Non-asymptotic rate for Random Shuffling for Quadratic functions

Raghav Somani

July 12, 2018

There have been many recent efforts understanding why Random Shuffling empirically converges faster than SGD. At each iteration, SGD samples a uniform index $i \in [n]$ and uses the stochastic gradient ∇f_i to compute its update. Uniformly sampling an index makes the stochastic gradient an unbiased estimate of the true gradient ∇f. However what actually is used in practice, and is computationally more practical, is a similar variant where at each epoch an index is uniformly sampled without replacement from a random permutation instead of sampling an index uniformly with replacement.

Empirically Random Shuffling is known to be faster than vanilla SGD [1] and understanding this discrepancy in theory and practice has been an open problem since long. Some known results that are available, show that Random Shuffle is not much worse than SGD [5] provided the number of epochs is not too large, while it has also been shown that Random Shuffle is faster than SGD asymptotically at the rate $\mathcal{O}\left(\frac{1}{T^2}\right)$ [2] compared to SGD that has an asymptotic rate of $\mathcal{O}\left(\frac{1}{T}\right)$, where T is the number of iterations. Under small fixed step size conditions, it has also been shown that the Random Shuffling converges to a smaller neighborhood compared to SGD [6], also discussed in one of my blogs here. Inspired from all the past works, there has been a recent attempt [3] where the authors come up with a non-asymptotic rate $\mathcal{O}\left(\frac{1}{T^2}+\frac{n^3}{T^3}\right)$, omitting constants and logarithmic factors, with n being the number of components in the function f, which is strictly better than SGD under reasonable conditions. The focus of this article is to throw light on this result by considering a quadratic objective function.

We consider minimization of the finite sum problem

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \tag{0.0.1}$$

It is assumed that the functions f and f_i's are μ-strongly convex and L-smooth respectively, i.e.,

$$f(\mathbf{x}) \geq f(y) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{\mu}{2}\left\|\mathbf{x}-\mathbf{y}\right\|_2^2 \tag{0.0.2}$$

$$\left\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\right\|_2 \leq L\left\|\mathbf{x}-\mathbf{y}\right\|_2 \tag{0.0.3}$$

Because $f(\mathbf{x})$ is strongly convex, let $\mathbf{x}^*$ be its minimizer. Also, let $\kappa = \frac{L}{\mu}$ denote the condition number of the function f.
The Random Shuffling update can be written as

$$\mathbf{x}_k^t = \mathbf{x}_{k-1}^t - \eta_t \nabla f_{\sigma_t(k)}(\mathbf{x}_{k-1}^t) \tag{0.0.4}$$

where $\mathbf{x}_k^t = \mathbf{x}_{(t-1)n+k}$ represents the k-th iteration within the t-th epoch. The step size for the t-th epoch is denoted by η_t and σ_t denotes the random permutation of $[n]$ with $\sigma_t(k)$ being its k-th index. For two consecutive epochs, $\mathbf{x}_n^t = x_0^{t+1}$ and $\mathbf{x}_0^1 = \mathbf{x}_0$.

1 Random Shuffle for Quadratic

For a simplistic start, let us first deal with the simple quadratic case of (0.0.1) where

$$f_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}_i\mathbf{x} + \mathbf{b}_i^T\mathbf{x} \qquad i = 1,\ldots,n \tag{1.0.1}$$

where $\mathbf{A}_i \in \mathbb{R}^{d\times d}$ is positive semi-definite and $\mathbf{b}_i \in \mathbb{R}^d$. The Hessians of quadratic functions are constant, which makes the analysis simpler. We further assume that the domain and the norm of the gradient in the domain is bounded as

$$\begin{aligned}\|\mathbf{x}-\mathbf{x}^*\|_2 &\leq D,\\ \|\nabla f_i(\mathbf{x})\|_2 &\leq G \qquad i=1,\dots,n\end{aligned}$$

1.1 Convergence Analysis

Considering the t^{th} epoch as a whole

$$\begin{aligned}\mathbf{x}_i^t &= \mathbf{x}_{i-1}^t - \eta_t \nabla f_{\sigma_t(i)}(\mathbf{x}_{i-1}^t)\\ \implies \mathbf{x}_n^t &= \mathbf{x}_0^t - \eta_t \sum_{i=1}^n \nabla f_{\sigma_t(i)}(\mathbf{x}_{i-1}^t)\\ \mathbf{x}_n^t - \mathbf{x}^* &= \mathbf{x}_0^t - \mathbf{x}^* - \eta_t \sum_{i=1}^n \nabla f_{\sigma_t(i)}(\mathbf{x}_{i-1}^t) \qquad (1.1.1)\end{aligned}$$

Taking Euclidean norm on both sides and squaring we get

$$\begin{aligned}\left\|\mathbf{x}_n^t - \mathbf{x}^*\right\|_2^2 &= \left\|\mathbf{x}_0^t - \mathbf{x}^* - \eta_t \sum_{i=1}^n \nabla f_{\sigma_t(i)}(\mathbf{x}_{i-1}^t)\right\|_2^2\\ &= \left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 - 2\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \sum_{i=1}^n \nabla f_{\sigma_t(i)}(\mathbf{x}_{i-1}^t)\right\rangle + \eta_t^2 \left\|\sum_{i=1}^n \nabla f_{\sigma_t(i)}(\mathbf{x}_{i-1}^t)\right\|_2^2 \qquad (1.1.2)\end{aligned}$$

Defining the error term R^t as

$$\begin{aligned}\mathbf{r}^t &= \sum_{i=1}^n \nabla f_{\sigma_t(i)}(\mathbf{x}_{i-1}^t) - \sum_{i=1}^n \nabla f_{\sigma_t(i)}(\mathbf{x}_0^t)\\ &= \sum_{i=1}^n \nabla f_{\sigma_t(i)}(\mathbf{x}_{i-1}^t) - n\nabla f(\mathbf{x}_0^t) \qquad (1.1.3)\end{aligned}$$

From (1.1.2) and (1.1.3) we get

$$\begin{aligned}\left\|\mathbf{x}_n^t - \mathbf{x}^*\right\|_2^2 &= \left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 - 2n\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \nabla f(\mathbf{x}_0^t)\right\rangle - 2\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbf{r}^t\right\rangle + \eta_t^2 \left\|n\nabla F(\mathbf{x}_0^t) + \mathbf{r}^t\right\|_2^2\\ &\leq \left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 - 2n\eta_t \left[\frac{L\mu}{L+\mu}\left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 + \frac{1}{L+\mu}\left\|\nabla f(\mathbf{x}_0^t)\right\|_2^2\right] - 2\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbf{r}^t\right\rangle + \eta_t^2 \left\|n\nabla F(\mathbf{x}_0^t) + \mathbf{r}^t\right\|_2^2\\ &\qquad\qquad \text{(From Theorem 2.1.11 in [4])}\\ &\leq \left(1 - 2n\eta_t \frac{L\mu}{L+\mu}\right)\left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 - \left(\frac{2n\eta_t}{L+\mu} - 2\eta_t^2 n^2\right)\left\|\nabla f(\mathbf{x}_0^t)\right\|_2^2 - 2\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbf{r}^t\right\rangle + 2\eta_t^2 \left\|\mathbf{r}^t\right\|_2^2 \qquad (1.1.4)\end{aligned}$$

Taking expectations with respect to σ_t on both side of (1.1.4) we have

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{x}_n^t - \mathbf{x}^*\right\|_2^2\right] &\leq \left(1 - 2n\eta_t \frac{L\mu}{L+\mu}\right)\left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 - \left(\frac{2n\eta_t}{L+\mu} - 2\eta_t^2 n^2\right)\left\|\nabla f(\mathbf{x}_0^t)\right\|_2^2\\ &\quad - 2\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}\left[\mathbf{r}^t\right]\right\rangle + 2\eta_t^2 \mathbb{E}\left[\left\|\mathbf{r}^t\right\|_2^2\right] \qquad (1.1.5)\end{aligned}$$

We can separately analyze the two terms with expectations in (1.1.5)

$$\left\|\mathbf{r}^t\right\|_2 = \left\|\sum_{i=1}^n \nabla f_{\sigma_t(i)}(\mathbf{x}_{i-1}^t) - \sum_{i=1}^n \nabla f_{\sigma_t(i)}(\mathbf{x}_0^t)\right\|_2$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{n} \left\| f_{\sigma_t(i)}(\mathbf{x}_{i-1}^t) - f_{\sigma_t(i)}(\mathbf{x}_0^t) \right\|_2 \\
&= \sum_{i=1}^{n} \left\| \sum_{j=1}^{i-1} f_{\sigma_t(i)}(\mathbf{x}_j^t) - f_{\sigma_t(i)}(\mathbf{x}_{j-1}^t) \right\|_2 \\
&\leq \sum_{i=1}^{n} \sum_{j=1}^{i-1} L \left\| \mathbf{x}_j^t - \mathbf{x}_{j-1}^t \right\|_2 \\
&\leq \sum_{i=1}^{n} \sum_{j=1}^{i-1} L \left\| -\eta_t \nabla f_{\sigma_t(j)}(\mathbf{x}_{j-1}^t) \right\|_2 \\
&\leq \eta_t L \sum_{i=1}^{n} \sum_{j=1}^{i-1} G \\
&= \frac{n(n-1)}{2} \eta_t L G
\end{aligned} \tag{1.1.6}
$$

It is to note that the above bound is a deterministic bound independent of the random permutation, therefore there might be a scope of tightness while talking about the expectation.
From (1.1.6) we have

$$
\mathbb{E}\left[\left\| \mathbf{r}^t \right\|_2^2 \right] \leq \frac{n^4}{4} \eta_t^2 G^2 L^2 \tag{1.1.7}
$$

And,

$$
\begin{aligned}
\mathbf{r}^t &= \sum_{i=1}^{n} \left[\nabla f_{\sigma_t(i)}(\mathbf{x}_{i-1}^t) - \nabla f_{\sigma_t(i)}(\mathbf{x}_0^t) \right] \\
&= \sum_{i=1}^{n} \left[\mathbf{H}_{\sigma_t(i)} (\mathbf{x}_{i-1}^t - \mathbf{x}_0^t) \right] \\
&= \sum_{i=1}^{n} \left[\mathbf{H}_{\sigma_t(i)} \sum_{j=1}^{i-1} \left[-\eta_t \nabla f_{\sigma_t(j)}(\mathbf{x}_{j-1}^t) \right] \right] \\
&= \sum_{i=1}^{n} \left[-\eta_t \mathbf{H}_{\sigma_t(i)} \sum_{j=1}^{i-1} \left[\nabla f_{\sigma_t(j)}(\mathbf{x}_0^t) + \left(\nabla f_{\sigma_t(j)}(\mathbf{x}_{j-1}^t) - \nabla f_{\sigma_t(j)}(\mathbf{x}_0^t) \right) \right] \right] \\
&= -\eta_t \sum_{i=1}^{n} \mathbf{H}_{\sigma_t(i)} \sum_{j=1}^{i-1} \nabla f_{\sigma_t(j)}(\mathbf{x}_0^t) - \eta_t \sum_{i=1}^{n} \left[\mathbf{H}_{\sigma_t(i)} \sum_{j=1}^{i-1} \left[\nabla f_{\sigma_t(j)}(\mathbf{x}_{j-1}^t) - \nabla f_{\sigma_t(j)}(\mathbf{x}_0^t) \right] \right]
\end{aligned} \tag{1.1.8}
$$

Separately analyzing the two terms in (1.1.8)

$$
\begin{aligned}
\mathbf{a}^t &:= -\eta_t \sum_{i=1}^{n} \mathbf{H}_{\sigma_t(i)} \sum_{j=1}^{i-1} \nabla f_{\sigma_t(j)}(\mathbf{x}_0^t) \\
\mathbf{b}^t &:= -\eta_t \sum_{i=1}^{n} \left[\mathbf{H}_{\sigma_t(i)} \sum_{j=1}^{i-1} \left[\nabla f_{\sigma_t(j)}(\mathbf{x}_{j-1}^t) - \nabla f_{\sigma_t(j)}(\mathbf{x}_0^t) \right] \right]
\end{aligned}
$$

Then we have

$$
\mathbb{E}\left[\mathbf{a}^t \right] = -\frac{n(n-1)}{2} \eta_t \mathbb{E}_{i \neq j} \left[\mathbf{H}_i \nabla f_j(\mathbf{x}_0^t) \right], \tag{1.1.9}
$$

$$
\left\| \mathbf{b}^t \right\|_2 \leq \eta_t \sum_{i=1}^{n} \left[\left\| \mathbf{H}_{\sigma_t(i)} \right\|_2 \sum_{j=1}^{i-1} \left\| \nabla f_{\sigma_t(j)}(\mathbf{x}_{j-1}^t) - \nabla f_{\sigma_t(j)}(\mathbf{x}_0^t) \right\|_2 \right]
$$

$$
\begin{aligned}
&\leq \eta_t \sum_{i=1}^{n} \left[L \sum_{j=1}^{i-1} (j-1)\eta_t G L \right] \\
&\leq \eta_t^2 L^2 G \sum_{i=1}^{n} \frac{(i-1)(i-2)}{2} \\
&\leq \frac{1}{2}\eta_t^2 L^2 G n^3 \qquad (1.1.10)
\end{aligned}
$$

Now expanding the first term with expectation in (1.1.5), we have

$$
\begin{aligned}
-2\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}\left[\mathbf{r}^t\right] \right\rangle &= -2\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}\left[\mathbf{a}^t\right] \right\rangle - 2\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}\left[\mathbf{b}^t\right] \right\rangle \\
&= \eta_t^2 n(n-1) \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}_{i\neq j}\left[\mathbf{H}_i \nabla f_j(\mathbf{x}_0^t)\right] \right\rangle - 2\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}\left[\mathbf{b}^t\right] \right\rangle \qquad (1.1.11)
\end{aligned}
$$

The first term in (1.1.11)

$$
\begin{aligned}
&\eta_t^2 n(n-1) \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}_{i\neq j}\left[\mathbf{H}_i \nabla f_j(\mathbf{x}_0^t)\right] \right\rangle \\
=&\eta_t^2 n(n-1) \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}_{i\neq j}\left[\mathbf{H}_i \left[\nabla f_j(\mathbf{x}_0^t) - \nabla f_j(\mathbf{x}^*)\right]\right] \right\rangle + \eta_t^2 n(n-1) \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}_{i\neq j}\left[\mathbf{H}_i \nabla f_j(\mathbf{x}^*)\right] \right\rangle \\
\leq&\eta_t^2 n(n-1) \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}_{i,j}\left[\mathbf{H}_i \mathbf{H}_j\right] \left(\mathbf{x}_0^t - \mathbf{x}^*\right) \right\rangle + \eta_t^2 n(n-1) \left[\frac{\lambda_1}{2} \left\| \mathbf{x}_0^t - \mathbf{x}^* \right\|_2^2 + \frac{1}{2\lambda_1} \left\| \boldsymbol{\delta} \right\|_2^2 \right] \\
=&\eta_t^2 n(n-1) \left\| \nabla f(\mathbf{x}_0^t) \right\|_2^2 + \frac{1}{4}\eta_t \mu (n-1) \left\| \mathbf{x}_0^t - \mathbf{x}^* \right\|_2^2 + \eta_t^3 \mu^{-1} n^2 (n-1) \left\| \boldsymbol{\delta} \right\|_2^2 \qquad (1.1.12)
\end{aligned}
$$

Here $\boldsymbol{\delta} := \mathbb{E}_{i\neq j}\left[\mathbf{H}_i \nabla f_j(\mathbf{x}^*)\right]$ and $\lambda_1 := \frac{1}{2}\mu\eta_t^{-1}n^{-1}$. Bounding the norm of $\boldsymbol{\delta}$ we get

$$
\begin{aligned}
\left\| \boldsymbol{\delta} \right\|_2 &= \left\| \mathbb{E}_{i\neq j}\left[\mathbf{H}_i \nabla f_j(\mathbf{x}^*)\right] \right\|_2 \\
&= \left\| \frac{1}{n(n-1)} \left[\sum_{i=1}^{n} \mathbf{H}_i \sum_{j=1}^{n} \nabla f_j(\mathbf{x}^*) - \sum_{i=1}^{n} \mathbf{H}_i \nabla f_i(\mathbf{x}^*) \right] \right\|_2 \\
&= \left\| -\frac{1}{n(n-1)} \left[\sum_{i=1}^{n} \mathbf{H}_i \nabla f_i(\mathbf{x}^*) \right] \right\|_2 \\
&= \frac{1}{n-1} \left\| \mathbb{E}_i\left[\mathbf{H}_i \nabla f_i(\mathbf{x}^*)\right] \right\|_2 \\
&\leq \frac{1}{n-1} L G \qquad (1.1.13)
\end{aligned}
$$

Using (1.1.13) in (1.1.12) we get

$$
\begin{aligned}
\eta_t^2 n(n-1) \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}_{i\neq j}\left[\mathbf{H}_i \nabla f_j(\mathbf{x}_0^t)\right] \right\rangle &\leq \eta_t^2 n(n-1) \left\| \nabla f(\mathbf{x}_0^t) \right\|_2^2 + \frac{1}{4}\eta_t \mu (n-1) \left\| \mathbf{x}_0^t - \mathbf{x}^* \right\|_2^2 + \frac{\eta_t^3 \mu^{-1} n^2 L^2 G^2}{n-1} \\
&\leq \eta_t^2 n(n-1) \left\| \nabla f(\mathbf{x}_0^t) \right\|_2^2 + \frac{1}{4}\eta_t \mu (n-1) \left\| \mathbf{x}_0^t - \mathbf{x}^* \right\|_2^2 + 2\eta_t^3 \mu^{-1} n L^2 G^2 \qquad (1.1.14)
\end{aligned}
$$

The second term in (1.1.11)

$$
-2\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}\left[\mathbf{b}^t\right] \right\rangle \leq 2\eta_t \left[\frac{\lambda_2}{2} \left\| \mathbf{x}_0^t - \mathbf{x}^* \right\|_2^2 + \frac{1}{2\lambda_2} \left\| \mathbb{E}\left[\mathbf{b}^t\right] \right\|_2^2 \right]
$$

Setting $\lambda_2 := \frac{1}{4}\mu(n-1)$, we get

$$
\begin{aligned}
-2\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}\left[\mathbf{b}^t\right] \right\rangle &\leq \frac{1}{4}\eta_t \mu (n-1) \left\| \mathbf{x}_0^t - \mathbf{x}^* \right\|_2^2 + 4\eta_t \mu^{-1} (n-1)^{-1} \left\| \mathbb{E}\left[\mathbf{b}^t\right] \right\|_2^2 \\
&\leq \frac{1}{4}\eta_t \mu (n-1) \left\| \mathbf{x}_0^t - \mathbf{x}^* \right\|_2^2 + \mu^{-1} (n-1)^{-1} \eta_t^5 L^4 G^2 n^6 \\
&\leq \frac{1}{4}\eta_t \mu (n-1) \left\| \mathbf{x}_0^t - \mathbf{x}^* \right\|_2^2 + 2\mu^{-1} \eta_t^5 L^4 G^2 n^5 \qquad (1.1.15)
\end{aligned}
$$

Plugging in (1.1.14) and (1.1.15) in (1.1.11) we get

$$-2\eta_t \left\langle \mathbf{x}_0^t - \mathbf{x}^*, \mathbb{E}\left[\mathbf{r}^t\right] \right\rangle \leq \eta_t^2 n^2 \left\|\nabla f(\mathbf{x}_0^t)\right\|_2^2 + \frac{1}{2}\eta_t \mu (n-1) \left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 + 2\eta_t^3 \mu^{-1} n L^2 G^2 + 2\mu^{-1}\eta_t^5 L^4 G^2 n^5 \quad (1.1.16)$$

Plugging back (1.1.16) and (1.1.7) back in (1.1.5) we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{x}_n^t - \mathbf{x}^*\right\|_2^2\right] &\leq \left(1 - 2n\eta_t \frac{L\mu}{L+\mu}\right) \left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 - \left(\frac{2n\eta_t}{L+\mu} - 2\eta_t^2 n^2\right) \left\|\nabla f(\mathbf{x}_0^t)\right\|_2^2 \\
&\quad \eta_t^2 n^2 \left\|\nabla f(\mathbf{x}_0^t)\right\|_2^2 + \frac{1}{2}\eta_t \mu (n-1) \left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 + 2\eta_t^3 \mu^{-1} n L^2 G^2 + 2\mu^{-1}\eta_t^5 L^4 G^2 n^5 + \frac{n^4}{2}\eta_t^4 G^2 L^2 \\
&= \left(1 - 2n\eta_t \frac{L\mu}{L+\mu} + \frac{1}{2}\eta_t \mu (n-1)\right) \left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 - \left(\frac{2n\eta_t}{L+\mu} - 3\eta_t^2 n^2\right) \left\|\nabla f(\mathbf{x}_0^t)\right\|_2^2 \\
&\quad + 2\eta_t^3 \mu^{-1} n L^2 G^2 + 2\mu^{-1}\eta_t^5 L^4 G^2 n^5 + \frac{n^4}{2}\eta_t^4 G^2 L^2 \quad (1.1.17)
\end{aligned}$$

Note that because $\kappa \geq 1$, we always have

$$n\eta_t \frac{L\mu}{L+\mu} > \frac{1}{2}\eta_2 \mu (n-1) \quad (1.1.18)$$

We also assume that η_t is such that the coefficient of $\left\|\nabla f(\mathbf{x}_0^t)\right\|_2^2$ is always non-positive, that is

$$\begin{aligned}
&\frac{2n\eta_t}{L+\mu} \geq 3\eta_t^2 n^2 \\
&\implies \eta_t \leq \frac{2}{3n(L+\mu)} \qquad \forall\ t \geq 1 \quad (1.1.19)
\end{aligned}$$

Again, we assume that the coefficient of $\left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2$ is positive, which would imply

$$n\eta_t \frac{L\mu}{L+\mu} < 1 \qquad \forall\ t \geq 1$$

Using (1.1.18) and (1.1.19) in (1.1.17) we get

$$\mathbb{E}\left[\left\|\mathbf{x}_n^t - \mathbf{x}^*\right\|_2^2\right] \leq \left(1 - n\eta_t \frac{L\mu}{L+\mu}\right) \left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 + \eta_t^3 n C_1 + \eta_t^5 n^5 C_2 + \eta_t^4 n^4 C_3 \quad (1.1.20)$$

where $C_1 = 2\frac{L^2G^2}{\mu}$, $C_2 = 2\frac{L^4G^2}{\mu}$ and $C_3 = \frac{1}{2}L^2G^2$. Unrolling (1.1.20) for all epochs, we get

$$\mathbb{E}\left[\left\|\mathbf{x}_n^t - \mathbf{x}^*\right\|_2^2\right] \leq \prod_{i=1}^{t}\left(1 - n\eta_i \frac{L\mu}{L+\mu}\right) \left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 + \sum_{i=0}^{t-1}\prod_{j=0}^{i-1}\left(1 - n\eta_{t-j}\frac{L\mu}{L+\mu}\right)\left(\eta_{t-i}^3 n C_1 + \eta_{t-i}^5 n^5 C_2 + \eta_{t-i}^4 n^4 C_3\right) \quad (1.1.21)$$

Setting $\eta_t = \eta = \frac{4\log nt}{\mu nt}$, and let $T = nt$, we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{x}_n^t - \mathbf{x}^*\right\|_2^2\right] &\leq \left(1 - n\frac{4\log nt}{\mu nt}\frac{L\mu}{L+\mu}\right)^t \left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 + t\left(\eta^3 n C_1 + \eta^5 n^5 C_2 + \eta^4 n^4 C_3\right) \\
&= \left(1 - \frac{2\log nt}{t}\right)^{\frac{t}{2\log nt}2\log nt} \left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 + \frac{T}{n}\left(\frac{64(\log T)^3}{\mu^3 T^3} n C_1 + \frac{1024(\log T)^5}{\mu^5 T^5} n^5 C_2 + \frac{256(\log T)^4}{\mu^4 T^4} n^4 C_3\right) \\
&\leq \frac{1}{T^2}\left\|\mathbf{x}_0^t - \mathbf{x}^*\right\|_2^2 + C_4\frac{(\log T)^3}{T^2} + C_5 n^3 \frac{(\log T)^4}{T^3} + C_6 n^4 \frac{(\log T)^5}{T^4} \quad (1.1.22)
\end{aligned}$$

Bringing back the assumptions we have made till now, we get the following constraints on the step size

$$\frac{4\log T}{\mu T} \leq \frac{2}{3n(L+\mu)}$$

$$\implies \frac{T}{\log T} \geq 6n(1+\kappa) \tag{1.1.23}$$

$$\text{And, } n\frac{4\log T}{\mu T}\frac{L\mu}{L+\mu} < 1$$

$$\implies \frac{T}{\log T} > 4n\frac{L}{L+\mu} \qquad \text{(True if (1.1.23) holds)}$$

Therefore, with step-size chosen as $\eta_t = \frac{4\log T}{\mu T}$, after $\mathcal{O}(\kappa)$ number of epochs, we can guarantee that the error in the parameter space is of the order $\mathcal{O}\left(\frac{(\log T)^3}{T^2} + n^3\frac{(\log T)^4}{T^3} + n^4\frac{(\log T)^5}{T^4}\right)$ which is strictly better than that of SGD. From the analysis, we need at least $\mathcal{O}(\kappa)$ iterations to come up with a guarantee, that might not be feasible when the problem is ill-conditioned. We also see that we need to know the number of iterations before hand to set the step size, making it at-least $\mathcal{O}(n\kappa)$ times lower than that required by SGD under similar settings.

References

[1] Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. 2009.

[2] M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo. Why Random Reshuffling Beats Stochastic Gradient Descent. *ArXiv e-prints*, October 2015.

[3] J. Z. HaoChen and S. Sra. Random Shuffling Beats SGD after Finite Epochs. *ArXiv e-prints*, June 2018.

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

[5] Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pages 46–54, 2016.

[6] B. Ying, K. Yuan, S. Vlaski, and A. H. Sayed. Stochastic Learning under Random Reshuffling. *ArXiv e-prints*, March 2018.