Accelerated Stochastic Gradient Descent

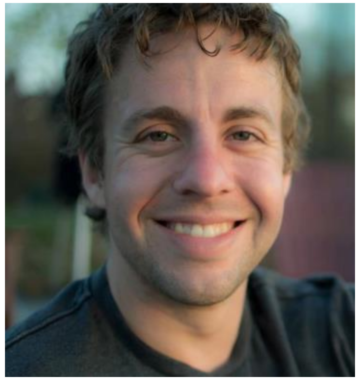
Praneeth Netrapalli

MSR India

Presented at **OSL workshop, Les Houches, France.**

Joint work with

Prateek Jain, Sham M. Kakade, Rahul Kidambi and Aaron Sidford



Linear regression

$$\min_{x} f(x) = \|Ax - b\|_2^2$$

$$x \in \mathbb{R}^d, A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$$

- Basic problem and arises pervasively in applications
- Deeply studied in literature

Gradient descent for linear regression

$$x_{t+1} = x_t - \delta \cdot \underbrace{A^\top (Ax_t - b)}_{\text{Gradient}}$$

- Convergence rate: $O\left(\kappa \log \frac{f(x_0) - f^*}{\epsilon}\right)$
- $f^* = \min_x f(x)$; $\epsilon =$ Target suboptimality
- Condition number: $\kappa = \frac{\sigma_{max}(A^\top A)}{\sigma_{min}(A^\top A)}$

Question: Is it possible to do better?

- Hope: GD does not reuse past gradients

Gradient descent:

$$x_{t+1} = x_t - \delta \cdot \nabla f(x_t)$$

- Answer: Yes!

  - Conjugate gradient (Hestenes and Stiefel 1952)
  - Heavy ball method (Polyak 1964)
  - Accelerated gradient descent (Nemirovsky and Yudin 1977, Nesterov 1983)

Accelerated gradient descent (AGD)

$$x_{t+1} = y_t - \delta \nabla f(y_t)$$
$$y_{t+1} = x_{t+1} + \gamma(x_{t+1} - x_t)$$

- Convergence rate: $O\left(\sqrt{\kappa} \log \frac{f(x_0) - f^*}{\epsilon}\right)$

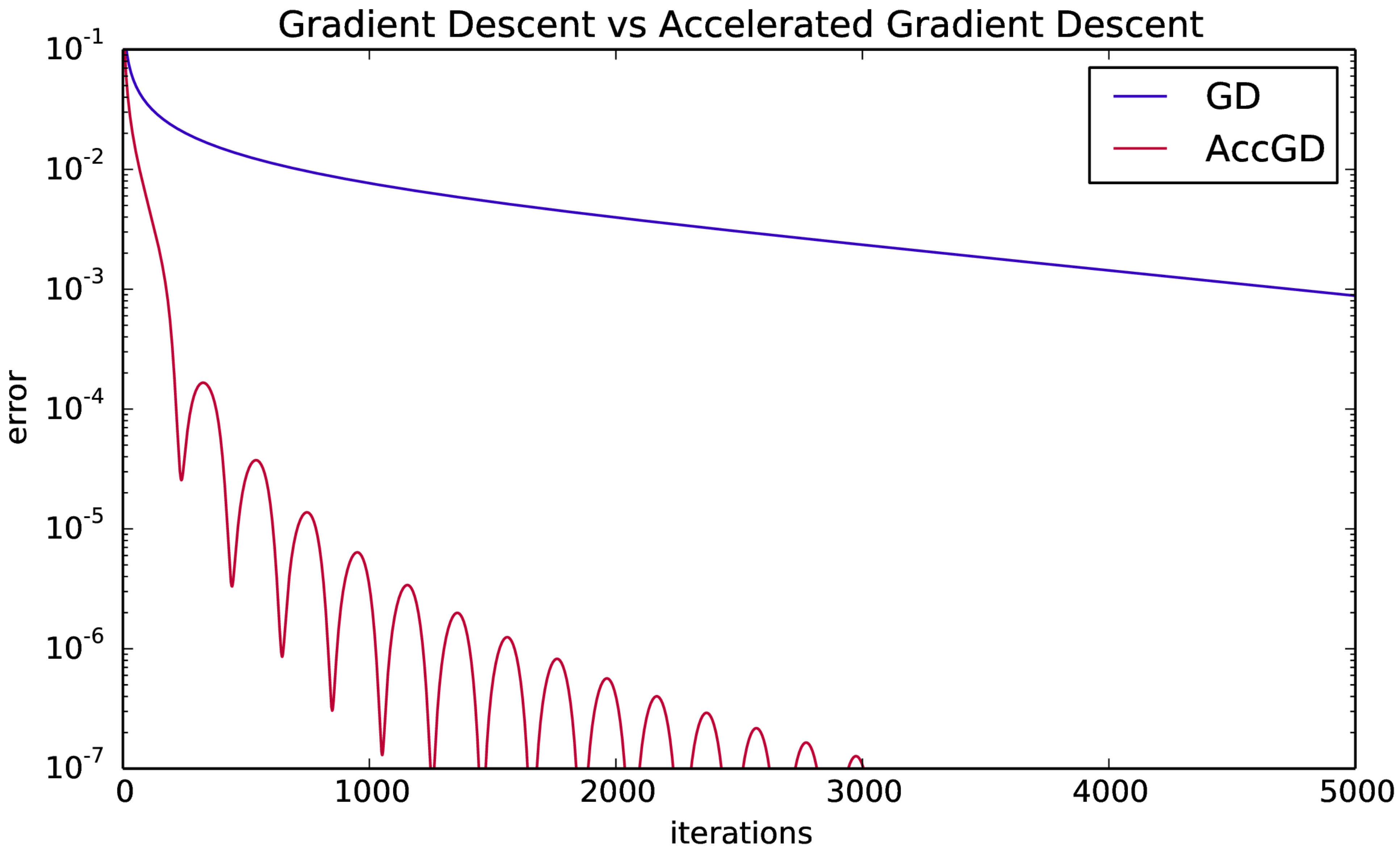
- $f^* = \min_x f(x)$; ϵ = Target suboptimality

- Condition number: $\kappa = \frac{\sigma_{max}(A^\top A)}{\sigma_{min}(A^\top A)}$

Accelerated gradient descent (AGD)

<u>Compared to:</u> $O\left(\kappa \log\frac{f(x_0)-f^*}{\epsilon}\right)$ for GD

- Convergence rate: $O\left(\sqrt{\kappa} \log\frac{f(x_0)-f^*}{\epsilon}\right)$
- $f^* = \min_x f(x)$; $\epsilon =$ Target suboptimality
- Condition number: $\kappa = \frac{\sigma_{max}(A^\top A)}{\sigma_{min}(A^\top A)}$



Source: http://blog.mrtz.org/2014/08/18/robustness-versus-acceleration.html

Stochastic approximation (Robbins and Monro 1951)

- Distribution $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$
- $f(x) \triangleq \mathbb{E}_{(a,b)\sim\mathcal{D}}[(a^\top x - b)^2]$
- Equivalent to $||Ax - b||_2^2$ where A has infinite rows
- Observe n pairs $(a_1, b_1), \cdots, (a_n, b_n)$
- Interested in entire distribution $\mathcal{D}$ rather than data points like in ML
- Fit a linear model to the distribution
- Cannot compute exact gradients

Stochastic gradient descent (SGD) (Robbins and Monro 1951)

- $x_{t+1} = x_t - \delta \cdot \hat{\nabla} f(x_t)$ where $\mathbb{E}\left[\hat{\nabla} f(x_t)\right] = \nabla f(x_t)$

- Return $\frac{1}{n} \sum_i x_i$ (Polyak and Juditsky 1992)

- Is gradient descent in expectation

- For linear regression, SGD: $x_{t+1} = x_t - \delta \cdot \left(a_t^\top x_t - b_t\right) a_t$

- **Streaming algorithm**: extremely efficient and widely used in practice

Best possible rate

- Consider $b = a^\top x^* + \text{noise}$; $\quad \text{noise} \sim \mathcal{N}(0, \sigma^2)$

- Recall: $f(x) \triangleq \mathbb{E}_{(a,b)\sim\mathcal{D}}[(a^\top x - b)^2]$

- $\hat{x} \triangleq \underset{x}{\operatorname{argmin}} \sum_{i=1}^{n} \left(a_i^\top x - b_i\right)^2$

- $\mathbb{E}[f(\hat{x})] - f(x^*) = \big(1 + o(1)\big)\frac{\sigma^2 d}{n}$ (van der Vaart, 2000)

Best possible rate

- In general: $x^* \triangleq \underset{x}{\operatorname{argmin}} f(x)$

- $\mathbb{E}[(a^\top x^* - b)^2 a a^\top] \preccurlyeq \sigma^2 \mathbb{E}[a a^\top]$

- $\hat{x} \triangleq \underset{x}{\operatorname{argmin}} \sum_{i=1}^{n} \left(a_i^\top x - b\right)^2$

- $\mathbb{E}[f(\hat{x})] - f(x^*) \leq \left(1 + o(1)\right) \frac{\sigma^2 d}{n}$ (van der Vaart, 2000)

<u>Equivalently</u>

$$n \leq \left(1 + o(1)\right) \frac{\sigma^2 d}{\epsilon}$$

Convergence rate of SGD

- Convergence rate: $\tilde{O}\left(\kappa \log \frac{f(x_0)-f^*}{\epsilon} + \frac{\sigma^2 d}{\epsilon}\right)$ (Jain et al. 2016)
- $f^* = \min_x f(x)$; $\epsilon =$ Target suboptimality
- Condition number: $\kappa \triangleq \frac{\max\|a\|_2^2}{\sigma_{min}(\mathbb{E}[aa^\top])}$
- Noise level: $\mathbb{E}[(a^\top x^* - b)^2 aa^\top] \preccurlyeq \sigma^2 \mathbb{E}[aa^\top]$

Recap

Deterministic case	Stochastic approximation
GD $O\left(\kappa \log \frac{f(x_0) - f^*}{\epsilon}\right)$	**SGD** $\tilde{O}\left(\kappa \log \frac{f(x_0) - f^*}{\epsilon} + \frac{\sigma^2 d}{\epsilon}\right)$
AGD $O\left(\sqrt{\kappa} \log \frac{f(x_0) - f^*}{\epsilon}\right)$	**Accelerated SGD?** Unknown

Question: Is accelerating SGD possible?

Is this really important?

- Extremely important in practice
  - ➢ As we saw, acceleration can really give orders of magnitude improvement
  - ➢ Neural network training uses Nesterov's AGD as well as Adam; but no theoretical understanding
  - ➢ Jain et al. 2016 shows acceleration leads to more parallelizability

- Existing results show AGD not robust to deterministic noise (d'Aspremont 2008, Devolder et al. 2014) but is robust to random additive noise (Ghadimi and Lan 2010, Dieuleveut et al. 2016)

- Stochastic approximation falls between the above two cases

- Key issue: mixes optimization and statistics (i.e., # iterations = #samples)

Is acceleration possible?

- $b = a^\top x^*$ $\Rightarrow$ Noise level: $\sigma^2 = 0$
- SGD convergence rate: $\tilde{O}\left(\kappa \log \frac{f(x_0) - f^*}{\epsilon}\right)$
- Accelerated rate: $\tilde{O}\left(\sqrt{\kappa} \log \frac{f(x_0) - f^*}{\epsilon}\right)$?

Example I: Discrete distribution

- $a = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ with probability p_i
- In this case, $\kappa \triangleq \frac{\max \|a\|_2^2}{\sigma_{min}(\mathbb{E}[aa^\top])} = \frac{1}{p_{\min}}$
- Is $\tilde{O}\left(\sqrt{\kappa} \log \frac{f(x_0) - f^*}{\epsilon}\right)$ possible?
- Or, halve the error using $\tilde{O}(\sqrt{\kappa})$ samples?

Example I: Discrete distribution

- $a = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ with probability p_i; $\kappa \triangleq \frac{1}{p_{\min}}$
- Fewer than κ samples $\Rightarrow$ do not observe $p_{\min}$ direction
  $\Rightarrow \sum_i a_i a_i^\top$ not invertible
- Cannot do better than $O(\kappa)$
- Acceleration not possible

Example II: Gaussian

- $a \sim \mathcal{N}(0, H)$, H is a PSD matrix

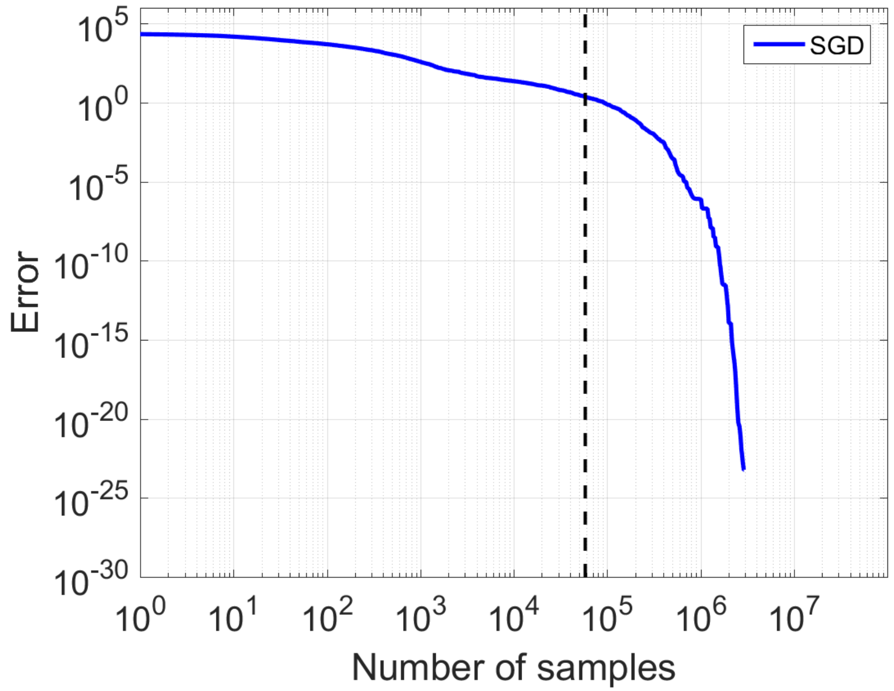
- In this case, $\kappa \sim \frac{\mathrm{Tr}(H)}{\sigma_{\min}(H)} \geq d$

- However, after $O(d)$ samples: $\frac{1}{n}\sum_i \; a_i a_i^\top \sim H$

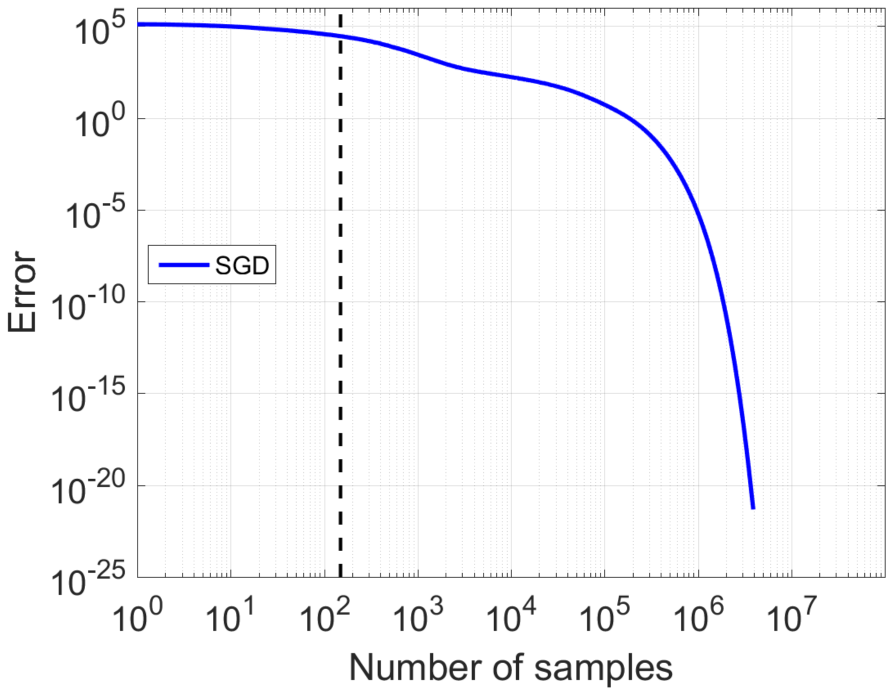
- Possible to solve $a_i^\top x^* = b_i$ after $\mathrm{O}(d)$ samples

- Acceleration *might be* possible

Discrete vs Gaussian



Discrete distribution



Gaussian distribution

Key issue: matrix spectral concentration

- Recall: $a_i \sim \mathcal{D}$. Let $H \stackrel{\text{def}}{=} \mathbb{E}\left[a_i a_i^\top\right]$.
- For $\hat{x} \triangleq \underset{x}{\operatorname{argmin}} \sum_{i=1}^{n}\left(a_i^\top x - b_i\right)^2$ to be good, need:

$$(1-\delta)H \preccurlyeq \frac{1}{n}\sum_{i=1}^{n} a_i a_i^\top \preccurlyeq (1+\delta)H$$

- How many samples are required for spectral concentration?

Separating optimization and statistics

Matrix variance (Tropp 2012): $\left\|\mathbb{E}\left[\|a\|_2^2 aa^\top\right]\right\|_2$

Recall $H \stackrel{\text{def}}{=} \mathbb{E}[aa^\top]$

Statistical condition number: $\tilde{\kappa} \stackrel{\text{def}}{=} \left\|\mathbb{E}\left[\left\|H^{-\frac{1}{2}}a\right\|_2^2 \left(H^{-\frac{1}{2}}a\right)\left(H^{-\frac{1}{2}}a\right)^\top\right]\right\|_2$

<u>Matrix Bernstein Theorem</u>

(Tropp 2015)

If $n > O(\tilde{\kappa})$, then $(1-\delta)H \preccurlyeq \frac{1}{n}\sum_{i=1}^{n} a_i a_i^\top \preccurlyeq (1+\delta)H$

Is acceleration possible?

- $O(\tilde{\kappa})$ samples sufficient

- Recall SGD convergence rate: $\tilde{O}\left(\kappa \log \frac{f(x_0)-f^*}{\epsilon}\right)$

- Always $\tilde{\kappa} \leq \kappa$. Acceleration might be possible if $\tilde{\kappa} \ll \kappa$

- Discrete case: $\tilde{\kappa} = \frac{1}{p_{\min}} = \kappa$; Gaussian case: $\tilde{\kappa} = O(d) \leq \kappa$

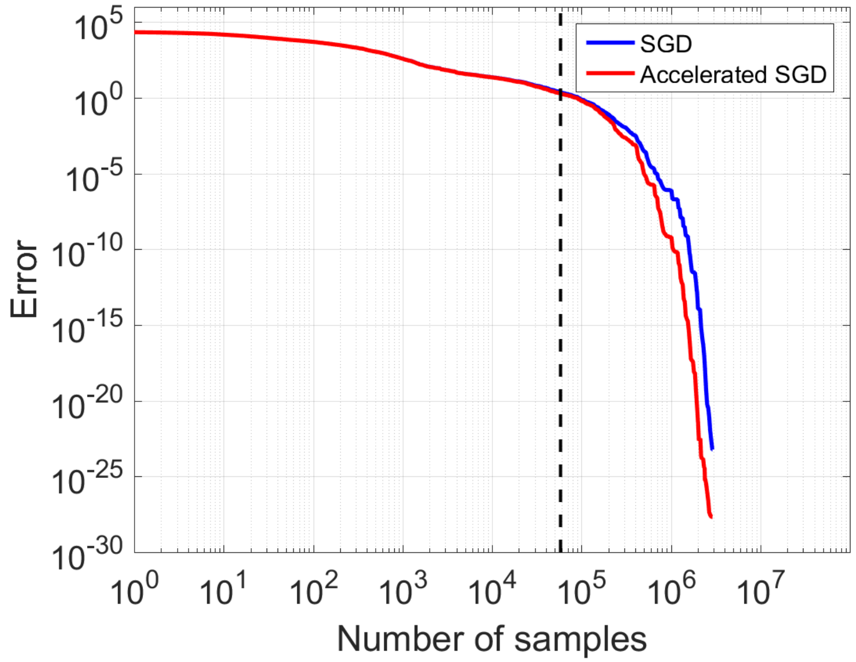
Result

- Convergence rate of ASGD: $\tilde{O}\left(\sqrt{\kappa\tilde{\kappa}}\log\frac{f(x_0)-f^*}{\epsilon}+\frac{\sigma^2 d}{\epsilon}\right)$
- Compared to SGD: $\tilde{O}\left(\kappa\log\frac{f(x_0)-f^*}{\epsilon}+\frac{\sigma^2 d}{\epsilon}\right)$
- Improvement since $\tilde{\kappa} \leq \kappa$
- Conjecture: lower bound $\Omega\left(\sqrt{\kappa\tilde{\kappa}}\log\frac{f(x_0)-f^*}{\epsilon}\right)$ (inspired by Woodworth and Srebro 2016)

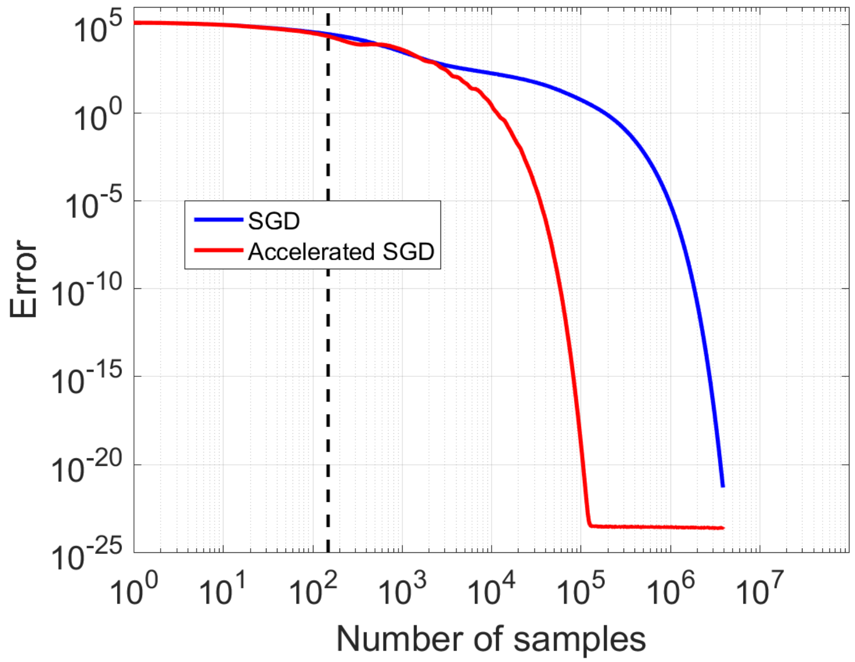
Key takeaway

- Acceleration possible!
- Gain depends on statistical condition number

Simulations – No noise

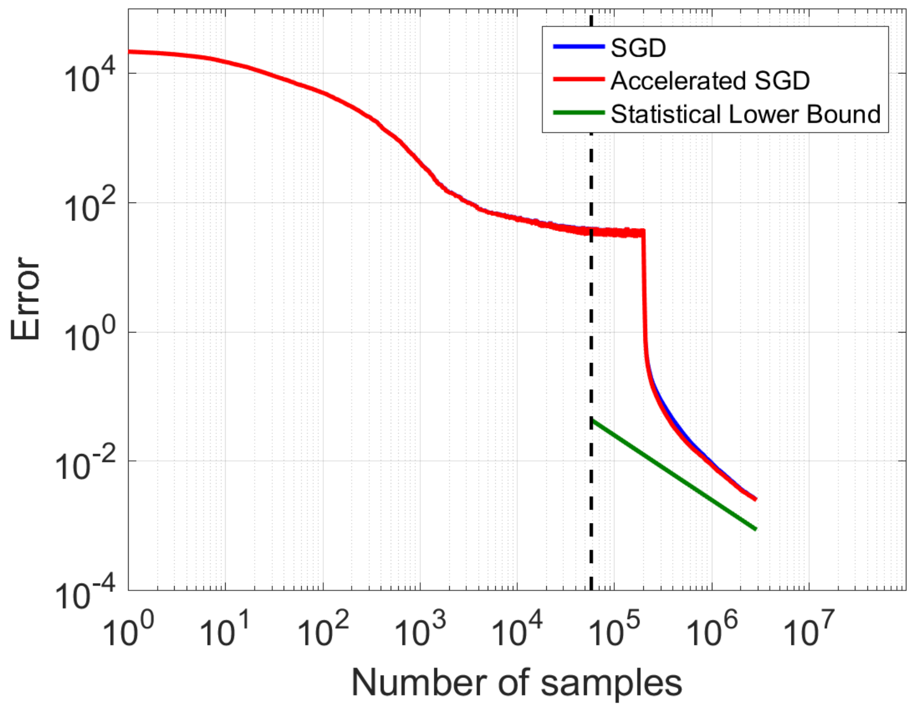


Discrete distribution

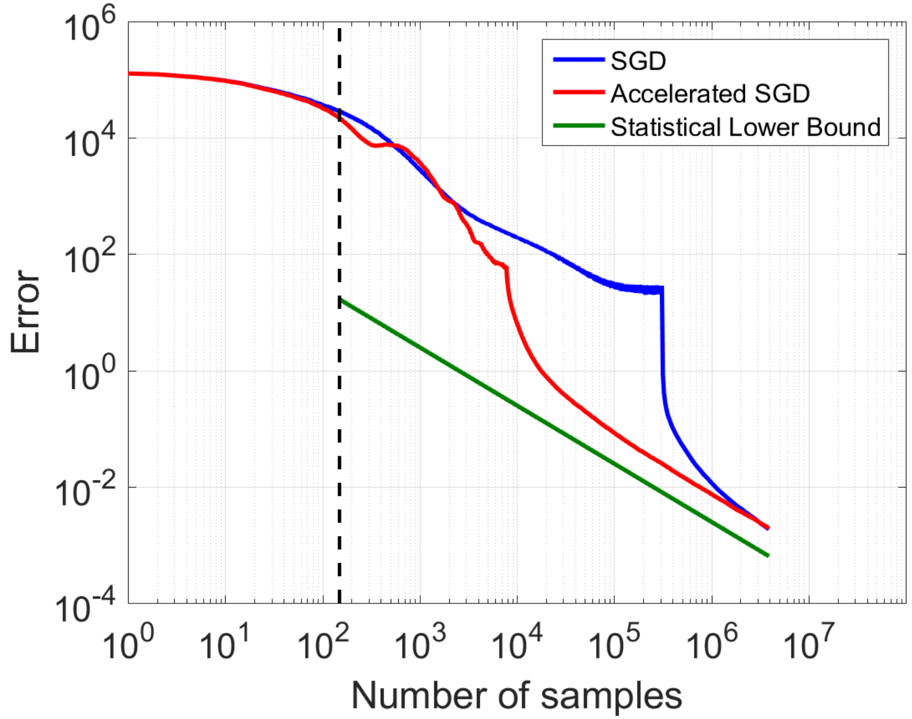


Gaussian distribution

Simulations – With noise



Discrete distribution



Gaussian distribution

High level challenges

- Several versions of accelerated algorithms known e.g., conjugate gradient 1952, heavy ball 1964, momentum methods 1983, accelerated coordinate descent 2012, linear coupling 2014

- Many of them are equivalent in deterministic setting but not in stochastic setting

- Many different analyses even for momentum methods: Nesterov's analysis 1983, coordinate descent 2012, ODE analysis 2013, linear coupling 2014

Algorithm

Parameters: $\alpha, \beta, \gamma, \delta$

1. $v_0 = x_0$
2. $y_{t-1} = \alpha x_{t-1} + (1-\alpha)v_{t-1}$
3. $x_t = y_{t-1} - \delta \nabla f(y_{t-1})$
4. $z_{t-1} = \beta y_{t-1} + (1-\beta)v_{t-1}$
5. $v_t = z_{t-1} - \gamma \nabla f(y_{t-1})$

Nesterov 2012

Parameters: $\alpha, \beta, \gamma, \delta$

1. $v_0 = x_0$
2. $y_{t-1} = \alpha x_{t-1} + (1-\alpha)v_{t-1}$
3. $x_t = y_{t-1} - \delta \hat{\nabla}_t f(y_{t-1})$
4. $z_{t-1} = \beta y_{t-1} + (1-\beta)v_{t-1}$
5. $v_t = z_{t-1} - \gamma \hat{\nabla}_t f(y_{t-1})$

Our algorithm

Proof overview

- Recall our guarantee: $\tilde{O}\left(\sqrt{\kappa\tilde{\kappa}}\log\frac{f(x_0)-f^*}{\epsilon}+\frac{\sigma^2 d}{\epsilon}\right)$
- First term depends on initial error; second is statistical error
- Different analyses for the two terms
- For the first term, analyze assuming $\sigma = 0$
- For the second term, analyze assuming $x_0 = x^*$

Part I: Potential function

- Iterates x_t, v_t of ASGD. $H \triangleq \mathbb{E}[aa^\top]$.

- Existing analyses use potential function

$$\|x_t - x^*\|_H^2 + \sigma_{\min}(H) \cdot \|v_t - x^*\|_2^2$$

- We use $$\|x_t - x^*\|_2^2 + \sigma_{\min}(H) \cdot \|v_t - x^*\|_{H^{-1}}^2$$

- We show $$\|x_t - x^*\|_2^2 + \sigma_{\min}(H) \cdot \|v_t - x^*\|_{H^{-1}}^2$$

$$\leq \left(1 - \frac{1}{\sqrt{\kappa\tilde{\kappa}}}\right) \|x_{t-1} - x^*\|_2^2 + \sigma_{\min}(H) \cdot \|v_{t-1} - x^*\|_{H^{-1}}^2$$

Part II: Stochastic process analysis

$$\begin{bmatrix} x_{t+1} - x^* \\ y_{t+1} - x^* \end{bmatrix} = C \begin{bmatrix} x_t - x^* \\ y_t - x^* \end{bmatrix} + \text{noise}$$

Let $\theta_t \triangleq \mathbb{E}\begin{bmatrix} x_t - x^* \\ y_t - x^* \end{bmatrix}\begin{bmatrix} x_t - x^* \\ y_t - x^* \end{bmatrix}^\top$

$$\theta_{t+1} = \mathfrak{B}\theta_t + \text{noise} \cdot \text{noise}^\top$$

$$\theta_n \to \sum_i \mathfrak{B}^i(\text{noise} \cdot \text{noise}^\top)$$
$$= (\mathbb{I} - \mathfrak{B})^{-1}(\text{noise} \cdot \text{noise}^\top)$$

Parameters: $\alpha, \beta, \gamma, \delta$

1. $v_0 = x_0$
2. $y_{t-1} = \alpha x_{t-1} + (1-\alpha)v_{t-1}$
3. $x_t = y_{t-1} - \delta\hat{\nabla} f(y_{t-1})$
4. $z_{t-1} = \beta y_{t-1} + (1-\beta)v_{t-1}$
5. $v_t = z_{t-1} - \gamma\hat{\nabla} f(y_{t-1})$

Our algorithm

Part II: Stochastic process analysis

- Need to understand $(\mathbb{I} - \mathfrak{B})^{-1}(\mathrm{noise} \cdot \mathrm{noise}^{\top})$

- $\mathfrak{B}$ has singular values > 1, but fortunately eigenvalues < 1

- Solve the 1-dim version of $(\mathbb{I} - \mathfrak{B})^{-1}(\mathrm{noise} \cdot \mathrm{noise}^{\top})$ via explicit computations

- Combine the 1-dim bounds with (statistical) condition number bounds

- $(\mathbb{I} - \mathfrak{B})^{-1}(\mathrm{noise} \cdot \mathrm{noise}^{\top}) \preccurlyeq \tilde{\kappa} H^{-1} + \delta \cdot I$

Recap

Deterministic case	Stochastic approximation
GD $O\left(\kappa \log \frac{f(x_0) - f^*}{\epsilon}\right)$	**SGD** $\tilde{O}\left(\kappa \log \frac{f(x_0) - f^*}{\epsilon} + \frac{\sigma^2 d}{\epsilon}\right)$
AGD $O\left(\sqrt{\kappa} \log \frac{f(x_0) - f^*}{\epsilon}\right)$	**ASGD** $\tilde{O}\left(\sqrt{\kappa\tilde{\kappa}} \log \frac{f(x_0) - f^*}{\epsilon} + \frac{\sigma^2 d}{\epsilon}\right)$

- Acceleration possible – depends on statistical condition number
- Techniques: new potential function, stochastic process analysis
- **Conjecture**: Our result is tight

Streaming optimization for ML

- Streaming algorithms are very powerful for ML applications
- SGD and variants widely used in practice
- Classical stochastic approximation focuses on asymptotic rates
- Tools from optimization help obtain strong finite sample guarantees
- Have implications for parallelization as well

Some examples

- Linear regression
  - ➢Finite sample guarantees: Moulines and Bach 2011, Defossez and Bach 2015
  - ➢Parallelization: Jain et al. 2016
  - ➢Acceleration: This talk
- Smooth convex functions:
  - ➢Finite sample guarantees: Bach and Moulines 2013
- PCA: Oja’s algorithm
  - ➢Rank-1: Balsubramani et al. 2013, Jain et al. 2016
  - ➢Higher rank: Allen-Zhu and Li 2016

Open problems

- **Linear regression**: Parameter free algorithm e.g., conjugate gradient
- **General convex functions**: Acceleration, parallelization?
- **Non-convex functions**: Streaming algorithms, acceleration, parallelization?
- **PCA**: Tight finite sample guarantees?
- **Quasi-Newton methods**

Thank you!

Questions?

praneeth@microsoft.com