Random Reshuffling: Simple Analysis with Vast Improvements

Konstantin Mishchenko [1] Ahmed Khaled [2] Peter Richtárik [1]

[1] KAUST [2] Cairo University

The problem

We consider the finite-sum minimization problem

$$\text{find } x_* = \arg\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right\}, \qquad (1)$$

where each f_i is L-smooth function (potentially non-convex).
Our goal is to explain convergence of stochastic algorithms.
Assumption 1: For every i, f_i is L-smooth, that is, for all $x, y \in \mathbb{R}^d$ we have

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x - y\|.$$

Motivation

- Huge dimension $d \implies$ first-order methods are more efficient;
- Large dataset size $n \implies$ stochastic updates are necessary;
- Fast convergence to approximate solution is preferred $\implies$ large stepsizes are paramount.

Algorithms for Problem (1)

Algorithm 1 SGD

Input: $x_0 \in \mathbb{R}^d$, $\gamma > 0$

```
for t = 0, 1, ... do
    Sample i uniformly from {1, ..., n}
    x_{t+1} = x_t − γ∇f_i(x_t)
end for
```

Algorithm 2 IG

Input: $x_0^0 = x_0 \in \mathbb{R}^d$, $\gamma > 0$

```
for t = 0, 1, ... do
    for i = 0, ..., n − 1 do
        x_t^{i+1} = x_t^i − γ∇f_i(x_t^i)
    end for
    x_{t+1}^0 = x_t^n
end for
```

Algorithm 3 RR

Input: $x_0^0 = x_0 \in \mathbb{R}^d$, $\gamma > 0$

```
for t = 0, 1, ... do
    Sample a permutation π_0, ..., π_{n−1} of {1, ..., n}
    for i = 0, ..., n − 1 do
        x_t^{i+1} = x_t^i − γ∇f_{π_i}(x_t^i)
    end for
    x_{t+1}^0 = x_t^n
end for
```

Algorithm 4 SO

Input: $x_0^0 = x_0 \in \mathbb{R}^d$, $\gamma > 0$

```
Sample a permutation π_0, ..., π_{n−1} of {1, ..., n}
for t = 0, 1, ... do
    for i = 0, ..., n − 1 do
        x_t^{i+1} = x_t^i − γ∇f_{π_i}(x_t^i)
    end for
    x_{t+1}^0 = x_t^n
end for
```

SGD

Stochastic Gradient Descent (SGD) is one of the most popular algorithms that samples functions uniformly at each iteration.
Pros: unbiased update, $\mathbb{E}_i[x_{t+1}] = x_t - \gamma\nabla f(x_t)$; easy to analyze.
Cons: does not use the finite-sum structure; access to arbitrary sample is expensive (cache misses).
Rate of convergence:[a] $\mathcal{O}\left(\frac{1}{T}\right)$

IG

Incremental Gradient (IG) is an alternative to SGD that performs cyclic data passes.
Pros: each function gets used exactly once per epoch; fast sequential access to the memory.
Cons: slow if the data are structured/sorted; always slower than gradient descent.
Rate of convergence: $\mathcal{O}\left(\frac{n^2}{T^2}\right)$ (better than SGD when $T \ge n^2$)

RR/SO

Random Reshuffling (RR) and **Shuffle-Once (SO)** improve upon IG by sampling a permutation each epoch (RR) or just shuffling the data once (SO).
Pros: faster rate than that of IG.
Cons: hard to analyze as $\mathbb{E}_\pi[x_t^{i+1}] \neq x_t^i - \gamma\nabla f(x_t^i)$.
Rate of convergence (new!): $\mathcal{O}\left(\frac{n}{T^2}\right)$ (better than SGD when $T \ge n$)

[a] For all methods, the rate is provided in the strongly convex case and in terms of full number of computed stochastic gradients T.

Key contributions

1. Tight rates for RR and SO;
2. First result that allows for $\gamma = \frac{1}{L}$;
3. New insight into convergence within each epoch;
4. Improved estimate of shuffling variance.

New complexities for RR/SO

Let $\sigma_*^2 \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_*)\|^2$ be the variance at the optimum and $\kappa \overset{\text{def}}{=} \frac{L}{\mu}$ (convex f) or $\sigma^2 = \sup_x \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f(x)\|^2$ (non-convex f). New complexities:

- If all $f_1, \ldots, f_n$ are μ-strongly convex: $\mathcal{O}\left(\kappa \log\frac{1}{\varepsilon} + \frac{\sqrt{\kappa n}\sigma_*}{\mu\sqrt{\varepsilon}}\right)$;
- If only $f = \frac{1}{n}\sum_{i=1}^n f_i$ is μ-strongly convex: $\mathcal{O}\left(\kappa n \log\frac{1}{\varepsilon} + \frac{\sqrt{\kappa n}\sigma_*}{\mu\sqrt{\varepsilon}}\right)$;
- If f is convex: $\mathcal{O}\left(\frac{n}{\varepsilon} + \frac{\sqrt{n}\sigma_*}{\varepsilon^{3/2}}\right)$;
- If f is non-convex (RR only): $\mathcal{O}\left(\frac{n}{\varepsilon^2} + \frac{\sqrt{n}\sigma}{\varepsilon^3}\right)$.

New theoretical insights

Definition 1. For any i, we define the Bregman divergence of f_i as

$$D_{f_i}(x, y) \overset{\text{def}}{=} f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y\rangle.$$

f_i is called μ-strongly convex if $D_{f_i}(x, y) \ge \frac{\mu}{2}\|x - y\|^2$ for any $x, y \in \mathbb{R}^d$.
Definition 2. Given a permutation $\pi_0, \ldots, \pi_{n-1}$ and stepsize $\gamma > 0$, we let

$$x_*^i \overset{\text{def}}{=} x_* - \gamma\sum_{j=0}^{i-1} \nabla f_{\pi_j}(x_*).$$

Clearly, by optimality of x_*, we have $x_*^n = x_*$.
Lemma 1. [Key recursion] It holds

$$\|x_t^{i+1} - x_*^{i+1}\|^2 = \|x_t^i - x_*^i\|^2 + \gamma^2\|\nabla f_{\pi_i}(x_t^i) - \nabla f_{\pi_i}(x_*)\|^2 - 2\gamma[D_{f_{\pi_i}}(x_*^i, x_t^i) + D_{f_{\pi_i}}(x_t^i, x_*) - D_{f_{\pi_i}}(x_*^i, x_*)].$$

Lemma 1 is used to obtain the following theorem.
Theorem 1. If $f_1, \ldots, f_n$ are μ-strongly convex and $\gamma \le \frac{1}{L}$, then

$$\mathbb{E}[\|x_t^{i+1} - x_*^{i+1}\|^2] \le (1 - \gamma\mu)\|x_t^i - x_*^i\|^2 + 2\gamma^2\sigma_{\text{Shuffle}}^2,$$

where

$$\sigma_{\text{Shuffle}}^2 \overset{\text{def}}{=} \max_{i=1,\ldots,n}\left[\frac{1}{\gamma}\mathbb{E}[D_{f_{\pi_i}}(x_*^i, x_*)]\right].$$

To compare this to convergence of SGD, we prove the following upper and lower bounds.
Theorem 2. It holds

$$\frac{\gamma\mu n}{8}\sigma_*^2 \le \sigma_{\text{Shuffle}}^2 \le \frac{\gamma L n}{4}\sigma_*^2.$$

Experiments

We run experiments on ℓ_2 regularized logistic regression problem and set ℓ_2 penalty to be $\frac{L}{\sqrt{N}}$, where N is the dataset size.

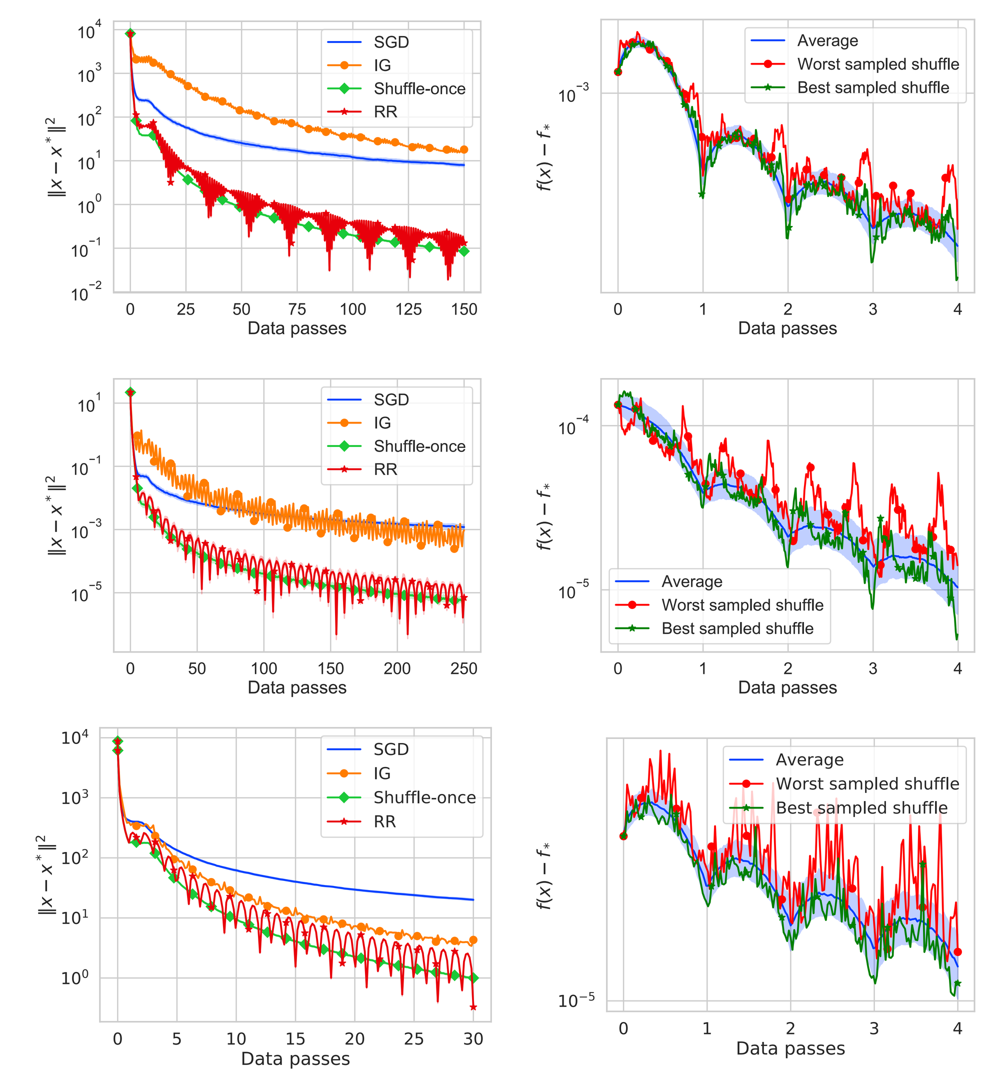

Figure 1: Top: **real-sim** dataset ($N = 72{,}309$; $d = 20{,}958$), middle row: **w8a** dataset ($N = 49{,}749$; $d = 300$), bottom: **RCV1** dataset ($N = 804{,}414$; $d = 47{,}236$). Left: convergence of $\|x_t^i - x_*\|^2$, right: convergence of SO with different permutations.

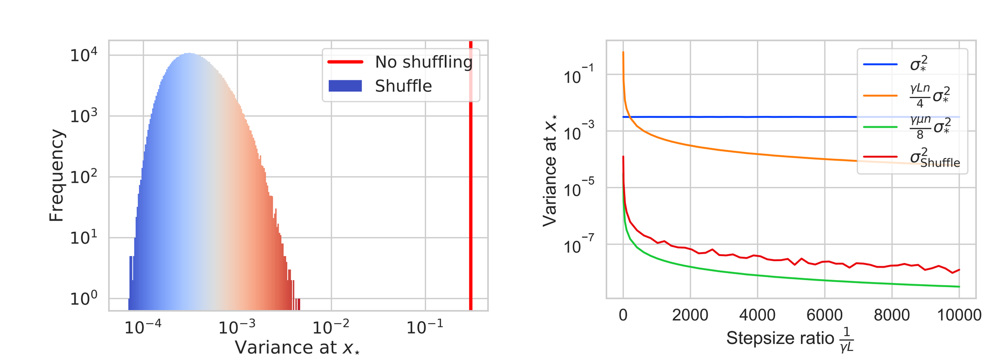

Figure 2: Left: histogram of values of $\sigma_{\text{Shuffle}}^2$ evaluated on 500,000 sampled permutation. Right: values of $\sigma_{\text{Shuffle}}^2$ for different values of γ. Both plots are computed for **w8a** dataset.