Accelerating Stochastic Gradient Decent for Least Squares Regression

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli and Aaron Sidford
Microsoft Research, India; University of Washington, Seattle; Stanford University

Goal and Motivation

- **Goal:** provably speedup SGD as implemented in practice.
- SGD [Robbins & Monro 1951]: simplest streaming algorithm.
  - Backbone of practical large-scale ML [Bottou & Bousquet 2008].
  - Iterate averaged SGD; asymptotically optimal [Polyak & Juditsky 1992].
- Many attempts to speed up SGD using curvature, momentum.
  - Constant factor improvement only (see for e.g. [Kidambi et al 2018]).

This work: presents the first non-asymptotic speedup of SGD on every problem while retaining its asymptotic optimality [Polyak & Juditsky 1992].

Problem Setup

- **Goal:** $w^* = \arg\min L(w) = 0.5 \cdot E_{(x,y)\sim D}[(y - \langle w, x\rangle)^2]$.
- **Hessian:** $H = E[xx^\top] \succ 0$; $\kappa_{GD} = \frac{\lambda_{max}[H]}{\lambda_{min}[H]}$; $\kappa = \frac{\max ||x||^2}{\lambda_{min}(H)}$.
- **Noise Model:** $y = \langle w^*, x\rangle + \epsilon$; $\Sigma = E[\epsilon^2 xx^\top] \preccurlyeq \sigma^2 H$.
- **SGD**: $w_t = w_{t-1} - \gamma\, \widehat{\nabla} L(w_{t-1}), \widehat{\nabla} L(w_{t-1}) = -(y_t - \langle w_{t-1}, x_t\rangle) \cdot x_t$.

Computations to achieve minimax error $O(d\sigma^2/n)$

		Vanilla Gradient	Fast Gradient
Offline (storage $O(nd)$)	Deterministic	[Cauchy, 1847] $\tilde{O}(nd \cdot \kappa_{GD})$	[Polyak, 1964] [Nesterov, 1983] $\tilde{O}(nd \cdot \sqrt{\kappa_{GD}})$
	Stochastic	[Johnson & Zhang 2013] $\tilde{O}((n+\kappa)\cdot d)$	[Frostig et al. 2015] [Allen-Zhu 2016] $\tilde{O}((n+\sqrt{n\kappa})\cdot d)$
Streaming (storage $O(d)$)		[Frostig et al. 2015] [Jain et al. 2016] $\tilde{O}(\kappa \cdot d)$	???

Related Work – I (Negative Results)

- Several efforts from Optimization, Controls, Signal Processing and Machine Learning tried to accelerate SGD.
- All efforts yielded negative results.
- Numerical errors: Paige (1971), Greenbaum (1989).
- Statistical errors: Proakis (1974), Polyak (1987), Roy et al (1990),….
- Adversarial errors: d'Aspremont (2008), Devolder et al. (2013, 2014).

Reason: (a) Inability to sharply characterize error accumulation of fast gradient methods.
(b) Inability to decouple Optimization from Statistics .

What does accelerating SGD even mean??

- Tail-averaged SGD [Jain et al. 2016]:
$$E[L(\overline{w})] - L(w^*) \le \underbrace{\exp(-n/\kappa)\cdot\Delta_0}_{Bias} + \underbrace{2d\sigma^2/n}_{Variance}$$
- Variance: minimax optimal. Unimprovable.
- Bias: Decays after κ steps. Is it improvable to $\sqrt{\kappa}$?? No!!

The Statistical Condition Number $\tilde{\kappa}$

- Assume $\boldsymbol{x}^\top \boldsymbol{H}^{-1}\boldsymbol{x} < \tilde{\boldsymbol{\kappa}}$. Once $n > \tilde{\kappa}$,
$$\frac{1}{c}\cdot\widehat{H} \preccurlyeq H \preccurlyeq c\cdot\widehat{H},\ c > 1, \text{ with } \widehat{H} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top.$$
- **Discrete case:** $x \sim e_i$ with probability p_i; $\tilde{\kappa} = \kappa = 1/p_{min}$.
  - $\widehat{H}$ invertible after $\tilde{\kappa} = \kappa$ samples! No improvement over SGD.
- **Gaussian case:** $x \sim N(0, H)$, $\tilde{\kappa} = d < \kappa$.
  - $O(d)$ samples suffice for inverting $\widehat{H}$. SGD appears improvable.
- What is this improvement even going to resemble?

This paper's Main Result

Theorem: Assume $n > \tilde{O}(\sqrt{\kappa\tilde{\kappa}})$. Running Accelerated SGD with $\beta = \frac{0.9c}{\sqrt{\kappa\tilde{\kappa}}}$, $\alpha = \frac{c}{c+\beta}$, $\delta = \frac{1}{\max ||x||^2}$ returns $\overline{\boldsymbol{w}}$ that satisfies:
$$E[L(\overline{w})] - L(w^*) \le exp\left(-\frac{n}{\sqrt{\kappa\tilde{\kappa}}}\right)\Delta_0 + 11\frac{d\sigma^2}{n}.$$

Related Work-II (Additive noise oracle model)

- **Bounded Noise**, i.e. $||\widehat{\nabla} L(\cdot) - \nabla L(\cdot)||^2 \le \sigma^2$: textbook assumption for analyzing SGD ($\approx 990/1000$ papers).
- Accelerating SGD - positive results in this additive noise oracle
  - Lan (2008), Ghadimi & Lan (2012,13), Dieuleveut, Flammarion and Bach (2017), Dieuleveut et al (2017b).
- Reasonable, but not reflective of SGD's implementations in ML:
  - Requires compactness of parameter set (enforced via projections).
  - No input dependent characterization (e.g. Gaussian versus Discrete inputs)
  - Requires $O(d^2)$ computation per iteration [Flammarion, thesis 2017].
  - Worst case upper bounds! **This paper's bounds** hold on every problem.
- SGD in practice: Multiplicative noise oracle (for e.g.: this paper).

Algorithm 1: Tail-Averaged Accelerated SGD

Start with $w_0 = v_0 = z_0$. Repeat for $t = 1, 2, \cdots, n$

- $w_t \leftarrow z_{t-1} - \delta \cdot \widehat{\nabla} L_t(z_{t-1})$ /* SGD step */
- $v_t \leftarrow \beta\left(z_{t-1} - \frac{1}{\lambda_{min}(H)}\cdot\widehat{\nabla} L_t(z_{t-1})\right) + (1-\beta)v_{t-1}$ /* discounted average of long steps */
- $z_t \leftarrow \alpha w_t + (1-\alpha)v_t$ /*linear combination of steps*/
- Return $\overline{w} \leftarrow \frac{1}{n/2}\sum_{i>n/2} w_i$. /*return tail-averaged iterate*/

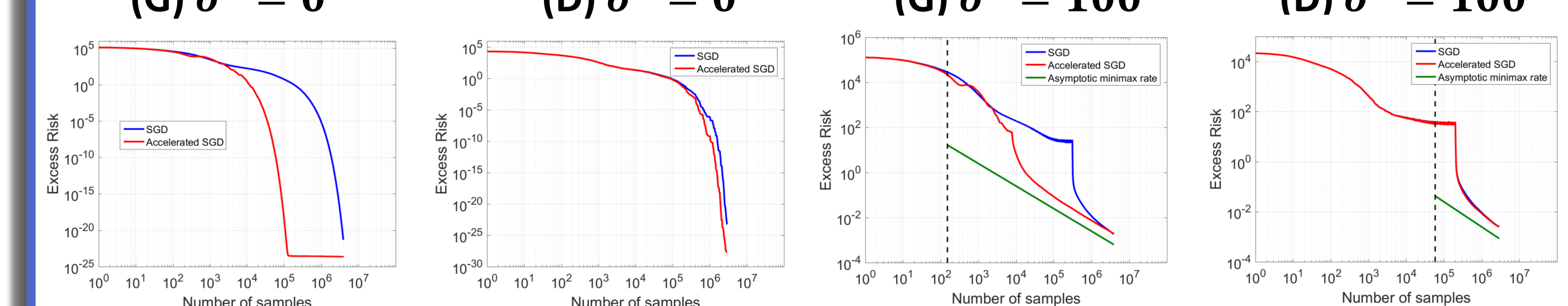
Techniques

- Centered estimate $\theta_t = [w_t - w^*; z_t - w^*]$.
- Proof goes via bias-variance decomposition:
  - Bias: $\theta_t^{bias} = A_t \theta_{t-1}^{bias}$ (running with no additive noise).
  - Variance: $\theta_t^{var} = A_t\theta_{t-1}^{var} + \zeta_t$ ($\theta_0^{var} = 0$, run SGD while starting at the solution).
- New potential function $P_t = ||w_t - w^*||^2 + \lambda_{min}(H)||v_t - w^*||_{H^{-1}}^2$
$$E[P_{t+1}] \le \left(1 - 1/\sqrt{\tilde{\kappa}\kappa}\right)\cdot P_t$$
- Stochastic process view: tight bound on steady state covariance of θ_t
$$\lim_{t\to\infty} E[\theta_t \otimes \theta_t] \preccurlyeq \sigma^2(\mathrm{H}^{-1}/\tilde{\kappa} + \delta I) \otimes I_{2\times 2}$$
- Implying final iterate has excess risk $O(\sigma^2)$, avg. iterate: minmax optimal.

Simulations

Synthetic ex. $d = 50, \kappa \approx 10^5$. (G)-Gaussian, (D)-Discrete.

(G) $\sigma^2 = 0$ (D) $\sigma^2 = 0$ (G) $\sigma^2 = 100$ (D) $\sigma^2 = 100$



Conclusions

- Acceleration of SGD indeed possible: gains – distribution dependent.
  - Gains formalized through the statistical condition number $\tilde{\kappa}$.
- The first accelerated SGD result with a multiplicative noise oracle.
  - Practically relevant gains [Kidambi et al. 2018], beyond least squares.
- Accelerated SGD improves on SGD ≡ heavy ball does over Gradient Descent.

Microsoft Research | UNIVERSITY of WASHINGTON | ADSI Algorithmic Foundations for Data Science Institute | Stanford University