Random Shuffling Beats SGD after Finite Epochs: Supplementary Material

A. Proof of Theorem 1

Proof. Assume $T = nl$ where l is positive integer. Notate x_i^t as the ith iteration for tth epoch. There is $x_0^1 = x_0$, $x_n^t = x_0^{t+1}$, $x_n^l = x_T$. Assume the permutation used in tth epoch is $\sigma_t(\cdot)$. Define error term

$$R^t = \sum_{i=1}^{n} \nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right) - \sum_{i=1}^{n} \nabla f_{\sigma_t(i)}\left(x_0^t\right).$$

For one epoch of RANDOMSHUFFLE, We have the following inequality

$$\begin{aligned}
\left\|x_n^t - x^*\right\|^2 &= \left\|x_0^t - x^*\right\|^2 - 2\gamma\left\langle x_0^t - x^*, \sum_{i=1}^{n} \nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right)\right\rangle + \gamma^2\left\|\sum_{i=1}^{n} \nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right)\right\|^2 \\
&= \left\|x_0^t - x^*\right\|^2 - 2\gamma\left\langle x_0^t - x^*, n\nabla F\left(x_0^t\right)\right\rangle - 2\gamma\left\langle x_0^t - x^*, R^t\right\rangle + \gamma^2\left\|n\nabla F\left(x_0^t\right) + R^t\right\|^2 \\
&\le \left\|x_0^t - x^*\right\|^2 - 2n\gamma\left[\frac{L\mu}{L+\mu}\left\|x_0^t - x^*\right\|^2 + \frac{1}{L+\mu}\left\|\nabla F\left(x_0^t\right)\right\|^2\right] \\
&\quad - 2\gamma\left\langle x_0^t - x^*, R^t\right\rangle + 2\gamma^2 n^2\left\|\nabla F\left(x_0^t\right)\right\|^2 + 2\gamma^2\left\|R^t\right\|^2 \\
&= \left(1 - 2n\gamma\frac{L\mu}{L+\mu}\right)\left\|x_0^t - x^*\right\|^2 - \left(2n\gamma\frac{1}{L+\mu} - 2\gamma^2 n^2\right)\left\|\nabla F\left(x_0^t\right)\right\|^2 \\
&\quad - 2\gamma\left\langle x_0^t - x^*, R^t\right\rangle + 2\gamma^2\left\|R^t\right\|^2,
\end{aligned} \tag{A.1}$$

where the inequality is due to Theorem 2.1.11 in (Nesterov, 2013).

Take the expectation of (A.1) over randomness of permutation $\sigma_t(\cdot)$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|x_n^t - x^*\right\|^2\right] \le \left(1 - 2n\gamma\frac{L\mu}{L+\mu}\right)\left\|x_0^t - x^*\right\|^2 - \left(2n\gamma\frac{1}{L+\mu} - 2n^2\gamma^2\right)\left\|\nabla F\left(x_0^t\right)\right\|^2 \\
- 2\gamma\left\langle x_0^t - x^*, \mathbb{E}\left[R^t\right]\right\rangle + 2\gamma^2\,\mathbb{E}\left[\left\|R^t\right\|^2\right].
\end{aligned} \tag{A.2}$$

What remains to be done is to bound the two terms with R^t dependence. Firstly, we give a bound on the norm of R^t:

$$\begin{aligned}
\left\|R^t\right\| &= \left\|\sum_{i=1}^{n} \nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right) - \sum_{i=1}^{n} \nabla f_{\sigma_t(i)}\left(x_0^t\right)\right\| \\
&\le \sum_{i=1}^{n}\left\|\nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right) - \nabla f_{\sigma_t(i)}\left(x_0^t\right)\right\| \\
&= \sum_{i=1}^{n}\left\|\sum_{j=1}^{i-1}\left(\nabla f_{\sigma_t(i)}\left(x_j^t\right) - \nabla f_{\sigma_t(i)}\left(x_{j-1}^t\right)\right)\right\| \\
&\le \sum_{i=1}^{n}\sum_{j=1}^{i-1}\left\|\nabla f_{\sigma_t(i)}\left(x_j^t\right) - \nabla f_{\sigma_t(i)}\left(x_{j-1}^t\right)\right\| \\
&\le \sum_{i=1}^{n}\sum_{j=1}^{i-1} L\left\|x_j^t - x_{j-1}^t\right\|
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{j=1}^{i-1} L\left\|-\gamma\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right)\right\| \\
&\le \sum_{i=1}^{n}\sum_{j=1}^{i-1} L\gamma G \\
&= \frac{n\left(n-1\right)}{2}\gamma GL,
\end{aligned}
$$

where the first and second inequality is by triangle inequality of vector norm, the third inequality is by definition of L, the fourth inequality is by definition of G. By this result, we have

$$
\mathbb{E}\left[\left\|R^t\right\|^2\right] \le \frac{n^4}{4}\gamma^2 G^2 L^2. \tag{A.3}
$$

For the $\mathbb{E}\left[R^t\right]$ term, we need more careful bound. Since the Hessian is constant for quadratic functions, we use H_i to denote the Hessian matrix of function $f_i(\cdot)$. We begin with the following decomposition:

$$
\begin{aligned}
R^t &= \sum_{i=1}^{n}\left[\nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right) - \nabla f_{\sigma_t(i)}\left(x_0^t\right)\right] \\
&= \sum_{i=1}^{n}\left[H_{\sigma_t(i)}\left(x_{i-1}^t - x_0^t\right)\right] \\
&= \sum_{i=1}^{n}\left\{H_{\sigma_t(i)}\sum_{j=1}^{i-1}\left[-\gamma\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right)\right]\right\} \\
&= \sum_{i=1}^{n}\left\{-\gamma H_{\sigma_t(i)}\sum_{j=1}^{i-1}\left[\nabla f_{\sigma_t(j)}\left(x_0^t\right) + \left(\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right) - \nabla f_{\sigma_t(j)}\left(x_0^t\right)\right)\right]\right\} \\
&= -\gamma\sum_{i=1}^{n}\left[H_{\sigma_t(i)}\sum_{j=1}^{i-1}\nabla f_{\sigma_t(j)}\left(x_0^t\right)\right] - \gamma\sum_{i=1}^{n}\left\{H_{\sigma_t(i)}\sum_{j=1}^{i-1}\left[\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right) - \nabla f_{\sigma_t(j)}\left(x_0^t\right)\right]\right\} \\
&= A^t + B^t.
\end{aligned} \tag{A.4}
$$

Here we define random variables

$$
\begin{aligned}
A^t &= -\gamma\sum_{i=1}^{n}\left[H_{\sigma_t(i)}\sum_{j=1}^{i-1}\nabla f_{\sigma_t(j)}\left(x_0^t\right)\right], \\
B^t &= -\gamma\sum_{i=1}^{n}\left\{H_{\sigma_t(i)}\sum_{j=1}^{i-1}\left[\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right) - \nabla f_{\sigma_t(j)}\left(x_0^t\right)\right]\right\}.
\end{aligned}
$$

There is

$$
\mathbb{E}\left[A^t\right] = -\frac{n\left(n-1\right)}{2}\gamma\,\mathbb{E}_{i\neq j}\left[H_i\nabla f_j\left(x_0^t\right)\right], \tag{A.5}
$$

$$
\begin{aligned}
\left\|B^t\right\| &\le \gamma\sum_{i=1}^{n}H_{\sigma_t(i)}\sum_{j=1}^{i-1}\left\|\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right) - \nabla f_{\sigma_t(j)}\left(x_0^t\right)\right\| \\
&\le \gamma\sum_{i=1}^{n}L\sum_{j=1}^{i-1}\left(j-1\right)\gamma GL \\
&= \gamma^2 L^2 G\sum_{i=1}^{n}\frac{\left(i-1\right)\left(i-2\right)}{2}
\end{aligned}
$$

$$\leq \frac{1}{2}\gamma^2 L^2 G n^3. \tag{A.6}$$

Using (A.4) and (A.5), we can decompose the inner product of $x_0^t - x^*$ and $\mathbb{E}\left[R^t\right]$ into:

$$\begin{aligned}
-2\gamma\left\langle x_0^t - x^*, \mathbb{E}\left[R^t\right]\right\rangle &= -2\gamma\left\langle x_0^t - x^*, \mathbb{E}\left[A^t\right] + \mathbb{E}\left[B^t\right]\right\rangle \\
&= -2\gamma\left\langle x_0^t - x^*, \mathbb{E}\left[A^t\right]\right\rangle - 2\gamma\left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right]\right\rangle \\
&= \gamma^2 n\left(n-1\right)\left\langle x_0^t - x^*, \mathbb{E}_{i\neq j} H_i \nabla f_j\left(x_0^t\right)\right\rangle - 2\gamma\left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right]\right\rangle. \qquad (A.7)
\end{aligned}$$

For the first term in (A.7), there is

$$\begin{aligned}
&\gamma^2 n\left(n-1\right)\left\langle x_0^t - x^*, \mathbb{E}_{i\neq j} H_i \nabla f_j\left(x_0^t\right)\right\rangle \\
&= \gamma^2 n\left(n-1\right)\left\langle x_0^t - x^*, \mathbb{E}_{i\neq j} H_i\left[\nabla f_j\left(x_0^t\right) - \nabla f_j\left(x^*\right)\right]\right\rangle + \gamma^2 n\left(n-1\right)\left\langle x_0^t - x^*, \mathbb{E}_{i\neq j} H_i \nabla f_j\left(x^*\right)\right\rangle \\
&\leq \gamma^2 n^2\left\langle x_0^t - x^*, \mathbb{E}_{i,j} H_i H_j\left(x_0^t - x^*\right)\right\rangle + \gamma^2 n\left(n-1\right)\left[\frac{\lambda_1}{2}\left\|x_0^t - x^*\right\|^2 + \frac{1}{2\lambda_1}\left\|\Delta\right\|^2\right] \\
&\leq \gamma^2 n^2\left\|\nabla F\left(x_0^t\right)\right\|^2 + \frac{1}{4}\gamma\mu\left(n-1\right)\left\|x_0^t - x^*\right\|^2 + \gamma^3\mu^{-1}n^2\left(n-1\right)\left\|\Delta\right\|^2. \qquad (A.8)
\end{aligned}$$

Here we introduce variable $\Delta = \mathbb{E}_{i\neq j}\left[H_i \nabla f_j(x^*)\right]$ for simplicity of notation, with i, j uniformly sampled from all pairs of different indices. The first inequality is by $\left\langle x_0^t - x^*, H_i H_i\left(x_0^t - x^*\right)\right\rangle \geq 0$ and AM–GM inequality, where λ_1 is any positive number. The second inequality comes from noticing that $\mathbb{E}_{i,j} H_i H_j = H^2$ (with i, j uniformly sampled from all pairs of indices), and let $\lambda_1 = \frac{1}{2}\mu\gamma^{-1}n^{-1}$.

For the second term in (A.7), we use the bound

$$-2\gamma\left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right]\right\rangle \leq 2\gamma\left[\frac{\lambda_2}{2}\left\|x_0^t - x^*\right\|^2 + \frac{1}{2\lambda_2}\left\|\mathbb{E}\left[B^t\right]\right\|^2\right]. \tag{A.9}$$

Set $\lambda_2 = \frac{1}{4}\mu\left(n-1\right)$ in (A.9) and using (A.6), there is

$$\begin{aligned}
-2\gamma\left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right]\right\rangle &\leq \frac{1}{4}\gamma\mu\left(n-1\right)\left\|x_0^t - x^*\right\|^2 + 4\gamma\mu^{-1}\left(n-1\right)^{-1}\left\|\mathbb{E}\left[B^t\right]\right\|^2 \\
&\leq \frac{1}{4}\gamma\mu\left(n-1\right)\left\|x_0^t - x^*\right\|^2 + \mu^{-1}\left(n-1\right)^{-1}\gamma^5 L^4 G^2 n^6 \\
&\leq \frac{1}{4}\gamma\mu\left(n-1\right)\left\|x_0^t - x^*\right\|^2 + 2\mu^{-1}\gamma^5 L^4 G^2 n^5. \qquad (A.10)
\end{aligned}$$

Substituting (A.8) and (A.10) back to (A.7), we get

$$\begin{aligned}
-2\gamma\left\langle x_0^t - x^*, \mathbb{E}\left[R^t\right]\right\rangle \leq\; &\gamma^2 n^2\left\|\nabla F\left(x_0^t\right)\right\|^2 + \frac{1}{2}\gamma\mu\left(n-1\right)\left\|x_0^t - x^*\right\|^2 \\
&+ \gamma^3\mu^{-1}n^2\left(n-1\right)\left\|\Delta\right\|^2 + 2\mu^{-1}\gamma^5 L^4 G^2 n^5. \qquad (A.11)
\end{aligned}$$

The next step requires to bound the $\|\Delta\|$ term. Toward this end, we use the following important fact:

$$\begin{aligned}
\left\|\Delta\right\| &= \left\|\mathbb{E}_{i\neq j} H_i \nabla f_j\left(x^*\right)\right\| \\
&= \left\|\frac{1}{n\left(n-1\right)}\sum_{i\neq j} H_i \nabla f_j\left(x^*\right)\right\| \\
&= \left\|\frac{-1}{n\left(n-1\right)}\sum_{i} H_i \nabla f_i\left(x^*\right)\right\| \\
&= \frac{1}{n-1}\left\|\mathbb{E}_i\left[H_i \nabla f_i\left(x^*\right)\right]\right\| \\
&\leq \frac{1}{n-1}LG. \qquad (A.12)
\end{aligned}$$

This fact captures the importance of randomly drawing a permutation instead of using a fixed one. Substituting (A.3) (A.11) back to (A.2) and using (A.12) , we finally get a recursion bound for one epoch:

$$\begin{aligned}
&\mathbb{E}\left\|x_n^t - x^*\right\|^2 \\
\leq &\left(1 - 2n\gamma\frac{L\mu}{L+\mu} + \frac{1}{2}\gamma\mu\left(n-1\right)\right)\left\|x_0^t - x^*\right\|^2 - \left(2n\gamma\frac{1}{L+\mu} - 3\gamma^2 n^2\right)\left\|\nabla F\left(x_0^t\right)\right\|^2 \\
&+ \gamma^3\mu^{-1}n^2\left(n-1\right)\left\|\Delta\right\|^2 + 2\mu^{-1}\gamma^5 L^4 G^2 n^5 + \frac{1}{2}n^4\gamma^4 G^2 L^2 \\
\leq &\left(1 - 2n\gamma\frac{L\mu}{L+\mu} + \frac{1}{2}\gamma\mu\left(n-1\right)\right)\left\|x_0^t - x^*\right\|^2 - \left(2n\gamma\frac{1}{L+\mu} - 3\gamma^2 n^2\right)\left\|\nabla F\left(x_0^t\right)\right\|^2 \\
&+ 2\gamma^3\mu^{-1}nL^2G^2 + 2\mu^{-1}\gamma^5 L^4 G^2 n^5 + \frac{1}{2}n^4\gamma^4 G^2 L^2
\end{aligned} \tag{A.13}$$

Now assume

$$n\gamma\frac{L\mu}{L+\mu} > \frac{1}{2}\gamma\mu\left(n-1\right),$$

and

$$2n\gamma\frac{1}{L+\mu} - 3\gamma^2 n^2 > 0,$$

which we call assumption 1 and assumption 2, (A.13) can be further turned into:

$$\mathbb{E}\left[\left\|x_n^t - x^*\right\|^2\right] \leq \left(1 - n\gamma\frac{L\mu}{L+\mu}\right)\left\|x_0^t - x^*\right\|^2 + \gamma^3 nC_1 + \gamma^5 n^5 C_2 + \gamma^4 n^4 C_3, \tag{A.14}$$

where $C_1 = 2\mu^{-1}L^2G^2$, $C_2 = 2\mu^{-1}L^4G^2$, $C_3 = \frac{1}{2}G^2L^2$. Now assume $n\gamma\frac{L\mu}{L+\mu} < 1$, which we call assumption 3. Expanding (A.14) over all epochs leads to a final bound of RANDOMSHUFFLE:

$$\mathbb{E}\left[\left\|x_T - x^*\right\|^2\right] \leq \left(1 - n\gamma\frac{L\mu}{L+\mu}\right)^{\frac{T}{n}}\left\|x_0 - x^*\right\|^2 + \frac{T}{n}\left(\gamma^3 nC_1 + \gamma^5 n^5 C_2 + \gamma^4 n^4 C_3\right). \tag{A.15}$$

Not substituting $\gamma = \frac{4\log T}{T\mu}$ into (A.15), we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|x_T - x^*\right\|^2\right] &\leq \left(1 - \frac{2n\log T}{T}\right)^{\frac{T}{2n\log T}2\log T}\left\|x_0 - x^*\right\|^2 + \frac{T}{n}\left(\gamma^3 nC_1 + \gamma^5 n^5 C_2 + \gamma^4 n^4 C_3\right) \\
&\leq \frac{1}{T^2}\left\|x_0 - x^*\right\|^2 + \frac{1}{T^2}\left(\log T\right)^3 C_4 + \frac{n^3}{T^3}\left(\log T\right)^4 C_5 + \frac{n^4}{T^4}\left(\log T\right)^5 C_6,
\end{aligned} \tag{A.16}$$

where $C_4 = \frac{64C_1}{\mu^3}$, $C_5 = \frac{256C_3}{\mu^4}$, $C_6 = \frac{1024C_2}{\mu^5}$. The first inequality uses the fact that

$$n\frac{4\log T}{T\mu}\frac{L\mu}{L+\mu} \geq \frac{2n\log T}{T}.$$

The second inequality comes from $(1-x)^{\frac{1}{x}} \leq \frac{1}{e}$ for $0 < x < 1$. Obviously, (A.16) is a result of the form $\mathcal{O}\left(\frac{1}{T^2} + \frac{n^3}{T^3}\right)$. Or in the expanding version with constant dependence, we have

$$\mathbb{E}\left[\left\|x_T - x^*\right\|^2\right] \leq \frac{\left(\log T\right)^2}{T^2}\left(D^2 + 128\frac{L^2G^2}{\mu^4}\right) + \frac{n^3\left(\log T\right)^4}{T^3}128\frac{L^2G^2}{\mu^4} + \frac{n^4\left(\log T\right)^5}{T^4}2048\frac{L^4G^2}{\mu^6}. \tag{A.17}$$

What remains to determine is to satisfy the three assumptions: (1) $n\gamma\frac{L\mu}{L+\mu} > \frac{1}{2}\gamma\mu\left(n-1\right)$, (2) $2n\gamma\frac{1}{L+\mu} - 3\gamma^2n^2 > 0$, and (3) $n\gamma\frac{L\mu}{L+\mu} < 1$. The first is naturally satisfied since $\frac{L}{L+\mu} \geq \frac{1}{2}$ and $n > n-1$. The second assumption is equivalent to

$$\frac{T}{\log T} > 6\left(1 + \frac{L}{\mu}\right)n.$$

Assumption 3 is equivalent to

$$\frac{T}{\log T} > \frac{4L}{L+\mu} n,$$

which is obviously satisfied when

$$\frac{T}{\log T} > 4n.$$

So we only need

$$\frac{T}{\log T} > 6\left(1 + \frac{L}{\mu}\right) n.$$

So whenever $\frac{T}{\log T} > 6\left(1 + \frac{L}{\mu}\right) n$, the three assumptions hold. Therefore the theorem is proved. $\square$

B. Proof of Theorem 2

Proof. The idea is similar to the proof of Theorem 1, with a slightly different analysis on the R^t term capturing the changing Hessian. For any i, we use H_i to denote $H_i\left(x^*\right)$. For any vector v not being zero, define vector value directional function

$$dir\left(v\right) = \frac{v}{\|v\|},$$

with norm being ℓ_2 norm. For the convenience of notation, we define $dir\left(\vec{0}\right) = \vec{0}$, where $\vec{0}$ is the zero vector. For any two points $a, b \in \mathbb{R}^d$, and a matrix function $g\left(\cdot\right) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, define line integral:

$$\int_a^b g\left(x\right) dx := \int_0^{\|b-a\|} g\left(a + t\frac{b-a}{\|b-a\|}\right) dir\left(b-a\right) dt,$$

where the integral on the right hand side is integral of vector valued function over real number interval. This integral represents integrating the matrix values function along the line from a to b. Again, define error term

$$R^t = \sum_{i=1}^n \nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right) - \sum_{i=1}^n \nabla f_{\sigma_t(i)}\left(x_0^t\right).$$

We have the following decomposition for the error term:

$$\begin{aligned}
R^t &= \sum_{i=1}^n \left[\nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right) - \nabla f_{\sigma_t(i)}\left(x_0^t\right)\right] \\
&= \sum_{i=1}^n \left[\int_{x_0^t}^{x_{i-1}^t} H_{\sigma_t(i)}\left(x\right) dx\right] \\
&= \sum_{i=1}^n \left[\int_{x_0^t}^{x_{i-1}^t} H_{\sigma_t(i)} dx\right] + \sum_{i=1}^n \left[\int_{x_0^t}^{x_{i-1}^t} \left(H_{\sigma_t(i)}\left(x\right) - H_{\sigma_t(i)}\right) dx\right] \\
&= \sum_{i=1}^n \left[H_{\sigma_t(i)}\left(x_{i-1}^t - x_0^t\right)\right] + \sum_{i=1}^n \left[\int_{x_0^t}^{x_{i-1}^t} \left(H_{\sigma_t(i)}\left(x\right) - H_{\sigma_t(i)}\right) dx\right] \\
&= \sum_{i=1}^n \left[H_{\sigma_t(i)} \sum_{j=1}^{i-1} \left(-\gamma \nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right)\right)\right] + \sum_{i=1}^n \left[\int_{x_0^t}^{x_{i-1}^t} \left(H_{\sigma_t(i)}\left(x\right) - H_{\sigma_t(i)}\right) dx\right] \\
&= -\gamma \sum_{i=1}^n \left[H_{\sigma_t(i)} \sum_{j=1}^{i-1} \nabla f_{\sigma_t(j)}\left(x_0^t\right)\right] - \gamma \sum_{i=1}^n \left\{H_{\sigma_t(i)} \sum_{j=1}^{i-1} \left[\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right) - \nabla f_{\sigma_t(j)}\left(x_0^t\right)\right]\right\} \\
&\quad + \sum_{i=1}^n \left[\int_{x_0^t}^{x_{i-1}^t} \left(H_{\sigma_t(i)}\left(x\right) - H_{\sigma_t(i)}\right) dx\right]
\end{aligned}$$

$$= A^t + B^t + C^t. \tag{B.1}$$

Here we define random variables

$$A^t = -\gamma \sum_{i=1}^{n} \left[H_{\sigma_t(i)} \sum_{j=1}^{i-1} \nabla f_{\sigma_t(j)} \left(x_0^t\right) \right],$$

$$B^t = -\gamma \sum_{i=1}^{n} \left\{ H_{\sigma_t(i)} \sum_{j=1}^{i-1} \left[\nabla f_{\sigma_t(j)} \left(x_{j-1}^t\right) - \nabla f_{\sigma_t(j)} \left(x_0^t\right) \right] \right\},$$

$$C^t = \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} \left(H_{\sigma_t(i)}(x) - H_{\sigma_t(i)} \right) dx \right].$$

Compared with quadratic case, C^t is the new term capturing the difference introduced by a changing Hessian. There is

$$\mathbb{E}\left[A^t\right] = -\frac{n(n-1)}{2} \gamma \, \mathbb{E}_{i \neq j} \left[H_i \nabla f_j \left(x_0^t\right) \right], \tag{B.2}$$

$$\begin{aligned}
\left\|B^t\right\| &\leq \gamma \sum_{i=1}^{n} H_{\sigma_t(i)} \sum_{j=1}^{i-1} \left(\nabla f_{\sigma_t(j)} \left(x_{j-1}^t\right) - \nabla f_{\sigma_t(j)} \left(x_0^t\right) \right) \\
&\leq \gamma \sum_{i=1}^{n} L \sum_{j=1}^{i-1} (j-1) \gamma G L \\
&= \gamma^2 L^2 G \sum_{i=1}^{n} \frac{(i-1)(i-2)}{2} \\
&\leq \frac{1}{2} \gamma^2 L^2 G n^3.
\end{aligned} \tag{B.3}$$

$$\begin{aligned}
\left\|C^t\right\| &\leq \sum_{i=1}^{n} \left[\int_0^{\left\|x_{i-1}^t - x_0^t\right\|} \left\| H_{\sigma_t(i)} \left(a + t \frac{x_{i-1}^t - x_0^t}{\left\|x_{i-1}^t - x_0^t\right\|} \right) - H_{\sigma_t(i)} \right\| dt \right] \\
&\leq \sum_{i=1}^{n} \left[L_H \max\left\{ \left\|x_{i-1}^t - x^*\right\|, \left\|x_0^t - x^*\right\| \right\} \left\|x_{i-1}^t - x_0^t\right\| \right] \\
&\leq n \left[\left(\left\|x_0^t - x^*\right\| + n\gamma G \right) L_H n \gamma G \right] \\
&= n^2 \gamma L_H G \left\|x_0^t - x^*\right\| + n^3 \gamma^2 L_H G^2.
\end{aligned} \tag{B.4}$$

Using (B.1) (B.2), we can decompose the innerproduct of $x_0^t - x^*$ and $\mathbb{E}\left[R^t\right]$ as following:

$$\begin{aligned}
-2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[R^t\right] \right\rangle &= -2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[A^t\right] + \mathbb{E}\left[B^t\right] + \mathbb{E}\left[C^t\right] \right\rangle \\
&= -2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[A^t\right] \right\rangle - 2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right] \right\rangle - 2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[C^t\right] \right\rangle \\
&= \gamma^2 n(n-1) \left\langle x_0^t - x^*, \mathbb{E}_{i \neq j} H_i \nabla f_j \left(x_0^t\right) \right\rangle - 2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right] \right\rangle - 2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[C^t\right] \right\rangle.
\end{aligned} \tag{B.5}$$

For the first term in the (B.5), we have further bound:

$$\begin{aligned}
&\gamma^2 n(n-1) \left\langle x_0^t - x^*, \mathbb{E}_{i \neq j} H_i \nabla f_j \left(x_0^t\right) \right\rangle \\
&= \gamma^2 n(n-1) \mathbb{E}_{i \neq j} \left\langle H_i \left(x_0^t - x^*\right), \nabla f_j \left(x_0^t\right) - \nabla f_j (x^*) \right\rangle + \gamma^2 n(n-1) \left\langle x_0^t - x^*, \mathbb{E}_{i \neq j} H_i \nabla f_j (x^*) \right\rangle \\
&\leq \gamma^2 n^2 \, \mathbb{E}_{i,j} \left\langle \nabla f_i \left(x_0^t\right) - \nabla f_i (x^*), \nabla f_j \left(x_0^t\right) - \nabla f_j (x^*) \right\rangle + \gamma^2 n(n-1) \left[\frac{\lambda}{2} \left\|x_0^t - x^*\right\|^2 + \frac{1}{2\lambda} \left\|\Delta\right\|^2 \right] \\
&\quad + \gamma^2 n(n-1) \mathbb{E}_{i \neq j} \left\langle H_i \left(x_0^t - x^*\right) - \left(\nabla f_i \left(x_0^t\right) - \nabla f_i (x^*) \right), \nabla f_j \left(x_0^t\right) - \nabla f_j (x^*) \right\rangle
\end{aligned}$$

$$\leq \gamma^2 n^2 \left\|\nabla F\left(x_0^t\right)\right\|^2 + \frac{1}{4}\gamma\mu\left(n-1\right)\left\|x_0^t - x^*\right\|^2 + \gamma^3\mu^{-1}n^2\left(n-1\right)\left\|\Delta\right\|^2 + \gamma^2 n\left(n-1\right)L_H L\left\|x_0^t - x^*\right\|^3. \quad \text{(B.6)}$$

Note that here $H_i\left(x_0^t - x^*\right)$ is the matrix $H_i(x^*)$ times vector $x_0^t - x^*$, not the Hessian at point $x_0^t - x^*$. The last inequality is because of

$$\begin{aligned}
\left\|H_i\left(x_0^t - x^*\right) - \left(\nabla f_i\left(x_0^t\right) - \nabla f_i\left(x^*\right)\right)\right\| &= \left\|H_i\left(x_0^t - x^*\right) - \int_{x^*}^{x_0^t} H_i\left(x\right) dx\right\| \\
&= \left\|\int_{x^*}^{x_0^t}\left(H_i - H_i\left(x\right)\right) dx\right\| \\
&\leq \int_0^{\left\|x_0^t - x^*\right\|}\left\|H_i - H_i\left(x^* + t\frac{x_0^t - x^*}{\left\|x_0^t - x^*\right\|}\right)\right\| dt \\
&\leq L_H\left\|x_0^t - x^*\right\|^2.
\end{aligned}$$

For the second term in (B.5), we use the bound

$$-2\gamma\left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right]\right\rangle \leq \frac{1}{4}\gamma\mu\left(n-1\right)\left\|x_0^t - x^*\right\|^2 + 2\mu^{-1}\gamma^5 L^4 G^2 n^5. \quad \text{(B.7)}$$

For the third term in (B.5), we use the bound

$$\begin{aligned}
-2\gamma\left\langle x_0^t - x^*, \mathbb{E}\left[C^t\right]\right\rangle &\leq 2\gamma\left\|x_0^t - x^*\right\| \cdot \left(n^2\gamma L_H G\left\|x_0^t - x^*\right\| + n^3\gamma^2 L_H G^2\right) \\
&= 2n^2\gamma^2 L_H G\left\|x_0^t - x^*\right\|^2 + \gamma^3 n^3 2\left\|x_0^t - x^*\right\| L_H G^2 \\
&\leq 3n^2\gamma^2 L_H G\left\|x_0^t - x^*\right\|^2 + \gamma^4 n^4 G^3 L_H. \quad \text{(B.8)}
\end{aligned}$$

Substituting (B.6) (B.7) (B.8) back to (B.5), we get

$$\begin{aligned}
-2\gamma\left\langle x_0^t - x^*, \mathbb{E}\left[R^t\right]\right\rangle &\leq \gamma^2 n^2\left\|\nabla F\left(x_0^t\right)\right\|^2 + \frac{1}{2}\gamma\mu n\left(n-1\right)\left\|x_0^t - x^*\right\|^2 + \gamma^3\mu^{-1}n^2\left(n-1\right)\left\|\Delta\right\|^2 \\
&\quad + 2\mu^{-1}\gamma^5 L^4 G^2 n^5 + \gamma^4 n^4 G^3 L_H + \gamma^2 n^2\left(L_H L D + 3L_H G\right)\left\|x_0^t - x^*\right\|^2. \quad \text{(B.9)}
\end{aligned}$$

Substituting (B.9) to (A.2), for one epoch we get recursion bound:

$$\begin{aligned}
&\mathbb{E}\left\|x_n^t - x^*\right\|^2 \\
&\leq \left(1 - 2n\gamma\frac{L\mu}{L+\mu} + \frac{1}{2}\gamma\mu\left(n-1\right) + \gamma^2 n^2\left(L_H L D + 3L_H G\right)\right)\left\|x_0^t - x^*\right\|^2 - \left(2n\gamma\frac{1}{L+\mu} - 3\gamma^2 n^2\right)\left\|\nabla F\left(x_0^t\right)\right\|^2 \\
&\quad + \gamma^3\mu^{-1}n^2\left(n-1\right)\left\|\Delta\right\|^2 + 2\mu^{-1}\gamma^5 L^4 G^2 n^5 + \gamma^4 n^4 G^3 L_H + \frac{1}{2}n^4\gamma^4 G^2 L^2. \quad \text{(B.10)}
\end{aligned}$$

Now assume

$$\frac{3}{2}n\gamma\frac{L\mu}{L+\mu} > \frac{1}{2}\gamma\mu\left(n-1\right) + \gamma^2 n^2\left(L_H L D + 3L_H G\right),$$

and

$$2n\gamma\frac{1}{L+\mu} - 3\gamma^2 n^2 > 0,$$

which we call assumption 1 and assumption 2, (B.10) can be further turned into:

$$\mathbb{E}\left[\left\|x_n^t - x^*\right\|^2\right] \leq \left(1 - \frac{1}{2}n\gamma\frac{L\mu}{L+\mu}\right)\left\|x_0^t - x^*\right\|^2 + \gamma^3 n C_1 + \gamma^4 n^4 C_2 + \gamma^5 n^5 C_3, \quad \text{(B.11)}$$

where $C_1 = 2\mu^{-1}L^2G^2$, $C_2 = G^3 L_H + \frac{1}{2}G^2L^2$, $C_3 = 2\mu^{-1}L^4G^2$. Further assume $n\gamma\frac{L\mu}{L+\mu} < 1$, which we call assumption 3, expanding (B.11) over all the epochs we finally get a bound for RANDOMSHUFFLE:

$$\mathbb{E}\left\|x_T - x^*\right\|^2 \leq \left(1 - \frac{1}{2}n\gamma\frac{L\mu}{L+\mu}\right)^{\frac{T}{n}}\left\|x_0 - x^*\right\|^2 + \frac{T}{n}\left(\gamma^3 n C_1 + \gamma^4 n^4 C_2 + \gamma^5 n^5 C_3\right).$$

Let $\gamma = \frac{8 \log T}{T\mu}$, there is

$$\begin{aligned} \mathbb{E}\left\|x_T - x^*\right\|^2 &\le \left(1 - \frac{2n\log T}{T}\right)^{\frac{T}{2n\log T} 2\log T} \left\|x_0 - x^*\right\|^2 + \frac{T}{n}\left(\gamma^3 n C_1 + \gamma^4 n^4 C_2 + \gamma^5 n^5 C_3\right) \\ &\le \frac{1}{T^2}\left\|x_0 - x^*\right\|^2 + \frac{1}{T^2}\left(\log T\right)^3 C_4 + \frac{n^3}{T^3}\left(\log T\right)^4 C_5 + \frac{n^4}{T^4}\left(\log T\right)^5 C_6, \end{aligned} \tag{B.12}$$

where $C_4 = \frac{512 C_1}{\mu^3}$, $C_5 = \frac{4096 C_2}{\mu^4}$, $C_6 = \frac{8^5 C_2}{\mu^5}$.The second inequality comes from $(1-x)^{\frac{1}{x}} \le \frac{1}{e}$ for $0 < x < 1$. Obviously, this is a result of the form $\mathcal{O}\left(\frac{1}{T^2} + \frac{n^3}{T^3}\right)$.

What remains to determine is to satisfy the three assumptions: (1) $\frac{3}{2} n\gamma \frac{L\mu}{L+\mu} > \frac{1}{2}\gamma\mu\left(n-1\right) + \gamma^2 n^2 \left(L_H L D + 3 L_H G\right)$, (2) $2n\gamma\frac{1}{L+\mu} - 3\gamma^2 n^2 > 0$, and (3) $n\gamma\frac{L\mu}{L+\mu} < 1$. The first is satisfied when

$$n\gamma\frac{L\mu}{L+\mu} > \frac{1}{2}\gamma\mu\left(n-1\right),$$

which is naturally satisfied and

$$\frac{1}{2} n\gamma\frac{L\mu}{L+\mu} > \gamma^2 n^2 \left(L_H L D + 3 L_H G\right),$$

which is equivalent to

$$\frac{T}{\log T} > 16\frac{L+\mu}{L\mu^2}\left(L_H L D + 3 L_H G\right) n,$$

which is obviously satisfied if we assume

$$\frac{T}{\log T} > \frac{32}{\mu^2}\left(L_H L D + 3 L_H G\right) n.$$

The second assumption is equivalent to

$$\frac{T}{\log T} > 12\left(1 + \frac{L}{\mu}\right) n.$$

Assumption 3 is equivalent to

$$\frac{T}{\log T} > \frac{8L}{L+\mu} n,$$

which is satisfied when

$$\frac{T}{\log T} > 8n.$$

Since $12\left(1+\frac{L}{\mu}\right) > 8$, we only need

$$\frac{T}{\log T} > \max\left\{\frac{32}{\mu^2}\left(L_H L D + 3 L_H G\right) n, 12\left(1 + \frac{L}{\mu}\right) n\right\}.$$

So whenever $\frac{T}{\log T} > \max\left\{\frac{32}{\mu^2}\left(L_H L D + 3 L_H G\right), 12\left(1+\frac{L}{\mu}\right)\right\} n$, the three assumptions hold. Therefore the theorem is proved. □

C. Proof of Theorem 3

Proof. We only need to show that when $T = n$ (i.e., one epoch is run for each problem) and n is even, no such step size schedule exists. We note the random permutation of this single epoch as $\sigma(\cdot)$. For n even, consider the following quadratic problem:

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x),$$

where

$$f_i(x) = \begin{cases} \frac{1}{2}(x-b)'A(x-b) & i\ odd, \\ \\ \frac{1}{2}(x+b)'A(x+b) & i\ even, \end{cases}$$

where A is some $d \times d$ positive definite matrix with minimal eigenvalue μ and maximal eigenvalue L, b is a d dimensional vector. We use $(\cdot)'$ to notate the transpose, so as to distinguish from exponential T. The exact value of A and b will be determined later. Obviously, $x^* = 0$ is the minimizer. In this setting, we have:

$$\begin{aligned} x_t &= x_{t-1} - \gamma A(x_{t-1} + (-1)^{\sigma(t)} b) \\ &= (I - \gamma A)x_{t-1} - (-1)^{\sigma(t)}\gamma A b. \end{aligned} \tag{C.1}$$

Expanding (C.1) over iterations leads to:

$$x_T = (I-\gamma A)^T x_0 - \sum_{t=1}^{T} (-1)^{\sigma(t)} \gamma (I - \gamma A)^{T-t} A b. \tag{C.2}$$

Taking expectation of (C.2) over the randomness of σ, there is

$$\mathbb{E}\left[x_T\right] = (I - \gamma A)^T x_0. \tag{C.3}$$

With (C.2) (C.3), we have close-formed expression on the final error:

$$\begin{aligned} \mathbb{E}\left[\|x_T - x^*\|^2\right] &= \|\mathbb{E}\left[x_T\right] - x^*\|^2 + \mathbb{E}\left[\|x_T - \mathbb{E}\left[x_T\right]\|^2\right] \\ &= \left\|(I-\gamma A)^T(x_0 - x^*)\right\|^2 + \mathbb{E}\left[\left\|\sum_{t=1}^{T}(-1)^{\sigma(t)}\gamma(I-\gamma A)^{T-t}Ab\right\|^2\right]. \end{aligned} \tag{C.4}$$

Assume the eigenvalues of A are $\lambda_1, \lambda_2, \cdots, \lambda_d$, there is an orthogonal basis $e_1, \cdots, e_d$ for $\mathbb{R}^d$ such that e_k is eigenvector of A with eigenvalue λ_k. We can write

$$b = \sum_{i=1}^{d} b_i e_i.$$

Since $\langle e_i, e_j \rangle = 0$ for $i \neq j$, we can simplify the last term in (C.4):

$$\begin{aligned} \left\|\sum_{t=1}^{T}(-1)^{\sigma(t)}\gamma(I-\gamma A)^{T-t}Ab\right\|^2 &= \left\|\sum_{t=1}^{T}(-1)^{\sigma(t)}\gamma(I-\gamma A)^{T-t}A(\sum_{i=1}^{d} b_i e_i)\right\|^2 \\ &= \left\|\sum_{i=1}^{d}\left[\sum_{t=1}^{T}(-1)^{\sigma(t)}\gamma(I-\gamma A)^{T-t}A(b_i e_i)\right]\right\|^2 \\ &= \left\|\sum_{i=1}^{d}\left[\sum_{t=1}^{T}(-1)^{\sigma(t)}\gamma(1-\gamma \lambda_i)^{T-t}\lambda_i(b_i e_i)\right]\right\|^2 \\ &= \sum_{i=1}^{d}\left[\sum_{t=1}^{T}(-1)^{\sigma(t)}\gamma(1-\gamma\lambda_i)^{T-t}\lambda_i b_i\right]^2 \\ &= \gamma^2 \sum_{i=1}^{d} b_i^2 \lambda_i^2 \left[\sum_{t=1}^{T}(-1)^{\sigma(t)}(1-\gamma\lambda_i)^{T-t}\right]^2. \end{aligned} \tag{C.5}$$

Substituting (C.5) to (C.4), we have

$$\mathbb{E}\left[\|x_T - x^*\|^2\right] = \left\|(I-\gamma A)^T(x_0 - x^*)\right\|^2 + \gamma^2 \sum_{i=1}^d b_i^2 \lambda_i^2\, \mathbb{E}\left[\left[\sum_{t=1}^T (-1)^{\sigma(t)}(1-\gamma\lambda_i)^{T-t}\right]^2\right] \tag{C.6}$$

Once again, we can write

$$x_0 - x^* = \sum_{i=1}^d a_i e_i.$$

Then (C.6) can simplified as

$$\mathbb{E}\left[\|x_T - x^*\|^2\right] = \sum_{i=1}^d (1-\gamma\lambda_i)^{2T} a_i^2 + \gamma^2 \sum_{i=1}^d b_i^2 \lambda_i^2\, \mathbb{E}\left[\left[\sum_{t=1}^T (-1)^{\sigma(t)}(1-\gamma\lambda_i)^{T-t}\right]^2\right] \tag{C.7}$$

Define random variables $s_t = (-1)^{\sigma(t)}$ for $t = 1, \cdots, T$. Then for any index pair $t \neq u$, over randomness of σ, there is

$$\begin{aligned}\mathbb{E}\left[s_t s_u\right] &= \frac{2\frac{(\frac{T}{2})(\frac{T}{2}-1)}{2}}{\frac{T(T-1)}{2}} - \frac{(\frac{T}{2})(\frac{T}{2})}{\frac{T(T-1)}{2}} \\ &= -\frac{1}{T-1}.\end{aligned}$$

Using this fact, we can simplify the last term in (C.7) as:

$$\begin{aligned}\mathbb{E}\left[\left[\sum_{t=1}^T (-1)^{\sigma(t)}(1-\gamma\lambda_i)^{T-t}\right]^2\right] &= \sum_{t=1}^T (1-\gamma\lambda_i)^{2(T-t)} + \sum_{t\neq u}(1-\gamma\lambda_i)^{2T-t-u}\,\mathbb{E}\left[s_t s_u\right] \\ &= \sum_{t=0}^{T-1}(1-\gamma\lambda_i)^{2t} - \frac{1}{T-1}\sum_{t=0}^{T-1}\sum_{u=0,u\neq t}^{T-1}(1-\gamma\lambda_i)^{t+u} \\ &= \sum_{t=0}^{T-1}(1-\gamma\lambda_i)^{2t} + \frac{1}{T-1}\sum_{t=0}^{T-1}(1-\gamma\lambda_i)^{2t} - \frac{1}{T-1}\left[\sum_{t=0}^{T-1}(1-\gamma\lambda_i)^t\right]^2 \\ &= \frac{T}{T-1}\frac{1-(1-\gamma\lambda_i)^{2T}}{1-(1-\gamma\lambda_i)^2} - \frac{1}{T-1}\left[\frac{1-(1-\gamma\lambda_i)^T}{\gamma\lambda_i}\right]^2.\end{aligned} \tag{C.8}$$

For contradiction, we assume for any T, there is a γ dependent on T such that

$$\mathbb{E}\left[\|x_T - x^*\|^2\right] \leq o(1/T). \tag{C.9}$$

Now we determine the specific requirement of A and b. The only requirement is: A has at least three different positive eigenvalues $\lambda_1 > \lambda_2 > \lambda_3$, and $b_i \neq 0$ for any i. Furthermore, we assume $a_i \neq 0$ for any i. Now for the faster convergence rate (C.9) to hold, from (C.7) we know there must be

$$(1-\gamma\lambda_i)^{2T} = o(\frac{1}{T}), \tag{C.10}$$

$$\gamma^2\,\mathbb{E}\left[\left[\sum_{t=1}^T (-1)^{\sigma(t)}(1-\gamma\lambda_i)^{T-t}\right]^2\right] = o(\frac{1}{T}), \tag{C.11}$$

hold for any i.

However with (C.8), we know:

$$\begin{aligned}
&\gamma^2\,\mathbb{E}\left[\left[\sum_{t=1}^{T}(-1)^{\sigma(t)}(1-\gamma\lambda_i)^{T-t}\right]^2\right]\\
&=\gamma^2\left\{\frac{T}{T-1}\frac{1-(1-\gamma\lambda_i)^{2T}}{1-(1-\gamma\lambda_i)^2}-\frac{1}{T-1}\left[\frac{1-(1-\gamma\lambda_i)^T}{\gamma\lambda_i}\right]^2\right\}\\
&=\gamma^2\left[\frac{T}{T-1}\frac{1}{1-(1-\gamma\lambda_i)^2}-\frac{1}{T-1}\frac{1}{\gamma^2\lambda_i^2}\right]+\gamma^2\left[\frac{T}{T-1}\frac{(1-\gamma\lambda_i)^{2T}}{1-(1-\gamma\lambda_i)^2}-\frac{1}{T-1}\frac{-2(1-\gamma\lambda_i)^T+(1-\gamma\lambda_i)^{2T}}{\gamma^2\lambda_i^2}\right].
\end{aligned}\tag{C.12}$$

So by (C.11), there must be (C.12) is $o(\frac{1}{T})$. We now analyze the terms in (C.12). There must be $|1-\gamma\lambda_1|<1$ for convergence, so $|\gamma|$ is no more than $\frac{2}{\lambda_1}$ which is constant. Since (C.10), there is $(1-\gamma\lambda_i)^T=o(1)$, so

$$\gamma^2\left[-\frac{1}{T-1}\frac{-2(1-\gamma\lambda_i)^T+(1-\gamma\lambda_i)^{2T}}{\gamma^2\lambda_i^2}\right]=o(\frac{1}{T}).$$

Again, since $|1-\gamma\lambda_1|<1$, for $i=2,3$ there is

$$|\frac{\gamma^2}{2\gamma\lambda_i-\gamma^2\lambda_i^2}|\le\frac{\frac{2}{\lambda_1}}{(2-\frac{2\lambda_i}{\lambda_1})\lambda_i}$$

which is constant. Therefore by (C.10),

$$\gamma^2\left[\frac{T}{T-1}\frac{(1-\gamma\lambda_i)^{2T}}{1-(1-\gamma\lambda_i)^2}\right]=o(\frac{1}{T})$$

for $i=2,3$. So for what remains in (C.12),

$$\gamma^2\left[\frac{T}{T-1}\frac{(1-\gamma\lambda_i)^{2T}}{1-(1-\gamma\lambda_i)^2}-\frac{1}{T-1}\frac{-2(1-\gamma\lambda_i)^T+(1-\gamma\lambda_i)^{2T}}{\gamma^2\lambda_i^2}\right]=o(\frac{1}{T})$$

for $i=2,3$. Therefore,

$$\gamma^2\left[\frac{T}{T-1}\frac{1}{1-(1-\gamma\lambda_i)^2}-\frac{1}{T-1}\frac{1}{\gamma^2\lambda_i^2}\right]=o(\frac{1}{T}),$$

so

$$\gamma\frac{T}{T-1}\frac{1}{\lambda_i(2-\gamma\lambda_i)}=\frac{1}{T-1}\frac{1}{\lambda_i^2}+o(\frac{1}{T}),$$

which means

$$\frac{\gamma T}{2-\gamma\lambda_i}=\frac{1}{\lambda_i}+o(1).$$

Since $2-\frac{2}{\lambda_1}\lambda_i\le 2-\gamma\lambda_i\le 2$ for $i=2,3$, there must be

$$\sup\lim_{T\to\infty}\gamma T<C$$

for some $C>0$, so $\gamma\to 0$ as $T\to\infty$. Therefore, $(2-\gamma\lambda_i)\to 2$. So there has to be

$$\lim_{T\to\infty}\gamma T=\frac{2}{\lambda_i}.$$

However, this cannot be true for $\lambda_2\neq\lambda_3$ at the same time, contradiction. As a result, no step size can leads to convergence of $o(\frac{1}{T})$. $\square$

D. Proof of Theorem 4

Proof. The idea is similar to the proof of Theorem 2, with a slightly different analysis on the R^t term adopting the sparsity parameter. For any i, we use H_i to denote $H_i(x^*)$. Again, we have the following decomposition for the error term:

$$\begin{aligned}
R^t &= \sum_{i=1}^{n} \left[\nabla f_{\sigma_t(i)}(x_{i-1}^t) - \nabla f_{\sigma_t(i)}(x_0^t)\right] \\
&= \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} H_{\sigma_t(i)}(x)dx\right] \\
&= \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} H_{\sigma_t(i)}dx\right] + \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} (H_{\sigma_t(i)}(x) - H_{\sigma_t(i)})dx\right] \\
&= \sum_{i=1}^{n} \left[H_{\sigma_t(i)}(x_{i-1}^t - x_0^t)\right] + \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} (H_{\sigma_t(i)}(x) - H_{\sigma_t(i)})dx\right] \\
&= \sum_{i=1}^{n} \left[H_{\sigma_t(i)} \sum_{j=1}^{i-1}(-\gamma\nabla f_{\sigma_t(j)}(x_{j-1}^t))\right] + \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} (H_{\sigma_t(i)}(x) - H_{\sigma_t(i)})dx\right] \\
&= -\gamma\sum_{i=1}^{n} \left[H_{\sigma_t(i)} \sum_{j=1}^{i-1}\nabla f_{\sigma_t(j)}(x_0^t)\right] - \gamma\sum_{i=1}^{n} \left\{H_{\sigma_t(i)} \sum_{j=1}^{i-1}\left[\nabla f_{\sigma_t(j)}(x_{j-1}^t) - \nabla f_{\sigma_t(j)}(x_0^t)\right]\right\} \\
&\quad + \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} (H_{\sigma_t(i)}(x) - H_{\sigma_t(i)})dx\right] \\
&= A^t + B^t + C^t.
\end{aligned} \tag{D.1}$$

Here we define random variables

$$A^t = -\gamma\sum_{i=1}^{n} \left[H_{\sigma_t(i)} \sum_{j=1}^{i-1}\nabla f_{\sigma_t(j)}(x_0^t)\right],$$

$$B^t = -\gamma\sum_{i=1}^{n} \left\{H_{\sigma_t(i)} \sum_{j=1}^{i-1}\left[\nabla f_{\sigma_t(j)}(x_{j-1}^t) - \nabla f_{\sigma_t(j)}(x_0^t)\right]\right\},$$

$$C^t = \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} (H_{\sigma_t(i)}(x) - H_{\sigma_t(i)})dx\right].$$

This time, we have bounds for these three terms adopting sparsity information:

$$\mathbb{E}\left[A^t\right] = -\frac{n(n-1)}{2}\gamma\,\mathbb{E}_{i\neq j}\left[H_{\sigma_t(i)}\nabla f_{\sigma_t(j)}(x_0^t)\right], \tag{D.2}$$

$$\begin{aligned}
\left\|B^t\right\| &\le \gamma\sum_{i=1}^{n} H_{\sigma_t(i)} \sum_{j=1}^{i-1}(\nabla f_{\sigma_t(j)}(x_{j-1}^t) - \nabla f_{\sigma_t(j)}(x_0^t)) \\
&\le \gamma\sum_{i=1}^{n} L\sum_{j=1}^{i-1} \rho n\gamma GL \\
&\le n^3\gamma^2\rho GL^2.
\end{aligned} \tag{D.3}$$

$$\left\|C^t\right\| \le \sum_{i=1}^{n}\sum_{j=1}^{i-1}\left\|\int_{x_{j-1}^t}^{x_j^t} (H_{\sigma_t(i)}(x) - H_{\sigma_t(i)})dx\right\|$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{n} \rho n \left[\max\left\{\left\|x_j^t - x^*\right\| | 0 \leq j \leq i-1\right\} L_H \gamma G\right] \\
&\leq \rho n^2 \left[(\left\|x_0^t - x^*\right\| + n\gamma G) L_H \gamma G\right] \\
&= \rho n^2 \gamma L_H G \left\|x_0^t - x^*\right\| + \rho n^3 \gamma^2 L_H G^2.
\end{aligned} \tag{D.4}
$$

Here the introduction of ρ in (D.3) is because: if $f_{\sigma_t(k)}$ and $f_{\sigma_t(j)}$ depend on disjoint dimensions of variables and $k < j$, then there must be $\nabla f_{\sigma_t(j)}(x_k^t) = \nabla f_{\sigma_t(j)}(x_{k-1}^t)$. The introduction of ρ in (D.4) is similar: if $f_{\sigma_t(i)}$ and $f_{\sigma_t(j)}$ depend on disjoint dimensions of variables and $j < i$, then there must be $\int_{x_{j-1}^t}^{x_j^t} (H_{\sigma_t(i)}(x) - H_{\sigma_t(i)}) dx = 0$.

With (D.1) (D.2), we can decompose the innerproduct of $x_0^t - x^*$ and $\mathbb{E}\left[R^t\right]$ into:

$$
\begin{aligned}
-2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[R^t\right]\right\rangle &= -2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[A^t\right] + \mathbb{E}\left[B^t\right] + \mathbb{E}\left[C^t\right]\right\rangle \\
&= -2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[A^t\right]\right\rangle - 2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right]\right\rangle - 2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[C^t\right]\right\rangle \\
&= \gamma^2 n(n-1) \left\langle x_0^t - x^*, \mathbb{E}_{i \neq j} H_i \nabla f_j(x_0^t)\right\rangle - 2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right]\right\rangle - 2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[C^t\right]\right\rangle.
\end{aligned} \tag{D.5}
$$

For the first term in the (D.5), there is

$$
\begin{aligned}
&\gamma^2 n(n-1) \left\langle x_0^t - x^*, \mathbb{E}_{i \neq j} H_i \nabla f_j(x_0^t)\right\rangle \\
&= \gamma^2 n(n-1)\, \mathbb{E}_{i \neq j} \left\langle H_i(x_0^t - x^*), \nabla f_j(x_0^t) - \nabla f_j(x^*)\right\rangle + \gamma^2 n(n-1) \left\langle x_0^t - x^*, \mathbb{E}_{i \neq j} H_i \nabla f_j(x^*)\right\rangle \\
&\leq \gamma^2 n^2\, \mathbb{E}_{i,j} \left\langle \nabla f_i(x_0^t) - \nabla f_i(x^*), \nabla f_j(x_0^t) - \nabla f_j(x^*)\right\rangle + \gamma^2 n(n-1) \left[\frac{\lambda}{2} \left\|x_0^t - x^*\right\|^2 + \frac{1}{2\lambda} \|\Delta\|^2\right] \\
&\quad + \gamma^2 n(n-1)\, \mathbb{E}_{i \neq j} \left\langle H_i(x_0^t - x^*) - (\nabla f_i(x_0^t) - \nabla f_i(x^*)), \nabla f_j(x_0^t) - \nabla f_j(x^*)\right\rangle \\
&\leq \gamma^2 n^2 \left\|\nabla F(x_0^t)\right\|^2 + \frac{1}{4}\gamma\mu(n-1) \left\|x_0^t - x^*\right\|^2 + \gamma^3 \mu^{-1} n^2 (n-1) \|\Delta\|^2 + \gamma^2 n(n-1) L_H L \left\|x_0^t - x^*\right\|^3.
\end{aligned} \tag{D.6}
$$

Where the last inequality is because of

$$
\begin{aligned}
\left\|H_i(x_0^t - x^*) - (\nabla f_i(x_0^t) - \nabla f_i(x^*))\right\| &= \left\|H_i(x_0^t - x^*) - \int_{x^*}^{x_0^t} H_i(x) dx\right\| \\
&= \left\|\int_{x^*}^{x_0^t} (H_i - H_i(x)) dx\right\| \\
&\leq \int_0^{\left\|x_0^t - x^*\right\|} \left\|H_i - H_i\left(x^* + t\frac{x_0^t - x^*}{\|x_0^t - x^*\|}\right)\right\| dt \\
&\leq L_H \left\|x_0^t - x^*\right\|^2.
\end{aligned}
$$

For the second term in (D.5), we use the bound

$$
-2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right]\right\rangle \leq \frac{1}{4}\gamma\mu(n-1) \left\|x_0^t - x^*\right\|^2 + 2\mu^{-1}\gamma^5 \rho^2 L^4 G^2 n^5. \tag{D.7}
$$

For the third term in (D.5), we use the bound

$$
\begin{aligned}
-2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[C^t\right]\right\rangle &\leq 2\gamma \left\|x_0^t - x^*\right\| \cdot (\rho n^2 \gamma L_H G \left\|x_0^t - x^*\right\| + \rho n^3 \gamma^2 L_H G^2) \\
&= 2n^2 \rho \gamma^2 L_H G \left\|x_0^t - x^*\right\|^2 + \rho \gamma^3 n^3 2 \left\|x_0^t - x^*\right\| L_H G^2 \\
&\leq (2\rho + 1) n^2 \gamma^2 L_H G \left\|x_0^t - x^*\right\|^2 + \rho^2 \gamma^4 n^4 G^3 L_H \\
&\leq 3n^2 \gamma^2 L_H G \left\|x_0^t - x^*\right\|^2 + \rho^2 \gamma^4 n^4 G^3 L_H.
\end{aligned} \tag{D.8}
$$

Substituting (D.6) (D.7) (D.8) back to (D.5), we get

$$
-2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[R^t\right]\right\rangle \leq \gamma^2 n^2 \left\|\nabla F(x_0^t)\right\|^2 + \frac{1}{2}\gamma\mu(n-1) \left\|x_0^t - x^*\right\|^2 + \gamma^3 \mu^{-1} n^2 (n-1) \|\Delta\|^2
$$

$$+ 2\mu^{-1}\gamma^5\rho^2L^4G^2n^5 + \rho^2\gamma^4n^4G^3L_H + \gamma^2n^2(L_HLD + 3L_HG)\left\|x_0^t - x^*\right\|^2. \tag{D.9}$$

Substituting (D.9) to (A.2), we have recursion bound for one epoch:

$$\begin{aligned}
&\mathbb{E}\left\|x_n^t - x^*\right\|^2 \\
&\leq (1 - 2n\gamma\frac{L\mu}{L+\mu} + \frac{1}{2}\gamma\mu(n-1) + \gamma^2n^2(L_HLD + 3L_HG))\left\|x_0^t - x^*\right\|^2 - (2n\gamma\frac{1}{L+\mu} - 3\gamma^2n^2)\left\|\nabla F(x_0^t)\right\|^2 \\
&\quad + \gamma^3\mu^{-1}n^2(n-1)\left\|\Delta\right\|^2 + \rho^2\gamma^4n^4G^3L_H + 2\mu^{-1}\gamma^5\rho^2L^4G^2n^5 + 2\rho^2n^4\gamma^4G^2L^2.
\end{aligned}$$

Here the last inequality is because

$$\begin{aligned}
\left\|R^t\right\| &= \left\|\sum_{i=1}^n \nabla f_{\sigma_t(i)}(x_{i-1}^t) - \sum_{i=1}^n \nabla f_{\sigma_t(i)}(x_0^t)\right\| \\
&\leq \sum_{i=1}^n \left\|\nabla f_{\sigma_t(i)}(x_{i-1}^t) - \nabla f_{\sigma_t(i)}(x_0^t)\right\| \\
&= \sum_{i=1}^n \left\|\sum_{j=1}^{i-1}(\nabla f_{\sigma_t(i)}(x_j^t) - \nabla f_{\sigma_t(i)}(x_{j-1}^t))\right\| \\
&\leq \sum_{i=1}^n\sum_{j=1}^{i-1}\left\|\nabla f_{\sigma_t(i)}(x_j^t) - \nabla f_{\sigma_t(i)}(x_{j-1}^t)\right\| \\
&\leq n^2\rho L\gamma G.
\end{aligned}$$

Finally, we again use the fact

$$\left\|\Delta\right\| \leq \frac{1}{n-1}LG.$$

The remaining process is same as proof of Theorem 2, leading to a bound $\mathcal{O}(\frac{1}{T^2} + \frac{\rho^2n^3}{T^3})$.

□

E. Proof of Theorem 5

Proof. The idea is similar to the proof of Theorem 2. For any vector v not being zero, define vector value directional function

$$dir\,(v) = \frac{v}{\|v\|},$$

with norm being ℓ_2 norm. For the convenience of notation, we define $dir\left(\vec{0}\right) = \vec{0}$, where $\vec{0}$ is the zero vector. For any two points $a, b \in \mathbb{R}^d$, and a matrix function $g\left(\cdot\right) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, define line integral:

$$\int_a^b g\left(x\right)dx := \int_0^{\|b-a\|} g\left(a + t\frac{b-a}{\|b-a\|}\right) dir\left(b-a\right)dt,$$

where the integral on the right hand side is integral of vector valued function over real number interval. This integral represents integrating the matrix values function along the line from a to b. Again, define error term

$$R^t = \sum_{i=1}^n \nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right) - \sum_{i=1}^n \nabla f_{\sigma_t(i)}\left(x_0^t\right).$$

Assume F^* being the minimum of function $F(\cdot)$. For one epoch of RANDOMSHUFFLE, we have

$$F(x_0^{t+1}) - F^* \leq F(x_0^t) - F^* - \gamma\left\langle\nabla F(x_0^t), n\nabla F(x_0^t) + R^t\right\rangle + \frac{L}{2}\gamma^2\left\|n\nabla F(x_0^t) + R^t\right\|^2$$

$$
\begin{aligned}
&\leq (1-2n\mu\gamma)\left[F(x_0^t)-F^*\right]-\gamma\left\langle \nabla F(x_0^t), R^t\right\rangle+\frac{L}{2}\gamma^2\left[2n^2\left\|\nabla F(x_0^t)\right\|^2+2\left\|R^t\right\|^2\right]\\
&\leq (1-2n\mu\gamma+2L^2n^2\gamma^2)\left[F(x_0^t)-F^*\right]-\gamma\left\langle \nabla F(x_0^t), R^t\right\rangle+L\gamma^2\left\|R^t\right\|^2. \qquad \text{(E.1)}
\end{aligned}
$$

Here the second inequality is by the definition of Polyak-Łojasiewicz condition, the last inequality uses the fact

$$
2L[F(x_0^t)-F^*]\geq \left\|\nabla F(x_0^t)\right\|^2.
$$

We have the following decomposition for the error term:

$$
\begin{aligned}
R^t &= \sum_{i=1}^{n}\left[\nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right)-\nabla f_{\sigma_t(i)}\left(x_0^t\right)\right]\\
&= \sum_{i=1}^{n}\left[\int_{x_0^t}^{x_{i-1}^t} H_{\sigma_t(i)}\left(x\right)dx\right]\\
&= \sum_{i=1}^{n}\left[\int_{x_0^t}^{x_{i-1}^t} H_{\sigma_t(i)}(x_0^t)dx\right]+\sum_{i=1}^{n}\left[\int_{x_0^t}^{x_{i-1}^t}\left(H_{\sigma_t(i)}\left(x\right)-H_{\sigma_t(i)}(x_0^t)\right)dx\right]\\
&= \sum_{i=1}^{n}\left[H_{\sigma_t(i)}(x_0^t)\left(x_{i-1}^t-x_0^t\right)\right]+\sum_{i=1}^{n}\left[\int_{x_0^t}^{x_{i-1}^t}\left(H_{\sigma_t(i)}\left(x\right)-H_{\sigma_t(i)}(x_0^t)\right)dx\right]\\
&= \sum_{i=1}^{n}\left[H_{\sigma_t(i)}(x_0^t)\sum_{j=1}^{i-1}\left(-\gamma\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right)\right)\right]+\sum_{i=1}^{n}\left[\int_{x_0^t}^{x_{i-1}^t}\left(H_{\sigma_t(i)}\left(x\right)-H_{\sigma_t(i)}(x_0^t)\right)dx\right]\\
&= -\gamma\sum_{i=1}^{n}\left[H_{\sigma_t(i)}(x_0^t)\sum_{j=1}^{i-1}\nabla f_{\sigma_t(j)}\left(x_0^t\right)\right]-\gamma\sum_{i=1}^{n}\left\{H_{\sigma_t(i)}(x_0^t)\sum_{j=1}^{i-1}\left[\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right)-\nabla f_{\sigma_t(j)}\left(x_0^t\right)\right]\right\}\\
&\quad+\sum_{i=1}^{n}\left[\int_{x_0^t}^{x_{i-1}^t}\left(H_{\sigma_t(i)}\left(x\right)-H_{\sigma_t(i)}(x_0^t)\right)dx\right]\\
&= A^t+B^t+C^t. \qquad \text{(E.2)}
\end{aligned}
$$

Here we define random variables

$$
A^t=-\gamma\sum_{i=1}^{n}\left[H_{\sigma_t(i)}(x_0^t)\sum_{j=1}^{i-1}\nabla f_{\sigma_t(j)}\left(x_0^t\right)\right],
$$

$$
B^t=-\gamma\sum_{i=1}^{n}\left\{H_{\sigma_t(i)}(x_0^t)\sum_{j=1}^{i-1}\left[\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right)-\nabla f_{\sigma_t(j)}\left(x_0^t\right)\right]\right\},
$$

$$
C^t=\sum_{i=1}^{n}\left[\int_{x_0^t}^{x_{i-1}^t}\left(H_{\sigma_t(i)}\left(x\right)-H_{\sigma_t(i)}(x_0^t)\right)dx\right].
$$

There is

$$
\mathbb{E}\left[A^t\right]=-\frac{n\left(n-1\right)}{2}\gamma\,\mathbb{E}_{i\neq j}\left[H_i(x_0^t)\nabla f_j\left(x_0^t\right)\right], \qquad \text{(E.3)}
$$

$$
\begin{aligned}
\left\|B^t\right\| &\leq \gamma\sum_{i=1}^{n}H_{\sigma_t(i)}(x_0^t)\sum_{j=1}^{i-1}\left(\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right)-\nabla f_{\sigma_t(j)}\left(x_0^t\right)\right)\\
&\leq \gamma\sum_{i=1}^{n}L\sum_{j=1}^{i-1}\left(j-1\right)\gamma GL
\end{aligned}
$$

$$
\begin{aligned}
&= \gamma^2 L^2 G \sum_{i=1}^{n} \frac{(i-1)(i-2)}{2} \\
&\leq \frac{1}{2}\gamma^2 L^2 G n^3. \qquad \text{(E.4)}
\end{aligned}
$$

$$
\begin{aligned}
\left\|C^t\right\| &\leq \sum_{i=1}^{n}\left[\int_0^{\left\|x_{i-1}^t - x_0^t\right\|}\left\|H_{\sigma_t(i)}\left(x_0^t + t\frac{x_{i-1}^t - x_0^t}{\left\|x_{i-1}^t - x_0^t\right\|}\right) - H_{\sigma_t(i)}(x_0^t)\right\| dt\right] \\
&\leq \sum_{i=1}^{n}\left[L_H \left\|x_{i-1}^t - x_0^t\right\|^2\right] \\
&\leq n^3\gamma^2 L_H G^2. \qquad \text{(E.5)}
\end{aligned}
$$

Using (E.2) (E.3), we can decompose the innerproduct of $\nabla F(x_0^t)$ and $\mathbb{E}\left[R^t\right]$ as following:

$$
\begin{aligned}
-\gamma\left\langle \nabla F(x_0^t), \mathbb{E}\left[R^t\right]\right\rangle &= -\gamma\left\langle \nabla F(x_0^t), \mathbb{E}\left[A^t\right] + \mathbb{E}\left[B^t\right] + \mathbb{E}\left[C^t\right]\right\rangle \\
&= -\gamma\left\langle \nabla F(x_0^t), \mathbb{E}\left[A^t\right]\right\rangle - \gamma\left\langle \nabla F(x_0^t), \mathbb{E}\left[B^t\right]\right\rangle - \gamma\left\langle \nabla F(x_0^t), \mathbb{E}\left[C^t\right]\right\rangle \\
&= \frac{1}{2}\gamma^2 n(n-1)\left\langle \nabla F(x_0^t), \mathbb{E}_{i\neq j} H_i(x_0^t)\nabla f_j\left(x_0^t\right)\right\rangle - \gamma\left\langle \nabla F(x_0^t), \mathbb{E}\left[B^t\right]\right\rangle - \gamma\left\langle \nabla F(x_0^t), \mathbb{E}\left[C^t\right]\right\rangle. \qquad \text{(E.6)}
\end{aligned}
$$

For the first term in the (E.6), we have further bound:

$$
\begin{aligned}
&\frac{1}{2}\gamma^2 n(n-1)\left\langle \nabla F(x_0^t), \mathbb{E}_{i\neq j} H_i(x_0^t)\nabla f_j\left(x_0^t\right)\right\rangle \\
=& \frac{1}{2}\gamma^2 n^2\left\langle \nabla F(x_0^t), \mathbb{E}_{i,j} H_i(x_0^t)\nabla f_j\left(x_0^t\right)\right\rangle - \frac{1}{2}\gamma^2 n\left\langle \nabla F(x_0^t), \mathbb{E}_{i} H_i(x_0^t)\nabla f_i\left(x_0^t\right)\right\rangle \\
\leq& \frac{1}{2}\gamma^2 n^2 L\left\|\nabla F(x_0^t)\right\|^2 + \frac{1}{8}\gamma n\frac{\mu}{L}\left\|\nabla F(x_0^t)\right\|^2 + \frac{1}{2}\gamma^3 n\mu^{-1}L^3G^2 \\
\leq& (\gamma^2 n^2 L^2 + \frac{1}{4}\gamma n\mu)[F(x_0^t) - F^*] + \frac{1}{2}\gamma^3 n\mu^{-1}L^3G^2. \qquad \text{(E.7)}
\end{aligned}
$$

For the second term in (E.6), we use the bound

$$
\begin{aligned}
-\gamma\left\langle \nabla F(x_0^t), \mathbb{E}\left[B^t\right]\right\rangle &\leq \frac{1}{8}\gamma\frac{\mu}{L}n\left\|\nabla F(x_0^t)\right\|^2 + \frac{1}{2}\mu^{-1}\gamma^5 n^5 L^5 G^2 \\
&\leq \frac{1}{4}\gamma\mu n[F(x_0^t) - F^*] + \frac{1}{2}\mu^{-1}\gamma^5 n^5 L^5 G^2. \qquad \text{(E.8)}
\end{aligned}
$$

For the third term in (E.6), we use the bound

$$
\begin{aligned}
-\gamma\left\langle \nabla F(x_0^t), \mathbb{E}\left[C^t\right]\right\rangle &\leq \gamma\left\|\nabla F(x_0^t)\right\| \cdot \left(n^3\gamma^2 L_H G^2\right) \\
&= \gamma^3 n^3\left\|\nabla F(x_0^t)\right\| L_H G^2 \\
&\leq \frac{1}{2}n^2\gamma^2 L\left\|\nabla F(x_0^t)\right\|^2 + \frac{1}{2L}n^4\gamma^4 L_H^2 G^4 \\
&\leq n^2\gamma^2 L^2[F(x_0^t) - F^*] + \frac{1}{2L}n^4\gamma^4 L_H^2 G^4. \qquad \text{(E.9)}
\end{aligned}
$$

Substituting (E.7) (E.8) (E.9) back to (E.6), we get

$$
-\gamma\left\langle \nabla F(x_0^t), \mathbb{E}\left[R^t\right]\right\rangle \leq (\frac{1}{2}\gamma n\mu + 2n^2\gamma^2 L^2)[F(x_0^t) - F^*] + \frac{1}{2}\gamma^3 n\mu^{-1}L^3G^2 + \frac{1}{2}\mu^{-1}\gamma^5 n^5 L^5 G^2 + \frac{1}{2L}n^4\gamma^4 L_H^2 G^4. \qquad \text{(E.10)}
$$

Substituting (E.10) to (E.1), for one epoch we get recursion bound:

$$
\mathbb{E}[F(x_0^{t+1}) - F^*]
$$

$$\leq (1 - \frac{3}{2}n\mu\gamma + 4L^2n^2\gamma^2)\left[F(x_0^t) - F^*\right] + \frac{1}{2}\gamma^3 n\mu^{-1}L^3G^2 + \frac{1}{2}\mu^{-1}\gamma^5 n^5 L^5 G^2 + \frac{1}{2L}n^4\gamma^4 L_H^2 G^4 + \frac{1}{4}n^4\gamma^4 G^2 L^3. \tag{E.11}$$

Now assume

$$\frac{1}{2}n\mu\gamma > 4L^2n^2\gamma^2,$$

which we call assumption 1, (E.11) can be further turned into:

$$\begin{aligned}&\mathbb{E}[F(x_0^{t+1}) - F^*]\\ &\leq (1 - n\mu\gamma)\left[F(x_0^t) - F^*\right] + \gamma^3 nC_1 + n^4\gamma^4 C_2 + n^5\gamma^5 C_3.\end{aligned} \tag{E.12}$$

where $C_1 = \frac{1}{2}\mu^{-1}L^3G^2$, $C_2 = \frac{1}{2L}L_H^2G^4 + \frac{1}{4}G^2L^3$, $C_3 = \frac{1}{2}\mu^{-1}L^5G^2$. Further assume $n\gamma\mu < 1$, which we call assumption 2, expanding (E.12) over all the epochs we finally get a bound for RANDOMSHUFFLE:

$$\mathbb{E}[F(x_T) - F^*] \leq (1 - n\gamma\mu)^{\frac{T}{n}}\left[F(x_0) - F^*\right] + \frac{T}{n}\left(\gamma^3 nC_1 + \gamma^4 n^4 C_2 + \gamma^5 n^5 C_3\right).$$

Let $\gamma = \frac{2\log T}{T\mu}$, there is

$$\begin{aligned}\mathbb{E}[F(x_T) - F^*] &\leq \left(1 - \frac{2n\log T}{T}\right)^{\frac{T}{2n\log T}2\log T}\left[F(x_0) - F^*\right] + \frac{T}{n}\left(\gamma^3 nC_1 + \gamma^4 n^4 C_2 + \gamma^5 n^5 C_3\right)\\ &\leq \frac{1}{T^2}[F(x_0) - F^*] + \frac{1}{T^2}\left(\log T\right)^3 C_4 + \frac{n^3}{T^3}\left(\log T\right)^4 C_5 + \frac{n^4}{T^4}\left(\log T\right)^5 C_6,\end{aligned} \tag{E.13}$$

where $C_4 = \frac{8C_1}{\mu^3}$, $C_5 = \frac{16C_2}{\mu^4}$, $C_6 = \frac{32C_2}{\mu^5}$.The second inequality comes from $(1-x)^{\frac{1}{x}} \leq \frac{1}{e}$ for $0 < x < 1$. Obviously, this is a result of the form $\mathcal{O}\left(\frac{1}{T^2} + \frac{n^3}{T^3}\right)$.

What remains to determine is to satisfy the two assumptions: (1) $\frac{1}{2}n\mu\gamma > 4L^2n^2\gamma^2$, (2) $n\gamma\mu < 1$. The first is satisfied when

$$\frac{T}{\log T} > 16\frac{L^2}{\mu^2}n.$$

The second assumption is satisfied when

$$\frac{T}{\log T} > 2n.$$

Since $2 < \frac{L}{\mu}$, the theorem is proved. □

F. Proof of Theorem 6

Proof. For one epoch of RANDOMSHUFFLE, We have the following inequality

$$\begin{aligned}\left\|x_n^t - x^*\right\|^2 &= \left\|x_0^t - x^*\right\|^2 - 2\gamma\left\langle x_0^t - x^*, \sum_{i=1}^n \nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right)\right\rangle + \gamma^2\left\|\sum_{i=1}^n \nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right)\right\|^2\\ &= \left\|x_0^t - x^*\right\|^2 - 2\gamma\left\langle x_0^t - x^*, n\nabla F\left(x_0^t\right)\right\rangle - 2\gamma\left\langle x_0^t - x^*, R^t\right\rangle + \gamma^2\left\|n\nabla F\left(x_0^t\right) + R^t\right\|^2\\ &\leq \left\|x_0^t - x^*\right\|^2 - 2n\gamma\left[F(x_0^t) - F(x^*)\right] - 2\gamma\left\langle x_0^t - x^*, R^t\right\rangle + 2\gamma^2n^2\left\|\nabla F\left(x_0^t\right)\right\|^2 + 2\gamma^2\left\|R^t\right\|^2\\ &\leq \left\|x_0^t - x^*\right\|^2 - \left(2n\gamma - 2n^2\gamma^2L\right)\left[F(x_0^t) - F(x^*)\right] - 2\gamma\left\langle x_0^t - x^*, R^t\right\rangle + 2\gamma^2\left\|R^t\right\|^2,\end{aligned} \tag{F.1}$$

where the first inequality is because of

$$\left\langle x_0^t - x^*, \nabla F(x_0^t)\right\rangle \geq F(x_0^t) - F(x^*),$$

the second inequality is because of

$$\left\|\nabla F(x_0^t)\right\|^2 \leq L\left[F(x) - F(x^*)\right].$$

Therefore, taking expectation of (F.1) leads to:

$$\mathbb{E}[\left\|x_n^t - x^*\right\|^2] \le \left\|x_0^t - x^*\right\|^2 - (2n\gamma - 2n^2\gamma^2 L)\left[F(x_0^t) - F(x^*)\right] - 2\gamma\,\mathbb{E}\left\langle x_0^t - x^*, R^t\right\rangle + 2\gamma^2\,\mathbb{E}\left[\left\|R^t\right\|^2\right], \quad \text{(F.2)}$$

Define random variables

$$R_k^t = \sum_{i=1}^{k}\left[\nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right) - \nabla f_{\sigma_t(i)}\left(x_0^t\right)\right],$$

where $1 \le k \le n$. Obviously $R_n^t = R^t$. We firstly show that $\|R_k^t\| \le 3n^2 L\gamma(\|\nabla F(x_0^t)\| + \delta)$, which is an important fact to be used in further analysis.

For any index $1 \le id \le n$, there is

$$\left\|\nabla f_{id}(x_1^t) - \nabla f_{id}(x_0^t)\right\| \le L\gamma(\left\|\nabla F(x_0^t)\right\| + \delta).$$

Assume for any $1 \le id \le n$ and some i, there is (which is obviously true when $i = 1$)

$$\left\|\nabla f_{id}(x_i^t) - \nabla f_{id}(x_0^t)\right\| \le \left[\sum_{j=0}^{i-1}(1+L\gamma)^j\right] L\gamma(\left\|\nabla F(x_0^t)\right\| + \delta).$$

Then for $i+1$, there is

$$\begin{aligned}
\left\|\nabla f_{id}(x_{i+1}^t) - \nabla f_{id}(x_0^t)\right\| &\le \left\|\nabla f_{id}(x_i^t) - \nabla f_{id}(x_0^t)\right\| + \left\|\nabla f_{id}(x_{i+1}^t) - \nabla f_{id}(x_i^t)\right\| \\
&\le \left\|\nabla f_{id}(x_i^t) - \nabla f_{id}(x_0^t)\right\| + L\gamma(\left\|\nabla F(x_i^t)\right\| + \delta) \\
&\le \left\|\nabla f_{id}(x_i^t) - \nabla f_{id}(x_0^t)\right\| + L\gamma(\left\|\nabla F(x_0^t)\right\| + \delta) + L\gamma(\left\|\nabla F(x_i^t) - \nabla F(x_0^t)\right\|) \\
&\le (1+L\gamma)\left[\sum_{j=0}^{i-1}(1+L\gamma)^j\right] L\gamma(\left\|\nabla F(x_0^t)\right\| + \delta) + L\gamma(\left\|\nabla F(x_0^t)\right\| + \delta) \\
&= \left[\sum_{j=0}^{i}(1+L\gamma)^j\right] L\gamma(\left\|\nabla F(x_0^t)\right\| + \delta).
\end{aligned}$$

So by induction, there is

$$\left\|\nabla f_{id}(x_i^t) - \nabla f_{id}(x_0^t)\right\| \le \left[\sum_{j=0}^{i-1}(1+L\gamma)^j\right] L\gamma(\left\|\nabla F(x_0^t)\right\| + \delta)$$

for all $1 \le i \le n$. Since $\gamma \le \frac{1}{16nL} \le \frac{1}{nL}$, there is $1 + \gamma L \le \frac{1}{n}$. Therefore, we have

$$\begin{aligned}
\left\|\nabla f_{id}(x_i^t) - \nabla f_{id}(x_0^t)\right\| &\le \left[\sum_{j=0}^{i-1}(1+L\gamma)^j\right] L\gamma(\left\|\nabla F(x_0^t)\right\| + \delta) \\
&\le \left[n(1+\frac{1}{n})^n\right] L\gamma(\left\|\nabla F(x_0^t)\right\| + \delta) \\
&\le 3nL\gamma(\left\|\nabla F(x_0^t)\right\| + \delta).
\end{aligned}$$

Therefore, for any $1 \le k \le n$, there is

$$\begin{aligned}
\left\|R_k^t\right\| &\le \sum_{i=1}^{k}\left\|\nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right) - \nabla f_{\sigma_t(i)}\left(x_0^t\right)\right\| \\
&\le \sum_{i=1}^{k} 3nL\gamma(\left\|\nabla F(x_0^t)\right\| + \delta)
\end{aligned}$$

$$\leq 3n^2 L\gamma(\left\|\nabla F(x_0^t)\right\| + \delta).$$

Similar to the previous proof, we have the following decomposition for the error term:

$$\begin{aligned}
R^t &= \sum_{i=1}^{n} \left[\nabla f_{\sigma_t(i)}\left(x_{i-1}^t\right) - \nabla f_{\sigma_t(i)}\left(x_0^t\right)\right] \\
&= \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} H_{\sigma_t(i)}\left(x\right) dx\right] \\
&= \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} H_{\sigma_t(i)} dx\right] + \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} \left(H_{\sigma_t(i)}\left(x\right) - H_{\sigma_t(i)}\right) dx\right] \\
&= \sum_{i=1}^{n} \left[H_{\sigma_t(i)}\left(x_{i-1}^t - x_0^t\right)\right] + \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} \left(H_{\sigma_t(i)}\left(x\right) - H_{\sigma_t(i)}\right) dx\right] \\
&= \sum_{i=1}^{n} \left[H_{\sigma_t(i)} \sum_{j=1}^{i-1} \left(-\gamma \nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right)\right)\right] + \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} \left(H_{\sigma_t(i)}\left(x\right) - H_{\sigma_t(i)}\right) dx\right] \\
&= -\gamma \sum_{i=1}^{n} \left[H_{\sigma_t(i)} \sum_{j=1}^{i-1} \nabla f_{\sigma_t(j)}\left(x_0^t\right)\right] - \gamma \sum_{i=1}^{n} \left\{H_{\sigma_t(i)} \sum_{j=1}^{i-1} \left[\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right) - \nabla f_{\sigma_t(j)}\left(x_0^t\right)\right]\right\} \\
&\quad + \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} \left(H_{\sigma_t(i)}\left(x\right) - H_{\sigma_t(i)}\right) dx\right] \\
&= A^t + B^t + C^t.
\end{aligned} \tag{F.3}$$

Here we define random variables

$$A^t = -\gamma \sum_{i=1}^{n} \left[H_{\sigma_t(i)} \sum_{j=1}^{i-1} \nabla f_{\sigma_t(j)}\left(x_0^t\right)\right],$$

$$B^t = -\gamma \sum_{i=1}^{n} \left\{H_{\sigma_t(i)} \sum_{j=1}^{i-1} \left[\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right) - \nabla f_{\sigma_t(j)}\left(x_0^t\right)\right]\right\},$$

$$C^t = \sum_{i=1}^{n} \left[\int_{x_0^t}^{x_{i-1}^t} \left(H_{\sigma_t(i)}\left(x\right) - H_{\sigma_t(i)}\right) dx\right].$$

There is

$$\mathbb{E}\left[A^t\right] = -\frac{n\left(n-1\right)}{2} \gamma\, \mathbb{E}_{i \neq j}\left[H_i \nabla f_j\left(x_0^t\right)\right], \tag{F.4}$$

$$\begin{aligned}
\left\|B^t\right\| &\leq \gamma \sum_{i=1}^{n} H_{\sigma_t(i)} \sum_{j=1}^{i-1} \left\|\nabla f_{\sigma_t(j)}\left(x_{j-1}^t\right) - \nabla f_{\sigma_t(j)}\left(x_0^t\right)\right\| \\
&\leq \gamma \sum_{i=1}^{n} L \sum_{j=1}^{i-1} 3nL\gamma(\left\|\nabla F(x_0^t)\right\| + \delta) \\
&\leq 3\gamma^2 L^2 n^3 (\left\|\nabla F(x_0^t)\right\| + \delta).
\end{aligned} \tag{F.5}$$

$$\left\|C^t\right\| \leq \sum_{i=1}^{n} \left[\int_0^{\left\|x_{i-1}^t - x_0^t\right\|} \left\|H_{\sigma_t(i)}\left(x_0^t + t\frac{x_{i-1}^t - x_0^t}{\left\|x_{i-1}^t - x_0^t\right\|}\right) - H_{\sigma_t(i)}\right\| dt\right]$$

$$\leq \sum_{i=1}^{n} \left[L_H \max\left\{ \left\| x_{i-1}^t - x^* \right\|, \left\| x_0^t - x^* \right\| \right\} \left\| x_{i-1}^t - x_0^t \right\| \right]$$
$$\leq n L_H D n \gamma G. \tag{F.6}$$

Using (F.3) and (F.4), we can decompose the inner product of $x_0^t - x^*$ and $\mathbb{E}\left[R^t\right]$ into:

$$\begin{aligned} -2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[R^t\right] \right\rangle &= -2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[A^t\right] + \mathbb{E}\left[B^t\right] + \mathbb{E}\left[C^t\right] \right\rangle \\ &= -2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[A^t\right] \right\rangle - 2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right] \right\rangle - 2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[C^t\right] \right\rangle \\ &= \gamma^2 n (n-1) \left\langle x_0^t - x^*, \mathbb{E}_{i \neq j} H_i \nabla f_j \left(x_0^t\right) \right\rangle - 2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right] \right\rangle - 2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[C^t\right] \right\rangle. \end{aligned} \tag{F.7}$$

For the first term in (F.7), there is

$$\begin{aligned} &\gamma^2 n (n-1) \left\langle x_0^t - x^*, \mathbb{E}_{i \neq j} H_i \nabla f_j \left(x_0^t\right) \right\rangle \\ &= \gamma^2 n (n-1) \mathbb{E}_{i \neq j} \left\langle H_i \left(x_0^t - x^*\right), \nabla f_j \left(x_0^t\right) - \nabla f_j \left(x^*\right) \right\rangle + \gamma^2 n (n-1) \left\langle x_0^t - x^*, \mathbb{E}_{i \neq j} H_i \nabla f_j \left(x^*\right) \right\rangle \\ &\leq \gamma^2 n^2 \mathbb{E}_{i,j} \left\langle \nabla f_i \left(x_0^t\right) - \nabla f_i \left(x^*\right), \nabla f_j \left(x_0^t\right) - \nabla f_j \left(x^*\right) \right\rangle + \gamma^2 n (n-1) D \left\| \Delta \right\| \\ &\quad + \gamma^2 n (n-1) \mathbb{E}_{i \neq j} \left\langle H_i \left(x_0^t - x^*\right) - \left(\nabla f_i \left(x_0^t\right) - \nabla f_i \left(x^*\right)\right), \nabla f_j \left(x_0^t\right) - \nabla f_j \left(x^*\right) \right\rangle \\ &\leq \gamma^2 n^2 \left\| \nabla F \left(x_0^t\right) \right\|^2 + \gamma^2 n (n-1) D \left\| \Delta \right\| + \gamma^2 n (n-1) L_H L \left\| x_0^t - x^* \right\|^3. \end{aligned} \tag{F.8}$$

Here we introduce variable $\Delta = \mathbb{E}_{i \neq j}\left[H_i \nabla f_j(x^*)\right]$ for simplicity of notation, with i, j uniformly sampled from all pairs of different indices. The last inequality is because of

$$\begin{aligned} \left\| H_i \left(x_0^t - x^*\right) - \left(\nabla f_i \left(x_0^t\right) - \nabla f_i \left(x^*\right)\right) \right\| &= \left\| H_i \left(x_0^t - x^*\right) - \int_{x^*}^{x_0^t} H_i(x)\, dx \right\| \\ &= \left\| \int_{x^*}^{x_0^t} \left(H_i - H_i(x)\right) dx \right\| \\ &\leq \int_0^{\left\| x_0^t - x^* \right\|} \left\| H_i - H_i \left(x^* + t \frac{x_0^t - x^*}{\left\| x_0^t - x^* \right\|} \right) \right\| dt \\ &\leq L_H \left\| x_0^t - x^* \right\|^2. \end{aligned}$$

For the second term in (F.7), we use (F.5) and have the bound

$$-2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[B^t\right] \right\rangle \leq 6\gamma^3 n^3 L^2 D (\left\| \nabla F(x_0^t) \right\| + \delta). \tag{F.9}$$

For the third term in (F.7), we use (F.6) and have the bound

$$-2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[C^t\right] \right\rangle \leq 2\gamma^2 n^2 L_H D^2 G. \tag{F.10}$$

Substituting (F.8) (F.9) and (F.10) back to (F.7), we get

$$\begin{aligned} -2\gamma \left\langle x_0^t - x^*, \mathbb{E}\left[R^t\right] \right\rangle \leq\ & \gamma^2 n^2 \left\| \nabla F \left(x_0^t\right) \right\|^2 + \gamma^2 n (n-1) D \left\| \Delta \right\| + 6\gamma^3 n^3 L^2 D (\left\| \nabla F(x_0^t) \right\| + \delta) \\ & + \gamma^2 n^2 L_H (L D^3 + 2 D^2 G). \end{aligned} \tag{F.11}$$

Furthermore, we have

$$\begin{aligned} \mathbb{E}\left[\left\| R^t \right\|^2 \right] &\leq \left[3n^2 L \gamma (\left\| \nabla F(x_0^t) \right\| + \delta) \right]^2 \\ &\leq 18 n^4 L^2 \gamma^2 (\left\| \nabla F(x_0^t) \right\|^2 + \delta^2). \end{aligned}$$

Inequality (F.2) can be simplified to:

$$\begin{aligned}
\mathbb{E}[\left\|x_n^t - x^*\right\|^2] &\leq \left\|x_0^t - x^*\right\|^2 - (2n\gamma - 3n^2\gamma^2 L)\left[F(x_0^t) - F(x^*)\right] + \gamma^2 n\,(n-1)\,D\left\|\Delta\right\| + \gamma^2 n^2 L_H (LD^3 + 2D^2 G) \\
&\quad + 6\gamma^3 n^3 L^2 D(\left\|\nabla F(x_0^t)\right\| + \delta) + 36 n^4 L^2 \gamma^4 (\left\|\nabla F(x_0^t)\right\|^2 + \delta^2). \\
&\leq \left\|x_0^t - x^*\right\|^2 - (2n\gamma - 3n^2\gamma^2 L)\left[F(x_0^t) - F(x^*)\right] + \gamma^2 n\,(n-1)\,D\left\|\Delta\right\| + \gamma^2 n^2 L_H (LD^3 + 2D^2 G) \\
&\quad + 12\gamma^2 n^2 \left\|\nabla F(x_0^t)\right\|^2 + 12\gamma^4 n^4 L^4 D^2 + 6\gamma^3 n^3 L^2 D\delta + 36 n^4 L^2 \gamma^4 (\left\|\nabla F(x_0^t)\right\|^2 + \delta^2).
\end{aligned} \tag{F.12}$$

By the definition of γ, there is

$$36 n^4 L^2 \gamma^4 \leq n^2 \gamma^2,$$
$$16 n^2 \gamma^2 L \leq n\gamma.$$

So there is

$$\begin{aligned}
\mathbb{E}[\left\|x_n^t - x^*\right\|^2] &\leq \left\|x_0^t - x^*\right\|^2 - (2n\gamma - 16n^2\gamma^2 L)\left[F(x_0^t) - F(x^*)\right] + \gamma^2 n\,(n-1)\,D\left\|\Delta\right\| \\
&\quad + \gamma^2 n^2 L_H (LD^3 + 2D^2 G) + 12\gamma^4 n^4 L^4 D^2 + 6\gamma^3 n^3 L^2 D\delta + 36 n^4 L^2 \gamma^4 \delta^2. \\
&\leq \left\|x_0^t - x^*\right\|^2 - n\gamma\left[F(x_0^t) - F(x^*)\right] + \gamma^2 n^2 D\left\|\Delta\right\| \\
&\quad + \gamma^2 n^2 L_H (LD^3 + 2D^2 G) + 12\gamma^4 n^4 L^4 D^2 + 6\gamma^3 n^3 L^2 D\delta + 36 n^4 L^2 \gamma^4 \delta^2.
\end{aligned}$$

Furthermore, there is

$$\begin{aligned}
n\gamma\left[F(x_0^t) - F(x^*)\right] &\leq \left\|x_0^t - x^*\right\|^2 - \mathbb{E}[\left\|x_n^t - x^*\right\|^2] + \gamma^2 n^2\left(D\left\|\Delta\right\| + L_H L D^3 + 2 L_H D^2 G\right) \\
&\quad + 6\gamma^3 n^3 L^2 D\delta + n^4\gamma^4(12 L^4 D^2 + 36 L^2 \delta^2).
\end{aligned} \tag{F.13}$$

Taking expectation of (F.13) and summing over all epochs, we have:

$$T\gamma\left[F(\bar{x}) - F(x^*)\right] \leq D^2 + \gamma^2 T n (D\left\|\Delta\right\| + L_H L D^3 + 2L_H D^2 G) + T\gamma^3 n^2 L^2 6D\delta + T\gamma^4 n^3 (12 L^4 D^2 + 36 L^2 \delta^2). \tag{F.14}$$

Substituting the step size into (F.14), we have

$$\begin{aligned}
F(\bar{x}) - F(x^*) &\leq \frac{D^2}{T}\max\left\{16nL, \sqrt{\frac{Tn\left(\left\|\Delta\right\| + L_H L D^2 + 2L_H D G\right)}{D}}, \left(\frac{Tn^2L^2\delta}{D}\right)^{\frac{1}{3}}, (Tn^3L^4)^{\frac{1}{4}}\right\} \\
&\quad + \frac{D\sqrt{nD\left(\left\|\Delta\right\| + L_H L D^2 + 2L_H D G\right)}}{\sqrt{T}} + \frac{6D(D^2n^2L^2\delta)^{\frac{1}{3}}}{T^{\frac{2}{3}}} + \frac{n^{\frac{3}{4}}}{T^{\frac{3}{4}}}\left(12LD^2 + \frac{36\delta^2}{L}\right) \\
&\leq \frac{2D\sqrt{nD\left(\left\|\Delta\right\| + L_H L D^2 + 2L_H D G\right)}}{\sqrt{T}} + \frac{7D(D^2n^2L^2\delta)^{\frac{1}{3}}}{T^{\frac{2}{3}}} + \frac{n^{\frac{3}{4}}}{T^{\frac{3}{4}}}\left(13LD^2 + \frac{36\delta^2}{L}\right) + \frac{16D^2 nL}{T}.
\end{aligned}$$

Obviously, this result is of the form

$$\frac{2D\sqrt{nD\left(\left\|\Delta\right\| + L_H L D^2 + 2L_H D G\right)}}{\sqrt{T}} + \mathcal{O}\left(\left(\frac{n}{T}\right)^{\frac{2}{3}}\delta^{\frac{1}{3}} + \left(\frac{n}{T}\right)^{\frac{3}{4}}\right)$$

□

G. Proof of Theorem 7

Proof. For both SGD and RANDOMSHUFFLE, we use $s(i)$ to denote the index of component function picked in the ith iteration. We have the following inequality

$$||x_t - x^*||^2 = ||x_{t-1} - x^*||^2 - 2\gamma\langle x_{t-1} - x^*, \nabla f_{s(t)}(x_{t-1})\rangle + \gamma^2||\nabla f_{s(t)}(x_{t-1})||^2$$

$$
\begin{aligned}
&= ||x_{t-1} - x^*||^2 - 2\gamma\langle x_{t-1} - x^*, \nabla f_{s(t)}(x_{t-1}) - \nabla f_{s(t)}(x^*)\rangle + \gamma^2||\nabla f_{s(t)}(x_{t-1})||^2 \\
&\leq ||x_{t-1} - x^*||^2 - 2\gamma(\frac{||\nabla f_{s(t)}(x_{t-1}) - \nabla f_{s(t)}(x^*)||^2}{L_{s(t)} + \mu_{s(t)}} + \frac{L_{s(t)}\mu_{s(t)}}{L_{s(t)} + \mu_{s(t)}}||x_{t-1} - x^*||^2) + \gamma^2||\nabla f_{s(t)}(x_{t-1})||^2 \\
&= (1 - 2\gamma\frac{L_{s(t)}\mu_{s(t)}}{L_{s(t)} + \mu_{s(t)}})||x_{t-1} - x^*||^2 - \gamma(\frac{2}{L_{s(t)} + \mu_{s(t)}} - \gamma)||\nabla f_{s(t)}(x_{t-1})||^2 \\
&\leq (1 - 2\gamma\frac{L_{s(t)}\mu_{s(t)}}{L_{s(t)} + \mu_{s(t)}} + \mu_{s(t)}^2\gamma^2 - 2\gamma\frac{\mu_{s(t)}^2}{L_{s(t)} + \mu_{s(t)}})||x_{t-1} - x^*||^2 \\
&= (1 - 2\gamma\mu_{s(t)} + \mu_{s(t)}^2\gamma^2)||x_{t-1} - x^*||^2 \\
&= (1 - \gamma\mu_{s(t)})^2||x_{t-1} - x^*||^2.
\end{aligned}
$$

The first inequality is by Theorem 2.1.11 in (Nesterov, 2013), the second inequality is by the definition of strongly convexity. So we have

$$\mathbb{E}\,||x_T - x^*||^2 \leq \mathbb{E}[\prod_{i=1}^{T}(1 - \gamma\mu_{s(t)})^2]||x_{t-1} - x^*||^2.$$

By the AM-GM inequality, we know the term $\mathbb{E}[\prod_{i=1}^{T}(1 - \gamma\mu_{s(t)})^2]$ for RANDOMSHUFFLE is no larger than that of SGD. Also, this bound is tight when we consider $f_i(x) = \frac{\mu_i}{2}||x - x^*||^2$, which completes the proof.

□

H. SGD under Polyak-Łojasiewicz condition

For the completeness of the paper, we include the following analysis of SGD under Polyak-Łojasiewicz condition.

Theorem 1. *For finite sum problem satisfying Polyak-Łojasiewicz condition with parameter μ, Lipschitz constant L, setting step size*

$$\gamma = \frac{\log T}{\mu T},$$

there is

$$F(x_T) - F^* \leq \mathcal{O}(\frac{1}{T}).$$

Proof. We have the following one iteration for SGD with step size γ:

$$x_{t+1} = x_t - \gamma\nabla f_{s(t)}(x_t). \tag{H.1}$$

Given x_t, there is randomness over index

$$
\begin{aligned}
\mathbb{E}[F(x_{t+1})] - F^* &\leq F(x_t) - \gamma\,\mathbb{E}[\langle\nabla F(x_t), \nabla f_{s(t)}(x_t)\rangle] + \frac{L}{2}\gamma^2\,\mathbb{E}[\|\nabla f_i(x_t)\|^2] - F^* && \text{(H.2)}\\
&= F(x_t) - \gamma\langle\nabla F(x_t), \nabla F(x_t)\rangle + \frac{L}{2}\gamma^2\,\mathbb{E}[\|\nabla f_i(x_t)\|^2] - F^* && \text{(H.3)}\\
&\leq F(x_t) - \gamma\mu[F(x_t) - F^*] + \frac{L}{2}\gamma^2 G^2 - F^* && \text{(H.4)}\\
&= (1 - 2\gamma\mu)[F(x_t) - F^*] + \frac{L}{2}\gamma^2 G^2. && \text{(H.5)}
\end{aligned}
$$

The first inequality is because

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

The second inequality is because of the definition of PL condition.

Expanding over iterations leads to

$$\mathbb{E}[F(x_T) - F^*] \leq (1 - 2\gamma\mu)^T [F(x_0) - F^*] + \frac{L}{2} T \gamma^2 G^2.$$

Setting $\gamma = \frac{\log T}{\mu T}$ leads to a $O(\frac{1}{T})$ convergence of $F(x_T) - F^*$. $\square$

References

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.