Adaptivity of Stochastic Gradient Methods for Nonconvex Optimization

Samuel Horváth[1], Lihua Lei[2], Peter Richtárik[1] and Michael I. Jordan[3]

KAUST[1] Stanford University[2] University of California, Berkeley[3]

Problem setup

We study **smooth non-convex** problems of the form

$$\min_{x\in\mathbb{R}^d}\left\{f(x)\stackrel{\text{def}}{=}\mathbb{E}f_\xi(x)\right\}. \qquad (1)$$

- the randomness comes from the selection of data points and is represented by the index ξ,
- the number of indices $n\in\{1,2,\dots,\infty\}$,
- optimal solution $x^\star$ of (1) exists and its value is finite: $f(x^\star)>-\infty$.

Dilemma of Parameter Tuning

A major drawback of many SGD methods to solve (1) is their **dependence on parameters** that are **unlikely to be known in a real-world machine-learning setting**, e.g.

- a uniform bound on the variance or second moment of the stochastic estimators of the gradient,
- required knowledge of final precision,
- lack of adaptivity of many SGD variants to different modelling regimes, for instance, if the function is known to satisfy some extra assumptions such as the Polyak-Łojasiewicz (PL) inequality.

We review two fundamental definitions introduced by Lei et al., 2019 that serve as a building block for desirable "parameter-free" optimization algorithms.

Definition

An algorithm is **ϵ-independent** if it guarantees convergence at all accuracies $\epsilon>0$.

Definition

An algorithm is **almost universal** if it only requires the knowledge of the smoothness L.

Complexity to reach an $\mathbb{E}\|\nabla f(x)\|^2\le\epsilon^2$ with $L,\sigma^2,\Delta_f=O(1)$.

Method	Complexity	Knowledge
SVRG (non-cvx) Reddi et al., 2016	$\mathcal{O}\left(n+\frac{n^{2/3}}{\epsilon^2}\right)$	L
SCSG (non-cvx) Lei et al., 2017	$\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^{10/3}}\wedge\frac{n^{2/3}}{\epsilon^2}\right)$	L
SNVRG (non-cvx) Zhou et al., 2018	$\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^3}\wedge\frac{\sqrt{n}}{\epsilon^2}\right)$ $\tilde{\mathcal{O}}\left(n+\frac{\sqrt{n}}{\epsilon^2}\right)$	L,σ^2,ϵ L
SARAH (non-cvx) Nguyen et al., 2019	$\mathcal{O}\left(n+\frac{\sqrt{n}}{\epsilon^2}\right)$	L
Q-Geom-SARAH	$\tilde{\mathcal{O}}\left(\left\{n^{3/2}+\frac{\sqrt{n}}{\mu}\right\}\wedge\frac{1}{\epsilon^3}\wedge\frac{\sqrt{n}}{\epsilon^2}\right)$	L
E-Geom-SARAH	$\tilde{\mathcal{O}}\left(\left(\frac{1}{\mu\wedge\epsilon}\right)^{2(1+\delta)}\wedge\left\{n+\frac{\sqrt{n}}{\mu}\right\}\wedge\frac{1}{\epsilon^4}\wedge\frac{\sqrt{n}}{\epsilon^2}\right)$	L
Non-adapt. Geom-SARAH	$\mathcal{O}\left(\left\{\frac{1}{\epsilon^{4/3}(\mu\wedge\epsilon)^{2/3}}\wedge n\right\}+\frac{1}{\mu}\left\{\frac{1}{\epsilon^{4/3}(\mu\wedge\epsilon)^{2/3}}\wedge n\right\}^{1/2}\right)$	L,σ^2,ϵ,μ

Contributions

- we present a new method—the *geometrized stochastic recursive gradient* (`Geom-SARAH`) algorithm—that exhibits adaptivity to the PL constant, target accuracy and to the variance of stochastic gradients.
- our algorithm does not require the computation of the full gradient in the outer loop as performed by other methods, but makes use of stochastic estimates of gradients in both the outer loop and the inner loop.
- by exploiting a randomization technique "geometrization" that allows certain terms to telescope across the outer loop and the inner loop, we obtain a significantly simpler analysis. As a byproduct, this allows us to obtain adaptivity, and our rates either match the known lower bounds (Fang et al., 2018) or achieve the same rates as existing state-of-the-art specialized methods
- for $\epsilon\sim\mu$, our complexity even beats the best available rate for strongly convex functions (Allen-Zhu, 2018).

Geom-SARAH

Input: stepsizes $\{\eta_j\}_{j=1}^{(1+\delta)T}$, big-batch sizes $\{B_j\}_{j=1}^{(1+\delta)T}$, expected inner-loop queries $\{m_j\}_{j=1}^{(1+\delta)T}$, mini-batch sizes $\{b_j\}_{j=1}^{(1+\delta)T}$, initializer $\tilde{x}_0$, tail-randomized fraction δ

for $j=1,\dots(1+\delta)T$ **do**
  $x_0^{(j)}=\tilde{x}_{j-1}$
  Sample J_j, $|J_j|=B_j$
  $v_0^{(j)}=\frac{1}{B_j}\sum_{i\in J_j}\nabla f_i(x_0^{(j)})$
  Sample $N_j\sim\text{Geom}(\gamma_j)$ s.t. $\mathbb{E}N_j=m_jb_j$
  for $k=0,\dots,N_j-1$ **do**
    $x_{k+1}^{(j)}=x_k^{(j)}-\eta_jv_k^{(j)}$
    Sample $I_k^{(j)}$, $|I_k^{(j)}|=b_j$
    $v_{k+1}^{(j)}=\frac{1}{b_j}\sum_{i\in I_k^{(j)}}(\nabla f_i(x_{k+1}^{(j)})-\nabla f_i(x_k^{(j)}))+v_k^{(j)}$
  end for
end for

Generate $\mathcal{R}(T)$ supported on $\{T,\dots,(1+\delta)T\}$ with $\text{Prob}(\mathcal{R}(T)=j)=\eta_jm_j\Sigma_{j=T}^{\lceil(1+\delta)T\rceil}\eta_jm_j$

Output: $\tilde{x}_{\mathcal{R}(T)}$

Numerical Experiments

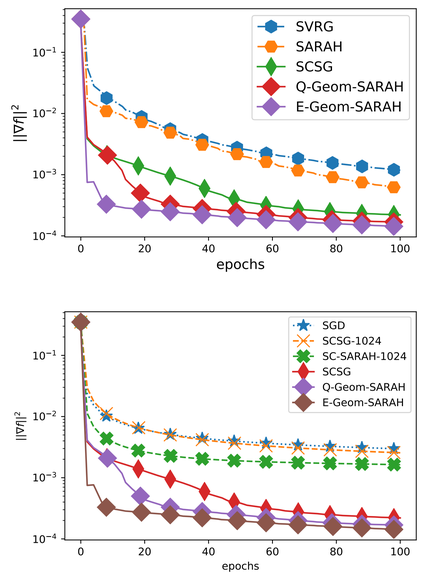

Figure 1:Comparison of convergence with respect to norm of the gradient for different high (top row) low precision (bottom row) VR methods. Dataset: `mushrooms`.

Contact Information

- Paper: https://arxiv.org/pdf/2002.05359.pdf
- Email: samuel.horvath@kaust.edu.sa
- Email: lihua.lei@stanford.edu

NEURAL INFORMATION PROCESSING SYSTEMS