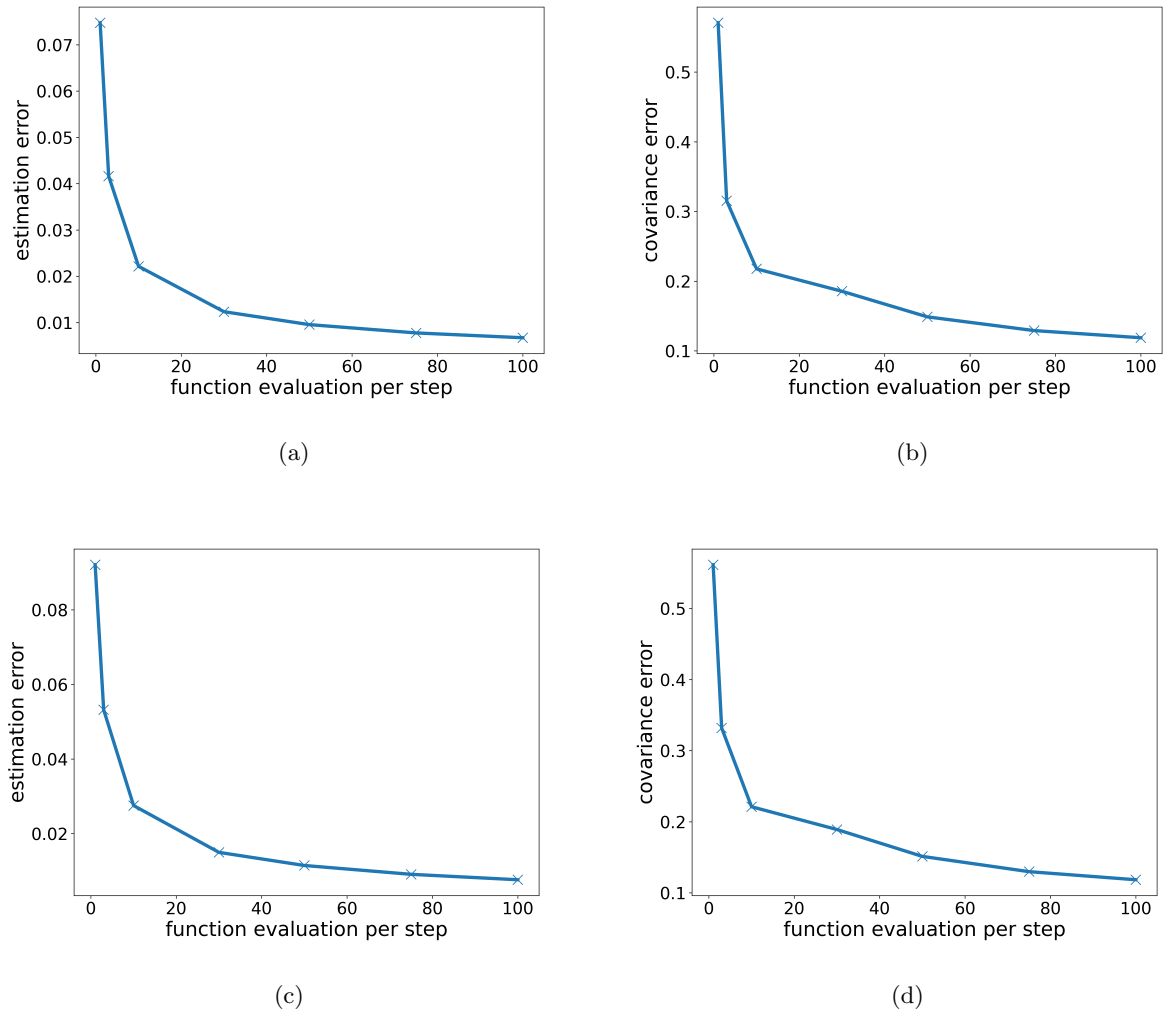
Table C.3: Comparison among different direction distributions $\mathcal{P}_{\boldsymbol{v}}$ (Detailed specification of `(I),(S),(G)` can be referred to Section 3.1). We consider the linear regression model with identity covariance design, and the `(AKW)` estimators are computed based on the case of two function queries ($m = 1$). Corresponding standard errors are reported in the brackets.

d	$\mathcal{P}_{\boldsymbol{v}}$	Estimation error		Average coverage rate		Average length	
		Parameter	Plug-in Cov	Plug-in	Oracle	Plug-in	Oracle
5	(I)	0.0035	0.0342	0.9428	0.9412	2.0109	2.0078
		(0.0012)	(0.0092)	(0.1096)	(0.1102)	(0.0097)	-
	(S)	0.0034	0.0348	0.9464	0.9456	1.9664	1.9636
		(0.0012)	(0.0082)	(0.1051)	(0.1070)	(0.0095)	-
	(G)	0.0040	0.0535	0.9464	0.9460	2.3274	2.3233
		(0.0014)	(0.0145)	(0.1117)	(0.1119)	(0.0184)	-
20	(I)	0.0172	0.1124	0.9194	0.9170	4.3140	4.2753
		(0.0029)	(0.0199)	(0.0644)	(0.0656)	(0.0207)	-
	(S)	0.0170	0.1116	0.9182	0.9165	4.2769	4.2374
		(0.0028)	(0.0126)	(0.0602)	(0.0608)	(0.0212)	-
	(G)	0.0177	0.1278	0.9216	0.9188	4.4885	4.4443
		(0.0029)	(0.0167)	(0.0598)	(0.0610)	(0.0264)	-
100	(I)	0.0921	0.5615	0.9331	0.9044	10.7701	9.7508
		(0.0076)	(0.0647)	(0.0250)	(0.0320)	(0.1400)	-
	(S)	0.0927	0.5445	0.9323	0.9000	10.7712	9.7318
		(0.0072)	(0.0487)	(0.0240)	(0.0321)	(0.1358)	-
	(G)	0.0933	0.5668	0.9336	0.9026	10.8925	9.8286
		(0.0073)	(0.0597)	(0.0243)	(0.0321)	(0.1403)	-

Table C.4: Comparison among different direction distributions $\mathcal{P}_{\boldsymbol{v}}$ (Detailed specification of `(I),(S),(G)` can be referred to Section 3.1). We consider the linear regression model with equicorrelation covariance design, and the `(AKW)` estimators are computed based on the case of two function queries ($m = 1$). Corresponding standard errors are reported in the brackets.

$m;\Sigma$	$\mathcal{P}_{\boldsymbol{v}}$	Estimation error		Average coverage rate		Average length	
		Parameter	Plug-in Cov	Plug-in	Oracle	Plug-in	Oracle
10; Identity	(I+WOR)	0.0221	0.2176	0.9274	0.9221	2.5225	2.4791
		(0.0017)	(0.0079)	(0.0265)	(0.0274)	(0.0078)	-
	(I+WR)	0.0233	0.2262	0.9251	0.9192	2.6366	2.5883
		(0.0017)	(0.0101)	(0.0259)	(0.0269)	(0.0081)	-
	(S)	0.0232	0.2257	0.9277	0.9222	2.6372	2.5883
		(0.0017)	(0.0081)	(0.0255)	(0.0270)	(0.0080)	-
10; Equicorr	(I+WOR)	0.0275	0.2210	0.9258	0.9206	3.1104	3.0554
		(0.0019)	(0.0083)	(0.0264)	(0.0270)	(0.0099)	-
	(I+WR)	0.0285	0.2291	0.9270	0.9210	3.2535	3.1926
		(0.0020)	(0.0097)	(0.0258)	(0.0273)	(0.0106)	-
	(S)	0.0285	0.2299	0.9291	0.9229	3.2487	3.1868
		(0.0019)	(0.0084)	(0.0248)	(0.0255)	(0.0107)	-
100; Identity	(I+WOR)	0.0067	0.1189	0.9405	0.9395	0.7909	0.7882
		(0.0005)	(0.0034)	(0.0287)	(0.0285)	(0.0020)	-
	(I+WR)	0.0093	0.1686	0.9407	0.9393	1.1130	1.1059
		(0.0007)	(0.0055)	(0.0240)	(0.0240)	(0.0029)	-
	(S)	0.0093	0.1683	0.9410	0.9389	1.1130	1.1059
		(0.0007)	(0.0054)	(0.0231)	(0.0232)	(0.0030)	-
100; Equicorr	(I+WOR)	0.0076	0.1183	0.9375	0.9364	0.8800	0.8770
		(0.0006)	(0.0033)	(0.0264)	(0.0270)	(0.0023)	-
	(I+WR)	0.0111	0.1727	0.9378	0.9361	1.3141	1.3054
		(0.0008)	(0.0061)	(0.0253)	(0.0257)	(0.0036)	-
	(S)	0.0110	0.1722	0.9399	0.9383	1.3126	1.3040
		(0.0008)	(0.0056)	(0.0231)	(0.0235)	(0.0036)	-

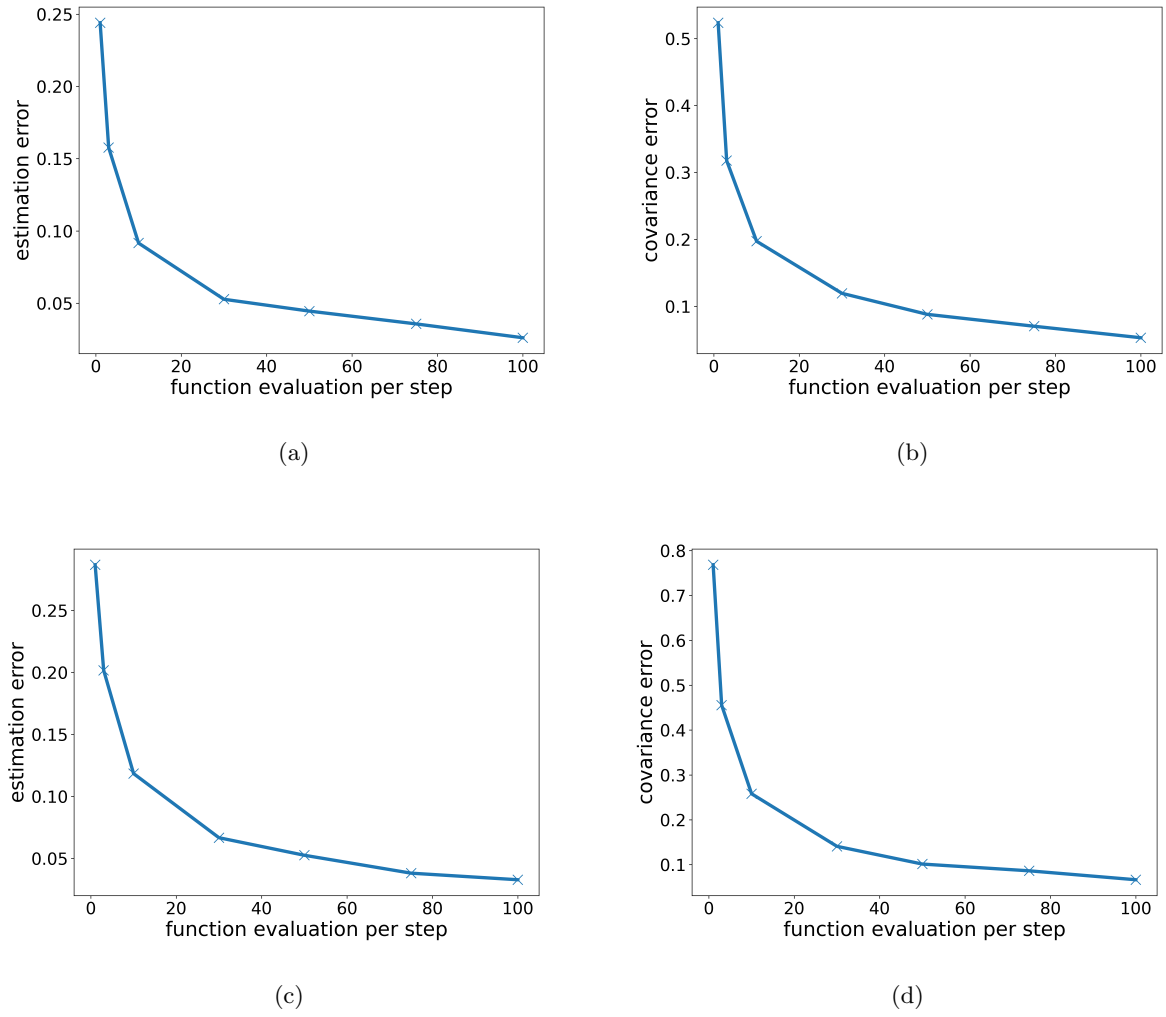
Table C.5: Comparison among different sampling schemes for multi-query algorithms under linear regression model with dimension $d = 100$ and $m = 10, 100$, respectively (Detailed specification of `(I+WOR),(I+WR),(S)` can be referred to Section 3.1). Corresponding standard errors are reported in the brackets.



(a) (b)

(c) (d)

Figure C.6: The parameter estimation error and the relative covariance estimation error (see (24)) for multiple function-value evaluations. The x-axis is the number of function evaluations per step (i.e., $m+1$). Here, we consider the linear regression model with $n = 10^5$ and $d = 100$. Plots (a) to (b) show the case of identity covariance matrix and plots (c) to (d) show the case of equicorrelation covariance matrix.



(a) (b)

(c) (d)

Figure C.7: The parameter estimation error and the relative covariance estimation error (see (24)) for multiple function-value evaluations. The x-axis is the number of function evaluations per step (i.e., $m+1$). Here, we consider the logistic regression model with $n = 10^5$ and $d = 100$. Plots (a) to (b) show the case of identity covariance matrix and plots (c) to (d) show the case of equicorrelation covariance matrix.

C.1 Numerical Experiments on Non-smooth Loss Function

In this section, we provide simulation studies to illustrate the performance of our `(AKW)` estimator as well as our inference procedure, i.e., the plug-in covariance estimator and the fixed-b HAR inference, on quantile regression. Our data is generated from a linear regression model,

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}^* + \epsilon_i,$$

where $\{\boldsymbol{\zeta}_i = (\boldsymbol{x}_i, y_i),\ i = 1, 2, \ldots, n\}$ is an *i.i.d.* sample of $\boldsymbol{\zeta} = (\boldsymbol{x}, y)$ with the covariate $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and the noise $\{\epsilon_i\}$ follows an *i.i.d.* normal distribution such that

$$\epsilon_i \sim \mathcal{N}(-\sigma\Phi^{-1}(\tau), \sigma^2), \quad \Pr\left(\epsilon_i \le 0 \mid \boldsymbol{x}_i\right) = \tau.$$

Here $\Phi(\cdot)$ is the cumulative density function of standard normal distribution and $\Phi^{-1}(\cdot)$ is its inverse function. For each quantile level $\tau \in (0, 1)$, the individual loss is $f(\boldsymbol{\theta}; \zeta) = \rho_\tau\left(y - \boldsymbol{x}^\top\boldsymbol{\theta}\right)$, where $\rho_\tau(z) = z\left(\tau - 1\{z < 0\}\right)$.

From Theorem A.4, we know that the `(AKW)` estimator of the above quantile regression model is asymptotically normal with asymptotic covariance matrix $H^{-1}QH^{-1}$ and

$$H = \frac{1}{\sigma}\phi(\Phi^{-1}(\tau))\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \frac{1}{\sigma}\phi(\Phi^{-1}(\tau))\Sigma,$$
$$S = \tau(1-\tau)\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \tau(1-\tau)\Sigma.$$

Here S is the matrix used to construct Q under different sampling direction (see Proposition 3.5). For example, if we sample uniformly from the canonical basis with two function queries ($m = 1$), then $Q^{(\mathtt{I})} = d\,\mathrm{diag}(S)$.

In the numerical experiments below, we fix sample size $n = 10^5$, dimension $d = 20$, and the noise variance $\sigma^2 = 0.2$. We consider two different covariance structure for Σ: identity matrix I_d and equicorrelation covariance matrix, i.e., $\Sigma_{k\ell} = 0.2$ for all $k \neq \ell$ and $\Sigma_{kk} = 1$. We present our results below in Table C.8 with three quantile level $\tau = 0.1, 0.5, 0.9$. As can be inferred from the table, plug-in estimators have a good coverage rate closed to the oracle ones. Our fixed-b HAR inference structure provide coverage around 90% under a much faster speed and no additional function queries condition.

τ	Σ	Estimation error		Average coverage rate			Average length		
		Parameter	Plug-in Cov.	Plug-in	Fixed-b	Oracle	Plug-in	Fixed-b	Oracle
0.1	Identity	0.0578	0.2223	0.9390	0.9040	0.9470	15.6048	18.6994	14.9837
		(0.0090)	(0.0403)	(0.0605)	(0.0714)	(0.0581)	(0.1278)	(2.5821)	-
	Equicorr	0.0726	0.2281	0.9245	0.8935	0.9405	19.0581	21.0679	18.2687
		(0.0120)	(0.0385)	(0.0634)	(0.0726)	(0.0578)	(0.1667)	(2.5188)	-
0.5	Identity	0.0387	0.0493	0.9500	0.9180	0.9495	10.9806	13.0493	10.9858
		(0.0062)	(0.0116)	(0.0510)	(0.0585)	(0.0517)	(0.0249)	(1.4260)	-
	Equicorr	0.0464	0.0511	0.9470	0.9100	0.9515	13.3991	15.5384	13.3943
		(0.0082)	(0.0110)	(0.0529)	(0.0612)	(0.0526)	(0.0320)	(1.9022)	-
0.9	Identity	0.0536	0.1993	0.9415	0.8995	0.9475	15.3729	16.4834	14.9837
		(0.0086)	(0.0422)	(0.0512)	(0.0630)	(0.0471)	(0.1179)	(1.8594)	-
	Equicorr	0.0652	0.1963	0.9460	0.9020	0.9505	18.8055	19.7132	18.2687
		(0.0105)	(0.0399)	(0.0569)	(0.0663)	(0.0502)	(0.1387)	(2.2770)	-

Table C.8: Estimation errors, averaged coverage rates, and average lengths of the proposed algorithm with search direction `(I)` and two function queries ($m = 1$), under quantile regression model. Sample size $n = 10^5$, dimension $d = 20$. Corresponding standard errors are reported in the brackets. We compare the plug-in covariance estimator (plug-in) based inference (17) and fixed-b HAR (fixed-b) based inference (22).

References

Assouad, Patrice (1975). Espaces p-lisses et q-convexes, inégalités de Burkholder. In *Séminaire Maurey-Schwartz 1974–1975: Espaces L^p, applications radonifiantes et géométrie des espaces de Banach, Exp. No. XV.*

Chen, Haoyu, Wenbin Lu, and Rui Song (2021). Statistical inference for online decision making via stochastic gradient descent. *Journal of the American Statistical Association 116*(534), 708–719.

Duflo, Marie (1997). *Random iterative models, volume 34 of Applications of Mathematics (New York)*. Springer-Verlag, Berlin.

Durrett, Rick (2019). *Probability: theory and examples*. Cambridge University Press, Cambridge. Fifth edition.

Hall, Peter and Christopher C Heyde (1980). *Martingale limit theory and its application*. Academic press.

Moulines, Eric and Francis Bach (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *24*, 451–459.

Polyak, Boris T and Anatoli B Juditsky (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization 30*(4), 838–855.

Robbins, Herbert and Sutton Monro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics 22*(3), 400–407.

Ruppert, David (1985). A newton-raphson version of the multivariate robbins-monro procedure. *The Annals of Statistics 13*(1), 236–245.

Su, Weijie J and Yuancheng Zhu (2018). Uncertainty quantification for online learning and stochastic approximation via hierarchical incremental gradient descent. *arXiv preprint arXiv:1802.04876*.