Appendix

The appendix provides additional lemmas, proofs of results in the main body, and additional results that are complementary to the main body.

Appendix A: Lemma 1

LEMMA 1. *Suppose that Assumption 1 holds for $\Theta = \mathbb{R}^d$. Consider a sequence $\{\theta_t\}_{t\in\mathbb{N}}$ iteratively defined by*

$$\theta_{t+1} = \theta_t - \gamma_t(\nabla_\theta g(\theta_t) + B_t + V_t). \tag{24}$$

Define $\mathcal{F}_t := \mathcal{F}(\theta_0, \theta_1, ...\theta_{t-1}, \theta_t)$. Assume that B_t is $\mathcal{F}_t$ measurable and $\mathbb{E}[V_t|\mathcal{F}_t] = 0$. Furthermore, assume that there exist constants $\beta \in (\frac{1}{2}, 1]$, $r \in \mathbb{R}$, $\rho, \gamma_0, K_1, K_2 \in (0,\infty)$ such that

$$\begin{aligned} \gamma_t &= \gamma_0 t^{-\beta}, \\ \|B_t\| &\leq K_1 t^{-\rho}, \\ \mathbb{E}[\|V_t\|^2|\mathcal{F}_t] &\leq K_2 t^{-r}. \end{aligned} \tag{25}$$

We require that $\beta + r > 0$. If $\beta = 1$, require additionally that $\gamma_0 \in (\max\{\frac{2\rho}{\mu}, \frac{1+r}{\mu}\}, \infty)$. Then there exists a $\kappa \in (0,\infty)$ such that for all $t \in \mathbb{N}$,

$$\mathbb{E}\left[\|\theta_t - \theta^*\|^2\right] \leq \kappa t^{-(2\rho)\wedge(\beta+r)}, \tag{26}$$

and

$$\mathbb{E}\left[g(\theta_t) - g(\theta^*)\right] \leq \frac{1}{2} L\kappa\, t^{-(2\rho)\wedge(\beta+r)}. \tag{27}$$

Proof of Lemma 1

$$\begin{aligned} &\mathbb{E}[\|\theta_{t+1} - \theta^*\|^2|\theta_t] - \|\theta_t - \theta^*\|^2 \\ =&2(\theta_t - \theta^*)^\top \mathbb{E}[\theta_{t+1} - \theta_t|\theta_t] + \mathbb{E}[\|\theta_{t+1} - \theta_t\|^2|\theta_t] \\ =&-2\gamma_t(\theta_t - \theta^*)^\top \mathbb{E}[\nabla_\theta g(\theta_t) + B_t|\theta_t] + \gamma_t^2 \mathbb{E}[\|\nabla_\theta g(\theta_t) + B_t + V_t\|^2|\theta_t] \\ \leq&-2\mu\gamma_t\|\theta_t - \theta^*\|^2 - 2\gamma_t(\theta_t - \theta^*)^\top B_t + 3\gamma_t^2(\|\nabla_\theta g(\theta_t)\|^2 + \|B_t\|^2 + \mathbb{E}[\|V_t\|^2|\theta_t]) \\ \leq&-2\mu\gamma_t\|\theta_t - \theta^*\|^2 - 2\gamma_t(\theta_t - \theta^*)^\top B_t + 3L^2\gamma_t^2\|\theta_t - \theta^*\|^2 + 3\gamma_t^2(\|B_t\|^2 + \mathbb{E}[\|V_t\|^2|\theta_t]), \end{aligned} \tag{28}$$

where we use the fact that $(\theta - \theta^*)^\top \nabla_\theta g(\theta) \geq \mu\|\theta - \theta^*\|^2$ and $\|\nabla_\theta g(\theta)\| \leq L\|\theta - \theta^*\|$. Using the inequality that $2a^\top b \leq (\|a\|^2 + \|b\|^2)$ with $a = \sqrt{\frac{\mu}{2}}(\theta_t - \theta^*)$ and $b = \sqrt{\frac{2}{\mu}}B_t$, we arrive at

$$-2(\theta_t - \theta^*)^\top B_t \leq \frac{\mu}{2}\|\theta_t - \theta^*\|^2 + \frac{2}{\mu}\|B_t\|^2.$$

Combining it with (28), we have

$$\begin{aligned}
&\mathbb{E}[\|\theta_{t+1}-\theta^*\|^2|\theta_t]-\|\theta_t-\theta^*\|^2\\
\le&-2\mu\gamma_t\|\theta_t-\theta^*\|^2+\frac{\mu}{2}\gamma_t\|\theta_t-\theta^*\|^2+\frac{2\gamma_t}{\mu}\|B_t\|^2+3L^2\gamma_t^2\|\theta_t-\theta^*\|^2+3\gamma_t^2(\|B_t\|^2+\mathbb{E}[\|V_t\|^2|\theta_t])\\
=&-\frac{3}{2}\mu\gamma_t\|\theta_t-\theta^*\|^2+\frac{2\gamma_t}{\mu}\|B_t\|^2+3L^2\gamma_t^2\|\theta_t-\theta^*\|^2+3\gamma_t^2(\|B_t\|^2+\mathbb{E}[\|V_t\|^2|\theta_t]).
\end{aligned}$$

Now consider our assumption that $\gamma_t=\gamma_0 t^{-\beta}$, $\|B_t\|\le K_1t^{-\rho}$ and $\mathbb{E}[\|V_t\|^2|\theta_t]\le K_2t^{-r}$. Since $\beta>0$, we can find t_0 such that $3L^2\gamma_0t_0^{-\beta}\le\frac{1}{2}\mu$, and $3\gamma_0t_0^{-\beta}\le\frac{1}{\mu}$. Therefore, for $t\ge t_0$,

$$\begin{aligned}
&\mathbb{E}[\|\theta_{t+1}-\theta^*\|^2|\theta_t]-\|\theta_t-\theta^*\|^2\\
\le&-\mu\gamma_t\|\theta_t-\theta^*\|^2+\frac{3\gamma_t}{\mu}\|B_t\|^2+3\gamma_t^2\,\mathbb{E}[\|V_t\|^2|\theta_t]\\
\le&-\mu\gamma_0t^{-\beta}\|\theta_t-\theta^*\|^2+\frac{3\gamma_0K_1^2}{\mu}t^{-2\rho-\beta}+3K_2\gamma_0^2t^{-2\beta-r}.
\end{aligned}$$

By taking expectation on both sides and moving $\mathbb{E}\,\|\theta_t-\theta^*\|^2$ to the right side, we get the recursion equation

$$\mathbb{E}[\|\theta_{t+1}-\theta^*\|^2]\le(1-\mu\gamma_0t^{-\beta})\,\mathbb{E}[\|\theta_t-\theta^*\|^2]+\frac{3\gamma_0K_1^2}{\mu}t^{-2\rho-\beta}+3K_2\gamma_0^2t^{-2\beta-r},\tag{29}$$

which holds for $t\ge t_0$.

By unrolling the recursion, we have

$$\mathbb{E}[\|\theta_{t+1}-\theta^*\|^2]\le\quad A_{t,t_0-1}\,\mathbb{E}[\|\theta_{t_0}-\theta^*\|^2]+\frac{3\gamma_0K_1^2}{\mu}\sum_{i=t_0}^{t}i^{-2\rho-\beta}A_{n,i}+3K_2\gamma_0^2\sum_{i=t_0}^{t}i^{-2\beta-r}A_{n,i},\tag{30}$$

where

$$A_{tj}=\begin{cases}\prod_{k=j+1}^{t}(1-\mu\gamma_0k^{-\beta}), & j<t,\\ 1, & j=t.\end{cases}$$

First we discuss the case when $\frac{1}{2}<\beta<1$. Without loss of generality we assume that t_0 also satisfies $2\rho t_0^{\beta-1}<\frac{\mu\gamma_0}{2}$. Notice that when $\frac{1}{2}<\beta<1$,

$$\begin{aligned}
|A_{tj}|\le&\exp\left(-\mu\gamma_0\sum_{k=j+1}^{t}k^{-\beta}\right)\\
\le&\exp\left(\mu\gamma_0j^{-\beta}-\mu\gamma_0\int_j^t x^{-\beta}dx\right)\\
=&\exp\left(\mu\gamma_0j^{-\beta}-\frac{\mu\gamma_0\,(t^{1-\beta}-j^{1-\beta})}{1-\beta}\right).
\end{aligned}$$

So the first term on the R.H.S. of (30) is $O(\exp(-\mu\gamma_0\frac{t^{1-\beta}}{1-\beta}))$. Besides, for the second term on the R.H.S. of (30), we have that

$$\sum_{i=t_0}^{t}i^{-\beta-2\rho}A_{ti}=t^{-2\rho}\sum_{i=t_0}^{t}i^{-\beta}A_{ti}+\sum_{i=t_0}^{t-1}\left(\frac{1}{i^{2\rho}}-\frac{1}{(i+1)^{2\rho}}\right)\sum_{j=t_0}^{i}j^{-\beta}A_{tj}.\tag{31}$$

For $t_0 \leq i \leq t$,

$$\sum_{j=t_0}^{i} j^{-\beta} A_{tj} = \frac{1}{\mu\gamma_0} \sum_{j=t_0}^{i} \left(A_{tj} - A_{t,j-1}\right) = \frac{1}{\mu\gamma_0} \left(A_{t,i} - A_{t,t_0-1}\right). \tag{32}$$

Notice that

$$\frac{1}{i^{2\rho}} - \frac{1}{(i+1)^{2\rho}} = i^{-2\rho}\left(2\rho i^{-1} + O\left(i^{-2}\right)\right) = 2\rho i^{-2\rho-1} + O\left(i^{-2\rho-2}\right), \tag{33}$$

So we have

$$\begin{aligned}\sum_{i=t_0}^{t-1} \left(\frac{1}{i^{2\rho}} - \frac{1}{(i+1)^{2\rho}}\right) A_{ti} &= \sum_{i=t_0}^{t-1} \left(2\rho i^{-2\rho-1} + O\left(i^{-2\rho-2}\right)\right) A_{ti} \\ &\leq \frac{\mu\gamma_0}{2} \sum_{i=t_0}^{t} i^{-2\rho-\beta} A_{ti} + O\left(t^{-2\rho-1}\right).\end{aligned} \tag{34}$$

The last inequality comes from the choice of t_0 such that $2\rho t_0^{\beta-1} < \frac{\mu\gamma_0}{2}$. Combining (32) and (34) with (31), we have

$$\begin{aligned}\sum_{i=t_0}^{t} i^{-\beta-2\rho} A_{ti} &\leq \frac{1}{\mu\gamma_0} \left(A_{t,t} - A_{t,t_0-1}\right) t^{-2\rho} + \left(\frac{\mu\gamma_0}{2} \sum_{i=t_0}^{t} i^{-2\rho-\beta} A_{ti}\right)\left(\frac{1}{\mu\gamma_0}\left(A_{t,t} - A_{t,t_0-1}\right)\right) + O\left(t^{-2\rho-1}\right) \\ &\leq \frac{1}{\mu\gamma_0} t^{-2\rho} + \frac{1}{2} \sum_{i=t_0}^{t} i^{-2\rho-\beta} A_{ti} + O\left(t^{-2\rho-1}\right)\end{aligned} \tag{35}$$

Therefore,

$$\sum_{i=t_0}^{t} i^{-\beta-2\rho} A_{ti} \leq \frac{2}{\mu\gamma_0} t^{-2\rho} + O\left(t^{-2\rho-1}\right). \tag{36}$$

Repeating the argument above, we have

$$\begin{aligned}\sum_{i=t_0}^{t} i^{-r-2\beta} A_{ti} &\leq \frac{1}{\mu\gamma_0} \left(A_{t,t} - A_{t,t_0-1}\right) t^{-r-\beta} + \left(\frac{\mu\gamma_0}{2} \sum_{i=t_0}^{t} i^{-r-2\beta} A_{ti}\right)\left(\frac{1}{\mu\gamma_0}\left(A_{t,t} - A_{t,t_0-1}\right)\right) + O\left(t^{-r-\beta-1}\right) \\ &\leq \frac{1}{\mu\gamma_0} t^{-r-\beta} + \frac{1}{2} \sum_{i=t_0}^{t} i^{-2\beta-r} A_{ti} + O\left(t^{-r-\beta-1}\right) \\ &\leq \frac{2}{\mu\gamma_0} t^{-r-\beta} + O\left(t^{-r-\beta-1}\right)\end{aligned} \tag{37}$$

Combining (36), (37) with (30), we conclude that $\mathbb{E}[\|\theta_{t+1} - \theta^*\|^2] = O(t^{-(2\rho)\wedge(\beta+r)})$ for $t \geq t_0$.

Now we consider the case when $\beta = 1$. For $t_0 - 1 \leq j \leq t$,

$$\begin{aligned}A_{tj} &\leq \exp\left(\mu\gamma_0/j - \mu\gamma_0 \int_j^t x^{-1} dx\right) \\ &= \exp\left(\mu\gamma_0/j - \mu\gamma_0 \ln(t/j)\right) \\ &= \exp\left(\mu\gamma_0/j\right)(j/t)^{\mu\gamma_0}\end{aligned} \tag{38}$$

For the second and the third term on the R.H.S. of (30), we have

$$\begin{aligned}\sum_{i=t_0}^{t} i^{-2\rho-1}\,|A_{ti}| &\leq \exp\left(\mu\gamma_0/t_0\right)\big(\sum_{i=t_0}^{t-1} i^{-2\rho-1}(i/t)^{\mu\gamma_0}+t^{-2\rho-1}\big)\\ &\leq \exp\left(\mu\gamma_0/t_0\right)t^{-\mu\gamma_0}\int_{t_0}^{t} x^{\mu\gamma_0-2\rho-1}dx+O(t^{-2\rho-1})\\ &\leq \exp\left(\mu\gamma_0/t_0\right)\left(\mu\gamma_0-2\rho\right)^{-1}t^{-2\rho}+O(t^{-2\rho-1}),\\ \sum_{i=t_0}^{t-1} i^{-2-r}\,|A_{ti}| &\leq \exp\left(\mu\gamma_0/t_0\right)\sum_{i=t_0}^{t-1} i^{-2-r}(i/t)^{\mu\gamma_0}+O(t^{-2-r})\\ &\leq \exp\left(\mu\gamma_0/t_0\right)\left(\mu\gamma_0-r-1\right)^{-1}t^{-r-1}+O(t^{-2-r}).\end{aligned} \tag{39}$$

Therefore, by combining (39) with (30), we have $\mathbb{E}[\|\theta_{t+1}-\theta^*\|^2]=O(t^{-(2\rho)\wedge(\beta+r)})$ for $t\geq t_0$. (27) holds because $g(\cdot)$ is L-smooth. □

If Θ is a closed and convex set of $\mathbb{R}^d$, and the iteration (24) changes to $\theta_{t+1}=\mathrm{pr}_\Theta(\theta_t-\gamma_t(\nabla_\theta g(\theta_t)+B_t+V_t))$, the conclusion of Lemma 1 still holds and the discussions are the same as before.

Appendix B: Proofs in Section 3

B.1. Proof of Theorem 1

Proof of Theorem 1 Define $B_t=\mathbb{E}[\frac{G_{m_t}(\theta_t+h_tZ,Y_{m_t})-G_{m_t}(\theta_t,Y_{m_t})}{h_t}Z]-\nabla g(\theta_t)$ and $V_t=\frac{1}{N_t}\sum_{l=1}^{N_t}\frac{G_{m_t}(\theta_t+h_tZ_{t,l},Y_{m_t,l})-G_{m_t}(\theta_t,Y_{m_t,l})}{h_t}Z_{t,l}-\mathbb{E}[\frac{G_{m_t}(\theta_t+h_tZ,Y_{m_t})-G_{m_t}(\theta_t,Y_{m_t})}{h_t}Z]$. We want to show that B_t and V_t satisfy conditions of Lemma 1.

Under Assumption 1, 2, and 4, for $k\geq 1$, we have

$$\begin{aligned}&\|\,\mathbb{E}[\frac{G_k(\theta+hZ,Y_k)-G_k(\theta,Y_k)}{h}Z]-\nabla g(\theta)\|\\ \leq&\|\,\mathbb{E}[\frac{G_k(\theta+hZ,Y_k)-G_k(\theta,Y_k)}{h}Z]-\mathbb{E}[\frac{g(\theta+hZ)-g(\theta)}{h}Z]\|+\|\,\mathbb{E}[\frac{g(\theta+hZ)-g(\theta)}{h}Z]-\nabla g(\theta)\|.\end{aligned}$$

Since

$$\begin{aligned}&\|\,\mathbb{E}[\frac{G_k(\theta+hZ,Y_k)-G_k(\theta,Y_k)}{h}Z]-\mathbb{E}[\frac{g(\theta+hZ)-g(\theta)}{h}Z]\|\\ =&\|\,\mathbb{E}[\mathbb{E}[\frac{G_k(\theta+hZ,Y_k)-G_k(\theta,Y_k)}{h}Z-\frac{g(\theta+hZ)-g(\theta)}{h}Z|Z]]\|\\ =&\|\,\mathbb{E}[\frac{g_k(\theta+hZ)-g_k(\theta)}{h}Z-\frac{g(\theta+hZ)-g(\theta)}{h}Z]\|\\ \leq&\mathbb{E}[(|\frac{g_k(\theta+hZ)-g(\theta+hZ)}{h}|+|\frac{g_k(\theta)-g(\theta)}{h}|)\|Z\|]\\ \leq&\frac{2}{kh}\,\mathbb{E}[\|Z\|]\leq\frac{2c_1d^{\frac{1}{2}}}{kh},\end{aligned}$$

and

$$\|\,\mathbb{E}[\frac{g(\theta+hZ)-g(\theta)}{h}Z]-\nabla g(\theta)\|$$

$$
\begin{aligned}
&= \| \mathbb{E}[ZZ^\top \nabla g(\theta + \bar{h}Z) - ZZ^\top \nabla g(\theta) \| \\
&\leq \mathbb{E}[\|ZZ^\top\| \|\nabla g(\theta + \bar{h}Z) - \nabla g(\theta)\|] \\
&\leq L\,\mathbb{E}[\|ZZ^\top\| \|\bar{h}Z\|] \\
&\leq Lh\,\mathbb{E}[\|Z\|^3] \leq c_3 L h d^{\frac{3}{2}},
\end{aligned}
$$

where $\bar{h}$ satisfies $0 \leq \bar{h} \leq h$ according to mean value theorem.

Therefore, by setting $m_t = \lceil m_0 d^2 t^{2\rho} \rceil$ and $h_t = d^{-\frac{3}{2}} t^{-\rho}$, we have

$$
\begin{aligned}
&\| \mathbb{E}[\frac{G_{m_t}(\theta_t + h_t Z, Y_{m_t}) - G_{m_t}(\theta_t, Y_{m_t})}{h_t} Z | \theta_t] - \nabla g(\theta_t) \| \\
\leq &\frac{2c_1 d^{\frac{1}{2}}}{m_t h_t} + c_3 L h_t d^{\frac{3}{2}} = O(t^{-\rho}).
\end{aligned}
$$

On the other hand, by Assumption 3,

$$
\begin{aligned}
\mathbb{E}[\|V_t\|^2 | \theta_t] = &\frac{1}{N_t} \mathbb{E}[\| \frac{G_{m_t}(\theta_t + h_t Z, Y_{m_t}) - G_{m_t}(\theta_t, Y_{m_t})}{h_t} Z - \mathbb{E}[\frac{G_{m_t}(\theta_t + h_t Z, Y_{m_t}) - G_{m_t}(\theta_t, Y_{m_t})}{h_t} Z]\|^2 | \theta_t] \\
\leq &\frac{1}{N_t} \mathbb{E}[\| \frac{G_{m_t}(\theta_t + h_t Z, Y_{m_t}) - G_{m_t}(\theta_t, Y_{m_t})}{h_t} Z \|^2 | \theta_t] = O(t^{-r}).
\end{aligned}
$$

According to Lemma 1, the conclusion of Theorem 1 holds. □

B.2. Proof of Theorem 2

Proof of Theorem 2 First, suppose that Assumption 5 (i) holds, so $d^{-2p} k^{-2\rho p} \mathbb{E}[C_{m_k}] \to m_0^p \kappa_1$ as $k \to \infty$. The expected cumulative computation cost by the t-th iteration is

$$
\begin{aligned}
\mathbb{E}\, T_t &= 2\,\mathbb{E} \sum_{j=1}^{t} N_j C_{m_j} \\
&= 2 \sum_{j=1}^{t} N_j\, \mathbb{E}\, C_{m_j} \\
&= 2N_0 \sum_{j=1}^{t} d^{5-\frac{3}{2}q} j^{r+\rho(2-q)}\, \mathbb{E}\, C_{m_j} \\
&= O(d^{2p+5-\frac{3}{2}q} t^{r+2\rho p+\rho(2-q)+1}).
\end{aligned}
$$

According to Theorem 1, $\tau(\epsilon) = O(\epsilon^{-\frac{2}{(2\rho)\wedge(\beta+r)}})$. Therefore,

$$
\mathbb{E}\, T_{\tau(\epsilon)} = O(d^{2p+5-\frac{3}{2}q} \epsilon^{-\frac{2(r+2\rho p+\rho(2-q)+1)}{(2\rho)\wedge(\beta+r)}}).
$$

Suppose that Assumption 5 (ii) holds. The expected cumulative computation cost by the t-th iteration is

$$
\mathbb{E}\, T_t = O(d^{5-\frac{3}{2}q} t^{r+\rho(2-q)+1} \exp((m_0 d^2 + 1)\log(\alpha) t^{2\rho})).
$$

According to Theorem 7, $\tau(\epsilon) \leq \kappa^{\frac{1}{(2\rho)\wedge(\beta+r)}} \epsilon^{-\frac{2}{(2\rho)\wedge(\beta+r)}}$. Therefore,

$$
\mathbb{E}\, T_{\tau(\epsilon)} = O(d^{5-\frac{3}{2}q} \epsilon^{-\frac{2(r+\rho(2-q)+1)}{(2\rho)\wedge(\beta+r)}} \exp((m_0 d^2 + 1) \kappa^{\frac{2\rho}{(2\rho)\wedge(\beta+r)}} \epsilon^{-\frac{4\rho}{(2\rho)\wedge(\beta+r)}} \log \alpha)).
$$

□

B.3. Proof of Proposition 1

Proof of Proposition 1 For $\beta \in (\frac{1}{2}, 1]$, we have that

$$\frac{r+\rho(2p+2-q)+1}{(2\rho)\wedge(\beta+r)} \geq \frac{r+\rho(2p+2-q)+\beta}{(2\rho)\wedge(\beta+r)} = \begin{cases} p+1-\frac{q}{2}+\frac{r+\beta}{2\rho}, & \text{if } 2\rho < \beta+r \\ 1+\frac{\rho(2p+2-q)}{r+\beta}, & \text{if } 2\rho \geq \beta+r \end{cases}$$

Given fixed β, the optimal value of $\frac{r+\rho(2p+2-q)+\beta}{(2\rho)\wedge(\beta+r)}$ is obtained when $2\rho = \beta + r$ and is $p+2-\frac{q}{2}$. Note that the equation $\frac{r+\rho(2p+2-q)+1}{(2\rho)\wedge(\beta+r)} = \frac{r+\rho(2p+2-q)+\beta}{(2\rho)\wedge(\beta+r)}$ holds only if $\beta = 1$. Therefore, $\beta = 1$, $2\rho = r+1$ represents the set of optimal parameters that minimize the computational cost needed for the algorithm to achieve a given precision level. □

Appendix C: Proofs in Section 4

C.1. Proof of Theorem 5

Proof of Theorem 5 According to the proof of Theorem 1,

$$\begin{aligned} &\| \mathbb{E}[\frac{G_{m_t}(\theta_t + h_t Z, Y) - G_{m_t}(\theta_t, Y)}{h_t} Z|\theta_t] - \nabla g(\theta_t)\| \\ \leq &\frac{2c_1 d^{\frac{1}{2}}}{h_t} M^{-m_t\alpha} + c_3 L h_t d^{\frac{3}{2}} = O(t^{-\rho}). \end{aligned}$$

Therefore,

$$\begin{aligned} \|B_t\| &= \| \mathbb{E}[\sum_{k=1}^{m_t} \frac{1}{N_{t,k}} \sum_{\ell=1}^{N_{t,k}} (F_k^{\mathrm{sm}}(\theta_t; h_t, Z_{t,k,l}, Y_{t,k,l}) - F_{k-1}^{\mathrm{sm}}(\theta_t; h_t, Z_{t,k,l}, Y_{t,k,l}))|\theta_t] - \nabla g(\theta_t)\| \\ &= \| \mathbb{E} \frac{G_{m_t}(\theta_t + h_t Z, Y) - G_{m_t}(\theta_t, Y)}{h_t} Z|\theta_t] - \nabla g(\theta_t)\| = O(t^{-\rho}). \end{aligned}$$

On the other hand,

$$\begin{aligned} &\mathbb{E}[\|F_k^{\mathrm{sm}}(\theta; h, Z, Y) - F_{k-1}^{\mathrm{sm}}(\theta; h, Z, Y) - \mathbb{E}\, F_k^{\mathrm{sm}}(\theta; h, Z, Y) + \mathbb{E}\, F_{k-1}^{\mathrm{sm}}(\theta; h, Z, Y)\|^2]^{1/2} \\ \leq &\mathbb{E}[\|F_k^{\mathrm{sm}}(\theta; h, Z, Y) - F_{k-1}^{\mathrm{sm}}(\theta; h, Z, Y) - \mathbb{E}[F_k^{\mathrm{sm}}(\theta; h, Z, Y)|Z] + \mathbb{E}[F_{k-1}^{\mathrm{sm}}(\theta; h, Z, Y)|Z]\|^2]^{1/2} \\ &+ \mathbb{E}[\| \mathbb{E}[F_k^{\mathrm{sm}}(\theta; h, Z, Y)|Z] - \mathbb{E}[F_{k-1}^{\mathrm{sm}}(\theta; h, Z, Y)|Z] - \mathbb{E}\, F_k^{\mathrm{sm}}(\theta; h, Z, Y) + \mathbb{E}\, F_{k-1}^{\mathrm{sm}}(\theta; h, Z, Y)\|^2]^{1/2} \\ = &\mathbb{E}[\mathbb{E}[\|F_k^{\mathrm{sm}}(\theta; h, Z, Y) - F_{k-1}^{\mathrm{sm}}(\theta; h, Z, Y) - \frac{g_k(\theta + hZ) - g_k(\theta)}{h} Z - \frac{g_{k-1}(\theta + hZ) - g_{k-1}(\theta)}{h} Z\|^2|Z]]^{1/2} \\ &+ \mathbb{E}[\|\frac{g_k(\theta + hZ) - g_k(\theta)}{h} Z - \frac{g_{k-1}(\theta + hZ) - g_{k-1}(\theta)}{h} Z - \mathbb{E}\, F_k^{\mathrm{sm}}(\theta; h, Z, Y) + \mathbb{E}\, F_{k-1}^{\mathrm{sm}}(\theta; h, Z, Y)\|^2]^{1/2} \\ \leq &\mathbb{E}[\|Z\|^2\, \mathbb{E}[\|\frac{G_k(\theta + hZ, Y) - G_{k-1}(\theta + hZ, Y) - g_k(\theta + hZ) + g_{k-1}(\theta + hZ)}{h}\|^2|Z]]^{1/2} \\ &+ \mathbb{E}[\|Z\|^2\, \mathbb{E}[\|\frac{G_k(\theta, Y) - G_{k-1}(\theta, Y) - g_k(\theta) + g_{k-1}(\theta)}{h}\|^2|Z]]^{1/2} \\ &+ \mathbb{E}[\|\frac{g_k(\theta + hZ) - g_k(\theta)}{h} Z - \frac{g_{k-1}(\theta + hZ) - g_{k-1}(\theta)}{h} Z - \mathbb{E}\, F_k^{\mathrm{sm}}(\theta; h, Z, Y) + \mathbb{E}\, F_{k-1}^{\mathrm{sm}}(\theta; h, Z, Y)\|^2]^{1/2} \\ \leq &\mathbb{E}[\|Z\|^2\, \mathbb{E}[\|\frac{G_k(\theta + hZ, Y) - G_{k-1}(\theta + hZ, Y) - g_k(\theta + hZ) + g_{k-1}(\theta + hZ)}{h}\|^2|Z]]^{1/2} \end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}[\|Z\|^2\,\mathbb{E}[\|\frac{G_k(\theta,Y)-G_{k-1}(\theta,Y)-g_k(\theta)+g_{k-1}(\theta)}{h}\|^2|Z]]^{1/2}\\
&+\mathbb{E}[\|\frac{g_k(\theta+hZ)-g_k(\theta)}{h}Z-\frac{g_{k-1}(\theta+hZ)-g_{k-1}(\theta)}{h}Z\|^2]^{1/2}\\
\le&\,\mathbb{E}[\frac{\|Z\|^2}{h^2}M^{-2k\eta}]^{1/2}+\mathbb{E}[\frac{4M\|Z\|^2}{h^2}M^{-2k\alpha}]^{1/2}\\
\le&\,\frac{\tilde{C}'\sqrt{d}M^{-k\eta}}{h},
\end{aligned}
$$

where the last inequality comes from the assumption that $\alpha \ge \eta$.

Under given parameters h_t, $N_{t,k}$ and m_t, we calculate that

$$
\begin{aligned}
\mathbb{E}[\|V_t\|^2|\theta_t]=&\sum_{k=1}^{m_t}\frac{1}{N_{t,k}^2}\sum_{\ell=1}^{N_{t,k}}\mathbb{E}[\|F_k^{\mathrm{sm}}(\theta_t;h_t,Z_{t,k,l},Y_{t,k,l})-F_{k-1}^{\mathrm{sm}}(\theta_t;h_t,Z_{t,k,l},Y_{t,k,l})\\
&-\mathbb{E}[F_k^{\mathrm{sm}}(\theta_t;h_t,Z_{t,k,l},Y_{t,k,l})|\theta_t]+\mathbb{E}[F_{k-1}^{\mathrm{sm}}(\theta_t;h_t,Z_{t,k,l},Y_{t,k,l})|\theta_t]\|^2|\theta_t]\\
\le&\sum_{k=1}^{m_t}\frac{C'd}{N_0\kappa_t h_t^2}M^{k(1/2-\eta)}\le\frac{C'}{N_0(M-1)}t^{-r}.
\end{aligned}
$$

According to Lemma 1, the conclusion of Theorem 5 is proved. □

C.2. Proof of Theorem 6

Proof of Theorem 6 The expected cumulative computation cost by the t-th iteration is

$$
\mathbb{E}\,T_t=2\,\mathbb{E}\sum_{j=1}^{t}\sum_{k=1}^{m_j}N_{j,k}C_k.
$$

The expected cost for the j-th iteration is

$$
\begin{aligned}
2\,\mathbb{E}\sum_{k=1}^{m_j}N_{j,k}C_k\le&2\sum_{k=1}^{m_j}\left(1+N_0\kappa_j M^{-k(\eta+1/2)}\right)M^k\\
\le&2\sum_{k=1}^{m_j}M^k+2\sum_{k=1}^{m_j}N_0\kappa_j M^{k(1/2-\eta)}.
\end{aligned}
$$

If $\eta\ne\frac{1}{2}$,

$$
\begin{aligned}
\mathbb{E}\sum_{k=1}^{m_j}N_{j,k}C_k\le&\frac{M^{m_j}}{1-M^{-1}}+N_0\kappa_j\frac{M^{m_j(1/2-\eta)_+}}{1-M^{-|1/2-\eta|}}\\
\le&c(M^{m_j}+d^4j^{r+2\rho}M^{m_j(1-2\eta)_+}).
\end{aligned}
$$

According to definition of m_j, $M^{m_j}\le d^{\frac{2}{\alpha}}j^{\frac{2\rho}{\alpha}}$, so

$$
\mathbb{E}\sum_{k=1}^{m_j}N_{j,k}C_k\le c(d^{\frac{2}{\alpha}}j^{\frac{2\rho}{\alpha}}+d^{4+\frac{2(1-2\eta)_+}{\alpha}}j^{r+2\rho+\frac{2\rho(1-2\eta)_+}{\alpha}}).
$$

If $\eta=\frac{1}{2}$,

$$
\begin{aligned}
\mathbb{E}\sum_{k=1}^{m_j}N_{j,k}C_k\le&\frac{M^{m_j}}{1-M^{-1}}+N_0\kappa_j m_j\\
\le&c(M^{m_j}+d^4j^{r2\rho}m_j^2)\\
\le&c(d^{\frac{2}{\alpha}}j^{\frac{2\rho}{\alpha}}+d^4j^{r+2\rho}\log_M(j)^2).
\end{aligned}
$$

Using the fact that $r = 2\rho - 1$, we have

$$\mathbb{E}\,T_t \le c_2 \begin{cases} \sum_{j=1}^{t} \left(d^{\frac{2}{\alpha}} j^{\frac{2\rho}{\alpha}} + d^{4+\frac{2(1-2\eta)_+}{\alpha}} j^{4\rho+\frac{2\rho(1-2\eta)_+}{\alpha}-1}\right), & \text{if } \eta \neq 1/2 \\ \sum_{j=1}^{t} \left(d^{\frac{2}{\alpha}} j^{\frac{2\rho}{\alpha}} + d^4 j^{4\rho-1} \log_M(j)^2\right), & \text{if } \eta = 1/2. \end{cases}$$

Hence there exists $c' \in (0, \infty)$ such that

$$\mathbb{E}\,T_t \le c' \begin{cases} \left(d^{\frac{2}{\alpha}} t^{\frac{2\rho}{\alpha}+1} + d^{4+\frac{2(1-2\eta)_+}{\alpha}} t^{4\rho+\frac{2\rho(1-2\eta)_+}{\alpha}}\right), & \text{if } \eta \neq 1/2 \\ \left(d^{\frac{2}{\alpha}} t^{\frac{2\rho}{\alpha}+1} + d^4 t^{4\rho} \log_M(t)^2\right), & \text{if } \eta = 1/2. \end{cases}$$

According to Theorem 5, $\tau(\epsilon) = O(\epsilon^{-\frac{1}{\rho}})$. Therefore,

$$\mathbb{E}\,T_{\tau(\epsilon)} = \begin{cases} O\left(d^{\frac{2}{\alpha}} \epsilon^{-\frac{2}{\alpha}-\frac{1}{\rho}} + d^{4+\frac{2(1-2\eta)_+}{\alpha}} \epsilon^{-4-\frac{2(1-2\eta)_+}{\alpha}}\right), & \text{if } \eta \neq 1/2 \\ O\left(d^{\frac{2}{\alpha}} \epsilon^{-\frac{2}{\alpha}-\frac{1}{\rho}} + d^4 \epsilon^{-4} \log_M(\epsilon^{-1})^2\right), & \text{if } \eta = 1/2. \end{cases}$$

If additionally $\alpha > \frac{1}{2}$ and $\rho > \frac{\alpha}{4\alpha+2(1-2\eta)_+-2}$, then $\frac{2}{\alpha} + \frac{1}{\rho} \le 4 + \frac{2(1-2\eta)_+}{\alpha}$. In this case,

$$\mathbb{E}\,T_{\tau(\epsilon)} = \begin{cases} O(d^4 \epsilon^{-4}), & \text{if } \eta > \frac{1}{2}, \\ O(d^4 \epsilon^{-4} (\ln(\epsilon^{-1}))^2), & \text{if } \eta = \frac{1}{2}, \\ O(d^{4+\frac{2(1-2\eta)_+}{\alpha}} \epsilon^{-\left(4+\frac{2-4\eta}{\alpha}\right)}), & \text{if } \eta < \frac{1}{2}. \end{cases}$$

□

Appendix D: Additional Results

D.1. Simulation Algorithms with IPA/AD/BP Gradient Estimators

In this section, we propose gradient-based simulation-optimization algorithms that take advantage of the infinitesimal perturbation analysis (IPA) gradient estimators, the automatic differentiation (AD) gradient estimators, or the backpropagation (BP) gradient estimators of the approximating systems. All three classes of gradient estimators can enjoy computational efficiency for gradient evaluation at high-dimensional decision variables. For simplicity, we use IPA gradient estimators in this section to represent all three classes. For the approximating system $G_k(\theta, Y_k)$ with index k, the IPA gradient estimator is given by

$$\nabla_\theta G_k(\theta, Y_k). \tag{40}$$

The main advantages of IPA gradient estimators are twofold. First, when the system performance function $G_k(\theta, Y_k)$ is differentiable and is Lipschitz continuous in θ, the IPA gradient estimator is unbiased, in the sense that

$$\mathbb{E}[\nabla_\theta G_k(\theta, Y_k)] = \nabla_\theta\, \mathbb{E}[G_k(\theta, Y_k)]. \tag{41}$$

Second, when the dominant computation cost is from running the simulation logic (i.e., evaluating $G_k(\theta, Y_k)$ given θ and Y_k), the IPA gradient estimator with respect to a high dimensional variable

$\theta \in \mathbb{R}^d$, can be simultaneously obtained from a *single* simulation run of $G_k(\theta, Y_k)$. That is, when the computation cost of a single simulation run of $G_k(\theta, Y_k)$ is C_k, the computation cost of obtaining the gradient vector $\nabla_\theta G_k(\theta, Y_k)$ is given by $R \cdot C_k$, where $R > 1$ is a constant multiplier that does not increase linearly with dimension d; see Griewank et al. (1989) and Fu and Hu (2012).

Consider an increasing positive integer sequence $(m_t : t \geq 1)$. At the t-th step, the algorithm updates the θ_{t-1} from the previous step by using a gradient estimator constructed from the m_t-th approximating system. Specifically, the algorithm generates N_t independent copies of the random input Y_{m_t}, noted as $\{Y_{m_t,l}\}_{l=1}^{N_t}$. Correspondingly, the algorithm runs N_t independent simulation copies of the simulation logic of the m_t-th system and obtains N_t copies of the IPA gradient estimator

$$\nabla_\theta G_{m_t}(\theta_{t-1}, Y_{m_t,1}), \nabla_\theta G_{m_t}(\theta_{t-1}, Y_{m_t,2}), \ldots, \nabla_\theta G_{m_t}(\theta_{t-1}, Y_{m_t,N_t}). \tag{42}$$

The algorithm then averages the N_t independent IPA gradient estimator as

$$H_t(\theta_{t-1}) := \frac{1}{N_t} \sum_{l=1}^{N_t} \nabla_\theta G_{m_t}(\theta_{t-1}, Y_{m_t,l}) \tag{43}$$

and then updates θ_{t-1} as

$$\theta_t = \mathrm{pr}_\Theta(\theta_{t-1} - \gamma_t H_t(\theta_{t-1})). \tag{44}$$

For a set of initialization parameters $\gamma_0, N_0, m_0 > 0$, $\beta, r, \rho \geq 0$, we consider algorithm parameters given by

$$m_t = \lceil m_0 t^{2\rho} \rceil, \quad N_t = \lceil N_0 t^r \rceil, \quad \gamma_t = \gamma_0 \frac{1}{t^\beta} \tag{45}$$

for $t \geq 1$. We impose the following assumption on the sequence of approximating systems.

Assumption 8. *There exists a positive constant M_0 such that for all $k \in \mathbb{N}$ and $\theta \in \mathbb{R}^d$,*

(i) The expected performance g_k is L_k-smooth, and $L^* := \sup_{k \geq 1} L_k < \infty$;

(ii) $G_k(\cdot, Y_k)$ is $\Psi_k(Y_k)$-Lipschitz continuous and $\mathbb{E}[|\Psi_k(Y_k)|^2] < \infty$;

(iii) $\mathbb{E}[\|\nabla_\theta G_k(\theta, Y_k) - \mathbb{E}\,\nabla_\theta G_k(\theta, Y_k)\|^2] \leq M_0$.

Theorem 7. *Suppose that $\beta \in (\frac{1}{2}, 1]$, $r \geq 0$, $\rho, \gamma_0, N_0, m_0 \in (0, \infty)$. If $\beta = 1$, suppose additionally that $\gamma_0 \in (\max\{\frac{2\rho}{\mu}, \frac{1+r}{\mu}\}, \infty)$. Under Assumption 1, 2, and 8, with θ_t defined in the scheme* (43) *and* (44)*, there exists a $\kappa \in (0, \infty)$ such that for all $t \in \mathbb{N}$,*

$$\mathbb{E}\left[\|\theta_t - \theta^*\|^2\right] \leq \kappa t^{-(2\rho) \wedge (\beta + r)}, \tag{46}$$

and

$$\mathbb{E}\left[g(\theta_t) - g(\theta^*)\right] \leq \frac{1}{2} L \kappa t^{-(2\rho) \wedge (\beta + r)}, \tag{47}$$

where κ only depends on $\beta, r, \rho, \gamma_0, m_0, N_0, L^*, L$, μ and M_0.

Proof of Theorem 7. First, we show that under conditions of Theorem 7, $\|\mathbb{E}\left[\nabla_\theta G_{m_t}(\theta, Y_{m_t}) - \nabla_\theta g(\theta)\right]\| \le (L^* + L + 2)m_t^{-\frac{1}{2}} \le \frac{L^*+L+2}{\sqrt{m_0}}t^{-\rho}$.

Denote $u_n := \frac{1}{\|\nabla_\theta g_n(\theta) - \nabla_\theta g(\theta)\|}(\nabla_\theta g_n(\theta) - \nabla_\theta g(\theta))$. That is, u_n is the unit vector in $\mathbb{R}^d$ that shares the same direction with $\nabla_\theta g_n(\theta) - \nabla_\theta g(\theta)$. For any $h_n > 0$ and $\theta \in \Theta$, we have

$$\begin{aligned}
&|\frac{g_n(\theta + h_n u_n) - g_n(\theta)}{h_n} - \nabla_\theta g_n(\theta)^\top u_n| \\
\le& |(\nabla_\theta g_n(\theta + \xi_n(\theta) h_n u_n) - \nabla_\theta g_n(\theta))^\top u_n| \\
\le& \|\nabla_\theta g_n(\theta + \xi_n(\theta) h_n u_n) - \nabla_\theta g_n(\theta)\|,
\end{aligned}$$

where $\xi_n(\theta) \in (0,1)$ and its value depends on n and θ. Using the fact that $g_n(\cdot)$ is L_n- smooth, we have

$$|\frac{g_n(\theta + h_n u_n) - g_n(\theta)}{h_n} - \nabla_\theta g_n(\theta)^\top u_n| \le L_n h_n \le L^* h_n.$$

So

$$\begin{aligned}
&\|\mathbb{E}\left[\nabla_\theta G_n(\theta, Y_n) - \nabla_\theta g(\theta)\right]\| \\
=& |(\nabla_\theta g_n(\theta)) - \nabla_\theta g(\theta))^\top u_n| \\
\le& |\nabla_\theta g_n(\theta)^\top u_n - \frac{g_n(\theta + h_n u_n) - g_n(\theta)}{h_n}| + |\frac{g_n(\theta + h_n u_n) - g_n(\theta)}{h_n} - \frac{g(\theta + h_n u_n) - g(\theta)}{h_n}| + |\frac{g(\theta + h_n u_n) - g(\theta)}{h_n} - \nabla_\theta g(\theta)^\top u_n| \\
\le& |\nabla_\theta g_n(\theta)^\top u_n - \frac{g_n(\theta + h_n u_n) - g_n(\theta)}{h_n}| + |\frac{g_n(\theta + h_n u_n) - g(\theta + h_n u_n)}{h_n}| + |\frac{g_n(\theta) - g(\theta)}{h_n}| + |\frac{g(\theta + h_n u_n) - g(\theta)}{h_n} - \nabla_\theta g(\theta)^\top u_n| \\
\le& L^* h_n + \frac{2}{n h_n} + L h_n.
\end{aligned}$$

If we set $h_n = n^{-\frac{1}{2}}$, $\|\mathbb{E}\left[\nabla_\theta G_n(\theta, Y_n) - \nabla_\theta g(\theta)\right]\|$ is bounded by $(L^* + L + 2)n^{-\frac{1}{2}}$. Therefore, we have that $\|\mathbb{E}\left[\nabla_\theta G_{m_t}(\theta, Y_{m_t}) - \nabla_\theta g(\theta)\right]\| \le (L^* + L + 2)m_t^{-\frac{1}{2}} \le \frac{L^*+L+2}{\sqrt{m_0}}t^{-\rho}$.

On the other hand, under Assumption 8 (iii), $\mathbb{E}[\|\frac{1}{N_t}\sum_{l=1}^{N_t} \nabla_\theta G_{m_t}(\theta_t, Y_{m_t,l}) - \nabla g_{m_t}(\theta_t)\|^2 | \theta_t] = \frac{1}{N_t}\mathbb{E}[\|\nabla_\theta G_{m_t}(\theta_t, Y_{m_t}) - \nabla g_{m_t}(\theta_t)\|^2 | \theta_t] \le \frac{M_0}{N_0}t^{-r}$.

Let $B_t = \mathbb{E}\left[\nabla_\theta G_{m_t}(\theta, Y_{m_t}) - \nabla_\theta g(\theta)\right]$ and $V_t = \frac{1}{N_t}\sum_{l=1}^{N_t} \nabla_\theta G_{m_t}(\theta_t, Y_{m_t,l}) - \nabla g_{m_t}(\theta_t)$. The iteration scheme (43) and (44) can be written as

$$\theta_{t+1} = \theta_t - \gamma_t(\nabla g(\theta_t) + B_t + V_t),$$

and B_t and V_t satisfy all conditions of Lemma 1, according to the analysis above. Therefore, according to Lemma 1, the conclusion of Theorem 7 holds. □

The cumulative computation cost for the algorithm by the $\tau(\epsilon)$-th iteration, is denoted as

$$T_{\tau(\epsilon)} = R\sum_{j=1}^{\tau(\epsilon)} N_j C_{m_j}. \tag{48}$$

Theorem 8. *Under Assumption 5 (i),*

$$\mathbb{E}\, T_{\tau(\epsilon)} = O(\epsilon^{-\frac{2(r+2\rho p+1)}{(2\rho)\wedge(\beta+r)}}).$$

Under Assumption 5 (ii),

$$\mathbb{E}\, T_{\tau(\epsilon)} = O(\epsilon^{-\frac{2(r+1)}{(2\rho)\wedge(\beta+r)}} \exp(\epsilon^{-\frac{4\rho}{(2\rho)\wedge(\beta+r)}} \kappa^{\frac{2\rho}{(2\rho)\wedge(\beta+r)}} (m_0+1)\log\alpha)),$$

in which $O(g(\epsilon))$ denotes a function of ϵ that is bounded by a constant multiplied by $g(\epsilon)$.

Proof of Theorem 8. First, suppose that Assumption 5 (i) holds, so $k^{-2\rho p}\,\mathbb{E}[C_{m_k}] \to m_0^p \kappa_1$ as $k \to \infty$. The expected cumulative computation cost by the t-th iteration is

$$\begin{aligned}
\mathbb{E}\, T_t &= \mathbb{E}\, R \sum_{j=1}^{t} N_j C_{m_j} \\
&= R \sum_{j=1}^{t} N_j\, \mathbb{E}\, C_{m_j} \\
&= R N_0 \sum_{j=1}^{t} t^r\, \mathbb{E}\, C_{m_j} \\
&= O(t^{r+2\rho p+1}).
\end{aligned}$$

According to Theorem 7, $\tau(\epsilon) = O(\epsilon^{-\frac{2}{(2\rho)\wedge(\beta+r)}})$. Therefore,

$$\mathbb{E}\, T_{\tau(\epsilon)} = O(\epsilon^{-\frac{2(r+2\rho p+1)}{(2\rho)\wedge(\beta+r)}}).$$

Suppose that Assumption 5 (i) holds. The expected cumulative computation cost by the t-th iteration is

$$\begin{aligned}
\mathbb{E}\, T_t &= R \sum_{j=1}^{t} N_j\, \mathbb{E}\, C_{m_j} \\
&= R N_0 \sum_{j=1}^{t} t^r\, \mathbb{E}\, C_{m_j} \\
&= O(t^{r+1} \exp((m_0+1)\log(\alpha) t^{2\rho})).
\end{aligned}$$

According to Theorem 7, $\tau(\epsilon) \le \kappa^{\frac{1}{(2\rho)\wedge(\beta+r)}} \epsilon^{-\frac{2}{(2\rho)\wedge(\beta+r)}}$. Therefore,

$$\mathbb{E}\, T_{\tau(\epsilon)} = O(\epsilon^{-\frac{2(r+1)}{(2\rho)\wedge(\beta+r)}} \exp(\epsilon^{-\frac{4\rho}{(2\rho)\wedge(\beta+r)}} \kappa^{\frac{2\rho}{(2\rho)\wedge(\beta+r)}} (m_0+1)\log\alpha)).$$

□

We next devote our attention to studying the asymptotic distribution of θ_t when t tends to infinity and obtain a central limit theorem (CLT). We then change our lens to the available computation budget C and derive a central limit theorem for the best estimator available with the given budget. Both two results of CLT are under suitable regularity assumptions:

ASSUMPTION 9. *$H(\theta) := \nabla^2 g(\theta)$ exists for every $\theta \in \Theta$ and is continuous with respect to θ. Denote $H^* := H(\theta^*)$. All eigenvalues of $H^* - \frac{1+r}{2\gamma_0} I$ have positive real parts.*

ASSUMPTION 10. *$\|\theta_n\| < \infty$ a.s. $\forall n$.*

ASSUMPTION 11. *θ^* is an asymptotically stable solution of the following ordinary differential equation*

$$\frac{dx(t)}{dt} = -\nabla g(x).$$

Define $D(\theta^*) = \{x_0 : \lim_{t\to\infty} x(t|x_0) = \theta^*\}$, where $x(t|x_0)$ denotes the solution to the ordinary differential equation based on initial condition x_0. There exists a compact $S \subset D(\theta^*)$ such that $\theta_n \in S$ infinitely often for almost all sample points.

ASSUMPTION 12. *There exists $\delta > 0$ such that $\sup_{n\in\mathbb{N},\theta\in\Theta} \mathbb{E}[\|\nabla_\theta G_n(\theta, Y_n)\|^{2+\delta}] < \infty$.*

ASSUMPTION 13. *There exists a continuous function $\mathcal{E}(\cdot)$ such that $n^{1/2}(\nabla_\theta g_n(\theta) - \nabla_\theta g(\theta))$ converges to $\mathcal{E}(\theta)$ uniformly for every $\theta \in \Theta$. Especially, denote $C' := \mathcal{E}(\theta^*)$.*

ASSUMPTION 14. *$\mathbb{E}[\sup_{n\in\mathbb{N},\theta\in\Theta} \|\nabla_\theta G_n(\theta, Y_n)\|^2] < \infty$.*

ASSUMPTION 15. *There exists a constant $b > 0$ such that*

$$\lim_{n\to\infty} \frac{\mathrm{Var}(C_n)}{n^{p(1-b)}} = 0.$$

THEOREM 9. *Denote H^* as the Hessian matrix for $g(\theta)$ at θ^* and $\tilde{H} := H^* - \frac{1+r}{2\gamma_0} I$. Denote $C' := \lim_{n\to\infty} n^{1/2}(\nabla_\theta g_n(\theta^*) - \nabla_\theta g(\theta^*))$. Under suitable regularity assumptions, for the optimal algorithm,*

$$n^{(1+r)/2}(\theta_n - \theta^*) \xrightarrow{d} N(-m_0^{-\frac{1}{2}} \tilde{H}^{-1} C', \Sigma) \quad as \quad n \to \infty$$

where

$$\Sigma = \frac{\gamma_0}{N_0} \int_0^\infty \exp(-\tilde{H}u)\, \mathbb{E}[\nabla_\theta G(\theta^*, Y) \nabla_\theta G(\theta^*, Y)^\top] \exp(-\tilde{H}^\top u) du. \tag{49}$$

THEOREM 10. *Suppose that all conditions of Theorem 9 are satisfied. Let C be the computation budget and $n(C) := \sup\{n \geq 1 : \sum_{j=1}^n N_j R C_{m_j} \leq C\}$. If $n^{-p}\,\mathbb{E}[C_n] \to \kappa_1$ for some $p > 0$ and $\kappa_1 > 0$, then*

$$(\frac{C}{\kappa_1 N_0 R m_0^p (p+1)})^{1/2(1+p)} (\theta_{n(C)} - \theta^*) \xrightarrow{d} N(-m_0^{-\frac{1}{2}} \tilde{H}^{-1} C', \Sigma) \quad as \quad C \to \infty$$

with Σ defined the same as in Theorem 9.

Proof of Theorem 9 Suppose that Assumption 1, 2, 8, 9, 10, 11, 12, 13 and 14 hold.

Under Assumption 1, 2, 8, 9, 10, and 11, according to Kushner and Clark (2012) Theorem 2.3.1, (see also Theorem 1 of Ljung (1977) and Proposition 1 of Spall et al. (1992)), $\theta_t \to \theta^*$ w.p.1 as $t \to \infty$.

We show that conditions (2.2.1), (2.2.2), and (2.2.3) in Fabian et al. (1968) hold. First, we observe that

$$\nabla_\theta g(\theta_t) = \nabla_\theta g(\theta^*) + H(\bar{\theta}_t)(\theta_t - \theta^*) = H(\bar{\theta}_t)(\theta_t - \theta^*),$$

where $\bar{\theta}_t$ lies on the line segment between θ_t and θ^*. Then we have

$$\begin{aligned}\theta_{t+1} - \theta^* =& \theta_t - \theta^* - \gamma_t(\nabla_\theta g(\theta_t) + B_t + V_t)\\ =&(I - \gamma_0 t^{-1} H(\bar{\theta}_t))(\theta_t - \theta^*) - \gamma_0 t^{-1} B_t + \gamma_0 t^{-1}\Phi_t V_t\\ =&(I - \gamma_0 t^{-1} H(\bar{\theta}_t))(\theta_t - \theta^*) - \gamma_0 t^{-1-\rho}(t^\rho B_t) + \gamma_0 t^{-1-\rho+\frac{1}{2}}(t^{\rho-\frac{1}{2}}\Phi_t V_t),\end{aligned}$$

where $B_t = \nabla_\theta g_{m_t}(\theta_t) - \nabla_\theta g(\theta_t)$, $\Phi_t = -I$ and $V_t = \frac{1}{N_t}\sum_{l=1}^{N_t} \nabla_\theta G_{m_t}(\theta_t, Y_{m_t,l}) - \nabla_\theta g_{m_t}(\theta_t)$.

Since $\theta_t \to \theta^*$ w.p.1 as $t \to \infty$ and by the continuity of $H(\cdot)$, we have $H(\bar{\theta}_t) \to H(\theta^*)$ w.p.1. According to Assumption 13, $t^\rho B_t = t^\rho(\nabla_\theta g_{\lceil m_0 t^{2\rho}\rceil}(\theta_t) - \nabla_\theta g(\theta_t)) \to m_0^{-\frac{1}{2}} C'$ w.p. 1, so condition (2.2.1) of Fabian et al. (1968) holds.

Under Assumption 14, by dominated convergence theorem and the fact that $\theta_t \to \theta^*$ w.p.1 as $t \to \infty$,

$$\begin{aligned} t^r\,\mathbb{E}[V_t V_t^\top | \theta_t] =& \frac{t^r}{N_t}\,\mathbb{E}[(\nabla_\theta G_{m_t}(\theta_t, Y_{m_t}) - \nabla_\theta g_{m_t}(\theta_t))(\nabla_\theta G_{m_t}(\theta_t, Y_{m_t}) - \nabla_\theta g_{m_t}(\theta_t))^\top | \theta_t]\\ &\xrightarrow{p} \frac{1}{N_0}\,\mathbb{E}[(\nabla_\theta G(\theta^*, Y) - \nabla g(\theta^*))(\nabla_\theta G(\theta^*, Y)^\top - \nabla g(\theta^*))^\top]\\ &= \frac{1}{N_0}\,\mathbb{E}[\nabla_\theta G(\theta^*, Y)\nabla_\theta G(\theta^*, Y)^\top]\end{aligned}$$

as $t \to \infty$. So condition (2.2.2) of Fabian et al. (1968) holds.

For $0 < \delta' < \delta/2$, and any $\lambda > 0$, we have

$$\lim_{k\to\infty} \mathbb{E}\left[1_{\left\{\left\|k^{\frac{r}{2}} V_k\right\|^2 \geq \lambda k\right\}} \left\|k^{\frac{r}{2}} V_k\right\|^2\right] \quad \leq \lim_{k\to\infty}\sup \left(\frac{\mathbb{E}\left\|k^{\frac{r}{2}} V_k\right\|^2}{\lambda k}\right)^{\delta'/\left(1+\delta'\right)} \left(\mathbb{E}\left\|k^{\frac{r}{2}} V_k\right\|^{2\left(1+\delta'\right)}\right)^{1/\left(1+\delta'\right)}$$

By Burkholder-Davis-Gundy inequality and the triangle inequality on the $L^{\frac{1}{2(1+\delta')}}$ space, there exists a constant $c_{\delta'}$ which only depends on δ', such that

$$\begin{aligned}\left(\mathbb{E}\left\|k^{\frac{r}{2}}V_k\right\|^{2\left(1+\delta'\right)}\right)^{1/\left(1+\delta'\right)} &= \left(\mathbb{E}\left\|\frac{1}{N_0 k^{\frac{r}{2}}}\sum_{l=1}^{N_k}(\nabla_\theta G_{m_k}(\theta_k, Y_{m_k,l}) - \nabla_\theta g_{m_k}(\theta_k))\right\|^{2\left(1+\delta'\right)}\right)^{1/\left(1+\delta'\right)} \\ &\le c_{\delta'}\left(\mathbb{E}[(\frac{1}{N_0^2 k^r}\sum_{l=1}^{N_k}\|\nabla_\theta G_{m_k}(\theta_k, Y_{m_k,l}) - \nabla_\theta g_{m_k}(\theta_k)\|^2)^{\left(1+\delta'\right)}]\right)^{1/\left(1+\delta'\right)} \\ &\le c_{\delta'}\frac{1}{N_0^2 k^r}\sum_{l=1}^{N_k}\left(\mathbb{E}(\|\nabla_\theta G_{m_k}(\theta_k, Y_{m_k,l}) - \nabla_\theta g_{m_k}(\theta_k)\|^{2\left(1+\delta'\right)}\right)^{1/\left(1+\delta'\right)} \\ &\le \frac{c_{\delta'}}{N_0}(\sup_{k\in\mathbb{N},\theta\in\Theta}\mathbb{E}[\|\nabla_\theta G_k(\theta, Y_k)\|^{2+\delta}])^{1/\left(1+\delta'\right)} < \infty.\end{aligned}$$

So $\lim_{k\to\infty}\mathbb{E}\left[1_{\left\{\left\|k^{\frac{r}{2}}V_k\right\|^2\ge\lambda k\right\}}\left\|k^{\frac{r}{2}}V_k\right\|^2\right] \to 0$ for every $\lambda > 0$. Therefore, (2.2.3) of Fabian et al. (1968) holds and CLT is proved. □

Proof of Theorem 10 Suppose that all conditions of Theroem 9 hold and additionally Assumption 15 holds. Since $n^{-p}\mathbb{E}C_n \to \kappa_1$, we have $N_0^{-1}R^{-1}j^{-r}m_j^{-p}\mathbb{E}[N_jRC_{m_j}] = N_0^{-1}R^{-1}m_0^{-p}j^{-r-2\rho p}\mathbb{E}[N_jRC_{m_j}] \to \kappa_1$.

For the cumulative computation cost by iteration n, denoted as T_n, we have

$$\frac{\mathbb{E}T_n}{\kappa_1 N_0 R m_0^p(p+1)n^{r+1+2\rho p}} \xrightarrow{p} 1 \tag{50}$$

as $n\to\infty$. Under Assumption 15, $\frac{\mathrm{Var}(N_nRC_{m_n})}{\mathbb{E}T_n} = o(n^{-1-2\rho pb})$, so

$$\sum_{j=1}^{\infty}\frac{\mathrm{Var}(N_jRC_{m_j})}{\mathbb{E}T_j} < \infty.$$

According to Kronecker's Law of Large Numbers,

$$\frac{T_n - \mathbb{E}T_n}{\mathbb{E}T_n} \xrightarrow{p} 0.$$

Therefore,

$$\frac{T_n}{\kappa_1 N_0 R m_0^p(p+1)n^{r+1+2\rho p}} \xrightarrow{p} 1. \tag{51}$$

Combining (51) with the conclusion of Theorem 9, we have

$$(\frac{T_n}{\kappa_1 N_0 R m_0^p(p+1)})^{\frac{r+1}{2(r+1+2\rho p)}}(\theta_n - \theta^*) \xrightarrow{d} N(-\tilde{H}^{-1}C', \Sigma) \quad as \quad n\to\infty$$

Changing T_n to C and n to $n(C)$, and using the fact that $r+1 = 2\rho$, the conclusion in Theorem 10 is derived. □

D.2. CLT for Finite Difference Gradient Estimator

Assumption 16. *There exists* $\delta > 0$ *such that* $\sup_{t\in\mathbb{N},\theta\in\Theta}\mathbb{E}[\|\frac{G_{m_t}(\theta+h_tZ,Y)-G_{m_t}(\theta,Y)}{h_t}Z\|^{2+\delta}] < +\infty$.

Assumption 17. *Denote* $H_n(\cdot)$ *as the Hessian matrix of* $g_n(\cdot)$*. There exists a constant* $\tilde{L} > 0$ *such that* $\sup_{n\in\mathbb{N},\theta\in\Theta}\|H_n(\theta)\| < \tilde{L}$.

Proof of Theorem 3 Suppose that Assumption 1, 2, 9, 10, 11, 12, 13, 14, 16 and 17 hold.

For any $k\in\mathbb{N}$, $G_k(\cdot,Y)$ is Lipschitz continuous. Because of Assumption 12, we set $q=2$ in the parameter setting.

We show that conditions (2.2.1), (2.2.2), and (2.2.3) in Fabian et al. (1968) hold. For simplicity, denote $\mathbb{E}_t(\cdot) := \mathbb{E}[\cdot|\theta_t]$, $H_t^{mc}(\theta) = \frac{G_{m_t}(\theta+h_tZ,Y)-G_{m_t}(\theta,Y)}{h_t}Z$, and $F(\theta,Y) = \nabla_\theta G(\theta,Y)$. We want to show that $t^\rho(\mathbb{E}_t[H_t^{mc}(\theta_t)] - \nabla_\theta g(\theta_t)) \xrightarrow{p} \frac{1}{2}\mathbb{E}[ZZ^\top HZ] + m_0^{-\frac{1}{2}}d^{-1}C'$ and $\mathbb{E}_t[(H_t(\theta_t) - \mathbb{E}_t[H_t^{mc}(\theta_t)])(H_t(\theta_t) - \mathbb{E}_t[H_t^{mc}(\theta_t)])^\top] \xrightarrow{p} \Omega$. First, by Taylor expansion we have

$$\begin{aligned}
&\mathbb{E}_t[H_t^{mc}(\theta_t)] - \nabla g(\theta_t)\\
=&\mathbb{E}_t[\frac{g_{m_t}(\theta_t+h_tZ) - g_{m_t}(\theta_t)}{h_t}Z] - \nabla g(\theta_t)\\
=&\mathbb{E}_t[\nabla g_{m_t}(\theta_t)ZZ^\top + \frac{1}{2}h_tZZ^\top H_{m_t}(\theta_t+\xi(m_t,h_t,Z)h_tZ)Z] - \nabla g(\theta_t)\\
=&\nabla g_{m_t}(\theta_t) - \nabla g(\theta_t) + \frac{1}{2}h_t\,\mathbb{E}_t[ZZ^\top H_{m_t}(\theta_t+\xi(m_t,h_t,Z)h_tZ)Z],
\end{aligned}$$

where $\xi(m_t,h_t,Z)\in(0,1)$ and $H_{m_t}(\cdot)$ is the Hessian matrix of $g_{m_t}(\cdot)$.

Since $\theta_t\to\theta^*$ w.p.1 and $h_t\to 0$ as $t\to\infty$, $H_{m_t}(\theta_t+\xi(m_t,h_t,Z)h_tZ)\to H$ w.p.1 as $t\to\infty$. According to Assumption 17, for any $t\in\mathbb{N}$ and any $\theta\in\mathbb{R}^d$, $\|H_{m_t}(\theta)\|\le\tilde{L}$. Therefore, $\|ZZ^\top H_{m_t}(\theta_t+\xi(m_t,h_t,Z)h_tZ)Z\|\le L\|Z\|^3$. By dominated convergence theorem, $\mathbb{E}_t[ZZ^\top H_{m_t}(\theta_t+\xi(m_t,h_t,Z)h_tZ)Z] \xrightarrow{p} \mathbb{E}[ZZ^\top HZ]$ as $t\to\infty$.

On the other hand, $t^\rho(\nabla_\theta g_{\lceil m_0d^2t^{2\rho}\rceil}(\theta_t) - \nabla_\theta g(\theta_t)) \to m_0^{-\frac{1}{2}}d^{-1}C'$ w.p.1. So we have

$$t^\rho(\mathbb{E}_t[H_t^{mc}(\theta_t)] - \nabla_\theta g(\theta_t)) \xrightarrow{p} \frac{1}{2}\mathbb{E}[ZZ^\top HZ] + m_0^{-\frac{1}{2}}d^{-1}C'. \tag{52}$$

By Taylor expansion $\frac{G_{m_t}(\theta_t+h_tZ,Y)-G_{m_t}(\theta_t,Y)}{h_t}Z = ZZ^\top F_{m_t}(\theta_t+\lambda(Y,Z,h_t,\theta_t)h_tZ,Y)$ with $0<\lambda(Y,Z,h_t,\theta_t)<1$. Therefore,

$$\begin{aligned}
&\mathbb{E}_t[(H_t^{mc}(\theta_t) - \mathbb{E}_t[H_t^{mc}(\theta_t)])(H_t^{mc}(\theta_t) - \mathbb{E}_t[H_t^{mc}(\theta_t)])^\top]\\
=&\mathbb{E}_t[(Z^\top F_{m_t}(\theta_t+\lambda(Y,Z,h_t,\theta_t)h_tZ,Y))^2ZZ^\top] - \mathbb{E}_t[H_t^{mc}(\theta_t)]\,\mathbb{E}_t[H_t^{mc}(\theta_t)]^\top\\
=&\mathbb{E}_t[(Z^\top F_{m_t}(\theta_t+\lambda(Y,Z,h_t,\theta_t)h_tZ,Y))^2ZZ^\top] - \mathbb{E}_t[H_t^{mc}(\theta_t)]\,\mathbb{E}_t[H_t^{mc}(\theta_t)]^\top.
\end{aligned}$$

Since $\mathbb{E}_t[H_t^{mc}(\theta_t)] = \nabla g(\theta_t) + o_p(1)$ and $\nabla g(\theta_t) \xrightarrow{p} \nabla g(\theta^*) = 0$, we have $\mathbb{E}_t[H_t^{mc}(\theta_t)]\,\mathbb{E}_t[H_t^{mc}(\theta_t)]^\top \xrightarrow{p} 0$. By the fact that $\theta_t\to\theta^*$ w.p.1, we have $F_{m_t}(\theta_t +$

$\lambda(Y,Z,h_t.\theta_t)h_tZ,Y) = F(\theta^*,Y)+o_p(1)$. We can bound $\|(Z^\top F_{m_t}(\theta_t+\lambda(Y,Z,h_t,\theta_t)h_tZ,Y))^2ZZ^\top\|$ by

$$(Z^\top F_{m_t}(\theta_t+\lambda(Y,Z,h_t,\theta_t)h_tZ,Y))^2ZZ^\top \le \|Z\|^4 \sup_{n\in\mathbb{N},\theta\in\Theta}\|\nabla_\theta G_n(\theta,Y_n)\|^2.$$

so by Assumption 14 and dominated convergence theorem, $\mathbb{E}_t[(Z^\top F(\theta_t + \lambda(Y,Z,h_t,\theta_t)h_tZ,Y))^2ZZ^\top] = \mathbb{E}[(Z^\top F(\theta^*,Y))^2ZZ^\top]+o_p(1)$. Therefore,

$$\begin{aligned}&\mathbb{E}_t[(H_t^{mc}(\theta_t)-\mathbb{E}_t[H_t^{mc}(\theta_t)])(H_t^{mc}(\theta_t)-\mathbb{E}_t[H_t^{mc}(\theta_t)])^\top]\\ =&\mathbb{E}[(Z^\top F(\theta^*,Y))^2ZZ^\top]+o_p(1).\end{aligned}$$

Under Assumption 16, (2.2.3) of Fabian et al. (1968) can be checked to stand using the same technique as in the proof of Theorem 9.

Therefore, according to Fabian et al. (1968), the central limit theorem in Theorem 3 is proved.

Proof of Theorem 4 Suppose that Assumption 15 holds. Since $n^{-p}\mathbb{E}C_n \to \kappa_1$, we have $N_0^{-1}R^{-1}d^{-2}j^{-r}m_j^{-p}\mathbb{E}[N_jC_{m_j}] = N_0^{-1}d^{-2-2p}m_0^{-p}j^{-r-2\rho p}\mathbb{E}[N_jC_{m_j}] \to \kappa_1$. Then, for the cumulative computation cost by iteration n, denoted as T_n, similarly to the proof of Theorem 10, we have

$$\frac{T_n}{2\kappa_1N_0d^{2(p+1)}m_0^p(p+1)n^{r+1+2\rho p}} \xrightarrow{p} 1 \tag{53}$$

as $n\to\infty$. Combining (53) with the conclusion of Theorem 9, we have

$$(\frac{T_n}{2\kappa_1N_0d^{2(p+1)}m_0^p(p+1)})^{\frac{r+1}{2(r+1+2\rho p)}}(\theta_n-\theta^*) \xrightarrow{d} N(-\tilde{H}^{-1}(\frac{1}{2}\mathbb{E}[ZZ^\top H^*Z]+m_0^{-\frac{1}{2}}d^{-1}C'),\Sigma) \quad as \quad n\to\infty$$

Changing T_n to C and n to $n(C)$, and using the fact that $r+1=2\rho$, the conclusion in Theorem 4 is derived. □

D.3. Optimal Set of Algorithm Parameters for Multilevel FD Estimator

Given parameter r,ρ and β, according to the proof of Theorem 6,

$$\mathbb{E}T_t \le c\begin{cases}(d^{\frac{2}{\alpha}}t^{\frac{2\rho}{\alpha}+1}+d^{4+\frac{2(1-2\eta)_+}{\alpha}}t^{r+1+2\rho+\frac{2\rho(1-2\eta)_+}{\alpha}}), & \text{if } \eta\neq 1/2\\ (d^{\frac{2}{\alpha}}t^{\frac{2\rho}{\alpha}+1}+d^4t^{r+1+2\rho}\log_M(t)^2), & \text{if } \eta=1/2\end{cases}$$

where c is a constant independent with r, ρ and β.

According to Theorem 5, $\tau(\epsilon)=O(\epsilon^{-\frac{2}{(2\rho)\wedge(\beta+r)}})$. Therefore,

$$\mathbb{E}T_{\tau(\epsilon)} = \begin{cases}O\left(d^{\frac{2}{\alpha}}\epsilon^{-(\frac{2\rho}{\alpha}+1)\frac{2}{(2\rho)\wedge(\beta+r)}}+d^{4+\frac{2(1-2\eta)_+}{\alpha}}\epsilon^{-\frac{2(r+1+2\rho+\frac{2\rho(1-2\eta)_+}{\alpha})}{(2\rho)\wedge(\beta+r)}}\right), & \text{if } \eta\neq 1/2\\ O\left(d^{\frac{2}{\alpha}}\epsilon^{-(\frac{2\rho}{\alpha}+1)\frac{2}{(2\rho)\wedge(\beta+r)}}+d^4\frac{4}{((2\rho)\wedge(\beta+r))^2}\epsilon^{-\frac{2(r+1+2\rho)}{(2\rho)\wedge(\beta+r)}}\log_M(\epsilon^{-1})^2\right), & \text{if } \eta=1/2.\end{cases}$$

Now we suppose that ρ is fixed. When $\eta\neq 1/2$, $(\frac{2\rho}{\alpha}+1)\frac{2}{(2\rho)\wedge(\beta+r)}$ is minimized when $\beta+r\ge 2\rho$. Consider the minimization of $\frac{2(r+1+2\rho+\frac{2\rho(1-2\eta)_+}{\alpha})}{(2\rho)\wedge(\beta+r)}$. For $\beta\in(\frac{1}{2},1]$, we have that

$$\frac{r+1+\rho(2+\frac{2(1-2\eta)_+}{\alpha})}{(2\rho)\wedge(\beta+r)} \ge \frac{r+\beta+\rho(2+\frac{2(1-2\eta)_+}{\alpha})}{(2\rho)\wedge(\beta+r)} = \begin{cases}1+\frac{(1-2\eta)_+}{\alpha}+\frac{r+\beta}{2\rho}, & \text{if } 2\rho<\beta+r\\ 1+\frac{\rho(2+\frac{2(1-2\eta)_+}{\alpha})}{r+\beta}, & \text{if } 2\rho\ge\beta+r\end{cases}$$

The minimal value of $\frac{r+\beta+\rho(2+\frac{2(1-2\eta)_+}{\alpha})}{(2\rho)\wedge(\beta+r)}$ is obtained when $2\rho=\beta+r$. Note that the equation $\frac{r+1+\rho(2+\frac{2(1-2\eta)_+}{\alpha})}{(2\rho)\wedge(\beta+r)}=\frac{r+\beta+\rho(2+\frac{2(1-2\eta)_+}{\alpha})}{(2\rho)\wedge(\beta+r)}$ holds only if $\beta=1$. Therefore, given fixed ρ and for the case when $\eta\neq 1/2$, $\beta=1$, $r+1=2\rho$ represents the set of optimal parameters that minimize the computational cost needed for the algorithm to achieve a given precision level.

If $\eta=1/2$, $\frac{4}{((2\rho)\wedge(\beta+r))^2}$ is minimized when $\beta+r\geq 2\rho$. Given fixed ρ, $\frac{2(r+1+2\rho+\frac{2\rho(1-2\eta)_+}{\alpha})}{(2\rho)\wedge(\beta+r)}$ and $(\frac{2\rho}{\alpha}+1)\frac{2}{(2\rho)\wedge(\beta+r)}$ are minimized when $\beta=1$ and $r+1=2\rho$. Therefore, when $\eta=1/2$, the order of expected computation cost is minimized when $r+1=2\rho$ and $\beta=1$.

Appendix E: Algorithm Details

E.1. Algorithm for Multilevel Finite Difference Gradient Estimator

Algorithm 2 Simulation optimization with multilevel finite difference estimator

```
Input:Number of iterations $N$, initial point $\theta_0$, parameters $N_0\in(0,\infty)$, $r\geq 0$, $\rho=\frac{1+r}{2}$, $\gamma_0\in(\frac{1+r}{\mu},\infty)$
Output: $\theta_N$
for $t=1$ to $N-1$ do
    Set $m_t=\lceil\frac{2}{\alpha}\log_M(dt^\rho)\rceil$, $h_t=d^{-\frac{3}{2}}t^{-\rho}$ and $N_{t,k}=\kappa_t M^{-k(\eta+1/2)}$, where
        $$\kappa_t=\begin{cases} d^4t^{r+2\rho}M^{m_t\left(\frac{1}{2}-\eta\right)+}, & \text{if } \eta\neq\frac{1}{2}\\ d^4t^{r+2\rho}m_t, & \text{if } \eta=\frac{1}{2}\end{cases}$$
    generates $N_{t,k}$ independent copies of $Y$ for $k=1,...m_t$, denoted by $\{(Y_{t,k,l})_{l=1}^{N_{t,k}}\}_{k=1}^{m_t}$
    Update $\theta_{t+1}=\mathrm{pr}_\Theta(\theta_t-\frac{\gamma_0}{t}H_t(\theta_t))$, where
        $$H_t(\theta)=\sum_{k=1}^{m_t}\frac{1}{N_{t,k}}\sum_{\ell=1}^{N_{t,k}}(F_k^{\mathrm{sm}}(\theta;h_t,Z_{t,k,l},Y_{t,k,l})-F_{k-1}^{\mathrm{sm}}(\theta;h_t,Z_{t,k,l},Y_{t,k,l}))$$
end for
return $\theta_N$
```