Federated Learning Algorithm

Yida Lyu

2024 年 11 月 24 日

§ 1 SGD vs SVRG(convex)

1.1 SGD

1.1.1

We consider

$$x^* = \arg\min_{x\in\mathbb{R}^d}[f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)] \tag{1}$$

$f_i : \mathbb{R}^d \to \mathbb{R}$ is smooth. Further, we assume that f has a unique global minimizer x^* and is μ–strongly quasi-convex:

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x^* - x\|^2 \tag{2}$$

1.1.2

Introduce a sampling vector v:$\mathbb{E}_{\mathcal{D}}[v_i] = 1, \forall \in [n]$. So (1) is equivalent of:

$$\min_{x\in\mathbb{R}^d} \mathbb{E}_{\mathcal{D}}[f_v(x) := \frac{1}{n}\sum_{i=1}^{n} v_i f_i(x)] \tag{3}$$

$f_v(x)$ and $\nabla f_v(x)$ are unbiased estimators of $f(x)$ and $\nabla f(x)$,so (3)is equal to (1).

(3) can be solved using SGD:

$$x^{k+1} = x^k - \gamma^k \nabla f_{v^k}(x^k) \tag{4}$$

where v^k is sampled i.i.d. at each iteration and $\gamma^k > 0$ is a stepsize.

Algorithm 1 SGD

Input: initial point x^0, stepsize γ/γ^k, sampling vector v

1: **for** $k = 0, 1, 2, \cdots$ **do**
2: $\quad g^k = \frac{1}{n}\sum_{i=1}^{n} v_i \nabla f_i(x^k)$
3: $\quad x^{k+1} = x^k - \gamma^k g^k$
4: **end for**

1.1.3

Assumption 1: f is $\mathcal{L}$-smooth in expectation w.r.t $\mathcal{D}$:

$$\mathbb{E}_D[\|\nabla f_v(x) - \nabla f_v(x^*)\|^2] \leq 2\mathcal{L}(f(x) - f(x^*)), \forall x \in \mathbb{R}^d \tag{5}$$

we will write $(f, \mathcal{D}) \sim ES(\mathcal{L})$

(Convexity and $\mathcal{L}_i$–smoothness of f_i implies expected smoothness, but the opposite implication does not hold.)

Assumption 2(Finite Gradient Noise):

$$\sigma^2 := \mathbb{E}_{\mathcal{D}}[\|\nabla f_v(x^*)\|^2] \tag{6}$$

is finite.

Lemma. If$(f, \mathcal{D}) \sim ES(\mathcal{L})$,then

$$\mathbb{E}_D[\|\nabla f_v(x)\|^2] \leq 4\mathcal{L}(f(x) - f(x^*)) + 2\sigma^2 \tag{7}$$

Proof.

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\|\nabla f_v(x)\|^2 &= \mathbb{E}_{\mathcal{D}}\|\nabla f_v(x) - \nabla f_v(x^*) + \nabla f_v(x^*)\|^2 \\
&\leq 2\mathbb{E}_{\mathcal{D}}\|\nabla f_v(x) - \nabla f_v(x^*)\|^2 + 2\mathbb{E}_{\mathcal{D}}\|\nabla f_v(x^*)\|^2 \\
&\leq 4\mathcal{L}[f(x) - f(x^*)] + 2\mathbb{E}_{\mathcal{D}}\|\nabla f_v(x^*)\|^2.
\end{aligned}$$

When $\sigma = 0$,(7) is known as the week growth condition.

1.1.4 Analysis

Thm. Assume f is μ-quasi-strongly convex and that $(f, \mathcal{D}) \sim ES(\mathcal{L})$. Choose $\gamma^k = \gamma \in (0, \frac{1}{2\mathcal{L}}]\forall k$. Then iterates of SGD given by (4) satisfy:

$$E\|x^k - x^*\|^2 \leq (1-\gamma\mu)^k\|x^0 - x^*\|^2 + \frac{2\gamma\sigma^2}{\mu}. \tag{8}$$

Hence,$\forall \epsilon > 0$,choosing

$$\gamma = \min\{\frac{1}{2\mathcal{L}}, \frac{\epsilon\mu}{4\sigma^2}\} \tag{9}$$

and

$$k \geq \max\{\frac{2\mathcal{L}}{\mu}, \frac{4\sigma^2}{\epsilon\mu^2}\}\log(\frac{2\|x^0 - x^*\|^2}{\epsilon}) \tag{10}$$

implies $\mathbb{E}\|x^k - x^*\| \leq \epsilon$.

Proof. Let $r^k = x^k - x^*$. From (4), we have

$$\begin{aligned}
\|r^{k+1}\|^2 &= \|x^k - x^* - \gamma\nabla f_v(x^k)\|^2 \\
&= \|r^k\|^2 - 2\gamma\langle r^k, \nabla f_{v^k}(x^k)\rangle + \gamma^2\|\nabla f_{v^k}(x^k)\|^2
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\|r^{k+1}\|^2 &= \|r^k\|^2 - 2\gamma\langle r^k, \nabla f(x^k)\rangle + \gamma^2\mathbb{E}_{\mathcal{D}}\|\nabla f_{v^k}(x^k)\|^2 \\
&\leq (1-\gamma\mu)\|r^k\|^2 - 2\gamma[f(x^k) - f(x^*)] + \gamma^2\mathbb{E}_{\mathcal{D}}\|\nabla f_{v^k}(x^k)\|^2 \\
&\quad (\langle\nabla f(x^k), r^k\rangle \geq f(x^k) - f(x^*) + \frac{\mu}{2}\|r^k\|^2)
\end{aligned}$$

Taking expectations again and using Lemma:

$$\begin{aligned}
\mathbb{E}\|r^{k+1}\|^2 &\leq (1-\gamma\mu)\mathbb{E}\|r^k\|^2 + 2\gamma^2\sigma^2 + 2\gamma(2\gamma\mathcal{L} - 1)\mathbb{E}[f(x^k) - f(x^*)] \\
&\leq (1-\gamma\mu)\mathbb{E}\|r^k\|^2 + 2\gamma^2\sigma^2
\end{aligned}$$

At last,

$$\begin{aligned}
\mathbb{E}\|r^k\|^2 &\leq (1-\gamma\mu)^k\|r^0\|^2 + 2\sum_{j=0}^{k-1}(1-\gamma\mu)^j\gamma^2\sigma^2 \\
&\leq (1-\gamma\mu)^k\|r^0\|^2 + \frac{2\gamma\sigma^2}{\mu} \qquad (11)
\end{aligned}$$

Furthermore, we can control this additive constant by carefully choosing the stepsize.

1.2 SVRG

1.2.1

One practical issue for SGD is that in order to ensure convergence the learning rate t has to decay to zero. This leads to slower convergence.

At each time, we keep a version of estimated x as $\tilde{x}$ that is close to the optimal x.

We can keep a snapshot of $\tilde{w}$ after every m SGD iterations.

Moreover, we maintain the average gradient

$$\tilde{\mu} = \nabla f(\tilde{x}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x})$$

Algorithm 2 SVRG

Input: initial point $\tilde{x}^0$, stepsize γ, update frequency m

for $s = 0, 1, 2, \cdots$ **do**
  $\tilde{x} = \tilde{x}^s$
  $\tilde{\mu} = \nabla f(\tilde{x}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x})$
  $x^0 = \tilde{x}$
  for $k = 0, 1, 2, \cdots, m-1$ **do**
    Randomly pick $i_k \in [n]$
    $g^k = \nabla f_{i_k}(x_k) - \nabla f_{i_k}(\tilde{x}) + \tilde{\mu}$
    $x^{k+1} = x^k - \gamma g^k$
  end for
  Randomly pick $t \in \{0, 1, \cdots, m-1\}$
  $\tilde{x}^{s+1} = x^t$
end for

1.2.2 Analysis

Thm. f_i are smooth:

$$f_i(x^*) \leq f_i(x) + \langle \nabla f_i(x), x^* - x \rangle + \frac{L}{2}\|x^* - x\|^2 \tag{12}$$

and convex, f is quasi-convex;

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\mu}{2}\|x^* - x\|^2 \tag{13}$$

$L \geq \mu \geq 0, \gamma > 0$, m is sufficiently large s.t.

$$\alpha = \frac{1}{\mu\gamma(1 - 2L\gamma)m} + \frac{2L\gamma}{1 - 2L\gamma} < 1$$

then we have geometric convergence in expectation for SVRG:

$$\mathbb{E}[f(\tilde{x}^s) - f(x^*)] \leq \alpha^s [f(\tilde{x}^0) - f(x^*)]$$

Proof. By (12)

$$\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \leq 2L[f_i(x) - f_i(x^*) - \nabla f_i(x^*)^\top (x - x^*)]$$

By suming,

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \leq 2L(f(x) - f(x^*)), \forall x \in \mathbb{R}^d \tag{14}$$

$$
\begin{aligned}
&\mathbb{E}\|g^k\|^2\\
\leq&2\mathbb{E}\|\nabla f_{i_k}(x^k)-\nabla f_{i_k}(x^*)\|^2+2\mathbb{E}\|[\nabla f_{i_k}(\tilde{x})-\nabla f_{i_k}(x^*)]-\nabla f(\tilde{x})\|^2\\
=&2\mathbb{E}\|\nabla f_{i_k}(x^k)-\nabla f_{i_k}(x^*)\|^2+2\mathbb{E}\|[\nabla f_{i_k}(\tilde{x})-\nabla f_{i_k}(x^*)]-\mathbb{E}[\nabla f_{i_k}(\tilde{x})-\nabla f_{i_k}(x^*)]\|^2\\
\leq&2\mathbb{E}\|\nabla f_{i_k}(x^k)-\nabla f_{i_k}(x^*)\|^2+2\mathbb{E}\|\nabla f_{i_k}(\tilde{x})-\nabla f_{i_k}(x^*)\|^2\\
\leq&4L[f(x^k)-f(x^*)+f(\tilde{x})-f(x^*)]\ (14)
\end{aligned}
$$

Let $r^k=x^k-x^*$, we have

$$
\begin{aligned}
\|r^{k+1}\|^2&=\|x^k-x^*-\gamma g^k\|^2\\
&=\|r^k\|^2-2\gamma\langle r^k,g^k\rangle+\gamma^2\|g^k\|^2
\end{aligned}
$$

Then,

$$
\begin{aligned}
\mathbb{E}\|r^{k+1}\|^2&=\|r^k\|^2-2\gamma\langle r^k,\nabla f(x^k)\rangle+\gamma^2\mathbb{E}\|g^k\|^2\\
&\leq\|r^k\|^2-2\gamma[f(x^k)-f(x^*)]+4L\gamma^2[f(x^k)-f(x^*)+f(\tilde{x})-f(x^*)]\ (13)\\
&=\|r^k\|^2-2\gamma(1-2L\gamma)[f(x^k)-f(x^*)]+4L\gamma^2[f(\tilde{x})-f(x^*)]
\end{aligned}
$$

We fixed s, so that $\tilde{x}=\tilde{x}^s$ and $\tilde{x}^{s+1}$ is selected after all of the updates have completed. By suming $k\in\{0,1,\cdots,m-1\}$ we can get:

$$
\begin{aligned}
&\mathbb{E}\|r^m\|^2+2\gamma(1-2L\gamma)m\mathbb{E}[f(\tilde{x}^{s+1})-f(x^*)]\\
\leq&\mathbb{E}\|r^0\|^2+4Lm\gamma^2\mathbb{E}[f(\tilde{x})-f(x^*)]\\
=&\mathbb{E}\left\|\tilde{x}-x^*\right\|^2+4Lm\gamma^2\mathbb{E}[f(\tilde{x})-f(x^*)]\\
\leq&\frac{2}{\mu}\mathbb{E}[f(\tilde{x})-f(x^*)]+4Lm\gamma^2\mathbb{E}[f(\tilde{x})-f(x^*)]\\
=&2(\frac{1}{\mu}+2Lm\gamma^2)\mathbb{E}[f(\tilde{x})-f(x^*)].
\end{aligned}
$$

By $\mathbb{E}\|r^m\|^2\geq 0$ and let $0<\gamma<\frac{1}{4L},m>\frac{1}{\mu\gamma(1-4L\gamma)}$,

$$
\mathbb{E}\left[f(\tilde{x}^{s+1})-f(x^*)\right]\leq\left[\frac{1}{\mu\gamma(1-2L\gamma)m}+\frac{2L\gamma}{1-2L\gamma}\right]\mathbb{E}[f(\tilde{x}^s)-f(w_*)]
$$

Then $\alpha<1$ and

$$
\mathbb{E}\left[f(\tilde{x}^s)-f(x^*)\right]\leq\alpha^s\mathbb{E}\left[f(\tilde{x}^0)-f(x^*)\right]
$$

Let $C=\mathbb{E}\left[f(\tilde{x}^0)-f(x^*)\right]$. Next we choose s s.t.

$$
\alpha^sC\leq\epsilon
$$

i.e.

$$
\log(\frac{C}{\epsilon})\leq s\log(\frac{1}{\alpha})
$$

By $\log(\frac{1}{\alpha})\geq 1-\alpha$, $0<\alpha<1$, we only need to have

$$
s\geq\frac{1}{1-\alpha}\log\left(\frac{C}{\epsilon}\right)=\frac{1}{\frac{1-4L\gamma}{1-2L\gamma}-\frac{1}{\mu\gamma(1-2L\gamma)m}}\log\left(\frac{C}{\epsilon}\right)
$$

Let $\gamma=\frac{1}{(k+4)L},\quad m=\frac{\rho+1}{\mu\gamma(1-4L\gamma)}=\frac{(k+4)^2L(\rho+1)}{\mu k},\quad k>0,\rho>0,$

Then

$$s \geq \frac{\mu(k+2)m}{\mu mk - (k+4)^2 L} \log\left(\frac{C}{\epsilon}\right) = \frac{(\rho+1)(k+2)}{\rho k} \log\left(\frac{C}{\epsilon}\right)$$

$$sm \geq \frac{(\rho+1)^2}{\rho} \frac{(k+2)(k+4)^2}{k^2} \frac{L}{\mu} \log\left(\frac{C}{\epsilon}\right)$$

Let $\rho = 1, k = 2 + 2\sqrt{5}$,

$$sm \geq \frac{(4+2\sqrt{5})(6+2\sqrt{5})^2}{(1+\sqrt{5})^2} \frac{L}{\mu} \log\left(\frac{\mathbb{E}\left[f(\tilde{x}^0) - f(x^*)\right]}{\epsilon}\right)$$

$$(\gamma = \frac{1}{(6+2\sqrt{5})L}, m = \frac{(6+2\sqrt{5})^2 L}{(1+\sqrt{5})\mu}, s \geq \frac{4+2\sqrt{5}}{1+\sqrt{5}} \log\left(\frac{\mathbb{E}\left[f(\tilde{x}^0) - f(x^*)\right]}{\epsilon}\right))$$

Finally,

$$s(m+n) \geq (C_1 \frac{L}{\mu} + C_2 n) \log(\frac{C}{\epsilon})$$

§ 2 Direct Compression vs Shift Compression(nonconvex)

2.1 Problem

We consider the more general nonconvex distributed/federated problem with online form or finite-sum form, i.e.

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{m} \sum_{i=1}^{m} f_i(x) \right\}$$

where

$$f_i(x) := \mathbb{E}_{\zeta \sim \mathcal{D}_i}[f_i(x, \zeta)], \quad \text{or} \quad f_i(x) := \frac{1}{n} \sum_{j=1}^{n} f_{i,j}(x).$$

Def. (Compression operator) A randomized map $\mathcal{C} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is an ω-compression operator if

$$\mathbb{E}[\mathcal{C}(x)] = x, \quad \mathbb{E}[\|\mathcal{C}(x) - x\|^2] \leq \omega \|x\|^2, \quad \forall x \in \mathbb{R}^d.$$

In particular, no compression ($\mathcal{C}(x) \equiv x$) implies $\omega = 0$.

Assumption 1 (Gradient estionator) $\mathbb{E}_k[g^k] = \nabla f(x^k)$, and $\exists$ non-ngative constants $A_1, A_2, B_1, B_2, C_1, C_2, D_1, \rho$ and a random sequence $\{\sigma_k^2\}$ s.t.

$$\mathbb{E}_k[\|g^k\|^2] \leq 2A_1(f(x^k) - f^*) + B_1\|\nabla f(x^k)\|^2 + D_1\sigma_k^2 + C_1$$

$$\mathbb{E}_k[\sigma_{k+1}^2] \leq (1-\rho)\sigma_k^2 + 2A_2(f(x^k) - f^*) + B_2\|\nabla f(x^k)\|^2 + C_2$$

Assumption 2 (L-smoothness) For each work $i \in [m]$, the function $f_i(x)$ is L_i-smooth if

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i\|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

2.2 DC framework

2.2.1 Algorithm

Algorithm 3 DC

Input: initial point x^0, stepsize η

1: **for** $k = 0, 1, 2, \cdots$ **do**

2:     **for** all machines $i = 0, 1, 2, \cdots, m$ in parallel **do**

Compute local stochastic gradient $\tilde{g}_i^k$
$\widehat{\Delta}_i^k = \mathcal{C}_i^k(\tilde{g}_i^k)$
end for
$g^k = \frac{1}{m}\sum_{i=1}^m \widehat{\Delta}_i^k$
$x^{k+1} = x^k - \eta g^k$
end for

2.2.2 Analysis

Theorem. (DC framework) If the local stochastic gradient $\widetilde{g}_i^k$ satisfes the recursions:

$$\mathbb{E}_k[\|\widetilde{g}_i^k\|^2] \le 2A_{1,i}(f_i(x^k) - f_i^*) + B_{1,i}\|\nabla f_i(x^k)\|^2 + D_{1,i}\sigma_{k,i}^2 + C_{1,i},$$
$$\mathbb{E}_k[\sigma_{k+1,i}^2] \le (1-\rho_i)\sigma_{k,i}^2 + 2A_{2,i}(f(x^k) - f^*) + B_{2,i}\|\nabla f(x^k)\|^2 + D_{2,i}\mathbb{E}_k[\|g^k\|^2] + C_{2,i},$$

then g^k satisfes the unified Assumption with

$$A_1 = \frac{(1+\omega)A}{m}, \quad B_1 = 1, \quad D_1 = \frac{1+\omega}{m}, \quad \sigma_k^2 = \frac{1}{m}\sum_{i=1}^m D_{1,i}\sigma_{k,i}^2, \quad C_1 = \frac{(1+\omega)C}{m}$$
$$\rho = \min_i \rho_i - \tau, \quad A_2 = D_A + \tau A, \quad B_2 = D_B + D_D, \quad C_2 = D_C + \tau C$$

where

$$A := \max_i(A_{1,i} + B_{1,i}L_i - L_i/(1+\omega)), \quad C := \frac{1}{m}\sum_{i=1}^m C_{1,i} + 2A\Delta_f^*,$$
$$\Delta_f^* := f^* - \frac{1}{m}\sum_{i=1}^m f_i^*, \quad \tau := \frac{(1+\omega)D_D}{m},$$
$$D_A := \frac{1}{m}\sum_{i=1}^m D_{1,i}A_{2,i}, \quad D_B := \frac{1}{m}\sum_{i=1}^m D_{1,i}B_{2,i},$$
$$D_C := \frac{1}{m}\sum_{i=1}^m D_{1,i}C_{2,i}, \quad D_D := \frac{1}{m}\sum_{i=1}^m D_{1,i}D_{2,i}.$$

2.2.3 DC-GD

Assume that

$$\|\nabla f_i(x) - \nabla f(x)\|^2 \le \zeta^2, \quad \forall x \in \mathbb{R}^d$$

$\widetilde{g}_i^k = \nabla f_i(x^k)$. Let $\Delta_f^* := f^* - \frac{1}{m}\sum_{i=1}^m f_i^*$, then,

$$\mathbb{E}_k[g^k] = \mathbb{E}_k\left[\frac{1}{m}\sum_{i=1}^m \mathcal{C}_i^k(\widetilde{g}_i^k)\right] = \frac{1}{m}\sum_{i=1}^m \nabla f_i(x^k) = \nabla f(x^k)$$

$$\begin{aligned}
\mathbb{E}_k[\|g^k\|^2] &= \mathbb{E}_k\left[\left\|\frac{1}{m}\sum_{i=1}^m \mathcal{C}_i^k(\widetilde{g}_i^k) - \frac{1}{m}\sum_{i=1}^m \widetilde{g}_i^k + \frac{1}{m}\sum_{i=1}^m \widetilde{g}_i^k\right\|^2\right] \\
&= \mathbb{E}_k\left[\left\|\frac{1}{m}\sum_{i=1}^m \left(\mathcal{C}_i^k(\widetilde{g}_i^k) - \widetilde{g}_i^k\right)\right\|^2\right] + \left\|\frac{1}{m}\sum_{i=1}^m \widetilde{g}_i^k\right\|^2 \\
&\le \frac{\omega}{m^2}\sum_{i=1}^m \|\widetilde{g}_i^k\|^2 + \|\nabla f(x^k)\|^2 \\
&= \frac{\omega}{m^2}\sum_{i=1}^m \|\nabla f_i(x^k)\|^2 + \|\nabla f(x^k)\|^2 (*) \\
&\le \frac{\omega\zeta^2}{m} + (\frac{\omega}{m} + 1)\|\nabla f(x^k)\|^2
\end{aligned}$$

$$\mathbb{E}_k[\langle \nabla f(x^k), x^{k+1} - x^k \rangle] = -\eta \|\nabla f(x^k)\|^2$$

$$\mathbb{E}_k \|x^{k+1} - x^k\|^2 \leq \eta^2 \left[\frac{\omega \zeta^2}{m} + (\frac{\omega}{m} + 1) \|\nabla f(x^k)\|^2 \right]$$

so

$$\begin{aligned}
\mathbb{E}_k[f(x^{k+1})] &\leq f(x^k) + \mathbb{E}_k[\langle \nabla f(x^k), x^{k+1} - x^k \rangle] + \frac{L}{2} \mathbb{E}_k \|x^{k+1} - x^k\|^2 \\
&\leq f(x^k) - \eta \|\nabla f(x^k)\|^2 + \frac{L\eta^2}{2} \left[\frac{\omega \zeta^2}{m} + (\frac{\omega}{m} + 1) \|\nabla f(x^k)\|^2 \right] \\
&\leq f(x^k) - (\eta - \frac{L\eta^2}{2}(\frac{\omega}{m} + 1)) \|\nabla f(x^k)\|^2 + \frac{L\eta^2 \omega \zeta^2}{2m} \\
\mathbb{E}_k[f(x^{k+1}) - f^*] &\leq (f(x^k) - f^*) - (\eta - \frac{L\eta^2}{2}(\frac{\omega}{m} + 1)) \|\nabla f(x^k)\|^2 + \frac{L\eta^2 \omega \zeta^2}{2m}
\end{aligned}$$

Let $C = \frac{L\eta^2 \omega \zeta^2}{2m}$, $\eta' = \eta - \frac{L\eta^2}{2}(\frac{\omega}{m} + 1)$, $\quad \Delta^k = f(x^k) - f^*$,

$$\mathbb{E}[\Delta^{k+1}] \leq \mathbb{E}[\Delta^k] - \eta' \mathbb{E} \|\nabla f(x^k)\|^2 + C$$

$$i.e. \quad \eta' \mathbb{E}[\|\nabla f(x^k)\|^2] \leq \mathbb{E}[\Delta^k] - \mathbb{E}[\Delta^{k+1}] + C, \quad \forall 0 \leq k \leq K - 1$$

By suming up,

$$\sum_{k=0}^{K-1} \eta' [\|\nabla f(x^k)\|^2] \leq \mathbb{E}[\Delta^0] - \mathbb{E}[\Delta^K] + KC \leq \mathbb{E}[\Delta^0] + KC$$

$$\eta' \mathbb{E}[\|\nabla f(\widehat{x})\|^2] \leq \frac{\Delta^0}{K} + C$$

where $\widehat{x}$ randomly chosen from $\{x^k\}_{k=0}^{K-1}$ with probability $p_k = \frac{1}{K}$ for x^k.

$$\begin{aligned}
\eta' =& \eta - \frac{L\eta^2}{2}(\frac{\omega}{m} + 1) \\
\geq& \frac{\eta}{2} \quad (c)
\end{aligned}$$

we need $\eta \leq \frac{1}{L(1 + \frac{\omega}{m})}$, then (c) holds.

$$\begin{aligned}
\mathbb{E}[\|\nabla f(\widehat{x})\|^2] &\leq \frac{2\Delta^0}{\eta K} + \frac{L\eta \omega \zeta^2}{m} \\
&\leq \mathcal{O}\left(\frac{\Delta_0 L}{K}(1 + \frac{\omega}{m}) + \frac{\omega \zeta^2}{m + \omega} \right)
\end{aligned}$$

2.3 DIANA framework

2.3.1 Algorithm

Considering any stationary point $\widehat{x}$ such that $\nabla f(\widehat{x}) = \sum_{i=1}^m \nabla f_i(\widehat{x}) = 0$, the aggregated compressed gradient (even if the full gradient is used locally, i.e., $\widetilde{g}_i^k = \nabla f_i(x^k)$), is not equal to zero 0, i.e., $g(\widehat{x}) = \frac{1}{m} \sum_{i=1}^m \mathcal{C}_i(\nabla f_i(\widehat{x})) \neq 0$. This effect slows down convergence of the methods in DC framework. To address this issue, we use the DIANA framework to compress the gradient differences instead.

Algorithm 4 DIANA

Input: initial point x^0, $\{h_i^0\}_{i=1}^m, h^0 = \frac{1}{m} \sum_{i=1}^m h_i^0$, stepsize η, α

1: **for** $k = 0, 1, 2, \cdots$ **do**

for all machines $i = 0, 1, 2, \cdots, m$ in parallel **do**
  Compute local stochastic gradient $\tilde{g}_i^k$
  $\widehat{\Delta}_i^k = \mathcal{C}_i^k(\widetilde{g}_i^k - h_i^k)$
  $h_i^{k+1} = h_i^k + \alpha \widehat{\Delta}_i^k$
end for
$g^k = h^k + \frac{1}{m}\sum_{i=1}^m \widehat{\Delta}_i^k$
$x^{k+1} = x^k - \eta g^k$
$h^{k+1} = h^k + \alpha \frac{1}{m}\sum_{i=1}^m \widehat{\Delta}_i^k$
end for

2.3.2 Analysis

Theorem. (DIANA framework) If the local stochastic gradient $\widetilde{g}_i^k$ satisfes the recursions:

$$\mathbb{E}_k[\|\widetilde{g}_i^k\|^2] \leq 2A_{1,i}(f_i(x^k) - f_i^*) + B_{1,i}\|\nabla f_i(x^k)\|^2 + D_{1,i}\sigma_{k,i}^2 + C_{1,i},$$
$$\mathbb{E}_k[\sigma_{k+1,i}^2] \leq (1-\rho_i)\sigma_{k,i}^2 + 2A_{2,i}(f(x^k) - f^*) + B_{2,i}\|\nabla f(x^k)\|^2 + D_{2,i}\mathbb{E}_k[\|g^k\|^2] + C_{2,i},$$

then g^k satisfes the unified Assumption with

$$A_1 = \frac{(1+\omega)A}{m}, \quad B_1 = 1, \quad D_1 = \frac{1+\omega}{m}, \quad \sigma_k^2 = \frac{1}{m}\sum_{i=1}^m D_{1,i}\sigma_{k,i}^2 + \frac{\omega}{(1+\omega)m}\sum_{i=1}^m \|\nabla f_i(x^k) - h_i^k\|^2,$$

$$C_1 = \frac{(1+\omega)C}{m}, \quad \rho = \min\left\{\min_i \rho_i - \tau,\ 2\alpha - (1-\alpha)\beta^{-1} - \alpha^2 - \tau\right\},$$
$$A_2 = D_A + \tau A, \quad B_2 = D_B + B, \quad C_2 = D_C + \tau C$$

where

$$A := \max_i(A_{1,i} + (B_{1,i} - 1)L_i), \quad B := \frac{\omega(1+\beta)L^2\eta^2}{1+\omega} + D_D, \quad C := \frac{1}{m}\sum_{i=1}^m C_{1,i} + 2A\Delta_f^*,$$

$$\Delta_f^* := f^* - \frac{1}{m}\sum_{i=1}^m f_i^*, \quad \tau := \alpha^2\omega + \frac{(1+\omega)B}{m}, \quad \forall \beta > 0$$

$$D_A := \frac{1}{m}\sum_{i=1}^m D_{1,i}A_{2,i}, \quad D_B := \frac{1}{m}\sum_{i=1}^m D_{1,i}B_{2,i},$$

$$D_C := \frac{1}{m}\sum_{i=1}^m D_{1,i}C_{2,i}, \quad D_D := \frac{1}{m}\sum_{i=1}^m D_{1,i}D_{2,i}.$$

2.3.3 DIANA-GD

$\widetilde{g}_i^k = \nabla f_i(x^k)$, so

$$\mathbb{E}_k[\widetilde{g}_i^k] = \nabla f_i(x^k),$$

$$\mathbb{E}_k[\|\widetilde{g}_i^k\|^2] \leq \|\nabla f_i(x^k)\|^2$$

Then,

$$\begin{aligned}
\mathbb{E}_k[g^k] &= \mathbb{E}_k\left[h^k + \frac{1}{m}\sum_{i=1}^m \widehat{\Delta}_i^k\right] \\
&= \mathbb{E}_k\left[\frac{1}{m}\sum_{i=1}^m h_i^k + \frac{1}{m}\sum_{i=1}^m \mathcal{C}_i^k(\widetilde{g}_i^k - h_i^k)\right] \\
&= \mathbb{E}_k\left[\frac{1}{m}\sum_{i=1}^m h_i^k + \frac{1}{m}\sum_{i=1}^m (\widetilde{g}_i^k - h_i^k)\right] \\
&= \mathbb{E}_k\left[\frac{1}{m}\sum_{i=1}^m \widetilde{g}_i^k\right] \\
&= \nabla f(x^k)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_k[\|g^k\|^2] &= \mathbb{E}_k\left[\left\|\frac{1}{m}\sum_{i=1}^m h_i^k + \frac{1}{m}\sum_{i=1}^m \mathcal{C}_i^k(\widetilde{g}_i^k - h_i^k)\right\|^2\right] \\
&= \mathbb{E}_k\left[\left\|\frac{1}{m}\sum_{i=1}^m h_i^k + \frac{1}{m}\sum_{i=1}^m \mathcal{C}_i^k(\widetilde{g}_i^k - h_i^k) - \frac{1}{m}\sum_{i=1}^m \widetilde{g}_i^k + \frac{1}{m}\sum_{i=1}^m \widetilde{g}_i^k\right\|^2\right] \\
&= \mathbb{E}_k\left[\left\|\frac{1}{m}\sum_{i=1}^m \left(\mathcal{C}_i^k(\widetilde{g}_i^k - h_i^k) - (\widetilde{g}_i^k - h_i^k)\right)\right\|^2\right] + \mathbb{E}_k\left[\left\|\frac{1}{m}\sum_{i=1}^m \widetilde{g}_i^k\right\|^2\right] \\
&\leq \frac{\omega}{m^2}\mathbb{E}_k\left[\sum_{i=1}^m \|\widetilde{g}_i^k - h_i^k\|^2\right] + \mathbb{E}_k\left[\left\|\frac{1}{m}\sum_{i=1}^m (\widetilde{g}_i^k - \nabla f_i(x^k)) + \nabla f(x^k)\right\|^2\right] \\
&= \frac{\omega}{m^2}\mathbb{E}_k\left[\sum_{i=1}^m \|\widetilde{g}_i^k - h_i^k\|^2\right] + \|\nabla f(x^k)\|^2 \\
&= \frac{\omega}{m^2}\mathbb{E}_k\left[\sum_{i=1}^m \|\widetilde{g}_i^k - \nabla f_i(x^k) + \nabla f_i(x^k) - h_i^k\|^2\right] + \|\nabla f(x^k)\|^2 \\
&= \frac{\omega}{m^2}\left(\sum_{i=1}^m \|\nabla f_i(x^k) - h_i^k\|^2\right) + \|\nabla f(x^k)\|^2 (*)
\end{aligned}$$

Let $\sigma_k^2 = \dfrac{\omega}{m^2}\sum_{i=1}^m \|\nabla f_i(x^k) - h_i^k\|^2$,

$$\mathbb{E}_k[\|g^k\|^2] \leq \|\nabla f(x^k)\|^2 + \sigma_k^2$$

$$
\begin{aligned}
&\mathbb{E}_k[\sigma_{k+1}^2] \\
=&\frac{\omega}{m^2}\mathbb{E}_k\left[\sum_{i=1}^m \|\nabla f_i(x^{k+1}) - h_i^{k+1}\|^2\right] \\
=&\frac{\omega}{m^2}\mathbb{E}_k\left[\sum_{i=1}^m \|\nabla f_i(x^{k+1}) - \nabla f_i(x^k) + \nabla f_i(x^k) - h_i^{k+1}\|^2\right] \\
=&\frac{\omega}{m^2}\mathbb{E}_k\left[\sum_{i=1}^m \|\nabla f_i(x^{k+1}) - \nabla f_i(x^k) + \nabla f_i(x^k) - h_i^k - \alpha\mathcal{C}_i^k(\widetilde{g}_i^k - h_i^k)\|^2\right] \\
=&\frac{\omega}{m^2}\sum_{i=1}^m \mathbb{E}_k\left[\|\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\|^2 + \|\nabla f_i(x^k) - h_i^k - \alpha\mathcal{C}_i^k(\widetilde{g}_i^k - h_i^k)\|^2\right. \\
&\left.+2\langle \nabla f_i(x^{k+1}) - \nabla f_i(x^k), \nabla f_i(x^k) - h_i^k - \alpha\mathcal{C}_i^k(\widetilde{g}_i^k - h_i^k)\rangle\right] \\
=&\frac{\omega}{m^2}\sum_{i=1}^m \mathbb{E}_k\left[\|\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\|^2 + \|\nabla f_i(x^k) - h_i^k - \alpha\mathcal{C}_i^k(\widetilde{g}_i^k - h_i^k)\|^2\right. \\
&\left.+2\langle \nabla f_i(x^{k+1}) - \nabla f_i(x^k), (1-\alpha)(\nabla f_i(x^k) - h_i^k)\rangle\right] \\
=&\frac{\omega}{m^2}\sum_{i=1}^m \mathbb{E}_k\left[\|\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\|^2 + (1-2\alpha)\|\nabla f_i(x^k) - h_i^k\|^2 + \alpha^2\|\mathcal{C}_i^k(\widetilde{g}_i^k - h_i^k)\|^2\right. \\
&\left.+2\langle \nabla f_i(x^{k+1}) - \nabla f_i(x^k), (1-\alpha)(\nabla f_i(x^k) - h_i^k)\rangle\right] \\
\leq&\frac{\omega}{m^2}\sum_{i=1}^m \mathbb{E}_k\Big[\|\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\|^2 + (1-2\alpha)\|\nabla f_i(x^k) - h_i^k\|^2 + \alpha^2\|\mathcal{C}_i^k(\widetilde{g}_i^k - h_i^k)\|^2 \\
&+\beta\|\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\|^2 + \frac{(1-\alpha)^2}{\beta}\|\nabla f_i(x^k) - h_i^k\|^2\Big] \quad \forall\beta > 0 \\
=&\frac{\omega}{m^2}\sum_{i=1}^m \mathbb{E}_k\Big[(1+\beta)\|\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\|^2 + (1-2\alpha+\frac{(1-\alpha)^2}{\beta})\|\nabla f_i(x^k) - h_i^k\|^2 \\
&+\alpha^2\|\mathcal{C}_i^k(\widetilde{g}_i^k - h_i^k)\|^2\Big] \\
\leq&\frac{\omega}{m^2}\sum_{i=1}^m \mathbb{E}_k\Big[(1+\beta)\|\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\|^2 + (1-2\alpha+\frac{(1-\alpha)^2}{\beta})\|\nabla f_i(x^k) - h_i^k\|^2 \\
&+\alpha^2(1+\omega)\|\widetilde{g}_i^k - h_i^k\|^2\Big] \\
=&\frac{\omega}{m^2}\sum_{i=1}^m \mathbb{E}_k\Big[(1+\beta)\|\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\|^2 + (1-2\alpha+\frac{(1-\alpha)^2}{\beta})\|\nabla f_i(x^k) - h_i^k\|^2 \\
&+\alpha^2(1+\omega)\|\widetilde{g}_i^k - \nabla f_i(x^k) + \nabla f_i(x^k) - h_i^k\|^2\Big] \\
=&\left(1-2\alpha+\frac{(1-\alpha)^2}{\beta}+\alpha^2(1+\omega)\right)\sigma_k^2 + \frac{\omega(1+\beta)}{m^2}\sum_{i=1}^m \mathbb{E}_k[\|\nabla f_i(x^{k+1}) - \nabla f_i(x^k)\|^2] \\
\leq&\left(1-2\alpha+\frac{(1-\alpha)^2}{\beta}+\alpha^2(1+\omega)\right)\sigma_k^2 + \frac{\omega(1+\beta)}{m^2}\sum_{i=1}^m L_i^2\mathbb{E}_k[\|x^{k+1} - x^k\|^2] \\
=&\left(1-2\alpha+\frac{(1-\alpha)^2}{\beta}+\alpha^2(1+\omega)\right)\sigma_k^2 + \frac{\omega(1+\beta)}{m^2}\sum_{i=1}^m L_i^2\eta^2\mathbb{E}_k[\|g^k\|^2] \\
\leq&\left(1-2\alpha+\frac{(1-\alpha)^2}{\beta}+\alpha^2(1+\omega)+\omega(1+\beta)A^2\eta^2\right)\sigma_k^2 + \frac{\omega(1+\beta)}{m}A^2\eta^2\|\nabla f(x^k)\|^2 \\
\leq&(1-\rho)\sigma_k^2 + B\|\nabla f(x^k)\|^2
\end{aligned}
$$

where $A = \max_i(L_i), B = \frac{\omega(1+\beta)}{m}A^2\eta^2, \rho \leq 2\alpha - \frac{(1-\alpha)^2}{\beta} - \alpha^2(1+\omega) - mB$

Now we have

$$\mathbb{E}_k[\|g^k\|] = \nabla f(x^k)$$

$$\mathbb{E}_k[\|g^k\|^2] \leq \|\nabla f(x^k)\|^2 + \sigma_k^2$$

$$\mathbb{E}_k[\sigma_{k+1}^2] \leq (1-\rho)\sigma_k^2 + B\|\nabla f(x^k)\|^2$$

Then,

$$\mathbb{E}_k[\langle \nabla f(x^k), x^{k+1} - x^k\rangle] = -\eta\|\nabla f(x^k)\|^2$$

$$\mathbb{E}_k\|x^{k+1} - x^k\|^2 \leq \eta^2[\|\nabla f(x^k)\|^2 + \sigma_k^2]$$

so

$$\begin{aligned}\mathbb{E}_k[f(x^{k+1})] &\leq f(x^k) + \mathbb{E}_k[\langle \nabla f(x^k), x^{k+1} - x^k\rangle] + \frac{L}{2}\mathbb{E}_k\|x^{k+1} - x^k\|^2 \\ &\leq f(x^k) - (\eta - \frac{L\eta^2}{2})\|\nabla f(x^k)\|^2 + \frac{L\eta^2}{2}\sigma_k^2\end{aligned}$$

$$\mathbb{E}_k[f(x^{k+1}) - f^* + \alpha\sigma_{k+1}^2] \leq (f(x^k) - f^*) + \left(\frac{L\eta^2}{2} + \alpha(1-\rho)\right)\sigma_k^2 - \left(\eta - \frac{L\eta^2}{2} - \alpha B\right)\|\nabla f(x^k)\|^2$$

Let $\eta' = \eta - \frac{L\eta^2}{2} - \alpha B, \quad \Delta^k = f(x^k) - f^* + \alpha\sigma_k^2, \quad \alpha = \frac{L\eta^2}{2\rho}$

$$\begin{aligned}\mathbb{E}[\Delta^{k+1}] \leq& \mathbb{E}[f(x^k) - f^* + \left(\frac{L\eta^2}{2} + \alpha(1-\rho)\right)\sigma_k^2] - \eta'\mathbb{E}\|\nabla f(x^k)\|^2 \\ =& \mathbb{E}[\Delta^k] - \eta'\mathbb{E}\|\nabla f(x^k)\|^2\end{aligned}$$

$$i.e. \quad \eta'\mathbb{E}[\|\nabla f(x^k)\|^2] \leq \mathbb{E}[\Delta^k] - \mathbb{E}[\Delta^{k+1}], \quad \forall 0 \leq k \leq K-1$$

By suming up,

$$\sum_{k=0}^{K-1}\eta'[\|\nabla f(x^k)\|^2] \leq \mathbb{E}[\Delta^0] - \mathbb{E}[\Delta^K] \leq \mathbb{E}[\Delta^0]$$

$$\eta'\mathbb{E}[\|\nabla f(\widehat{x})\|^2] \leq \frac{\Delta_0}{K}$$

where $\widehat{x}$ randomly chosen from $\{x^k\}_{k=0}^{K-1}$ with probability $p_k = \frac{1}{K}$ for x^k.

$$\begin{aligned}\eta' =& \eta - \frac{L\eta^2}{2} - \alpha B \\ =& \eta - \frac{L\eta^2}{2} - \frac{L\eta^2 B}{2\rho} \\ =& \eta - \frac{\eta}{2}(L\eta + L\eta B\rho^{-1}) \\ \geq& \frac{\eta}{2} \quad (c)\end{aligned}$$

we need $\eta \leq \frac{1}{L(1+B\rho^{-1})}$, then (c) holds, and then

When $K \geq \frac{2\Delta^0}{\eta\epsilon^2}$, $\mathbb{E}[\|\nabla f(\widehat{x})\|] \leq \sqrt{\mathbb{E}[\|\nabla f(\widehat{x})\|^2]} \leq \epsilon$.

We need to satisify:

$$1.\rho \leq \alpha \iff \alpha \geq \sqrt{\frac{L}{2}}\eta$$

$$2.\alpha \geq \frac{(1-\alpha)^2}{\beta} - \alpha^2(1+w) - w(1+\beta)A^2\eta^2$$

$$3.\eta \leq \frac{1}{L + \frac{2\alpha w(1+\beta)A^2}{m}}$$

Let $\alpha \geq \sqrt{\frac{L}{2}}\eta, \beta \geq \max\{\frac{1-\alpha}{\sqrt{2w}A\eta}, \frac{2\alpha}{(1-\alpha)^2}\}$, 1,2 are satisfied.

Let $\alpha = \sqrt{\frac{L}{2}}\eta, 1+\beta = \frac{1-\alpha}{\sqrt{2wA\eta}} + \frac{1+\alpha^2}{(1-\alpha)^2}, \eta = \min\{\frac{m}{2(mL+\sqrt{wL}A)}, \frac{mL+2\sqrt{wL}A}{4wA^2\sqrt{\frac{L}{2}}}, \frac{1}{\sqrt{8L}}\}$, 3 is satisfied.

$$Km \sim \mathcal{O}\left(\frac{\Delta^0}{\epsilon^2}(mL + \sqrt{wL}A + \frac{mwA^2}{m\sqrt{L}+\sqrt{w}A})\right)$$

§ 3 FedAvg, SCAFFOLD, SAGA(nonconvex)

3.1 FedAvg

3.1.1 Problem

We consider

$$x^* = \arg\min_{x\in\mathbb{R}^d}[f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)] \tag{15}$$

$f_i(x) = \mathbb{E}_{\zeta_i}[f_i(x;\zeta_i)] : \mathbb{R}^d \to \mathbb{R}$ is β-smooth:

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le \beta\|x - y\|, \quad \forall i, x, y \tag{16}$$

$g_i(x) = \nabla f_i(x;\zeta_i)$ is unbiased with variance bounded by σ^2:

$$\mathbb{E}_{\zeta_i}[\|g_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x})\|^2] \le \sigma^2 \text{ , for any } i, \boldsymbol{x} \tag{17}$$

(G,B)-BGD or bounded gradient dissimilarity: there exist constants $G \ge 0$ and $B \ge 1$ such that

$$\frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(\boldsymbol{x})\|^2 \le G^2 + B^2\|\nabla f(\boldsymbol{x})\|^2, \forall \boldsymbol{x} \tag{18}$$

$\mathbb{E}_r$ represents the conditional expectation of the information known at the beginning of round r, given that round $r-1$ has just ended and round r is just starting. Given $y^r_{i,k-1}$ and x^r, $g_i(y^r_{i,k-1})$ is unknown.

3.1.2 Algorithm

Algorithm 5 FedAvg

Input: initial point x^0, stepsize η, γ

for each round $r = 0, \cdots, R-1$ **do**
- sample clients $\mathcal{S}^r \subseteq [N]$, $|\mathcal{S}^r| = S$
- **for** client $i \in \mathcal{S}^r$ in parallel **do**
  - $y^r_{i,0} = x^r$
  - **for** $k = 1, \cdots, K$ **do**
    - Compute local stochastic gradient $g_i(y^r_{i,k-1})$
    - $y^r_{i,k} = y^r_{i,k-1} - \eta g_i(y^r_{i,k-1})$
  - **end for**
  - $\Delta y^r_i = y^r_{i,K} - x^{r-1}$
- **end for**
- $\Delta x^r = \frac{\gamma}{S}\sum_{i\in\mathcal{S}^r}\Delta y^r_i$
- $x^{r+1} = x^r + \Delta x^r$

end for

3.1.3 Analysis

Let $\gamma \ge 1$, $\tilde{\eta} = K\gamma\eta$,

$$\tilde{\eta} \le \frac{1}{(1+B^2)8\beta}, \ \beta\tilde{\eta} \le \frac{1}{8} \quad (*)$$

$$\mathcal{E}^r = \frac{1}{KN}\sum_{i,k}\mathbb{E}_r\|y^r_{i,k-1} - x^r\|^2$$

When $K, S, N > 1$,

$$
\begin{aligned}
\mathbb{E}_r\|y_{i,k}^r - x^r\|^2 &= \mathbb{E}_r\|y_{i,k-1}^r - x^r - \eta g_i(y_{i,k-1}^r)\|^2 \\
&\leq \mathbb{E}_r\|y_{i,k-1} - x - \eta\nabla f_i(y_{i,k-1})\|^2 + \eta^2\sigma^2 \\
&\leq (1 + \frac{1}{K-1})\mathbb{E}_r\|y_{i,k-1} - x\|^2 + K\eta^2\|\nabla f_i(y_{i,k-1})\|^2 + \eta^2\sigma^2 \\
&= (1 + \frac{1}{K-1})\mathbb{E}_r\|y_{i,k-1} - x\|^2 + K\eta^2\|\nabla f_i(y_{i,k-1})\|^2 + \eta^2\sigma^2 \\
&\leq (1 + \frac{1}{K-1})\mathbb{E}_r\|y_{i,k-1} - x\|^2 + 2K\eta^2\|\nabla f_i(y_{i,k-1}) - \nabla f_i(x)\|^2 + 2K\eta^2\|\nabla f_i(x)\|^2 + \eta^2\sigma^2 \\
&\leq (1 + \frac{1}{K-1} + K\eta^2\beta^2)\mathbb{E}_r\|y_{i,k-1} - x\|^2 + 2K\eta^2\|\nabla f_i(x)\|^2 + \eta^2\sigma^2
\end{aligned}
$$

Then,

$$
\begin{aligned}
\mathbb{E}_r\|y_{i,k}^r - x^r\|^2 &\leq (2K\eta^2\|\nabla f_i(x)\|^2 + \eta^2\sigma^2)\sum_{\tau=1}^{k-1}(1 + \frac{1}{K-1} + K\eta^2\beta^2)^\tau \\
&\leq (\frac{2\tilde{\eta}^2}{K\gamma^2}\|\nabla f_i(x)\|^2 + \frac{\tilde{\eta}^2\sigma^2}{K^2\gamma^2})3K \\
&= \frac{6\tilde{\eta}^2}{\gamma^2}\|\nabla f_i(x)\|^2 + \frac{3\tilde{\eta}^2\sigma^2}{K\gamma^2} \\
&\leq \frac{6\tilde{\eta}^2}{\gamma^2}\|\nabla f_i(x)\|^2 + \frac{3\tilde{\eta}\sigma^2}{8\beta K\gamma^2}
\end{aligned}
$$

$$
\begin{aligned}
\beta\tilde{\eta}\mathcal{E}_r &\leq \frac{6\beta\tilde{\eta}^3}{\gamma^2}\frac{1}{N}\sum_i\|\nabla f_i(x)\|^2 + \frac{3\tilde{\eta}^2\sigma^2}{8K\gamma^2} \\
&\leq \frac{6\beta\tilde{\eta}^3}{\gamma^2}(G^2 + B^2\|\nabla f(x)\|^2) + \frac{3\tilde{\eta}^2\sigma^2}{8K\gamma^2} \\
&= \frac{6\beta\tilde{\eta}^3G^2}{\gamma^2} + \frac{3\tilde{\eta}^2\sigma^2}{8K\gamma^2} + \frac{6\beta\tilde{\eta}^3B^2}{\gamma^2}\|\nabla f(x)\|^2
\end{aligned}
$$

some r not exist above.

Then,

$$
\Delta x^r = -\frac{\tilde{\eta}}{KS}\sum_{i\in\mathcal{S}^r}\sum_{k=1}^{K} g_i(y_{i,k-1}^r)
$$

$$
\mathbb{E}_r[\Delta x^r] = -\frac{\tilde{\eta}}{KN}\sum_{k,i}\mathbb{E}_r[\nabla f_i(y_{i,k-1}^r)]
$$

$$
\begin{aligned}
\mathbb{E}_r[\langle\nabla f(x^r), \Delta x^r\rangle] &= -\frac{\tilde{\eta}}{2}\left[\|\nabla f(x^r)\|^2 + \mathbb{E}_r\|\Delta x^r\|^2 - \mathbb{E}_r\|\frac{1}{KS}\sum_{i\in\mathcal{S}^r}\sum_{k=1}^{K}\left(\nabla f_i(y_{i,k-1}^r) - \nabla f(x^r)\right)\|^2\right] \\
&\leq -\frac{\tilde{\eta}}{2}\left[\|\nabla f(x^r)\|^2 + \mathbb{E}_r\|\Delta x^r\|^2\right] + \frac{\tilde{\eta}\beta^2\mathcal{E}^r}{2} \\
&\leq -\frac{\tilde{\eta}}{2}\|\nabla f(x^r)\|^2 + \frac{\tilde{\eta}\beta^2\mathcal{E}^r}{2}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_r\|\Delta x^r\|^2 &= \tilde{\eta}^2\mathbb{E}_r\|\frac{1}{KS}\sum_{i\in\mathcal{S}^r}\sum_{k=1}^{K} g_i(y_{i,k-1}^r)\|^2 \\
&= \tilde{\eta}^2\mathbb{E}_r\|\frac{1}{KS}\sum_{i\in\mathcal{S}^r}\sum_{k=1}^{K}\nabla f_i(y_{i,k-1}^r)\|^2 + \frac{\tilde{\eta}^2\sigma^2}{KS} \\
&\le 2\tilde{\eta}^2\mathbb{E}_r\|\frac{1}{KS}\sum_{k,i}\nabla f_i(y_{i,k-1}^r)-\nabla f_i(x^r)\|^2 + 2\tilde{\eta}^2\mathbb{E}_r\|\frac{1}{S}\sum_{i\in\mathcal{S}}\nabla f_i(x^r)\|^2 + \frac{\tilde{\eta}^2\sigma^2}{KS} \\
&\le \frac{2\tilde{\eta}^2}{KN}\sum_{i,k}\mathbb{E}_r\|\nabla f_i(y_{i,k-1}^r)-\nabla f_i(x^r)\|^2 + 2\tilde{\eta}^2\mathbb{E}_r\|\frac{1}{S}\sum_{i\in\mathcal{S}}\nabla f_i(x^r)-\nabla f(x^r)+\nabla f(x^r)\|^2 + \frac{\tilde{\eta}^2\sigma^2}{KS} \\
&\le \frac{2\tilde{\eta}^2\beta^2}{KN}\sum_{i,k}\mathbb{E}_r\|y_{i,k-1}^r - x^r\|^2 + 2\tilde{\eta}^2\|\nabla f(x^r)\|^2 + (1-\frac{S}{N})4\tilde{\eta}^2\frac{1}{SN}\sum_i\|\nabla f_i(x^r)\|^2 + \frac{\tilde{\eta}^2\sigma^2}{KS} \\
&\le 2\tilde{\eta}^2\beta^2\mathcal{E}^r + 2\tilde{\eta}^2\|\nabla f(x^r)\|^2 + (1-\frac{S}{N})\frac{4\tilde{\eta}^2}{S}(G^2+B^2\|\nabla f(x^r)\|^2) + \frac{\tilde{\eta}^2\sigma^2}{KS} \\
&= 2\tilde{\eta}^2\beta^2\mathcal{E}^r + 2\tilde{\eta}^2(1+(\frac{2}{S}-\frac{2}{N})B^2)\|\nabla f(x^r)\|^2 + (1-\frac{S}{N})\frac{4\tilde{\eta}^2}{S}G^2 + \frac{\tilde{\eta}^2\sigma^2}{KS} \\
&\le 2\tilde{\eta}^2\beta^2\mathcal{E}^r + 2\tilde{\eta}^2(1+B^2)\|\nabla f(x^r)\|^2 + (1-\frac{S}{N})\frac{4\tilde{\eta}^2}{S}G^2 + \frac{\tilde{\eta}^2\sigma^2}{KS}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_r[f(x^{r+1})] \le& f(x^r) + \mathbb{E}_r[\langle\nabla f(x^r),\Delta x^r\rangle] + \frac{\beta}{2}\mathbb{E}_r\|\Delta x^r\|^2 \\
\le& f(x^r) - \frac{\tilde{\eta}}{2}\|\nabla f(x^r)\|^2 + \frac{\tilde{\eta}\beta^2\mathcal{E}^r}{2} + \frac{\beta}{2}\mathbb{E}_r\|\Delta x^r\|^2 \\
\le& f(x^r) - \frac{\tilde{\eta}}{2}\|\nabla f(x^r)\|^2 + \frac{\tilde{\eta}\beta^2\mathcal{E}^r}{2} \\
&+ \frac{\beta}{2}\left[2\tilde{\eta}^2\beta^2\mathcal{E}^r + 2\tilde{\eta}^2(1+B^2)\|\nabla f(x^r)\|^2 + (1-\frac{S}{N})\frac{4\tilde{\eta}^2}{S}G^2 + \frac{\tilde{\eta}^2\sigma^2}{KS}\right] \\
\le& f(x^r) + \tilde{\eta}\left[-\frac{1}{2}+\beta\tilde{\eta}(1+B^2)\right]\|\nabla f(x^r)\|^2 \\
&+ \beta\left[\frac{1}{2}+\beta\tilde{\eta}\right](\frac{6\beta\tilde{\eta}^3G^2}{\gamma^2} + \frac{3\tilde{\eta}^2\sigma^2}{8K\gamma^2} + \frac{6\beta\tilde{\eta}^3B^2}{\gamma^2}\|\nabla f(x)\|^2) \\
&+ \frac{\beta}{2}\left[(1-\frac{S}{N})\frac{4\tilde{\eta}^2}{S}G^2 + \frac{\tilde{\eta}^2\sigma^2}{KS}\right] \\
=& f(x^r) - \eta'\|\nabla f(x^r)\|^2 + C
\end{aligned}
$$

$$
\begin{aligned}
\eta' =& \tilde{\eta}\left[\frac{1}{2}-\beta\tilde{\eta}(1+B^2)\right] - \beta\left[\frac{1}{2}+\beta\tilde{\eta}\right]\frac{6\beta\tilde{\eta}^3B^2}{\gamma^2} \\
\ge& \frac{3}{8}\tilde{\eta} - \left[\frac{1}{2}+\beta\tilde{\eta}\right]\frac{3\tilde{\eta}B^2}{32\gamma^2} \\
\ge& \frac{3}{16}\tilde{\eta}
\end{aligned}
$$

$$
C = \beta\left[\frac{1}{2}+\beta\tilde{\eta}\right](\frac{6\beta\tilde{\eta}^2G^2}{\gamma^2} + \frac{3\tilde{\eta}\sigma^2}{8K\gamma^2}) + \frac{\beta}{2}\left[(1-\frac{S}{N})\frac{4\tilde{\eta}^2}{S}G^2 + \frac{\tilde{\eta}^2\sigma^2}{KS}\right]
$$

$$
\mathbb{E}_r[f(x^{r+1})-f^*] \le (f(x^r)-f^*) - \eta'\|\nabla f(x^r)\|^2 + C
$$

By suming $r\in\{0,1,\cdots,R-1\}$ and let $F=f(x^0)-f^*$,

$$
\eta'\mathbb{E}[\frac{1}{R}\sum_{r=0}^{R-1}\|\nabla f(x^r)\|^2] \le \frac{F}{R} + C
$$

i.e.

$$
\begin{aligned}
\frac{3}{16}\mathbb{E}[\frac{1}{R}\sum_{r=0}^{R-1}\|\nabla f(x^r)\|^2] \leq & \frac{F}{\tilde\eta R} + \beta\left[\frac{1}{2}+\beta\tilde\eta\right](\frac{6\beta\tilde\eta^2 G^2}{\gamma^2}+\frac{3\tilde\eta\sigma^2}{8K\gamma^2}) + \frac{\beta}{2}\left[(1-\frac{S}{N})\frac{4\tilde\eta}{S}G^2+\frac{\tilde\eta\sigma^2}{KS}\right] \\
\leq & \frac{F}{\tilde\eta R} + \beta\frac{5}{8}(\frac{6\beta\tilde\eta^2 G^2}{\gamma^2}+\frac{3\tilde\eta\sigma^2}{8K\gamma^2}) + \frac{\beta}{2}\left[(1-\frac{S}{N})\frac{4\tilde\eta}{S}G^2+\frac{\tilde\eta\sigma^2}{KS}\right] \\
= & \frac{F}{\tilde\eta R} + (\frac{15\beta^2\tilde\eta^2 G^2}{4\gamma^2}+\frac{15\beta\tilde\eta\sigma^2}{64K\gamma^2}) + \frac{\beta}{2}\left[(1-\frac{S}{N})\frac{4\tilde\eta}{S}G^2+\frac{\tilde\eta\sigma^2}{KS}\right]
\end{aligned}
$$

By

Lemma 2 (sub-linear convergence rate). *For every non-negative sequence* $\{d_{r-1}\}_{r\geq 1}$ *and any parameters* $\eta_{\max} \geq 0$, $c \geq 0$, $R \geq 0$, *there exists a constant step-size* $\eta \leq \eta_{\max}$ *and weights* $w_r = 1$ *such that,*

$$
\Psi_R := \frac{1}{R+1}\sum_{r=1}^{R+1}\left(\frac{d_{r-1}}{\eta} - \frac{d_r}{\eta} + c_1\eta + c_2\eta^2\right) \leq \frac{d_0}{\eta_{\max}(R+1)} + \frac{2\sqrt{c_1 d_0}}{\sqrt{R+1}} + 2\left(\frac{d_0}{R+1}\right)^{\frac{2}{3}} c_2^{\frac{1}{3}} .
$$

Proof. Unrolling the sum, we can simplify

$$
\Psi_R \leq \frac{d_0}{\eta(R+1)} + c_1\eta + c_2\eta^2 .
$$

Similar to the strongly convex case (Lemma 1), we distinguish the following cases:

- When $R+1 \leq \frac{d_0}{c_1\eta_{\max}^2}$, and $R+1 \leq \frac{d_0}{c_2\eta_{\max}^3}$ we pick $\eta = \eta_{\max}$ to claim

$$
\Psi_R \leq \frac{d_0}{\eta_{\max}(R+1)} + c_1\eta_{\max} + c_2\eta_{\max}^2 \leq \frac{d_0}{\eta_{\max}(R+1)} + \frac{\sqrt{c_1 d_0}}{\sqrt{R+1}} + \left(\frac{d_0}{R+1}\right)^{\frac{2}{3}} c_2^{\frac{1}{3}} .
$$

- In the other case, we have $\eta_{\max}^2 \geq \frac{d_0}{c_1(R+1)}$ or $\eta_{\max}^3 \geq \frac{d_0}{c_2(R+1)}$. We choose $\eta = \min\left\{\sqrt{\frac{d_0}{c_1(R+1)}}, \sqrt[3]{\frac{d_0}{c_2(R+1)}}\right\}$ to prove

$$
\Psi_R \leq \frac{d_0}{\eta(R+1)} + c\eta = \frac{2\sqrt{c_1 d_0}}{\sqrt{R+1}} + 2\sqrt[3]{\frac{d_0^2 c_2}{(R+1)^2}} .
$$

□

Let $d_0 = F,\ c_1 = \frac{\beta}{2}\left[(1-\frac{S}{N})\frac{4}{S}G^2 + \frac{\sigma^2}{KS}\right] + \frac{15\beta\sigma^2}{64K\gamma^2},\ c_2 = \frac{15\beta^2 G^2}{4\gamma^2}$

Finally, we get

$$
\mathbb{E}[\|\nabla f(\bar x^R)\|^2] \leq \mathcal{O}\left(\frac{M\sqrt{\beta F}}{\sqrt{RKS}} + \frac{F^{2/3}(\beta^2 G^2)^{1/3}}{R^{2/3}} + \frac{(B^2+1)\beta F}{R}\right),
$$

where $M^2 = \sigma^2(1+\frac{S}{\gamma^2}) + K(1-\frac{S}{N})G^2$.

3.2 SCAFFOLD

3.2.1 Algorithm

Algorithm 6 SCAFFOLD

Input: initial point $x^0, c_i, i \in [N], c = \frac{1}{N}\sum_{i=1}^N c_i$, stepsize η_l, η_g

```
for each round r = 1, ··· , R do
    sample clients S^r ⊆ [N]
    for client i ∈ S^r in parallel do
        y^r_{i,0} = x^{r-1}
        for k = 1, ··· , K do
            Compute local stochastic gradient g_i(y^r_{i,k-1})
            y^r_{i,k} = y^r_{i,k-1} − η_l(g_i(y^r_{i,k-1}) − c^r_i + c^r)
        end for
        c^+_i = c^r_i − c^r + 1/(Kη_l) (x^{r-1} − y^r_{i,K})
        (Δy^r_i, Δc^r_i) = (y^r_{i,K} − x^{r-1}, c^+_i − c^r_i)
        c^r_i = c^+_i
    end for
    (Δx^r, Δc^r) = 1/|S^r| Σ_{i∈S^r} (Δy^r_i, Δc^r_i)
    x^r = x^{r-1} + η_g Δx^r and c^r = c^{r-1} + |S^r|/N Δc^r
end for
```

3.2.2 Analysis

Additional notation.

In round r,

$$c_i^r = \begin{cases} \frac{1}{K}\sum_{k=1}^K g_i(y_{i,k-1}^r) & \text{if } i \in \mathcal{S}^r, \\ c_i^{r-1} & \text{otherwise.} \end{cases}$$

$$\alpha_{i,k-1}^0 := x^0, \quad \alpha_{i,k-1}^r := \begin{cases} y_{i,k-1}^r & \text{if } i \in \mathcal{S}^r, \\ \alpha_{i,k-1}^{r-1} & \text{otherwise.} \end{cases}$$

$$\Xi_r := \frac{1}{KN}\sum_{k=1}^K\sum_{i=1}^N \mathbb{E}\|\alpha_{i,k-1}^r - x^r\|^2.$$

$$\mathcal{E}_r := \frac{1}{KN}\sum_{k=1}^K\sum_{i=1}^N \mathbb{E}\|y_{i,k}^r - x^{r-1}\|^2.$$

Lemma 1. $\forall \tilde{\eta} := \eta_l \eta_g K \in [0, \frac{1}{\beta}]$,

$$\mathbb{E}\|\mathbb{E}_{r-1}[x^r] - x^{r-1}\|^2 \leq 2\tilde{\eta}^2\beta^2\mathcal{E}_r + 2\tilde{\eta}^2\mathbb{E}\|\nabla f(x^{r-1})\|^2,$$

$$\mathbb{E}\|x^r - x^{r-1}\|^2 \leq 4\tilde{\eta}^2\beta^2\mathcal{E}_r + 8\tilde{\eta}^2\beta^2\Xi_{r-1} + 4\tilde{\eta}^2\mathbb{E}\|\nabla f(x^{r-1})\|^2 + \frac{9\tilde{\eta}^2\sigma^2}{KS}\,.$$

Proof.

$$\mathbb{E}[\Delta x] = -\frac{\tilde{\eta}}{KN}\sum_{k,i}\mathbb{E}[g_i(y_{i,k-1})].$$

$$\Delta x = -\frac{\tilde{\eta}}{KS}\sum_{k,i\in\mathcal{S}}(g_i(y_{i,k-1}) + c - c_i) \text{ where } c_i = \frac{1}{K}\sum_k g_i(\alpha_{i,k-1}).$$

Then,

$$\begin{aligned}
\mathbb{E}\|\Delta x\|^2 &= \mathbb{E}\| - \frac{\eta}{KS}\sum_{k,i\in\mathcal{S}}(g_i(\boldsymbol{y}_{i,k-1}) - g_i(\boldsymbol{\alpha}_{i,k-1}) + \boldsymbol{c} - \boldsymbol{c}_i)\|^2 \\
&\leq \mathbb{E}\left\|-\frac{\tilde{\eta}}{KS}\sum_{k,i\in\mathcal{S}}(\nabla f_i(\boldsymbol{y}_{i,k-1}) + \mathbb{E}[\boldsymbol{c}] - \mathbb{E}[\boldsymbol{c}_i])\right\|^2 + \frac{9\tilde{\eta}^2\sigma^2}{KS} \\
&\leq \mathbb{E}\left[\frac{\tilde{\eta}^2}{KS}\sum_{k,i\in\mathcal{S}}\left\|\nabla f_i(\boldsymbol{y}_{i,k-1}) + \mathbb{E}[\boldsymbol{c}] - \mathbb{E}[\boldsymbol{c}_i]\right\|^2\right] + \frac{9\tilde{\eta}^2\sigma^2}{KS} \\
&= \frac{\tilde{\eta}^2}{KN}\sum_{k,i}\mathbb{E}\,\|(\nabla f_i(y_{i,k-1}) - \nabla f_i(x)) + (\mathbb{E}[c] - \nabla f(x)) + \nabla f(x) - (\mathbb{E}[c_i] - \nabla f_i(x))\|^2 + \frac{9\tilde{\eta}^2\sigma^2}{KS} \\
&\leq \frac{4\tilde{\eta}^2}{KN}\sum_{k,i}\mathbb{E}\|\nabla f_i(y_{i,k-1}) - \nabla f_i(x)\|^2 + \frac{8\tilde{\eta}^2}{KN}\sum_{k,i}\mathbb{E}\|\nabla f_i(\alpha_{i,k-1}) - \nabla f_i(x)\|^2 + 4\tilde{\eta}^2\,\mathrm{E}\,\|\nabla f(x)\|^2 + \frac{9\tilde{\eta}^2\sigma^2}{KS} \\
&\leq 4\tilde{\eta}^2\beta^2\mathcal{E}_r + 8\beta^2\tilde{\eta}^2\Xi_{r-1} + 4\tilde{\eta}^2\mathbb{E}\|\nabla f(x)\|^2 + \frac{9\tilde{\eta}^2\sigma^2}{KS}.
\end{aligned}$$

Lemma 2. the following holds true for any $\tilde{\eta} \leq \frac{1}{24\beta}(\frac{S}{N})^{\alpha}, \forall \alpha \in [\frac{1}{2}, 1]$, where $\tilde{\eta} := \eta_l \eta_g \dot{K}$:

$$\Xi_r \leq (1 - \frac{17S}{36N})\Xi_{r-1} + \frac{1}{48\beta^2}(\frac{S}{N})^{2\alpha-1}\|\nabla f(x^{r-1})\|^2 + \frac{97}{48}(\frac{S}{N})^{2\alpha-1}\mathcal{E}_r + (\frac{S}{N\beta^2})\frac{\sigma^2}{32KS}.$$

Proof.

$$\mathbb{E}_{\mathcal{S}^r}[\alpha^r_{i,k-1}] = (1 - \frac{S}{N})\alpha^{r-1}_{i,k-1} + \frac{S}{N}y^r_{i,k-1}.$$

$$\begin{aligned}
\Xi_r &= \frac{1}{KN}\sum_{i,k}\mathbb{E}\|\alpha^r_{i,k-1} - x^r\|^2 \\
&= \left(1 - \frac{S}{N}\right) \cdot \underbrace{\frac{1}{KN}\sum_{i}\mathbb{E}\|\alpha^{r-1}_{i,k-1} - x^r\|^2}_{\mathcal{T}_5} + \underbrace{\frac{S}{N} \cdot \frac{1}{KN}\sum_{k,i}\mathbb{E}\|y^r_{i,k-1} - x^r\|^2}_{\mathcal{T}_6}.
\end{aligned}$$

$$\begin{aligned}
\mathcal{T}_5 &= \frac{1}{KN}\sum_{i}\mathbb{E}(\|\alpha^{r-1}_{i,k-1} - x^{r-1}\|^2 + \|\Delta x^r\|^2 + \mathbb{E}_{r-1}\left\langle \Delta x^r, \alpha^{r-1}_{i,k-1} - x^{r-1}\right\rangle) \\
&\leq \frac{1}{KN}\sum_{i}\mathbb{E}(\|\alpha^{r-1}_{i,k-1} - x^{r-1}\|^2 + \|\Delta x^r\|^2 + \frac{1}{b}\|\mathbb{E}_{r-1}[\Delta x^r]\|^2 + b\|\alpha^{r-1}_{i,k-1} - x^{r-1}\|^2)
\end{aligned}$$

$$\mathcal{T}_6 \leq 2(\mathcal{E}_r + \mathbb{E}\|\Delta x^r\|^2)$$

By lemma 1,

$$\begin{aligned}
\Xi_r &\leq \left(1 - \frac{S}{N}\right)(1+b)\Xi_{r-1} + 2\frac{S}{N}\mathcal{E}_r + 2\mathbb{E}\|\Delta x^r\|^2 + \frac{1}{b}\mathbb{E}\|\mathbb{E}_{r-1}[\Delta x^r]\|^2 \\
&\leq (\left(1 - \frac{S}{N}\right)(1+b) + 16\tilde{\eta}^2\beta^2)\Xi_{r-1} + (\frac{2S}{N} + 8\tilde{\eta}^2\beta^2 + 2\frac{1}{b}\tilde{\eta}^2\beta^2)\mathcal{E}_r + (8 + 2\frac{1}{b})\tilde{\eta}^2\mathbb{E}\|\nabla f(x)\|^2) + \frac{18\tilde{\eta}^2\sigma^2}{KS}
\end{aligned}$$

Let $b = \frac{S}{2(N-S)}$, we have

$$(1 - \frac{S}{N})(1+b) \leq (1 - \frac{S}{2N}), \quad \frac{1}{b} \leq \frac{2N}{S}.$$

Plugging these values along with the bound on the step-size

$$\beta^2\tilde{\eta}^2 \leq \frac{1}{36}(\frac{S}{N})^{2\alpha} \leq \frac{S}{36N}$$

completes the lemma.

Lemma 3. Suppose our step-sizes satisfy $n_l \leq \dfrac{1}{24\beta K\eta_g}$. Then, for any global $\eta_g \geq 1$ we can bound the drift as

$$\frac{5}{3}\beta^2\tilde{\eta}\mathcal{E}_r \leq \frac{5}{3}\beta^3\tilde{\eta}^2\Xi_{r-1} + \frac{\tilde{\eta}}{24\eta_g^2}\mathbb{E}\|\nabla f(x^{r-1})\|^2 + \frac{\tilde{\eta}^2\beta}{4K\eta_g^2}\sigma^2.$$

Proof. For $K \geq 2$,

$$\begin{aligned}
\mathbb{E}\|y_{i,k} - x\|^2 =& \mathbb{E}\|y_{i,k-1} - \eta_l(g_i(y_{i,k-1}) + c - c_i) - x\|^2 \\
\leq& \mathbb{E}\|\boldsymbol{y}_{i,k-1} - \eta_l(\nabla f_i(\boldsymbol{y}_{i,k-1}) + c - \boldsymbol{c}_i) - \boldsymbol{x}\|^2 + \eta_l^2\sigma^2 \\
\leq& (1 + \frac{1}{K-1})\,\mathrm{E}\,\|y_{i,k-1} - x\|^2 + K\eta_l^2\,\mathrm{E}\,\|\nabla f_i(y_{i,k-1}) + c - c_i\|^2 + \eta_l^2\sigma^2 \\
=& (1 + \frac{1}{K-1})\mathbb{E}\|y_{i,k-1} - x\|^2 + \eta_l^2\sigma^2 \\
&+ K\eta_l^2\mathbb{E}\|\nabla f_i(\boldsymbol{y}_{i,k-1}) - \nabla f_i(x) + (c - \nabla f(\boldsymbol{x})) + \nabla f(\boldsymbol{x}) - (c_i - \nabla f_i(\boldsymbol{x})\|^2 \\
\leq& (1 + \frac{1}{K-1})\mathbb{E}\|y_{i,k-1} - x\|^2 + 4K\eta_l^2\mathbb{E}\|\nabla f_i(y_{i,k-1}) - \nabla f_i(\boldsymbol{x})\|^2 + \eta_l^2\sigma^2 \\
&+ 4K\eta_l^2\mathbb{E}\|c - \nabla f(x)\|^2 + 4K\eta_l^2\mathbb{E}\|\nabla f(x)\|^2 + 4K\eta_l^2\mathbb{E}\|c_i - \nabla f_i(\boldsymbol{x})\|^2 \\
\leq& (1 + \frac{1}{K-1} + 4K\beta^2\eta_l^2)\mathbb{E}\|y_{i,k-1} - x\|^2 + \eta_l^2\sigma^2 + 4K\eta_l^2\mathbb{E}\|\nabla f(x)\|^2 \\
&+ 4K\eta_l^2\,\mathrm{E}\,\|c - \nabla f(x)\|^2 + 4K\eta_l^2\,\mathrm{E}\,\|c_i - \nabla f_i(x)\|^2
\end{aligned}$$

$$\begin{aligned}
\frac{1}{N}\sum_i \mathbb{E}\|y_{i,k} - x\|^2 \leq& (1 + \frac{1}{K-1} + 4K\beta^2\eta_l^2)\frac{1}{N}\sum_i \mathbb{E}\|y_{i,k-1} - x\|^2 \\
&+ \eta_l^2\sigma^2 + 4K\eta_l^2\ \mathrm{E}\|\nabla f(x)\|^2 + 8K\eta_l^2\beta^2\Xi_{r-1} \\
\leq& \left(\eta_l^2\sigma^2 + 4K\eta_l^2\mathbb{E}\|\nabla f(x)\|^2 + 8K\eta_l^2\beta^2\Xi_{r-1}\right)\left(\sum_{\tau=0}^{k-1}(1 + \frac{1}{K-1} + 4K\beta^2\eta_l^2)^\tau\right) \\
=& \left(\frac{\tilde{\eta}^2\sigma^2}{K^2\eta_g^2} + \frac{4\tilde{\eta}^2}{K\eta_g^2}\mathbb{E}\|\nabla f(x)\|^2 + \frac{8\tilde{\eta}^2\beta^2}{K\eta_g^2}\Xi_{r-1}\right)\left(\sum_{\tau=0}^{k-1}(1 + \frac{1}{K-1} + \frac{4\beta^2\tilde{\eta}^2}{K\eta_g^2})^\tau\right) \\
\leq& \left(\frac{\tilde{\eta}\sigma^2}{24\beta K^2\eta_g^2} + \frac{1}{144\beta^2 K\eta_g^2}\mathbb{E}\|\nabla f(x)\|^2 + \frac{\tilde{\eta}\beta}{3K\eta_g^2}\Xi_{r-1}\right)3K.
\end{aligned}$$

Lemma 4. If $\tilde{\eta} \leq \frac{1}{24\beta}\left(\frac{S}{N}\right)^{\frac{2}{3}}$,

$$\left(\mathbb{E}[f(x^r)] + 12\beta^3\tilde{\eta}^2\frac{N}{S}\Xi_r\right) \leq \left(\mathbb{E}[f(x^{r-1})] + 12\beta^3\tilde{\eta}^2\frac{N}{S}\Xi_{r-1}\right) + \frac{5\beta\tilde{\eta}^2\sigma^2}{KS}(1+\frac{S}{\eta_g^2}) - \frac{\tilde{\eta}}{14}\mathbb{E}\|\nabla f(x^{r-1})\|^2$$

Proof.

$$\begin{aligned}
\mathbb{E}[f(x+\Delta x)] - f(x) &\leq -\frac{\widetilde{\eta}}{KN}\sum_{k,i}\langle \nabla f(\boldsymbol{x}), \mathbb{E}[\nabla f_i(\boldsymbol{y}_{i,k-1})]\rangle + \frac{\beta}{2}\,\mathrm{E}\,\|\Delta \boldsymbol{x}\|^2 \\
&\leq -\frac{\tilde{\eta}}{KN}\sum_{k,i}\langle \nabla f(x), \mathbb{E}[\nabla f_i(\boldsymbol{y}_{i,k-1})]\rangle + \\
&\quad 2\tilde{\eta}^2\beta^3\mathcal{E}_r + 4\tilde{\eta}^2\beta^3\Xi_{r-1} + 2\beta\tilde{\eta}^2\mathbb{E}\|\nabla f(\boldsymbol{x})\|^2 + \frac{9\beta\tilde{\eta}^2\sigma^2}{2KS} \\
&\leq -\frac{\widetilde{\eta}}{2}\|\nabla f(x)\|^2 + \frac{\widetilde{\eta}}{2}\mathbb{E}\left\|\frac{1}{KN}\sum_{i,k}\nabla f_i(\boldsymbol{y}_{i,k-1}) - \nabla f(\boldsymbol{x})\right\|^2 + \\
&\quad 2\tilde{\eta}^2\beta^3\mathcal{E}_r + 4\tilde{\eta}^2\beta^3\Xi_{r-1} + 2\beta\tilde{\eta}^2\mathbb{E}\|\nabla f(\boldsymbol{x})\|^2 + \frac{9\beta\tilde{\eta}^2\sigma^2}{2KS} \\
&\leq -\frac{\tilde{\eta}}{2}\|\nabla f(x)\|^2 + \frac{\tilde{\eta}}{2KN}\sum_{i,k}\mathbb{E}\left\|\nabla f_i(\boldsymbol{y}_{i,k-1}) - \nabla f_i(\boldsymbol{x})\right\|^2 + \\
&\quad 2\tilde{\eta}^2\beta^3\mathcal{E}_r + 4\tilde{\eta}^2\beta^3\Xi_{r-1} + 2\beta\tilde{\eta}^2\mathbb{E}\|\nabla f(x)\|^2 + \frac{9\beta\tilde{\eta}^2\sigma^2}{2KS} \\
&\leq -(\frac{\tilde{\eta}}{2} - 2\beta\tilde{\eta}^2)\|\nabla f(x)\|^2 + (\frac{\tilde{\eta}}{2} + 2\beta\tilde{\eta}^2)\beta^2\mathcal{E}_r + 4\beta^3\tilde{\eta}^2\Xi_{r-1} + \frac{9\beta\tilde{\eta}^2\sigma^2}{2KS}
\end{aligned}$$

By Lemma 2,

$$\begin{aligned}
12\beta^3\tilde{\eta}^2\frac{N}{S}\Xi_r &\leq 12\beta^3\tilde{\eta}^2\frac{N}{S}\left((1-\frac{17S}{36N})\Xi_{r-1} + \frac{1}{48\beta^2}(\frac{S}{N})^{2\alpha-1}\|\nabla f(x^{r-1})\|^2 + \frac{97}{48}(\frac{S}{N})^{2\alpha-1}\mathcal{E}_r + (\frac{S}{N\beta^2})\frac{\sigma^2}{32KS}\right) \\
&= 12\beta^3\tilde{\eta}^2\frac{N}{S}\Xi_{r-1} - \frac{17}{3}\beta^3\tilde{\eta}^2\Xi_{r-1} + \frac{1}{4}\beta\tilde{\eta}^2(\frac{N}{S})^2 - 2\alpha\|\nabla f(x)\|^2 + \frac{97}{4}\beta^3\tilde{\eta}^2(\frac{N}{S})^{2-2\alpha}\mathcal{E}_r + \frac{3\beta\tilde{\eta}^2\sigma^2}{8KS}
\end{aligned}$$

By Lemma 3,

$$\frac{5}{3}\beta^2\tilde{\eta}\mathcal{E}_r \leq \frac{5}{3}\beta^3\tilde{\eta}^2\Xi_{r-1} + \frac{\tilde{\eta}}{24\eta_g^2}\mathbb{E}\|\nabla f(x^{r-1})\|^2 + \frac{\tilde{\eta}^2\beta}{4K\eta_g^2}\sigma^2.$$

At last,

$$\begin{aligned}
\mathbb{E}[f(x+\Delta x)] + 12\beta^3\tilde{\eta}^2\frac{N}{S}\Xi_r &\leq (\mathbb{E}[f(x)] + 12\beta^3\tilde{\eta}^2\frac{N}{S}\Xi_{r-1}) + (\frac{5}{3} - \frac{17}{3})\beta^3\tilde{\eta}^2\Xi_{r-1} \\
&\quad - (\frac{\tilde{\eta}}{2} - 2\beta\tilde{\eta}^2 - \frac{1}{4}\beta\tilde{\eta}^2(\frac{N}{S})^{2-2\alpha})\|\nabla f(\boldsymbol{x})\|^2 \\
&\quad + (\frac{\tilde{\eta}}{2} - \frac{5\tilde{\eta}}{3} + 2\beta\tilde{\eta}^2 + \frac{97}{4}\beta\tilde{\eta}^2(\frac{N}{S})^{2-2\alpha})\beta^2\mathcal{E}_r + \frac{39\beta\tilde{\eta}^2\sigma^2}{8KS}(1 + \frac{S}{\eta_g^2})
\end{aligned}$$

Let $\alpha = \frac{2}{3}$, then $\beta\tilde{\eta}(\frac{N}{S})^{2-2\alpha} \leq \frac{1}{24}$ proves the lemma.

Thus, by lemma 4,

$$\left(\mathbb{E}[f(x^R)] + 12\beta^3\tilde{\eta}^2\frac{N}{S}\Xi_R\right) \leq \mathbb{E}[f(x^0)] + \sum_{r=1}^{R}\left(\frac{5\beta\tilde{\eta}^2\sigma^2}{KS}(1+\frac{S}{\eta_g^2}) - \frac{\tilde{\eta}}{14}\mathbb{E}\|\nabla f(x^{r-1})\|^2\right)$$

so let $F = f(x^0) - f^*$,

$$0 \leq \frac{F}{R} + \frac{1}{R}\sum_{r=1}^{R}\left(\frac{5\beta\tilde{\eta}^2\sigma^2}{KS}(1+\frac{S}{\eta_g^2}) - \frac{\tilde{\eta}}{14}\mathbb{E}\|\nabla f(x^{r-1})\|^2\right)$$

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(x^{r-1})\|^2 \leq \mathcal{O}\left(\frac{F}{\tilde{\eta}R} + \frac{\beta\tilde{\eta}\sigma^2}{KS}(1+\frac{S}{\eta_g^2})\right)$$

Since $\tilde{\eta} \leq \tilde{\eta}_{\max} = \frac{1}{24\beta}\left(\frac{S}{N}\right)^{\frac{2}{3}}$, we need consider two conditions below:

1) $\frac{F}{\tilde{\eta}_{\max}R} \geq \frac{\beta\tilde{\eta}_{\max}\sigma^2}{KS}(1+\frac{S}{\eta_g^2})$, let $\tilde{\eta} = \tilde{\eta}_{\max}$,

$$\frac{F}{\tilde{\eta}R} + \frac{\beta\tilde{\eta}\sigma^2}{KS}(1+\frac{S}{\eta_g^2}) \leq \frac{F}{\tilde{\eta}_{\max}R} + \sqrt{\frac{F\beta\sigma^2}{RKS}(1+\frac{S}{\eta_g^2})}$$

2) $\frac{F}{\tilde{\eta}_{\max}R} \leq \frac{\beta\tilde{\eta}_{\max}\sigma^2}{KS}(1+\frac{S}{\eta_g^2})$, i.e. $\tilde{\eta}_{\max} \geq \tilde{\eta}_m$, let $\tilde{\eta} = \tilde{\eta}_m$, i.e. $\frac{F}{\tilde{\eta}R} = \frac{\beta\tilde{\eta}\sigma^2}{KS}(1+\frac{S}{\eta_g^2})$,

$$\frac{F}{\tilde{\eta}R} + \frac{\beta\tilde{\eta}\sigma^2}{KS}(1+\frac{S}{\eta_g^2}) = \frac{2\beta\tilde{\eta}_m\sigma^2}{KS}(1+\frac{S}{\eta_g^2}) \leq 2\sqrt{\frac{F\beta\sigma^2}{RKS}(1+\frac{S}{\eta_g^2})}$$

In summary,

$$\mathcal{O}\left(\frac{F}{\tilde{\eta}R} + \frac{\beta\tilde{\eta}\sigma^2}{KS}(1+\frac{S}{\eta_g^2})\right) \leq \mathcal{O}\left(\frac{F}{\tilde{\eta}_{\max}R} + \sqrt{\frac{F\beta\sigma^2}{RKS}(1+\frac{S}{\eta_g^2})}\right) = \mathcal{O}\left(\frac{\beta F}{R}\left(\frac{N}{S}\right)^{\frac{2}{3}} + \sqrt{\frac{\beta F\sigma^2}{RKS}(1+\frac{S}{\eta_g^2})}\right)$$

#

3.3 SAGA

3.3.1 Algorithm

Algorithm 7 SAGA

Input: initial point x^0, $w_i^0 = x^0, \forall i$, stepsize γ

for $k = 0, 1, 2, \cdots$ **do**

Compute $f_i'(w_i^k)$ for all i

Randomly pick $j \in [N]$

$x^{k+1} = x^k - \gamma \left[f_j'(x^k) - f_j'(w_j^k) + \frac{1}{n} \sum_{i=1}^{n} f_i'(w_i^k) \right]$

$w_i^{k+1} = \begin{cases} x^k, & i = j \\ w_i^k, & i \neq j \end{cases}$

end for

§ 4 Appendix

4.1

Thm. Let $f : \mathbb{E} \to (-\infty, \infty]$ be an L-smooth function ($L \geq 0$) over a given convex set D. Then for any $x, y \in D$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|^2$$

Proof.

$$\begin{aligned}
|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| &= \left| \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \right| \\
&\leq \int_0^1 |\langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle| dt \\
&\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \cdot \|y - x\| dt \\
&\leq \int_0^1 tL \|y - x\|^2 dt \\
&= \frac{L}{2} \|y - x\|^2
\end{aligned}$$

Remark. We can also get:

$$\begin{aligned}
f(x^*) &= \min_{y \in \mathbb{R}^n} f(y) \\
&\leq \min_{y \in \mathbb{R}^n} \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 \right\} \\
&= \min_{r \geq 0} \left\{ f(x) - r \|\nabla f(x)\| + \frac{L}{2} r^2 \right\} \\
&= f(x) - \frac{1}{2L} \|\nabla f(x)\|^2
\end{aligned}$$

i.e.

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - f^*)$$