Random Reshuffling converges to a smaller neighborhood than SGD

Raghav Somani

July 4, 2018

It has been empirically observed that the performance of Stochastic Gradient Algorithms depend on the sampling method involved in the iterations. The literature on the analysis of SGD based methods consider an unbiased estimator of the gradient at each step and there are decent tight bounds that agree with empirical observations in a good number of the practical scenarios.

When the empirical risk function $F(\mathbf{w}, \mathcal{X})$ of the data $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$, can be written as an average of the sample risk $f(\mathbf{w}, \mathbf{x}_i)$, a naive way to pick an unbiased estimator of the gradient $\nabla F(\mathbf{w}, \mathcal{X})$, is to uniformly pick a data point $\mathbf{x}_{\xi_i}$ where $\xi_i \sim U[n]$ and compute the gradient of its risk $\nabla f(\mathbf{w}, \mathbf{x}_{\xi_i})$. This estimator of the gradient is unbiased by the below argument

$$F(\mathbf{w}, \mathcal{X}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}, \mathbf{x}_i) \tag{0.0.1}$$

$$\therefore \nabla F(\mathbf{w}, \mathcal{X}) = \frac{1}{n}\sum_{i=1}^{n} \nabla f(\mathbf{w}, \mathbf{x}_i)$$

$$\mathbb{E}_{\xi_i}\left[\nabla f(\mathbf{w}, \mathbf{x}_{\xi_i})\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla f(\mathbf{w}, \mathbf{x}_i) \qquad \text{(Because } \xi_i \sim U[n]\text{)}$$

$$= \nabla F(\mathbf{w}, \mathcal{X})$$

Because the expectation of $\nabla f(\mathbf{w}, \mathbf{x}_{\xi_i})$ is the exact gradient of the empirical risk function, therefore it is an unbiased estimator.

It has been observed that incorporating random reshuffling into stochastic gradient implementation helps achieving better performance. To elaborate further, the algorithm is run multiple times on the data where each run is indexed by $k \geq 1$ and is refereed to as an epoch. The data is then randomly reshuffled via a random permutation σ^k for the next epoch. Therefore the i^{th} iteration of the k^{th} epoch samples $\mathbf{x}_{\sigma^k(i)}$ for computing the gradient estimate. This method of sampling data points is very efficient in practice as it reduces the random access overheads compared to the i.i.d. based sampling method.

This article mostly relies on the recent work [1], in which the authors show that Stochasic Gradient Descent algorithm with random reshuffling outperforms independent sampling with replacement by showing that the mean square error of the iterates at the end of each epoch is of the order $O\left(\eta^2\right)$. This is a significant improvement compared to the traditional Stochastic Gradient Descent with i.i.d. sampling where the same quantity is of the order $O\left(\eta\right)$.

Let us consider minimizing an empirical risk function $F(\mathbf{w}, \mathcal{X})$ as defined in (0.0.1). Also let us assume that F is strongly convex allowing us to ensure that its minimizer $\mathbf{w}^*$ is unique. We also assume that $f(\mathbf{w}, \mathbf{x}_i)$ is smooth and continuously differentiable for all i.
Let $F(\mathbf{w}, \mathcal{X})$ be μ-strongly convex, and $\nabla f(\mathbf{w}, \mathbf{x}_i)$ be L-Lipschitz continuous, i.e.,

$$\left\|\nabla f(\mathbf{w}_1, \mathbf{x}_i) - \nabla f(\mathbf{w}_1, \mathbf{x}_i)\right\|_2 \leq L\left\|\mathbf{w}_1 - \mathbf{w}_2\right\|_2 \qquad \forall\ i \in [n] \text{ and } \forall\ \mathbf{w}_i, \mathbf{w}_2 \in \mathbb{R}^d \tag{0.0.2}$$

$$\left\langle \nabla F(\mathbf{w}_1, \mathcal{X}) - \nabla F(\mathbf{w}_2, \mathcal{X}), \mathbf{w}_1 - \mathbf{w}_2 \right\rangle \geq \mu\left\|\mathbf{w}_1 - \mathbf{w}_2\right\|_2^2 \qquad \forall\ \mathbf{w}_i, \mathbf{w}_2 \in \mathbb{R}^d \tag{0.0.3}$$

Before directly analyzing the random reshuffling for SGD, we will first see how the traditional SGD algorithm with constant step-size using i.i.d samples of the data points, converges to a neighborhood.

1 Stochastic Gradient Descent with i.i.d. sampling

The traditional SGD algorithm translates into a repeated update strategy as below

$$\mathbf{w}_i = \mathbf{w}_{i-1} - \eta_t \nabla f(\mathbf{w}_{i-1}, \mathbf{x}_{\xi_i}) \tag{1.0.1}$$

When the step-size η_t constant equal to η, (1.0.1) becomes

$$\mathbf{w}_i = \mathbf{w}_{i-1} - \eta \nabla f(\mathbf{w}_{i-1}, \mathbf{x}_{\xi_i}) \tag{1.0.2}$$

1.1 Convergence Analysis

From (1.0.2) we obtain

$$\begin{aligned}
\mathbf{w}_i - \mathbf{w}^* &= \mathbf{w}_{i-1} - \mathbf{w}^* - \eta \nabla f(\mathbf{w}_{i-1}, \mathbf{x}_{\xi_i}) \\
\mathbb{E}\left[\|\mathbf{w}_i - \mathbf{w}^*\|_2^2\right] &= \mathbb{E}\left[\|\mathbf{w}_{i-1} - \mathbf{w}^* - \eta \nabla f(\mathbf{w}_{i-1}, \mathbf{x}_{\xi_i})\|_2^2\right] \\
&= \mathbb{E}\left[\|\mathbf{w}_{i-1} - \mathbf{w}^*\|_2^2\right] + \eta^2 \mathbb{E}\left[\|\nabla f(\mathbf{w}_{i-1}, \mathbf{x}_{\xi_i})\|_2^2\right] - 2\eta \mathbb{E}\left[\langle \mathbf{w}_{i-1} - \mathbf{w}^*, \nabla f(\mathbf{w}_{i-1}, \mathbf{x}_{\xi_i})\rangle\right] \\
&= \mathbb{E}\left[\|\mathbf{w}_{i-1} - \mathbf{w}^*\|_2^2\right] + \eta^2 \mathbb{E}\left[\|\nabla f(\mathbf{w}_{i-1}, \mathbf{x}_{\xi_i})\|_2^2\right] - 2\eta \mathbb{E}\left[\langle \mathbf{w}_{i-1} - \mathbf{w}^*, \nabla F(\mathbf{w}_{i-1}, \mathcal{X})\rangle\right]
\end{aligned} \tag{1.1.1}$$

From the strong convexity of F at $\mathbf{w}^*$, using (0.0.3), we have

$$\begin{aligned}
&\langle \nabla F(\mathbf{w}_{i-1}, \mathcal{X}) - \nabla F(\mathbf{w}^*, \mathcal{X}), \mathbf{w}_{i-1} - \mathbf{w}^* \rangle \geq \mu \|\mathbf{w}_{i-1} - \mathbf{w}^*\|_2^2 \\
\implies &2\eta \mathbb{E}\left[\langle \nabla F(\mathbf{w}_{i-1}, \mathcal{X}), \mathbf{w}_{i-1} - \mathbf{w}^* \rangle\right] \geq 2\mu\eta \mathbb{E}\left[\|\mathbf{w}_{i-1} - \mathbf{w}^*\|_2^2\right]
\end{aligned} \tag{1.1.2}$$

Plugging (1.1.2) in (1.1.1) we get

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{w}_i - \mathbf{w}^*\|_2^2\right] &\leq \mathbb{E}\left[\|\mathbf{w}_{i-1} - \mathbf{w}^*\|_2^2\right] + \eta^2 \mathbb{E}\left[\|\nabla f(\mathbf{w}_{i-1}, \mathbf{x}_{\xi_i})\|_2^2\right] - 2\mu\eta \mathbb{E}\left[\|\mathbf{w}_{i-1} - \mathbf{w}^*\|_2^2\right] \\
&= (1 - 2\mu\eta) \mathbb{E}\left[\|\mathbf{w}_{i-1} - \mathbf{w}^*\|_2^2\right] + \eta^2 \mathbb{E}\left[\|\nabla f(\mathbf{w}_{i-1}, \mathbf{x}_{\xi_i})\|_2^2\right]
\end{aligned}$$

Assuming that the stochastic gradient norm is upper bounded by a constant G in the domain of optimization,

$$\mathbb{E}\left[\|\mathbf{w}_i - \mathbf{w}^*\|_2^2\right] \leq (1 - 2\mu\eta) \mathbb{E}\left[\|\mathbf{w}_{i-1} - \mathbf{w}^*\|_2^2\right] + \eta^2 G^2 \tag{1.1.3}$$

Let $r_i := \mathbb{E}\left[\|\mathbf{w}_i - \mathbf{w}^*\|_2^2\right]$, we have

$$\begin{aligned}
r_i &\leq (1 - 2\mu\eta) r_{i-1} + \eta^2 G^2 \\
&\leq (1 - 2\mu\eta)^i r_0 + \sum_{j=0}^{i} (1 - 2\eta\mu)^j \eta^2 G^2 \\
&\leq (1 - 2\mu\eta)^i r_0 + \eta^2 G^2 \sum_{j=0}^{\infty} (1 - 2\eta\mu)^j \\
&= (1 - 2\mu\eta)^i r_0 + \frac{\eta^2 G^2}{2\eta\mu} \qquad (\text{Assuming } \mu < \frac{1}{2\mu}) \\
&= (1 - 2\mu\eta)^i r_0 + \frac{\eta}{2} \frac{G^2}{\mu}
\end{aligned} \tag{1.1.4}$$

Therefore as $i \to \infty$, $\mathbb{E}\left[\|\mathbf{w}_i - \mathbf{w}^*\|_2^2\right]$ converges to a neighborhood of size upper bounded by $\frac{\eta}{2}\frac{G^2}{\mu} = O(\eta)$.

2 Stochastic Gradient Descent with Random Reshuffling

The iteration update for SGD with Random Reshuffling for k^{th} epoch is

$$\mathbf{w}_i^k = \mathbf{w}_{i-1}^k - \eta \nabla f(\mathbf{w}_{i-1}^k, \mathbf{x}_{\sigma^k(i)}) \qquad i = 1, \ldots, n \tag{2.0.1}$$

with the boundary condition of

$$\mathbf{w}_0^k = \mathbf{w}_n^{k-1}$$

It is to note that the gradient estimator $\nabla f(\mathbf{w}_{i-1}^k, \mathbf{x}_{\sigma^k(i)})$ is a biased estimator of $\nabla F(\mathbf{w}_{i-1}^k, \mathcal{X})$ due to the properties of a random permutation. Let us see how.

The samples are now no longer picked independently from each other. This is because of the constraint that each data point has to be picked exactly once. To proof that this estimate is a biased estimate, we will first have a closer look at the properties of the permutation random variable σ^k.

$$\begin{aligned}
\sigma^k(i) &\neq \sigma^k(j) \qquad 1 \leq i \neq j \leq n \\
P\left\{\sigma^k(i) = j\right\} &= \frac{1}{n} \qquad 1 \leq j \leq n \\
P\left\{\sigma^k(i) = j \mid \sigma^k(1:i)\right\} &= \begin{cases} \frac{1}{n-i}, & j \notin \sigma^k(1:i) \\ 0, & j \in \sigma^k(1:i) \end{cases}
\end{aligned}$$

where $\sigma^k(1:i)$ is the collection of all the indexes of the permutation σ^k from 1 to i.

Now let us see why $\nabla f(\mathbf{w}_{i-1}^k, \mathbf{x}_{\sigma^k(i)})$ is a biased estimate of $\nabla F(\mathbf{w}_{i-1}^k, \mathcal{X})$.

$$\begin{aligned}
\mathbb{E}\left[\nabla f(\mathbf{w}_{i-1}^k, \mathbf{x}_{\sigma^k(i)}) \mid \mathbf{w}_{i-1}^k, \sigma^k(1:i-1)\right] &= \frac{1}{n-i+1} \sum_{j \notin \sigma^k(1:i-1)} \nabla f(\mathbf{w}_{i-1}^k, \mathbf{x}_j) \\
&\begin{cases} \neq \nabla F(\mathbf{w}_0^k, \mathcal{X}) & i > 1 \\ = \nabla F(\mathbf{w}_0^k, \mathcal{X}) & i = 1 \end{cases}
\end{aligned}$$

But because every sample is picked exactly once, therefore the sampled average of the estimators within an epoch is an unbiased estimator.

$$\frac{1}{n} \sum_{i=1}^{n} \nabla f(\mathbf{w}_{i-1}^k, \mathbf{x}_{\sigma^k(i)}) = \nabla F(\mathbf{w}_{i-1}^k, \mathcal{X})$$

2.1 Convergence Analysis

Let us define gradient noise variance at $\mathbf{w}^*$ as $\mathcal{V}$, i.e.,

$$\mathcal{V} := \frac{1}{n} \sum_{i=1}^{n} \|\nabla f(\mathbf{w}^*, \mathbf{x}_i)\|_2^2 \tag{2.1.1}$$

We will first look at the convergence of the first iterates of each epoch, followed by the convergence of all the iterates within an epoch.

For this, we will express the first iterate of the immediate next epoch $k+1$ in the terms of the iterates in the current epoch k.

$$\begin{aligned}
\mathbf{w}_0^{k+1} &= \mathbf{w}_n^k \\
&= \mathbf{w}_{n-1}^k - \eta \nabla f(\mathbf{w}_{n-1}^k, \mathbf{x}_{\sigma^k(n)}) \\
&= \mathbf{w}_0^k - \eta \sum_{i=1}^{n} \nabla f(\mathbf{w}_{i-1}^k, \mathbf{x}_{\sigma^k(i)}) \\
&= \mathbf{w}_0^k - \eta n \nabla F(\mathbf{w}_0^k, \mathcal{X}) - \eta \sum_{i=1}^{n} \left(\nabla f(\mathbf{w}_{i-1}^k, \mathbf{x}_{\sigma^k(i)}) - \nabla f(\mathbf{w}_0^k, \mathbf{x}_{\sigma^k(i)})\right)
\end{aligned} \tag{2.1.2}$$

Defining $\mathbf{g}_{\sigma^k(i)}(\mathbf{w}_{i-1}^k)$ as the incremental gradient noise, or the mismatch of the gradient approximations evaluated at $\mathbf{w}_0^k$ and $\mathbf{w}_{i-1}^k$,

$$\mathbf{g}_{\sigma^k(i)}(\mathbf{w}_{i-1}^k) := \nabla f(\mathbf{w}_{i-1}^k, \mathbf{x}_{\sigma^k(i)}) - \nabla f(\mathbf{w}_0^k, \mathbf{x}_{\sigma^k(i)}) \tag{2.1.3}$$

Let us also define the error vector as

$$\tilde{\mathbf{w}}_0^k := \mathbf{w}^* - \mathbf{w}_0^k \tag{2.1.4}$$

Simplifying (2.1.2) using (2.1.3) and (2.1.4), we get

$$\begin{aligned}
\tilde{\mathbf{w}}_0^{k+1} &= \tilde{\mathbf{w}}_0^k + \eta n \nabla F(\mathbf{w}_0^k, \mathcal{X}) + \eta \sum_{i=1}^{n} \mathbf{g}_{\sigma^k(i)}(\mathbf{w}_{i-1}^k) \\
\left\|\tilde{\mathbf{w}}_0^{k+1}\right\|_2^2 &= \left\|\tilde{\mathbf{w}}_0^k + \eta n \nabla F(\mathbf{w}_0^k, \mathcal{X}) + \eta \sum_{i=1}^{n} \mathbf{g}_{\sigma^k(i)}(\mathbf{w}_{i-1}^k)\right\|_2^2 \\
&\leq \frac{1}{t}\left\|\tilde{\mathbf{w}}_0^k + \eta n \nabla F(\mathbf{w}_0^k, \mathcal{X})\right\|_2^2 + \frac{\eta^2}{1-t}\left\|\sum_{i=1}^{n} \mathbf{g}_{\sigma^k(i)}(\mathbf{w}_{i-1}^k)\right\|_2^2 \qquad \text{(Using Jensen's inequality, } 0 < t < 1) \\
&\leq \frac{1}{t}\left\|\tilde{\mathbf{w}}_0^k + \eta n \nabla F(\mathbf{w}_0^k, \mathcal{X})\right\|_2^2 + \frac{\eta^2 n}{1-t}\sum_{i=1}^{n}\left\|\mathbf{g}_{\sigma^k(i)}(\mathbf{w}_{i-1}^k)\right\|_2^2 \qquad \text{(Sub-additivity of } \ell_2\text{-norms)}
\end{aligned} \tag{2.1.5}$$

Now we will upper bound both the terms in (2.1.5).

$$\begin{aligned}
\sum_{i=1}^{n}\left\|\mathbf{g}_{\sigma^k(i)}(\mathbf{w}_{i-1}^k)\right\|_2^2 &\leq \sum_{i=1}^{n} L^2 \left\|\mathbf{w}_{i-1}^k - \mathbf{w}_0^k\right\|_2^2 \\
&= L^2 \sum_{i=1}^{n}\left\|\sum_{j=1}^{i-1}(\mathbf{w}_j^k - \mathbf{w}_{j-1}^k)\right\|_2^2 \\
&\leq L^2 \sum_{i=1}^{n}(i-1)\sum_{j=1}^{i-1}\left\|\mathbf{w}_j^k - \mathbf{w}_{j-1}^k\right\|_2^2 \qquad \text{(Sub-additivity of } \ell_2\text{-norm)} \\
&= L^2 \sum_{j=1}^{n}\sum_{i=j+1}^{n}(i-1)\left\|\mathbf{w}_j^k - \mathbf{w}_{j-1}^k\right\|_2^2 \qquad \text{(Equating lower triangular sums in 2 ways)} \\
&\leq \frac{n^2 L^2}{2}\sum_{j=1}^{n}\left\|\mathbf{w}_j^k - \mathbf{w}_{j-1}^k\right\|_2^2
\end{aligned} \tag{2.1.6}$$

It remains to bound $\sum_{j=1}^{n}\left\|\mathbf{w}_j^k - \mathbf{w}_{j-1}^k\right\|_2^2$ now

$$\begin{aligned}
\left\|\mathbf{w}_j^k - \mathbf{w}_{j-1}^k\right\|_2^2 &= \eta^2\left\|\nabla f(\mathbf{w}_{j-1}^k, \mathbf{x}_{\sigma^k(j)})\right\|_2^2 \\
&\leq 2\eta^2\left\|\nabla f(\mathbf{w}_{j-1}^k, \mathbf{x}_{\sigma^k(j)}) - \nabla f(\mathbf{w}^*, \mathbf{x}_{\sigma^k(j)})\right\|_2^2 + 2\eta^2\left\|\nabla f(\mathbf{w}^*, \mathbf{x}_{\sigma^k(j)})\right\|_2^2 \\
&\leq 2\eta^2 L^2\left\|\tilde{\mathbf{w}}_{j-1}^k\right\|_2^2 + 2\eta^2\left\|\nabla f(\mathbf{w}^*, \mathbf{x}_{\sigma^k(j)})\right\|_2^2 \qquad \text{(Using the smoothness of } f\text{, (0.0.2))} \\
&\leq 2\eta^2 L^2\left(2\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + 2\left\|\mathbf{w}_{j-1}^k - \mathbf{w}_0^k\right\|_2^2\right) + 2\eta^2\left\|\nabla f(\mathbf{w}^*, \mathbf{x}_{\sigma^k(j)})\right\|_2^2 \\
&= 4\eta^2 L^2\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + 4\eta^2 L^2\left\|\mathbf{w}_{j-1}^k - \mathbf{w}_0^k\right\|_2^2 + 2\eta^2\left\|\nabla f(\mathbf{w}^*, \mathbf{x}_{\sigma^k(j)})\right\|_2^2
\end{aligned} \tag{2.1.7}$$

Summing up (2.1.7) over j we get

$$\begin{aligned}
\sum_{j=1}^{n}\left\|\mathbf{w}_j^k - \mathbf{w}_{j-1}^k\right\|_2^2 &\leq 4\eta^2 L^2 n\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + 2\eta^2 n\mathcal{V} + 4\eta^2 L^2\sum_{j=1}^{n}\left\|\mathbf{w}_{j-1}^k - \mathbf{w}_0^k\right\|_2^2 \\
&= 4\eta^2 L^2 n\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + 2\eta^2 n\mathcal{V} + 4\eta^2 L^2\sum_{j=1}^{n}\left\|\sum_{i=1}^{j-1}(\mathbf{w}_i^k - \mathbf{w}_{i-1}^k)\right\|_2^2 \\
&= 4\eta^2 L^2 n\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + 2\eta^2 n\mathcal{V} + 4\eta^2 L^2\sum_{j=1}^{n}\sum_{i=1}^{j-1}(j-1)\left\|\mathbf{w}_i^k - \mathbf{w}_{i-1}^k\right\|_2^2
\end{aligned}$$

$$= 4\eta^2 L^2 n \left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + 2\eta^2 n\mathcal{V} + 4\eta^2 L^2 \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} (j-1)\left\|\mathbf{w}_i^k - \mathbf{w}_{i-1}^k\right\|_2^2$$

(Equating lower trangular sums in 2 ways)

$$\leq 4\eta^2 L^2 n \left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + 2\eta^2 n\mathcal{V} + 2\eta^2 L^2 n^2 \sum_{j=1}^{n-1}\left\|\mathbf{w}_j^k - \mathbf{w}_{j-1}^k\right\|_2^2$$

$$\leq 4\eta^2 L^2 n \left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + 2\eta^2 n\mathcal{V} + 2\eta^2 L^2 n^2 \sum_{j=1}^{n}\left\|\mathbf{w}_j^k - \mathbf{w}_{j-1}^k\right\|_2^2$$

$$\implies \sum_{j=1}^{n}\left\|\mathbf{w}_j^k - \mathbf{w}_{j-1}^k\right\|_2^2 \leq \frac{4\eta^2 L^2 n \left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + 2\eta^2 n\mathcal{V}}{1 - 2\eta^2 L^2 n^2} \qquad \left(\text{Rearranging terms for } \eta \leq \frac{1}{\sqrt{2}Ln}\right) \tag{2.1.8}$$

Plugging back (2.1.8) in (2.1.6) we get

$$\sum_{i=1}^{n}\left\|\mathbf{g}_{\sigma^k(i)}(\mathbf{w}_{i-1}^k)\right\|_2^2 \leq \frac{\eta^2 L^2 n^3}{1 - 2\eta^2 L^2 n^2}(2L^2\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + \mathcal{V}) \tag{2.1.9}$$

Now we are left to bound the first term in (2.1.5)

$$\begin{aligned}
\left\|\tilde{\mathbf{w}}_0^k + \eta n \nabla F(\mathbf{w}_0^k, \mathcal{X})\right\|_2^2 &= \left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + 2\eta n\left\langle \tilde{\mathbf{w}}_0^k, \nabla F(\mathbf{w}_0^k, \mathcal{X})\right\rangle + \eta^2 n^2\left\|\nabla F(\mathbf{w}_0^k, \mathcal{X})\right\|_2^2 \\
&= \left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 - 2\eta n\left\langle \tilde{\mathbf{w}}_0^k, \nabla F(\mathbf{w}^*, \mathcal{X}) - \nabla F(\mathbf{w}_0^k, \mathcal{X})\right\rangle + \eta^2 n^2\left\|\nabla F(\mathbf{w}^*, \mathcal{X}) - \nabla F(\mathbf{w}_0^k, \mathcal{X})\right\|_2^2 \\
&\leq (1 - 2\eta\mu n + \eta^2 n^2 L^2)\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 \qquad \text{(Using (0.0.2) and (0.0.3))} \\
&\leq \left(1 - \frac{2\eta\mu n}{3}\right)^2\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 \qquad \left(\text{For } \eta \leq \frac{2\mu}{3L^2 n}\right)
\end{aligned} \tag{2.1.10}$$

Plugging back (**??**) and (2.1.10) in (2.1.5), we get

$$\left\|\tilde{\mathbf{w}}_0^{k+1}\right\|_2^2 \leq \frac{1}{t}\left(1 - \frac{2\eta\mu n}{3}\right)^2\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + \frac{\eta^2 n}{1-t}\frac{\eta^2 L^2 n^3}{1 - 2\eta^2 L^2 n^2}(2L^2\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + \mathcal{V}) \tag{2.1.11}$$

For $t = 1 - \frac{2\eta\mu n}{3}$ and assuming η is sufficiently small such that $1 - 2\eta^2 L^2 n^2 \geq \frac{3}{4}$ or $\eta \leq \frac{1}{\sqrt{8}Ln}$, we get

$$\begin{aligned}
\left\|\tilde{\mathbf{w}}_0^{k+1}\right\|_2^2 &\leq \left(1 - \frac{2\eta\mu n}{3}\right)\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + \frac{2\eta^3 L^2 n^3}{\mu}(2L^2\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + \mathcal{V}) \\
&\leq \left(1 - \frac{2\eta\mu n}{3} + \frac{4\eta^3 L^4 n^3}{\mu}\right)\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + \frac{2\eta^3 L^2 n^3}{\mu}\mathcal{V} \\
&\leq \left(1 - \frac{\eta n\mu}{2}\right)\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2 + \frac{2\eta^3 L^2 n^3}{\mu}\mathcal{V} \\
&\qquad \left(\text{Assuming } \left(1 - \frac{2\eta\mu n}{3} + \frac{4\eta^3 L^4 n^3}{\mu}\right) \leq \left(1 - \frac{\eta n\mu}{2}\right) \text{ or } \eta \leq \frac{\mu}{\sqrt{24}L^2 n}\right)
\end{aligned} \tag{2.1.12}$$

Unrolling (2.1.12) we get

$$\begin{aligned}
\left\|\tilde{\mathbf{w}}_0^{k+1}\right\|_2^2 &\leq \left(1 - \frac{\eta n\mu}{2}\right)^k\left\|\tilde{\mathbf{w}}_0^0\right\|_2^2 + \frac{2\eta^3 L^2 n^3}{\mu}\mathcal{V}\sum_{j=1}^{k}\left(1 - \frac{\eta\mu n}{2}\right)^j \\
&\leq \left(1 - \frac{\eta n\mu}{2}\right)^k\left\|\tilde{\mathbf{w}}_0^0\right\|_2^2 + \frac{4\eta^2 L^2 n^2}{\mu^2}\mathcal{V} \\
\implies \mathbb{E}\left[\left\|\tilde{\mathbf{w}}_0^{k+1}\right\|_2^2\right] &\leq \left(1 - \frac{\eta n\mu}{2}\right)^k\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_0^0\right\|_2^2\right] + \frac{4\eta^2 L^2 n^2}{\mu^2}\mathcal{V}
\end{aligned} \tag{2.1.13}$$

Combining all the assumptions on η we have considered so far, we have $\eta \leq \min\left\{\frac{2\mu}{3L^2 n}, \frac{1}{\sqrt{8}Ln}, \frac{\mu}{\sqrt{24}L^2 n}\right\}$. Therefore it suffices to have $\eta \leq \frac{\mu}{5L^2 n}$.

From (2.1.13) we see that the first iterate of each epoch converges linearly to a neighborhood of of size $\frac{4\eta^2L^2n^2}{\mu^2}\mathcal{V} = O\left(\eta^2\right)$. We can now proceed by proving the same for any iterate within any epoch.

$$\begin{aligned}
\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_i^k\right\|_2^2\right] &\leq 2\mathbb{E}\left[\left\|\mathbf{w}_i^k - \mathbf{w}_0^k\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\sum_{j=0}^{i-1}(\mathbf{w}_{j+1}^k - \mathbf{w}_j^k)\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2\right] \\
&\leq 2\sum_{j=0}^{i-1}\mathbb{E}\left[\left\|\mathbf{w}_{j+1}^k - \mathbf{w}_j^k\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2\right] \\
&= 2\eta^2\sum_{j=0}^{i-1}\mathbb{E}\left[\left\|\nabla f(\mathbf{w}_j^k, \mathbf{x}_{\sigma^k(j)})\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2\right] \\
&\leq 2\eta^2L^2\sum_{j=0}^{i-1}\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_j^k\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2\right]
\end{aligned}$$

Summing (2.1.14) over i, we get

$$\begin{aligned}
\sum_{i=1}^{n-1}\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_i^k\right\|_2^2\right] &\leq 2\eta^2L^2\sum_{i=1}^{n-1}\sum_{j=0}^{i-1}\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_j^k\right\|_2^2\right] + 2n\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2\right] \\
&= 2\eta^2L^2\sum_{j=0}^{n-1}\sum_{i=j+1}^{n-1}\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_j^k\right\|_2^2\right] + 2n\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2\right] \\
&\leq 2\eta^2L^2n\sum_{j=0}^{n-1}\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_j^k\right\|_2^2\right] + 2n\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2\right] \\
&= 2\eta^2L^2n\sum_{j=1}^{n-1}\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_j^k\right\|_2^2\right] + (2n + 2\eta^2L^2n)\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2\right] \\
\implies \sum_{i=1}^{n-1}\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_i^k\right\|_2^2\right] &\leq \frac{2n(1+\eta^2L^2)}{1-2\eta^2L^2n}\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_0^k\right\|_2^2\right] \qquad (2.1.14)
\end{aligned}$$

Letting $k \to \infty$, we have

$$\limsup_{k\to\infty}\sum_{i=1}^{n-1}\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_i^k\right\|_2^2\right] = O\left(\eta^2\right) \qquad (2.1.15)$$

Since every term in the sum in (2.1.15) is non-negative, we have

$$\limsup_{k\to\infty}\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_j^k\right\|_2^2\right] \leq \limsup_{k\to\infty}\sum_{i=1}^{n-1}\mathbb{E}\left[\left\|\tilde{\mathbf{w}}_i^k\right\|_2^2\right] = O\left(\eta^2\right) \qquad (2.1.16)$$

Now if we compare SGD with i.i.d. sampling and with random reshuffling, we could still see a potential looseness. Comparing (1.1.4) and (2.1.13) we see that due to the small step-size requirement in random reshuffling, the rate of convergence to the neighborhood is off by a factor of n. The linear term in (1.1.4) is upper bounded by $\exp\{-2\mu\eta i\}$ where i is the iteration index, whereas the linear convergence term in (2.1.13) is upper bounded by $\exp\left\{-\frac{\eta n\mu k}{2}\right\}$ where k is the epoch index. So the iteration index of the k^{th} epoch is $i = kn$. Substituting this we get the linear convergence rate as $\exp\left\{-\frac{\eta\mu i}{2}\right\}$. The step-size restriction in traditional SGD is of the order $\frac{1}{L}$ whereas the restriction that random reshuffling has in its analysis is of the order $\frac{\mu}{L^2n}$ which is lower by an order of $\frac{Ln}{\mu}$. It will be interesting to see if the restriction on the step size can be improved.

References

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