SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient

Lam M Nguyen · Jie Liu · Katya Scheinberg · Martin Takáč (Lehigh University)

The Problem

Goal: minimize the finite-sum problem

$$\min_{x \in \mathbb{R}^d} \left\{ P(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

- each $f_i(x)$ is convex and has Lipschitz continuous gradient with parameter L

Two special cases

- **CASE-A:** function P is μ strongly convex
- **CASE-B:** each f_i is μ strongly convex

The Algorithm

Algorithm: SARAH vs. SVRG

choose x_0
for $s = 0, 1, 2, \dots$ **do**
  $\tilde{x}_0 = x_s$
  for $t = 0, 1, 2, \dots, m$ **do**
    choose random $i_t \sim U[\{1, 2, \dots, d\}]$
    compute stochastic gradient v_t
    update $\tilde{x}_{t+1} = \tilde{x}_t - \eta v_t$
  end for
  choose x_{s+1} randomly from $\{\tilde{x}_0, \dots, \tilde{x}_m\}$
end for

Stochastic gradient of SARAH

This gradient is defined recursively as

- $v_0 = \nabla P(\tilde{x}_0)$
- $v_t = v_{t-1} + \nabla f_{i_t}(\tilde{x}_t) - \nabla f_{i_t}(\tilde{x}_{t-1})$

Remarks:

1. $\mathbf{E}_{i_t}[v_t] \neq \nabla P(\tilde{x}_t)$, but $\mathbf{E}[v_t] = \nabla P(\tilde{x}_t)$!
2. No need for **extra storage as in SAG/SAGA!**

Stochastic gradient of SVRG

- $v_t = \nabla f_{i_t}(\tilde{x}_t) - \nabla f_{i_t}(\tilde{x}_0) + \nabla P(\tilde{x}_0)$

References

[1] Lam Nguyen, Jie Liu, Katya Scheinberg, Martin Takáč: Stochastic Recursive Gradient Algorithm for Nonconvex Optimization, arXiv:1705.07261.

[2] Rie Johnson, Tong Zhang: Accelerating stochastic gradient descent using predictive variance reduction, NIPS 2013.

Stochastic gradient of SARAH

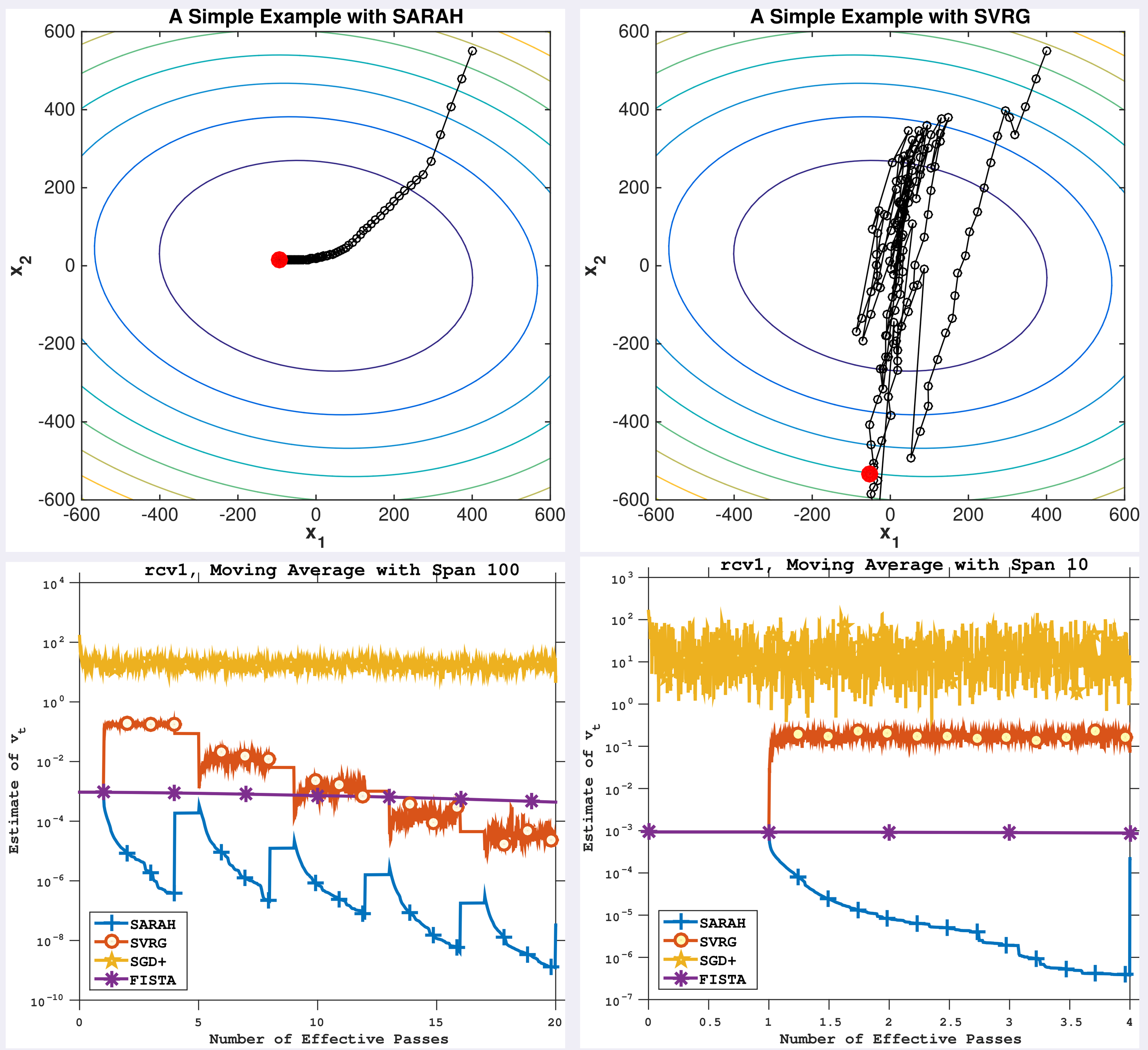
CASE-A

$$\mathbf{E}[\|v_t\|^2] \leq \left(1 - \left(\tfrac{2}{\eta L} - 1\right)\mu^2\eta^2\right)^t \|\nabla P(\tilde{x}_0)\|^2$$

CASE-B

$$\mathbf{E}[\|v_t\|^2] \leq \left(1 - \tfrac{2\mu\eta L}{\mu + L}\right)^t \|\nabla P(\tilde{x}_0)\|^2$$

SARAH is converging in each inner-loop



Convergence Analysis

Theorem: (CASE-A). For $\eta \in (0, 2/L)$ it holds

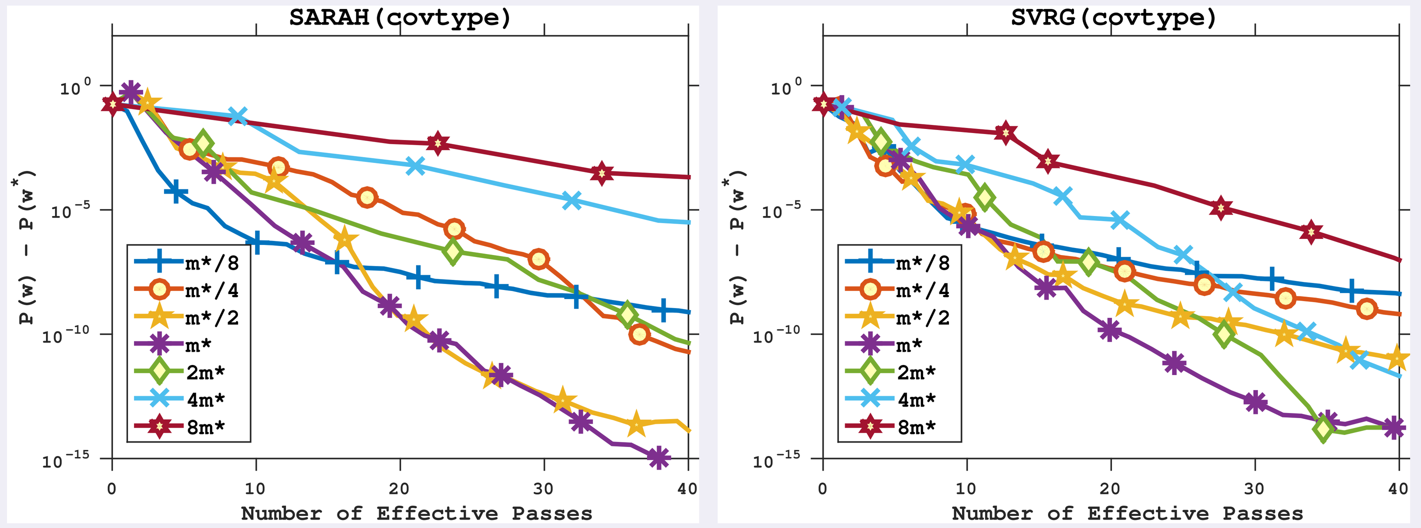
$$\mathbf{E}[\|\nabla P(x_s)\|^2] \leq \left(\frac{1}{\mu\eta(m+1)} + \frac{\eta L}{2 - \eta L}\right)^s \|\nabla P(x_0)\|^2$$

Remarks:

- This is **better** than convergence of SVRG.
- **CASE-B** have a slightly better convergence rate.
- In paper we also analyze convex case.
- We have extened it to non-convex case [1].

Practical variant SARAH+

Both SVRG and SARAH **need** m as an input! The performance is very sensitive on this choice.

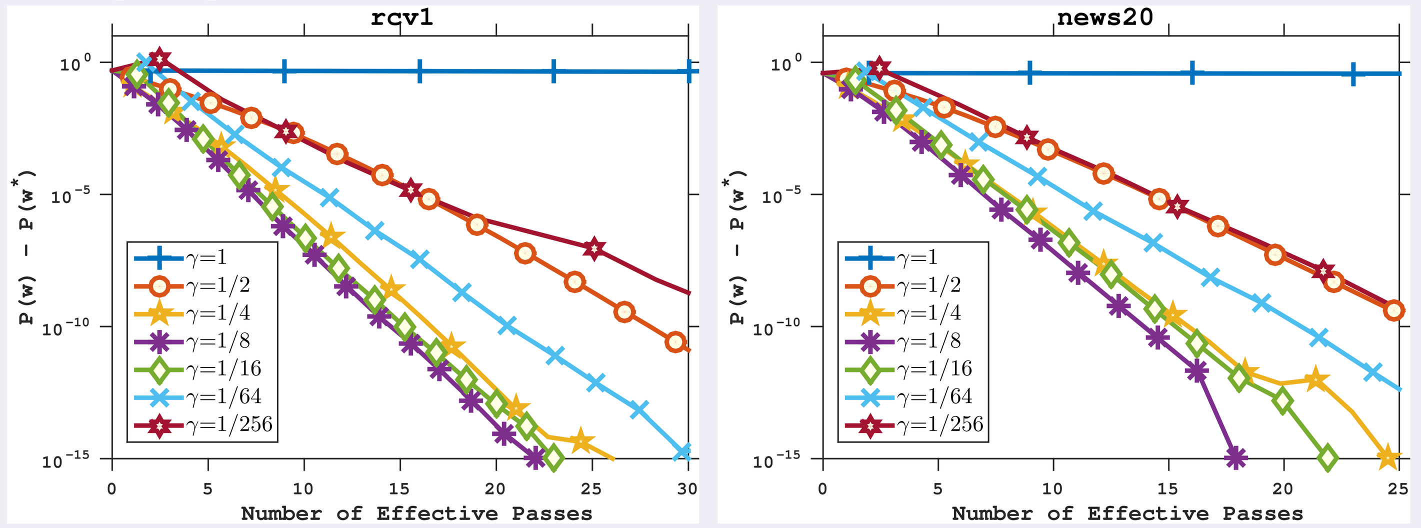


Facts:

- SARAH is converging in each outerloop.
- It would not be efficient to take many tiny steps.

SARAH+ Algorithm

Let's break the inner loop when $\|v_t\|^2 > \gamma\|v_0\|^2$ for some $\gamma \in [0, 1]$ (usually $\gamma = 0.1$ is a good choice).



Benefit: No need to tune parameter m!

Numerical Experiments