Introduction
Problem Formulation

We study the least-squares regression problem:

\[
\min_{x \in \mathbb{R}^d} P(x); \quad P(x) \overset{\text{def.}}{=} \frac{1}{2} \mathbb{E} \left[ (b - \langle x, a \rangle)^2 \right]
\] (1)

\[a \in \mathbb{R}^d, \quad b \in \mathbb{R}; \quad a \sim D; \quad b = \langle x^*, a \rangle + \varepsilon; \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \perp a\]

Performance metric: Generalization error

\[
P(x) - P(x^*) = \frac{1}{2} (x - x^*)^T \mathbb{E}[aa^T] (x - x^*) = \frac{1}{2} \|x - x^*\|_H^2
\] (2)

\[
H \overset{\text{def.}}{=} \mathbb{E}[aa^T] = \nabla^2 P(x) \succ 0
\]

Stochastic oracle:

\[x \rightarrow \hat{\nabla} P(x) = a(\langle x, a \rangle - b)\]

We focus on online (streaming) algorithms, which do not require storage of data
Stochastic Gradient Descent: Asymptotic Performance

- SGD with step-size $\gamma/n$: convergence rate $O(1/n)$ (this is also the lower bound)
- Least squares regression: Optimal rate = $\sigma^2 d/n$
  - Optimal rate achieved by SGD in asymptotic limit!
  - Implementation: Constant step-size followed by tail-averaging
    \[
    x_n = x_{n-1} - \alpha \hat{\nabla} P(x_{n-1}); \quad \bar{x}_n = \frac{1}{n-t} \sum_{i=t+1}^{n} x_i
    \]
  - Same rate achieved by
    \[
    \hat{x}_{ERM} = \arg \min \frac{1}{n} \sum_{i} (b_i - \langle x, a_i \rangle)^2
    \]
- Non-asymptotic analysis: How fast does SGD reach the optimal rate?
Non-asymptotic analysis of SGD has been covered extensively in recent work; e.g., Defossez and Bach (2015) [2], Frostig et. al. (2015) [3], and Jain et. al. (2016) [4]

Performance of ‘streaming SVRG’ algorithm [3]:

\[
\mathbb{E}[P(x_n)] - P(x^*) \leq Ce^{-n/\kappa}(P(x_0) - P(x^*)) + \frac{\sigma^2 d}{n} \tag{3}
\]

Hypothesis proposed in ‘Accelerating Stochastic Gradient Descent’ by Jain et. al. (2017) [5]:

can one improve the exponent (to \(n/\sqrt{\kappa}\)) via acceleration?
Intuition behind exponential convergence rate

- SGD with **constant step-size** achieves exponential decay

\[
\mathbb{E} [P(x_n)] - P(x^*) \leq \frac{Lc\alpha \sigma^2}{\mu} + (1 - \mu\alpha)^n(P(x_0) - P(x^*))
\]  

(4)

where \(\mu\) is strong convexity parameter; \(L\) is smoothness parameter (see Thm 4.6 of the review by Bottou(2016) [1])

- Higher rate of convergence obtained by larger step-size

- Maximum step-size \(\sim 1/L\), \(\Rightarrow e^{-n/\kappa}\) convergence rate

**Converging to 0 error**: After error has converged to a constant, one can start the tail-averaging phase, which makes error decrease as \(O(1/n)\)
Experimental results for SGD
Effect of tail-averaging phase

Figure 1: Tail-averaged SGD with step-size $\alpha = 0.02$. $d = 50$, $a \sim \mathcal{N}(0, I)$  
$\Rightarrow L \sim 50$, $\mu = 1$

It is optimal to begin averaging once error of constant step-size converges.
Effect of strong convexity

Figure 2: Tail-averaged SGD with step-size $\alpha = 0.02$. $d = 50$, $\mu = 1, 0.1, 0.01$; $a \sim \mathcal{N}(0, \text{diag}(\mu, \mu, \mu, \mu, \mu, 1, \ldots, 1)) \Rightarrow L \sim 50$

For small values of $\mu$, it is possible to achieve faster convergence compared to theoretical guarantees.
Accelerated SGD
Acceleration in SGD

- We can accelerate GD quite well: Heavy ball, momentum etc.
- Acceleration in SGD is not easy: instability and error accumulation (Paige, 1971; Greenbaum, 1989; Devolder et al., 2014 etc)
- Recent acceleration methods like Katyusha work in the finite sum setting
- Main Question: can we accelerate under the stochastic first order oracle model, using a streaming algorithm?
- That is, given access to only one gradient at a time, can you still reach $O(\sqrt{\kappa})$ samples?
The Bad Example

- Consider \( a = e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \) w.p. \( p_i \).
- Then to optimize, I need to look at all basis vectors and so \( n = \Omega(1/p_{\text{min}}) \)
- Here, condition number is \( \kappa = 1/p_{\text{min}} \) — can’t really accelerate!

But this is really just one really bad corner case...
Accelerated SGD

**Algorithm 1: Accelerated Stochastic Gradient Descent**

**input**: Initial point $x_0 = v_0$, average phase time $t$, Step-size parameters $\alpha, \beta, \gamma, \delta$

for $j = 1, \ldots, n$ do

\[
y_{j-1} \leftarrow \alpha x_{j-1} + (1 - \alpha)v_{j-1};
\]

\[
x_j \leftarrow y_{j-1} - \delta \nabla P(y_{j-1});
\]

\[
z_{j-1} \leftarrow \beta y_{j-1} + (1 - \beta)v_{j-1};
\]

\[
v_j \leftarrow z_{j-1} - \gamma \nabla P(y_{j-1});
\]

**output**: $\bar{x}_{t,n} \leftarrow \frac{1}{n-t} \sum_{j=t+1}^{n} x_j$

**Interpretation:**

$y_j$ is a convex combination of $x_j$ and $v_j$.

$x_j$ are iterates obtained by taking small step-sizes $\delta$.

$v_j$ are the iterates obtained by taking large step-sizes $\gamma$. 
Algorithm 2: Accelerated Stochastic Gradient Descent

input : Initial point $x_0$, parameters $\alpha, \beta, \mu, L$

for $t = 1, \ldots, n$ do

\[
\begin{align*}
    y_t & \leftarrow (1 - \beta)y_{t-1} + \beta(x_t - \frac{1}{\mu} \hat{\nabla} P(x_t)); \\
    x_{t+1} & \leftarrow \alpha(x_t - \frac{1}{L} \hat{\nabla} P(x_t)) + (1 - \alpha)y_t;
\end{align*}
\]

output: $\frac{1}{n-t} \sum_{j=t+1}^{n} x_j$

- Accumulate weighted sum of old steps (if you were aggressive)
- Mix between a conservative step in gradient direction and accumulated aggressive step
Main Theorem

Theorem (Acceleration of ASGD)

Suppose $H = \mathbb{E}[aa^T]$ and fourth moment tensor of $a$ are finite, and $H \succ 0$. Then, for $n > C' \sqrt{\kappa \tilde{\kappa}} \log(d \kappa \tilde{\kappa})$, $t = n/2$,

$$\mathbb{E}[P(\bar{x}_{t,n})] - P(x^*) \leq C \exp \left( -\frac{n}{20 \sqrt{\kappa \tilde{\kappa}}} \right) (P(x_0) - P(x^*)) + 11 \frac{\sigma^2 d}{n},$$

where $C$ is a universal constant, $\kappa, \tilde{\kappa}$ are conditional number and statistical conditional number, and $\sigma^2$ is the noise level.
Terms Defining Convergence Rate

- Noise Level, $\sigma^2$: minimum value such that
  \[ \Sigma \preceq \sigma^2 H, \quad \text{where} \quad \Sigma = \mathbb{E}[\hat{\nabla}P(x^*)P(x^*)^T] \]

- Condition Number, $\kappa = \frac{R^2}{\mu}$, where
  \[ \mu = \lambda_{\min}(H), \quad \mathbb{E}[\|a\|^2 aa^T] \preceq R^2 H \]

  In essence $\kappa \leq \frac{L}{\mu}$, $\|a\| \leq L$.

- Statistical Conditional Number, $\tilde{\kappa}$: minimum value such that
  \[ \mathbb{E}[\|a\|^2_{H^{-1}} aa^T] \preceq \tilde{\kappa} H. \]

  In essence, $\tilde{\kappa} \leq \max \|a\|^2_{H^{-1}}$. 
How many samples are needed to (spectrally) approximate Hessian $H$, i.e., $\frac{1}{c} H \preceq \frac{1}{n} \sum_{i=1}^{n} a_i a_i^T \preceq cH$? — $\tilde{\kappa}$ governs this number!

Why is that important?

And,

$E[\|a\|_2^2 H^{-1} a a^T] \preceq E[\|a\|_2^2 a a^T] \preceq \kappa H$

Thus, $\tilde{\kappa} \leq \kappa$!

Examples,

1. Discrete: $\kappa = \min_i p_i = \tilde{\kappa}$ – No Acceleration!

2. Gaussian: $\kappa = O(\text{tr}(H)/\mu)$, $\tilde{\kappa} = O(d)$ – Considerable Acceleration feasible!
How many samples are needed to (spectrally) approximate Hessian $H$, i.e., $\frac{1}{c}H \preceq \frac{1}{n} \sum_{i=1}^{n} a_i a_i^T \preceq cH$? — $\tilde{\kappa}$ governs this number!

Why is that important? $Hx^* = \mathbb{E}[ba]$.

And,

$$\mathbb{E}[\|a\|_{H^{-1}}^2 aa^T] \leq \frac{1}{\mu} \mathbb{E}[\|a\|^2 aa^T] \leq \kappa H$$

Thus, $\tilde{\kappa} \leq \kappa$!
How many samples are needed to (spectrally) approximate Hessian $H$, i.e., $\frac{1}{c} H \preceq \frac{1}{n} \sum_{i=1}^{n} a_i a_i^T \preceq cH$? — $\tilde{\kappa}$ governs this number!

Why is that important? $Hx^* = \mathbb{E}[ba]$.

And,

$$\mathbb{E}[\|a\|_H^2 aa^T] \leq \frac{1}{\mu} \mathbb{E}[\|a\|^2 aa^T] \leq \kappa H$$

Thus, $\tilde{\kappa} \leq \kappa$!

Examples,

1. Discrete: $\kappa = \frac{1}{\min_i p_i} = \tilde{\kappa} –$ No Acceleration!

2. Gaussian: $\kappa = O(\text{tr}(H)/\mu), \quad \tilde{\kappa} = O(d) –$ Considerable Acceleration feasible!
Theoretical Analysis
Let deviation vector be $\theta_j = \begin{bmatrix} x_j - x^* \\ y_j - x^* \end{bmatrix}$

ASGD update: $\theta_j = \hat{A}_j \theta_{j-1} + \zeta_j$, where

$$\hat{A}_j = \begin{bmatrix} 0 & (I - \delta a_j a_j^T) \\ -\alpha \bar{\beta} I & (1 + \alpha \bar{\beta}) I - (\alpha \delta + \bar{\alpha} \gamma) a_j a_j^T \end{bmatrix}, \quad \zeta_j = \begin{bmatrix} \delta \hat{\nabla} P(x_j) \\ (\alpha \delta + \bar{\alpha} \gamma) \hat{\nabla} P(x_j) \end{bmatrix}$$

Treat as a linear dynamical system where $\zeta_j$ represents the noise term.
Bias-variance Decomposition

- Decompose $\theta_j = \theta_j^{\text{bias}} + \theta_j^{\text{var}}$.
- Bias: noiseless evolution of the system
  \[ \theta_j^{\text{bias}} = \hat{A}_j \theta_{j-1}^{\text{bias}}, \quad \theta_0^{\text{bias}} = \theta_0. \]
- Variance: Progression of pure noise through the system
  \[ \theta_j^{\text{var}} = \hat{A}_j \theta_{j-1}^{\text{var}} + \zeta_j, \quad \theta_0^{\text{var}} = 0. \]
- Generalization error can be evaluated in terms of covariance of bias and variance components
  \[ \mathbb{E}[P(\bar{x}_{t,n})] - P(x^*) \leq \left\langle \begin{bmatrix} H & 0 \\ 0 & 0 \end{bmatrix}, \mathbb{E}[\bar{\theta}_t^{\text{bias}} \bar{\theta}_t^{\text{bias}}, T] + \mathbb{E}[\bar{\theta}_t^{\text{var}} \bar{\theta}_t^{\text{var}}, T] \right\rangle \]
Bias Contraction - the linear rate

- Recursively compute the covariance matrix of $\theta^t_{bias}$.
- Establish contraction in bias term:

$$\langle G, \mathbb{E}[\hat{A}_j \theta \theta^T \hat{A}_j^T] \rangle \leq \left(1 - \frac{1}{9\sqrt{\tilde{\kappa}}}\right) \langle G, \theta \theta^T \rangle \leq \exp \left(-\frac{1}{9\sqrt{\tilde{\kappa}}}\right) \langle G, \theta \theta^T \rangle,$$

where

$$G = \begin{bmatrix} I & \frac{-\mu \alpha}{\alpha^2} H^{-1} \\ \frac{-\mu \alpha}{\alpha^2} H^{-1} & \frac{\mu}{\alpha^2} H^{-1} \end{bmatrix}$$

- To prove claim, consider the potential function

$$\|x - x^*\|_2^2 + \mu \|v - x^*\|_{H^{-1}}^2$$

(instead of $\|x - x^*\|_2^2 + \mu \|v - x^*\|_{H}^2$) and show contraction with iterations.
An Intuitive Understanding of Variance

- SGD is only as accurate as the per-iteration noise in the system.
- Example, consider $b = a^T x + Z$, where $Z \sim \mathcal{N}(0, \sigma^2 I)$. Then, generalization error in conventional SGD is $\sigma^2 d$.
- Tail averaging $\implies$ reduction in noise in the estimate - $\frac{\sigma^2 d}{n}$.
- Here same effect as observed in tail-averaged SGD (Jain et al., 2016) and streaming SVRG (Frostig et al., 2015b).
Experimental Results of Acceleration of SGD
Figure 3: Comparison of SGD and ASGD: figure from paper ($\kappa = 10^5$, $\tilde{\kappa} = 50$)

Is there scope to improve the performance of SGD by changing step-size or averaging phase?
Accelerated SGD: Our simulations

Figure 4: Comparison of SGD and ASGD: our simulations

Parameters: $\mu = 0.001$, $d = 50$; $a \sim \mathcal{N}(0, \text{diag}(\mu, \mu, \mu, \mu, \mu, 1, \ldots, 1))$
Conclusion

- Accelerating streaming SGD is highly non-trivial.
- We were able to gain a clearer understanding of the analysis and acceleration tools used for streaming SGD.
- Constructing explicit methods require a simpler problem framework like the least-squares, and fairly specific and difficult analysis techniques.
- Can we achieve better performance by slightly increasing memory?
- Is there a broader, more unified analysis framework?
L. Bottou, F. E. Curtis, and J. Nocedal.  
**Optimization methods for large-scale machine learning.**  

A. Défossez and F. Bach.  
**Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions.**  

R. Frostig, R. Ge, S. M. Kakade, and A. Sidford.  
**Competing with the empirical risk minimizer in a single pass.**  

The Connection to the ODE Method

• Reformulating the recursive expression

\[ \theta_{j+1} = \theta_j + \eta(h(\theta_j) + M(\theta_j, W_{j+1})) , \]

where \( h \) is a simple linear map, and \( M \) is a first order Markov noise.

• Borkar and Meyn (2000) establish rate of convergence result for the stochastic approximation problem:

\[ \mathbb{P}[\|x_n - x^*\|_2^2 \geq \epsilon | X_0 = x_0] \leq A_1(\eta) + A_2(1 + \|x_0 - x^*\|_2^2) \exp(-\epsilon(\eta)n) , \]

where \( A_1(\eta) \to 0, \epsilon(\eta) \to 0 \) for \( \eta \to 0 \).