Distributed Optimization for Machine Learning

Lecture 9 - Variance reduction and momentum

Tianyi Chen

School of Electrical and Computer Engineering
Cornell Tech, Cornell University

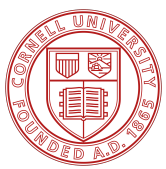
September 24, 2025

Recall stochastic programming

We view both the model training and testing problems as

$$\min_{\mathbf{x}} \quad \underbrace{F(\mathbf{x}) = \mathbb{E}[f(\mathbf{x}; \xi)]}_{\text{expected risk, popular risk, ...}}$$

- ξ: training or testing data in ML problems
- suppose $f(\cdot; \xi)$ is convex for every ξ (and hence $F(\cdot)$ is convex)



Example: empirical risk minimization

Let $\{\mathbf{a}_i, y_i\}_{i=1}^n$ be n random samples, and consider

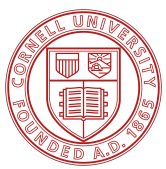
$$\min_{\mathbf{x}} \underbrace{F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}; \{\mathbf{a}_i, y_i\})}_{\text{empirical risk}}$$

e.g. quadratic loss $f(\mathbf{x}; \{\mathbf{a}_i, y_i\}) = (\mathbf{a}_i^\top \mathbf{x} - y_i)^2$.

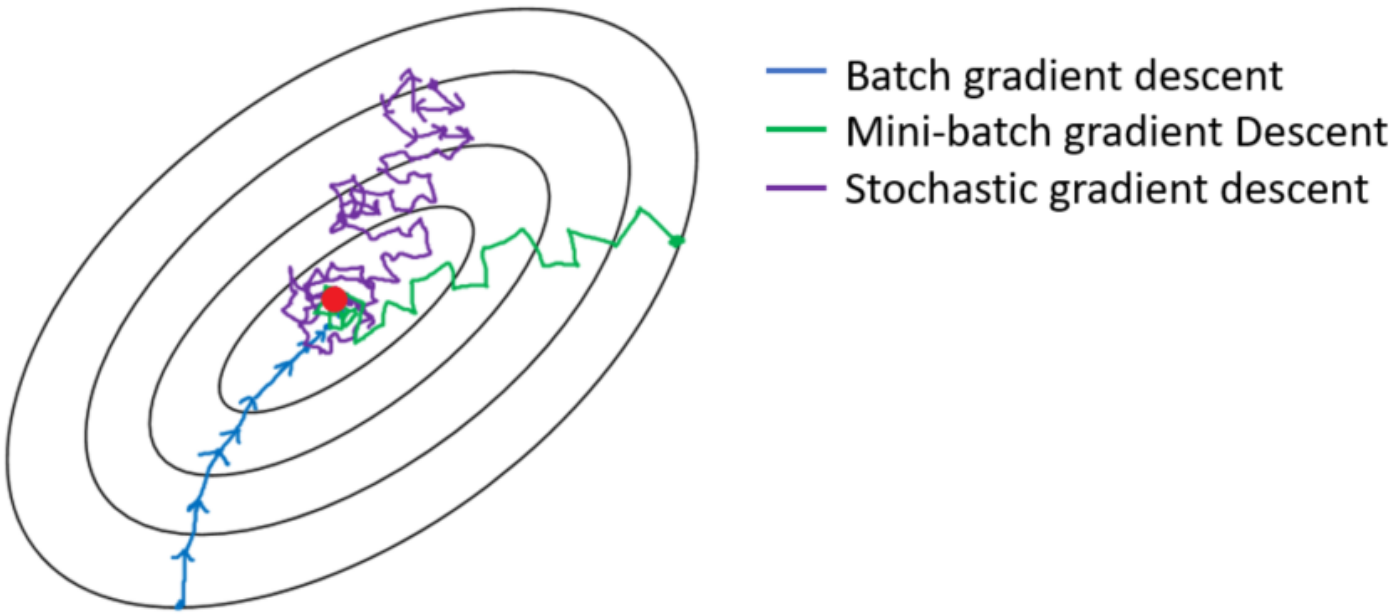
If one draws index $j \sim \text{Unif}(1, \ldots, n)$ uniformly at random, then

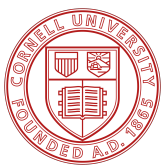
$$F(\mathbf{x}) = \mathbb{E}_j[f(\mathbf{x}; \{\mathbf{a}_j, y_j\})]$$

which is the model training problem we aim to tackle.

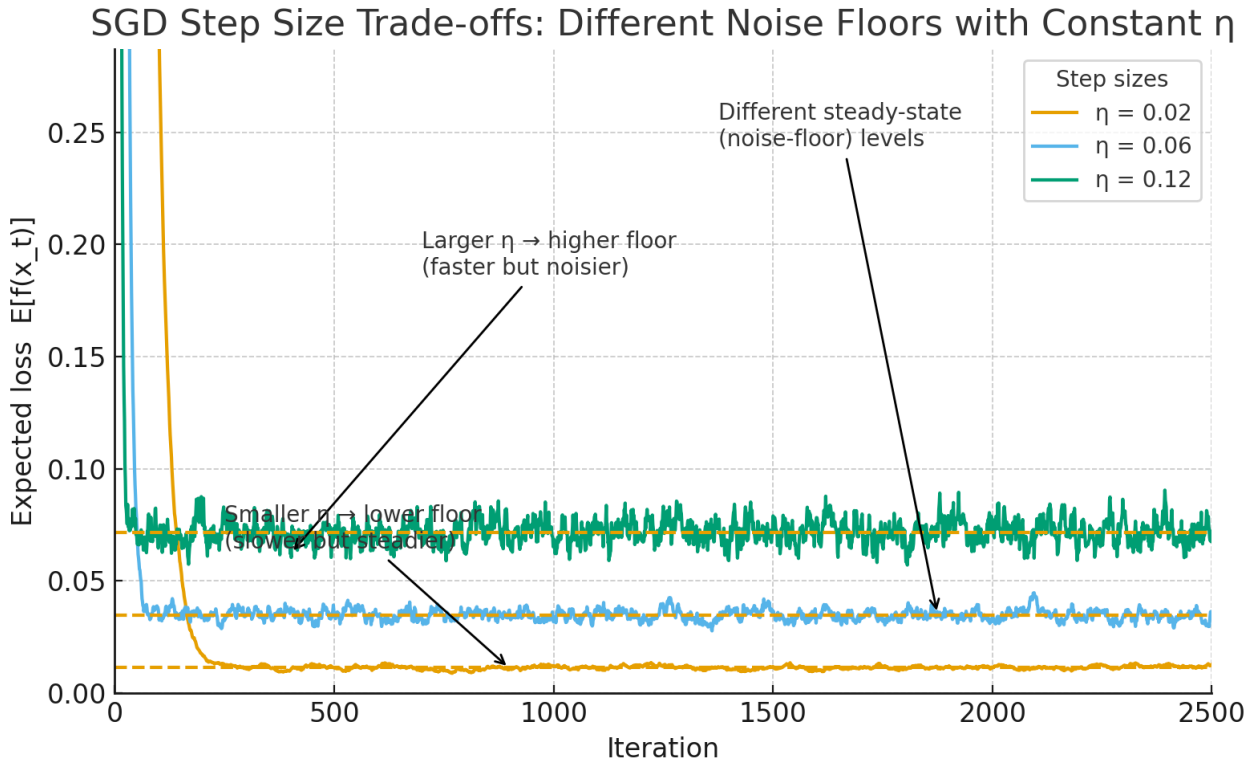


Compare GD, SGD and mini-batch SGD trajectories

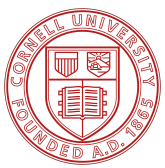




Tradeoffs in large and small stepsizes



Small η converges slowly to a low floor, moderate η balances speed and floor, large η drops fast but stabilizes at a higher floor.



Convergence with diminishing stepsizes

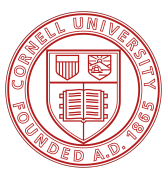
Theorem 2 (Strong convexity and diminishing stepsizes)

Suppose F is μ-strongly convex, and (2) holds with $c_g = 0$. If $\eta_t = \frac{\theta}{t+1}$ for some $\theta > \frac{1}{2\mu}$, then SGD (1) achieves

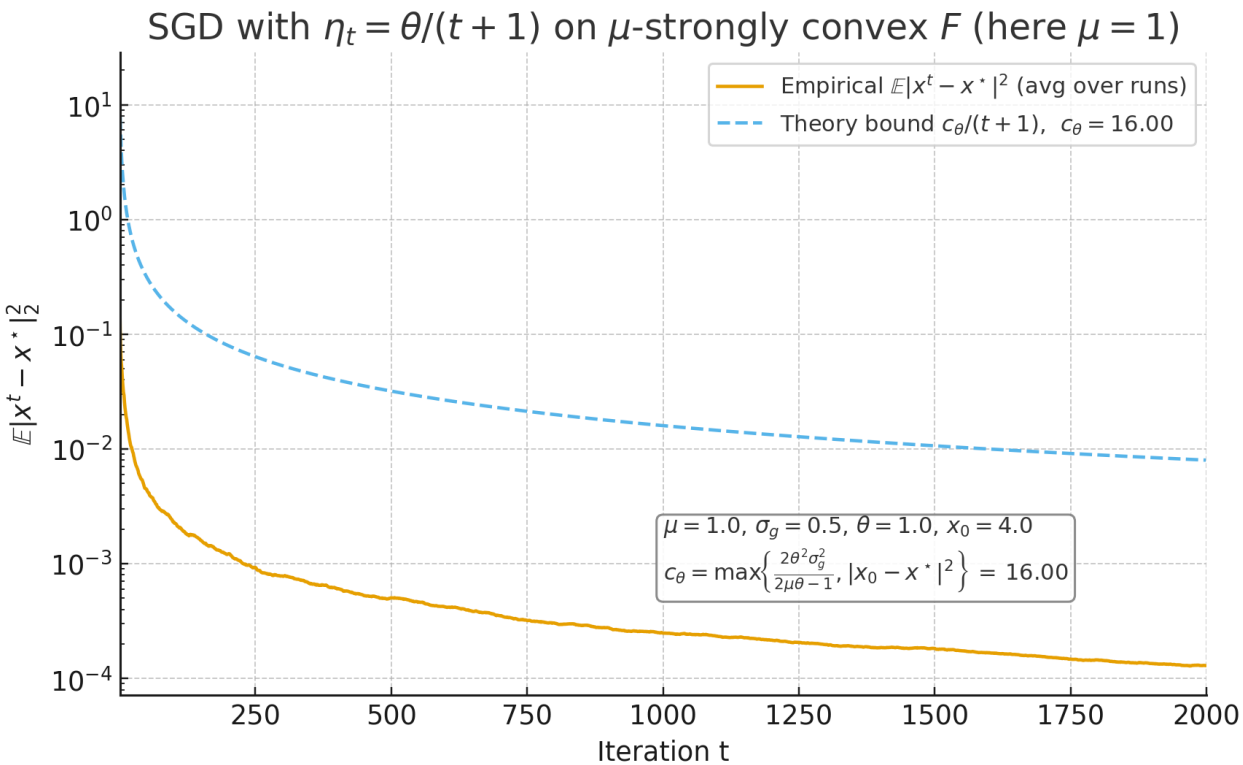
$$\mathbb{E}[||\mathbf{x}^t - \mathbf{x}^*||_2^2] \leq \frac{c_\theta}{t+1}$$

where $c_\theta = \max\left\{\frac{2\theta^2\sigma_g^2}{2\mu\theta - 1}, ||\mathbf{x}_0 - \mathbf{x}^*||_2^2\right\}$.

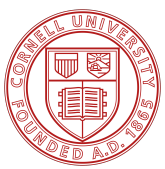
- convergence rate $\mathcal{O}(1/t)$ with diminishing stepsize $\eta_t \approx 1/t$



Simulations of SGD with diminishing stepsizes



The dashed curve is the theoretical $\frac{c_\theta}{t+1}$ bound; the solid line is the empirical mean-squared error.



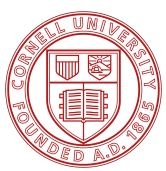
Proof of Theorem 2

Using the SGD update rule, we have (compare with GD proof steps)

$$\begin{aligned}||\mathbf{x}^{t+1} - \mathbf{x}^*||_2^2 &= ||\mathbf{x}^t - \eta_t g(\mathbf{x}^t; \xi^t) - \mathbf{x}^* (+\eta_t g(\mathbf{x}^*; \xi^t))?||_2^2 \\ &= ||\mathbf{x}^t - \mathbf{x}^*||_2^2 - 2\eta_t (\mathbf{x}^t - \mathbf{x}^*)^\top g(\mathbf{x}^t; \xi^t) + \eta_t^2 ||g(\mathbf{x}^t; \xi^t)||_2^2 \quad (\star)\end{aligned}$$

Since $\mathbf{x}^t$ is independent of ξ_t, apply the law of total expectation to obtain

$$\begin{aligned}\mathbb{E}[(\mathbf{x}^t - \mathbf{x}^*)^\top g(\mathbf{x}^t; \xi^t)] &= \mathbb{E}[\mathbb{E}[(\mathbf{x}^t - \mathbf{x}^*)^\top g(\mathbf{x}^t; \xi^t) | \xi_1, \ldots, \xi_{t-1}]] \\ &= \mathbb{E}[(\mathbf{x}^t - \mathbf{x}^*)^\top \mathbb{E}[g(\mathbf{x}^t; \xi^t) | \xi_1, \ldots, \xi_{t-1}]] \\ &= \mathbb{E}[(\mathbf{x}^t - \mathbf{x}^*)^\top \nabla F(\mathbf{x}^t)] \quad (\diamond)\end{aligned}$$



Proof of Theorem 2 (cont.)

Furthermore, strong convexity gives

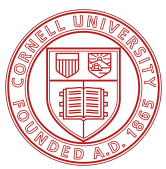
$$\langle \nabla F(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^* \rangle = \langle \nabla F(\mathbf{x}^t) - \underbrace{\nabla F(\mathbf{x}^*)}_{\mathbf{0}}, \mathbf{x}^t - \mathbf{x}^* \rangle \geq \mu ||\mathbf{x}^t - \mathbf{x}^*||_2^2$$

$$\implies \mathbb{E}[\langle \nabla F(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^* \rangle] \geq \mu \mathbb{E}[||\mathbf{x}^t - \mathbf{x}^*||_2^2]$$

Combine the above inequalities and (2) (with $c_g = 0$) to obtain

$$\mathbb{E}[||\mathbf{x}^{t+1} - \mathbf{x}^*||_2^2] \leq (1 - 2\mu\eta_t)\mathbb{E}[||\mathbf{x}^t - \mathbf{x}^*||_2^2] + \underbrace{\eta_t^2 \sigma_g^2}_{\text{does not vanish unless } \eta_t \to 0}$$

Take $\eta_t = \frac{\theta}{t+1}$ and use induction to conclude the proof (exercise!)



Optimality*

Whether $\mathcal{O}(1/t)$ convergence is the best we can hope for?

- Informally, when minimizing strongly convex functions, no algorithm performing t queries to noisy first-order oracles can achieve an accuracy better than the order of $1/t$.

  $\Rightarrow$ SGD with stepsizes $\eta_t \approx 1/t$ is optimal.

— *Nemirovski, Yudin '83, Agarwal et al. '11, Raginsky, Rakhlin '11*

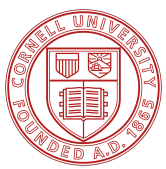
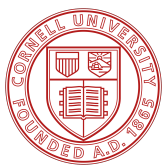


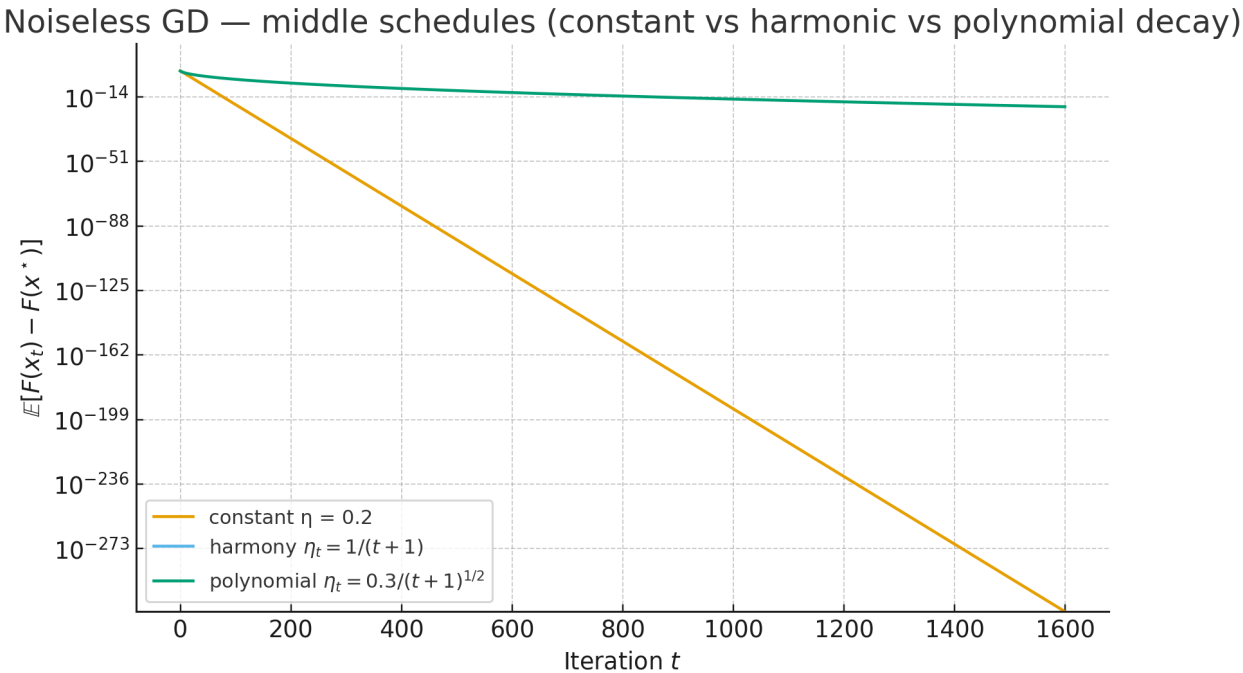
Table of Contents

Reducing variance via averaging

Heavy-ball type momentum methods

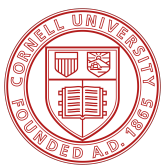


Stepsize choice $O(1/t)$?

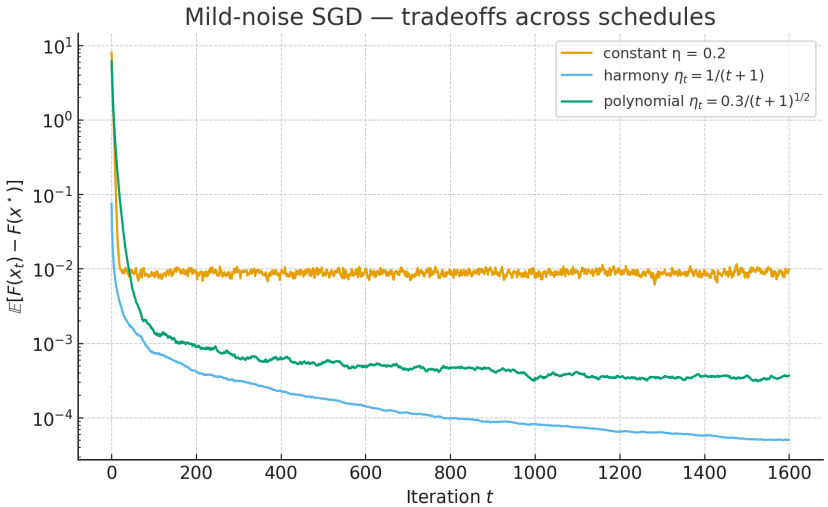


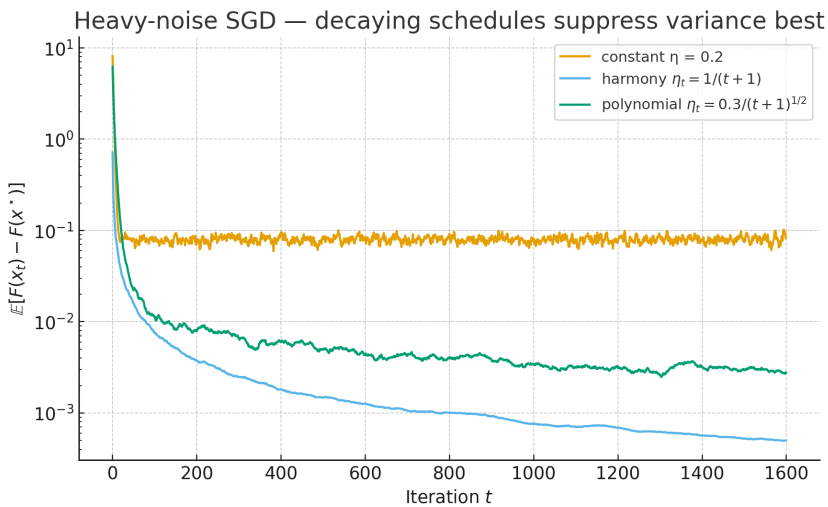
Two conflicting regimes

- the noiseless case (i.e. $g(\mathbf{x};\xi) = \nabla F(\mathbf{x})$): stepsizes $\eta_t \approx 1/t$ are way too conservative



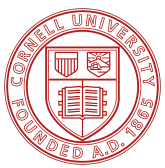
Stepsize choice $O(1/t)$?





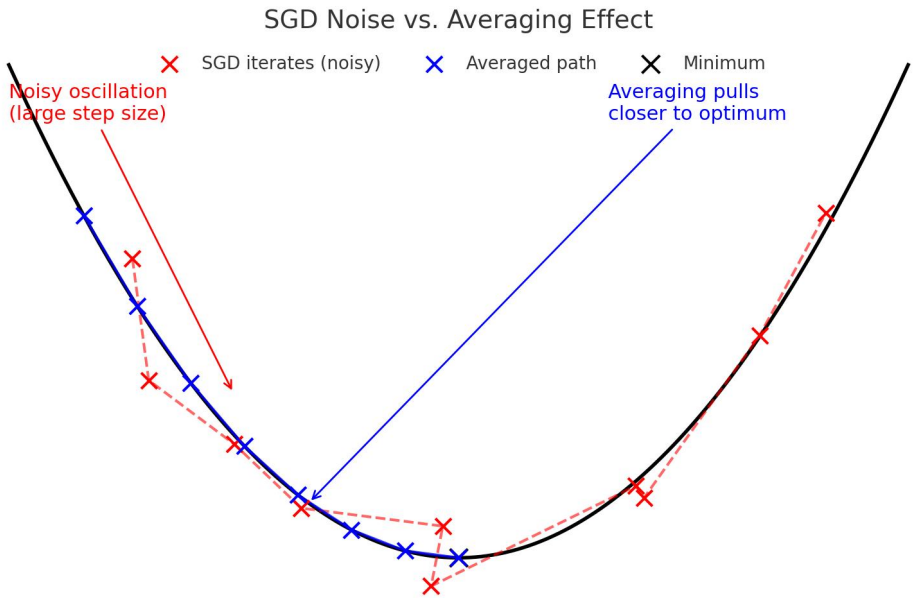
Two conflicting regimes

- the general noisy case: longer stepsizes ($\eta_t \gg 1/t$) might fail to suppress noise (and hence slow down convergence)

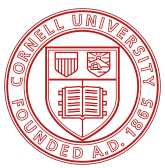


Motivation for iterate averaging

SGD with long stepsizes poorly suppresses noise, which tends to oscillate around the global minimizers due to the noisy gradient.



May average iterates to mitigate oscillation and reduce variance.



Acceleration by averaging the iterates

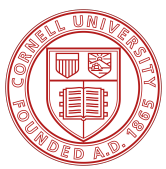
—*Ruppert '88, Polyak '90, Polyak, Juditsky '92*

Iterate averaging returns

$$\bar{\mathbf{x}}^t := \frac{1}{t}\sum_{i=0}^{t-1} \mathbf{x}^i$$

with larger stepsizes $\eta_t = t^{-\alpha}$, $\alpha < 1$.

Key idea: average the iterates (as the final output) to reduce variance and improve convergence.



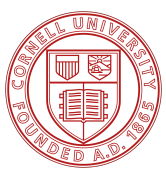
Example: a toy quadratic minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \ \frac{1}{2}||\mathbf{x}||_2^2$$

- constant stepsizes: $\eta_t \equiv \eta < 1$
- $g(\mathbf{x}^t;\xi^t) = \mathbf{x}^t + \xi^t$ with
  $\Rightarrow \mathbb{E}[\xi^t|\xi^0,\ldots,\xi^{t-1}] = 0$ and $\mathbb{E}[\xi^t(\xi^t)^\top|\xi^0,\ldots,\xi^{t-1}] = \mathbf{I}$

SGD iterates:

$$\begin{aligned}
\mathbf{x}^1 &= \mathbf{x}^0 - \eta(\mathbf{x}^0+\xi^0) = (1-\eta)\mathbf{x}^0 - \eta\xi^0 \\
\mathbf{x}^2 &= \mathbf{x}^1 - \eta(\mathbf{x}^1+\xi^1) = (1-\eta)^2\mathbf{x}^0 - \eta(1-\eta)\xi^0 - \eta\xi^1 \\
&\vdots \\
\mathbf{x}^t &= (1-\eta)^t\mathbf{x}^0 - \eta(1-\eta)^{t-1}\xi^0 - \eta(1-\eta)^{t-2}\xi^1 - \cdots
\end{aligned}$$



Example: a toy quadratic minimization problem

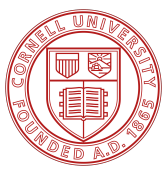
$$\min_{\mathbf{x}\in\mathbb{R}^d} \ \frac{1}{2}||\mathbf{x}||_2^2$$

Iterate averaging gives

$$\bar{\mathbf{x}}^t \approx \underbrace{\frac{1}{t}\sum_{k=0}^{t-1}(1-\eta)^k\mathbf{x}^0}_{=\frac{1}{t}\frac{1-(1-\eta)^t}{\eta}\mathbf{x}^0 \xrightarrow{t\to\infty} 0} \underbrace{-\eta\{1+(1-\eta)+\dots\}\frac{1}{t}\sum_{k=0}^{t-1}\xi^k}_{\text{imprecise; but close enough for large } t}$$

$$= -\frac{1}{t}\sum_{k=0}^{t-1}\xi^k \qquad \text{(since } 1+(1-\eta)+\cdots=\eta^{-1}\text{)}$$

$$\xrightarrow{t\to\infty} \frac{1}{\sqrt{t}}\mathcal{N}(0,\mathbf{I}) \qquad \text{(the central limit theorem)}$$



Last iterate vs. Averaged iterates in SGD

Last iterate. $\mathbf{x}^t = (1-\eta)^t \mathbf{x}^0 - \eta \sum_{k=0}^{t-1} (1-\eta)^{t-1-k} \xi^k$

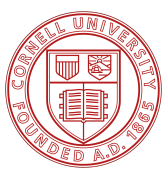
$$\lim_{t\to\infty} \mathbb{E}\|\mathbf{x}^t\|^2 = \frac{\eta}{2-\eta} \qquad \Rightarrow \text{ with } \eta = 1, \text{ variance floor } \mathcal{O}(1).$$

Averaged iterate. $\bar{\mathbf{x}}^t = \frac{1}{t}\sum_{j=0}^{t-1} \mathbf{x}^j \;\approx\; -\frac{1}{t}\sum_{k=0}^{t-1} \xi^k$

$$\sqrt{t}\,\bar{\mathbf{x}}^t \xrightarrow{d} \mathcal{N}(0, \mathbf{I}) \quad \Rightarrow \quad \mathbb{E}\|\bar{\mathbf{x}}^t\|^2 \approx \tfrac{d}{t}.$$

Takeaway:

- Last iterate: variance $\approx \mathcal{O}(1)$ (does not vanish).
- Averaged iterate: variance $\approx \mathcal{O}(1/t)$ (vanishes).



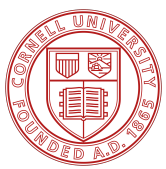
Example: more general quadratic problems

$$\min_{\mathbf{x}} \ \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$$

- $\mathbf{A} \succeq \mu \mathbf{I} \succ 0$ (strongly convex)
- constant stepsizes: $\eta_t \equiv \eta < 1/\mu$
- $g(\mathbf{x}^t; \xi^t) = \mathbf{A}\mathbf{x}^t - \mathbf{b} + \xi^t$ with

  $\Rightarrow \mathbb{E}[\xi^t | \xi^0, \ldots, \xi^{t-1}] = 0$

  $\Rightarrow \mathbf{S} := \lim_{t\to\infty} \mathbb{E}[\xi^t (\xi^t)^\top | \xi^0, \ldots, \xi^{t-1}]$ is finite



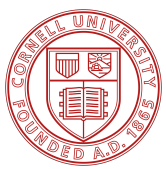
Example: more general quadratic problems

$$\min_{\mathbf{x}} \ \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$$

Theorem 4 Fix d. Then as $t \to \infty$, the iterate average $\bar{\mathbf{x}}^t$ obeys

$$\sqrt{t}(\bar{\mathbf{x}}^t - \mathbf{x}^*) \xrightarrow{D} \mathcal{N}(0, \mathbf{A}^{-1}\mathbf{S}\mathbf{A}^{-1})$$

where $\xrightarrow{D}$ denotes convergence in distribution.

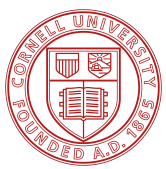


Example: quadratic problems

$$\sqrt{t}(\bar{\mathbf{x}}^t - \mathbf{x}^*) \xrightarrow{D} \mathcal{N}(0, \mathbf{A}^{-1}\mathbf{S}\mathbf{A}^{-1}), \quad t \to \infty$$

- asymptotically, $||\bar{\mathbf{x}}^t - \mathbf{x}^*||_2^2 \asymp 1/t$, matching the rate in Theorem 2
- much longer stepsizes ($\eta_t \not\approx 1/t$)

  $\Rightarrow$ faster convergence for less noisy cases (e.g. $\xi^t = 0$)



Proof sketch of Theorem 4*

(1) Let $\Delta^t = \mathbf{x}^t - \mathbf{x}^*$ and $\bar{\Delta}^t = \bar{\mathbf{x}}^t - \mathbf{x}^*$. SGD update rule gives

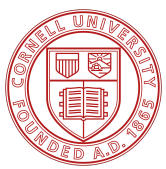
$$\Delta^{t+1} = \Delta^t - \eta(\mathbf{A}\Delta^t + \xi^t) = (\mathbf{I} - \eta\mathbf{A})\Delta^t - \eta\xi^t$$

$$\Delta^{t+1} = (\mathbf{I} - \eta\mathbf{A})^{t+1}\Delta^0 - \eta\sum_{k=0}^{t}(\mathbf{I} - \eta\mathbf{A})^{t-k}\xi^k$$

(2) Simple calculation gives (check Polyak, Juditsky '92)

$$\bar{\Delta}^t = \frac{1}{t\eta}G_0^t\Delta^0 + \frac{1}{t}\sum_{j=0}^{t-2}\mathbf{A}^{-1}\xi^j + \frac{1}{t}\sum_{j=0}^{t-2}(G_j^t - \mathbf{A}^{-1})\xi^j$$

where $G_j^t := \eta\sum_{i=0}^{t-1-j}(\mathbf{I} - \eta\mathbf{A})^i$.



Proof sketch of Theorem 4 (cont.)*

(3) From the central limit theorem for martingales,

$$\frac{1}{\sqrt{t}} \sum_{j=0}^{t-2} \mathbf{A}^{-1} \xi^j \xrightarrow{D} \mathcal{N}(0, \mathbf{A}^{-1} \mathbf{S} \mathbf{A}^{-1})$$

(4) With proper stepsizes, one has (check Polyak, Juditsky '92)

$$||G_0^t|| < \infty, \quad ||G_j^t - \mathbf{A}^{-1}|| < \infty \quad \text{and} \quad \lim_{t \to \infty} \frac{1}{t} \sum_{j=0}^{t-1} ||G_j^t - \mathbf{A}^{-1}||^2 = 0$$

(5) Combining these bounds establishes Theorem 4

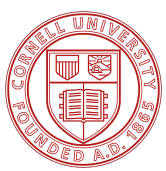
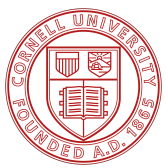


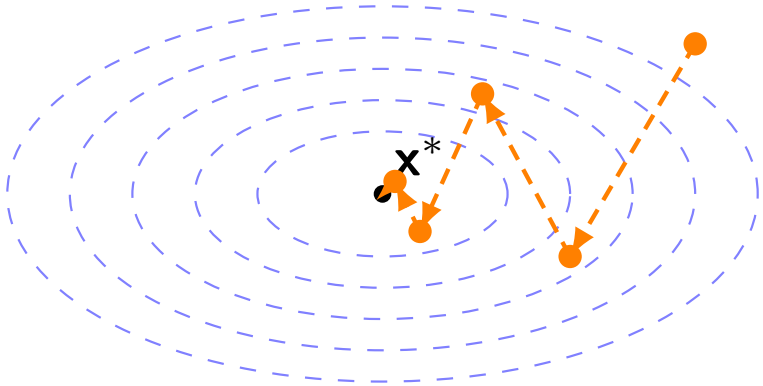
Table of Contents

Reducing variance via averaging

Heavy-ball type momentum methods



Why we want momentum for GD?

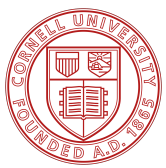


gradient descent

Iteration complexities of (projected) gradient methods under strongly convex and smooth problems (κ can be large)

$$\mathcal{O}\left(\kappa \log \frac{1}{\epsilon}\right)$$

Can one still hope to further accelerate convergence?



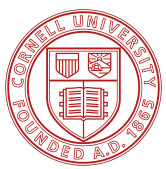
Issues of GD and possible solutions

Issues:

- GD focuses on improving the cost per iteration, which might sometimes be too "short-sighted"
- GD might sometimes zigzag or experience abrupt changes

Solutions:

- exploit information from the history (i.e. past iterates)
- add buffers (like momentum) to yield smoother trajectory



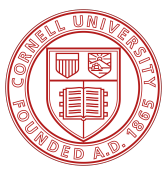
Heavy-ball or Polyak's momentum method

B. Polyak

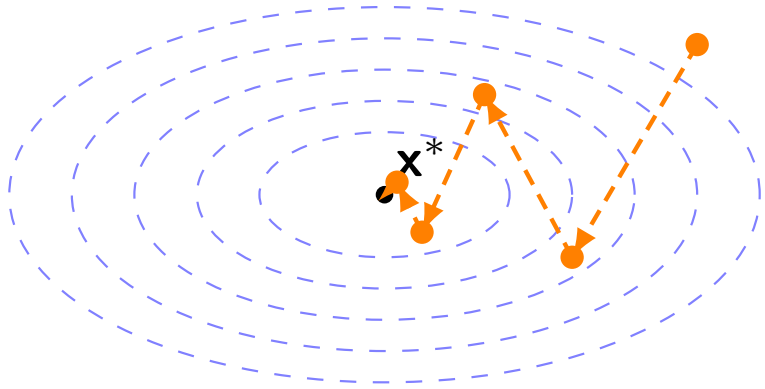
$$\text{minimize}_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x})$$

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) + \underbrace{\theta_t(\mathbf{x}^t - \mathbf{x}^{t-1})}_{\text{momentum term}}$$

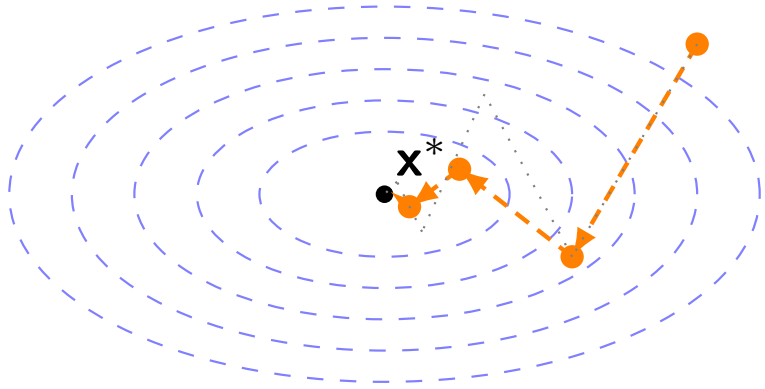
- add inertia to the "ball" (i.e., a momentum) to mitigate zigzagging



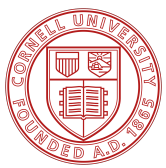
Heavy-ball vs Gradient descent method



gradient descent



heavy-ball method



Gradient descent with Polyak's momentum

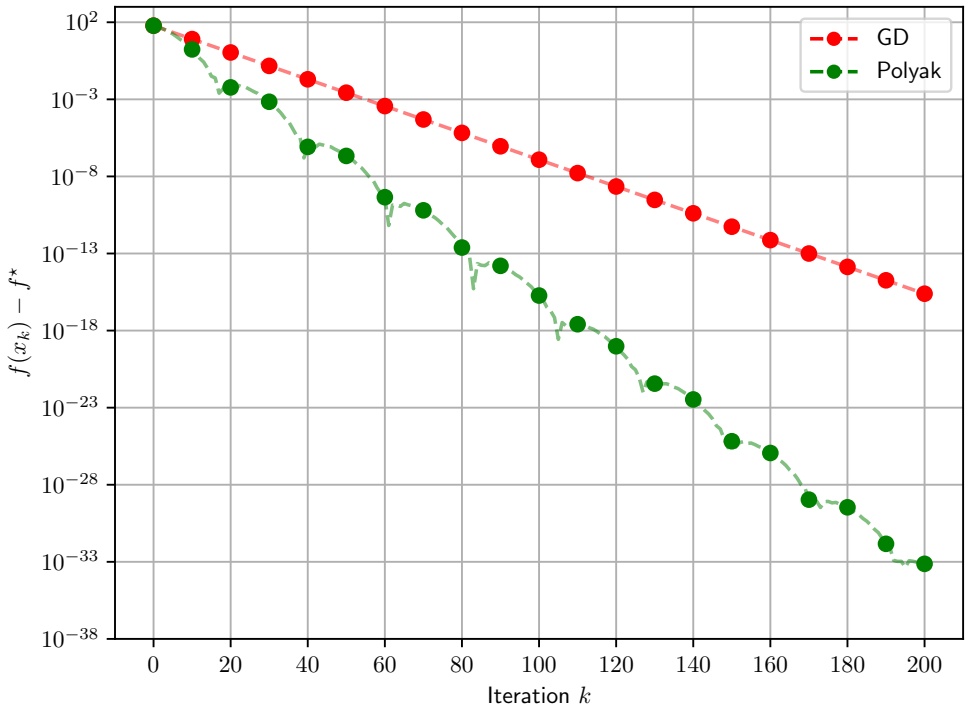
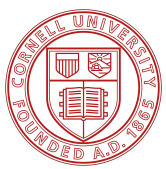


Figure: GD v.s. Polayk momentum

How to theoretically explain this phenomenon?



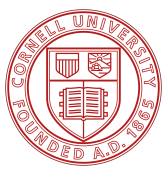
Heavy-ball methods for quadratic minimization

Consider the quadratic minimization problem

$$\text{minimize}_{\mathbf{x}} \quad \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^\top \mathbf{Q}(\mathbf{x}-\mathbf{x}^*)$$

where $\mathbf{Q} \succ 0$ has a condition number κ

One can understand heavy-ball methods through dynamical systems



Heavy-ball method as a linear dynamical system

Heavy-ball update rule:

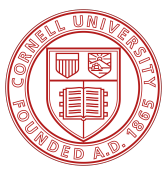
$$\mathbf{x}^{t+1} = \mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) + \theta_t(\mathbf{x}^t - \mathbf{x}^{t-1})$$

Equivalent augmented state representation:

$$\begin{bmatrix} \mathbf{x}^{t+1} \\ \mathbf{x}^t \end{bmatrix} = \begin{bmatrix} (1+\theta_t)I & -\theta_t I \\ I & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}^t \\ \mathbf{x}^{t-1} \end{bmatrix} - \begin{bmatrix} \eta_t \nabla f(\mathbf{x}^t) \\ 0 \end{bmatrix}.$$

Interpretation:

- Dynamics as a $2d$-dimensional linear system with input $-\eta_t \nabla f(\mathbf{x}^t)$.
- Matrix captures the *momentum carryover* between $(\mathbf{x}^t, \mathbf{x}^{t-1})$.
- Useful for analyzing stability via spectral radius.



Heavy-ball method as a linear dynamical system

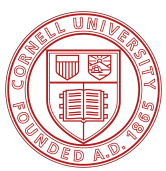
Consider the following dynamical system

$$\begin{bmatrix} \mathbf{x}^{t+1} \\ \mathbf{x}^{t} \end{bmatrix} = \begin{bmatrix} (1+\theta_t)I & -\theta_t I \\ I & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}^{t} \\ \mathbf{x}^{t-1} \end{bmatrix} - \begin{bmatrix} \eta_t \nabla f(\mathbf{x}^t) \\ 0 \end{bmatrix}$$

or equivalently,

$$\underbrace{\begin{bmatrix} \mathbf{x}^{t+1} - \mathbf{x}^* \\ \mathbf{x}^{t} - \mathbf{x}^* \end{bmatrix}}_{\text{state}} = \begin{bmatrix} (1+\theta_t)I & -\theta_t I \\ I & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}^{t} - \mathbf{x}^* \\ \mathbf{x}^{t-1} - \mathbf{x}^* \end{bmatrix} - \begin{bmatrix} \eta_t \nabla f(\mathbf{x}^t) \\ 0 \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} (1+\theta_t)I - \eta_t \mathbf{Q} & -\theta_t I \\ I & 0 \end{bmatrix}}_{\text{system matrix}} \begin{bmatrix} \mathbf{x}^{t} - \mathbf{x}^* \\ \mathbf{x}^{t-1} - \mathbf{x}^* \end{bmatrix}$$

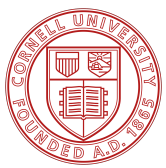


System matrix

$$\begin{bmatrix} \mathbf{x}^{t+1} - \mathbf{x}^* \\ \mathbf{x}^t - \mathbf{x}^* \end{bmatrix} = \underbrace{\begin{bmatrix} (1+\theta_t)I - \eta_t \mathbf{Q} & -\theta_t I \\ I & 0 \end{bmatrix}}_{=:\mathbf{H}_t \text{ (system matrix)}} \begin{bmatrix} \mathbf{x}^t - \mathbf{x}^* \\ \mathbf{x}^{t-1} - \mathbf{x}^* \end{bmatrix} \tag{1}$$

Implication: convergence of heavy-ball methods depends on the spectrum of the system matrix $\mathbf{H}_t$

Key idea: find appropriate stepsizes η_t and momentum coefficients θ_t to control the spectrum of $\mathbf{H}_t$



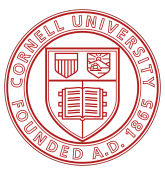
Convergence for quadratic problems

Theorem 1 (Convergence for quadratic functions)

Suppose f is an L-smooth and μ-strongly convex quadratic function. Set $\eta_t = 4/(\sqrt{L}+\sqrt{\mu})^2$, $\theta_t = \max\{|1-\sqrt{\eta_t L}|, |1-\sqrt{\eta_t \mu}|\}^2$, and $\kappa = L/\mu$. Then

$$\left\| \begin{bmatrix} \mathbf{x}^{t+1} - \mathbf{x}^* \\ \mathbf{x}^t - \mathbf{x}^* \end{bmatrix} \right\|_2 \approx \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^t \left\| \begin{bmatrix} \mathbf{x}^1 - \mathbf{x}^* \\ \mathbf{x}^0 - \mathbf{x}^* \end{bmatrix} \right\|_2$$

- iteration complexity: $O(\sqrt{\kappa}\log\frac{1}{\varepsilon})$
- significant improvement over GD: $O(\sqrt{\kappa}\log\frac{1}{\varepsilon})$ vs. $O(\kappa\log\frac{1}{\varepsilon})$
- relies on knowledge of both L and μ



Proof of Theorem 1*

In view of (1), it suffices to control the spectrum of $\mathbf{H}_t$ (which is time-invariant). Let λ_i be the ith eigenvalue of $\mathbf{Q}$ and set

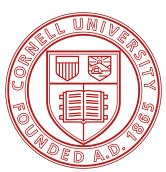
$$\mathbf{\Lambda} := \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

then the spectral radius (denoted by $\rho(\cdot)$) of $\mathbf{H}_t$ obeys

$$\begin{aligned} \rho(\mathbf{H}_t) &= \rho\left(\begin{bmatrix} (1+\theta_t)I - \eta_t \mathbf{\Lambda} & -\theta_t I \\ I & 0 \end{bmatrix}\right) \\ &\leq \max_{1 \leq i \leq n} \rho\left(\begin{bmatrix} 1+\theta_t - \eta_t \lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix}\right) \end{aligned}$$

To finish the proof, it suffices to show

$$\max_i \rho\left(\begin{bmatrix} 1+\theta_t - \eta_t \lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix}\right) \leq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \tag{2}$$



Proof of Theorem 1*

To show (2), the two eigenvalues of $\begin{bmatrix} 1+\theta_t-\eta_t\lambda_i & -\theta_t \\ 1 & 0 \end{bmatrix}$ are the roots of

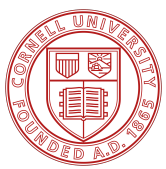
$$z^2-(1+\theta_t-\eta_t\lambda_i)z+\theta_t=0 \tag{3}$$

If $(1+\theta_t-\eta_t\lambda_i)^2 \le 4\theta_t$, then the roots of (3) have the same magnitudes $\sqrt{\theta_t}$ (either they are conjugates of each other or only one root).

In addition, one can easily check that $(1+\theta_t-\eta_t\lambda_i)^2 \le 4\theta_t$ is satisfied if

$$\theta_t \in \left[(1-\sqrt{\eta_t\lambda_i})^2,(1+\sqrt{\eta_t\lambda_i})^2\right], \tag{4}$$

which would hold if one picks $\theta_t=\max\{(1-\sqrt{\eta_t L})^2,(1-\sqrt{\eta_t\mu})^2\}$



Proof of Theorem 1*

With this choice of θ_t, we have (from (3) and the fact that two eigenvalues have identical magnitudes)

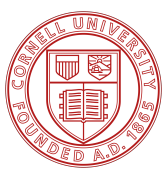
$$\rho(\mathbf{H}_t) \leq \sqrt{\theta_t}.$$

Setting $\eta_t = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ ensures $1 - \sqrt{\eta_t L} = -(1 - \sqrt{\eta_t \mu})$, which yields

$$\theta_t = \max\left\{\left(1 - \frac{2\sqrt{L}}{\sqrt{L}+\sqrt{\mu}}\right)^2, \left(1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2\right\} = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2.$$

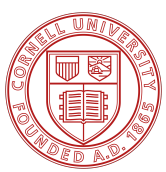
This in turn establishes

$$\rho(\mathbf{H}_t) \leq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$$



Drawbacks in heavy-ball method

- The accelerated rate of heavy-ball method can only be theoretically established for quadratic optimization algorithms
- It is unknown whether heavy-ball can theoretically outperform gradient descent in problems other than quadratic optimization
- In practice, heavy ball is always faster than gradient descent



Recap and fine-tuning

- What we have talked about today?
  - $\Rightarrow$ The impact of noise variance on SGD convergence?
  - $\Rightarrow$ How to reduce variance by averaging iterates?
  - $\Rightarrow$ What is momentum and how it helps convergence?

Welcome anonymous survey!