ShuffleSGD: noise of machine learning type

Aishwarya Hiwrale [1] Javed Akhtar [1] Tom Jacobs [1, 2]

[1]University of Saarland [2]CISPA Helmholtz Center

Introduction and Background

First-order sub-gradient methods are essential in machine learning.
SGD variants have distinct convergence rates. Faster convergence rates are beneficial for computational savings.
This study examines these rates for both PL inequality and convex problems.
The problems at hand are linear and logistic regression tasks.

SGD (Stochastic Gradient Descent): A standard method with random sampling at each step.
Single Shuffle (SS): A fixed order is chosen randomly and iterated through completely.
Random Reshuffle (RR): Similar to SS but the order changes each iteration.

Convergence rate

The convergence rate for L-smooth objective functions is derived as:

$$O\left(\frac{F_0}{\gamma T} + L^2\gamma^2\sigma_\tau^2\right)$$

- F_0: Initial function value.
- γ: Learning rate.
- T: Number of iterations.
- σ_τ^2: Noise strength parameter.

Noise Assumptions: For SGD, the noise is defined as:

$$\sigma_{\text{SGD}}^2 = \sup_{w\in\mathbb{R}^d} \mathbb{E}_i||\nabla f_i(w) - \nabla f(w)||^2$$

Algorithm	σ_τ^2 upperbound
SGD	$\tau\sigma_{\text{SGD}}^2$
SS	$\min\{\tau, n\}n\sigma_{\text{SGD}}^2$
RR	$4\min\{\tau, n\}n\sigma_{\text{SGD}}^2$

Table 1. Upperbound on σ_τ^2.

Noise of Machine Learning Type

In [2] it is proposed that regression tasks in ML can satisfy the noise constraint:

$$\mathbb{E}_i||\nabla f_i(w) - \nabla f(w)||^2 \leq M_{\text{SGD}} f(w) \qquad \forall w \in \mathbb{R}^d \tag{1}$$

which we refer to as noise of machine learning type.

Main insights

- In regression tasks, noise decays and may satisfy the Polyak-Lojasiewicz (PL) inequality locally, affecting convergence differently than L-smooth loss functions.
- The noise strength parameter σ_τ in SS and RR methods manages index correlation, enhancing convergence compared to plain SGD.
- Experiments with linear and logistic regression validate these insights, highlighting differences in noise structure between classification and regression tasks.
- We propose a sequential correlation analogue to capture this alternative noise structure.

Linear Regression

- **Objective**: Minimize the loss function $f_i(w) = \frac{1}{2}(w^T x_i - y_i)^2$.
- **Properties**: Each f_i is L_i-smooth with $L_i = \lambda_{\max}(x_i x_i^T)$, satisfying the PL-inequality with $\Lambda = \lambda_{\min}(x_i x_i^T)$.
- **Convergence Rate Analysis:** The convergence rate for SS and RR variants is improved due to reduced sequence correlation in the gradient updates. And emmpirical results confirm faster convergence compared to plain SGD, leveraging the structured noise assumption.

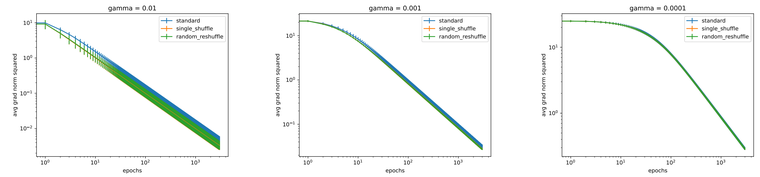

Figure 1. Linear regression with sparse ground truth

Logistic Regression

- **Objective**: Minimize the loss function $f_i(w) = \log(1 + e^{-y_i w^T x_i})$.
- **Properties**: Each f_i is L_i-smooth with $L_i = \frac{1}{4}\lambda_{\max}(x_i x_i^T)$.
- **Convergence Rate Analysis:** Investigate convergence rates across different datasets (e.g., Australian credit approval, Breast Cancer). And evaluate convergence based on Lipschitz constants and noise characteristics, showing improvements with SS and RR over SGD.

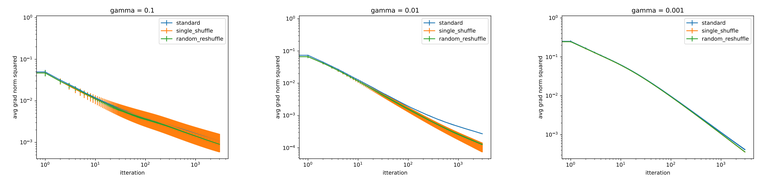

Figure 2. Logistic regression on the *Breast Cancer* dataset.

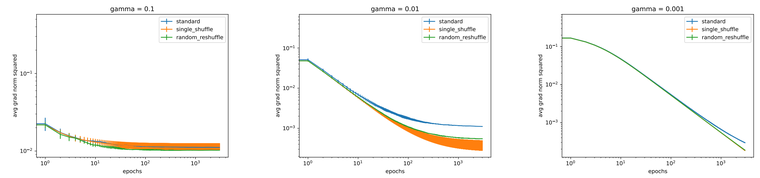

Figure 3. Logistic regression on the *Australian* dataset.

Analysis of Results

Regression Type	Linear	Logistic
Data Set(s)	Synthetic data with sparse ground truth	Australian credit approval and Breast Cancer
Lipschitz Constant(s)	111.1	3.1 (Australian), 5.5 (Breast Cancer)
Convergence	SS and RR outperform plain SGD	SS and RR show faster convergence compared to plain SGD

Conjecture

We propose a generalization of sequence correlation tailored for machine learning noise:

$$\max_{k=0,\dots,\lfloor\frac{T}{\tau}\rfloor, j=0,\dots,\tau-1} \mathbb{E}\left[\phi_{k\tau+j}(x) \mid i_0, \dots, i_{k\tau-1}\right] \leq M_\tau f(x), \tag{2}$$

where $\phi_{k\tau+j}(x)$ is defined as

$$\phi_{k\tau+j}(x) = \left\| \sum_{t=k\tau}^{\min\{k\tau+j, T\}} \nabla f_{i_t}(x) - \nabla f(x) \right\|^2.$$

Difference from Original Sequence Correlation

In contrast to the original sequence correlation, equation (2) substitutes the right-hand side with σ_τ^2.

Discussion and Summary

- **Performance**: SS and RR outperform SGD in convergence rate for both linear and logistic regression.
- **Generalization**: Practical performance often exceeds theoretical bounds due to conservative Lipschitz constant estimates.
- **Convergence Analysis**: Empirical results validate the theoretical convergence rates for SS and RR, showing significant improvement over plain SGD.
- **Noise Structure**: Noise in regression tasks aligns with decaying noise conditions.
- **Future Work**: Further investigate noise conditions and their impact, and explore other machine learning applications.

References

[1] Anastasia Koloskova, Nikita Doikov, Sebastian U. Stich, and Martin Jaggi.
On convergence of incremental gradient for non-convex smooth functions, 2024.

[2] Stephan Wojtowytsch.
Stochastic gradient descent with noise of machine learning type. part i: Discrete time analysis, 2021.

aihi00001@uni-saarland.de & jaak00001@stud.uni-saarland.de & tom.jacobs@cispa.de