

Convergence Analysis of Random Reshuffling Under Strongly-Convex, Convex and Nonconvex Conditions

Qianxing Jia(SSE) 121090223

Abstract—In this report, we investigate the convergence properties of random reshuffling algorithms under various conditions, including strong convexity, convexity, and non-convexity. The focus of our study is on sample selection methods without replacement, which are commonly used in a wide range of applications. We analyze the behavior of these algorithms and provide theoretical insights into their convergence rates. Under the assumption of strong convexity, we establish rigorous convergence guarantees for random reshuffling algorithms, showing that they are capable of efficiently finding optimal solutions within a finite number of iterations. We derive explicit bounds on the convergence rates and explore the impact of different sampling strategies on the algorithm's performance. Moving beyond strong convexity, we extend our analysis to scenarios where the objective function is merely convex. We investigate the convergence behavior under these less restrictive conditions, highlighting the trade-offs between convergence speed and optimality. We provide insights into how the convexity of the objective function affects the performance of random reshuffling algorithms and discuss practical implications. Finally, we explore the challenging case of non-convex optimization, where the objective function may exhibit multiple local optima. We analyze the behavior of random reshuffling algorithms in this context, shedding light on their ability to escape local optima and converge to global solutions. We discuss the impact of sample selection without replacement on the exploration-exploitation trade-off and propose strategies to enhance the algorithm's performance in non-convex settings.

I. INTRODUCTION

We study the finite-sum minimization problem:

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

where each $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is differentiable and smooth, and are particularly interested in the big data machine learning setting where the number of functions n is large. Stochastic first order algorithms avoid huge computation cost and thus attracting machine learning community. In particular, the content of the paper is to obtain theoretical understanding of methods of solving (1) which rely on random permutation/shuffling of the data $\{1, 2, 3, \dots, n\}$ and perform incremental gradient updates following the permuted order.(see Algorithm 1). This method is the **Random Reshuffling(RR)**. In each epoch t of RR, we sample indices $\{\pi_0, \pi_1, \dots, \pi_{n-1}\}$ *without replacement* from $\{1, 2, 3, \dots, n\}$, i.e., $\{\pi_0, \pi_1, \dots, \pi_{n-1}\}$ is a random permutation of the set $\{1, 2, \dots, n\}$, and proceed with n iterates of the form

$$x_t^{i+1} = x_t^i - \gamma \nabla f_{\pi_i}(x_t^i)$$

where γ is the stepsize. We then set $x_{t+1} = x_t^n$, and repeat the process for a total of T epoch, which is why the term *reshuffling* is used.

Algorithm 1 Random Reshuffling(RR)

Input: Stepsize $\gamma > 0$, initial vector $x_0 = x_0^0 \in \mathbb{R}^d$, number of epochs T

- 1: **for** epochs $t = 0, 1, \dots, T - 1$ **do**
- 2: Sample a permutation $\pi_0, \pi_1, \dots, \pi_{n-1}$ of $\{1, 2, \dots, n\}$
- 3: **for** $i = 0, 1, \dots, n - 1$ **do**
- 4: $x_t^{i+1} = x_t^i - \gamma \nabla f_{\pi_i}(x_t^i)$
- 5: $x_{t+1} = x_t^n$

II. PRELIMINARIES AND ASSUMPTIONS

A. Sampling without replacement

The following algorithm-independent lemma characterizes the variance of sampling a number of vectors from a finite set of vectors, without replacement. It is a key ingredient in our results on the convergence of the RR.

Let $X_1, \dots, X_n \in \mathbb{R}^d$ be fixed vectors, $\bar{X} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n X_i$ be their average and $\sigma^2 \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \|X_i - \bar{X}\|^2$ be the population variance. Fix any $k \in \{1, \dots, n\}$, let $X_{\pi_1}, \dots, X_{\pi_k}$ be sampled uniformly without replacement from $\{X_1, \dots, X_n\}$ and \bar{X}_π be their average. The expectation is as followed:

$$\mathbb{E}[\bar{X}_\pi] = \frac{1}{k} \sum_{i=1}^k \mathbb{E}[X_{\pi_i}] = \frac{1}{k} \sum_{i=1}^k \bar{X} = \bar{X}.$$

To derive the variance, let us first establish that the covariance $\text{cov}(X_{\pi_i}, X_{\pi_j}) = -\frac{\sigma^2}{n-1}$ holds for any $i \neq j$. Indeed,

$$\begin{aligned} \text{cov}(X_{\pi_i}, X_{\pi_j}) &= \mathbb{E}[\langle X_{\pi_i} - \bar{X}, X_{\pi_j} - \bar{X} \rangle] \\ &= \frac{1}{n(n-1)} \sum_{l=1}^n \sum_{m \neq l}^n \langle X_l - \bar{X}, X_m - \bar{X} \rangle \\ &= \frac{1}{n(n-1)} \sum_{l=1}^n \left\langle X_l - \bar{X}, \sum_{m=1}^n (X_m - \bar{X}) \right\rangle - \frac{\sigma^2}{n-1} \\ &= -\frac{\sigma^2}{n-1}. \end{aligned}$$

This identity helps us to establish the formula for sample variance:

$$\begin{aligned}\mathbb{E}[\|\bar{X}_\pi - \bar{X}\|^2] &= \frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k \text{cov}(X_{\pi_i}, X_{\pi_j}) \\ &= \frac{1}{k^2} \mathbb{E} \left[\sum_{i=1}^k \|X_{\pi_i} - \bar{X}\|^2 \right] + \sum_{i=1}^k \sum_{j \neq i}^k \text{cov}(X_{\pi_i}, X_{\pi_j}) \\ &= \frac{1}{k^2} \left(k\sigma^2 - k(k-1) \frac{\sigma^2}{n-1} \right) \\ &= \frac{n-k}{k(n-1)} \sigma^2.\end{aligned}$$

B. Lipschitz gradient assumption

Throughout the paper, we impose the following Lipschitz gradient assumption on each component function $f_i(x)$ in problem (1). For all $i \in \{1, 2, \dots, n\}$, f_i is bounded from below and the gradient ∇f_i is Lipschitz continuous with parameter L . Suppose that the gradient of a continuously differentiable function $h: \mathbb{R}^n \rightarrow \mathbb{R}$ is Lipschitz continuous with parameter L . Then, applying the so-called descent lemma, it follows:

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

Let us set $h \equiv f_i$ for some $i \in \{1, 2, \dots, n\}$. Choosing $y = x - \frac{1}{L} \nabla f_i(x)$, we can infer that:

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

This variance-type bound will play a big role in our convergence analysis under nonconvex conditions.

III. μ -STRONGLY CONVEX CASE

A. Strong limit point convergence

The analysis of many stochastic methods, including SGD, rely on the fact that the iterates converge to x_* up to some noise. This is exactly where we part ways with the standard analysis techniques, since, it turns out, the intermediate iterates of shuffling algorithms converge to some other points. Given a permutation π , the real limit points are defined below,

$$x_*^i \doteq x_* - \gamma \sum_{j=0}^{i-1} \nabla f_{\pi_j}(x_*), i = 1, \dots, n-1 \quad (3)$$

Suppose that the functions f_1, \dots, f_n are μ -strongly convex and that Assumption 1 holds. Then for Algorithms 1 run with a constant stepsize $\gamma \leq \frac{1}{L}$,

$$E[\|x_T - x_*\|^2] \leq (1 - \gamma\mu)^{nT} \|x_0 - x_*\|^2 +$$

The key insight of the proof is that the intermediate iterates x_t^1, x_t^2, \dots do not converge to x_* , but rather converge to the sequence x_*^1, x_*^2, \dots defined by (2). Keeping this intuition in mind, it makes sense to study the following recursion:

$$\begin{aligned}\mathbb{E}[\|x_t^{i+1} - x_*^{i+1}\|^2] \\ = \mathbb{E}[\|x_t^i - x_*^i\|^2 - 2\gamma \langle \nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*), x_t^i - x_*^i \rangle \\ + \gamma^2 \|\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*)\|^2].\end{aligned} \quad (4)$$

Once we have this recursion, it is useful to notice that the scalar product can be decomposed as

$$\begin{aligned}\langle \nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*), x_t^i - x_*^i \rangle \\ = [f_{\pi_i}(x_*^i) - f_{\pi_i}(x_t^i) - \langle \nabla f_{\pi_i}(x_t^i), x_*^i - x_t^i \rangle] \\ + [f_{\pi_i}(x_t^i) - f_{\pi_i}(x_*) - \langle \nabla f_{\pi_i}(x_*), x_t^i - x_* \rangle] \\ - [f_{\pi_i}(x_*^i) - f_{\pi_i}(x_*) - \langle \nabla f_{\pi_i}(x_*), x_*^i - x_* \rangle]\end{aligned} \quad (5)$$

Utilizing μ -convexity of each component, we can derive an lower bound of the first block:

$$[f_{\pi_i}(x_*^i) - f_{\pi_i}(x_t^i) - \langle \nabla f_{\pi_i}(x_t^i), x_*^i - x_t^i \rangle] \geq \frac{\mu}{2} \|x_t^i - x_*^i\|^2 \quad (6)$$

According to Sham M. Kakade and Shai Shalev-Shwartz and Ambuj Tewari[2009], Assume that f is a closed and convex function. Then f is β -strongly convex w.r.t. a norm $\|\cdot\|$ if and only if f^* is $\frac{1}{\beta}$ -strongly smooth w.r.t. the dual norm $\|\cdot\|_*$.

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Utilizing the lemma above, we can derive a lower bound of the second block, that is:

$$\begin{aligned}f_{\pi_i}(x_t^i) - f_{\pi_i}(x_*) - \langle \nabla f_{\pi_i}(x_*) , x_t^i - x_* \rangle \geq \\ \frac{1}{2L} \|\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*)\|^2\end{aligned} \quad (7)$$

Utilizing randomness, smoothness and convexity of each component, we derive an upper bound of the third block. That is,

$$\begin{aligned}E[f_{\pi_i}(x_*^i) - f_{\pi_i}(x_*) - \langle \nabla f_{\pi_i}(x_*) , x_*^i - x_* \rangle] \\ \leq \frac{L}{2} E[\|x_*^i - x_*\|^2] = \frac{L}{2} E[\|\sum_{k=0}^{i-1} \gamma \nabla f_{\pi_k}(x_*)\|^2] \\ = \frac{\gamma^2 L i(n-i)}{2(n-1)} \sigma_*^2 \\ \leq \frac{\gamma^2 L n}{4} \sigma_*^2\end{aligned} \quad (8)$$

Notice that $\sigma_*^2 = \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x_*)\|^2$. Combining (3),(4),(5),(6),(7), we derive an upper bound of expected distant within one iteration.

$$\begin{aligned}\mathbb{E}[\|x_t^{i+1} - x_*^{i+1}\|^2] \leq \mathbb{E}[(1 - \gamma\mu)\|x_t^i - x_*^i\|^2] - \\ (\frac{\gamma}{L} - \gamma^2) \|\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*)\|^2 + \frac{\gamma^3 L n}{2} \sigma_*^2\end{aligned} \quad (9)$$

By utilizing (8) within one epoch, we derive the upper bound of expected distance:

$$\begin{aligned}\mathbb{E}[\|x_{t+1} - x_*\|^2] \leq (1 - \gamma\mu)^n \mathbb{E}[\|x_t - x_*\|^2] \\ + \frac{\gamma^3 L n}{2} \sigma_*^2 \left(\sum_{i=0}^{n-1} (1 - \gamma\mu)^i \right).\end{aligned}$$

Unrolling this recursion across T epochs, we obtain

$$\begin{aligned}\mathbb{E}[\|x_T - x_*\|^2] \leq (1 - \gamma\mu)^{nT} \|x_0 - x_*\|^2 + \\ \frac{\gamma^3 L n}{2} \sigma_*^2 \left(\sum_{i=0}^{n-1} (1 - \gamma\mu)^i \right) \left(\sum_{j=0}^{T-1} (1 - \gamma\mu)^{nj} \right).\end{aligned}$$

The product of the two sums can be bounded as follows:

$$\begin{aligned}
& \left(\sum_{j=0}^{T-1} (1 - \gamma\mu)^{nj} \right) \left(\sum_{i=0}^{n-1} (1 - \gamma\mu)^i \right) \\
&= \sum_{j=0}^{T-1} \sum_{i=0}^{n-1} (1 - \gamma\mu)^{nj+i} \\
&= \sum_{k=0}^{nT-1} (1 - \gamma\mu)^k \\
&\leq \sum_{k=0}^{\infty} (1 - \gamma\mu)^k = \frac{1}{\gamma\mu}.
\end{aligned}$$

Plugging this bound back, we finally obtain the bound

$$\mathbb{E} [\|x_T - x_*\|^2] \leq (1 - \gamma\mu)^{nT} \|x_0 - x_*\|^2 + \frac{\gamma^2 L n}{2\mu} \sigma_*^2. \quad (10)$$

B. Convergence rate

It's noticeable that we have a noise term in equation (9). If we choose our γ fixed, then the bias will not decrease to zero. On the other hand, if we choose our γ decreases linearly with $\frac{1}{T}$, the first term of equation (9) will not decrease to zero. Therefore, our perfect stepsize is:

$$\gamma = \min\left\{\frac{1}{L}, \frac{\log T}{T}\right\}$$

Case 1: when $\frac{1}{L} \leq \frac{\log T}{T}$

$$\begin{aligned}
\mathbb{E} [\|x_T - x_*\|^2] &\leq (1 - \frac{\mu}{L})^{nT} \|x_0 - x_*\|^2 + \frac{n}{2L\mu} \sigma_*^2 \\
&\leq (1 - \frac{\mu}{L})^{nT} \|x_0 - x_*\|^2 + \frac{n \log T}{2T\mu} \sigma_*^2
\end{aligned} \quad (11)$$

Using that $1 - x \leq e^{-x}$ in the previous inequality, we get

$$\mathbb{E} [\|x_T - x_*\|^2] = O(\exp(-\frac{\mu n T}{L}) \|x_0 - x_*\|^2 + \frac{\log T}{T})$$

Case 2: when $\frac{1}{L} \geq \frac{\log T}{T}$

$$\begin{aligned}
\mathbb{E} [\|x_T - x_*\|^2] &\leq (1 - \frac{\mu}{L})^{nT} \|x_0 - x_*\|^2 + \frac{n}{2L\mu} \sigma_*^2 \\
&\leq \exp(-\mu n \log T) \|x_0 - x_*\|^2 + \frac{n \log T}{2T\mu} \sigma_*^2 \\
&\leq \frac{\exp(-\mu n)}{T} \|x_0 - x_*\|^2 + \frac{n \log T}{2T\mu} \sigma_*^2
\end{aligned} \quad (12)$$

then, it's obvious that

$$\mathbb{E} [\|x_T - x_*\|^2] = O(\frac{\log T}{T})$$

Combining these two cases, we get the convergence rate of **Random Reshuffling** under μ -strongly convex case is $O(\exp(-\frac{\mu n T}{L}) \|x_0 - x_*\|^2 + \frac{\log T}{T})$

IV. CONVEX CASE

Now we take a look at the case when each component is convex. That is there may not exist a finite L to bound the gradient of each f_i .

A. Bregman divergence

We define the *Bregman divergence* between x and y associated with f_i as $D_{f_i}(x, y) = f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle$. It's well-known that if f_i is L -smooth, then for all $x, y \in \mathbb{R}^d$

$$D_{f_i}(x, y) \leq \frac{L}{2} \|x - y\|^2 \quad (13)$$

Moreover, the difference between the gradients of a convex and L -smooth f_i is related to its Bregman divergence by

$$\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L \cdot D_{f_i}(x, y)$$

B. Bounding forward per-epoch deviation

Let $x_t^0, x_t^1, \dots, x_t^n$ be iterates generated by Algorithms 1. We define the forward per-epoch deviation over the t -th epoch \mathcal{V}_t as

$$\mathcal{V}_t \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} \|x_t^i - x_{t+1}\|^2$$

If the functions f_1, \dots, f_n are convex and each of them is μ -smooth. Then for any fixed $k \in \{0, \dots, n-1\}$, by definition of x_t^k and x_{t+1} we get the decomposition

$$\begin{aligned}
x_t^k - x_{t+1} &= \gamma \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_t^i) \\
&= \gamma \sum_{i=k}^{n-1} (\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*)) + \gamma \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*)
\end{aligned} \quad (14)$$

Applying Young's inequality to the sums above yields

$$\begin{aligned}
\|x_t^k - x_{t+1}\|^2 &\leq 2\gamma^2 \left\| \sum_{i=k}^{n-1} (\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*)) \right\|^2 \\
&\quad + 2\gamma^2 \left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2 \\
&\leq 2\gamma^2 n \sum_{i=k}^{n-1} \|\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*)\|^2 \\
&\quad + 2\gamma^2 \left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2 \\
&\leq 4\gamma^2 L n \sum_{i=k}^{n-1} D_{f_{\pi_i}}(x_*, x_t^i) + 2\gamma^2 \left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2 \\
&\leq 4\gamma^2 L n \sum_{i=0}^{n-1} D_{f_{\pi_i}}(x_*, x_t^i) + 2\gamma^2 \left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2
\end{aligned} \quad (15)$$

Summing up and taking expectations leads to

$$\begin{aligned}
\sum_{k=0}^{n-1} \mathbb{E} [\|x_t^k - x_{t+1}\|^2] &\leq 4\gamma^2 L n^2 \sum_{i=0}^{n-1} \mathbb{E} [D_{f_{\pi_i}}(x_*, x_t^i)] + \\
&\quad 2\gamma^2 \sum_{k=0}^{n-1} \mathbb{E} \left[\left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2 \right]
\end{aligned} \quad (16)$$

We now bound the second term in the right-hand side of (15). First, using sampling without replacement, we get

$$\begin{aligned}\mathbb{E} \left[\left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2 \right] &= (n-k)^2 \mathbb{E} \left[\left\| \frac{1}{n-k} \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2 \right] \\ &= (n-k)^2 \frac{k}{(n-k)(n-1)} \sigma_*^2 \\ &= \frac{k(n-k)}{n-1} \sigma_*^2.\end{aligned}$$

Next, by summing this for k from 0 to $n-1$, we obtain

$$\begin{aligned}\sum_{k=0}^{n-1} \mathbb{E} \left[\left\| \sum_{i=k}^{n-1} \nabla f_{\pi_i}(x_*) \right\|^2 \right] &= \sum_{k=0}^{n-1} \frac{k(n-k)}{n-1} \sigma_*^2 \\ &= \frac{1}{6} n(n+1) \sigma_*^2 \\ &\leq \frac{n^2 \sigma_*^2}{4}\end{aligned}\tag{17}$$

Plugging it back to equation (15), we derive an upper bound of forward per-epoch deviation.

$$\mathbb{E}[\mathcal{V}_t] \leq 4\gamma^2 n^2 L \sum_{i=0}^{n-1} \mathbb{E}[D_{f_{\pi_i}}(x_*, x_t^i)] + \frac{1}{2} \gamma^2 n^2 \sigma_*^2 \tag{18}$$

C. Finding a recursion

Assume that functions f_1, \dots, f_n are convex and that Assumption 1 is satisfied. If Random Reshuffling (Algorithm 1) is run with a stepsize satisfying $\gamma \leq \frac{1}{\sqrt{2Ln}}$, then

$$\begin{aligned}\mathbb{E}[\|x_{t+1} - x_*\|^2] &\leq \mathbb{E}[\|x_t - x_*\|^2] - 2\gamma n \mathbb{E}[f(x_{t+1}) - f_*] \\ &\quad + \frac{\gamma^3 L n^2 \sigma_*^2}{2}\end{aligned}\tag{19}$$

Define the sum of gradients used in the t -th epoch as $g_t \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} \nabla f_{\pi_i}(x_t^i)$. We will use g_t to relate the iterates x_t and x_{t+1} . By definition of x_{t+1} , we can write

$$\begin{aligned}x_{t+1} &= x_t^n = x_t^{n-1} - \gamma \nabla f_{\pi_{n-1}}(x_t^{n-1}) \\ &= \dots \\ &= x_t^0 - \gamma \sum_{i=0}^{n-1} \nabla f_{\pi_i}(x_t^i)\end{aligned}$$

Further, since $x_t^0 = x_t$, we see that $x_{t+1} = x_t - \gamma g_t$, which leads to

$$\begin{aligned}\|x_t - x_*\|^2 &= \|x_{t+1} + \gamma g_t - x_*\|^2 \\ &= \|x_{t+1} - x_*\|^2 + 2\gamma \langle g_t, x_{t+1} - x_* \rangle + \gamma^2 \|g_t\|^2 \\ &\geq \|x_{t+1} - x_*\|^2 + 2\gamma \langle g_t, x_{t+1} - x_* \rangle \\ &= \|x_{t+1} - x_*\|^2 + 2\gamma \sum_{i=0}^{n-1} \langle \nabla f_{\pi_i}(x_t^i), x_{t+1} - x_* \rangle.\end{aligned}\tag{20}$$

Observe that for any i , we have the following decomposition

$$\begin{aligned}\langle \nabla f_{\pi_i}(x_t^i), x_{t+1} - x_* \rangle &= [f_{\pi_i}(x_{t+1}) - f_{\pi_i}(x_*)] \\ &\quad + [f_{\pi_i}(x_*) - f_{\pi_i}(x_t^i) - \langle \nabla f_{\pi_i}(x_t^i), x_t^i - x_* \rangle] \\ &\quad - [f_{\pi_i}(x_{t+1}) - f_{\pi_i}(x_t^i) - \langle \nabla f_{\pi_i}(x_t^i), x_{t+1} - x_t^i \rangle] \\ &= [f_{\pi_i}(x_{t+1}) - f_{\pi_i}(x_*)] + D_{f_{\pi_i}}(x_*, x_t^i) - D_{f_{\pi_i}}(x_{t+1}, x_t^i).\end{aligned}\tag{21}$$

Summing the first term in the right in (20) over i from 0 to $n-1$ gives

$$\sum_{i=0}^{n-1} [f_{\pi_i}(x_{t+1}) - f_{\pi_i}(x_*)] = n(f(x_{t+1}) - f_*).$$

Now, we can bound the third term in the decomposition (20) using L -smoothness as follows:

$$D_{f_{\pi_i}}(x_{t+1}, x_t^i) \leq \frac{L}{2} \|x_{t+1} - x_t^i\|^2.$$

By summing the right-hand side over i from 0 to $n-1$ we get the forward deviation over an epoch \mathcal{V}_t , which we bound by preceding lemma to get

$$\begin{aligned}&\sum_{i=0}^{n-1} \mathbb{E}[D_{f_{\pi_i}}(x_{t+1}, x_t^i)] \\ &\leq \frac{L}{2} \mathbb{E}[\mathcal{V}_t] \\ &\leq 2\gamma^2 L^2 n^2 \sum_{i=0}^{n-1} \mathbb{E}[D_{f_{\pi_i}}(x_*, x_t^i)] + \frac{\gamma^2 L n^2 \sigma_*^2}{4}\end{aligned}\tag{22}$$

Therefore, we can lower-bound the sum of the second and the third term in (20) as

$$\begin{aligned}&\sum_{i=0}^{n-1} \mathbb{E}[D_{f_{\pi_i}}(x_*, x_t^i) - D_{f_{\pi_i}}(x_{t+1}, x_t^i)] \\ &\geq \sum_{i=0}^{n-1} \mathbb{E}[D_{f_{\pi_i}}(x_*, x_t^i)] - 2\gamma^2 L^2 n^2 \sum_{i=0}^{n-1} \mathbb{E}[D_{f_{\pi_i}}(x_*, x_t^i)] \\ &\quad - \frac{\gamma^2 L n^2 \sigma_*^2}{4} \\ &\geq (1 - 2\gamma^2 L^2 n^2) \sum_{i=0}^{n-1} \mathbb{E}[D_{f_{\pi_i}}(x_*, x_t^i)] - \frac{\gamma^2 L n^2 \sigma_*^2}{4} \\ &\geq - \frac{\gamma^2 L n^2 \sigma_*^2}{4}\end{aligned}\tag{23}$$

where in the third inequality we used that $\gamma \leq \frac{1}{\sqrt{2Ln}}$ and that $D_{f_{\pi_i}}(x_*, x_t^i)$ is nonnegative. Plugging this back into the lower-bound on $\|x_t - x_*\|^2$ yields

$$\begin{aligned}\mathbb{E}[\|x_t - x_*\|^2] &\geq \mathbb{E}[\|x_{t+1} - x_*\|^2] \\ &\quad + 2\gamma n \mathbb{E}[f(x_{t+1}) - f_*] - \frac{\gamma^3 L n^2 \sigma_*^2}{2}\end{aligned}\tag{24}$$

D. Strongly convex f

Since f is μ -convex, we can get:

$$\begin{aligned}\mathbb{E}[\|x_{t+1} - x_*\|^2] &\leq \mathbb{E}[\|x_t - x_*\|^2] - 2\gamma n \mathbb{E}[f(x_{t+1}) - f_*] \\ &\quad + \frac{\gamma^3 L n^2 \sigma_*^2}{2} \\ &\leq \mathbb{E}[\|x_t - x_*\|^2] - \gamma n \mu \mathbb{E}[\|x_{t+1} - x_*\|^2] \\ &\quad + \frac{\gamma^3 L n^2 \sigma_*^2}{2},\end{aligned}\tag{25}$$

whence

$$\begin{aligned}\mathbb{E}[\|x_{t+1} - x_*\|^2] &\leq \frac{1}{1 + \gamma \mu n} \left(\mathbb{E}[\|x_t - x_*\|^2] + \frac{\gamma^3 L n^2 \sigma_*^2}{2} \right) \\ &= \frac{1}{1 + \gamma \mu n} \mathbb{E}[\|x_t - x_*\|^2] + \frac{\gamma^3 L n^2 \sigma_*^2}{2} \\ &\leq \left(1 - \frac{\gamma \mu n}{2}\right) \mathbb{E}[\|x_t - x_*\|^2] + \frac{\gamma^3 L n^2 \sigma_*^2}{2}\end{aligned}$$

Recurring for T iterations, we get that the final iterate satisfies

$$\begin{aligned}\mathbb{E}[\|x_T - x_*\|^2] &\leq \left(1 - \frac{\gamma \mu n}{2}\right)^T \|x_0 - x_*\|^2 \\ &\quad + \frac{\gamma^3 L n^2 \sigma_*^2}{2} \left(\sum_{j=0}^{T-1} \left(1 - \frac{\gamma \mu n}{2}\right)^j \right) \\ &\leq \left(1 - \frac{\gamma \mu n}{2}\right)^T \|x_0 - x_*\|^2 + \frac{\gamma^3 L n^2 \sigma_*^2}{2} \left(\frac{2}{\gamma \mu n} \right) \\ &= \left(1 - \frac{\gamma \mu n}{2}\right)^T \|x_0 - x_*\|^2 + \gamma^2 \kappa n \sigma_*^2\end{aligned}\tag{26}$$

Just like the way we derive the convergence rate when each f_i is strongly convex, we can get the convergence rate directly from equation (25).

E. Convex f

If we remove the strongly convex condition of f , again with equation (23). Rearranging the result leads to

$$\begin{aligned}2\gamma n \mathbb{E}[f(x_{t+1}) - f(x_*)] \\ \leq \mathbb{E}[\|x_t - x_*\|^2] - \mathbb{E}[\|x_{t+1} - x_*\|^2] + \frac{\gamma^3 L n^2 \sigma_*^2}{2}\end{aligned}\tag{27}$$

Summing these inequalities for $t = 0, 1, \dots, T-1$ gives

$$\begin{aligned}2\gamma n \sum_{t=0}^{T-1} \mathbb{E}[f(x_{t+1}) - f(x_*)] \\ \leq \sum_{t=0}^{T-1} \left(\mathbb{E}[\|x_t - x_*\|^2] - \mathbb{E}[\|x_{t+1} - x_*\|^2] \right) + \frac{\gamma^3 L n^2 \sigma_*^2 T}{2} \\ = \|x_0 - x_*\|^2 - \mathbb{E}[\|x_T - x_*\|^2] + \frac{\gamma^3 L n^2 \sigma_*^2 T}{2} \\ \leq \|x_0 - x_*\|^2 + \frac{\gamma^3 L n^2 \sigma_*^2 T}{2}\end{aligned}\tag{28}$$

and dividing both sides by $2\gamma n T$, we get

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[f(x_{t+1}) - f(x_*)] \leq \frac{\|x_0 - x_*\|^2}{2\gamma n T} + \frac{\gamma^2 L n \sigma_*^2}{4}.$$

Finally, using convexity of f , the average iterate $\hat{x}_T \stackrel{\text{def}}{=} \frac{1}{T} \sum_{t=1}^T x_t$ satisfies

$$\begin{aligned}\mathbb{E}[f(\hat{x}_T) - f(x_*)] &\leq \frac{1}{T} \sum_{t=1}^T \mathbb{E}[f(x_t) - f(x_*)] \\ &\leq \frac{\|x_0 - x_*\|^2}{2\gamma n T} + \frac{\gamma^2 L n \sigma_*^2}{4}\end{aligned}\tag{29}$$

From equation (28), it's not difficult to find if we choose our stepsize to be the minimum of $\frac{1}{L}$ and $T^{-\frac{1}{3}}$ we can guarantee that $\mathbb{E}[f(\hat{x}_T) - f(x_*)] \leq \epsilon$ after $O(T^{-\frac{2}{3}})$ times.

V. NONCONVEX CASE

We now turn our look to when there are no assumptions about the convexity of each component. However, we have the following assumption on gradients assumption.

A. Gradients variance

We start with (2) (which does not require convexity) applied to each f_i :

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

Averaging, we derive

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x)\|^2 &\leq 2L \left(f(x) - \frac{1}{n} \sum_{i=1}^n f_i^* \right) \\ &= 2L(f(x) - f_*) + 2L \left(f_* - \frac{1}{n} \sum_{i=1}^n f_i^* \right)\end{aligned}\tag{30}$$

Note that because f_* is the infimum of $f(\cdot)$ and $\frac{1}{n} \sum_{i=1}^n f_i^*$ is a lower bound on f then $f_* - \frac{1}{n} \sum_{i=1}^n f_i^* \geq 0$. We may now use the variance decomposition

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x) - \nabla f(x)\|^2 &\stackrel{(21)}{=} \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x)\|^2 - \|\nabla f(x)\|^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x)\|^2 \\ &\leq 2L(f(x) - f_*) + 2L \left(f_* - \frac{1}{n} \sum_{i=1}^n f_i^* \right)\end{aligned}\tag{31}$$

B. Finding a per-epoch recursion

Recall that $x_{t+1} = x_t - \gamma g_t$, where $g_t = \sum_{i=0}^{n-1} \nabla f_{\pi_i}(x_t^i)$. Using L -smoothness of f , we get

$$\begin{aligned}f(x_{t+1}) &\leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\ &= f(x_t) - \gamma n \left\langle \nabla f(x_t), \frac{g_t}{n} \right\rangle + \frac{\gamma^2 L n^2}{2} \left\| \frac{g_t}{n} \right\|^2 \\ &= f(x_t) - \frac{\gamma n}{2} \left(\|\nabla f(x_t)\|^2 + \left\| \frac{g_t}{n} \right\|^2 - \left\| \nabla f(x_t) - \frac{g_t}{n} \right\|^2 \right) \\ &\quad + \frac{\gamma^2 L n^2}{2} \left\| \frac{g_t}{n} \right\|^2 \\ &= f(x_t) - \frac{\gamma n}{2} \|\nabla f(x_t)\|^2 - \frac{\gamma n}{2} (1 - L\gamma n) \left\| \frac{g_t}{n} \right\|^2 \\ &\quad + \frac{\gamma n}{2} \left\| \nabla f(x_t) - \frac{g_t}{n} \right\|^2\end{aligned}\tag{32}$$

By assumption, we have $\gamma \leq \frac{1}{Ln}$, and hence $1 - L\gamma n \geq 0$. Using this in (32), we get

$$f(x_{t+1}) \leq f(x_t) - \frac{\gamma n}{2} \|\nabla f(x_t)\|^2 + \frac{\gamma n}{2} \left\| \nabla f(x_t) - \frac{g_t}{n} \right\|^2 \quad (33)$$

For the last term in (33), we note

$$\begin{aligned} \left\| \nabla f(x_t) - \frac{g_t}{n} \right\|^2 &= \left\| \frac{1}{n} \sum_{i=0}^{n-1} [\nabla f_{\pi_i}(x_t) - \nabla f_{\pi_i}(x_t^i)] \right\|^2 \\ &\stackrel{(18)}{\leq} \frac{1}{n} \sum_{i=0}^{n-1} \|\nabla f_{\pi_i}(x_t) - \nabla f_{\pi_i}(x_t^i)\|^2 \\ &\stackrel{(14)}{\leq} \frac{1}{n} \sum_{i=0}^{n-1} L^2 \|x_t - x_t^i\|^2 = \frac{L^2}{n} V_t \end{aligned} \quad (34)$$

Plugging in (34) into (33) yields the claim:

$$f(x_{t+1}) \leq f(x_t) - \frac{\gamma n}{2} \|\nabla f(x_t)\|^2 + \frac{\gamma L^2}{2} V_t \quad (35)$$

Here $V_t = \frac{1}{n} \sum_{i=1}^n \|x_t^i - x_t\|^2$ is the backward per-epoch deviation.

C. Bounding backward per-epoch deviation

By Young's inequality, Jensen's inequality and gradient Lipschitzness,

$$\begin{aligned} \mathbb{E}_t \left[\|x_t^k - x_t\|^2 \right] &= \gamma^2 \mathbb{E}_t \left[\left\| \sum_{i=0}^{k-1} \nabla f_{\pi_i}(x_t^i) \right\|^2 \right] \\ &\leq 2\gamma^2 \mathbb{E}_t \left[\left\| \sum_{i=0}^{k-1} (\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_t)) \right\|^2 \right] \\ &\quad + 2\gamma^2 \mathbb{E}_t \left[\sum_{i=0}^{k-1} \|\nabla f_{\pi_i}(x_t)\|^2 \right] \\ &\leq 2\gamma^2 k \sum_{i=0}^{k-1} \mathbb{E}_t \left[\|\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_t)\|^2 \right] \\ &\quad + 2\gamma^2 \mathbb{E}_t \left[\left\| \sum_{i=0}^{k-1} \nabla f_{\pi_i}(x_t) \right\|^2 \right] \\ &\leq 2\gamma^2 L^2 k \sum_{i=0}^{k-1} \mathbb{E}_t \left[\|x_t^i - x_t\|^2 \right] \\ &\quad + 2\gamma^2 \mathbb{E}_t \left[\left\| \sum_{i=0}^{k-1} \nabla f_{\pi_i}(x_t) \right\|^2 \right] \end{aligned} \quad (36)$$

Let us bound the second term. For any i we have $\mathbb{E}_t [\nabla f_{\pi_i}(x_t)] = \nabla f(x_t)$, so using Lemma 1 (with vectors

$\nabla f_{\pi_0}(x_t), \nabla f_{\pi_1}(x_t), \dots, \nabla f_{\pi_{k-1}}(x_t)$) we obtain

$$\begin{aligned} &\mathbb{E}_t \left[\left\| \sum_{i=0}^{k-1} \nabla f_{\pi_i}(x_t) \right\|^2 \right] \\ &= k^2 \|\nabla f(x_t)\|^2 + k^2 \mathbb{E}_t \left[\left\| \frac{1}{k} \sum_{i=0}^{k-1} (\nabla f_{\pi_i}(x_t) - \nabla f(x_t)) \right\|^2 \right] \\ &\leq k^2 \|\nabla f(x_t)\|^2 + \frac{k(n-k)}{n-1} \sigma_t^2 \end{aligned} \quad (37)$$

where $\sigma_t^2 \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^n \|\nabla f_j(x_t) - \nabla f(x_t)\|^2$. Combining the produced bounds yields

$$\begin{aligned} &\mathbb{E}_t \left[\|x_t^k - x_t\|^2 \right] \\ &\leq 2\gamma^2 L^2 k \sum_{i=0}^{k-1} \mathbb{E}_t \left[\|x_t^i - x_t\|^2 \right] + 2\gamma^2 k^2 \|\nabla f(x_t)\|^2 + 2\gamma^2 \frac{k(n-k)}{n-1} \sigma_t^2 \\ &\leq 2\gamma^2 L^2 k \mathbb{E}[V_t] + 2\gamma^2 k^2 \|\nabla f(x_t)\|^2 + 2\gamma^2 \frac{k(n-k)}{n-1} \sigma_t^2, \end{aligned} \quad (38)$$

whence

$$\begin{aligned} \mathbb{E}[V_t] &= \sum_{k=0}^{n-1} \mathbb{E}_t \left[\|x_t^k - x_t\|^2 \right] \\ &\leq \gamma^2 L^2 n(n-1) \mathbb{E}[V_t] + \frac{1}{3} \gamma^2 (n-1)n(2n-1) \|\nabla f(x_t)\|^2 + \\ &\quad \frac{1}{3} \gamma^2 n(n+1) \sigma_t^2 \end{aligned} \quad (39)$$

Since $\mathbb{E}[V_t]$ appears in both sides of the equation, we rearrange and use that $\gamma \leq \frac{1}{2Ln}$ by assumption, which leads to

$$\begin{aligned} \mathbb{E}[V_t] &\leq \frac{4}{3} (1 - \gamma^2 L^2 n(n-1)) \mathbb{E}[V_t] \\ &\leq \frac{4}{9} \gamma^2 (n-1)n(2n-1) \|\nabla f(x_t)\|^2 + \frac{4}{9} \gamma^2 n(n+1) \sigma_t^2 \\ &\leq \gamma^2 n^3 \|\nabla f(x_t)\|^2 + \gamma^2 n^2 \sigma_t^2. \end{aligned} \quad (40)$$

D. Solving recursion

Combining (35) and (40), let $\delta_t = f(x_t) - f_*$, and take expectation at both sides:

$$\mathbb{E}[\delta_{t+1}] \leq (1 + \gamma^3 A L^2 n^2) \mathbb{E}[\delta_t] - \frac{\gamma n}{4} \mathbb{E}[\|\nabla f(x_t)\|^2] + \frac{\gamma^3 L^2 n^2 B^2}{2} \quad (41)$$

where $A = L$ and $B^2 = 2L(f_* - \frac{1}{n} \sum_{i=1}^n f_i^*)$

We then apply a very useful lemma. Suppose that there exist constants $a, b, c \geq 0$ and nonnegative sequences $(s_t)_{t=0}^T, (q_t)_{t=0}^T$ such that for any t satisfying $0 \leq t \leq T$ we have the recursion

$$s_{t+1} \leq (1 + a)s_t - bq_t + c.$$

Then, the following holds:

$$\min_{t=0, \dots, T-1} q_t \leq \frac{(1+a)^T}{bT} s_0 + \frac{c}{b}.$$

Here we can view s as the function value and q as the descent per iteration or epoch and c is some noise term. By using the above lemma, we get:

$$\min_{t=0,\dots,T-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \leq \frac{12(f(x_0) - f_*)}{\gamma n T} + 2\gamma^2 L^2 n B^2 \quad (42)$$

It's not difficult to find that if we choose our stepszie $O(T^{-\frac{1}{3}})$, $\mathbb{E} \left[\|\nabla f(x_t)\|^2 \right]$ converges with the rate $O(T^{-\frac{2}{3}})$

VI. FUTURE WORK: RR ON RIEMANNIAN MANIFOLDS

Algorithm 2 Random Reshuffling(RR) on Riemannian Manifolds

Input: Stepszie $\gamma > 0$, a retraction Retr on \mathcal{M} , initial vector $x_0 = x_0^0 \in \mathcal{M}^d$, number of epochs T

```

1:  for epochs  $t = 0, 1, \dots, T - 1$  do
2:      Sample a permutation  $\pi_0, \pi_1, \dots, \pi_{n-1}$  of  $\{1, 2, \dots, n\}$ 
3:      for  $i = 0, 1, \dots, n - 1$  do
4:           $\eta_k = -\text{grad}f_{\pi_k}$ 
4:           $x_t^{i+1} = \text{Retr}_{x_t^i}(\gamma\eta_k)$ 
5:       $x_{t+1} = x_t^n$ 

```

In the above algorithm, "grad" stands for the Riemannian gradient with respect to the given manifolds \mathcal{M} . The main difficulty of generalizing the convergence analysis of random reshuffling on manifolds is that riemannian gradients lie in different subspaces within one epoch. Therefore, it's difficult if we apply the preceding analysis structure that view one epoch as a whole.

REFERENCES

- [1] K. Mishchenko, K. Thuwal, S. Arabia, A. Khaled, and P. Richtárik, "Random Reshuffling: Simple Analysis with Vast Improvements," Apr. 2021.
- [2] S. Kakade, S. Shalev-Shwartz, and A. Tewari, "On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization," 2009.
- [3] S. Bonnabel, "Stochastic Gradient Descent on Riemannian Manifolds," *IEEE Transactions on Automatic Control*, vol. 58, no. 9, pp. 2217–2229, Sep. 2013, doi: <https://doi.org/10.1109/tac.2013.2254619>.
- [4] X. Li, A. Milzarek, and J. Qiu, "Convergence of Random Reshuffling under the Kurdyka–Łojasiewicz Inequality," *SIAM Journal on Optimization*, vol. 33, no. 2, pp. 1092–1120, Jun. 2023, doi: <https://doi.org/10.1137/21m1468048>.