Assumption 1. ***(Bounded variance)*** *Assume that the DSP stochastic gradient $\mathcal{G}(x;\xi)$ satisfies:*

$$Var\left[\mathcal{G}(x;\xi)\right] \leq \sigma^2.$$

Assumption 2. ***(Lipschitz continuous gradient)*** *Assume that the loss and the output of the blocks have Lipschitz continuous gradient, that is, $\forall k \in \{0,1,..,K-1\}$, and $\forall (x_{0,1},...,x_{k,1}),(x_{0,2},...,x_{k,2}) \in \mathbb{R}^{d_0+d_1+...+d_k}$,*

$$\left\|\nabla F\left(h_0;x_{0,1};...;x_{k,1}\right)-\nabla F(h_0;x_{0,2};...;x_{k,2})\right\| \leq L_k\left\|(x_{0,1},...,x_{k,1})-(x_{0,2},...,x_{k,2})\right\|,$$

and $\forall x_1,\ x_2 \in \mathbb{R}^d$,

$$\left\|\nabla f(x_1)-\nabla f(x_2)\right\| \leq L_K\left\|x_1-x_2\right\|.$$

Assumption 3. ***(Bounded error gradient)*** *Assume that the norm of the error gradient that a block receives is bounded, that is, for any $x \in \mathbb{R}^d$, $\forall k \in \{0,1,...,K-2\}$,*

$$\left\|\frac{\partial f_{k+1}(h_{k+1};x_{k+1})}{\partial h_{k+1}}...\frac{\partial f_{K-1}(h_{K-1};x_{K-1})}{\partial h_{K-1}}\frac{\partial \mathcal{L}(h_K,l)}{\partial h_K}\right\| \leq M \quad and \quad \left\|\frac{\partial \mathcal{L}(h_K,l)}{\partial h_K}\right\| \leq M.$$

1 Basic Lemmas

Lemma 1. *If Assumptions 2 and 3 hold, the difference between DSP gradient and BP gradient regarding the parameters of block k satisfies:*

$$\left\|\nabla_{x_k}\mathcal{L}(F(h_0;x_0^{t_0};...;x_{K-1}^{t_{K-1}}),y)-\mathcal{G}_{x_k}(x_0^{t_{2K-1}};...;x_{K-1}^{t_K})\right\| \leq LM\sum_{i=k}^{K-1}\left\|x_i^{t_{2K-1-i}}-x_i^{t_i}\right\|.$$

Proof. We gradually move the DSP gradient of the block k towards the BP gradient by replacing one block's backward parameters with its forward parameters at a time. $K-k$ steps in total are needed, and each step will introduce an error. After all the replacement is done, it becomes the BP gradient at the forward parameters. Firstly we replace $x_k^{t_{2K-1-k}}$ with $x_k^{t_k}$, and calculate the error introduced as follows,

$$\begin{aligned}
\|\Delta_k\| = &\left\|\left(\frac{\partial F(h_0;x_0^{t_0};...;x_{k-1}^{t_{k-1}};x_k^{t_{2K-1-k}})}{\partial x_k^{t_{2K-1-k}}}-\frac{\partial F(h_0;x_0^{t_0};...;x_{k-1}^{t_{k-1}};x_k^{t_k})}{\partial x_k^{t_k}}\right)\cdot\right.\\
&\frac{\partial F(h_0;x_0^{t_0};...;x_k^{t_k};x_{k+1}^{t_{2K-2-k}})}{\partial F(h_0;x_0^{t_0};...;x_k^{t_k})}\cdots\frac{\partial F(h_0;x_0^{t_0};...;x_{K-2}^{t_{K-2}};x_{K-1}^{t_K})}{\partial F(h_0;x_0^{t_0};...;x_{K-2}^{t_{K-2}})}\cdot\\
&\left.\frac{\partial \mathcal{L}\left(F(h_0;x_0^{t_0};...;x_{K-1}^{t_{K-1}}),l\right)}{\partial F(h_0;x_0^{t_0};...;x_{K-1}^{t_{K-1}})}\right\|\\
\leq &\left\|\frac{\partial F(h_0;x_0^{t_0};...;x_{k-1}^{t_{k-1}};x_k^{t_{2K-1-k}})}{\partial x_k^{t_{2K-1-k}}}-\frac{\partial F(h_0;x_0^{t_0};...;x_{k-1}^{t_{k-1}};x_k^{t_k})}{\partial x_k^{t_k}}\right\|\cdot\\
&\left\|\frac{\partial F(h_0;x_0^{t_0};...;x_k^{t_k};x_{k+1}^{t_{2K-2-k}})}{\partial F(h_0;x_0^{t_0};...;x_k^{t_k})}\cdots\frac{\partial F(h_0;x_0^{t_0};...;x_{K-2}^{t_{K-2}};x_{K-1}^{t_K})}{\partial F(h_0;x_0^{t_0};...;x_{K-2}^{t_{K-2}})}\cdot\right.\\
&\left.\frac{\partial \mathcal{L}\left(F(h_0;x_0^{t_0};...;x_{K-1}^{t_{K-1}}),l\right)}{\partial F(h_0;x_0^{t_0};...;x_{K-1}^{t_{K-1}})}\right\|\\
\leq &LM\left\|x_k^{t_{2K-1-k}}-x_k^{t_k}\right\|.
\end{aligned}$$

Secondly we replace $x_{k+1}^{t_{2K-2-k}}$ with $x_{k+1}^{t_{k+1}}$, and calculate the error introduced,

$$
\begin{aligned}
\|\Delta_{k+1}\| = & \left\| \left(\frac{\partial F(h_0; x_0^{t_0}; ...; x_k^{t_k}; x_{k+1}^{t_{2K-2-k}})}{\partial x_k^{t_k}} - \frac{\partial F(h_0; x_0^{t_0}; ...; x_k^{t_k}; x_{k+1}^{t_{k+1}})}{\partial x_k^{t_k}} \right) \cdot \right. \\
& \frac{\partial F(h_0; x_0^{t_0}; ...; x_{k+1}^{t_{k+1}}; x_{k+2}^{t_{2K-3-k}})}{\partial F(h_0; x_0^{t_0}; ...; x_{k+1}^{t_{k+1}})} \cdots \frac{\partial F(h_0; x_0^{t_0}; ...; x_{K-2}^{t_{K-2}}; x_{K-1}^{t_K})}{\partial F(h_0; x_0^{t_0}; ...; x_{K-2}^{t_{K-2}})} \cdot \\
& \left. \frac{\partial \mathcal{L}\left(F(h_0; x_0^{t_0}; ...; x_{K-1}^{t_{K-1}}), l\right)}{\partial F(h_0; x_0^{t_0}; ...; x_{K-1}^{t_{K-1}})} \right\| \\
\leq & \left\| \frac{\partial F(h_0; x_0^{t_0}; ...; x_k^{t_k}; x_{k+1}^{t_{2K-2-k}})}{\partial x_k^{t_k}} - \frac{\partial F(h_0; x_0^{t_0}; ...; x_k^{t_k}; x_{k+1}^{t_{k+1}})}{\partial x_k^{t_k}} \right\| \cdot \\
& \left\| \frac{\partial F(h_0; x_0^{t_0}; ...; x_{k+1}^{t_{k+1}}; x_{k+2}^{t_{2K-3-k}})}{\partial F(h_0; x_0^{t_0}; ...; x_{k+1}^{t_{k+1}})} \cdots \frac{\partial F(h_0; x_0^{t_0}; ...; x_{K-2}^{t_{K-2}}; x_{K-1}^{t_K})}{\partial F(h_0; x_0^{t_0}; ...; x_{K-2}^{t_{K-2}})} \right. \cdot \\
& \left. \frac{\partial \mathcal{L}\left(F(h_0; x_0^{t_0}; ...; x_{K-1}^{t_{K-1}}), l\right)}{\partial F(h_0; x_0^{t_0}; ...; x_{K-1}^{t_{K-1}})} \right\| \\
\leq & LM \left\| x_{k+1}^{t_{2K-2-k}} - x_{k+1}^{t_{k+1}} \right\|.
\end{aligned}
$$

We repeatedly perform the above procedure, until we get the error in the last step,

$$
\begin{aligned}
\|\Delta_{K-1}\| = & \left\| \left(\frac{\partial F(h_0; x_0^{t_0}; ...; x_{K-2}^{t_{K-2}}; x_{K-1}^{t_K})}{\partial x_k^{t_k}} - \frac{\partial F(h_0; x_0^{t_0}; ...; x_{K-2}^{t_{K-2}}; x_{K-1}^{t_{K-1}})}{\partial x_k^{t_k}} \right) \cdot \right. \\
& \left. \frac{\partial \mathcal{L}\left(F(h_0; x_0^{t_0}; ...; x_{K-1}^{t_{K-1}}), l\right)}{\partial F(h_0; x_0^{t_0}; ...; x_{K-1}^{t_{K-1}})} \right\| \\
\leq & \left\| \frac{\partial F(h_0; x_0^{t_0}; ...; x_{K-2}^{t_{K-2}}; x_{K-1}^{t_K})}{\partial x_k^{t_k}} - \frac{\partial F(h_0; x_0^{t_0}; ...; x_{K-2}^{t_{K-2}}; x_{K-1}^{t_{K-1}})}{\partial x_k^{t_k}} \right\| \cdot \\
& \left\| \frac{\partial \mathcal{L}\left(F(h_0; x_0^{t_0}; ...; x_{K-1}^{t_{K-1}}), l\right)}{\partial F(h_0; x_0^{t_0}; ...; x_{K-1}^{t_{K-1}})} \right\| \\
\leq & LM \left\| x_{K-1}^{t_K} - x_{K-1}^{t_{K-1}} \right\|.
\end{aligned}
$$

Add them together and we will have

$$
\begin{aligned}
& \left\| \nabla_{x_k} \mathcal{L}(F(h_0; x_0^{t_0}; x_1^{t_1}; ...; x_{K-1}^{t_{K-1}}), l) - \mathcal{G}_{x_k}(x_0^{t_{2K-1}}; x_1^{t_{2K-2}}; ...; x_{K-1}^{t_K}) \right\| \\
= & \|\Delta_k + \Delta_{k+1} + ... + \Delta_{K-1}\| \\
\leq & \|\Delta_k\| + \|\Delta_{k+1}\| + ... + \|\Delta_{K-1}\| \\
\leq & LM \sum_{i=k}^{K-1} \left\| x_i^{t_{2K-1-i}} - x_i^{t_i} \right\|.
\end{aligned}
$$

□

Lemma 2. *Assume Assumption 2 and 3 exist. The second moment of the difference between DSP and BP gradient satisfies,*

$$
\left\| \nabla f(x_0^{t_0}; ...; x_{K-1}^{t_{K-1}}) - \mathcal{G}(x_0^{t_{2K-1}}; ...; x_{K-1}^{t_K}) \right\|^2 \leq \frac{1}{2} L^2 c_0 \sum_{k=0}^{K-1} \frac{k+1}{K+1} \left\| x_k^{t_{2K-1-k}} - x_k^{t_k} \right\|^2.
$$

Proof. Via summation of Lemma 1 we can get,

$$\left\|\nabla f(x_0^{t_0};x_1^{t_1};...;x_{K-1}^{t_{K-1}})-\mathcal{G}(x_0^{t_{2K-1}};x_1^{t_{2K-2}};...;x_{K-1}^{t_K})\right\|\leq LM\sum_{k=0}^{K-1}(k+1)\left\|x_k^{t_{2K-1-k}}-x_k^{t_k}\right\|.$$

Then we have,

$$\begin{aligned}
&\left\|\nabla f(x_0^{t_0};x_1^{t_1};...;x_{K-1}^{t_{K-1}})-\mathcal{G}(x_0^{t_{2K-1}};x_1^{t_{2K-2}};...;x_{K-1}^{t_K})\right\|^2\\
&\leq L^2M^2\left(\sum_{k=0}^{K-1}(k+1)\left\|x_k^{t_{2K-1-k}}-x_k^{t_k}\right\|\right)^2\\
&=L^2M^2\left(\sum_{k=0}^{K-1}(k+1)\right)^2\left(\sum_{k=0}^{K-1}\frac{k+1}{\sum_{k=0}^{K-1}(k+1)}\left\|x_k^{t_{2K-1-k}}-x_k^{t_k}\right\|\right)^2\\
&\leq L^2M^2\left(\sum_{k=0}^{K-1}(k+1)\right)^2\sum_{k=0}^{K-1}\frac{k+1}{\sum_{k=0}^{K-1}(k+1)}\left\|x_k^{t_{2K-1-k}}-x_k^{t_k}\right\|^2\\
&=\frac{1}{2}L^2M^2K(K+1)\sum_{k=0}^{K-1}(k+1)\left\|x_k^{t_{2K-1-k}}-x_k^{t_k}\right\|^2.
\end{aligned}$$

□

2 DSP with SGD

Theorem 1. *Assume Assumptions 1, 2 and 3 hold. Let* $c_0=M^2K(K+1)^2$, *and* $c_1=-(\Delta t^2+2)+\sqrt{(\Delta t^2+2)^2+2c_0\Delta t^2}$. *If the learning rate* $\alpha_n\leq\frac{c_1}{Lc_0\Delta t^2}$, *then*

$$\frac{\sum_{n=0}^{N-1}\alpha_n\mathbb{E}\left\|\nabla f(x^{n'})\right\|^2}{\sum_{n=0}^{N-1}\alpha_n}\leq\frac{2\left[f(x^0)-f^*\right]}{\sum_{n=0}^{N-1}\alpha_n}+\frac{L\sigma^2(2+K\Delta t^2+\frac{1}{4}Kc_1)\sum_{n=0}^{N-1}\alpha_n^2}{\sum_{n=0}^{N-1}\alpha_n}.$$

Proof. According to Lipschitz continuous, we have

$$\begin{aligned}
&f(x^{n+1})-f(x^n)\leq\left\langle\nabla f(x^n),x^{n+1}-x^n\right\rangle+\frac{L}{2}\left\|x^{n+1}-x^n\right\|^2\\
&=-\alpha_n\left\langle\nabla f(x^n),\mathcal{G}(x^n;\xi)\right\rangle+\frac{L\alpha_n^2}{2}\left\|\mathcal{G}(x^n;\xi)\right\|^2\\
&=-\alpha_n\left\langle\nabla f(x^n)-\nabla f(x^{n'}),\mathcal{G}(x^n;\xi)\right\rangle-\alpha_n\left\langle\nabla f(x^{n'}),\mathcal{G}(x^n;\xi)\right\rangle+\frac{L\alpha_n^2}{2}\left\|\mathcal{G}(x^n;\xi)\right\|^2\\
&\leq\frac{1}{2L}\left\|\nabla f(x^n)-\nabla f(x^{n'})\right\|^2+\frac{L\alpha_n^2}{2}\left\|\mathcal{G}(x^n;\xi)\right\|^2-\alpha_n\left\langle\nabla f(x^{n'}),\mathcal{G}(x^n;\xi)\right\rangle\\
&\quad+\frac{L\alpha_n^2}{2}\left\|\mathcal{G}(x^n;\xi)\right\|^2\\
&\leq\frac{L}{2}\left\|x^n-x^{n'}\right\|^2-\alpha_n\left\langle\nabla f(x^{n'}),\mathcal{G}(x^n;\xi)\right\rangle+L\alpha_n^2\left\|\mathcal{G}(x^n;\xi)\right\|^2.
\end{aligned}$$

Take expectation regarding ξ on both sides,

$$
\begin{aligned}
&\mathbb{E}\left[f(x^{n+1})\right] - f(x^n) \leq \frac{L}{2}\left\|x^n - x^{n'}\right\|^2 - \alpha_n\left\langle \nabla f(x^{n'}), \mathcal{G}(x^n)\right\rangle + L\alpha_n^2\mathbb{E}\left\|\mathcal{G}(x^n;\xi)\right\|^2 \\
&= \frac{L}{2}\left\|x^n - x^{n'}\right\|^2 + \frac{\alpha_n}{2}\left(\left\|\nabla f(x^{n'}) - \mathcal{G}(x^n)\right\|^2 - \left\|\nabla f(x^{n'})\right\|^2 - \left\|\mathcal{G}(x^n)\right\|^2\right) \\
&\quad + L\alpha_n^2\left(\left\|\mathcal{G}(x^n)\right\|^2 + \mathrm{Var}\left[\mathcal{G}(x^n;\xi)\right]\right) \\
&\leq \frac{L}{2}\left\|x^n - x^{n'}\right\|^2 + \frac{\alpha_n}{2}\left\|\nabla f(x^{n'}) - \mathcal{G}(x^n)\right\|^2 - \left(\frac{\alpha_n}{2} - L\alpha_n^2\right)\left\|\mathcal{G}(x^n)\right\|^2 \\
&\quad - \frac{\alpha_n}{2}\left\|\nabla f(x^{n'})\right\|^2 + L\alpha_n^2\sigma^2 \\
&\leq \sum_{k=0}^{K-1}\left[\frac{L}{2} + \frac{1}{4}\alpha_n L^2 M^2 K(K+1)(k+1)\right]\left\|x_k^n - x_k^{n'}\right\|^2 - \left(\frac{\alpha_n}{2} - L\alpha_n^2\right)\left\|\mathcal{G}(x^n)\right\|^2 \\
&\quad - \frac{\alpha_n}{2}\left\|\nabla f(x^{n'})\right\|^2 + L\alpha_n^2\sigma^2.
\end{aligned}
$$

The last inequality utilizes Lemma 2. Consider the first term and take expectation,

$$
\begin{aligned}
\mathbb{E}\left\|x_k^n - x_k^{n'}\right\|^2 &= \mathbb{E}\left\|\sum_{i=n-\Delta t_k}^{n-1} -\alpha_i \mathcal{G}_{x_k}(x^i;\xi)\right\|^2 \\
&\leq \Delta t_k \sum_{i=n-\Delta t_k}^{n-1}\alpha_i^2\mathbb{E}\left\|\mathcal{G}_{x^k}(x^i;\xi)\right\|^2 \\
&\leq \Delta t \sum_{i=n-\Delta t}^{n-1}\alpha_i^2\left(\left\|\mathcal{G}_{x_k}(x^i)\right\|^2 + \sigma^2\right).
\end{aligned}
$$

Take the total expectation and perform summation for it,

$$
\begin{aligned}
&\sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\left(\frac{L}{2} + \frac{1}{4}\alpha_n L^2 M^2 K(K+1)(k+1)\right)\mathbb{E}\left\|x_k^n - x_k^{n'}\right\|^2 \\
&\leq \sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\left(\frac{L}{2} + \frac{1}{4}\alpha_n L^2 M^2 K(K+1)(k+1)\right)\Delta t\sum_{i=n-\Delta t}^{n-1}\alpha_i^2\left(\mathbb{E}\left\|\mathcal{G}_{x_k}(x^i)\right\|^2 + \sigma^2\right) \\
&\leq \sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\left(\frac{L}{2} + \frac{1}{4}\alpha_n L^2 M^2 K(K+1)(k+1)\right)\Delta t\cdot\Delta t\cdot\alpha_n^2\left(\mathbb{E}\left\|\mathcal{G}_{x_k}(x^n)\right\|^2 + \sigma^2\right).
\end{aligned}
$$

Take the total expectation and perform summation for all the terms,

$$
\begin{aligned}
&\mathbb{E}\left[f(x^N)\right] - f(x^0) \\
&\leq \sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\left(\frac{L}{2}+\frac{1}{4}\alpha_n L^2M^2K(K+1)(k+1)\right)\Delta t^2\alpha_n^2\left(\mathbb{E}\left\|\mathcal{G}_{x_k}(x^n)\right\|^2+\sigma^2\right) \\
&\quad - \sum_{n=0}^{N-1}\left(\frac{\alpha_n}{2}-L\alpha_n^2\right)\mathbb{E}\sum_{k=0}^{K-1}\left\|\mathcal{G}_{x_k}(x^n)\right\|^2 - \sum_{n=0}^{N-1}\frac{\alpha_n}{2}\mathbb{E}\left\|\nabla f(x^{n'})\right\|^2 + L\sigma^2\sum_{n=0}^{N-1}\alpha_n^2 \\
&= \sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\left(\left(\frac{L}{2}+\frac{1}{4}\alpha_n L^2M^2K(K+1)(k+1)\right)\Delta t^2\alpha_n^2 - \frac{\alpha_n}{2}+L\alpha_n^2\right)\mathbb{E}\left\|\mathcal{G}_{x_k}(x^n)\right\|^2 \\
&\quad + \sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\left(\frac{L}{2}+\frac{1}{4}\alpha_n L^2M^2K(K+1)(k+1)\right)\Delta t^2\alpha_n^2\sigma^2 - \sum_{n=0}^{N-1}\frac{\alpha_n}{2}\mathbb{E}\left\|\nabla f(x^{n'})\right\|^2 \\
&\quad + L\sigma^2\sum_{n=0}^{N-1}\alpha_n^2 \\
&\leq \sum_{n=0}^{N-1}\sum_{k=0}^{K-1}\frac{1}{4}\alpha_n\left(L^2M^2K(K+1)^2\Delta t^2\alpha_n^2 + \left(2\Delta t^2+4\right)L\alpha_n - 2\right)\mathbb{E}\left\|\mathcal{G}_{x^k}(x^n)\right\|^2 \\
&\quad + \sum_{n=0}^{N-1}\left(\frac{1}{2}LK+\frac{1}{8}\alpha_n L^2M^2K^2(K+1)^2\right)\Delta t^2\alpha_n^2\sigma^2 - \sum_{n=0}^{N-1}\frac{\alpha_n}{2}\mathbb{E}\left\|\nabla f(x^{n'})\right\|^2 + L\sigma^2\sum_{n=0}^{N-1}\alpha_n^2 \\
&\leq \sum_{n=0}^{N-1}\left(\frac{1}{2}LK+\frac{1}{8}\alpha_n L^2M^2K^2(K+1)^2\right)\Delta t^2\alpha_n^2\sigma^2 - \sum_{n=0}^{N-1}\frac{\alpha_n}{2}\mathbb{E}\left\|\nabla f(x^{n'})\right\|^2 + L\sigma^2\sum_{n=0}^{N-1}\alpha_n^2.
\end{aligned}
$$

The last inequality utilizes the restriction on the learning rate. Then we have

$$
\begin{aligned}
&\frac{\sum_{n=0}^{N-1}\alpha_n\mathbb{E}\left\|\nabla f(x^{n'})\right\|^2}{\sum_{n=0}^{N-1}\alpha_n} \\
&\leq \frac{2\left[f(x^0)-f^*\right]}{\sum_{n=0}^{N-1}\alpha_n} + \frac{L\sigma^2\sum_{n=0}^{N-1}\alpha_n^2\left[2+K\Delta t^2+\frac{1}{4}\alpha_n LM^2K^2(K+1)^2\Delta t^2\right]}{\sum_{n=0}^{N-1}\alpha_n}.
\end{aligned}
$$

□

3 DSP with Momentum SGD

The SUM method also implies the following recursions,

$$
\begin{aligned}
x^{n+1}+\frac{\beta}{1-\beta}v^{n+1} &= x^n+\frac{\beta}{1-\beta}v^n-\frac{\alpha}{1-\beta}\mathcal{G}(x^n;\xi), \quad n\geq 0 \\
v^{n+1} &= \beta v^n+((1-\beta)s-1)\alpha\mathcal{G}(x^n;\xi), \quad n\geq 0.
\end{aligned} \tag{1}
$$

where v^n is given by

$$
v^n=\begin{cases} x^n-x^{n-1}+s\alpha\mathcal{G}(x^{n-1};\xi), & n\geq 1 \\ 0, & n=0. \end{cases} \tag{2}
$$

Let $z^n = x^n+\frac{\beta}{1-\beta}v^n$.

Lemma 3. *Assume Assumption 1 exists. Let* $c_2=\frac{((1-\beta)s-1)^2}{(1-\beta)^2}$*, then*

$$
\sum_{n=0}^{N-1}\mathbb{E}\left\|v^n\right\|^2 \leq c_2\alpha^2\sum_{n=0}^{N-1}\mathbb{E}\left\|\mathcal{G}(x^n)\right\|^2 + c_2\sigma^2\alpha^2 N.
$$

Proof. Let $\hat{\alpha} = ((1-\beta)s - 1)\alpha$. From Eq. (1),

$$v^{n+1} = \beta v^n + \hat{\alpha}\mathcal{G}(x^n;\xi).$$

Note that $v^0 = 0$. Then

$$v^n = \hat{\alpha}\sum_{i=0}^{n-1}\beta^{n-1-i}\mathcal{G}(x^i;\xi).$$

Then we have,

$$\begin{aligned}
\mathbb{E}\left\|v^n\right\|^2 &= \hat{\alpha}^2\mathbb{E}\left\|\sum_{i=0}^{n-1}\beta^{n-1-i}\mathcal{G}(x^i;\xi)\right\|^2 = \hat{\alpha}^2\left(\sum_{i=0}^{n-1}\beta^{n-1-i}\right)^2\mathbb{E}\left\|\sum_{i=0}^{n-1}\frac{\beta^{n-1-i}}{\sum_{i=0}^{n-1}\beta^{n-1-i}}\mathcal{G}(x^i;\xi)\right\|^2 \\
&\le \hat{\alpha}^2\left(\sum_{i=0}^{n-1}\beta^{n-1-i}\right)^2\sum_{i=0}^{n-1}\frac{\beta^{n-1-i}}{\sum_{i=0}^{n-1}\beta^{n-1-i}}\mathbb{E}\left\|\mathcal{G}(x^i;\xi)\right\|^2 \\
&= \hat{\alpha}^2\sum_{i=0}^{n-1}\beta^{n-1-i}\sum_{i=0}^{n-1}\beta^{n-1-i}\left\|\mathcal{G}(x^i)\right\|^2 + \hat{\alpha}^2\sigma^2\left(\sum_{i=0}^{n-1}\beta^{n-1-i}\right)^2 \\
&\le \frac{\hat{\alpha}^2}{1-\beta}\sum_{i=0}^{n-1}\beta^{n-1-i}\left\|\mathcal{G}(x^i)\right\|^2 + \frac{\hat{\alpha}^2\sigma^2}{(1-\beta)^2} \\
&= (1-\beta)c_2\alpha^2\sum_{i=0}^{n-1}\beta^{n-1-i}\left\|\mathcal{G}(x^i)\right\|^2 + c_2\alpha^2\sigma^2.
\end{aligned}$$

Take the total expectation and perform summation,

$$\begin{aligned}
\sum_{n=0}^{N-1}\mathbb{E}\left[\left\|v^n\right\|^2\right] &\le (1-\beta)c_2\alpha^2\sum_{n=0}^{N-1}\sum_{i=0}^{n-1}\beta^{n-1-i}\mathbb{E}\left\|\mathcal{G}(x^i)\right\|^2 + c_2\alpha^2\sigma^2 N \\
&= (1-\beta)c_2\alpha^2\sum_{i=0}^{N-2}\sum_{n=i+1}^{N-1}\beta^{n-1-i}\mathbb{E}\left\|\mathcal{G}(x^i)\right\|^2 + c_2\alpha^2\sigma^2 N \\
&= (1-\beta)c_2\alpha^2\sum_{i=0}^{N-2}\frac{1-\beta^{N-1-i}}{1-\beta}\mathbb{E}\left\|\mathcal{G}(x^i)\right\|^2 + c_2\alpha^2\sigma^2 N \\
&\le c_2\alpha^2\sum_{n=0}^{N-2}\mathbb{E}\left\|\mathcal{G}(x^n)\right\|^2 + c_2\sigma^2\alpha^2 N \le c_2\alpha^2\sum_{n=0}^{N-1}\mathbb{E}\left\|\mathcal{G}(x^n)\right\|^2 + c_2\sigma^2\alpha^2 N.
\end{aligned}$$

□

Lemma 4. *Assume Assumption 1 exists, then*

$$\sum_{n=0}^{N-1}\mathbb{E}\left\|x^n - x^{n'}\right\|^2 \le 2\Delta t^2(c_2+s^2)\alpha^2\sum_{n=0}^{N-1}\mathbb{E}\left\|\mathcal{G}(x^n)\right\|^2 + 2\Delta t^2\sigma^2(c_2+s^2)\alpha^2 N.$$

Proof. First take expectation regarding ξ,

$$
\begin{aligned}
\mathbb{E}\left\|x^n - x^{n'}\right\|^2 &= \sum_{k=0}^{K-1} \mathbb{E}\left\|x_k^n - x_k^{n'}\right\|^2 = \sum_{k=0}^{K-1} \mathbb{E}\left\|\sum_{i=n-\Delta t_k}^{n-1} v_k^{i+1} - s\alpha \mathcal{G}_{x_k}(x^i;\xi)\right\|^2 \\
&\le \sum_{k=0}^{K-1} \Delta t_k \sum_{i=n-\Delta t_k}^{n-1} \mathbb{E}\left\|v_k^{i+1} - s\alpha \mathcal{G}_{x_k}(x^i;\xi)\right\|^2 \\
&\le \sum_{k=0}^{K-1} 2\Delta t_k \sum_{i=n-\Delta t_k}^{n-1} \left(\mathbb{E}\left\|v_k^{i+1}\right\|^2 + s^2\alpha^2 \mathbb{E}\left\|\mathcal{G}_{x_k}(x^i;\xi)\right\|^2\right) \\
&\le \sum_{k=0}^{K-1} 2\Delta t \sum_{i=n-\Delta t}^{n-1} \left(\mathbb{E}\left\|v_k^{i+1}\right\|^2 + s^2\alpha^2 \mathbb{E}\left\|\mathcal{G}_{x_k}(x^i;\xi)\right\|^2\right) \\
&= 2\Delta t \sum_{i=n-\Delta t}^{n-1} \left(\mathbb{E}\left\|v^{i+1}\right\|^2 + s^2\alpha^2 \mathbb{E}\left\|\mathcal{G}(x^i;\xi)\right\|^2\right) \\
&\le 2\Delta t \sum_{i=n-\Delta t}^{n-1} \left(\mathbb{E}\left\|v^{i+1}\right\|^2 + s^2\alpha^2 \left\|\mathcal{G}(x^i)\right\|^2 + s^2\alpha^2\sigma^2\right).
\end{aligned}
$$

Take total expectation on both sides and perform summation,

$$
\begin{aligned}
\sum_{n=0}^{N-1} \mathbb{E}\left\|x^n - x^{n'}\right\|^2 &\le 2\Delta t \sum_{n=0}^{N-1} \sum_{i=n-\Delta t}^{n-1} \left(\mathbb{E}\left\|v^{i+1}\right\|^2 + s^2\alpha^2 \mathbb{E}\left\|\mathcal{G}(x^i)\right\|^2 + s^2\alpha^2\sigma^2\right) \\
&\le 2\Delta t^2 \sum_{n=0}^{N-2} \left(\mathbb{E}\left\|v^{n+1}\right\|^2 + s^2\alpha^2 \mathbb{E}\left\|\mathcal{G}(x^n)\right\|^2 + s^2\alpha^2\sigma^2\right) \\
&\le 2\Delta t^2 \sum_{n=0}^{N-1} \mathbb{E}\left\|v^n\right\|^2 + 2\Delta t^2 s^2\alpha^2 \sum_{n=0}^{N-1} \mathbb{E}\left\|\mathcal{G}(x^n)\right\|^2 + 2\Delta t^2 s^2\alpha^2\sigma^2 N \\
&\le 2\Delta t^2 (c_2 + s^2)\alpha^2 \sum_{n=0}^{N-1} \mathbb{E}\left[\left\|\mathcal{G}(x^n)\right\|^2\right] + 2\Delta t^2\sigma^2(c_2+s^2)\alpha^2 N.
\end{aligned}
$$

□

Theorem 2. *Assume Assumption 1, 2 and 3 hold. Let* $c_2 = \frac{((1-\beta)s-1)^2}{(1-\beta)^2}$, $c_3 = M^2K(K+1)^2\Delta t^2(c_2+s^2)$, $c_4 = 3 + \beta^2 c_2 + 2(1-\beta)^2\Delta t^2(c_2+s^2)$, *and* $c_5 = \frac{2+\beta^2 c_2}{1-\beta} + 2(1-\beta)\Delta t^2(c_2+s^2) + \frac{-c_4+\sqrt{c_4^2+4(1-\beta)^2 c_3}}{2(1-\beta)}$. *If the learning rate* α *is fixed and satisfies* $\alpha \le \frac{-c_4+\sqrt{c_4^2+4(1-\beta)^2 c_3}}{2(1-\beta)c_3 L}$, *then*

$$
\frac{1}{N}\sum_{n=0}^{N-1} \mathbb{E}\left\|\nabla f(x^{n'})\right\|^2 \le \frac{2(1-\beta)(f(x^0)-f^*)}{N\alpha} + c_5\sigma^2 L\alpha.
$$

Proof. According to Lipschitz continuous gradient,

$$
\begin{aligned}
&f(z^{n+1}) - f(z^n) \\
&\leq \left\langle \nabla f(z^n), z^{n+1} - z^n \right\rangle + \frac{L}{2} \left\| z^{n+1} - z^n \right\|^2 \\
&= -\frac{\alpha}{1-\beta} \left\langle \nabla f(z^n), \mathcal{G}(x^n;\xi) \right\rangle + \frac{L\alpha^2}{2(1-\beta)^2} \left\| \mathcal{G}(x^n;\xi) \right\|^2 \\
&= -\frac{\alpha}{1-\beta} \left\langle \nabla f(z^n) - \nabla f(x^n), \mathcal{G}(x^n;\xi) \right\rangle - \frac{\alpha}{1-\beta} \left\langle \nabla f(x^n), \mathcal{G}(x^n;\xi) \right\rangle \\
&\quad + \frac{L\alpha^2}{2(1-\beta)^2} \left\| \mathcal{G}(x^n;\xi) \right\|^2 \\
&\leq \frac{1}{2} \left(\frac{1}{L} \left\| \nabla f(z^n) - \nabla f(x^n) \right\|^2 + \frac{L\alpha^2}{(1-\beta)^2} \left\| \mathcal{G}(x^n;\xi) \right\|^2 \right) \\
&\quad - \frac{\alpha}{1-\beta} \left\langle \nabla f(x^n), \mathcal{G}(x^n;\xi) \right\rangle + \frac{L\alpha^2}{2(1-\beta)^2} \left\| \mathcal{G}(x^n;\xi) \right\|^2 \\
&= \frac{1}{2L} \left\| \nabla f(z^n) - \nabla f(x^n) \right\|^2 - \frac{\alpha}{1-\beta} \left\langle \nabla f(x^n), \mathcal{G}(x^n;\xi) \right\rangle + \frac{L\alpha^2}{(1-\beta)^2} \left\| \mathcal{G}(x^n;\xi) \right\|^2 .
\end{aligned}
$$

Take expectation regarding ξ on both sides,

$$
\begin{aligned}
&\mathbb{E}\left[f(z^{n+1}) \right] - f(z^n) \\
&\leq \frac{1}{2L} \left\| \nabla f(z^n) - \nabla f(x^n) \right\|^2 - \frac{\alpha}{1-\beta} \left\langle \nabla f(x^n), \mathcal{G}(x^n) \right\rangle + \frac{L\alpha^2}{(1-\beta)^2} \left\| \mathcal{G}(x^n) \right\|^2 + \frac{L\alpha^2}{(1-\beta)^2} \sigma^2 \\
&= \frac{1}{2L} \left\| \nabla f(z^n) - \nabla f(x^n) \right\|^2 - \frac{\alpha}{1-\beta} \left\langle \nabla f(x^n) - \nabla f(x^{n'}), \mathcal{G}(x^n) \right\rangle \\
&\quad - \frac{\alpha}{1-\beta} \left\langle \nabla f(x^{n'}), \mathcal{G}(x^n) \right\rangle + \frac{L\alpha^2}{(1-\beta)^2} \left\| \mathcal{G}(x^n) \right\|^2 + \frac{L\alpha^2}{(1-\beta)^2} \sigma^2 \\
&\leq \frac{1}{2L} \left\| \nabla f(z^n) - \nabla f(x^n) \right\|^2 + \frac{1}{2} \left(\frac{1}{L} \left\| \nabla f(x^n) - \nabla f(x^{n'}) \right\|^2 + \frac{L\alpha^2}{(1-\beta)^2} \left\| \mathcal{G}(x^n) \right\|^2 \right) \\
&\quad + \frac{\alpha}{2(1-\beta)} \left(\left\| \nabla f(x^{n'}) - \mathcal{G}(x^n) \right\|^2 - \left\| \nabla f(x^{n'}) \right\|^2 - \left\| \mathcal{G}(x^n) \right\|^2 \right) \\
&\quad + \frac{L\alpha^2}{(1-\beta)^2} \left\| \mathcal{G}(x^n) \right\|^2 + \frac{L\alpha^2}{(1-\beta)^2} \sigma^2 \\
&= -\frac{\alpha}{2(1-\beta)} \left\| \nabla f(x^{n'}) \right\|^2 + \frac{1}{2L} \left\| \nabla f(z^n) - \nabla f(x^n) \right\|^2 + \frac{1}{2L} \left\| \nabla f(x^n) - \nabla f(x^{n'}) \right\|^2 \\
&\quad + \frac{\alpha}{2(1-\beta)} \left\| \nabla f(x^{n'}) - \mathcal{G}(x^n) \right\|^2 - \left(\frac{\alpha}{2(1-\beta)} - \frac{3L\alpha^2}{2(1-\beta)^2} \right) \left\| \mathcal{G}(x^n) \right\|^2 + \frac{L\alpha^2}{(1-\beta)^2} \sigma^2 .
\end{aligned}
$$

Take the total expectation and perform summation,

$$
\sum_{n=0}^{N-1} \mathbb{E} \left[\frac{1}{2L} \left\| \nabla f(z^n) - \nabla f(x^n) \right\|^2 \right] \leq \sum_{n=0}^{N-1} \frac{L}{2} \mathbb{E} \left\| z^n - x^n \right\|^2 = \sum_{n=0}^{N-1} \frac{L\beta^2}{2(1-\beta)^2} \mathbb{E} \left\| v^n \right\|^2 .
$$

$$
\begin{aligned}
&\sum_{n=0}^{N-1}\mathbb{E}\left[\frac{1}{2L}\left\|\nabla f(x^n)-\nabla f(x^{n'})\right\|^2+\frac{\alpha}{2(1-\beta)}\left\|\nabla f(x^{n'})-\mathcal{G}(x^n)\right\|^2\right]\\
&\le\sum_{n=0}^{N-1}\frac{L}{2}\mathbb{E}\left\|x^n-x^{n'}\right\|^2+\frac{\alpha}{4(1-\beta)}L^2M^2K(K+1)\sum_{k=0}^{K-1}(k+1)\sum_{n=0}^{N-1}\mathbb{E}\left\|x_k^n-x_k^{n'}\right\|^2\\
&\le\sum_{n=0}^{N-1}\frac{L}{2}\mathbb{E}\left\|x^n-x^{n'}\right\|^2+\frac{\alpha}{4(1-\beta)}L^2M^2K(K+1)^2\sum_{n=0}^{N-1}\mathbb{E}\left\|x^n-x^{n'}\right\|^2\\
&\le\sum_{n=0}^{N-1}\frac{L}{2}\left(1+\frac{\alpha}{2(1-\beta)}LM^2K(K+1)^2\right)\mathbb{E}\left\|x^n-x^{n'}\right\|^2.
\end{aligned}
$$

Then we have,

$$
\begin{aligned}
&\mathbb{E}\left[f(z^N)\right]-f(z^0)\\
&\le-\frac{\alpha}{2(1-\beta)}\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x^{n'})\right\|^2-\left(\frac{\alpha}{2(1-\beta)}-\frac{3L\alpha^2}{2(1-\beta)^2}\right)\sum_{n=0}^{N-1}\mathbb{E}\left\|\mathcal{G}(x^n)\right\|^2+\frac{L\sigma^2\alpha^2}{(1-\beta)^2}N\\
&\quad+\sum_{n=0}^{N-1}\frac{L\beta^2}{2(1-\beta)^2}\mathbb{E}\left\|v^n\right\|^2+\sum_{n=0}^{N-1}\frac{L}{2}\left(1+\frac{\alpha}{2(1-\beta)}LM^2K(K+1)^2\right)\mathbb{E}\left\|x^n-x^{n'}\right\|^2\\
&\le-\frac{\alpha}{2(1-\beta)}\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x^{n'})\right\|^2-\left(\frac{\alpha}{2(1-\beta)}-\frac{3L\alpha^2}{2(1-\beta)^2}\right)\sum_{n=0}^{N-1}\mathbb{E}\left\|\mathcal{G}(x^n)\right\|^2+\frac{L\sigma^2\alpha^2}{(1-\beta)^2}N\\
&\quad+\frac{L\beta^2}{2(1-\beta)^2}\left(c_2\alpha^2\sum_{n=0}^{N-1}\mathbb{E}\left\|\mathcal{G}(x^n)\right\|^2+c_2\sigma^2\alpha^2N\right)\\
&\quad+\frac{L}{2}\left(1+\frac{\alpha}{2(1-\beta)}LM^2K(K+1)^2\right)\cdot\\
&\qquad\left[2\Delta t^2(c_2+s^2)\alpha^2\sum_{n=0}^{N-1}\mathbb{E}\left\|\mathcal{G}(x^n)\right\|^2+2\Delta t^2\sigma^2(c_2+s^2)\alpha^2N\right]\\
&=-\frac{\alpha}{2(1-\beta)}\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x^{n'})\right\|^2-\left[\frac{\alpha}{2(1-\beta)}-\alpha^2\left(\frac{3L}{2(1-\beta)^2}+\frac{L\beta^2c_2}{2(1-\beta)^2}+\right.\right.\\
&\qquad\left.\left.L\left(1+\frac{\alpha}{2(1-\beta)}LM^2K(K+1)^2\right)\Delta t^2(c_2+s^2)\right)\right]\cdot\sum_{n=0}^{N-1}\mathbb{E}\left\|\mathcal{G}(x^n)\right\|^2\\
&\quad+\sigma^2\alpha^2N\left[\frac{L}{(1-\beta)^2}+\frac{L\beta^2c_2}{2(1-\beta)^2}+L\left(1+\frac{\alpha}{2(1-\beta)}LM^2K(K+1)^2\right)\Delta t^2(c_2+s^2)\right]\\
&=-\frac{\alpha}{2(1-\beta)}\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x^{n'})\right\|^2+\frac{\alpha}{2(1-\beta)^2}[(1-\beta)M^2K(K+1)^2\Delta t^2(c_2+s^2)L^2\alpha^2+\\
&\qquad\left(3+\beta^2c_2+2(1-\beta)^2\Delta t^2(c_2+s^2)\right)L\alpha-(1-\beta)]\cdot\sum_{n=0}^{N-1}\mathbb{E}\left\|\mathcal{G}(x^n)\right\|^2\\
&\quad+\sigma^2\alpha^2N\left[\frac{L}{(1-\beta)^2}+\frac{L\beta^2c_2}{2(1-\beta)^2}+L\left(1+\frac{\alpha}{2(1-\beta)}LM^2K(K+1)^2\right)\Delta t^2(c_2+s^2)\right].
\end{aligned}
$$

The second inequality utilizes Lemma 3 and 4. According to the restriction on the learning rate, we can remove the second term in the last equality,

$$
\begin{aligned}
f_*-f(x^0)\le-\frac{\alpha}{2(1-\beta)}\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x^{n'})\right\|^2+\sigma^2L\alpha^2N\left[\frac{1}{(1-\beta)^2}+\frac{\beta^2c_2}{2(1-\beta)^2}+\right.\\
\left.\left(1+\frac{\alpha}{2(1-\beta)}LM^2K(K+1)^2\right)\Delta t^2(c_2+s^2)\right].
\end{aligned}
$$

Therefore we have,

$$\frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}\left\|\nabla f(x^{n'})\right\|^2 \leq \frac{2(1-\beta)(f^*-f(x^0))}{N\alpha} + \sigma^2 L\alpha\left[\frac{2+\beta^2 c_2}{1-\beta} + \left(2(1-\beta)+\alpha L M^2 K(K+1)^2\right)\Delta t^2(c_2+s^2)\right].$$

□