Convergence Analysis

December 9, 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, $x_t \in \mathbb{R}^d$.
- We denote stochastic variables $\tilde{\ell}_t^i(x) := \ell_t^i(x,z)$, $\tilde{\nabla}\ell_{\mu,t}^i(x) := \nabla\ell_{\mu,t}^i(x,z)$, and $\tilde{g}_{\mu,t}^i(x) := g_{\mu,t}^i(x,z)$ for *i.i.d.* $z \sim P_z$, at time t, with the position of i^{th} agent as x for $x \in \mathbb{R}^d$ and $i \in \{1,...,N\}$.
- $\tilde{\ell}_{\mu,t}^i(x) := \mathbb{E}_u[\tilde{\ell}_t^i(x+\mu u)]$ for $x \in \mathbb{R}^d$, $u \sim \mathcal{N}(0, I_d)$ and $\mu \in \mathbb{R}$.
- $\tilde{\nabla}\ell_{\mu,t}^i(x) := \mathbb{E}_u\left[\tilde{g}_{\mu,t}^i(x)\right]$ where $\tilde{g}_{\mu,t}^i(x) := \dfrac{\tilde{\ell}_t^i(x+\mu u) - \tilde{\ell}_t^i(x)}{\mu}u$ for $x \in \mathbb{R}^d$, $u \sim \mathcal{N}(0, I_d)$ and $\mu \in \mathbb{R}$.

Assumptions

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

$$\mathbb{E}_z\left[\tilde{\ell}_t^i(x)\right] = \ell_t^i(x). \tag{1}$$

Assumption 2. *(Unbiased Stochastic First-Order Oracle) For any $t \in \mathbb{Z}^+$, $i \in \{1,\dots,N\}$ and $x \in \mathbb{R}^d$, we have*

$$\mathbb{E}_z\left[\nabla\tilde{\ell}_t^i(x)\right] = \nabla\ell_t^i(x) \tag{2}$$

Assumption 3. *(L-smoothness) Each $\tilde{\ell}_t^i(x)$ is continuously differentiable and L-smooth over x on $\mathbb{R}^d$, that is, there exists an $L \geq 0$ such that for all $x, y \in \mathbb{R}^d$, $t \in \mathbb{Z}^+$ and $i \in \{1,\dots,N\}$, we have*

$$\|\nabla\tilde{\ell}_t^i(x) - \nabla\tilde{\ell}_t^i(y)\| \leq L\|x-y\|. \tag{3}$$

We denote this by $\tilde{\ell}_t^i(x) \in C_L^{1,1}(\mathbb{R}^d)$. Note that this assumption implies $\ell_t^i(x) \in C_L^{1,1}(\mathbb{R}^d)$.

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

$$\mathbb{E}_{\mathcal{C}}\left[\|\mathcal{C}(x) - x\|^2 \mid x\right] \leq (1-\delta)\|x\|^2 \tag{4}$$

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 $t \in \mathbb{Z}^+$, $i \in \{1, \ldots, N\}$ and $x \in \mathbb{R}^d$, there exist $\sigma, M > 0$ such that*

$$\mathbb{E}_z \left[\|\nabla \tilde{\ell}_t^i(x)\|^2 \right] \leq \sigma^2 + M \|\nabla \ell_t^i(x)\|^2. \tag{5}$$

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

Lemmas

Suppose $f(x) \in C_L^{1,1}(\mathbb{R}^d)$. Then, we have the following results:

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

Lemma 2. *$f_\mu(x)$ has the following gradient with respect to x:*

$$\nabla f_\mu(x) = \frac{1}{(2\pi)^{d/2}} \int \frac{f(x + \mu u) - f(x)}{\mu} u e^{(-\frac{1}{2}\|u\|^2)} \mathrm{d}u, \tag{6}$$

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

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

$$|f_\mu(x) - f(x)| \leq \frac{\mu^2 L d}{2}. \tag{7}$$

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

$$\|\nabla f_\mu(x) - \nabla f(x)\| \leq \frac{\mu}{2} L (d+3)^{\frac{3}{2}}, \tag{8}$$

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

$$\mathbb{E}_u \left[\left\| \frac{f(x + \mu u) - f(x)}{\mu} u \right\|^2 \right] \leq \frac{\mu^2}{2} L^2 (d+6)^3 + 2(d+4) \|\nabla f(x)\|^2, \tag{9}$$

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

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

$$\langle x, y \rangle \leq \frac{\|x\|^2}{2\lambda} + \frac{\|y\|^2 \lambda}{2}. \tag{10}$$

EF-ZO-SGD Convergence Analysis for Single-Agent

We work with the following algorithm:

Algorithm 1 EF-ZO-SGD

Input: Number of time steps $T \in \mathbb{Z}^+$, smoothing parameter $\mu \in \mathbb{R}$, initial source position $x_0 \in \mathbb{R}^d$, learning rate $\eta \in \mathbb{R}$, sequence of target positions $\{z_t\}_{t=1}^T \subset \mathbb{R}^d$.

Output: Sequence of optimal source positions $\{x_t\}_{t=1}^T \subset \mathbb{R}^d$.

1: $e_0 = 0$
2: **for** $t = 1, \ldots, T$ **do**
3: $\quad u_t \sim \mathcal{N}(0, I_d)$
4: $\quad \tilde{g}_{\mu,t}(x_t) = \dfrac{\tilde{\ell}_t(x_t + \mu u_t) - \tilde{\ell}_t(x_t)}{\mu} u_t$
5: $\quad p_t = \tilde{g}_{\mu,t}(x_t) + e_t$
6: $\quad x_{t+1} = x_t - \eta \mathcal{C}(p_t)$
7: $\quad e_{t+1} = p_t - \mathcal{C}(p_t)$
8: **end for**

In the analysis, we assume that $z_t \in \mathbb{R}^d$ are *i.i.d.* random variables for all $t \in \mathbb{Z}^+$. Furthermore, we drop the superscript notation present in the assumptions, since i is always 1 for the single-agent case.

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

$$\tilde{x}_t := x_t - \eta e_t. \tag{11}$$

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

$$\begin{aligned} \tilde{x}_{t+1} &= x_{t+1} - \eta p_t + \eta \mathcal{C}(p_t) \\ &= x_t - \eta \mathcal{C}(p_t) - \eta \tilde{g}_{\mu,t}(x_t) - \eta e_t + \eta \mathcal{C}(p_t) \\ &= x_t - \eta e_t - \eta \tilde{g}_{\mu,t}(x_t) \\ &= \tilde{x}_t - \eta \tilde{g}_{\mu,t}(x_t), \end{aligned} \tag{12}$$

where $\tilde{g}_{\mu,t}(x_t) = \dfrac{\tilde{\ell}_t(x_t + \mu u_t) - \tilde{\ell}_t(x_t)}{\mu} u_t$ and $u_t \sim \mathcal{N}(0, I_d)$.

By assumption 3, we can write the following:

$$\ell_{\mu,t}(\tilde{x}_{t+1}) \le \ell_{\mu,t}(\tilde{x}_t) + \langle \nabla \ell_{\mu,t}(\tilde{x}_t), \tilde{x}_{t+1} - \tilde{x}_t \rangle + \frac{L}{2} \|\tilde{x}_{t+1} - \tilde{x}_t\|^2. \tag{13}$$

Now by assumption 6, we get:

$$\ell_{\mu,t+1}(\tilde{x}_{t+1}) \le \ell_{\mu,t}(\tilde{x}_t) - \eta \langle \tilde{g}_{\mu,t}(x_t), \nabla \ell_{\mu,t}(\tilde{x}_t) \rangle + \frac{L\eta^2}{2} \|\tilde{g}_{\mu,t}(x_t)\|^2 + \omega_t. \tag{14}$$

Since $\nabla \ell_{\mu,t}(x_t) = \mathbb{E}_{u_t,z_t}[\tilde{g}_{\mu,t}(x_t)]$, taking the expectation of both sides with respect to u_t and z_t, we have the following:

$$\begin{aligned} \mathbb{E}_{u_t,z_t}[\langle \tilde{g}_{\mu,t}(x_t), \nabla \ell_{\mu,t}(\tilde{x}_t) \rangle] &= \langle \nabla \ell_{\mu,t}(x_t), \nabla \ell_{\mu,t}(\tilde{x}_t) \rangle \\ &= \frac{1}{2} \|\nabla \ell_{\mu,t}(x_t)\|^2 + \frac{1}{2} \|\nabla \ell_{\mu,t}(\tilde{x}_t)\|^2 - \frac{1}{2} \|\nabla \ell_{\mu,t}(x_t) - \nabla \ell_{\mu,t}(\tilde{x}_t)\|^2. \end{aligned} \tag{15}$$

In the last step, we use the fact that $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$. Plugging this into (14), we get:

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

Note that $\|\nabla \ell_{\mu,t}(x_t) - \nabla \ell_{\mu,t}(\tilde{x}_t)\|^2 \leq L^2\|x_t - \tilde{x}_t\|^2$ by assumption 3, with subsequent application of lemma 1. Also, we can drop -$\frac{\eta}{2}\|\nabla \ell_{\mu,t}(\tilde{x}_t)\|^2$ because it is nonpositive. Using the fact that $\tilde{x}_t - x_t = \eta e_t$, we get:

$$\underbrace{\frac{\eta}{2}\|\nabla \ell_{\mu,t}(x_t)\|^2}_{\text{Term III}} \leq \underbrace{[\ell_{\mu,t}(\tilde{x}_t) - \ell_{\mu,t+1}(\tilde{x}_{t+1})]}_{\text{Term II}} + \underbrace{\frac{L\eta^2}{2}\mathbb{E}_{u_t,z_t}\left[\|\tilde{g}_{\mu,t}(x_t)\|^2\right]}_{\text{Term I}} + \underbrace{\frac{L^2\eta^3}{2}\|e_t\|^2}_{\text{Term IV}} + \omega_t. \tag{17}$$

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

$$\mathbb{E}_{u_t,z_{1:T}}\left[\|\tilde{g}_{\mu,t}(x_t)\|^2\right] \leq 2(d+4)\mathbb{E}_{z_{1:T}}\left[\|\tilde{\nabla}\ell_t(x_t)\|^2\right] + \frac{\mu^2L^2}{2}(d+6)^3, \tag{18}$$

where $\mathbb{E}_{z_{1:T}}[\|\tilde{\nabla}\ell_t(x_t)\|^2] \leq M\mathbb{E}_{z_{1:T}}\left[\|\nabla \ell_t(x_t)\|^2\right] + \sigma^2$ by assumption 5. Note that, in this step, we use the the principle of causality and the fact that z_t are *i.i.d.*.

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

$$\begin{aligned}\sum_{t=1}^{T}[\ell_{\mu,t}(\tilde{x}_t) - \ell_{\mu,t+1}(\tilde{x}_{t+1})] &= \ell_{\mu,1}(\tilde{x}_1) - \ell_{\mu,T+1}(\tilde{x}_{T+1}) \\ &\leq \mu^2Ld + \ell_1(\tilde{x}_1) - \ell_{T+1}(\tilde{x}_{T+1}) \\ &= \mu^2Ld + \ell_1(x_1) - \ell_{T+1}(\tilde{x}_{T+1}),\end{aligned} \tag{19}$$

where we use the fact that $\ell(x_1) = \ell_1(\tilde{x}_1)$ because $\tilde{x}_1 = x_1$ by definition. Then, we can do the following

$$\begin{aligned}\sum_{t=1}^{T}[\ell_{\mu,t}(\tilde{x}_t) - \ell_{\mu,t+1}(\tilde{x}_{t+1})] &\leq \mu^2Ld + \ell_1(x_1) - \ell_{T+1}(\tilde{x}_{T+1}) \\ &\leq \mu^2Ld + \ell_1(x_1) - \ell_{T+1}(x^*_{T+1}),\end{aligned} \tag{20}$$

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:

$$\frac{1}{2}\|\nabla \ell_t(x_t)\|^2 - \frac{\mu^2L^2}{4}(d+3)^3 \leq \|\nabla \ell_{\mu,t}(x_t)\|^2. \tag{21}$$

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

$$
\begin{aligned}
\mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}[\|e_{t+1}\|^2] &= \mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}\left[\|p_t - \mathcal{C}_t(p_t)\|^2\right] \leq (1-\delta)\mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}\left[\|p_t\|^2\right] \\
&= (1-\delta)\mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}\left[\|e_t + \tilde{g}_{\mu,t}(x_t)\|^2\right] \\
&\leq (1-\delta)(1+\varphi)\mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}\left[\|e_t\|^2\right] + (1-\delta)(1+\frac{1}{\varphi})\mathbb{E}_{u_{1:T},z_{1:T}}\left[\|\tilde{g}_{\mu,t}(x_t)\|^2\right] \\
&= \sum_{i=1}^{t}\left[(1-\delta)(1+\varphi)\right]^{t-i}(1-\delta)(1+\frac{1}{\varphi})\mathbb{E}_{u_i,z_{1:T}}\left[\|\tilde{g}_{\mu,i}(x_i)\|^2\right],
\end{aligned}
\tag{22}
$$

for some $\varphi > 0$, $z_t, u_t, \mathcal{C}_t$ 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 using lemma 5 and assumption 5,

$$
\mathbb{E}_{u_t,z_{1:T}}[\|\tilde{g}_{\mu,t}(x_t)\|^2] \leq A\mathbb{E}_{z_{1:T}}\left[\|\nabla \ell_t(x_t)\|^2\right] + B, \tag{23}
$$

where

$$
\begin{aligned}
&B = 2\sigma^2(d+4) + \frac{\mu^2 L^2}{2}(d+6)^3 \text{ and} \\
&A = 2M(d+4).
\end{aligned}
\tag{24}
$$

So we can rewrite (22) as follows:

$$
\mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}\left[\|e_{t+1}\|^2\right] \leq \sum_{i=1}^{t}\left[(1-\delta)(1+\varphi)\right]^{t-i}(1-\delta)(1+\frac{1}{\varphi})\left[A\mathbb{E}_{z_{1:T}}\left[\|\nabla \ell_i(x_i)\|^2\right] + B\right]. \tag{25}
$$

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:

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

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

$$
\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}_{u_{1:T},z_{1:T},\mathcal{C}_{1:T}}\left[\|e_t\|^2\right] &\leq \sum_{t=1}^{T}\sum_{i=1}^{t-1}\left(1-\frac{\delta}{2}\right)^{t-i}\left[A\mathbb{E}_{z_{1:T}}\left[\|\nabla \ell_i(x_i)\|^2\right] + B\right]\frac{2(1-\delta)}{\delta} \\
&\leq \sum_{t=1}^{T}\left[A\mathbb{E}_{z_{1:T}}\left[\|\nabla \ell_t(x_t)\|^2\right] + B\right]\sum_{i=0}^{\infty}\left(1-\frac{\delta}{2}\right)^{i}\frac{2(1-\delta)}{\delta} \\
&\leq \sum_{t=1}^{T}\left[A\mathbb{E}_{z_{1:T}}\left[\|\nabla \ell_t(x_t)\|^2\right] + B\right]C,
\end{aligned}
\tag{27}
$$

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 (18), (19), (22), and the lower bound derived in (21) and plug them into (17), we get the following:

$$\sum_{t=1}^{T}\frac{\eta}{4}\mathbb{E}_{z_{1:T}}\left[\|\nabla\ell_t(x_t)\|^2\right]-\frac{\eta\mu^2L^2}{8}(d+3)^3T$$
$$\begin{aligned}&\le \mu^2Ld+\Delta+\frac{T\mu^2L^3\eta^2}{4}(d+6)^3+\frac{L\eta^2}{2}\sigma^2T\times 2(d+4)\\&+\frac{L\eta^2}{2}\times 2M(d+4)\sum_{t=1}^{T}\mathbb{E}_{z_{1:T}}\left[\|\nabla\ell_t(x_t)\|^2\right]+\frac{\eta^3L^2}{2}\times\frac{4}{\delta^2}T\left[2\sigma^2(d+4)+\frac{\mu^2L^2}{2}(d+6)^3\right]\\&+\frac{\eta^3L^2}{2}\times\frac{4}{\delta^2}\sum_{t=1}^{T}2M(d+4)\mathbb{E}_{z_{1:T}}\left[\|\nabla\ell_t(x_t)\|^2\right]+\sum_{t=1}^{T}\omega_t.\end{aligned} \tag{28}$$

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

$$\begin{aligned}\frac{E}{T}\sum_{t=1}^{T}\mathbb{E}_{z_{1:T}}\left[\|\nabla\ell_t(x_t)\|^2\right]\le\frac{\mu^2Ld+\Delta}{T}+\frac{\eta^2L^3\mu^2(d+6)^3}{4}+L\eta^2\sigma^2(d+4)+\frac{\eta\mu^2L^2(d+3)^3}{8}\\+\frac{\eta^3L^2}{\delta^2}4\sigma^2(d+4)+\frac{\eta^3L^2}{\delta^2}\mu^2L^2(d+6)^3+\frac{1}{T}\sum_{t=1}^{T}\omega_t,\end{aligned} \tag{29}$$

where

$$\begin{aligned}E&=\frac{\eta}{4}-LM\eta^2(d+4)-\frac{L^2\eta^3}{\delta^2}4M(d+4)\\&=\eta\left[\frac{1}{4}-LM\eta(d+4)\left(1+\frac{4L\eta}{\delta^2}\right)\right].\end{aligned} \tag{30}$$

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

$$1+\frac{4L\eta}{\delta^2}\le 1+\frac{1}{\delta^2}=\frac{\delta^2+1}{\delta^2}\le\frac{2}{\delta^2}. \tag{31}$$

We proceed to find an η such that

$$\frac{2}{\delta^2}LM\eta(d+4)\le\frac{1}{8}. \tag{32}$$

Then, we get

$$\eta\le\frac{\delta^2}{16LM(d+4)}, \tag{33}$$

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

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{z_{1:T}}\left[\|\nabla\ell_t(x_t)\|^2\right]\le\frac{8\Delta}{(\eta T)}+\frac{8\mu^2Ld}{\eta T}+2\eta L^3\mu^2(d+6)^3\\+8L\eta\sigma^2(d+4)+\mu^2L^2(d+3)^3\\+\frac{32\eta^2L^2}{\delta^2}\sigma^2(d+4)+\frac{8\eta^2L^4\mu^2(d+6)^3}{\delta^2}+\frac{8}{\eta T}\sum_{t=1}^{T}\omega_t.\end{aligned} \tag{34}$$

Let

$$\eta=\frac{1}{\sigma\sqrt{(d+4)MTL}}\quad\text{and}\quad\mu=\frac{1}{(d+4)\sqrt{T}}. \tag{35}$$

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

$$\frac{1}{CT}\sum_{t=1}^{T}\mathbb{E}_{z_{1:T}}\left[\|\nabla \ell_t(x_t)\|^2\right] \leq \frac{1}{\delta^2}\frac{dL\Delta}{T} + \sigma\sqrt{\frac{d}{T}L\Delta M} + \frac{1}{\eta T}\sum_{t=1}^{T}\omega_t. \tag{36}$$

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

$$T = \mathcal{O}\left(\frac{d\sigma^2 L\Delta M}{\xi^2} + \frac{dL\Delta}{\delta^2\xi} + \frac{\bar{\omega}\sigma^2 dML}{\xi^2}\right). \tag{37}$$

Remark: In choosing $\eta = \frac{1}{\sigma\sqrt{(d+4)MTL}}$, we assumed that it satisfies (33). For this to hold, T can be made arbitrarily large as long as it does not exceed the bound we found in (37). (35) and (33) imply that

$$T = \Omega\left(\frac{dLM}{\delta^4\sigma^2}\right). \tag{38}$$

In practice, since $\xi \ll \delta$, this term is smaller than (36). 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:

$$T = \mathcal{O}\left(\frac{d\sigma^2 L\Delta M}{\xi^2} + \frac{dL\Delta}{\delta^2\xi}\right). \tag{39}$$

References

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