Appendix A. Technique lemmas

The first two lemmas we will used in our theorems are from Lemma A.1 and Lemma A.2 in Lei et al. (2017b).

Lemma A.1 *Let* $x_1, ..., x_M \in \mathbb{R}^d$ *be an arbitrary population of N vectors with*

$$\sum_{j=1}^{M} x_j = 0.$$

Further let $\mathcal{J}$ *be a uniform random subset of* $\{1, ...M\}$ *with size m. Then*

$$\mathbb{E} \parallel \frac{1}{m} \sum_{j \in \mathcal{J}} \parallel^2 = \frac{M-m}{(M-1)m} \frac{1}{M} \sum_{j=1}^{M} \parallel x_j \parallel^2 \leq \frac{I(m < M)}{m} \frac{1}{M} \sum_{j=1}^{M} \parallel x_j \parallel^2 .$$

The geometric random variable N_j has the key properties below.

Lemma A.2 *Let N Geom(γ) for some* $B > 0$*. Then for any sequence* $D_0,\ D_1, ...,\ D_N$ *with* $\mathbb{E}|D_N| < \infty$*,*

$$\mathbb{E}(D_N - D_{N+1}) = (\frac{1}{\gamma} - 1)(D_0 - \mathbb{E}D_N).$$

Appendix B. One-Epoch Analysis

B.1. Unbiased Estimator Version

Our algorithm is based on the SVRG method, thus the hyper-parameter λ should be within the range as $0 < \lambda < 1$ in both unbiased and biased cases. We start by bounding the gradient $\mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k^{(j)} \parallel^2$ in Lemma B.1 and the variance $\mathbb{E}_{\mathcal{I}_j} \parallel e_j \parallel^2$ in Lemma B.2.

Lemma B.1 *Under Definition 2.3,*

$$\mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k^{(j)} \parallel^2 \leq \frac{L^2}{b_j} \parallel (1-\lambda)x_k^{(j)} - \lambda x_0^{(j)} \parallel + 2(1-\lambda)^2 \parallel \nabla f(x_k^{(j)}) \parallel^2 + 2\lambda^2 \parallel e_j \parallel^2 .$$

Proof *Using the fact that for a random variable Z* $\mathbb{E} \parallel Z \parallel^2 = \|Z - \mathbb{E}Z \parallel^2 + \parallel \mathbb{E}Z \parallel^2$*, we have*

$$\begin{aligned}
\mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k^{(j)} \parallel^2 =& \mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k^{(j)} - \mathbb{E}_{\tilde{\mathcal{I}}_k} v_k^{(j)} \parallel^2 + \parallel \mathbb{E}_{\tilde{\mathcal{I}}_k} v_k^{(j)} \parallel^2 \\
=& \mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel (1-\lambda)\nabla f_{\tilde{\mathcal{I}}_k}(x_k^{(j)}) - \lambda \nabla f_{\tilde{\mathcal{I}}_k}(x_0^{(j)}) - ((1-\lambda)\nabla f(x_k^{(j)}) - \lambda \nabla f x_0^{(j)}) \parallel^2 \\
& + \parallel (1-\lambda)\nabla f(x_k^{(j)}) + \lambda e_j \parallel^2 \\
\leq& \mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel (1-\lambda)\nabla f_{\tilde{\mathcal{I}}_k}(x_k^{(j)}) - \lambda \nabla f_{\tilde{\mathcal{I}}_k}(x_0^{(j)}) - ((1-\lambda)\nabla f(x_k^{(j)}) - \lambda \nabla f x_0^{(j)}) \parallel^2 \\
& + 2 \parallel (1-\lambda)\nabla f(x_k^{(j)}) \parallel^2 + 2 \parallel \lambda e_j \parallel^2 .
\end{aligned} \tag{6}$$

By Lemma A.1,

$$
\begin{aligned}
&\mathbb{E}_{\tilde{\mathcal{I}}_k} \| (1-\lambda)\nabla f_{\tilde{\mathcal{I}}_k}(x_k^{(j)}) - \lambda\nabla f_{\tilde{\mathcal{I}}_k}(x_0^{(j)}) - ((1-\lambda)\nabla f(x_k^{(j)}) - \lambda\nabla f x_0^{(j)}) \|^2 \\
&\leq \frac{1}{b_j} \cdot \frac{1}{n} \sum_{i=1}^{n} \| (1-\lambda)\nabla f_i(x_k^{(j)}) - \lambda\nabla f_i(x_0^{(j)}) - ((1-\lambda)\nabla f(x_k^{(j)}) - \lambda\nabla f(x_0^{(j)})) \|^2 \\
&= \frac{1}{b_j} \cdot (\frac{1}{n} \sum_{i=1}^{n} \| (1-\lambda)\nabla f_i(x_k^{(j)}) - \lambda\nabla f_i(x_0^{(j)}) \|^2 - \| ((1-\lambda)\nabla f(x_k^{(j)}) - \lambda\nabla f(x_0^{(j)})) \|^2) \\
&\leq \frac{1}{b_j} \cdot \frac{1}{n} \sum_{i=1}^{n} \| (1-\lambda)\nabla f_i(x_k^{(j)}) - \lambda\nabla f_i(x_0^{(j)}) \|^2 \\
&\leq \frac{1}{b_j} \cdot L^2 \| (1-\lambda)x_k^{(j)} - \lambda x_0^{(j)} \|^2
\end{aligned} \tag{7}
$$

where the last line is based on Definition 2.3, then the bound of the gradient can be alternatively written as,

$$
\mathbb{E}_{\tilde{\mathcal{I}}_k} \| v_k^{(j)} \|^2 \leq \frac{L^2}{b_j} \| (1-\lambda)x_k^{(j)} - \lambda x_0^{(j)} \|^2 + 2(1-\lambda)^2 \| \nabla f(x_k^{(j)}) \|^2 + 2\lambda^2 \| e_j \|^2 . \tag{8}
$$

■

Lemma B.2

$$
\mathbb{E}_{\mathcal{I}_j} \| e_j \|^2 \leq \lambda^2 \frac{I(B_j < n)}{B_j} \cdot \mathcal{S}^*.
$$

Proof *Based on Lemma B.1 and the observation that* $\tilde{x}_{j-1}$ *is independent of* $\mathcal{I}_j$*, the bound of variance* e_j *can be expressed as*

$$
\begin{aligned}
\mathbb{E}_{\mathcal{I}_j} \| e_j \|^2 &= \frac{n - B_j}{(n-1)B_j} \cdot \frac{\lambda^2}{n} \sum_{i=1}^{n} \| \nabla f_i(\tilde{x}_{j-1}) - \nabla f(\tilde{x}_{j-1}) \|^2 \\
&\leq \lambda^2 \frac{n - B_j}{(n-1)B_j} \cdot \mathcal{S}^* \leq \lambda^2 \frac{I(B_j < n)}{B_j} \mathcal{S}^*
\end{aligned} \tag{9}
$$

where the upper bound of the variance of the stochastic gradients $\mathcal{S}^* = \frac{1}{n}\sum_{i=1}^{n} \| \nabla f_i(\tilde{x}_{j-1}) - \nabla f(\tilde{x}_{j-1}) \|^2$. ■

Theorem 3.1 below defines the bound of batch-size, B_j, for the unbiased estimator case.

Proof of Theorem 3.1

Theorem *If the expectation of the variance* $\mathbb{E}_{\mathcal{I}_j} \| e_j \|^2 \leq \sigma\rho^{2j}$ *in Alg 2 (*$\sigma \geq 0$ *is a constant for some* $\rho < 1$*), the lower bound of the batch-size,* B_j*, can be expressed as,*

$$
B_j \geq \frac{n\mathcal{S}^*}{\mathcal{S}^* + \lambda^2 n^{\frac{1}{2}} \sigma\rho^{2j}}.
$$

Proof *To define the bound of the batch-size,* B_j*, for the biased estimator case, we estimate the lower and upper bounds of the variance to control the size of the batch. Based on the result from Lemma B.2 and using the result that the norms of the gradients are bounded by* $\mathcal{K}^2$ *for all* x_j *(Babanezhad et al., 2015), we have*

$$\frac{1}{n-1}\sum_{i=1}^{n}[\| \nabla f_i(\tilde{x}_{j-1}) \|^2 - \| \nabla f(\tilde{x}_{j-1}) \|^2] \leq \mathcal{K}^2, \tag{10}$$

and using the inequality from (L. Lohr, 2000) we have

$$\mathbb{E}_{\mathcal{I}_j} \| e_j \|^2 \leq \lambda^2 \frac{n - B_j}{nB_j}\mathcal{K}^2. \tag{11}$$

If we want $\mathbb{E}_{\mathcal{I}_j} \| e_j \|^2 \leq \sigma\rho^{2j}$*, for a constant value* $\sigma \geq 0$ *and for some* $\rho^{2j} < 1$*, we need*

$$B_j \geq \frac{n\mathcal{K}^2}{\mathcal{K}^2 + n\lambda^2\sigma\rho^{2j}} \tag{12}$$

Using the Samuelson inequality (Niezgoda, 2007), $\mathcal{K}^2$ *satisfies*

$$\begin{aligned}&\sqrt{(n-1)\frac{1}{n-1}\sum_{i=1}^{n}[\| \nabla f_i(\tilde{x}_{j-1}) \|^2 - \| \nabla f(\tilde{x}_{j-1}) \|^2]} \\ &\geq n \cdot (\nabla f_i(\tilde{x}_{j-1}) - \nabla f(\tilde{x}_{j-1})).\end{aligned} \tag{13}$$

Inq. 13 can alternatively be written using Lemma B.2 as

$$\begin{aligned}&\sqrt{n-1}\mathbb{E}[\| \nabla f_i(\tilde{x}_{j-1}) \|^2 - \| \nabla f(\tilde{x}_{j-1}) \|^2 \\ &\geq n\mathbb{E}[\nabla f_i(\tilde{x}_{j-1}) - \nabla f(\tilde{x}_{j-1})]^2.\end{aligned} \tag{14}$$

Inq. 14 can be substituted by upper bounds $\mathcal{K}$ *and* $\mathcal{S}^*$ *giving*

$$\sqrt{n-1} \cdot \mathcal{K}^2 \geq n \cdot \mathcal{S}^*. \tag{15}$$

Thus, the result from Inq. 12 can be written as

$$\begin{aligned}B_j &\geq \frac{n\mathcal{K}^2}{\mathcal{K}^2 + n\lambda^2\sigma\rho^{2j}} \\ &\geq \frac{n\frac{n}{\sqrt{n-1}}\mathcal{S}^*}{\frac{n}{\sqrt{n-1}}\mathcal{S}^* + n\lambda^2\sigma\rho^{2j}}.\end{aligned} \tag{16}$$

■

Lemma B.3 *Suppose* $\eta_j L < 1$*, then under Definition 2.3,*

$$\begin{aligned}&(1-\lambda)\eta_j(1-(1-\lambda)L\eta_j)B_j\mathbb{E} \| \nabla f(\tilde{x}_j) \|^2 + \lambda\eta_j B_j \mathbb{E} < e_j, \nabla f(\tilde{x}_j) > \\ &\leq b_j\mathbb{E}(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)) + \frac{\eta_j^2 B_j L^3}{2b_j}\mathbb{E} \| \tilde{x}_j - \tilde{x}_{j-1} \|^2 + \lambda^2 L\eta_j^2 B_j \mathbb{E} \| e_j \|^2 .\end{aligned}$$

where $\mathbb{E}$ denotes the expectation with respect to all randomness.
Proof *By Definition 2.3, we have*

$$\begin{aligned}
&\mathbb{E}_{\tilde{\mathcal{I}}_k}[f(x_{k+1}^{(j)})] \leq f(x_k^{(j)}) - \eta_j < \mathbb{E}_{\tilde{\mathcal{I}}_k} v_k, \nabla f(x_k^{(j)}) > + \frac{L\eta_j^2}{2}\mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k \parallel^2 \\
&= f(x_k^{(j)}) - \eta_j < ((1-\lambda)\nabla f(x_k^{(j)}) + \lambda e_j), \nabla f(x)_k^{(j)}) > + \frac{L\eta_j^2}{2}\mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k \parallel^2 \\
&\leq f(x_k^{(j)}) - \eta_j(1-\lambda) \parallel \nabla f(x_k^{(j)}) \parallel^2 - \eta_j < \lambda e_j, \nabla f(x_k^{(j)}) > + \frac{L^3\eta_j^2}{2b_j} \parallel (1-\lambda)x_k^{(j)} - \lambda x_0^{(j)} \parallel^2 \\
&+ L\eta_j^2(1-\lambda)^2 \parallel \nabla f(x_k^{(j)}) \parallel^2 + L\eta_j^2\lambda^2 \parallel e_j \parallel^2 \\
&= f(x_k^{(j)}) - (\eta_j(1-\lambda) - L\eta_j^2(1-\lambda)^2) \parallel \nabla f(x_k^{(j)}) \parallel^2 - \lambda\eta_j < e_j, \nabla f(x_k^{(j)}) > \\
&+ \frac{L^3\eta_j^2}{2b_j} \parallel (1-\lambda)x_k^{(j)} - \lambda x_0^{(j)} \parallel^2 + L\eta_j^2\lambda^2 \parallel e_j \parallel^2 \\
&\leq f(x_k^{(j)}) - (\eta_j(1-\lambda) - L\eta_j^2(1-\lambda)^2) \parallel \nabla f(x_k^{(j)}) \parallel^2 - \lambda\eta_j < e_j, \nabla f(x_k^{(j)}) > \\
&+ \frac{L^3\eta_j^2}{2b_j} \parallel x_k^{(j)} - x_0^{(j)} \parallel^2 + L\eta_j^2\lambda^2 \parallel e_j \parallel^2
\end{aligned} \tag{17}$$

Let $\mathbb{E}_j$ denote the expectation $\tilde{\mathcal{I}}_0, \tilde{\mathcal{I}}_1, \ldots,$ given $\tilde{\mathcal{N}}_j$ since $\tilde{\mathcal{N}}_j$ is independent of them and let $k{=}\mathcal{N}_j$ in Inq. 17. As $\tilde{\mathcal{I}}_{k+1}, \tilde{\mathcal{I}}_{k+2}, \ldots$ are independent of $x_k^{(j)}$ and taking the expectation with respect to $\mathcal{N}_j$ and using Fubini's theorem, Inq. 17 implies that

$$\begin{aligned}
&\eta_j(1-\lambda)(1-(1-\lambda)L\eta_j)\mathbb{E}_{\mathcal{N}_j}\mathbb{E}_j[\parallel \nabla f(x_{\mathcal{N}_j}^{(j)}) \parallel^2] + \lambda\eta_j\mathbb{E}_{\mathcal{N}_j}\mathbb{E}_j < e_j, \nabla f(x_{\mathcal{N}_j}^{(j)}) > \\
&\leq \mathbb{E}_{\mathcal{N}_j}(\mathbb{E}_j[f(x_{\mathcal{N}_j}^{(j)})] - \mathbb{E}_j[f(x_{\mathcal{N}_{j+1}}^{(j)})]) + \frac{L^3\eta_j^2}{2b_j}\mathbb{E}_{\mathcal{N}_j}\mathbb{E}_j\mathbb{E}[\parallel (1-\lambda)x_{\mathcal{N}_j}^{(j)} - \lambda x_0^{(j)} \parallel^2] + L\lambda^2\eta_j^2 \parallel e_j \parallel^2 \\
&= \frac{b_j}{B_j}(f(x_0^{(j)}) - \mathbb{E}_j\mathbb{E}_{\mathcal{N}_j}[f_{\mathcal{N}_j}^{(j)}]) + \frac{L^3\eta_j^2}{2b_j}\mathbb{E}_j\mathbb{E}_{\mathcal{N}_j}[\parallel (1-\lambda)x_{\mathcal{N}_j}^{(j)} - \lambda x_0^{(j)} \parallel^2] + L\lambda^2\eta_j^2 \parallel e_j \parallel^2
\end{aligned} \tag{18}$$

where the last equation in Inq. 18 follows from Lemma A.2. The lemma substitutes $x_{\mathcal{N}_j}^{(j)}(x_0^{j})$ by $\tilde{x}_j(\tilde{x}_{j-1})$. ■

Lemma B.4 *Suppose $\eta_j^2L^2B_j < b_j^2$, then under Definition 2.3,*

$$\begin{aligned}
&(b_j - \frac{\eta_j^2L^2B_j}{b_j})\mathbb{E}[\parallel \tilde{x}_j - \tilde{x}_{j-1} \parallel^2] + 2\lambda\eta_jB_j\mathbb{E} < e_j, (\tilde{x}_j - \tilde{x}_{j-1}) > \\
&\leq -2\eta_j(1-\lambda)B_j\mathbb{E} < \nabla f(\tilde{x}_j), (\tilde{x}_j - \tilde{x}_{j-1}) > + 2(1-\lambda)^2\eta_j^2B_j\mathbb{E}[\parallel \nabla f(\tilde{x}_j) \parallel^2] + 2\lambda^2\eta_j^2B_j\mathbb{E}[\parallel e_j \parallel^2]
\end{aligned}$$

Proof *Since* $x^{(j)}_{k+1} = x^{(j)}_k - \eta_j v^{(j)}_k$, *we have*

$$
\begin{aligned}
&\mathbb{E}_{\tilde{\mathcal{I}}_k}[\| x^{(j)}_{k+1} - x^{(j)}_0 \|^2] \\
&= \| x^{(j)}_k - x^{(j)}_0 \|^2 - 2\eta_j < \mathbb{E}_{\tilde{\mathcal{I}}_k} v^{(j)}_k, (x^{(j)}_k - x^{(j)}_0) > + \eta_j^2 \mathbb{E}_{\tilde{\mathcal{I}}_k} \| v^{(j)}_k \|^2 \\
&= \| x^{(j)}_k - x^{(j)}_0 \|^2 - 2(1-\lambda)\eta_j < \nabla f(x^{(j)}_k), (x^{(j)}_k - x^{(j)}_0) > - 2\lambda\eta_j < e_j, (x^{(j)}_k - x^{(j)}_0) > + \eta_j^2 \mathbb{E}_{\tilde{\mathcal{I}}_k} \| v^{(j)}_k \|^2 \\
&\leq (1 + \frac{\eta_j^2 L^2}{b_j}) \| x^{(j)}_k - x^{(j)}_0 \|^2 - 2\eta_j(1-\lambda) < \nabla f(x^{(j)}_k), x^{(j)}_k - x^{(j)}_0 > \\
&- 2\lambda\eta_j < e_j, (x^{(j)}_k - x^{(j)}_0) > + 2(1-\lambda)^2\eta_j^2 \| \nabla f(x^{(j)}_k) \|^2 + 2\lambda^2\eta_j^2 \| e_j \|^2 .
\end{aligned}
\tag{19}
$$

where the last inequality follows from Lemma B.1. Using the same notation $\mathbb{E}_j$ *from Theorem 3.1 we have*

$$
\begin{aligned}
&2\eta_j(1-\lambda)\mathbb{E}_j < \nabla f(x^{(j)}_k), (x^{(j)}_k - x^{(j)}_0) > + 2\lambda\eta_j\mathbb{E}_j < e_j, (x^{(j)}_k - x^{(j)}_0) > \\
&\leq (1 + \frac{\eta_j^2 L^2}{b_j})\mathbb{E}_j \| x^{(j)}_k - x^{(j)}_0 \|^2 - \mathbb{E}_j \| x^{(j)}_{k+1} - x^{(j)}_0 \|^2 + 2(1-\lambda)^2\eta_j^2 \| \nabla f(x^{(j)}_k) \|^2 + 2\lambda\eta_j^2 \| e_j \|^2
\end{aligned}
\tag{20}
$$

Let $k = N_j$, *and using Fubini's theorem, we have,*

$$
\begin{aligned}
&2(1-\lambda)\eta_j\mathbb{E}_{N_j}\mathbb{E}_j < \nabla f(x^{(j)}_{N_j}), (x^{(j)}_{N_j} - x^{(j)}_0) > + 2\lambda\eta_j\mathbb{E}_{N_j}\mathbb{E}_j < e_j, (x^{(j)}_{N_j} - x^{(j)}_0) > \\
&\leq (1 + \frac{\eta_j L^2}{b_j})\mathbb{E}_{N_j}\mathbb{E}_j \| x^{(j)}_{N_j} - x^{(j)}_0 \|^2 - \mathbb{E}_{N_j}\mathbb{E}_j \| x^{(j)}_{N_j+1} - x^{(j)}_0 \|^2 \\
&+ 2(1-\lambda)^2\eta_j^2\mathbb{E}_{N_j} \| \nabla f(x^{(j)}_{N_j}) \|^2 + 2\lambda^2\eta_j^2 \| e_j \|^2 \\
&= (-\frac{b_j}{B_j} + \frac{\eta_j^2 L^2}{b_j})\mathbb{E}_{N_j}\mathbb{E}_j \| x^{(j)}_{N_j} - x^{(j)}_0 \|^2 + 2(1-\lambda)^2\eta_j^2\mathbb{E}_{N_j} \| \nabla f(x^{(j)}_{N_j}) \|^2 + 2\lambda^2\eta_j^2 \| e_j \|^2 .
\end{aligned}
\tag{21}
$$

The lemma is then proved by substituting $x^{(j)}_{N_j}(x^{(j)}_0)$ *by* $\tilde{x}_j(\tilde{x}_{j-1})$. ∎

Lemma B.5

$$b_j\mathbb{E} < e_j, (\tilde{x}_j - \tilde{x}_{j-1}) >= -\eta_j(1-\lambda)B_j\mathbb{E} < e_j, \nabla f(\tilde{x}_j) > - \lambda^2\eta_j B_j \mathbb{E} \| e_j \|^2$$

Proof *Let* $M^{(j)}_k =< e_j, (x^{(j)}_k - x^{(j)}_0) >$, *then we have*

$$\mathbb{E}_{N_j} < e_j, (\tilde{x}_j - \tilde{x}_{j-1}) >= \mathbb{E}_{N_j} M^{(j)}_{N_j}. \tag{22}$$

Since N_j *is independent of* $(x^{(j)}_0, e_j)$, *it has*

$$\mathbb{E} < e_j, (\tilde{x}_j - \tilde{x}_{j-1}) >= \mathbb{E} M^{(j)}_{N_j}. \tag{23}$$

Also $M^{(j)}_0 = 0$, *then we have*

$$
\begin{aligned}
&\mathbb{E}_{\tilde{\mathcal{I}}_k}(M^{(j)}_{k+1} - M^{(j)}_k) \\
&= \mathbb{E}_{\tilde{\mathcal{I}}_k} < e_j, (x^{(j)}_{k+1} - x^{(j)}_k) > \\
&= -\eta_j < e_j, \mathbb{E}_{\tilde{\mathcal{I}}_k}[v^{(j)}_k] > .
\end{aligned}
\tag{24}
$$

Using the same notation $\mathbb{E}_j$ *in Lemma B.3 and Lemma B.4, we have*

$$\mathbb{E}_j(M_{k+1}^{(j)} - M_k^{(j)}) = -\eta_j(1-\lambda) < e_j, \mathbb{E}_j \nabla f(x_k^{(j)}) > -\lambda^2 \eta_j \parallel e_j \parallel^2 . \tag{25}$$

Let $k = N_j$ *in Eq.25. Using Fubini's theorem and Lemma B.2, we have,*

$$\frac{b_j}{B_j}\mathbb{E}_{N_j} M_{N_j}^{(j)} = -\eta_j(1-\lambda) < e_j, \mathbb{E}_{N_j}\mathbb{E}_j \nabla f(x_k^{(j)}) > -\eta_j \parallel e_j \parallel^2 . \tag{26}$$

The lemma is then proved by substituting $x_{N_j}^{(j)}(x_0^{(j)})$ *by* $\tilde{x}_j(\tilde{x}_{j-1})$*.* ∎

Proof of Theorem 3.2

Theorem *Let* $\eta L = \gamma(\frac{b_j}{B_j})^{\alpha}$ *where* $0 \leq \alpha \leq 1$ *and* $\gamma \geq 0$*. Suppose* $B_j \geq b_j \geq B_j^{\beta}$ *(*$0 \leq \beta \leq 1$*) for all* j*, then under Definition 2.3, the output* $\tilde{x}_j$ *of Alg 2 satisfies*

$$\mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 \leq \frac{(\frac{2L}{\gamma})(\frac{b_j}{B_j})^{1-\alpha}\mathbb{E}(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)) + 2\lambda^4 \frac{I(B_j < n))}{B_j^{1-4\alpha}}\mathcal{S}^*}{2(1-\lambda) - (2\gamma B_j^{\alpha\beta-\alpha} + 2B_j^{\beta-1})(1-\lambda)^2 - 1.16(1-\lambda)^2}.$$

where $0 < \lambda < 1$ *and* $2(1-\lambda) - (2\gamma B_j^{\alpha\beta-\alpha} + 2B_j^{\beta-1})(1-\lambda)^2 - 1.16(1-\lambda)^2$ *is positive when* $B_j \leq 3$*,* $0 \leq \gamma \leq \frac{13}{50}$ *and* $0 < \lambda < 1$*.*

Proof *Multiplying Eq.B.3 by 2 and Eq.B.4 by* $\frac{b_j}{\eta_j B_j}$ *and summing them, then we have,*

$$\begin{aligned}
&2\eta_j B_j(1-\lambda)(1-(1-\lambda)L\eta_j - \frac{(1-\lambda)b_j}{B_j})\mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 + \frac{b_j^3 - \eta_j^2 L^2 b_j B_j - \eta_j^3 L^3 B_j^2}{b_j \eta_j B_j}\mathbb{E} \parallel \tilde{x}_j - \tilde{x}_{j-1} \parallel^2 \\
&+ 2\lambda\eta_j B_j \mathbb{E} < e_j, \nabla f(\tilde{x}_j) > + 2\lambda b_j \mathbb{E} < e_j, (\tilde{x}_j - \tilde{x}_{j-1}) > \\
&= 2\eta_j B_j(1-\lambda)(1-(1-\lambda)L\eta_j - \frac{(1-\lambda)b_j}{B_j})\mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 \\
&+ \frac{b_j^3 - (1-\lambda)^2\eta_j^2 L^2 b_j B_j - (1-\lambda)^2 \eta_j^3 L^3 B_j^2}{b_j \eta_j B_j}\mathbb{E} \parallel \tilde{x}_j - \tilde{x}_{j-1} \parallel^2 - 2\frac{\lambda^3}{(1-\lambda)}\eta_j B_j \mathbb{E} \parallel e_j \parallel^2 \text{ (Lemma B.5)} \\
&\leq -2(1-\lambda)b_j \mathbb{E} < \nabla f(\tilde{x}_j), (\tilde{x}_j - \tilde{x}_{j-1}) > + 2b_j \mathbb{E}(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)) + (2\lambda^2 L\eta_j^2 B_j + 2\lambda^2 \eta_j b_j)\mathbb{E} \parallel e_j \parallel^2
\end{aligned} \tag{27}$$

Using the fact that $2 < q, p > \leq \beta \parallel q \parallel^2 + \frac{1}{\beta} \parallel p \parallel^2$ *for any* $\beta > 0$*,* $-2(1-\lambda)b_j\mathbb{E} < \nabla f(\tilde{x}_j), (\tilde{x}_j - \tilde{x}_{j-1}) >$ *in Inq. 27 can be bounded as*

$$\begin{aligned}
&-2(1-\lambda)b_j \mathbb{E} < \nabla f(\tilde{x}_j), (\tilde{x}_j - \tilde{x}_{j-1}) > \\
&\leq (1-\lambda)(\frac{(1-\lambda)b_j \eta_j B_j}{b_j^3 - (1-\lambda)^2 \eta_j^2 L^2 b_j B_j - (1-\lambda)^2 \eta_j^3 L^3 B_j^2} b_j^2 \mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 \\
&+ \frac{b_j^3 - (1-\lambda)^2 \eta_j^2 L^2 b_j B_j - (1-\lambda)^2 \eta_j^3 L^3 B_j^2}{(1-\lambda) b_j \eta_j B_j}\mathbb{E} \parallel \tilde{x}_j - \tilde{x}_{j-1} \parallel^2)
\end{aligned} \tag{28}$$

Then Inq. 27 can be expressed as

$$\begin{aligned}&\frac{\eta_j B_j}{b_j}(2(1-\lambda)-2(1-\lambda)^2 L\eta_j-2(1-\lambda)^2\frac{b_j}{B_j}-\frac{(1-\lambda)^2 b_j^3}{b_j^3-(1-\lambda)^2\eta_j^2L^2b_jB_j-(1-\lambda)^2\eta_j^3L^3B_j^2})\\&\mathbb{E}\parallel\nabla f(\tilde{x}_j)\parallel^2\\&\leq 2\mathbb{E}(f(\tilde{x}_{j-1})-f(\tilde{x}_j))+\frac{2\eta_jB_j\lambda^2}{b_j}(\frac{\lambda^2}{(1-\lambda)}+\eta_jL+\frac{b_j}{B_j})\mathbb{E}\parallel e_j\parallel^2 .\end{aligned}\tag{29}$$

Since $\eta_j L=\gamma(\frac{b_j}{B_j})^\alpha$, $b_j\geq 1$ *and* $B_j\geq b_j\geq B_j^\beta$ *where* $\alpha>0$ *and* $\beta\geq 0$ *by Theorem 3.1, a one part in left hand side of above inequality can be simplified and positive as following:*

$$\begin{aligned}&b_j^3-(1-\lambda)^2\eta_j^2L^2b_jB_j-(1-\lambda)^2\eta_j^3L^3B_j^2\\&=b_j^3(1-(1-\lambda)^2\gamma^2\frac{b_j^{2\alpha-2}}{B_j^{2\alpha-1}}-(1-\lambda)^2\gamma^3\frac{b_j^{3\alpha-3}}{B_j^{3\alpha-2}})\\&\geq b_j^3(1-(1-\lambda)^2\gamma^2B_j^{-1}-(1-\lambda)^2\gamma^3B_j^{-1})\geq 0.86b_j^3\end{aligned}\tag{30}$$

By Eq.30, the left side of Inq. 29 can be simplified since the factor of geometry distribution $\gamma\geq 0$ *as*

$$\begin{aligned}&\frac{\eta_j B_j}{b_j}(2(1-\lambda)-2(1-\lambda)^2 L\eta_j-2(1-\lambda)^2\frac{b_j}{B_j}-\frac{(1-\lambda)^2 b_j^3}{b_j^3-(1-\lambda)^2\eta_j^2L^2b_jB_j-(1-\lambda)^2\eta_j^3L^3B_j^2})\\&\mathbb{E}\parallel\nabla f(\tilde{x}_j)\parallel^2\\&\geq\frac{\gamma}{L}B_j^{\alpha\beta-\alpha-\beta+1}\left(2(1-\lambda)-(2\gamma B_j^{\alpha\beta-\alpha}+2\frac{b_j}{B_j})(1-\lambda)^2-1.16(1-\lambda)^2\right)\mathbb{E}||\nabla f(\tilde{x}_j)||^2\\&\geq\frac{\gamma}{L}B_j^{\alpha\beta-\alpha-\beta+1}\left(2(1-\lambda)-(2\gamma+2)B_j^{-1}(1-\lambda)^2-1.16(1-\lambda)^2\right)\mathbb{E}||\nabla f(\tilde{x}_j)||^2\end{aligned}\tag{31}$$

Eq.31 is positive when $0\leq\gamma\leq\frac{13}{50}$ *and* $B_j\geq 3$. *Moreover, Lei et al. (2017a); Lei and Jordan (2017) determined the learning rate* $\eta=\frac{\gamma}{L}\frac{b_j}{B_j}\leq\frac{1}{3L}$ *that* $\gamma\leq\frac{1}{3}$ *which can guarantees the convergence in non-convex case. In our case,* $\gamma\leq\frac{13}{50}$ *satisfies within the range* $\gamma\leq\frac{1}{3}$. *Then Eq.29 can be simplified by Eq.31 as*

$$\begin{aligned}\mathbb{E}\parallel\nabla f(\tilde{x}_j)\parallel^2&\leq\frac{2\mathbb{E}[f(\tilde{x}_{j-1})-f(\tilde{x}_j)]+2\frac{\gamma}{L}B_j^{\alpha\beta-\alpha-\beta+1}\lambda^2(\frac{\lambda^2}{(1-\lambda)}+B_j^{\alpha\beta-\alpha}\gamma+B_j^{\beta-\alpha}L)\mathbb{E}||e_j||^2}{\frac{\gamma}{L}B_j^{\alpha\beta-\alpha-\beta+1}\left(2(1-\lambda)-(2\gamma B_j^{\alpha\beta-\alpha}+2B_j^{\beta-1})(1-\lambda)^2-1.16(1-\lambda)^2\right)}\\&\leq\frac{\overbrace{2\mathbb{E}(f(\tilde{x}_{j-1}-f(\tilde{x}_j)))}^{\text{positive by Lemma A.2}}+\overbrace{2\frac{\gamma}{L}\lambda^2B_j^{\alpha\beta-\alpha-\beta+1}B_j^{4\alpha}\mathbb{E}\parallel e_j\parallel^2}^{\text{positive}}}{\frac{\gamma}{L}B_j^{\alpha\beta-\alpha-\beta+1}\left(2(1-\lambda)-(2\gamma B_j^{\alpha\beta-\alpha}+2B_j^{\beta-1})(1-\lambda)^2-1.16(1-\lambda)^2\right)},\end{aligned}\tag{32}$$

Then, using Lemma B.2, Inq. 32 can be rewritten as

$$\mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 \leq \frac{2\mathbb{E}(f(\tilde{x}_{j-1} - f(\tilde{x}_j))) + 2\frac{\gamma}{L}\lambda^4 B_j^{\alpha\beta+3\alpha-\beta} I(B_j < n)\mathcal{S}^*}{\frac{\gamma}{L} B_j^{\alpha\beta-\alpha-\beta+1}\left(2(1-\lambda) - (2\gamma B_j^{\alpha\beta-\alpha} + 2B_j^{\beta-1})(1-\lambda)^2 - 1.16(1-\lambda)^2\right)}. \tag{33}$$

■

B.2. Biased Estimator Version

For the biased estimation version, we still start by bounding the gradient $\mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k^{(j)} \parallel^2$ in Lemma B.6 and the variance $\mathbb{E}_{\mathcal{I}_j} \parallel e_j \parallel^2$ in Lemma B.7.

Lemma B.6 *Under Definition 2.3,*

$$\mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k^{(j)} \parallel^2 \leq \frac{(1-\lambda)^2 L^2}{b_j} \parallel x_k^{(j)} - x_0^{(j)} \parallel + 2(1-\lambda)^2 \parallel \nabla f(x_k^{(j)}) \parallel^2 + 2 \parallel e_j \parallel^2 .$$

Proof *Using the fact that for a random variable Z* $\mathbb{E} \parallel Z \parallel^2 = \|Z - \mathbb{E}Z \parallel^2 + \parallel \mathbb{E}Z \parallel^2$, *we have*

$$\begin{aligned}
&\mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k^{(j)} \parallel^2 = \mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k^{(j)} - \mathbb{E}_{\tilde{\mathcal{I}}_k} v_k^{(j)} \parallel^2 + \parallel \mathbb{E}_{\tilde{\mathcal{I}}_k} v_k^{(j)} \parallel^2 \\
&= \mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel (1-\lambda)(\nabla f_{\tilde{\mathcal{I}}_k}(x_k^{(j)}) - \nabla f_{\tilde{\mathcal{I}}_k}(x_0^{(j)})) - (1-\lambda)(\nabla f(x_k^{(j)}) - \nabla f x_0^{(j)}) \parallel^2 \\
&+ \parallel (1-\lambda)\nabla f(x_k^{(j)}) + e_j \parallel^2 \\
&\leq (1-\lambda)^2 \mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel \nabla f_{\tilde{\mathcal{I}}_k}(x_k^{(j)}) - \nabla f_{\tilde{\mathcal{I}}_k}(x_0^{(j)}) - (\nabla f(x_k^{(j)}) - \nabla f x_0^{(j)}) \parallel^2 \\
&+ 2(1-\lambda)^2 \parallel \nabla f(x_k^{(j)}) \parallel^2 + 2 \parallel e_j \parallel^2 .
\end{aligned} \tag{34}$$

By Lemma A.1, the first part of inequality in Eq.34 can be rewritten as,

$$\begin{aligned}
&(1-\lambda)^2 \mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel \nabla f_{\tilde{\mathcal{I}}_k}(x_k^{(j)}) - \nabla f_{\tilde{\mathcal{I}}_k}(x_0^{(j)}) - (\nabla f(x_k^{(j)}) - \nabla f x_0^{(j)}) \parallel^2 \\
&\leq \frac{(1-\lambda)^2}{b_j} \cdot \frac{1}{n} \sum_{i=1}^{n} \parallel \nabla f_i(x_k^{(j)}) - \nabla f_i(x_0^{(j)}) - (\nabla f(x_k^{(j)}) - \nabla f(x_0^{(j)})) \parallel^2 \\
&= \frac{(1-\lambda)^2}{b_j} \cdot (\frac{1}{n} \sum_{i=1}^{n} \parallel \nabla f_i(x_k^{(j)}) - \nabla f_i(x_0^{(j)}) \parallel^2 - \parallel (\nabla f(x_k^{(j)}) - \nabla f(x_0^{(j)})) \parallel^2) \\
&\leq \frac{(1-\lambda)^2}{b_j} \cdot \frac{1}{n} \sum_{i=1}^{n} \parallel \nabla f_i(x_k^{(j)}) - \nabla f_i(x_0^{(j)}) \parallel^2 \\
&\leq \frac{(1-\lambda)^2}{b_j} \cdot L^2 \parallel x_k^{(j)} - x_0^{(j)} \parallel^2
\end{aligned} \tag{35}$$

where the last line is based on Definition 2.3, then the bound of the gradient can be written as,

$$\mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k^{(j)} \parallel^2 \leq \frac{(1-\lambda)^2 L^2}{b_j} \parallel x_k^{(j)} - x_0^{(j)} \parallel^2 + 2(1-\lambda)^2 \parallel \nabla f(x_k^{(j)}) \parallel^2 + 2 \parallel e_j \parallel^2 . \tag{36}$$

■

Lemma B.7

$$\mathbb{E}_{\mathcal{I}_j} \parallel e_j \parallel^2 \leq (1-\lambda)^2 \frac{I(B_j < n)}{B_j} \mathcal{S}^* + (1-2\lambda)^2 \mathbb{E}_{\mathcal{I}_j}[\nabla f_i(\tilde{x}_{j-1})]^2$$
$$= \mathbb{E}_{\mathcal{I}_j} \parallel \tilde{e}_j \parallel^2 + (1-2\lambda)^2 \mathbb{E}_{\mathcal{I}_j}[\nabla f_i(\tilde{x}_{j-1})]^2$$

where $(1-\lambda)^2 \frac{I(B_j < n)}{B_j} \mathcal{S}^* = \mathbb{E}_{\mathcal{I}_j} \parallel \tilde{e}_j \parallel^2$ *and* $0 < \lambda < 1$.

Proof *Based on Lemma A.1 and the observation that* $\tilde{x}_{j-1}$ *is independent of*

$$\begin{aligned}
\mathbb{E}_{\mathcal{I}_j} \parallel e_j \parallel^2 &= \frac{n-B_j}{(n-1)B_j} \cdot \frac{1}{n} \sum_{i=1}^{n} \parallel (1-\lambda)\nabla f_i(\tilde{x}_{j-1}) - \lambda \nabla f(\tilde{x}_{j-1}) \parallel^2 \\
&= \frac{n-B_j}{(n-1)B_j} \mathbb{E}_{\mathcal{I}_j} \parallel (1-\lambda)\nabla f_i(\tilde{x}_{j-1}) - \lambda \mathbb{E}_{\mathcal{I}_j}[\nabla f_i(\tilde{x}_{j-1})] \parallel^2 \\
&= \frac{n-B_j}{(n-1)B_j} \mathbb{E}_{\mathcal{I}_j} \left[(1-\lambda)^2 \nabla f_i(\tilde{x}_{j-1})^2 - (2\lambda - 3\lambda^2)\mathbb{E}_{\mathcal{I}_j}[\nabla f_i(\tilde{x}_{j-1})]^2\right] \\
&= \frac{n-B_j}{(n-1)B_j} \left[\underbrace{(1-\lambda)^2 \mathbb{E}_{\mathcal{I}_j}\left[\nabla f_i(\tilde{x}_{j-1})^2 - \mathbb{E}_{\mathcal{I}_j}[\nabla f_i(\tilde{x}_{j-1})]^2\right]}_{\text{Unbiased}} + \underbrace{(1-2\lambda)^2 \mathbb{E}_{\mathcal{I}_j}[\nabla f_i(\tilde{x}_{j-1})]^2}_{\text{Extra/term}}\right] \\
&= \frac{n-B_j}{(n-1)B_j} \cdot \left((1-\lambda)^2 \frac{1}{n} \sum_{i=1}^{n} \parallel \nabla f_i(\tilde{x}_{j-1}) - \nabla f(\tilde{x}_{j-1}) \parallel^2 + (1-2\lambda)^2 \mathbb{E}_{\mathcal{I}_j}[\nabla f_i(\tilde{x}_{j-1})]^2\right) \\
&\leq (1-\lambda)^2 \frac{n-B_j}{(n-1)B_j} \cdot \mathcal{S}^* + \frac{n-B_j}{(n-1)B_j}(1-2\lambda)^2 \mathbb{E}_{\mathcal{I}_j}[\nabla f_i(\tilde{x}_{j-1})]^2 \\
&\leq (1-\lambda)^2 \frac{I(B_j < n)}{B_j} \mathcal{S}^* + (1-2\lambda)^2 \mathbb{E}_{\mathcal{I}_j}[\nabla f_i(\tilde{x}_{j-1})]^2,
\end{aligned} \tag{37}$$

where the upper bound of the variance of the stochastic gradients $\mathcal{S}^* = \frac{1}{n}\sum_{i=1}^{n} \parallel \nabla f_i(\tilde{x}_{j-1}) - \nabla f(\tilde{x}_{j-1}) \parallel^2$*.In above function, as* $\nabla f(\tilde{x}_{j-1})$ *is the expectation value of* $\nabla f_i(\tilde{x}_{j-1})$*, we use* $\mathbb{E}_{\mathcal{I}_j}[\nabla f_i(\tilde{x}_{j-1})]$ *to alternative* $\nabla f(\tilde{x}_{j-1})$ *for easily understanding later proof. Meanwhile, We can achieve the third equation in above function since the fact that* $\mathbb{E}[(1-\lambda)Z - \lambda \mathbb{E}[Z]]^2 = (1-\lambda)^2 \mathbb{E}[Z^2] - (2\lambda - 3\lambda^2)\mathbb{E}[Z]^2 = \mathbb{E}[(1-\lambda)^2 Z^2 - (2\lambda - 3\lambda^2)\mathbb{E}[Z]^2]$. ■

Theorem 3.3 defines the bound of the batch-size, B_j, for the biased estimator case

Proof of Theorem 3.3

Theorem *If the expectation of the variance* $\mathbb{E}_{\mathcal{I}_j} \parallel e_j \parallel^2 \leq \sigma\rho^{2j}$ *in Alg 3* ($\sigma \geq 0$ *is a constant for some* $\rho < 1$) *and* $0 < \lambda < 1$*, the lower bound of the batch-size,* B_j*, can be expressed as,*

$$B_j \geq \frac{n\mathcal{S}^*}{\mathcal{S}^* + (1-\lambda)^2 n^{\frac{1}{2}} \sigma\rho^{2j}}$$

Proof *To define the bound of the batch-size,* B_j*, for the biased estimator case, we estimate the lower and upper bounds of the variance to control the size of the batch. Based on the* *result from Lemma B.7 and using the result that the norms of the gradients are bounded by* $\mathcal{K}^2$ *for all* x_j *(Babanezhad et al., 2015), we have*

$$\begin{aligned}
&\frac{1}{n-1}\sum_{i=1}^{n}[(1-\lambda)^2 \parallel \nabla f_i(\tilde{x}_{j-1}) \parallel^2 -\lambda^2 \parallel \nabla f(\tilde{x}_{j-1}) \parallel^2] \\
&\le (1-\lambda)^2\frac{1}{n-1}\sum_{i=1}^{n}[\parallel \nabla f_i(\tilde{x}_{j-1}) \parallel^2 - \parallel \nabla f(\tilde{x}_{j-1}) \parallel^2] + (1-2\lambda)^2\mathbb{E}_{\mathcal{I}_j}[\nabla f_i(\tilde{x}_{j-1})]^2 \\
&\le (1-\lambda)^2\mathcal{K}^2 + (1-2\lambda)^2\mathbb{E}_{\mathcal{I}_j}[\nabla f_i(\tilde{x}_{j-1})]^2,
\end{aligned} \tag{38}$$

and we use the same approach we applied in the unbiased case which is shown from Inq. 11 to 15 to achieve a bound of the batch size when $0 < \lambda < 1$*. The batch size can be bounded as,*

$$\begin{aligned}
B_j &\ge \frac{n\mathcal{K}^2}{\mathcal{K}^2 + (1-\lambda)^2 n\sigma\rho^{2j}} \ge \frac{n\frac{n}{\sqrt{n-1}}\mathcal{S}^*}{\frac{n}{\sqrt{n-1}}\mathcal{S}^* + (1-\lambda)^2 n\sigma\rho^{2j}} \\
&\ge \frac{n^2\mathcal{S}^*}{n\mathcal{S}^* + (1-\lambda)^2 n^{\frac{3}{2}}\sigma\rho^{2j}} = \frac{n\mathcal{S}^*}{\mathcal{S}^* + (1-\lambda)^2 n^{\frac{1}{2}}\sigma\rho^{2j}}.
\end{aligned} \tag{39}$$

■

Lemma B.8 *Suppose* $\eta_j L < 1$*, then under Definition 2.3,*

$$\begin{aligned}
&(1-\lambda)(1-(1-\lambda)L\eta_j)\eta_j B_j\mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 +\eta_j B_j \mathbb{E} < e_j, \nabla f(\tilde{x}_j) > \\
&\le b_j\mathbb{E}(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)) + \frac{(1-\lambda)^2\eta_j^2 B_j L^3}{2b_j}\mathbb{E} \parallel \tilde{x}_j - \tilde{x}_{j-1} \parallel^2 + L\eta_j^2 B_j \mathbb{E} \parallel e_j \parallel^2 .
\end{aligned}$$

where $\mathbb{E}$ *denotes the expectation with respect to all randomness.*

Proof *By Definition 2.3, we have*

$$\begin{aligned}
&\mathbb{E}_{\tilde{\mathcal{I}}_k}[f(x_{k+1}^{(j)})] \le f(x_k^{(j)}) - \eta_j < \mathbb{E}_{\tilde{\mathcal{I}}_k} v_k, \nabla f(x_k^{(j)}) > + \frac{L\eta_j^2}{2}\mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k \parallel^2 \\
&= f(x_k^{(j)}) - \eta_j < ((1-\lambda)\nabla f(x_k^{(j)}) + e_j), \nabla f(x)_k^{(j)}) > + \frac{L\eta_j^2}{2}\mathbb{E}_{\tilde{\mathcal{I}}_k} \parallel v_k \parallel^2 \\
&\le f(x_k^{(j)}) - \eta_j(1-\lambda) \parallel \nabla f(x_k^{(j)}) \parallel^2 - \eta_j < e_j, \nabla f(x_k^{(j)}) > \\
&+ \frac{L^3\eta_j^2(1-\lambda)^2}{2b_j} \parallel x_k^{(j)} - x_0^{(j)} \parallel^2 + L\eta_j^2(1-\lambda)^2 \parallel \nabla f(x_k^{(j)}) \parallel^2 + L\eta_j^2 \parallel e_j \parallel^2 \\
&= f(x_k^{(j)}) - (\eta_j(1-\lambda) - L\eta_j^2(1-\lambda)^2) \parallel \nabla f(x_k^{(j)}) \parallel^2 \\
&- \eta_j < e_j, \nabla f(x_k^{(j)}) > + \frac{L^3\eta_j^2(1-\lambda)^2}{2b_j} \parallel x_k^{(j)} - x_0^{(j)} \parallel^2 + L\eta_j^2 \parallel e_j \parallel^2
\end{aligned} \tag{40}$$

Let $\mathbb{E}_j$ *denote the expectation* $\tilde{\mathcal{I}}_0, \tilde{\mathcal{I}}_1, \ldots,$ *given* $\tilde{\mathcal{N}_j}$ *since* $\tilde{\mathcal{N}_j}$ *is independent of them and let* $k{=}\mathcal{N}_j$ *in Inq 40. As* $\tilde{\mathcal{I}}_{k+1}, \tilde{\mathcal{I}}_{k+2}, \ldots$ *are independent of* $x_k^{(j)}$ *and taking the expectation with* *respect to $\mathcal{N}_j$ and using Fubini's theorem, Inq. 40 implies that*

$$\begin{aligned}
&\eta_j(1-\lambda)(1-(1-\lambda)L\eta_j)\mathbb{E}_{\mathcal{N}_j}\mathbb{E}_j[\| \nabla f(x_{\mathcal{N}_j}^{(j)}) \|^2] + \eta_j\mathbb{E}_{\mathcal{N}_j}\mathbb{E}_j < e_j, \nabla f(x_{\mathcal{N}_j}^{(j)}) > \\
&\leq \mathbb{E}_{\mathcal{N}_j}(\mathbb{E}_j[f(x_{\mathcal{N}_j}^{(j)})] - \mathbb{E}_j[f(x_{\mathcal{N}_{j+1}}^{(j)})]) + \frac{L^3\eta_j^2(1-\lambda)^2}{2b_j}\mathbb{E}_{\mathcal{N}_j}\mathbb{E}_j\mathbb{E}[\| x_{\mathcal{N}_j}^{(j)} - x_0^{(j)} \|^2] + L\eta_j^2 \| e_j \|^2 \\
&= \frac{b_j}{B_j}(f(x_0^{(j)}) - \mathbb{E}_j\mathbb{E}_{\mathcal{N}_j}[f_{\mathcal{N}_j}^{(j)}]) + \frac{L^3\eta_j^2(1-\lambda)^2}{2b_j}\mathbb{E}_j\mathbb{E}_{\mathcal{N}_j}[\| x_{\mathcal{N}_j}^{(j)} - x_0^{(j)} \|^2] + L\eta_j^2 \| e_j \|^2
\end{aligned} \tag{41}$$

where the last equation in Inq. 41 follows from Lemma A.2. The lemma substitutes $x_{\mathcal{N}_j}^{(j)}(x_0^j)$ by $\tilde{x}_j(\tilde{x}_{j-1})$. ■

Lemma B.9 *Suppose $\eta_j^2L^2B_j < b_j^2$, then under Definition lsmooth1,*

$$\begin{aligned}
&(b_j - \frac{(1-\lambda)^2\eta_j^2L^2B_j}{b_j})\mathbb{E}[\| \tilde{x}_j - \tilde{x}_{j-1} \|^2] + 2\eta_jB_j\mathbb{E} < e_j, (\tilde{x}_j - \tilde{x}_{j-1}) > \\
&\leq -2(1-\lambda)\eta_jB_j\mathbb{E} < \nabla f(\tilde{x}_j), (\tilde{x}_j - \tilde{x}_{j-1}) > +2(1-\lambda)^2\eta_j^2B_j\mathbb{E}[\| \nabla f(\tilde{x}_j) \|^2] + 2\eta_j^2B_j\mathbb{E}[\| e_j \|^2]
\end{aligned}$$

Proof *Since $x_{k+1}^{(j)} = x_k^{(j)} - \eta_jv_k^{(j)}$, we have*

$$\begin{aligned}
&\mathbb{E}_{\tilde{\mathcal{I}}_k}[\| x_{k+1}^{(j)} - x_0^{(j)} \|^2] \\
&= \| x_k^{(j)} - x_0^{(j)} \|^2 - 2\eta_j < \mathbb{E}_{\tilde{\mathcal{I}}_k}v_k^{(j)}, (x_k^{(j)} - x_0^{(j)}) > +\eta_j^2\mathbb{E}_{\tilde{\mathcal{I}}_k} \| v_k^{(j)} \|^2 \\
&= \| x_k^{(j)} - x_0^{(j)} \|^2 - 2\eta_j(1-\lambda) < \nabla f(x_k^{(j)}), (x_k^{(j)} - x_0^{(j)}) > - 2\eta_j < e_j, (x_k^{(j)} - x_0^{(j)}) > +\eta_j^2\mathbb{E}_{\tilde{\mathcal{I}}_k} \| v_k^{(j)} \|^2 \\
&\leq (1 + \frac{(1-\lambda)^2\eta_j^2L^2}{b_j}) \| x_k^{(j)} - x_0^{(j)} \|^2 - 2\eta_j(1-\lambda) < \nabla f(x_k^{(j)}), x_k^{(j)} - x_0^{(j)} > - 2\eta_j < e_j, (x_k^{(j)} - x_0^{(j)}) > \\
&+ 2(1-\lambda)^2\eta_j^2 \| \nabla f(x_k^{(j)}) \|^2 + 2\eta_j^2 \| e_j \|^2 .
\end{aligned} \tag{42}$$

where the last inequality is based on Lemma B.6. Using the same notation $\mathbb{E}_j$ in Theorem 3.1 we have

$$\begin{aligned}
&2\eta_j(1-\lambda)\mathbb{E}_j < \nabla f(x_k^{(j)}), (x_k^{(j)} - x_0^{(j)}) > +2\eta_j\mathbb{E}_j < e_j, (x_k^{(j)} - x_0^{(j)}) > \\
&\leq (1 + \frac{(1-\lambda)^2\eta_j^2L^2}{b_j})\mathbb{E}_j \| x_k^{(j)} - x_0^{(j)} \|^2 - \mathbb{E}_j \| x_{k+1}^{(j)} - x_0^{(j)} \|^2 + 2(1-\lambda)^2\eta_j^2 \| \nabla f(x_k^{(j)}) \|^2 + 2\eta_j^2 \| e_j \|^2
\end{aligned} \tag{43}$$

Let $k = N_j$, and using Fubini's theorem, we have,

$$\begin{aligned}
&2\eta_j(1-\lambda)\mathbb{E}_{N_j}\mathbb{E}_j < \nabla f(x_{N_j}^{(j)}), (x_{N_j}^{(j)} - x_0^{(j)}) > +2\eta_j\mathbb{E}_{N_j}\mathbb{E}_j < e_j, (x_{N_j}^{(j)} - x_0^{(j)}) > \\
&\leq (1 + \frac{(1-\lambda)^2\eta_jL^2}{b_j})\mathbb{E}_{N_j}\mathbb{E}_j \| x_{N_j}^{(j)} - x_0^{(j)} \|^2 - \mathbb{E}_{N_j}\mathbb{E}_j \| x_{N_j+1}^{(j)} - x_0^{(j)} \|^2 \\
&+ 2(1-\lambda)^2\eta_j^2\mathbb{E}_{N_j} \| \nabla f(x_{N_j}^{(j)}) \|^2 + 2\eta_j^2 \| e_j \|^2 \\
&= (-\frac{b_j}{B_j} + \frac{(1-\lambda)^2\eta_j^2L^2}{b_j})\mathbb{E}_{N_j}\mathbb{E}_j \| x_{N_j}^{(j)} - x_0^{(j)} \|^2 + 2(1-\lambda)^2\eta_j^2\mathbb{E}_{N_j} \| \nabla f(x_{N_j}^{(j)}) \|^2 + 2\eta_j^2 \| e_j \|^2 .
\end{aligned} \tag{44}$$

The lemma is then proved by substituting $x_{N_j}^{(j)}(x_0^{(j)})$ *by* $\tilde{x}_j(\tilde{x}_{j-1})$. ■

Lemma B.10

$$b_j\mathbb{E} < e_j, (\tilde{x}_j - \tilde{x}_{j-1}) >= -\eta_j(1-\lambda)B_j\mathbb{E} < e_j, \nabla f(\tilde{x}_j) > -\eta_j B_j \mathbb{E} \parallel e_j \parallel^2$$

Proof *Let* $M_k^{(j)} =< e_j, (x_k^{(j)} - x_0^{(j)}) >$, *then we have*

$$\mathbb{E}_{N_j} < e_j, (\tilde{x}_j - \tilde{x}_{j-1}) >= \mathbb{E}_{N_j} M_{N_j}^{(j)}.$$

Since N_j *is independent of* $(x_0^{(j)}, e_j)$, *it has*

$$\mathbb{E} < e_j, (\tilde{x}_j - \tilde{x}_{j-1}) >= \mathbb{E}M_{N_j}^{(j)}. \tag{45}$$

Also $M_0^{(j)} = 0$, *then we have*

$$\begin{aligned} &\mathbb{E}_{\tilde{\mathcal{I}}_k}(M_{k+1}^{(j)} - M_k^{(j)}) \\ &= \mathbb{E}_{\tilde{\mathcal{I}}_k} < e_j, (x_{k+1}^{(j)} - x_k^{(j)}) >= -\eta_j < e_j, \mathbb{E}_{\tilde{\mathcal{I}}_k}[v_k^{(j)}] > \\ &= -\eta_j(1-\lambda) < e_j, \nabla f(x_k^{(j)}) > -\eta_j \parallel e_j \parallel^2 . \end{aligned} \tag{46}$$

Using the same notation $\mathbb{E}_j$ *in Theorem 3.1, we have*

$$\mathbb{E}_j(M_{k+1}^{(j)} - M_k^{(j)}) = -\eta_j(1-\lambda) < e_j, \mathbb{E}_j \nabla f(x_k^{(j)}) > -\eta_j \parallel e_j \parallel^2 . \tag{47}$$

Let $k = N_j$ *in Eq.47. Using Fubini's theorem and Lemma A.2, we have,*

$$\frac{b_j}{B_j}\mathbb{E}_{N_j} M_{N_j}^{(j)} = -\eta_j(1-\lambda) < e_j, \mathbb{E}_{N_j}\mathbb{E}_j \nabla f(x_k^{(j)}) > -\eta_j \parallel e_j \parallel^2 . \tag{48}$$

The lemma is then proved by substituting $x_{N_j}^{(j)}(x_0^{(j)})$ *by* $\tilde{x}_j(\tilde{x}_{j-1})$. ■

Proof of Theorem 3.4

Theorem *let* $\eta L = \gamma(\frac{b_j}{B_j})^\alpha$ $(0 < \alpha < 1)$ *and* $\gamma \leq \frac{1}{3}$. *Suppose* $\gamma \leq \frac{1}{3}$ *and* $B_j \geq b_j \geq B_j^\beta$ $(0 \leq \beta < 1)$ *for all* j, *then under Definition 2.3, the output* $\tilde{x}_j$ *of Alg 2 we have,*

$$\mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 \leq \frac{2\mathbb{E}[f(\tilde{x}_{j-1}) - f(\tilde{x}_j)] + 2(1-\lambda)^2 \frac{\gamma}{L} B_j^{\alpha\beta+3\alpha-\beta} I(B_j < n)\mathcal{S}^*}{\frac{\gamma}{L} B_j^{1-\alpha+\alpha\beta-\beta}\left(2(1-\lambda) - (2\gamma B_j^{\alpha\beta-\alpha} + 2B_j^{\beta-1} - 4LB_j^{2\alpha-2})(1-\lambda)^2 - 1.16(1-\lambda)^2\right)},$$

where $0 < \lambda < 1$.

Proof *Multiplying Eq.B.8 by 2 and Eq.B.9 by* $\frac{b_j}{\eta_j B_j}$ *and summing them, then we have,*

$$
\begin{aligned}
&2\eta_j B_j(1-\lambda)(1-(1-\lambda)L\eta_j - \frac{(1-\lambda)b_j}{B_j})\mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 \\
&+ \frac{b_j^3 - (1-\lambda)^2\eta_j^2 L^2 b_j B_j - (1-\lambda)^2\eta_j^3 L^3 B_j^2}{b_j\eta_j B_j}\mathbb{E} \parallel \tilde{x}_j - \tilde{x}_{j-1} \parallel^2 \\
&+ 2\eta_j B_j \mathbb{E} < e_j, \nabla f(\tilde{x}_j) > +2b_j\mathbb{E} < e_j, (\tilde{x}_j - \tilde{x}_{j-1}) > \\
&= 2\eta_j B_j(1-\lambda)(1-(1-\lambda)L\eta_j - \frac{(1-\lambda)b_j}{B_j} + \frac{(2\lambda-1)^2}{2\eta_j B_j(1-\lambda)})\mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 \\
&+ \frac{b_j^3 - (1-\lambda)^2\eta_j^2 L^2 b_j B_j - (1-\lambda)^2\eta_j^3 L^3 B_j^2}{b_j\eta_j B_j}\mathbb{E} \parallel \tilde{x}_j - \tilde{x}_{j-1} \parallel^2 - 2\eta_j B_j \mathbb{E} \parallel \tilde{e}_j \parallel^2 \ (\textit{ Lemma B.10}) \\
&\leq -2(1-\lambda)b_j\mathbb{E} < \nabla f(\tilde{x}_j), (\tilde{x}_j - \tilde{x}_{j-1}) > +2b_j\mathbb{E}(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)) + (2L\eta_j^2 B_j + 2\eta_j b_j)\mathbb{E} \parallel \tilde{e}_j \parallel^2
\end{aligned} \tag{49}
$$

Using the fact that $2 < q, p > \leq \beta \parallel q \parallel^2 + \frac{1}{\beta} \parallel p \parallel^2$ *for any* $\beta > 0$, $-2b_j\mathbb{E} < \nabla f(\tilde{x}_j), (\tilde{x}_j - \tilde{x}_{j-1}) >$ *in Inq. 49 can be bounded as*

$$
\begin{aligned}
&-2(1-\lambda)b_j\mathbb{E} < \nabla f(\tilde{x}_j), (\tilde{x}_j - \tilde{x}_{j-1}) > \\
&\leq (1-\lambda)(\frac{(1-\lambda)b_j\eta_j B_j}{b_j^3 - (1-\lambda)^2\eta_j^2 L^2 b_j B_j - (1-\lambda)^2\eta_j^3 L^3 B_j^2} b_j^2\mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 \\
&+ \frac{b_j^3 - (1-\lambda)^2\eta_j^2 L^2 b_j B_j - (1-\lambda)^2\eta_j^3 L^3 B_j^2}{(1-\lambda)b_j\eta_j B_j}\mathbb{E} \parallel \tilde{x}_j - \tilde{x}_{j-1} \parallel^2)
\end{aligned} \tag{50}
$$

Then Inq. 49 can be rewritten as

$$
\begin{aligned}
&\frac{\eta_j B_j}{b_j}(2(1-\lambda) - 2(1-\lambda)^2 L\eta_j - 2(1-\lambda)^2\frac{b_j}{B_j} + \frac{(2\lambda-1)^2}{\eta_j B_j} \\
&- \frac{(1-\lambda)^2 b_j^3}{b_j^3 - (1-\lambda)^2\eta_j^2 L^2 b_j B_j - (1-\lambda)^2\eta_j^3 L^3 B_j^2})\mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 \\
&\leq 2\mathbb{E}(f(\tilde{x}_{j-1}) - f(\tilde{x}_j)) + \frac{2\eta_j B_j}{b_j}(1 + \eta_j L + \frac{b_j}{B_j})\mathbb{E} \parallel \tilde{e}_j \parallel^2 .
\end{aligned} \tag{51}
$$

Since $\eta_j L = \gamma(\frac{b_j}{B_j})^{\alpha}$, $b_j \geq 1$ *and* $B_j \geq b_j \geq B_j^{\beta}$ *where* $0 < \alpha \leq 1, 0 \leq \beta \leq 1$, *we have*

$$
\begin{aligned}
&b_j^3 - (1-\lambda)^2\eta_j^2 L^2 b_j B_j - (1-\lambda)^2\eta_j^3 L^3 B_j^2 \\
&= b_j^3(1 - (1-\lambda)^2\gamma^2\frac{b_j^{2\alpha-2}}{B_j^{2\alpha-1}} - (1-\lambda)^2\gamma^3\frac{b_j^{3\alpha-3}}{B_j^{3\alpha-2}}) \\
&= b_j^3(1 - (1-\lambda)^2\gamma^2 B_j^{-1} - (1-\lambda)^2\gamma^3 B_j^{-1}) \geq 0.86b_j^3
\end{aligned} \tag{52}
$$

By Eq. 52, the left side of Inq. 51 can be simplified as

$$\frac{\eta_j B_j}{b_j}(2(1-\lambda)-2(1-\lambda)^2 L\eta_j-2(1-\lambda)^2\frac{b_j}{B_j}+\frac{(2\lambda-1)^2}{\eta_j B_j}-\frac{(1-\lambda)^2 b_j^3}{b_j^3-\eta_j^2L^2b_jB_j-\eta_j^3L^3B_j^2})\mathbb{E}\parallel\nabla f(\tilde{x}_j)\parallel^2$$
$$=\frac{\gamma}{L}B_j^{1-\alpha+\alpha\beta-\beta}\left(2(1-\lambda)-(2\gamma B_j^{\alpha\beta-\alpha}+2B_j^{\beta-1})(1-\lambda)^2+\frac{(2\lambda-1)^2}{\frac{\gamma}{L}B_j^{2\alpha-2}}-1.16(1-\lambda)^2\right)\mathbb{E}\parallel\nabla f(\tilde{x}_j)\parallel^2$$
$$\geq\frac{\gamma}{L}B_j^{\alpha\beta-\alpha-\beta+1}\left(2(1-\lambda)-(2\gamma B_j^{-1}+2B_j^{-1}-4)(1-\lambda)^2-1.16(1-\lambda)^2\right)\mathbb{E}\parallel\nabla f(\tilde{x}_j)\parallel^2. \tag{53}$$

Eq.53 is positive when $0\leq\gamma\leq 2.42B_j-1$ *and* $B_j\geq 1$*. Moreover, Lei et al. (2017a); Lei and Jordan (2017) determined the learning rate* $\eta=\frac{\gamma}{L}\frac{b_j}{B_j}\leq\frac{1}{3L}$ *that* $\gamma\leq\frac{1}{3}$ *which can guarantees the convergence in non-convex case. In our case,* γ *should satisfy the range* $0\leq\gamma\leq\frac{1}{3}\leq 2.42B_j-1$*, thus* $\gamma\leq\frac{1}{3}$.

Then Eq.51 can be simplified by Eq.53 as

$$\mathbb{E}\parallel\nabla f(\tilde{x}_j)\parallel^2\leq\frac{2\mathbb{E}[f(\tilde{x}_{j-1})-f(\tilde{x}_j)]+2\frac{\gamma}{L}B_j^{\alpha\beta-\alpha-\beta+1}(1+B_j^{\alpha\beta-\alpha}\gamma+B_j^{b-a}L)\mathbb{E}\parallel e_j\parallel^2}{\frac{\gamma}{L}B_j^{1-\alpha+\alpha\beta-\beta}\left(2(1-\lambda)-(2\gamma B_j^{\alpha\beta-\alpha}+2B_j^{\beta-1}-4LB_j^{2\alpha-2})(1-\lambda)^2-1.16(1-\lambda)^2\right)}$$
$$\leq\frac{\overbrace{2\mathbb{E}[f(\tilde{x}_{j-1})-f(\tilde{x}_j)]}^{\text{positive by Lemma A.2}}+\overbrace{2\frac{\gamma}{L}B_j^{\alpha\beta-\alpha-\beta+1}B_j^{4a}\mathbb{E}\parallel e_j\parallel^2}^{\text{positive}}}{\frac{\gamma}{L}B_j^{1-\alpha+\alpha\beta-\beta}\left(2(1-\lambda)-(2\gamma B_j^{\alpha\beta-\alpha}+2B_j^{\beta-1}-4LB_j^{2\alpha-2})(1-\lambda)^2-1.16(1-\lambda)^2\right)}. \tag{54}$$

Then, using Lemma B.7, Inq. 54 can be expressed as

$$\mathbb{E}\parallel\nabla f(\tilde{x}_j)\parallel^2\leq\frac{2\mathbb{E}[f(\tilde{x}_{j-1})-f(\tilde{x}_j)]+2(1-\lambda)^2\frac{\gamma}{L}B_j^{\alpha\beta+3\alpha-\beta}I(B_j<n)\mathcal{S}^*}{\frac{\gamma}{L}B_j^{1-\alpha+\alpha\beta-\beta}\left(2(1-\lambda)-(2\gamma B_j^{\alpha\beta-\alpha}+2B_j^{\beta-1}-4LB_j^{2\alpha-2})(1-\lambda)^2-1.16(1-\lambda)^2\right)}, \tag{55}$$

■

Appendix C. Convergence Analysis for L-smooth Objectives

Proof of Theorem 3.5

Theorem *Under the specifications of Theorem 3.2, Theorem 3.4 and Definition 2.3, the output* $\tilde{x}_T^*$ *can achieve its upper bound of gradients depending on two estimators.*

- *For the unbiased estimator (Alg. 2),* $0<\lambda<1$*. The upper bound is given by,*

$$\mathbb{E}\parallel\nabla f(\tilde{x}_T^*)\parallel^2\leq\frac{(\frac{2L}{\gamma})\triangle_f}{\theta\sum_{j=1}^T b_j^{\alpha-1}B_j^{1-\alpha}}+\frac{2\lambda^4 I(B_j<n))\mathcal{S}^*}{\theta B_j^{1-4\alpha}},$$

- *For the biased estimator (Alg. 3),* $0 < \lambda < 1$. *The upper bound is shown as,*

$$\mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 \leq \frac{(\frac{2L}{\gamma})\triangle_f}{\theta_{\text{biased}} \sum_{j=1}^{T} b_j^{\alpha-1} B_j^{1-\alpha}} + \frac{2(1-\lambda)^2 I(B_j < n))\mathcal{S}^*}{\theta_{\text{biased}} B_j^{1-4\alpha}},$$

where $\theta = 2(1-\lambda) - (2\gamma B_j^{\alpha\beta-\alpha} + 2B_j^{\beta-1})(1-\lambda)^2 - 1.16(1-\lambda)^2 > 0$, *and* $\theta_{\text{biased}} = 2(1-\lambda) - (2\gamma B_j^{\alpha\beta-\alpha} + 2B_j^{\beta-1} - 4LB_j^{2\alpha-2})(1-\lambda)^2 - 1.16(1-\lambda)^2$.
Proof *Since* $\tilde{x}_T^*$ *is a random element from* $(\tilde{x}_j)_{j=1}^T$ *with*

$$P(\tilde{x}_T^* = \tilde{x}_j) \propto \frac{\eta_j B_j}{b_j} \propto (\frac{B_j}{b_j})^\alpha, \tag{56}$$

Inq. 33 and 55 will be re-scaled as Inq. 57 and 58 respectively.

- *For the unbiased estimator (Alg. 2), the upper bound is shown as,*

$$\mathbb{E} \parallel \nabla f(\tilde{x}_T^*) \parallel^2 \leq \frac{(\frac{2L}{\gamma})\triangle_f}{\theta \sum_{j=1}^{T} b_j^{\alpha-1} B_j^{1-\alpha}} + \frac{2\lambda^4 I(B_j < n))\mathcal{S}^*}{\theta B_j^{1-4\alpha}}, \tag{57}$$

  where $\theta = 2(1-\lambda) - (2\gamma B_j^{\alpha\beta-\alpha} + 2B_j^{\beta-1})(1-\lambda)^2 - 1.16\lambda^2$.

- *For the biased estimator (Alg. 3), the upper bound is shown as,*

$$\mathbb{E} \parallel \nabla f(\tilde{x}_j) \parallel^2 \leq \frac{(\frac{2L}{\gamma})\triangle_f}{\theta_{\text{biased}} \sum_{j=1}^{T} b_j^{\alpha-1} B_j^{1-\alpha}} + \frac{(1-\lambda)^2 I(B_j < n))\mathcal{S}^*}{\theta_{\text{biased}} B_j^{1-4\alpha}}, \tag{58}$$

  where $\theta_{\text{biased}} = 2(1-\lambda) - (2\gamma B_j^{\alpha\beta-\alpha} + 2B_j^{\beta-1} - 4LB_j^{2\alpha-2})(1-\lambda)^2 - 1.16(1-\lambda)^2$.

■