Convergence Analysis

November 2, 2022

Notation & Definitions

- t : time index, $t \in \mathbb{Z}^+$.
- z_t : position of the target at time t, $z_t \in \mathbb{R}^d$.
- x_t : position of the agent at time t, $x_t \in \mathbb{R}^d$.
- $\ell_t\,(x, z)$: stochastic loss as evaluated by the zeroth-order oracle at time t, with the position of agent as x, and the position of target as z for $x, z \in \mathbb{R}^d$.
- $\ell_{\mu,t}(x, z) := \mathbb{E}_u[\ell_t(x + \mu u, z)]$ for $x, z \in \mathbb{R}^d$, $u \sim \mathcal{N}(0, I_d)$ and $\mu \in \mathbb{R}$.
- $\nabla \ell_{\mu,t}(x, z) := \mathbb{E}_u\left[g_{\mu,t}(x, z)\right]$ where $g_{\mu,t}(x, z) := \dfrac{\ell_t(x + \mu u, z) - \ell_t(x, z)}{\mu} u$ for $x, z \in \mathbb{R}^d$, $u \sim \mathcal{N}(0, I_d)$ and $\mu \in \mathbb{R}$.
- $\ell_t(x) := \mathbb{E}_z\left[\ell_t(x, z)\right]$ for $x, z \in \mathbb{R}^d$.

Assumptions

Assumption 1. *(Unbiased Stochastic Zeroth-Order Oracle) For any $t \in \mathbb{Z}^+$ and $x, z \in \mathbb{R}^d$, we have*

$$\mathbb{E}_z\left[\ell_t(x, z)\right] = \ell_t(x_t). \tag{1}$$

Assumption 2. *(Unbiased Stochastic First-Order Oracle with Bounded Variance) For any $t \in \mathbb{Z}^+$ and $x, z \in \mathbb{R}^d$, we have*

$$\mathbb{E}_z\left[\nabla \ell_t(x, z)\right] = \nabla \ell_t(x) \tag{2}$$

and

$$\mathbb{E}_z\left[\|\nabla \ell_t(x, z) - \nabla \ell_t(x)\|^2\right] \leq \sigma^2. \tag{3}$$

Assumption 3. *(L-smoothness) Each $\ell_t(x, z)$ is continuously differentiable and L-smooth over x on $\mathbb{R}^d$, that is, there exists an $L \geq 0$ such that for all $x, y, z \in \mathbb{R}^d$ and $t \in \mathbb{Z}^+$, we have*

$$\|\nabla \ell_t(x, z) - \nabla \ell_t(y, z)\| \leq L\|x - y\|. \tag{4}$$

We denote this by $\ell_t(x, z) \in C_L^{1,1}(\mathbb{R}^d)$.

Assumption 4. *(Contractive Compression) The compression function $\mathcal{C}$ is a contraction mapping, that is,*

$$\mathbb{E}_{\mathcal{C}}\left[\|\mathcal{C}(x)-x\|^2 \mid x\right] \leq (1-\delta)\|x\|^2 \tag{5}$$

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 5. *(Bounded Stochastic Gradients) For any $t \in \mathbb{Z}^+$ and $x, z \in \mathbb{R}^d$, there exist $\sigma, M > 0$ such that*

$$\mathbb{E}_z\left[\|\nabla \ell_t(x,z)\|^2\right] \leq \sigma^2 + M\|\nabla \ell_{\mu,t}(x)\|^2. \tag{6}$$

Assumption 6. *(Bounded Drift in Time) There exists a nonnegative sequence $\{\omega_t\}_{t=1}^T$ such that for all $t \in \mathbb{Z}^+$, $|\ell_t(x,z) - \ell_{t+1}(x,z)| \leq \omega_t$ for any $x, z \in \mathbb{R}^d$. Note that in the case where $\ell_{t+1} = \ell_t$, this assumption holds with $\omega_t = 0$.*

Lemmas

Suppose $\ell_t(x,z) \in C_L^{1,1}(\mathbb{R}^d)$ over x. We have the following results:

Lemma 1. *$\ell_{\mu,t}(x,z) \in C_{L_\mu}^{1,1}(\mathbb{R}^d)$ over x, where $L_\mu \leq L$.*

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

$$\nabla \ell_{\mu,t}(x,z) = \frac{1}{(2\pi)^{d/2}} \int \frac{\ell_t(x+\mu u, z) - \ell_t(x,z)}{\mu} u e^{(-\frac{1}{2}\|u\|^2)} \mathrm{d}u, \tag{7}$$

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

Lemma 3. *For any $x, z \in \mathbb{R}^d$, we have*

$$|\ell_{\mu,t}(x,z) - \ell_t(x,z)| \leq \frac{\mu^2 L d}{2}. \tag{8}$$

Lemma 4. *For any $x, z \in \mathbb{R}^d$, we have*

$$\|\nabla \ell_{\mu,t}(x,z) - \nabla \ell_t(x,z)\| \leq \frac{\mu}{2} L (d+3)^{\frac{3}{2}}, \tag{9}$$

where the gradient is with respect to x.

Lemma 5. *For any $x, z \in \mathbb{R}^d$, we have*

$$\mathbb{E}_u\left[\left\|\frac{\ell_t(x+\mu u, z) - \ell_t(x,z)}{\mu} u\right\|^2\right] \leq \frac{\mu^2}{2} L^2 (d+6)^3 + 2(d+4)\|\nabla \ell_t(x,z)\|^2, \tag{10}$$

where $u \sim \mathcal{N}(0, I_d)$ and the gradient is with respect to x.

Lemma 6. *(Young's inequality) For any $x, y \in \mathbb{R}^d$ and $\lambda > 0$, we have*

$$\langle x, y \rangle \leq \frac{\|x\|^2}{2\lambda} + \frac{\|y\|^2 \lambda}{2}. \tag{11}$$

Convergence Analysis

In the analysis, we assume that $z_t \in \mathbb{R}^d$ are *i.i.d.* random variables for all $t \in \mathbb{Z}^+$.

Let $\tilde{x}_t$ be defined as follows (following the analysis in [1]):

$$\tilde{x}_t = x_t - \eta e_t. \tag{12}$$

From the algorithm, we know that $e_{t+1} = p_t - \mathcal{C}(p_t)$ and $p_t = g_{\mu,t}(x_t, z_t) + e_t$, so we can rewrite $\tilde{x}_{t+1}$ as

$$\begin{aligned} \tilde{x}_{t+1} &= x_{t+1} - \eta p_t + \eta \mathcal{C}(p_t) \\ &= x_t - \eta \mathcal{C}(p_t) - \eta g_{\mu,t}(x_t, z_t) - \eta e_t + \eta \mathcal{C}(p_t) \\ &= x_t - \eta e_t - \eta g_{\mu,t}(x_t, z_t) \\ &= \tilde{x}_t - \eta g_{\mu,t}(x_t, z_t), \end{aligned} \tag{13}$$

where $g_{\mu,t}(x_t, z_t) := \dfrac{\ell_t(x_t + \mu u_t, z_t) - \ell_t(x_t, z_t)}{\mu} u_t$ and $u_t \sim \mathcal{N}(0, I_d)$.

By definition, $\ell_{\mu,t}(x_t, z_t) := \mathbb{E}_{u_t}\left[\ell_t(x_t + \mu u_t, z_t)\right]$, so by assumption 3, we can write the following:

$$\ell_{\mu,t}(\tilde{x}_{t+1}, z_{t+1}) \leq \ell_{\mu,t}(\tilde{x}_t, z_t) + \langle \nabla \ell_{\mu,t}(\tilde{x}_t, z_t), \tilde{x}_{t+1} - \tilde{x}_t \rangle + \frac{L}{2}\|\tilde{x}_{t+1} - \tilde{x}_t\|^2. \tag{14}$$

Now by assumption 6, we get:

$$\ell_{\mu,t+1}(\tilde{x}_{t+1}, z_{t+1}) \leq \ell_{\mu,t}(\tilde{x}_t, z_t) - \eta \langle g_{\mu,t}(x_t, z_t), \nabla \ell_{\mu,t}(\tilde{x}_t, z_t) \rangle + \frac{L\eta^2}{2}\|g_{\mu,t}(x_t, z_t)\|^2 + \omega_t \tag{15}$$

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

$$\begin{aligned} \mathbb{E}_{u_t}\left[\langle g_{\mu,t}(x_t, z_t), \nabla \ell_{\mu,t}(\tilde{x}_t, z_t) \rangle\right] &= \langle \nabla \ell_{\mu,t}(x_t, z_t), \nabla \ell_{\mu,t}(\tilde{x}_t, z_t) \rangle \\ &= \frac{1}{2}\|\nabla \ell_{\mu,t}(x_t, z_t)\|^2 + \frac{1}{2}\|\nabla \ell_{\mu,t}(\tilde{x}_t, z_t)\|^2 - \frac{1}{2}\|\nabla \ell_{\mu,t}(x_t, z_t) - \nabla \ell_{\mu,t}(\tilde{x}_t, z_t)\|^2. \end{aligned} \tag{16}$$

In the last step, we use the fact that $2\langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$. Plugging this into (15), we get:

$$\begin{aligned} \ell_{\mu,t+1}(\tilde{x}_{t+1}, z_{t+1}) \leq & \ell_{\mu,t}(\tilde{x}_t, z_t) - \frac{\eta}{2}\|\nabla \ell_{\mu,t}(x_t, z_t)\|^2 - \frac{\eta}{2}\|\nabla \ell_{\mu,t}(\tilde{x}_t, z_t)\|^2 \\ & + \frac{L^2\eta}{2}\|x_t - \tilde{x}_t\|^2 + \frac{L\eta^2}{2}\|g_{\mu,t}(x_t, z_t)\|^2 + \omega_t. \end{aligned} \tag{17}$$

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

$$\underbrace{\frac{\eta}{2}\|\nabla \ell_{\mu,t}(x_t, z_t)\|^2}_{\text{Term III}} \leq \underbrace{\left[\ell_{\mu,t}(\tilde{x}_t, z_t) - \ell_{\mu,t+1}(\tilde{x}_{t+1}, z_{t+1})\right]}_{\text{Term II}} + \underbrace{\frac{L\eta^2}{2}\|g_{\mu,t}(x_t, z_t)\|^2}_{\text{Term I}} + \underbrace{\frac{\eta^3 L^2}{2}\|e_t\|^2}_{\text{Term IV}} + \omega_t. \tag{18}$$

We will put an upper bound to the terms I, II, IV and a lower bound to term III. Starting with **term I**, by lemma 5, we know that

$$\mathbb{E}_{u_t,z_t}\left[\|g_{\mu,t}(x_t, z_t)\|^2\right] \leq 2(d+4)\mathbb{E}_{z_t}\left[\|\nabla \ell_t(x_t, z_t)\|^2\right] + \frac{\mu^2 L^2}{2}(d+6)^3, \tag{19}$$

where $\mathbb{E}_{z_t}[\|\nabla\ell_t(x_t,z_t)\|^2] \le \|\nabla\ell_t(x_t)\|^2 + \sigma^2$ by assumption 2.

We can put the following upper bound to **term II** by means of a telescoping sum and subsequently applying lemma 3:

$$\begin{aligned}\sum_{t=1}^{T}\left[\ell_{\mu,t}(\tilde{x}_t,z_t) - \ell_{\mu,t+1}(\tilde{x}_{t+1},z_{t+1})\right] &= \ell_{\mu,1}(\tilde{x}_1,z_1) - \ell_{\mu,T+1}(\tilde{x}_{T+1},z_{T+1}) \\ &\le \mu L^2 d + \ell_1(\tilde{x}_1,z_1) - \ell_{T+1}(\tilde{x}_{T+1},z_{T+1}) \\ &= \mu L^2 d + \ell_1(x_1,z_1) - \ell_{T+1}(\tilde{x}_{T+1},z_{T+1}),\end{aligned} \tag{20}$$

where we use the fact that $\ell(x_1,z_1) = \ell_1(\tilde{x}_1,z_1)$ because $\tilde{x}_1 = x_1$ by definition. If we take the expectation of both sides with respect to $z_{1:T+1} = \{z_1, z_2, ..., z_{T+1}\}$, owing to the fact that z_t's are *i.i.d.*, we get

$$\begin{aligned}\ell_{\mu,1}(\tilde{x}_1) - \ell_{\mu,T+1}(\tilde{x}^*) &\le \mu L^2 d + \ell_1(x_1) - \ell_{T+1}(\tilde{x}_{T+1}) \\ &\le \mu L^2 d + \ell_1(x_1) - \ell_{T+1}(x^*_{T+1}),\end{aligned} \tag{21}$$

where $x^*_{T+1} = \arg\min_x \ell_{T+1}(x)$.

We can put the following lower bound to **term III** by using lemma 4 and lemma 6:

$$\frac{1}{2}\|\nabla\ell_t(x_t,z_t)\|^2 - \frac{\mu^2L^2}{4}(d+3)^3 \le \|\nabla\ell_{\mu,t}(x_t,z_t)\|^2. \tag{22}$$

Lastly, we can put the following upper bound to **term IV** by assumption 4 and lemma 6:

$$\begin{aligned}\mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}[\|e_{t+1}\|^2] &= \|p_t - \mathcal{C}_t(p_t)\|^2 \le (1-\delta)\|p_t\|^2 = (1-\delta)\|e_t + g_{\mu,t}(x_t,z_t)\|^2 \\ &\le (1-\delta)(1+\varphi)\mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}\left[\|e_t\|^2\right] + (1-\delta)(1+\frac{1}{\varphi})\mathbb{E}_{u_{1:T},z_{1:T}}\left[\|g_{\mu,t}(x_t,z_t)\|^2\right] \\ &= \sum_{i=1}^{t}\left[(1-\delta)(1+\varphi)\right]^{t-i}(1-\delta)(1+\frac{1}{\varphi})\mathbb{E}_{u_i,z_i}\left[\|g_{\mu,i}(x_i,z_i)\|^2\right],\end{aligned} \tag{23}$$

for some $\varphi > 0$, $z_t, x_t, \mathcal{C}_t$ are *i.i.d.*, and $\mathbb{E}_{\mathcal{C}_t}[\,\cdot\,]$ denotes the expectation over the randomness at time t due to the compression used. Note that by assumption 5 and using lemma 5,

$$\mathbb{E}_{u_t,z_t}[\|g_{\mu,t}(x_t,z_t)\|^2] \le A\|\nabla\ell_t(x_t)\|^2 + B, \tag{24}$$

where

$$\begin{aligned}&B = 2\sigma^2(d+4) + \frac{\mu^2L^2}{2}(d+6)^3 \text{ and} \\ &A = 2M(d+4).\end{aligned} \tag{25}$$

So we can rewrite (23) as follows:

$$\mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}\left[\|e_{t+1}\|^2\right] \le \sum_{i=1}^{t}\left[(1-\delta)(1+\varphi)\right]^{t-i}(1-\delta)(1+\frac{1}{\varphi})\left[A\|\nabla\ell_i(x_i)\|^2 + B\right]. \tag{26}$$

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

$$\mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}\left[\|e_{t+1}\|^2\right] \leq \sum_{i=1}^{t}\left(1-\frac{\delta}{2}\right)^{t-i}\left[A\|\nabla\ell_i(x_i)\|^2+B\right]\frac{2(1-\delta)}{\delta}. \tag{27}$$

If we sum through all $\mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}[\|e_t\|^2]$, we get:

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}\left[\|e_t\|^2\right] &\leq \sum_{t=1}^{T}\sum_{i=1}^{t-1}\left(1-\frac{\delta}{2}\right)^{t-i}\left[A\|\nabla\ell_i(x_i)\|^2+B\right]\frac{2(1-\delta)}{\delta}\\ &\leq \sum_{t=1}^{T}\left[A\|\nabla\ell_t(x_t)\|^2+B\right]\sum_{i=0}^{\infty}\left(1-\frac{\delta}{2}\right)^{i}\frac{2(1-\delta)}{\delta}\\ &\leq \sum_{t=1}^{T}\left[A\|\nabla\ell_t(x_t)\|^2+B\right]C,\end{aligned} \tag{28}$$

where $C=\frac{2(1-\delta)}{\delta}\frac{2}{\delta}\leq\frac{4}{\delta^2}$. If we define $\Delta := \ell_1(x_1)-\ell_{T+1}(x^*_{T+1})$ and combine the upper bounds derived in (19), (20), (23), and the lower bound derived in (22) and plug them into (18), we get the following:

$$\begin{aligned}&\sum_{t=1}^{T}\frac{\eta}{4}\mathbb{E}_{z_t}\left[\|\nabla\ell_t(x_t,z_t)\|^2\right]-\frac{\eta\mu^2L^2}{8}(d+3)^3T\\ &\quad\leq \mu L^2 d+\Delta+\frac{T\mu^2L^3\eta^2}{4}(d+6)^3+\frac{L\eta^2}{2}\sigma^2T\times 2(d+4)\\ &\quad+\frac{L\eta^2}{2}\times 2M(d+4)\sum_{t=1}^{T}\mathbb{E}_{z_t}\left[\|\nabla\ell_t(x_t,z_t)\|^2\right]+\frac{\eta^3L^2}{2}\times\frac{4}{\delta^2}T\left[2\sigma^2(d+4)+\frac{\mu^2L^2}{2}(d+6)^3\right]\\ &\quad+\frac{\eta^3L^2}{2}\times\frac{4}{\delta^2}\sum_{t=1}^{T}2M(d+4)\mathbb{E}_{z_t}\left[\|\nabla\ell_t(x_t,z_t)\|^2\right]+\sum_{t=1}^{T}\omega_t.\end{aligned} \tag{29}$$

Now, since z_t's are *i.i.d.* for all $t\in\mathbb{Z}^+$, we have:

$$\begin{aligned}\frac{E}{T}\sum_{t=1}^{T}\mathbb{E}_{z_t}\left[\|\nabla\ell_t(x_t,z_t)\|^2\right] &\leq \frac{\mu L^2 d+\Delta}{T}+\frac{\eta^2L^3\mu^2(d+6)^3}{4}+L\eta^2\sigma^2(d+4)+\frac{\eta\mu^2L^2(d+3)^3}{8}\\ &\quad+\frac{\eta^3L^2}{\delta^2}4\sigma^2(d+4)+\frac{\eta^3L^2}{\delta^2}\mu^2L^2(d+6)^3+\frac{1}{T}\sum_{t=1}^{T}\omega_t,\end{aligned} \tag{30}$$

where

$$\begin{aligned}E&=\frac{\eta}{4}-LM\eta^2(d+4)-\frac{L^2\eta^3}{\delta^2}4M(d+4)\\ &=\eta\left[\frac{1}{4}-LM\eta(d+4)\left(1+\frac{4L\eta}{\delta^2}\right)\right].\end{aligned} \tag{31}$$

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

$$1+\frac{4L\eta}{\delta^2}\leq 1+\frac{1}{\delta^2}=\frac{\delta^2+1}{\delta^2}\leq\frac{2}{\delta^2}. \tag{32}$$

We proceed to find an η such that

$$\frac{2}{\delta^2} LM\eta(d+4) \leq \frac{1}{8}. \tag{33}$$

Then, we get

$$\eta \leq \frac{\delta^2}{16LM(d+4)}, \tag{34}$$

which implies $E \geq \frac{\eta}{8}$. Multiplying all terms in the bound by $\frac{8}{\eta}$,

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{z_t}\left[\|\nabla \ell_t(x_t, z_t)\|^2\right] \leq{}& \frac{8\Delta}{(\eta T)} + \frac{8\mu L^2 d}{\eta T} + 2\eta L^3 \mu^2 (d+6)^3 \\ &+ 8L\eta\sigma^2(d+4) + \mu^2 L^2 (d+3)^3 \\ &+ \frac{32\eta^2 L^2}{\delta^2}\sigma^2(d+4) + \frac{8\eta^2 L^4 \mu^2 (d+6)^3}{\delta^2} + \frac{8}{\eta T}\sum_{t=1}^{T}\omega_t. \end{aligned} \tag{35}$$

Let

$$\eta = \frac{1}{\sigma\sqrt{(d+4)MTL}} \quad \text{and} \quad \mu = \frac{1}{(d+4)\sqrt{T}}. \tag{36}$$

Then, for a numerical constant $C > 0$, we have

$$\frac{1}{CT}\sum_{t=1}^{T} \mathbb{E}_{z_t}\left[\|\nabla \ell_t(x_t, z_t)\|^2\right] \leq \frac{1}{\delta^2}\frac{dL\Delta}{T} + \sigma\sqrt{\frac{d}{T}L\Delta M} + \frac{1}{\eta T}\sum_{t=1}^{T}\omega_t. \tag{37}$$

Defining $\bar{\omega} := \sum_{t=1}^{T}\omega_t$, the number of times steps T to obtain a ξ-first order solution is

$$T = \mathcal{O}\left(\frac{d\sigma^2 L\Delta M}{\xi^2} + \frac{dL\Delta}{\delta^2\xi} + \frac{\bar{\omega}\sigma^2 dML}{\xi^2}\right). \tag{38}$$

Remark: In choosing $\eta = \frac{1}{\sigma\sqrt{(d+4)MTL}}$, we assumed that it satisfies (34). For this to hold, T can be made arbitrarily large as long as it does not exceed the bound we found in (38). (36) and (34) imply that

$$T = \Omega\left(\frac{dLM}{\delta^4\sigma^2}\right). \tag{39}$$

In practice, since $\xi \ll \delta$, this term is smaller than (37). This fact is also demonstrated by our experiments.

Lastly, if $\omega_t = 0$ for all $t \in \mathbb{Z}^+$, i.e., in the case where the loss function is time-invariant, the number of time steps T to obtain a ξ-first order solution is:

$$T = \mathcal{O}\left(\frac{d\sigma^2 L\Delta M}{\xi^2} + \frac{dL\Delta}{\delta^2\xi}\right). \tag{40}$$

References

[1] S. P. Karimireddy, Q. Rebjock, S. U. Stich, and M. Jaggi, “Error feedback fixes signsgd and other gradient compression schemes,” 2019.