Convergence Analysis

October 30, 2022

Assumptions

Assumption 1. *(L-smoothness) Each $\ell_t(x)$ is twice continuously differentiable and L-smooth, that is, there exists an $L \geq 0$ such that for all $x, y \in \mathbb{R}^d$, $\|\nabla \ell_t(x) - \nabla \ell_t(y)\| \leq L\|x - y\|$.*

Assumption 2. *(Contractive Compression) The compression operator $\mathcal{C}$ is a contraction mapping, that is*

$$\mathbb{E}_{\mathcal{C}}\left[\|\mathcal{C}(x) - x\|_2^2 \mid x\right] \leq (1 - \delta)\|x\|_2^2, \tag{1}$$

for all $x \in \mathbb{R}^d$ where $0 < \delta \leq 1$ and the expectation is over the randomness generated by compression $\mathcal{C}$.

Assumption 3. *(Bounded Stochastic Gradient) Any unbiased stochastic gradient $\tilde{\nabla}\ell_t$ satisfies*

$$\mathbb{E}\left[\|\tilde{\nabla}\ell_t\|^2\right] \leq \sigma^2 + M\|\nabla \ell_{\mu,t}(x_t)\|^2 \tag{2}$$

for all $t \in \mathbb{N}$, where $\sigma, M > 0$.

Assumption 4. *(Bounded Drift in Time) There exists $\omega_t \geq 0$ such that $|\ell_t(x) - \ell_{t+1}(x)| \leq \omega_t$ for all $x \in \mathbb{R}^d$. Note that in the case where $\ell_{t+1} = \ell_t$, this assumption holds with $\omega_t = 0$ for all $t \in \mathbb{N}$.*

Lemmas

Lemma 1. *If $\ell_t(x)$ is L-smooth, then $\ell_{\mu,t}(x)$ is $L_\mu - smooth$ where $L_\mu \leq L$.*

Lemma 2. *$\ell_{\mu,t}(x)$ has the following gradient:*

$$\nabla \ell_{\mu,t}(x) = \frac{1}{(2\pi)^{d/2}} \int \frac{\ell_t(x + \mu u) - \ell_t(x)}{\mu} u e^{(-\frac{1}{2}\|u\|^2)} \mathrm{d}u \tag{3}$$

where $u \sim \mathcal{N}(0, I_d)$.

Lemma 3.

$$|\ell_{\mu,t}(x) - \ell_t(x)| \leq \frac{\mu^2 L d}{2} \tag{4}$$

Lemma 4.

$$\|\nabla \ell_{\mu,t}(x) - \nabla \ell_t(x)\| \leq \frac{\mu}{2} L (d+3)^{\frac{3}{2}} \tag{5}$$

Lemma 5.

$$\mathbb{E}_u\left[\left\|\frac{\ell_t(x + \mu u) - \ell_t(x)}{\mu} u\right\|^2\right] \leq \frac{\mu^2}{2} L^2 (d+6)^3 + 2(d+4)\|\nabla \ell_t(x)\|^2 \tag{6}$$

Convergence Analysis

Let $\tilde{x}_t$ be defined as follows:

$$\tilde{x}_t = x_t - \eta e_t \tag{7}$$

from our algorithm we know that $e_{t+1} = p_t - \mathcal{C}(p_t)$ and $p_t = g_{\mu,t} + e_t$ so we can write $\tilde{x}_{t+1}$ as

$$\begin{aligned}
\tilde{x}_{t+1} &= x_{t+1} - \eta p_t + \eta \mathcal{C}(p_t) \\
&= x_t - \eta \mathcal{C}(p_t) - \eta g_{\mu,t} - \eta e_t + \eta \mathcal{C}(p_t) \\
&= x_t - \eta e_t - \eta g_{\mu,t} \\
&= \tilde{x}_t - \eta g_{\mu,t}
\end{aligned} \tag{8}$$

By definition $\ell_{\mu,t}(x_t) := \mathbb{E}_{u_t}\left[\ell_t(x_t + \mu u_t)\right]$ so using *L-smoothness* assumption, we can write the following:

$$\ell_{\mu,t}(\tilde{x}_{t+1}) \leq \ell_{\mu,t}(\tilde{x}_t) + \langle \nabla \ell_{\mu,t}(\tilde{x}_t), \tilde{x}_{t+1} - \tilde{x}_t \rangle + \frac{L}{2}\|\tilde{x}_{t+1} - \tilde{x}_t\|^2 \tag{9}$$

Using *Bounded Drift* assumption, we get the following:

$$\ell_{\mu,t+1}(\tilde{x}_{t+1}) \leq \ell_{\mu,t}(\tilde{x}_t) - \eta \langle g_{\mu,t}, \nabla \ell_{\mu,t}(\tilde{x}_t) \rangle + \frac{L\eta^2}{2}\|g_{\mu,t}\|^2 + \omega_t \tag{10}$$

Since $\nabla \ell_{\mu,t}(w_t) = \mathbb{E}_{u_t}\left[g_{\mu,t}\right]$, we have the following:

$$\begin{aligned}
\mathbb{E}_{u_t}\left[\langle g_{\mu,t}, \nabla \ell_{\mu,t}(\tilde{x}_t)\rangle\right] &= \langle \nabla \ell_{\mu,t}(x_t), \nabla \ell_{\mu,t}(\tilde{x}_t) \rangle \\
&= \frac{1}{2}\|\nabla \ell_{\mu,t}(x_t)\|^2 + \frac{1}{2}\|\nabla \ell_{\mu,t}(\tilde{x}_t)\|^2 - \frac{1}{2}\|\nabla \ell_{\mu,t}(x_t) - \nabla \ell_{\mu,t}(\tilde{x}_t)\|^2
\end{aligned} \tag{11}$$

In the last step, we use the fact that $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$. Putting this into (9), we get

$$\ell_{\mu,t+1}(\tilde{x}_{t+1}) \leq \ell_{\mu,t}(\tilde{x}_t) - \frac{\eta}{2}\|\nabla \ell_{\mu,t}(w_t)\|^2 - \frac{\eta}{2}\|\nabla \ell_{\mu,t}(\tilde{x}_t)\|^2 + \frac{L^2\eta}{2}\|x_t - \tilde{x}_t\|^2 + \frac{L\eta^2}{2}\|g_{\mu,t}\|^2 + \omega_t \tag{12}$$

Note that $\|\nabla \ell_{\mu,t}(x_t) - \nabla \ell_{\mu,t}(\tilde{x}_t)\|^2 \leq L^2\|x_t - \tilde{x}_t\|^2$ because of lemma 1. Also, we can drop -$\frac{\eta}{2}\|\nabla \ell_{\mu,t}(\tilde{x}_t)\|^2$ because it is always negative. Using the fact that $\tilde{x}_t - x_t = \eta e_t$, we get

$$\underbrace{\frac{\eta}{2}\|\nabla \ell_{\mu,t}(x_t)\|^2}_{\text{Term III}} \leq \underbrace{\left[\ell_{\mu,t}(\tilde{x}_t) - \ell_{\mu,t+1}(\tilde{x}_{t+1})\right]}_{\text{Term II}} + \underbrace{\frac{L\eta^2}{2}\|g_{\mu,t}\|^2}_{\text{Term I}} + \underbrace{\frac{\eta^3 L^2}{2}\|e_t\|^2}_{\text{Term IV}} \tag{13}$$

We will put an upper bound to the terms I, II, and IV and a lower bound to term III. Let's start with **term I**. By lemma 5, we know that

$$\mathbb{E}_{u_t}\left[\|g_{\mu,t}\|^2\right] \leq 2(d+4)\mathbb{E}\left[\|\tilde{\nabla}\ell_t(x_t)\|^2\right] + \frac{\mu^2 L^2}{2}(d+6)^3 \tag{14}$$

where $\mathbb{E}[\|\tilde{\nabla}\ell_t(x_t)\|^2] \leq \|\nabla \ell_t(x_t)\|^2 + \sigma^2$. After the telescoping sum, we can put the following upper bound to **term II** by lemma 3.

$$\begin{aligned}
\ell_{\mu,1}(\tilde{x}_1) - \ell_{\mu,T+1}(\tilde{x}^*) &\leq \mu L^2 d + \ell_1(\tilde{x}_1) - \ell_{T+1}(\tilde{x}_{T+1}) \\
&\leq \mu L^2 d + \ell_1(x_1) - \ell_{T+1}(x^*_{T+1})
\end{aligned} \tag{15}$$

where we use the two facts: $\ell_{T+1}(x^*_{T+1}) \leq \ell_{T+1}(\tilde{x}_{T+1})$ and $\ell_1(x_1) = \ell_1(\tilde{x}_1)$ because $x^*_{T+1} = \arg\min_x \ell_{T+1}(x)$. We can put the following lower bound to **term III** by using lemma 4 and Young's inequality.

$$\frac{1}{2}\|\nabla \ell_t(x_t)\|^2 - \frac{\mu^2 L^2}{4}(d+3)^3 \leq \|\nabla \ell_{\mu,t}(w_x)\|^2 \tag{16}$$

Lastly, we can put the following upper bound to **term IV** by using assumption 1 and Young's inequality.

$$\begin{aligned}
\mathbb{E}[\|e_{t+1}\|^2] = \|p_t - \mathcal{C}(p_t)\|^2 &\leq (1-\delta)\|p_t\|^2 = (1-\delta)\|e_t + g_{\mu,t}\|^2 \\
&\leq (1-\delta)(1+\varphi)\|e_t\|^2 + (1-\delta)(1+\frac{1}{\varphi})\|g_{\mu,t}\|^2 \\
&= \sum_{i=1}^{t} \left[(1-\delta)(1+\varphi)\right]^{t-i} (1-\delta)(1+\frac{1}{\varphi})\mathbb{E}\|g_{\mu,i}\|^2
\end{aligned} \tag{17}$$

And note that $\mathbb{E}[\|g_{\mu,t}\|^2] \leq A\|\nabla \ell_t(x_t)\|^2 + B$ where

$$\begin{aligned}
B &= 2\sigma^2(d+4) + \frac{\mu^2 L^2}{2}(d+6)^3 \\
A &= 2M(d+4)
\end{aligned} \tag{18}$$

and M come from the *Bounded Stochastic Gradient* assumption, and we used lemma 5. So we can write (17) as follows:

$$\mathbb{E}\left[\|e_{t+1}\|^2\right] \leq \sum_{i=1}^{t} \left[(1-\delta)(1+\varphi)\right]^{t-i} (1-\delta)(1+\frac{1}{\varphi})\left[A\|\nabla \ell_i(x_i)\|^2 + B\right] \tag{19}$$

If we set $\varphi = \frac{\delta}{2(1-\delta)}$, then $1 + \frac{1}{\varphi} \leq \frac{2}{\delta}$ and $(1-\delta)(1+\varphi) = (1-\frac{\delta}{2})$, so we get

$$\mathbb{E}\left[\|e_{t+1}\|^2\right] \leq \sum_{i=1}^{t} \left(1-\frac{\delta}{2}\right)^{t-i} \left[A\|\nabla \ell_i(x_i)\|^2 + B\right] \frac{2(1-\delta)}{\delta} \tag{20}$$

If we sum through all $\mathbb{E}[\|e_t\|^2]$, we get

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\|e_t\|^2\right] &\leq \sum_{t=1}^{T}\sum_{i=1}^{t-1} \left(1-\frac{\delta}{2}\right)^{t-i} \left[A\|\nabla \ell_i(x_i)\|^2 + B\right] \frac{2(1-\delta)}{\delta} \\
&\leq \sum_{t=1}^{T} (A\|\nabla \ell_t(x_t)\|^2 + B) \sum_{i=0}^{\infty} \left(1-\frac{\delta}{2}\right)^{i} \frac{2(1-\delta)}{\delta} \\
&\leq \sum_{t=1}^{T} (A\|\nabla \ell_t(x_t)\|^2 + B) C
\end{aligned} \tag{21}$$

where $C = \frac{2(1-\delta)}{\delta}\frac{2}{\delta} \leq \frac{4}{\delta^2}$. If we combine the upper bounds derived in (14), (15), (17), and lower bound derived in (16) and put them in (13), we get the following:

$$\begin{aligned} &\sum_{t=1}^{T}\frac{\eta}{4}\|\nabla\ell_t(x_t)\|^2 - \frac{\eta\mu^2L^2}{8}(d+3)^3T \\ &\quad\le \mu L^2 d + \ell_1(x_1) - \ell_{T+1}(x^*_{T+1}) + \frac{T\mu^2L^3\eta^2}{4}(d+6)^3 + \frac{L\eta^2}{2}\sigma^2 T\times 2(d+4) \\ &\quad + \frac{L\eta^2}{2}\times 2M(d+4)\sum_{t=1}^{T}\|\nabla\ell_t(x_t)\|^2 + \frac{\eta^3L^2}{2}\times\frac{4}{\delta^2}T\left[2\sigma^2(d+4) + \frac{\mu^2L^2}{2}(d+6)^3\right] \\ &\quad + \frac{\eta^3L^2}{2}\times\frac{4}{\delta^2}\sum_{t=1}^{T}2M(d+4)\|\nabla\ell_t(x_t)\|^2 + \sum_{t=1}^{T}\omega_t \end{aligned} \tag{22}$$

That implies

$$\begin{aligned} \frac{E}{T}\sum_{t=1}^{T}\|\nabla\ell_t(x_t)\|^2 \le{} & \frac{\mu L^2 d + \left[\ell_1(x_1) - \ell_{T+1}(x^*_{T+1})\right]}{T} + \frac{\eta^2L^3\mu^2(d+6)^3}{4} + L\eta^2\sigma^2(d+4) + \frac{\eta\mu^2L^2(d+3)^3}{8} \\ & + \frac{\eta^3L^2}{\delta^2}4\sigma^2(d+4) + \frac{\eta^3L^2}{\delta^2}\mu^2L^2(d+6)^3 + \frac{1}{T}\sum_{t=1}^{T}\omega_t \end{aligned} \tag{23}$$

where

$$\begin{aligned} E &= \frac{\eta}{4} - LM\eta^2(d+4) - \frac{L^2\eta^3}{\delta^2}4M(d+4) \\ &= \eta\left[\frac{1}{4} - LM\eta(d+4)\left[1 + \frac{4L\eta}{\delta^2}\right]\right] \end{aligned} \tag{24}$$

If $\eta \le \frac{1}{4L}$, instead first upper bound will be

$$1 + \frac{4L\eta}{\delta^2} \le 1 + \frac{1}{\delta^2} = \frac{\delta^2+1}{\delta^2} \le \frac{2}{\delta^2} \tag{25}$$

Find η such that

$$LM\eta(d+4)\times\frac{2}{\delta^2} \le \frac{1}{8} \tag{26}$$

Then, we get

$$\eta \le \frac{\delta^2}{16LM(d+4)} \tag{27}$$

which implies $E \ge \frac{\eta}{8}$. Multiply all terms in the bound by

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}\|\nabla\ell_t(x_t)\|^2 \le{} & \frac{8(\ell_1 - \ell^*)}{(\eta T)} + \frac{8\mu L^2 d}{\eta T} + 2\eta L^3\mu^2(d+6)^3 \\ & + 8L\eta\sigma^2(d+4) + \mu^2L^2(d+3)^3 \\ & + \frac{32\eta^2L^2}{\delta^2}\sigma^2(d+4) + \frac{8\eta^2L^2\mu^2L^2(d+6)^3}{\delta^2} + \frac{8}{\eta T}\sum_{t=1}^{T}\omega_t \end{aligned} \tag{28}$$

Let

$$\eta = \frac{1}{\sigma\sqrt{(d+4)MTL}} \quad \text{and} \quad \mu = \frac{1}{(d+4)\sqrt{T}} \tag{29}$$

Then, we have

$$\frac{1}{CT}\sum_{t=1}^{T}\|\nabla \ell_t(x_t)\|^2 \leq \frac{1}{\delta^2}\frac{dL\Delta}{T} + \sigma\sqrt{\frac{d}{T}L\Delta M} + \frac{1}{T}\sum_{t=1}^{T}\omega_t \tag{30}$$

for a numerical constant $C > 0$, where $\Delta = \ell_1(x_1) - \ell_{T+1}(x^*_{T+1})$ for $x^*_t = \arg\min_{x\in\mathbb{R}^d} \ell_t(x)$. Furthermore, if $\omega_t = 0$, i.e., in the case where $\ell_{t+1} = \ell_t$, the number of time steps T to obtain a ξ-first order solution is

$$T = \mathcal{O}\left(\frac{d\sigma^2 L\Delta M}{\xi^2} + \frac{dL\Delta}{\delta^2\xi}\right). \tag{31}$$

That is the end of the proof.