Random-reshuffled SARAH does not need full gradient computations

Aleksandr Beznosikov[1,2,3], Martin Takáč[3]

[1] MIPT, Russia [2] MBZUAI, UAE

Problem

Minimizing a finite-sum problem of the form

$$\min_{w \in \mathbb{R}^d} \left\{ P(w) := \frac{1}{n} \Sigma_{i=1}^n f_i(w) \right\}, \qquad (1)$$

Problems of this form are very common in e.g., supervised learning . Let a training dataset consists of n pairs, i.e., $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ is a feature vector for a datapoint i and y_i is the corresponding label. Then for example, the least squares regression problem corresponds to (1) with $f_i(w) = \frac{1}{2}(x_i^T w - y_i)^2$. If $y_i \in \{-1, 1\}$ would indicate a class, then a logistic regression is obtained by choosing $f_i(w) = \log(1 + \exp(-y_i x_i^T w))$.

- **SARAH.** SARAH [2] is a classical variance reduction method with recursive update of "gradient":

$$v_t = \nabla f_i(w_t) - \nabla f_i(w_{t-1}) + v_{t-1}, \quad w_{t+1} = w_t - \eta_t v_t.$$

- **Random Reshuffle.** Not to choose functions f_i randomly with replacement, but make a data permutation/shuffling and choose the f_is in a cyclic fashion. In [1] a few basic shuffling are discussed, including **Random Reshuffling (RR)**,**Shuffle-Once (SO)**, **Incremental Gradient (IG)**.

Assumptions

For problem (1) the following hold:

1. Each $f_i : \mathcal{R}^d \to \mathcal{R}$ is convex and twice differentiable, with L-smooth gradient:
$$\|\nabla f_i(w_1) - \nabla f_i(w_2)\| \le L\|w_1 - w_2\|,$$
for all $w_1, w_2 \in \mathcal{R}^d$;
2. $P(w)$ is μ-strongly convex function with minimizer x^* and optimal value P^*;
3. Each f_i is δ-similar with P, i.e. for all $w \in \mathcal{R}^d$ it holds that
$$\|\nabla^2 f_i(w) - \nabla^2 P(w)\| \le \delta/2.$$

The last assumption means the similarity of $\{f_i\}$. For example, this effect is observed when the data is divided uniformly across batches f_i, then with a high probability we have $\delta \sim \frac{L}{\sqrt{b}}$, where b is a size of local batch f_i (number of data points in f_i)

Method

Algorithm : `Shuffled-SARAH`

```
Input: 0 < η step-size
choose w⁻ ∈ ℝ^d
w = w⁻
v_0 = 0 ∈ ℝ^d
ṽ = &(v_0)      // ṽ will point to v_0
Δ = 0 ∈ ℝ^d
for s = 0, 1, 2, ... do
    define w_s := w
    w⁻ = w
    w = w − η v_s
    obtain permutation π_s = (π_s^1, ..., π_s^n) of [n] by some rule
    for i = 1, 2, ..., n do
        ṽ = (i−1)/i ṽ + 1/i ∇f_{π_s^i}(w)
        Δ = Δ + ∇f_{π_s^i}(w) − ∇f_{π_s^i}(w⁻)
        w⁻ = w
        w = w − η(v_s + Δ)
    end
    v_{s+1} = ṽ
    ṽ = 0 ∈ ℝ^d
    Δ = 0 ∈ ℝ^d
end
Return: w
```

Theorem. Suppose that Assumptions hold. Consider `Shuffled-SARAH` with the choice of η such that

$$\eta \le \min\left[\frac{1}{8nL}; \frac{1}{8n^2\delta}\right]. \qquad (2)$$

Then, we have convergence of $V_s := P(w_s) - P^* + \frac{\eta(n+1)}{16}\|v_{s-1}\|^2$ in the following form:

$$V_{s+1} \le \left(1 - \frac{\eta\mu(n+1)}{2}\right) V_s.$$

Corollary

Corollary. Fix ε, and let us run `Shuffled-SARAH` with η from (2). Then we can obtain an ε-accuracy solution on f after

$$S = \mathcal{O}\left(\max\left[\frac{L}{\mu}; \frac{\delta n}{\mu}\right] \log\frac{1}{\varepsilon}\right) \quad \text{iterations.}$$

Experiments

Trajectories. Compare the trajectories of the classical SARAH (two random and average), the average trajectory of the RR-SARAH, and the random trajectory `Shuffled-SARAH` with Random Reshuffling.

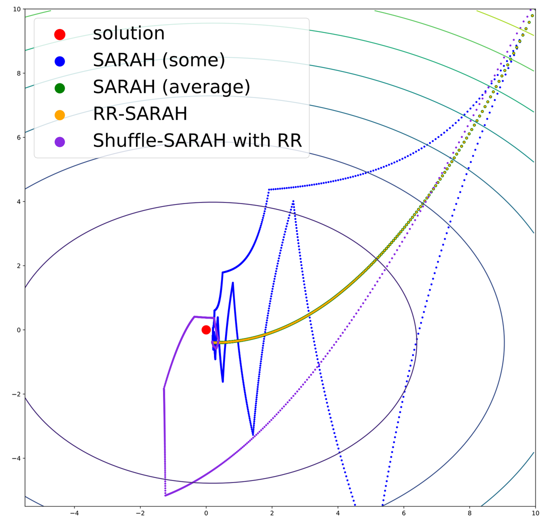
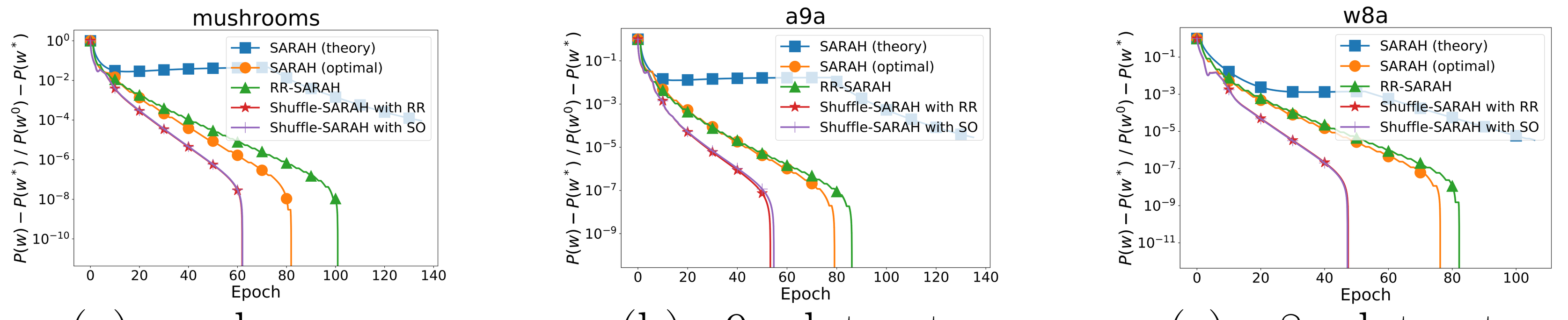


Figure 1:Trajectories on quadratic.

Logistic regression. Next, we consider the logistic regression problem with ℓ_2-regularization for binary classification with

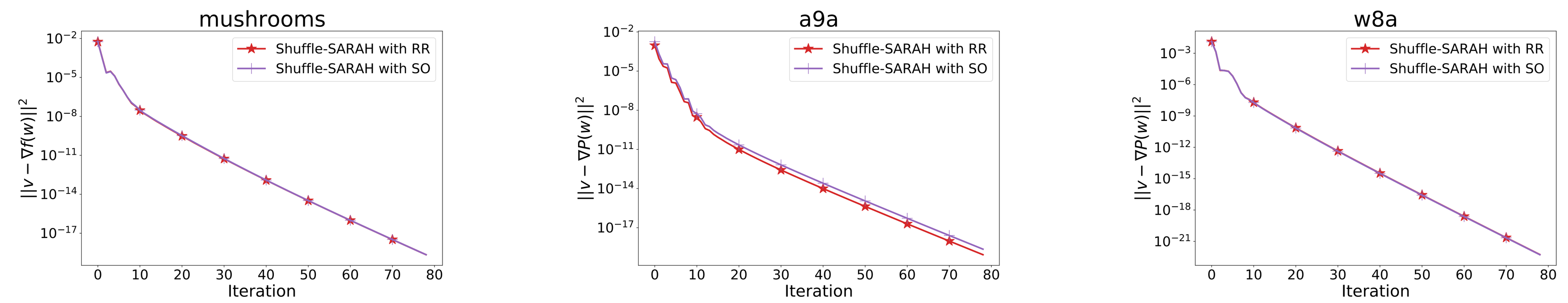
$$f_i(w) = \frac{1}{b}\Sigma_{k=1}^b \log\left(1 + \exp\left(-y_k \cdot (X_b w)_k\right)\right) + \frac{\lambda}{2}\|w\|^2,$$

where $X_b \in \mathbb{R}^{b \times d}$ is a matrix of objects, $y_1, \dots, y_b \in \{-1, 1\}$ are labels for these objects, b is the size of the local datasets and $w \in \mathbb{R}^d$ is a vector of weights. We optimize this problem for `mushrooms`, `a9a`, `w8a` datasets from LIBSVM.



(a) mushrooms (b) a9a dataset (c) w8a dataset

Figure 2:Convergence of SARAH-type methods on various LiBSVM datasets.



(a) mushrooms (b) a9a dataset (c) w8a dataset

Figure 3:$\|v_s - \nabla P(w_s)\|^2$ changes.

References

[1] Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements.

[2] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: a novel method for machine learning problems using stochastic recursive gradient.