ICML | 2020

Virtual Conference, July 2020

Almost Tune-Free Variance Reduction

Bingcong Li,* Lingda Wang,# and Georgios B. Giannakis*

*University of Minnesota
#University of Illinois at Urbana-Champaign

Acknowledgement: NSF 1711471 and 1901134

Context and motivation

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

Assumptions:

1. smoothness $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|, \forall i$
2. strong convexity $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \geq \mu\|\mathbf{x}-\mathbf{y}\|, \forall i$

   condition number $\kappa := L/\mu$

Complexity measure: number of $\nabla f_i(\mathbf{x})$ computed

- ❑ Solve ERM

GD

SGD

SVRG/SARAH

SARAH's gradient estimate

Algorithm 1 SARAH

```
Initialize: x̃^0, η, m, S
for s = 1, 2, ..., S do
    x_0^s = x̃^{s-1}, and v_0^s = ∇f(x_0^s)
    x_1^s = x_0^s − η v_0^s
    for k = 1, 2, ..., m − 1 do
        uniformly draw i_k ∈ [n]
        v_k^s = ∇f_{i_k}(x_k^s) − ∇f_{i_k}(x_{k−1}^s) + v_{k−1}^s
        x_{k+1}^s = x_k^s − η v_k^s
    end for
    draw x̃^s randomly from {x_k^s}_{k=0}^m according to p^s
end for
Output: x̃^S
```

- Identity of full gradient: Outer(s)-inner(k)

$$\nabla f(\mathbf{x}_k^s) = \nabla f(\mathbf{x}_k^s) - \sum_{\tau=0}^{k-1}\left[\nabla f(\mathbf{x}_\tau^s) - \nabla f(\mathbf{x}_\tau^s)\right] = \sum_{\tau=1}^{k}\left[\nabla f(\mathbf{x}_\tau^s) - \nabla f(\mathbf{x}_{\tau-1}^s)\right] + \nabla f(\mathbf{x}_0^s)$$

stochastic approximation

$$\mathbf{v}_k^s = \sum_{\tau=1}^{k}\left[\nabla f_{i_\tau}(\mathbf{x}_\tau^s) - \nabla f_{i_\tau}(\mathbf{x}_{\tau-1}^s)\right] + \nabla f(\mathbf{x}_0^s)$$

- Biased gradient estimate conditioning on $\mathcal{F}_{k-1}^s := \sigma(\tilde{\mathbf{x}}^{s-1}, i_0, i_1, \ldots, i_{k-1})$

$$\mathbb{E}[\mathbf{v}_k^s | \mathcal{F}_{k-1}^s] = \nabla f(\mathbf{x}_k^s) - \nabla f(\mathbf{x}_{k-1}^s) + \mathbf{v}_{k-1}^s \neq \nabla f(\mathbf{x}_k^s)$$

Nguyen LM, Liu J, Scheinberg K, Takáč M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. *In Proc. of International Conference on Machine Learning*. Vol. 70, pp. 2613-2621. Aug 2017.

SARAH recap

Algorithm 1 SARAH

Initialize: $\tilde{\mathbf{x}}^0$, η, m, S
for $s = 1, 2, \ldots, S$ **do**
  $\mathbf{x}_0^s = \tilde{\mathbf{x}}^{s-1}$, and $\mathbf{v}_0^s = \nabla f(\mathbf{x}_0^s)$
  $\mathbf{x}_1^s = \mathbf{x}_0^s - \eta \mathbf{v}_0^s$
  for $k = 1, 2, \ldots, m-1$ **do**
    uniformly draw $i_k \in [n]$
    $\mathbf{v}_k^s = \nabla f_{i_k}(\mathbf{x}_k^s) - \nabla f_{i_k}(\mathbf{x}_{k-1}^s) + \mathbf{v}_{k-1}^s$
    $\mathbf{x}_{k+1}^s = \mathbf{x}_k^s - \eta \mathbf{v}_k^s$
  end for
  draw $\tilde{\mathbf{x}}^s$ randomly from $\{\mathbf{x}_k^s\}_{k=0}^m$ according to $\mathbf{p}^s$
end for
Output: $\tilde{\mathbf{x}}^S$

- Uniform averaging (**U-Avg**) $p_m^s = 0, \text{ and } p_k^s = 1/m, \text{ for } k = \{0, 1, \ldots, m-1\}$
- Last iteration averaging (**L-Avg**) $p_{m-1}^s = 1 \text{ and } p_k^s = 0, \forall k \neq m-1$

Goal. Delve on averaging schemes to obtain tune-free algorithms

How about weighted averaging for SARAH?

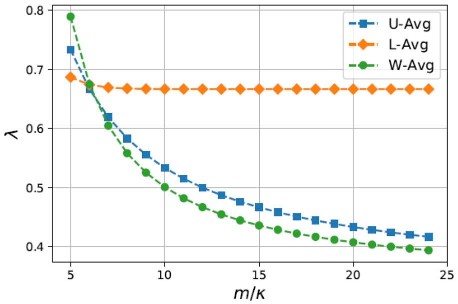
- Weighted averaging (W-Avg)

$$p_k \propto 1-(1-\mu\eta)^{m-k-1}, \forall k$$

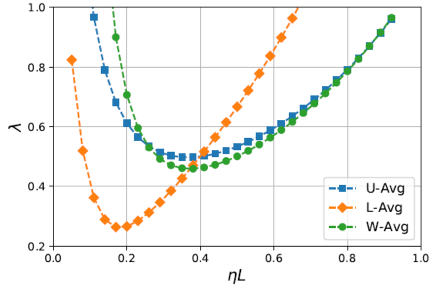
- Intuition

$$\mathbb{E}\big[\|\mathbf{v}_k - \nabla f(\mathbf{x}_k)\|^2\big] \leq \frac{\eta L}{2-\eta L}\Big(\mathbb{E}\big[\|\nabla f(\mathbf{x}_0)\|^2\big] - \underbrace{\mathbb{E}\big[\|\mathbf{v}_k\|^2\big]}_{\text{decreases as } k \text{ grows}}\Big)$$

- To ensure $\mathbb{E}\big[\|\nabla f(\mathbf{x})\|^2\big] \leq \epsilon$
  - Setting $\eta = \mathcal{O}(1/L)$ and $m = \mathcal{O}(\kappa)$, the complexity is $\mathcal{O}\big((n+\kappa)\ln\frac{1}{\epsilon}\big)$
  - Linear convergence $\mathbb{E}\big[\|\nabla f(\tilde{\mathbf{x}}^s)\|^2\big] \leq \lambda(\eta, m)\mathbb{E}\big[\|\nabla f(\tilde{\mathbf{x}}^{s-1})\|^2\big]$



fix η, change m



fix m, change η

Take home: W-Avg attractive if step size large or inner-loop large

Analysis highlights

- Estimate Sequence (ES)

$$w_k^\tau = (1-\mu\eta)^{k-\tau}$$

$$\textbf{GD:} \quad \Phi_k(\mathbf{x}) = w_k^0 \Phi_0(\mathbf{x}) + \sum_{\tau=0}^{k-1} w_k^\tau \Big[\underbrace{f(\mathbf{x}_\tau) + \langle \nabla f(\mathbf{x}_\tau), \mathbf{x} - \mathbf{x}_\tau \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}_\tau\|^2}_{\text{lower bound of } f(\mathbf{x}) \text{ due to strong convexity}} \Big]$$

$$\textbf{SGD/SVRG:} \quad \Phi_k(\mathbf{x}) = w_k^0 \Phi_0(\mathbf{x}) + \sum_{\tau=0}^{k-1} w_k^\tau \Big[\underbrace{f(\mathbf{x}_\tau) + \langle \mathbf{v}_\tau, \mathbf{x} - \mathbf{x}_\tau \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}_\tau\|^2}_{\text{lower bound of } f(\mathbf{x}) \text{ in expectation}} \Big]$$

- ES adapted for our context

$$\textbf{SARAH:} \quad \Phi_k(\mathbf{x}) = w_k^0 \Phi_0(\mathbf{x}) + \sum_{\tau=0}^{k-1} w_k^\tau \Big[f(\mathbf{x}_\tau) + \langle \mathbf{v}_\tau, \mathbf{x} - \mathbf{x}_\tau \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}_\tau\|^2 \Big]$$

not necessary a lower bound in expectation

Kulunchakov, A. and Mairal, J.,. Estimate Sequences for Variance-Reduced Stochastic Composite Optimization. *Proc. Intl. Conf. on Machine Learning*, pp. 3541-3550, May 2019.

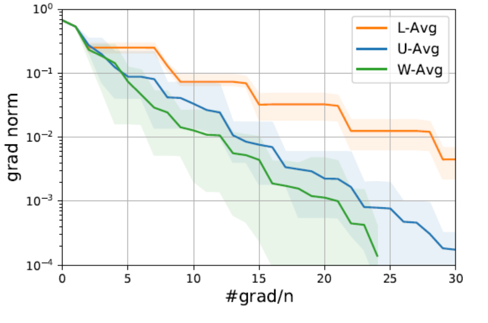
W-Avg on SARAH with BB step sizes

- SARAH with Barzilai-Borwein (BB) step sizes [Tan et al ‘16]

$$\eta^s = \frac{1}{\theta_\kappa} \frac{\|\tilde{\mathbf{x}}^{s-1} - \tilde{\mathbf{x}}^{s-2}\|^2}{\left\langle \tilde{\mathbf{x}}^{s-1} - \tilde{\mathbf{x}}^{s-2}, \nabla f(\tilde{\mathbf{x}}^{s-1}) - \nabla f(\tilde{\mathbf{x}}^{s-2}) \right\rangle}$$

  - Relying on L-Avg
  - Choosing $\theta_\kappa = m = \mathcal{O}(\kappa^2)$, the complexity is $\mathcal{O}\big((n+\kappa^2)\ln\frac{1}{\epsilon}\big)$

- W-Avg for BB step sizes
  - BB step sizes in L-Avg / U-Avg / W-Avg complexity $\mathcal{O}\big((n+\kappa^2)\ln\frac{1}{\epsilon}\big)$
  - negligible cost when $n \gg$
  - Observation: the choice of *m* can be very large



Tan, C., Ma, S., Dai, Y.H. and Qian, Y.. Barzilai-Borwein step size for stochastic gradient descent. In *Proc. Advances in Neural Information Processing Systems,* pp. 685-693, Dec. 2016.

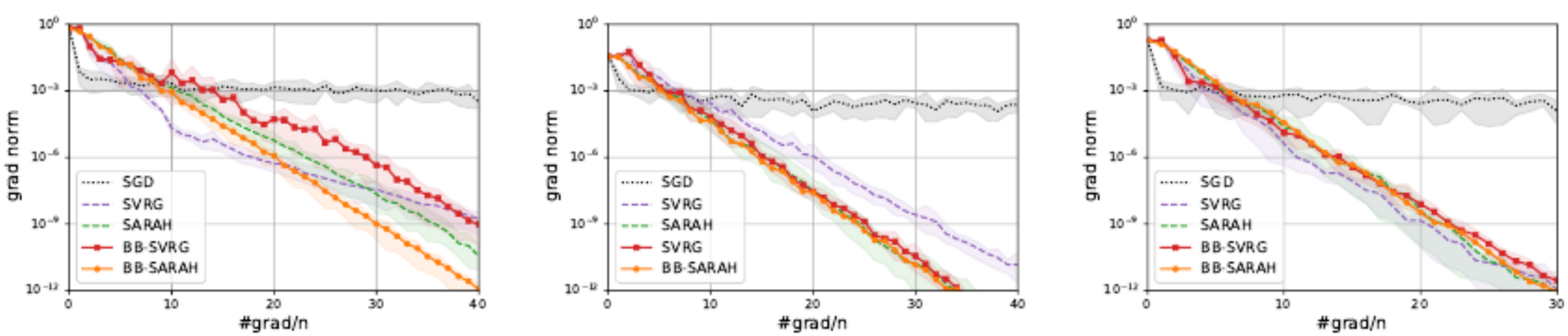
Almost tune free SARAH

- Inner loop length *m* still needs tuning
  - Solution: relying on a step-size-dependent inner loop length

$$m^s = \frac{c}{\mu \eta^s}$$

  - Principled guidelines to choose *c*
  - η^s and m inversely proportional ➡ **W-Avg**

- Numerical tests on regularized logistic regression



(a) *a9a* (b) *rcv1* (c) *real-sim*

All datasets are from LIBSVM; Online available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

Concluding remarks

ICML | 2020

- We talked about
  - Averaging schemes for SARAH
  - Almost tune free variance reduction with BB step sizes

- Future directions
  - Almost tune free variance reduction for (non)convex problems?
  - ES based analysis for ADAM type algorithms?

THANK YOU and STAY HEALTHY!