Random Reshuffling: Simple Analysis with Vast Improvements

Konstantin Mishchenko et al. Neurips 2020 accepted

Kyunghun Nam

09.13.2023

Table of Contents

Problem definition

$$\min_{x} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

where each f_i is differentiable and smooth.
Several methods belong to the class of data permutation methods, and this paper focuses on Random Reshuffling (**RR**) algorithm.

Algorithm

Algorithm 1 Random Reshuffling (RR)

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

for epochs $t = 0, 1, \ldots, T-1$ **do**

  Sample a permutation $\pi_0, \pi_1, \ldots, \pi_{n-1}$ of $\{1, 2, \ldots, n\}$

  for $i = 0, 1, \ldots, n-1$ **do**

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

  $x_{t+1} = x_t^n$

In each epoch t, we sample indices $\pi_0, \pi_1, \cdots, \pi_{n-1}$ without replacement from $\{1, 2, \cdots, n\}$, *i.e.*, $\{\pi_0, \pi_1, \cdots, \pi_{n-1}\}$ is a random permutation of the set $\{1, 2, \cdots, n\}$, and proceed with n iterates of the form

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

More challenging problems

Notice that in RR, a new permutation/shuffling is generated at the beginning of each epoch, which is why the term reshuffling is used.

Sampling without replacement allows RR to leverage the finite-sum structure by ensuring that each function contributes to the solution once per epoch.

On the other hand, it also introduces a significant complication: the steps are now biased.

$$\mathbb{E}[\nabla f_{\pi_i}(x_t^i)] \neq \nabla f(x_t^i)$$

Assumptions and well-known lemma

Assumption

1. *The objective f and the individual losses $f_1, \cdots, f_n$ are all L-smooth.*
2. *f is lower bounded by some $f^\star$. If f is convex, we also assume the existence of a minimizer $x^\star$ and $f^\star := f(x^\star)$.*

Lemma

1. *When f_i is L-smooth and μ-strongly convex, then*

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

2. *When f_i is convex and L-smooth, then*

$$\frac{1}{2L}\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq f_i(x) - f_i(y) - \langle \nabla f_i(y), x-y \rangle$$

Table of Contents

New notion of variance specific to RR

Given a permutation π, the real limit points are defined below,

$$x_\star^i := x_\star - \gamma \sum_{j=0}^{i-1} \nabla f_{\pi_j}(x_\star), \quad i = 1, 2, \cdots, n-1$$

Shuffling variance
Given a step size $\gamma > 0$ and a random permutation π. Then the shuffling variance is given by,

$$\sigma_{shuffle}^2 := \max_{i=1,\cdots,n-1} \left[\frac{1}{\gamma} \mathbb{E}[D_{f_{\pi_i}}(x_\star^i, x_\star)] \right]$$

where $D_{f_i}(x, y) := f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle$ is the Bregman divergence between x and y associated with f_i.

New notion of variance specific to RR

Proposition

Suppose that each $f_1, f_2, \cdots, f_n$ is μ-strongly convex and L-smooth. Then

$$\frac{\gamma \mu n}{8}\sigma_\star^2 \leq \sigma_{shuffle}^2 \leq \frac{\gamma L n}{4}\sigma_\star^2$$

where $\sigma_\star^2 := \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_\star)\|^2$.

Proof of Proposition

The proof of the proposition requires the following lemma.

Lemma

Let $X_1, \cdots, X_n$ be fixed vectors, $\bar{X}$ be their average and σ^2 be the population variance. Fix any $k \in \{1, 2, \cdots, n\}$, let $X_{\pi_1}, X_{\pi_2}, \cdots, X_{\pi_k}$ be sampled uniformly without replacement from $\{X_1, X_2, \cdots, X_n\}$ and $\bar{X}_\pi$ be their average.
Then the sample average and variance are given by

$$\mathbb{E}[\bar{X}_\pi] = \bar{X}, \quad \mathbb{E}[\|\bar{X}_\pi - \bar{X}\|^2] = \frac{n-k}{k(n-1)}\sigma^2$$

Proof of Proposition

Fixing any i such that $1 \leq i \leq n-1$, we have $i(n-i) \leq \frac{n^2}{4} \leq \frac{n(n-1)}{2}$ and using smoothness and Lemma leads to

$$\begin{aligned}
\mathbb{E}[D_{f_{\pi_i}}(x_\star^i, x_\star)] \leq \frac{L}{2}\mathbb{E}[\|x_\star^i - x_\star\|^2] &= \frac{L}{2}\mathbb{E}\left[\|\sum_{k=0}^{n-1}\gamma \nabla f_{\pi_k}(x_\star)\|^2\right] \\
&= \frac{\gamma^2 L i(n-i)}{2(n-1)}\sigma_\star^2 \\
&\leq \frac{\gamma^2 L n}{4}\sigma_\star^2
\end{aligned}$$

To obtain the upper bound, it remains to take the maximum with respect to i on both sides and divide by γ.

Proof of Proposition

To prove the lower bound, we use strong convexity and the fact that $\max_i i(n-i) \geq \frac{n(n-1)}{4}$ holds for any integer n.

$$\max_i \mathbb{E}[D_{f_{\pi_i}}(x_\star^i, x_\star)] \geq \max_i \frac{\mu}{2}\mathbb{E}[\|x_\star^i - x_\star\|^2] = \max_i \frac{\gamma^2 \mu i(n-i)}{2(n-1)}\sigma_\star^2 \geq \frac{\gamma^2 \mu n}{8}\sigma_\star^2$$

$\blacksquare$

Main Theorem 1

Theorem

Suppose that the functions $f_1, f_2, \cdots, f_n$ are μ-strongly convex and that assumptions hold. Then for RR run with a constant stepsize $\gamma \leq \frac{1}{L}$, the iterates generated by algorithm satisfy

$$\mathbb{E}[\|x_T - x_\star\|^2] \leq (1-\gamma\mu)^{nT}\|x_0 - x^\star\|^2 + \frac{2\gamma\sigma^2_{shuffle}}{\mu}$$

Main Theorem 1

Corollary

When we choose stepsize

$$\gamma = \min\{\frac{1}{L}, \frac{2}{\mu n T}\log(\frac{\|x_0 - x_\star\|\mu T\sqrt{n}}{\sqrt{\kappa}\sigma_\star})\}$$

The final iterate x_T then satisfies

$$\mathbb{E}[\|x_T - x_\star\|^2] = \tilde{O}\left(\exp\left(-\frac{\mu n T}{L}\right)\|x_0 - x_\star\|^2 + \frac{\kappa\sigma_\star^2}{\mu^2 n T^2}\right)$$

where $\tilde{O}$ denotes ignoring absolute constants and polylogarithmic factors and κ is the condition number.

Main Theorem 1

Corollary

Thus, in order to obtain an error (in squared distance to the optimum) less than ϵ, we require that the total number of iterations nT satisfies

$$nT = \Omega(\kappa + \frac{\sqrt{kn}\sigma_\star}{\mu\sqrt{\epsilon}})$$

Comparison with SGD

Several works (e.g. [Sti19]) have shown that for any $\gamma \leq \frac{1}{2L}$ the iterates of SGD satisfy

$$\mathbb{E}[\|x_{nT}^{SGD} - x_\star\|^2] \leq (1-\gamma\mu)^{nT}\|x_0 - x_\star\|^2 + \frac{2\gamma\sigma_\star^2}{\mu}$$

Which variance is smaller : $\sigma_{Shuffle}^2$ or $\sigma_\star^2$.
According to the proposition, it depends on both n and stepsize. Once the step size is sufficiently small, $\sigma_{Shuffle}^2$ becomes smaller than $\sigma_\star^2$, but this might not be practical. Similarly, if we partition n functions into $\frac{n}{b}$ groups, *i.e.*, use minibatch of size b, then $\sigma_\star^2$ decreases as $O(\frac{1}{b})$ and $\sigma_{Shuffle}^2$ as $O(\frac{1}{b^2})$, so RR can become faster even without decreasing step size.

Comparison with related works

Number of individual gradient evaluations needed by RR to reach an ϵ-accurate solution

- ([NJN19]) : $\kappa^2 n + \frac{\kappa\sqrt{n}G}{\mu\sqrt{\epsilon}}$
- ([Yin+18]) : $\kappa^2 n + \frac{\kappa n \sigma_\star}{\mu\sqrt{\epsilon}}$
- $\kappa + \frac{\sqrt{\kappa n}\sigma_\star}{\mu\sqrt{\epsilon}}$

The first work requires the Lipshitz function and bounded variance.
Second and this work doesn't require.

Proof of theorem

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

Then, we will use the following decomposition known as three-point identity.

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

Now, we bound each of the three Bregman divergence terms.

Proof of theorem

By μ-strong convexity of f_i,

$$\frac{\mu}{2}\|x_t^i - x_\star^i\|^2 \leq D_{f_{\pi_i}}(x_\star^i, x_t^i)$$

The second term can be bounded via

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

The expectation of the third term is trivially bounded as follows:

$$\mathbb{E}[D_{f_{\pi_i}}(x_\star^i, x_\star)] \leq \max_{i=1,\cdots,n-1}[\mathbb{E}[D_{f_{\pi_i}}(x_\star^i, x_\star)]] = \gamma\sigma_{Shuffle}^2$$

Plugging these three bounds back into inequality.

Proof of theorem

$$\mathbb{E}[\|x_t^{i+1} - x_\star^{i+1}\|^2] \leq \mathbb{E}[(1-\gamma\mu)\|x_t^i - x_\star^i\|^2 - 2\gamma(1-\gamma L)D_{f_{\pi_i}}(x_t^i, x_\star) + 2\gamma^2\sigma^2_{Shuffle}]$$
$$\leq (1-\gamma\mu)\mathbb{E}[\|x_t^i - x_\star^i\|^2] + 2\gamma^2\sigma^2_{Shuffle}$$

Since $x_{t+1} - x_\star = x_t^n - x_\star^n$ and $x_t - x_\star = x_t^0 - x_\star^0$, we can unroll the recursion, obtaining the epoch level recursion

$$\mathbb{E}[\|x_{t+1} - x_\star\|^2] \leq (1-\gamma\mu)^n \mathbb{E}[\|x_t - x_\star\|^2] + 2\gamma^2\sigma^2_{Shuffle}\left(\sum_{i=0}^{n-1}(1-\gamma\mu)^i\right)$$

Unrolling this recursion across T epochs, we obtain

$$\mathbb{E}[\|x_T - x_\star\|^2] \leq (1-\gamma\mu)^{nT}\|x_0 - x_\star\|^2 + 2\gamma^2\sigma^2_{Shuffle}\left(\sum_{i=0}^{n-1}(1-\gamma\mu)^i\right)\left(\sum_{j=0}^{T-1}(1-\gamma\mu)^{nj}\right)$$

Proof of theorem

$$\mathbb{E}[\|x_T - x_\star\|^2] \le (1-\gamma\mu)^{nT}\|x_0 - x_\star\|^2 + 2\gamma^2\sigma^2_{Shuffle}\left(\sum_{i=0}^{n-1}(1-\gamma\mu)^i\right)\left(\sum_{j=0}^{T-1}(1-\gamma\mu)^{nj}\right)$$

The product of two sums can be bounded as follows:

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

Plugging this bound back into inequality, we finally obtain the theorem. ■

Main Theorem 2

Theorem

Let functions $f_1, f_2 \cdots, f_n$ be convex. Suppose that assumptions hold. Then for RR runs with a step size $\gamma \leq \frac{1}{\sqrt{2}Ln}$, the average iterate $\hat{x}_T = \frac{1}{T}\sum_{j=1}^{T} x_j$ satisfies

$$\mathbb{E}[f(\hat{x}_T)] - f(x_\star)] \leq \frac{\|x_0 - x_\star\|^2}{2\gamma nT} + \frac{\gamma^2 Ln\sigma_\star^2}{4}$$

Main Theorem 2

Corollary

Under the same conditions as theorem 2, choose the step size

$$\gamma = \min\{\frac{1}{\sqrt{2}Ln}, \left(\frac{\|x_0 - x_\star\|^2}{Ln^2 T\sigma_\star^2}\right)^{\frac{1}{3}}\}$$

Then

$$\mathbb{E}[f(\hat{x}_T) - f(x_\star)] \leq \frac{L\|x_0 - x_\star\|^2}{\sqrt{2}T} + \frac{3L^{\frac{1}{3}}\|x_0 - x_\star\|^{\frac{4}{3}}\sigma_\star^{\frac{2}{3}}}{4n^{\frac{1}{3}}T^{\frac{2}{3}}}$$

Main Theorem 2

Corollary

We can guarantee that $\mathbb{E}[f(\hat{x}_T) - f(x_\star)] \leq \epsilon^2$ *provided*

$$nT \geq \frac{2\|x_0 - x_\star\|^2\sqrt{Ln}}{\epsilon^2}\max\{\sqrt{Ln}, \frac{\sigma_\star}{\epsilon}\}$$

Comparison with SGD

Compare to convergence upper bound of $O\left(\frac{L\|x_0-x_\star\|^2}{nT}+\frac{\sigma_\star\|x_0-x_\star\|}{\sqrt{nT}}\right)$ for SGD (e.g. [Sti19]).

Comparing upper bounds, RR beats SGD when the number of epochs satisfies $T\geq\frac{L^2\|x_0-x_\star\|^2 n}{\sigma_\star^2}$.

[NJN19] Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. “Sgd without replacement: Sharper rates for general smooth convex functions”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4703–4711.

[Sti19] Sebastian U Stich. “Unified optimal analysis of the (stochastic) gradient method”. In: *arXiv preprint arXiv:1907.04232* (2019).

[Yin+18] Bicheng Ying et al. “Stochastic learning under random reshuffling with constant step-sizes”. In: *IEEE Transactions on Signal Processing* 67.2 (2018), pp. 474–489.