High Probability Guarantees for Random Reshuffling*

Hengxu Yu
The Chinese University of Hong Kong, Shenzhen
`hengxuyu@link.cuhk.edu.cn`

Xiao Li
The Chinese University of Hong Kong, Shenzhen
`lixiao@cuhk.edu.cn`

Abstract

We study the probabilistic behaviors of stochastic gradient descent with *random reshuffling* (RR) on nonconvex problems. We prove that the same complexity (except for a logrithmic term) as that of in expectation case also holds with high probability, which characterizes the performance of RR for a single run instead of averaging infinitly many realizations. Our analysis does not impose any additional assumptions on the stochastic gradient errors, which admits heavy tails. This is in contrast to high probabiltiy analyses of SGD that rely on sub-Gaussian stochastic gradient errors or tricks like clipping, momentum, etc. Furthermore, leveraging the established high probability error bounds, we propose a simple stopping criterion for RR that introduces few computational costs. We prove that the function value strictly decreases with high probability before the stopping criterion is triggered, ensuring that the criterion will indeed be activated. Finally, a "last iterate" result is built for the iteration returned with this stopping criterion. We believe that our new developments for RR serve as a stepping stone towards enabling more refined analyses for characterizing its performance.

1 Introduction

In this work, we focus on the finite-sum optimization problem defined as follows:

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

where each component function f_i is continuously differentiable but not necessarily convex. Such problems arise in various engineering fields, including machine learning and signal processing [1]. In many modern applications, the number of component functions n is so large that applying stochastic optimization methods is necessary. In this work, we will study the stochastic gradient descent with *random reshuffling* (RR) for solving problem (1), which is a stochastic variant of the classical gradient descent method.

At iteration t, RR first samples a permutation π_t of $\{1, \ldots, n\}$ uniformly at random. Then, it starts with an initial inner iterate $x_t^0 = x_t$ and updates x_t to x_{t+1} by consecutively applying the gradient descent type steps as

$$x_t^i = x_t^{i-1} - \alpha \nabla f_{\pi_t^i}(x_t^{i-1}) \tag{2}$$

*This is an extended abstract of `https://arxiv.org/abs/2311.11841`.

for $i = 1, \ldots, n$, resulting in $x_t^n = x_{t+1}$. We display the pseudocode of RR in Algorithm 1.

Despite its widespread practical usage, the theoretical understanding of RR has been mainly limited to complexity bounds in expectation and almost sure asymptotic convergence results. In this study, our objective is to provide a complexity guarantee for RR with high probability, and further based on it develop a stopping criterion to obtain the "last iterate" result.

Algorithm 1: RR: Random Reshuffling

```
Input: Initial point $x_0 \in \mathbb{R}^d$ and number of epoches $T$;
for $t = 0, \ldots, T-1$ do
    Sample a permutation $\pi_t = \{\pi_t^1, \ldots, \pi_t^n\}$ of $\{1, \ldots, n\}$;
    Update the step size $\alpha_t$;
    Set $x_t^0 = x_t$;
    for $i = 1, \ldots, n$ do
        $x_t^i = x_t^{i-1} - \alpha \nabla f_{\pi_t^i}\left(x_t^{i-1}\right)$ ;                /* update */
    end
    Set $x_{t+1} = x_t^n$;
end
```

1.1 Related Works and Motivations

Finite-time complexity bounds in expectation. Various works have focused on deriving complexity bounds for RR, see, e.g., [12, 3, 7, 8, 11, 10, 13]. For instance, the work [7] establishes an $\mathcal{O}(\sqrt{n}/\varepsilon)$ iteration complexity under the assumptions that the objective function f is strongly convex and each f_i has Lipschitz continuous gradient. When f is nonconvex and each f_i is Lipschitz smooth, it was shown in [7, 8] that RR has an iteration complexity of $\mathcal{O}(\sqrt{n}\varepsilon^{-3})$. It is worth mentioning that all these complexity results for RR hold in the sense of expectation, characterizing the performance of the algorithm by averaging infinitely many runs.

Asymptotic convergence. For strongly convex functions with component-wise Lipschitz continuous Hessian, Gürbüzbalaban et al. demonstrated that iterate converges to the optimal solution at a rate of $\mathcal{O}(1/t^2)$ with high probability [2]. In the smooth nonconvex scenario, the asymptotic convergence of the gradient norm was derived using the unified convergence framework established in [5]. Furthermore, [6] provided rate results for this convergence. However, while these results offer significant theoretical assurances, they have some limitations in explaining the algorithm's practical performances, as they primarily offer insights into the long-term behavior of the algorithm when $t \to \infty$.

Motivations. The current complexity results in expectation may not effectively explain the performance of a single run of RR, and the almost sure asymptotic convergence results only characterize long-term behaviors. This observation motivates us to derive high probabiltiy finite-time guarantees for RR, without assuming strict restrictions on the stochastic gradient errors. Moreover, complexity results for nonconvex RR apply to the minimal expected gradient norm, i.e., $\min_{0 \le t \le T} \mathbb{E}\|\nabla f(x_t)\|^2$. While this measure provides certain insights, it is not computable furthermore used to determine which iterate to return. For nonconvex SGD-type methods, several stopping criteria have been proposed, see, e.g., [14, 9]. They either discuss statistical stationarity or propose asymptotic gradient-based stopping measure. However, as far as we know, a stopping criterion for RR remains unexplored, prompting our investigation.

1.2 Main contributions

Throughout this paper, we only make the assumption that each component function f_i is L-smooth and lower bounded, as specified in Assumption 2.1. Our main contributions are summarized below.

High probability sample complexity. We first enhance the analysis of RR by demonstrating that, with probability at least $1-\delta$, it can identify an ε-stationary point using at most $\tilde{\mathcal{O}}\left(\max\left\{n\varepsilon^{-2}, \sqrt{n}\varepsilon^{-3}\right\}\right)$ stochastic gradient evaluations, where $\tilde{\mathcal{O}}$ conceals a logarithmic term. It is worth mentioning that our high probability sample complexity matches the existing complexity in expectation [7, 8] up to a logarithmic term. In addition, our analysis does not impose any additional assumptions regarding the stochastic gradient errors. Technically speaking, the classical analysis of RR decomposes the stochastic gradient error into deterministically bounded terms and a term determined by the randomness of RR. Our main step lies in identifying that the latter term possesses a concentration property by applying the Bernstein's inequality over the random reshuffling sampling scheme of RR. This allows us to establish a standard approximate descent property that holds with high probability rather than in expectation, leading to the aforementioned complexity result.

Stopping Criterion. In the second part, we further leverage the concentration property of the stochastic gradient errors to design a new random reshuffling method with stopping criterion (RR-sc). This criterion terminates the algorithm and returns the starting inner iterate when the accumulated stochastic gradients in one epoch satisfy a preset tolerance $\gamma\varepsilon$, where $\gamma > 0$ is some constant. RR-sc introduces few additional computation loads compared to vanilla RR. We prove that the stopping criterion must be activated within $\tilde{\mathcal{O}}\left(\sqrt{n}\varepsilon^{-3}\right)$ stochastic gradient evaluations with probability at least $1-\delta$, which matches our previous complexity bound. The key step in establishing this result involve showing that RR-sc exhibits the so-called sufficient decrease property before the stopping criterion is triggered, closely resembling the deterministic gradient descent method. Furthermore, we establish the "last iterate" result, indicating that once the algorithm is terminated, the returned iterate x_τ satisfies $\|\nabla f(x_\tau)\| \leq \Theta(\varepsilon)$.

2 High Probability First-order Sample Complexity Guarantees

In this section, we establish high probability sample complexity guarantees for RR. We impose the following standard smoothness assumption on the component functions throughout this section.

Assumption 2.1. *For all $i \in [n]$, f_i in (1) is bounded from below by $\bar{f}_i$ and its gradient ∇f_i is Lipschitz continuous with parameter $\mathsf{L} \geq 0$.*

Let $\bar{f}$ be a lower bound of f in (1). It was established in [4, Proposition 3] that the following variance-type bound is true once Assumption 2.1 holds:

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f(x)\|^2 \leq \mathsf{A}(f(x) - \bar{f}) + \mathsf{B}, \tag{3}$$

where $\mathsf{A} = 2\mathsf{L} \geq 0$ and $\mathsf{B} = \frac{\mathsf{A}}{n}\sum_{i=1}^{n}(\bar{f} - \bar{f}_i) \geq 0$. The bound (3) plays a crucial role in our later analysis.

Lemma 2.2. *Suppose that Assumption 2.1 is valid and the step size $\alpha_t = \alpha, \forall 0 \leq t \leq T-1$ satisfies*

$$\alpha \leq \min\left\{\frac{1}{4n\mathsf{L}}, \frac{1}{(\mathsf{C}_1 n^2 T)^{1/3}}\right\}. \tag{4}$$

Then, with probability at least $1-\delta$, it holds that for all $0 \leq t \leq T-1$,

$$f(x_{t+1}) \leq f(x_t) - \frac{\alpha_t n}{8}\|\nabla f(x_t)\|^2 - \frac{\alpha_t n}{2}\|g_t\|^2 + \alpha_t^3 n^2 \mathsf{G}. \tag{5}$$

Here, $\mathsf{C}_1 = 32\mathsf{L}^2\mathsf{A}\log^2\left(\frac{8nT}{\delta}\right) \geq 0$ and $\mathsf{G} = \mathsf{C}_1\mathsf{F} + \mathsf{C}_2 \geq 0$ with $\mathsf{F} = 3(f(x_0) - \bar{f}) + 3\mathsf{B}/\mathsf{A} \geq 0$ and $\mathsf{C}_2 = 32\mathsf{L}^2\mathsf{B}\log^2\left(\frac{8nT}{\delta}\right) \geq 0$.

With this high probability approximate descent, we are now ready to establish the following complexity (without taking expectation) of RR for finding a stationary point of problem (1).

Theorem 2.3. *Under the setting of Lemma 2.2, with probability at least $1-\delta$, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|^2 \leq \mathcal{O}\left(\max\left\{\frac{1}{T}, \frac{\log^{2/3}(8nT/\delta)}{n^{1/3}T^{2/3}}\right\}\right), \tag{6}$$

where the "$\mathcal{O}$" hides constants that are independent of n, T, and δ. Consequently, to achieve $\sum_{t=0}^{T-1}\|\nabla f(x_t)\|^2/T \leq \varepsilon^2$, RR needs at most

$$nT = \tilde{\mathcal{O}}\left(\max\left\{n\varepsilon^{-2}, \sqrt{n}\varepsilon^{-3}\right\}\right) \tag{7}$$

stochastic gradient evaluations, where the "$\tilde{\mathcal{O}}$" hides an additional $\log(\sqrt{n}\varepsilon^{-3}/\delta)$ in the term $\sqrt{n}\varepsilon^{-3}$.

Algorithm 2: RR-sc: RR with Stopping Criterion

```
Input: Initial point $x_0 \in \mathbb{R}^d$, constant tolerance $\gamma$, target accuracy $\varepsilon$;
while true do
    Set $g_t = 0$;
    Update the step size $\alpha_t$;
    Sample a permutation $\pi_t = \{\pi_t^1, \ldots, \pi_t^n\}$ of $\{1, \ldots, n\}$;
    Set $x_t^0 = x_t$;
    for $i \leftarrow 1$ to $n$ do
        $x_t^i = x_t^{i-1} - \alpha_t \nabla f_{\pi_t^i}\left(x_t^{i-1}\right)$ ;                    /* update */
        $g_t = g_t + \nabla f_{\pi_t^i}\left(x_t^{i-1}\right)/n$ ;       /* accumulate stochastic gradients */
    end
    if $\|g_t\| \leq \gamma\varepsilon$ then                            /* check stopping criterion */
        Set $\tau = t$;
        return $x_\tau$;
    else
        Set $x_{t+1} = x_t^n$;
    end
    Set $t = t + 1$;
end
```

Remark

- The obtained high probabiltiy complexity bound is the same as that of in expectation case, with the only exception of dependence on a logrithmic term. We note that this logarithmic dependence is not intrinsic, but rather a result of our proof technique, which applies the union bound across epochs. Eliminating this reliance on the union-bound argument is one of the directions we plan to explore.
- Our proof techniques can be similarly applied to obtain high probabiltiy guarantees for RR for strongly convex and convex problems.

3 Stopping Criterion

Stopping criterion is often an important ingredient in algorithm design such as the gradient norm in deterministic gradient descent. However, computing the full gradient function for monitoring stationarity is not feasible in RR. Thus, it requires constructing certain new estimated stopping criterion for RR, which is the main topic of this section.

3.1 Random Reshuffling with Stopping Criterion

We observe that the accumulation of the stochastic gradients g_t plays almost the same role for descent as that of the true gradient. This motivates us to monitor g_t in the algorithmic procedure as a stopping criterion. It is worth noting that g_t is computable and increases almost no computational load compared to the base RR.

We design RR with stopping criterion (hereafter "RR-sc") in Algorithm 2. In this algorithm, we compute the accumulation of the stochastic gradients used in the update and stack it in g_t. After each epoch, we check the *stopping criterion*

$$\|g_t\| \leq \gamma\varepsilon \tag{8}$$

for some target accuracy level ε and some constant tolerance γ. Once this criterion is triggered, we stop the algorithm and return *the last* iterate x_τ that achieves it. In this subsection, we show that the stopping criterion must be triggered, and hence RR-sc is guaranteed to stop after a certain number of iterations. Now we define the maximum possible number of iterations T_1 through

$$nT_1 = 6\mathsf{F}(\gamma\varepsilon)^{-2} \max\left\{n\mathsf{L}, 2\sqrt{n\mathsf{A}\mathsf{F}}\mathsf{L}(\gamma\varepsilon)^{-1}\log\left(\frac{8nT_1}{\delta}\right)\right\} \sim \tilde{\mathcal{O}}(\max\{\sqrt{n}\varepsilon^{-3}, n\varepsilon^{-2}\}). \tag{9}$$

Proposition 3.1. *With probability at least* $1 - \delta$*,* RR-sc *terminates within* T_1 *number of iterations, i.e., in Algorithm 2 we have* $\tau \leq T_1$*.*

3.2 The Last Iterate Result

The fact that RR-sc must stop within T_1 number of iterations does not necessarily imply that *the underlying stopping criterion* $\|\nabla f(x_\tau)\| \leq \varepsilon$ holds true. We further show when the estimated stopping criterion $\|g_t\| \leq \gamma\epsilon$ triggers, we also have $\|\nabla f(x_\tau)\| \leq \Theta(\varepsilon)$, as stated in the following theorem.

Theorem 3.1. *With probability at least* $1 - \delta$*,* RR-sc *terminates at iteration* τ *satisfying* $\tau \leq T_1$*. Furthermore, we have* $\|\nabla f(x_\tau)\| \leq \Theta(\varepsilon)$*.*

We comment here on the number of iterations needed to stop the algorithm. The stopping criterion is indeed not a bad estimation. Namely, we will not have the case where the underlying criterion $\|\nabla f(x_t)\| = \Theta(\varepsilon)$ is already satisfied, but the stopping criterion is triggered much later. We can see this conclusion by repeating our arguments when approximating $\|\nabla f(x_t)\|$ with $\|g_t\|$ to similarly bound $\|g_t\|$ by $\|\nabla f(x_t)\|$.

References

[1] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[2] Mert Gürbüzbalaban, Asu Ozdaglar, and PA Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1-2):49–84, 2021.

[3] Jeff Haochen and Suvrit Sra. Random shuffling beats sgd after finite epochs. In *International Conference on Machine Learning*, pages 2624–2633, 2019.

[4] Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *Transactions on Machine Learning Research*, 2022.

[5] Xiao Li and Andre Milzarek. A unified convergence theorem for stochastic optimization methods. In *Advances in Neural Information Processing Systems*, volume 35, pages 33107–33119, 2022.

[6] Xiao Li, Andre Milzarek, and Junwen Qiu. Convergence of random reshuffling under the kurdyka-łojasiewicz inequality. *SIAM Journal on Optimization*, 33(2):1092–1120, 2023.

[7] Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. In *Advances in Neural Information Processing Systems*, volume 33, pages 17309–17320, 2020.

[8] Lam M Nguyen, Quoc Tran-Dinh, Dzung T Phan, Phuong Ha Nguyen, and Marten Van Dijk. A unified convergence analysis for shuffling-type gradient methods. *The Journal of Machine Learning Research*, 22(1):9397–9440, 2021.

[9] Vivak Patel. Stopping criteria for, and strong convergence of, stochastic gradient descent on bottou-curtis-nocedal functions. *Mathematical Programming*, 195(1-2):693–734, 2022.

[10] Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of sgd without replacement. *International Conference On Machine Learning*, 2020.

[11] Itay Safran and Ohad Shamir. How good is sgd with random shuffling? In *Conference on Learning Theory*, volume 125, pages 3250–3284, 2020.

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

[13] Trang H Tran, Katya Scheinberg, and Lam M Nguyen. Nesterov accelerated shuffling gradient method for convex optimization. In *International Conference on Machine Learning*, volume 162, pages 21703–21732, 2022.

[14] George Yin. A stopping rule for the robbins-monro method. *Journal of Optimization Theory and Applications*, 67(1):151–173, 1990.