Appendix

Lemma 3. *let f have Lipschitz gradient with a constant $L > 0$ and let $\{\mathbf{x}\}_{k\geq 0}$ be generated by* (8)*, we have*

$$\sup_k \|\mathbf{x}^k - \mathbf{x}^{k-1}\| \leq \frac{4\alpha\mu LR}{\delta(L+\mu)^2}, \tag{9}$$

and

$$\sum_{k=1}^{K} \|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2 \leq \frac{16KL^2\alpha^2\mu^2R^2}{\delta^2(L+\mu)^4}. \tag{10}$$

Proof. With the updating equation (4), we can derive that

$$\|\alpha(1-\beta_k)\mathbf{g}^k\| \geq |\ \|\mathbf{x}^{k+1} - \mathbf{x}^k\| - \|\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})\|\ | \geq |\ \|\mathbf{x}^{k+1} - \mathbf{x}^k\| - \|(1-\delta)(\mathbf{x}^k - \mathbf{x}^{k-1})\|\ |, \tag{11}$$

it can be inferred that $1 - \beta_k$ is decreasing with given β_k. Thus we have

$$1 - \beta_k \leq 1 - (\frac{1-\alpha\mu}{1+\alpha\mu})^2_{\alpha=\frac{1}{L}} = 1 - (\frac{L-\mu}{L+\mu})^2 = \frac{4L\mu}{(L+\mu)^2}, \tag{12}$$

then we derive

$$\|\alpha(1-\beta_k)\mathbf{g}^k\| \leq \frac{4\alpha\mu LR}{\delta(L+\mu)^2}.$$

Using the Mathematical Induction (MI) method, we have in the case of $k = 1$,

$$\|\mathbf{x}^1 - \mathbf{x}^0\| \leq \frac{4\alpha\mu LR}{\delta(L+\mu)^2}.$$

For any $k \geq 1$, we can infer that

$$\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \leq \frac{4\alpha\mu LR}{\delta(L+\mu)^2},$$

then we derive (9). Square on both sides of (11),

$$\begin{aligned}
\|\alpha(1-\beta_k)\mathbf{g}^k\|^2 &\geq \alpha\delta\|\mathbf{g}^k\|^2 \\
&\geq (\|\mathbf{x}^{k+1} - \mathbf{x}^k\| - \|(1-\delta)(\mathbf{x}^k - \mathbf{x}^{k-1})\|)^2 \\
&= \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 - 2(1-\delta)\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \\
&\times \|\mathbf{x}^k - \mathbf{x}^{k-1}\| + (1-\delta)^2\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2 \\
&\geq \delta\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 - \delta(1-\delta)\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2.
\end{aligned}$$

By summing the above from $k = 1$ to K, we then can get (10) □

Lemma 4. *Letf have Lipschitz gradient with a constant $L > 0$ and let $\{\mathbf{x}\}_{k\geq 0}$ be generated by* (8)*, we have*

$$\sum_{k=1}^{K} \beta_k \mathbb{E}\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \leq \frac{(1-\delta)L}{\delta} \sum_{k=1}^{K} \mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2. \tag{13}$$

Proof. We have

$$\begin{aligned}
&\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \\
&= \langle \nabla f(\mathbf{x}^{k-1}), \mathbf{x}^k - \mathbf{x}^{k-1}\rangle + \langle \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}), \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \\
&\leq \langle \nabla f(\mathbf{x}^{k-1}, \mathbf{x}^k - \mathbf{x}^{k-1}\rangle + L\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2 \\
&= \langle \nabla f(\mathbf{x}^{k-1}), -\alpha\mathbf{g}^{k-1} + \beta_{k-1}(\mathbf{x}^{k-1} - \mathbf{x}^{k-2})\rangle + L\|\mathbf{x}^k - \mathbf{x}^{k-1}\|.
\end{aligned}$$

Taking expectations on both sides, we then get

$$\begin{aligned}
&\mathbb{E}\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \\
&\leq -\alpha\mathbb{E}\|\nabla f(\mathbf{x}^{k-1})\|^2 + \beta_{k-1}\mathbb{E}\langle f(\mathbf{x}^{k-1}), \mathbf{x}^{k-1} - \mathbf{x}^{k-2}\rangle + L\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2 \\
&\leq \beta_{k-1}\mathbb{E}\langle f(\mathbf{x}^{k-1}), \mathbf{x}^{k-1} - \mathbf{x}^{k-2}\rangle + L\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2.
\end{aligned}$$

Multiplying β_k on both sides, we induct

$$\begin{aligned}
&\beta_k\mathbb{E}\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \\
&\leq \beta_k\beta_{k-1}\mathbb{E}\langle f(\mathbf{x}^{k-1}), \mathbf{x}^{k-1} - \mathbf{x}^{k-2}\rangle + L\beta_k\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2 \\
&\leq L\sum_{i=1}^{k-1}(\prod_{j=i}^{k-1}\beta_j)\mathbb{E}\|\mathbf{x}^i - \mathbf{x}^{i-1}\|^2,
\end{aligned}$$

with the fact that $\beta_k \leq 1-\delta$ for each $k = 0, 1, \dots K$, we derive

$$\beta_k \mathbb{E}\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \leq L \sum_{i=1}^{k} (1-\delta)^{k+1-i} \mathbb{E}\|\mathbf{x}^i - \mathbf{x}^{i-1}\|^2.$$

Thus, if we add the above term from $k = 1$ to K, we have

$$\begin{aligned}
&\sum_{k=1}^{K} \beta_k \mathbb{E}\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \\
&\leq L \sum_{k=1}^{K} \sum_{i=1}^{k} (1-\delta)^{k+1-i} \mathbb{E}\|\mathbf{x}^i - \mathbf{x}^{i-1}\|^2 \\
&\leq \frac{(1-\delta)L}{\delta} \sum_{k=1}^{K} \mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2,
\end{aligned}$$

then we get (13). □

Lemma 5. *Let f have Lipschitz gradient with a constant $L > 0$ and let $\{\mathbf{x}\}_{k\geq 0}$ be generated by* (8)*, we have*

$$\sum_{k=1}^{K} \mathbb{E}\langle \mathbf{x}^k - \mathbf{x}^*, \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \leq \frac{(1-\delta)}{\delta} \sum_{k=1}^{K} \mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2. \tag{14}$$

Proof. Using (4), we can have

$$\begin{aligned}
&\langle \mathbf{x}^k - \mathbf{x}^*, \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \\
&= \langle \mathbf{x}^{k-1} - \mathbf{x}^*, \mathbf{x}^k - \mathbf{x}^{k-1}\rangle + \langle \mathbf{x}^k - \mathbf{x}^{k-1}, \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \\
&= -\alpha(1-\beta_{k-1})\langle \mathbf{x}^{k-1} - \mathbf{x}^*, \mathbf{g}^{k-1}\rangle + \beta_{k-1}\langle \mathbf{x}^{k-1} - \mathbf{x}^*, \mathbf{x}^{k-1} - \mathbf{x}^{k-2}\rangle + \|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2.
\end{aligned}$$

Taking expectations on both sides, we derive that

$$\mathbb{E}\langle \mathbf{x}^k - \mathbf{x}^*, \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \leq \mathbb{E}\beta_{k-1}\langle \mathbf{x}^{k-1} - \mathbf{x}^*, \mathbf{x}^{k-1} - \mathbf{x}^{k-2}\rangle + \mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2.$$

Multiplying β_k on both sides and the fact that $\beta_k \leq 1-\delta$ for each k leads to

$$\begin{aligned}
&\mathbb{E}\beta_k\langle \mathbf{x}^k - \mathbf{x}^*, \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \\
&\leq \mathbb{E}\beta_k\beta_{k-1}\langle \mathbf{x}^{k-1} - \mathbf{x}^*, \mathbf{x}^{k-1} - \mathbf{x}^{k-2}\rangle + \mathbb{E}\beta_k\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2 \\
&\leq \sum_{i=1}^{k-1} (\prod_{j=i}^{k-1} \beta_j) \mathbb{E}\|\mathbf{x}^i - \mathbf{x}^{i-1}\|^2 \\
&\leq \sum_{i=1}^{k} (1-\delta)^{k+1-i} \mathbb{E}\|\mathbf{x}^i - \mathbf{x}^{i-1}\|^2.
\end{aligned}$$

Summing the above term from $k = 1$ to K, we have

$$\begin{aligned}
&\sum_{k=1}^{K} \beta_k \mathbb{E}\langle \mathbf{x}^k - \mathbf{x}^*, \mathbf{x}^k - \mathbf{x}^{k-1}\rangle \\
&\leq \sum_{k=1}^{K} \sum_{i=1}^{k} (1-\delta)^{k+1-i} \mathbb{E}\|\mathbf{x}^i - \mathbf{x}^{i-1}\|^2 \\
&\leq \frac{(1-\delta)}{\delta} \sum_{k=1}^{K} \mathbb{E}\|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2.
\end{aligned}$$

□

Proof of Theorem 1.

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2 = \mathbb{E}\langle \mathbf{x}^{k+1}-\mathbf{x}^*, \mathbf{x}^{k+1}-\mathbf{x}^*\rangle \\
&= \mathbb{E}\|\mathbf{x}^k - \alpha(1-\beta_k)\mathbf{g}^k + \beta_k(\mathbf{x}^k-\mathbf{x}^{k-1})\|^2 \\
&\leq \mathbb{E}(\|\mathbf{x}^k-\mathbf{x}^*\|^2 - 2\alpha(1-\beta_k)\langle \mathbf{x}^k-\mathbf{x}^*, \mathbf{g}^k\rangle + 2\beta_k\langle \mathbf{x}^k-\mathbf{x}^*, \mathbf{x}^k-\mathbf{x}^{k-1}\rangle + \|\mathbf{x}^k-\mathbf{x}^{k-1}\|^2 + \alpha^2\|\mathbf{g}^k\|^2).
\end{aligned} \tag{15}
$$

Using MI method from (5) we have

$$
\beta_k\langle \mathbf{x}^k-\mathbf{x}^*, \mathbf{x}^k-\mathbf{x}^{k-1}\rangle \leq \frac{16(1-\delta)L^2R^2\alpha^4}{\delta(L+\alpha)^2}.
$$

Leveraging convexity of f we derive

$$
\begin{aligned}
\mathbb{E}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2 &\leq \mathbb{E}(1-2\alpha\delta\mu)\|\mathbf{x}^k-\mathbf{x}^*\|^2 + \frac{32(1-\delta)L^2R^2\alpha^4}{\delta(L+\alpha)^2} + \frac{16L^2R^2\alpha^4}{(L+\alpha)^2} + \alpha^2R^2 \\
&= \mathbb{E}(1-2\alpha\delta\mu)^k\|\mathbf{x}^1-\mathbf{x}^*\|^2 + \mathcal{O}(\alpha).
\end{aligned}
$$

□

Proof of Theorem 2. Note that

$$
\mathbb{E}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2 \leq \mathbb{E}(\|\mathbf{x}^k-\mathbf{x}^*\|^2 - 2\alpha(1-\beta_k)(f(\mathbf{x}^k)-\min f) + 2\beta_k\langle \mathbf{x}^k-\mathbf{x}^*, \mathbf{x}^k-\mathbf{x}^{k-1}\rangle + \|\mathbf{x}^k-\mathbf{x}^{k-1}\|^2 + \alpha^2\|\mathbf{g}^k\|^2).
$$

We have that

$$
2\alpha(1-\beta_k)\mathbb{E}(f(\mathbf{x}^k)-\min f) \leq \mathbb{E}\|\mathbf{x}^k-\mathbf{x}^*\|^2 - \mathbb{E}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|^2 + 2\beta_k\langle \mathbf{x}^k-\mathbf{x}^*, \mathbf{x}^k-\mathbf{x}^{k-1}\rangle + \|\mathbf{x}^k-\mathbf{x}^{k-1}\|^2 + \alpha^2\|\mathbf{g}^k\|^2.
$$

Summing the above from $k=1$ to K then leverage (5) and (12) we can derive that

$$
\begin{aligned}
2\alpha(1-\beta_k)\sum_{k=1}^{K}\mathbb{E}(f(\mathbf{x}^k)-f(\mathbf{x}^*)) &\leq \frac{8L\alpha^2}{(L+\alpha)^2}\sum_{k=1}^{K}(\mathbb{E}f(\mathbf{x}^k)-f(\mathbf{x}^*)) \\
&\leq \mathbb{E}\|\mathbf{x}^1-\mathbf{x}^*\|^2 + 2LR^2\alpha^2 + \frac{(L+\alpha)^2KR^2}{8L}.
\end{aligned}
$$

From the property of f that convexity, we then derive

$$
\mathbb{E}f(\frac{\sum_{k=1}^{K}\mathbf{x}^k}{K}) - f(\mathbf{x}^*) = \frac{(L+\alpha)^2}{8KL\alpha^2}\mathbb{E}\|\mathbf{x}^1-\mathbf{x}^*\|^2 + \mathcal{O}(\alpha).
$$

□

Proof of Theorem 3. From the Lipschitz condition

$$
\begin{aligned}
f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1}-\mathbf{x}^k\rangle + \frac{L}{2}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2 \\
&= f(\mathbf{x}^k) - \langle \nabla f(\mathbf{x}^k), \alpha(1-\beta_k)\mathbf{g}^k\rangle + \beta_k\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k-\mathbf{x}^{k-1}\rangle + \frac{L}{2}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2.
\end{aligned}
$$

Taking expectations on both sides, we can get

$$
\mathbb{E}f(\mathbf{x}^{k+1}) \leq \mathbb{E}f(\mathbf{x}^k) - \alpha\delta\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 + \mathbb{E}\beta_k\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k-\mathbf{x}^{k-1}\rangle + \frac{L}{2}\mathbb{E}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2.
$$

Perform an equation transformation

$$
\alpha\delta\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \leq \mathbb{E}[f(\mathbf{x}^k)-f(\mathbf{x}^{k+1})] + \mathbb{E}\beta_k\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k-\mathbf{x}^{k-1}\rangle + \frac{L}{2}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2.
$$

Summing the above from $k=1$ to K then leverage Lemma 11 and Lemma 13

$$
\alpha\delta\sum_{k=1}^{K}\|\nabla f(\mathbf{x}^k)\|^2 = \mathbb{E}f(\mathbf{x}^1) - \min f + \mathcal{O}(\alpha^2).
$$

Therefore we lead to

$$
\min_{1\leq k\leq K}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 = \frac{f(\mathbf{x}^1)-\min f}{\alpha\delta K} + \mathcal{O}(\alpha).
$$

□