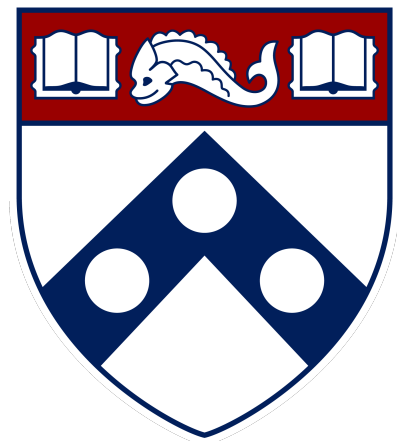
Stochastic gradient methods



Yuxin Chen

Wharton Statistics & Data Science, Fall 2023

Outline

- Stochastic gradient descent (stochastic approximation)
- Convergence analysis
- Reducing variance via iterate averaging

Stochastic programming

$$\text{minimize}_{\boldsymbol{x}} \qquad \underbrace{F\left(\boldsymbol{x}\right) = \mathbb{E}\left[f\left(\boldsymbol{x};\boldsymbol{\xi}\right)\right]}_{\text{expected risk, population risk, ...}}$$

- $\boldsymbol{\xi}$: randomness in problem
- suppose $f(\cdot, \boldsymbol{\xi})$ is convex for every $\boldsymbol{\xi}$ (and hence $F(\cdot)$ is convex)

Example: empirical risk minimization

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

$$\text{minimize}_{\boldsymbol{x}} \quad \underbrace{F(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} f\left(\boldsymbol{x}; \{\boldsymbol{a}_i, y_i\}\right)}_{\text{empirical risk}}$$

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

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

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

A natural solution

Under "mild" technical conditions

$$\begin{aligned} \boldsymbol{x}^{t+1} &= \boldsymbol{x}^t - \eta_t \nabla F(\boldsymbol{x}^t) \\ &= \boldsymbol{x}^t - \eta_t \nabla \, \mathbb{E}\big[f(\boldsymbol{x}^t; \boldsymbol{\xi})\big] \\ &= \boldsymbol{x}^t - \eta_t \mathbb{E}\big[\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^t; \boldsymbol{\xi})\big] \end{aligned}$$

issues:

- distribution of $\boldsymbol{\xi}$ may be unknown
- even if it is known, evaluating high-dimensional expectation is often expensive

Stochastic gradient descent (stochastic approximation)

Stochastic gradient descent (SGD)

— Robbins, Monro '51

stochastic approximation / stochastic gradient descent (SGD)

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \, \boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\xi}^t) \tag{11.1}$$

where $\boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\xi}^t)$ is *unbiased* estimate of $\nabla F(\boldsymbol{x}^t)$, i.e.

$$\mathbb{E}\big[\boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\xi}^t)\big] = \nabla F(\boldsymbol{x}^t)$$

Stochastic gradient descent (SGD)

— Robbins, Monro '51

stochastic approximation / stochastic gradient descent (SGD)

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \, \boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\xi}^t) \tag{11.1}$$

- a stochastic algorithm for finding a critical point $\boldsymbol{x}$ obeying $\nabla F(\boldsymbol{x}) = \boldsymbol{0}$
- more generally, a stochastic algorithm for finding the roots of $G(\boldsymbol{x}) := \mathbb{E}[\boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\xi})]$

Example: SGD for empirical risk minimization

$$\text{minimize}_{\boldsymbol{x}} \quad \underbrace{F(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} f\left(\boldsymbol{x}; \{\boldsymbol{a}_i, y_i\}\right)}_{\text{empirical risk}}$$

for $t = 0, 1, \ldots$

choose i_t uniformly at random, and run

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla_{\boldsymbol{x}} f_{i_t}(\boldsymbol{x}^t; \{\boldsymbol{a}_i, y_i\})$$

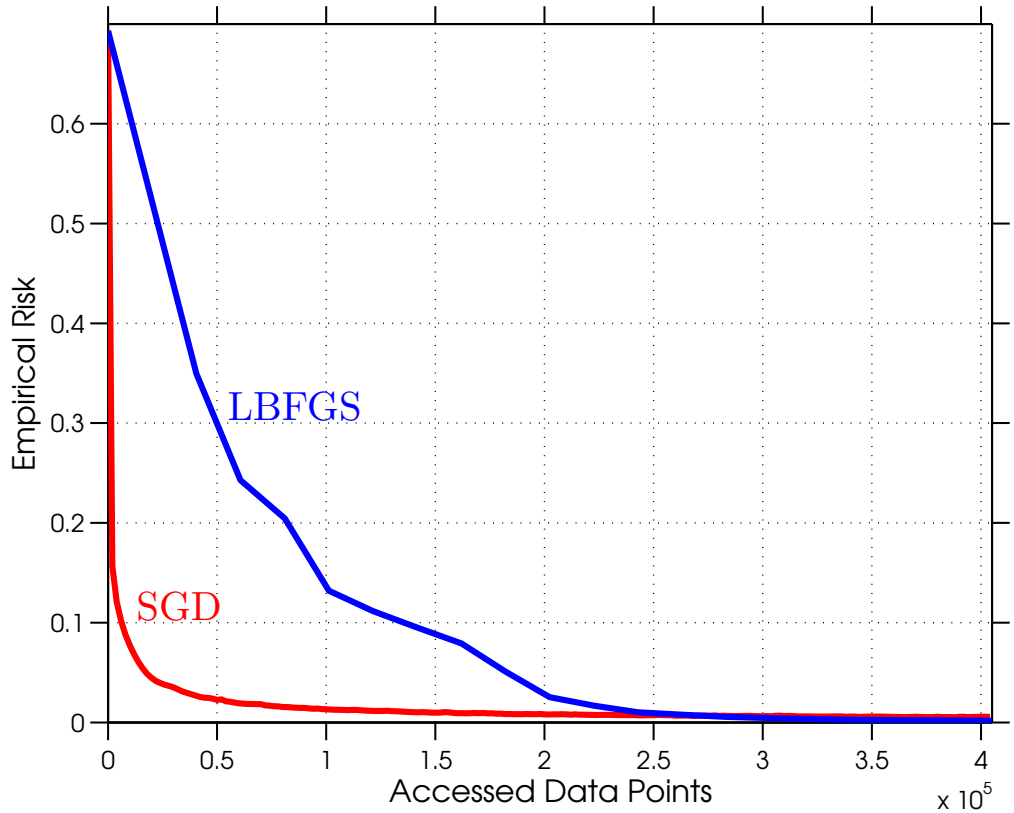
Example: SGD for empirical risk minimization

benefits: SGD exploits information more efficiently than batch methods

- practical data usually involve lots of redundancy; using all data simultaneously in each iteration might be inefficient
- SGD is particularly efficient at the very beginning, as it achieves fast initial improvement with very low per-iteration cost

Example: SGD for empirical risk minimization

— Bottou, Curtis, Nocedal '18



binary classification with logistic loss and RCV1 dataset ($\eta_t \equiv 4$)

Example: temporal difference (TD) learning

Reinforcement learning studies a Markov decision process (MDP) with unknown model

core problem: estimate the so-called "value function" under a stationary policy π

$$V^{\pi}(s) = \mathbb{E}\Big[r_0 + \gamma V^{\pi}(s_1) \mid s_0 = s\Big] \tag{11.2}$$

for all $s \in \mathcal{S}$, without knowing the transition probabilities of the MDP

Example: temporal difference (TD) learning

We won't explain what Equation (11.2) means, but remark that ...

- $V^{\pi}(\cdot)$: value function under policy π
- s_t: state at time t
- $\mathcal{S}$: state space
- $0 < \gamma < 1$: discount factor
- r_t: reward at time t

Example: temporal difference (TD) learning

The definition of the value function is equivalent to

$$\mathbb{E}\Big[\underbrace{V^\pi(s) - r_0 - \gamma V^\pi(s_1)}_{g(V^\pi)} \mid s_0 = s\Big] = 0$$

TD(0) algorithm: for $t = 0, 1, \ldots$

draw a new state s_{t+1}, collect a reward r_t, then update

$$\hat{V}^\pi(s_t) \leftarrow \hat{V}^\pi(s_t) - \eta_t \, g(\hat{V}^\pi) \qquad \text{or}$$

$$\hat{V}^\pi(s_t) \leftarrow \hat{V}^\pi(s_t) - \eta_t \Big\{\hat{V}^\pi(s_t) - r_t - \hat{V}^\pi(s_{t+1})\Big\}$$

Example: Q-learning

What if we also want to find an optimal policy?

core problem: solve the so-called "Bellman equation"

$$V(s) = \max_{a \in \mathcal{A}} \Big\{ R(s,a) + \gamma \, \mathbb{E}\big[V(s_1) \mid s_0 = s, a_0 = a \big] \Big\} \qquad (11.3)$$

for all $s \in \mathcal{S}$, without knowing the transition probabilities of the MDP

Example: Q-learning

Again we won't explain what the Bellman equation means, but remark that ...

- $V(\cdot)$: value function
- s_t: state at time t
- $\mathcal{S}$: state space
- a_t: action at time t
- $\mathcal{A}$: action space
- $0 < \gamma < 1$: discount factor
- $R(\cdot, \cdot)$: reward function

Example: Q-learning

$$V(s) = \max_{a \in \mathcal{A}} \Big\{ R(s,a) + \gamma \, \mathbb{E}\big[V(s_1) \mid s_0 = s, a_0 = a \big] \Big\}$$

- since the transition probabilities are unknown, it is natural to resort to stochastic approximation methods
- **issue:** the Bellman equation has $\mathbb{E}$ inside the max operator
- **a very cute idea:** introduce the so-called "Q function"

Example: Q-learning

$$V(s) = \max_{a \in \mathcal{A}} \Big\{ R(s,a) + \gamma \, \mathbb{E}\big[V(s_1) \mid s_0 = s, a_0 = a \big] \Big\}$$

Define the Q function as

$$\begin{aligned} Q(s,a) &:= R(s,a) + \gamma \, \mathbb{E}\left[V(s_1) \mid s_0 = s, a_0 = a \right] \\ &= R(s,a) + \gamma \, \mathbb{E}\Big[\underbrace{\max_{\tilde{a} \in \mathcal{A}} Q(s_1, \tilde{a})}_{= V(s_1)} \mid s_0 = s, a_0 = a \Big] \end{aligned} \tag{11.4}$$

- **Q learning:** use stochastic approximation methods to estimate the Q function (rather than the value function $V(\cdot)$)

Example: Q-learning

Definition of the Q-function is equivalent to

$$\mathbb{E}\Big[\underbrace{Q(s,a) - R(s,a) - \gamma \max_{\tilde{a}\in\mathcal{A}} Q(s_1,\tilde{a})}_{g(Q)} \mid s_0 = s, a_0 = a\Big] = 0$$

Q-learning algorithm: for $t = 0, 1, \ldots$

draw a new state s_{t+1} using an action a_t, then update

$$\hat{Q}(s_t,a_t) \leftarrow \hat{Q}(s_t,a_t) - \eta_t\, g(\hat{Q}) \qquad \text{or}$$

$$\hat{Q}(s_t,a_t) \leftarrow \hat{Q}(s_t,a_t) - \eta_t \left\{\hat{Q}(s_t,a_t) - R(s_t,a_t) - \gamma \max_{\tilde{a}\in\mathcal{A}} \hat{Q}(s_{t+1},\tilde{a})\right\}$$

Convergence analysis

Strongly convex and smooth problems

$$\text{minimize}_{\boldsymbol{x}} \quad F(\boldsymbol{x}) := \mathbb{E}\big[f(\boldsymbol{x};\boldsymbol{\xi})\big]$$

- F: μ-strongly convex, L-smooth
- $g(\boldsymbol{x}^t;\boldsymbol{\xi}^t)$: an **unbiased** estimate of $\nabla F(\boldsymbol{x}^t)$ given $\{\boldsymbol{\xi}^0,\cdots,\boldsymbol{\xi}^{t-1}\}$
- for all $\boldsymbol{x}$,

$$\mathbb{E}\big[\|g(\boldsymbol{x};\boldsymbol{\xi})\|_2^2\big] \le \sigma_{\mathrm{g}}^2 + c_{\mathrm{g}}\|\nabla F(\boldsymbol{x})\|_2^2 \tag{11.5}$$

Convergence: fixed stepsizes

Theorem 11.1 (Convergence of SGD for strongly convex problems; fixed stepsizes)

Under the assumptions in Page 11-20, if $\eta_t \equiv \eta \leq \frac{1}{Lc_{\mathrm{g}}}$, *then SGD* (11.1) *achieves*

$$\mathbb{E}\big[F(\boldsymbol{x}^t) - F(\boldsymbol{x}^*)\big] \leq \frac{\eta L \sigma_{\mathrm{g}}^2}{2\mu} + (1-\eta\mu)^t \big(F(\boldsymbol{x}^0) - F(\boldsymbol{x}^*)\big)$$

- check Bottou, Curtis, Nocedal '18 (Theorem 4.6) for the proof

Implications: SGD with fixed stepsizes

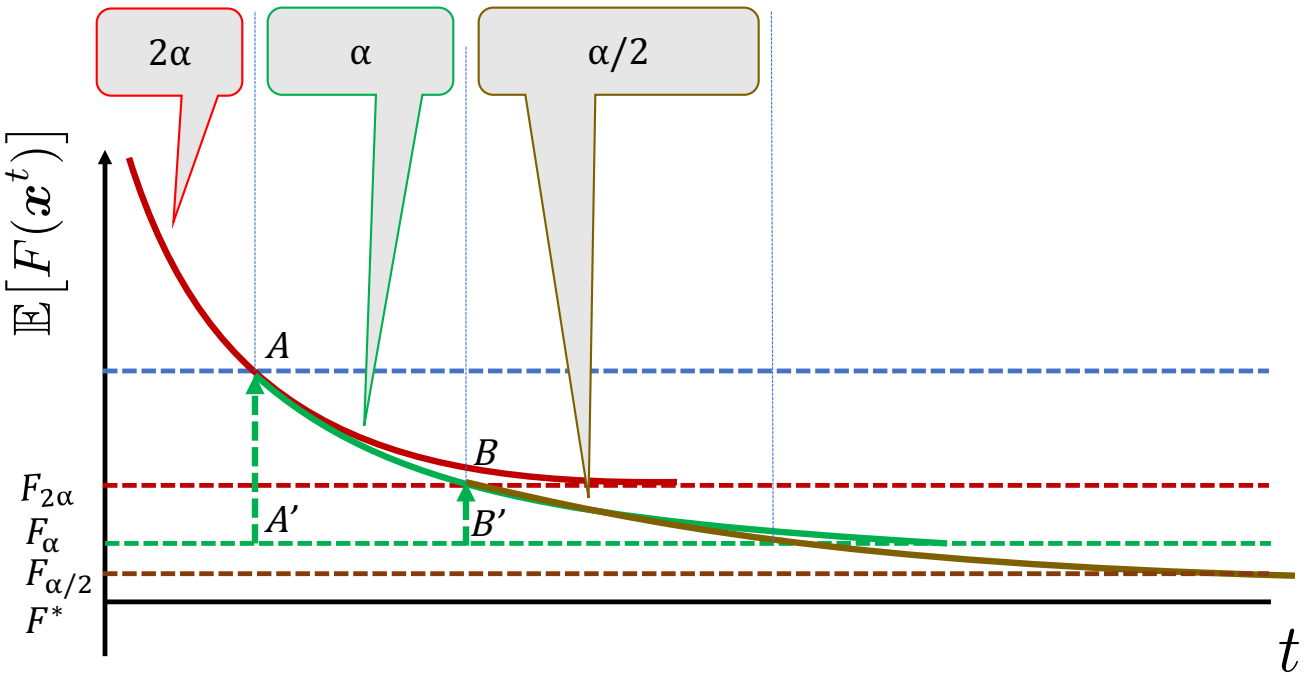
$$\mathbb{E}\big[F(\boldsymbol{x}^t) - F(\boldsymbol{x}^*)\big] \leq \frac{\eta L \sigma_{\mathrm{g}}^2}{2\mu} + (1-\eta\mu)^t \Big(F(\boldsymbol{x}^0) - F(\boldsymbol{x}^*)\Big)$$

- fast (linear) convergence at the very beginning
- converges to some neighborhood of $\boldsymbol{x}^*$ — variation in gradient computation prevents further progress
- when gradient computation is noiseless (i.e. $\sigma_{\mathrm{g}} = 0$), it converges linearly to optimal points
- smaller stepsizes η yield better converging points

One practical strategy

Run SGD with fixed stepsizes; whenever progress stalls, reduce stepsizes and continue SGD

— *Bottou, Curtis, Nocedal '18*



whenever progress stalls, we half the stepsizes and repeat

Convergence with diminishing stepsizes

Theorem 11.2 (Convergence of SGD for strongly convex problems; diminishing stepsizes)

Suppose F *is* μ*-strongly convex, and* (11.5) *holds with* $c_{\mathrm{g}} = 0$*. If* $\eta_t = \frac{\theta}{t+1}$ *for some* $\theta > \frac{1}{2\mu}$*, then SGD* (11.1) *achieves*

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

where $c_\theta = \max\left\{\frac{2\theta^2\sigma_{\mathrm{g}}^2}{2\mu\theta - 1},\ \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2\right\}$

- convergence rate $O(1/t)$ with diminishing stepsize $\eta_t \asymp 1/t$

Proof of Theorem 11.2

Using the SGD update rule, we have

$$\begin{aligned}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^*\|_2^2 &= \left\|\boldsymbol{x}^t-\eta_t\boldsymbol{g}(\boldsymbol{x}^t;\boldsymbol{\xi}^t)-\boldsymbol{x}^*\right\|_2^2 \\ &= \|\boldsymbol{x}^t-\boldsymbol{x}^*\|_2^2-2\eta_t(\boldsymbol{x}^t-\boldsymbol{x}^*)^\top\boldsymbol{g}(\boldsymbol{x}^t;\boldsymbol{\xi}^t)+\eta_t^2\|\boldsymbol{g}(\boldsymbol{x}^t;\boldsymbol{\xi}^t)\|_2^2 \qquad (11.6)\end{aligned}$$

Since $\boldsymbol{x}^t$ is indep. of $\boldsymbol{\xi}_t$, apply the law of total expectation to obtain

$$\begin{aligned}\mathbb{E}\big[(\boldsymbol{x}^t-\boldsymbol{x}^*)^\top\boldsymbol{g}(\boldsymbol{x}^t;\boldsymbol{\xi}^t)\big] &= \mathbb{E}\big[\mathbb{E}\big[(\boldsymbol{x}^t-\boldsymbol{x}^*)^\top\boldsymbol{g}(\boldsymbol{x}^t;\boldsymbol{\xi}^t)\,|\,\boldsymbol{\xi}_1,\cdots,\boldsymbol{\xi}_{t-1}\big]\big] \\ &= \mathbb{E}\big[(\boldsymbol{x}^t-\boldsymbol{x}^*)^\top\mathbb{E}[\boldsymbol{g}(\boldsymbol{x}^t;\boldsymbol{\xi}^t)\,|\,\boldsymbol{\xi}_1,\cdots,\boldsymbol{\xi}_{t-1}]\big] \\ &= \mathbb{E}\big[(\boldsymbol{x}^t-\boldsymbol{x}^*)^\top\nabla F(\boldsymbol{x}^t)\big] \qquad (11.7)\end{aligned}$$

Proof of Theorem 11.2 (cont.)

Furthermore, strong convexity gives

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

$$\implies \quad \mathbb{E}\big[\langle \nabla F(\boldsymbol{x}^t), \boldsymbol{x}^t - \boldsymbol{x}^* \rangle\big] \geq \mu \mathbb{E}\big[\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2\big] \tag{11.8}$$

Combine (11.6), (11.7), (11.8) and (11.5) (with $c_{\mathrm{g}} = 0$) to obtain

$$\mathbb{E}\big[\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2\big] \leq (1 - 2\mu\eta_t)\mathbb{E}\big[\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2\big] + \underbrace{\eta_t^2 \sigma_{\mathrm{g}}^2}_{\text{does not vanish unless } \eta_t \to 0} \tag{11.9}$$

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

Optimality

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

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 $1/t$

$\Longrightarrow$ SGD with stepsizes $\eta_t \asymp 1/t$ is optimal

Optimality

— Nemirovski, Yudin '83

More precisely, consider a class of problems in which f is μ-strongly convex and L-smooth, and $\mathrm{Var}(\|\boldsymbol{g}(\boldsymbol{x}^t;\boldsymbol{\xi}^t)\|_2) \le \sigma^2$. Then the worst-case iteration complexity for (stochastic) first-order methods:

$$\sqrt{\frac{L}{\mu}}\log\left(\frac{L\|\boldsymbol{x}_0-\boldsymbol{x}^*\|_2^2}{\varepsilon}\right)+\frac{\sigma^2}{\mu\varepsilon}$$

- for deterministic cases: $\sigma = 0$, and hence the lower bound is

$$\sqrt{\frac{L}{\mu}}\log\left(\frac{L\|\boldsymbol{x}_0-\boldsymbol{x}^*\|_2^2}{\varepsilon}\right)$$

(achievable by Nesterov's method)

Optimality

— Nemirovski, Yudin '83

More precisely, consider a class of problems in which f is μ-strongly convex and L-smooth, and $\mathrm{Var}(\|\boldsymbol{g}(\boldsymbol{x}^t;\boldsymbol{\xi}^t)\|_2) \leq \sigma^2$. Then the worst-case iteration complexity for (stochastic) first-order methods:

$$\sqrt{\frac{L}{\mu}}\log\left(\frac{L\|\boldsymbol{x}_0-\boldsymbol{x}^*\|_2^2}{\varepsilon}\right)+\frac{\sigma^2}{\mu\varepsilon}$$

- for noisy cases with large σ, the lower bound is dominated by

$$\frac{\sigma^2}{\mu}\cdot\frac{1}{\varepsilon}$$

Comparisons with batch GD

Empirical risk minimization with n samples:

	iteration complexity	per-iteration cost	total comput. cost
batch GD	$\log \frac{1}{\varepsilon}$	n	$n \log \frac{1}{\varepsilon}$
SGD	$\frac{1}{\varepsilon}$	1	$\frac{1}{\varepsilon}$

SGD is more appealing for large n and moderate accuracy ε (in which case $\frac{1}{\varepsilon} < n \log \frac{1}{\varepsilon}$)

— which often arises in the *big data* regime!

Convex problems

What if we lose strong convexity?

$$\text{minimize}_{\boldsymbol{x}} \quad F(\boldsymbol{x}) := \mathbb{E}\big[f(\boldsymbol{x};\boldsymbol{\xi})\big]$$

- F: convex
- $\mathbb{E}\big[\|g(\boldsymbol{x};\boldsymbol{\xi})\|_2^2\big] \leq \sigma_{\mathrm{g}}^2$ for all $\boldsymbol{x}$
- $g(\boldsymbol{x}^t;\boldsymbol{\xi}^t)$ is an unbiased estimate of $\nabla F(\boldsymbol{x}^t)$ given $\{\boldsymbol{\xi}^0,\cdots,\boldsymbol{\xi}^{t-1}\}$

Convex problems

Suppose we return a weighted average

$$\tilde{\boldsymbol{x}}^t := \sum_{k=0}^{t} \frac{\eta_k}{\sum_{j=0}^{t} \eta_j} \boldsymbol{x}^k$$

Theorem 11.3

Under the assumptions in Page 11-30, one has

$$\mathbb{E}[F(\tilde{\boldsymbol{x}}^t) - F(\boldsymbol{x}^*)] \leq \frac{\frac{1}{2}\mathbb{E}[\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2] + \frac{1}{2}\sigma_{\mathrm{g}}^2 \sum_{k=0}^{t} \eta_k^2}{\sum_{k=0}^{t} \eta_k}$$

- if $\eta_t \asymp 1/\sqrt{t}$, then

$$\mathbb{E}[F(\tilde{\boldsymbol{x}}^t) - F(\boldsymbol{x}^*)] \lesssim \frac{\log t}{\sqrt{t}}$$

Proof of Theorem 11.3

Remark: very similar to the convergence analysis for subgradient methods

By convexity of F, we have $F(\boldsymbol{x}) \geq F(\boldsymbol{x}^t) + (\boldsymbol{x} - \boldsymbol{x}^t)^\top \nabla F(\boldsymbol{x}^t)$

$$\Longrightarrow \quad \mathbb{E}[(\boldsymbol{x}^t - \boldsymbol{x}^*)^\top \nabla F(\boldsymbol{x}^t)] \geq \mathbb{E}[F(\boldsymbol{x}^t) - F(\boldsymbol{x}^*)]$$

This together with (11.6) and (11.7) implies

$$2\eta_k \mathbb{E}[F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*)] \leq \mathbb{E}\big[\|\boldsymbol{x}^k - \boldsymbol{x}^*\|_2^2\big] - \mathbb{E}\big[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^*\|_2^2\big] + \eta_k^2 \sigma_{\mathrm{g}}^2$$

Sum over $k = 0, \cdots, t$ to obtain

$$\sum_{k=0}^{t} 2\eta_k \mathbb{E}[F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*)] \leq \mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2\big] - \mathbb{E}\big[\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2\big] + \sigma_{\mathrm{g}}^2 \sum_{k=0}^{t} \eta_k^2$$

$$\leq \mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2\big] + \sigma_{\mathrm{g}}^2 \sum_{k=0}^{t} \eta_k^2$$

Proof of Theorem 11.3 (cont.)

Setting $v_t = \frac{\eta_t}{\sum_{k=0}^{t} \eta_k}$ yields

$$\sum_{k=0}^{t} v_k \mathbb{E}[F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*)] \leq \frac{\frac{1}{2}\mathbb{E}[\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2] + \frac{1}{2}\sigma_{\mathrm{g}}^2 \sum_{k=0}^{t} \eta_k^2}{\sum_{k=0}^{t} \eta_k}$$

By convexity of F, we arrive at

$$\mathbb{E}[F(\tilde{\boldsymbol{x}}^t) - F(\boldsymbol{x}^*)] \leq \frac{\frac{1}{2}\mathbb{E}[\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2] + \frac{1}{2}\sigma_{\mathrm{g}}^2 \sum_{k=0}^{t} \eta_k^2}{\sum_{k=0}^{t} \eta_k}$$

Reducing variance via iterate averaging

Stepsize choice $O(1/t)$?

Two conflicting regimes

- the noiseless case (i.e. $\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\xi}) = \nabla F(\boldsymbol{x})$): stepsizes $\eta_t \asymp 1/t$ are way too conservative
- the general noisy case: longer stepsizes ($\eta_t \gg 1/t$) might fail to suppress noise (and hence slow down convergence)

Can we modify SGD so as to allow for larger stepsizes without compromising convergence rates?

Motivation for iterate averaging

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

One may, however, average iterates to mitigate oscillation and reduce variance

Acceleration by averaging

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

$$\text{return} \quad \overline{\boldsymbol{x}}^t := \frac{1}{t}\sum_{i=0}^{t-1} \boldsymbol{x}^i \tag{11.10}$$

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

Key idea: average the iterates to reduce variance and improve convergence

Example: a toy quadratic problem

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \quad \frac{1}{2}\|\boldsymbol{x}\|_2^2$$

- constant stepsizes: $\eta_t \equiv \eta < 1$
- $\boldsymbol{g}(\boldsymbol{x}^t;\boldsymbol{\xi}^t) = \boldsymbol{x}^t + \boldsymbol{\xi}^t$ with
  - $\mathbb{E}\big[\boldsymbol{\xi}^t \mid \boldsymbol{\xi}^0,\cdots,\boldsymbol{\xi}^{t-1}\big] = \boldsymbol{0}$
  - $\mathbb{E}\big[\boldsymbol{\xi}^t\boldsymbol{\xi}^{t\top} \mid \boldsymbol{\xi}^0,\cdots,\boldsymbol{\xi}^{t-1}\big] = \boldsymbol{I}$

Example: a toy quadratic problem

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \quad \frac{1}{2}\|\boldsymbol{x}\|_2^2$$

SGD iterates:

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

Example: a toy quadratic problem

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \quad \frac{1}{2}\|\boldsymbol{x}\|_2^2$$

$$\overline{\boldsymbol{x}}^t \approx \underbrace{\frac{1}{t}\sum_{k=0}^{t-1}(1-\eta)^k \boldsymbol{x}^0}_{=\frac{1}{t}\frac{1-(1-\eta)^t}{\eta}\boldsymbol{x}^0 \overset{t\to\infty}{\to} \mathbf{0}} \underbrace{-\eta\left\{1+(1-\eta)+\cdots\right\}\frac{1}{t}\sum_{k=0}^{t-1}\boldsymbol{\xi}^k}_{\text{imprecise; but close enough for large } t}$$

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

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

Example: more general quadratic problems

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \quad \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$$

- $\boldsymbol{A} \succeq \mu \boldsymbol{I} \succ \boldsymbol{0}$ (strongly convex)
- constant stepsizes: $\eta_t \equiv \eta < 1/\mu$
- $\boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\xi}^t) = \boldsymbol{A}\boldsymbol{x}^t - \boldsymbol{b} + \boldsymbol{\xi}^t$ with
  - $\mathbb{E}\left[\boldsymbol{\xi}^t \mid \boldsymbol{\xi}^0, \cdots, \boldsymbol{\xi}^{t-1}\right] = \boldsymbol{0}$
  - $\boldsymbol{S} := \lim_{t\to\infty} \mathbb{E}\left[\boldsymbol{\xi}^t \boldsymbol{\xi}^{t\top} \mid \boldsymbol{\xi}^0, \cdots, \boldsymbol{\xi}^{t-1}\right]$ is finite

Example: more general quadratic problems

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \quad \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$$

Theorem 11.4

Fix d. Then as $t \to \infty$, the iterate average $\overline{\boldsymbol{x}}^t$ obeys

$$\sqrt{t}(\overline{\boldsymbol{x}}^t - \boldsymbol{x}^*) \quad \underbrace{\xrightarrow{\mathcal{D}}}_{\text{convergence in distribution}} \quad \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{S}\boldsymbol{A}^{-1})$$

Example: quadratic problems

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

- asymptotically, $\left\|\overline{\boldsymbol{x}}^t - \boldsymbol{x}^*\right\|_2^2 \asymp 1/t$, which matches the convergence rate in Theorem 11.2
- much longer stepsizes ($\eta_t \asymp 1$)

  $\Longrightarrow$ faster convergence for less noisy cases (e.g. $\boldsymbol{\xi}^t = \mathbf{0}$)

Proof sketch of Theorem 11.5

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

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

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

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

$$\overline{\boldsymbol{\Delta}}^t = \frac{1}{t\eta}\boldsymbol{G}_0^t\boldsymbol{\Delta}^0 + \frac{1}{t}\sum_{j=0}^{t-2}\boldsymbol{A}^{-1}\boldsymbol{\xi}^j + \frac{1}{t}\sum_{j=0}^{t-2}(\boldsymbol{G}_j^t - \boldsymbol{A}^{-1})\boldsymbol{\xi}^j$$

$$\text{where} \qquad \boldsymbol{G}_j^t := \eta\sum_{i=0}^{t-1-j}(\boldsymbol{I} - \eta\boldsymbol{A})^i$$

Proof sketch of Theorem 11.5 (cont.)

(3) From the central limit theorem for martingales,

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

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

$$\|\boldsymbol{G}_0^t\| < \infty, \ \|\boldsymbol{G}_j^t - \boldsymbol{A}^{-1}\| < \infty \text{ and } \lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{t-1} \|\boldsymbol{G}_j^t - \boldsymbol{A}^{-1}\|^2 = 0$$

(5) Combining these bounds establishes Theorem 11.5

More general strongly convex problems

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \quad F(\boldsymbol{x})$$

- F: strongly convex
- stepsizes: $\eta_t \asymp t^{-\alpha}$ with $\alpha \in (0.5, 1)$
- $\boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\xi}^t) = \nabla F(\boldsymbol{x}^t) + \boldsymbol{\xi}^t$
  - $\mathbb{E}\big[\boldsymbol{\xi}^t \mid \boldsymbol{\xi}^0, \cdots, \boldsymbol{\xi}^{t-1}\big] = \mathbf{0}$
  - $\boldsymbol{S} := \lim_{t\to\infty} \mathbb{E}\big[\boldsymbol{\xi}^t \boldsymbol{\xi}^{t\top} \mid \boldsymbol{\xi}^0, \cdots, \boldsymbol{\xi}^{t-1}\big]$ is finite

More general strongly convex problems

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \quad F(\boldsymbol{x})$$

Theorem 11.5 (informal, Polyak, Juditsky '92)

Fix d and let $t \to \infty$. For a large class of strongly convex problems, if $\eta_t \asymp t^{-\alpha}$ for some $1/2 < \alpha < 1$, then

$$\sqrt{t}(\overline{\boldsymbol{x}}^t - \boldsymbol{x}^*) \overset{\mathcal{D}}{\to} \mathcal{N}\Big(\mathbf{0}, \big(\nabla^2 F(\boldsymbol{x}^*)\big)^{-1} \boldsymbol{S} \big(\nabla^2 F(\boldsymbol{x}^*)\big)^{-1}\Big)$$

- depend on the local curvature at / around minimizers
- allow the stepsize η_t to be longer than $1/t$

Reference

- "*A stochastic approximation method*, H. Robbins, S. Monro, *the annals of mathematical statistics*, 1951.
- "*Robust stochastic approximation approach to stochastic programming*," A. Nemirovski et al., *SIAM Journal on optimization*, 2009.
- "*Optimization methods for large-scale machine learning*," L. Bottou et al., arXiv, 2016.
- "*New stochastic approximation type procedures*," B. Polyak, *Automat. Remote Control*, 1990.
- "*Acceleration of stochastic approximation by averaging*," B. Polyak, A. Juditsky, *SIAM Journal on Control and Optimization*, 1992.
- "*Efficient estimations from a slowly convergent Robbins-Monro process*," D. Ruppert, 1988.

Reference

- "*Problem complexity and method efficiency in optimization*," A. Nemirovski, D. Yudin, Wiley, 1983.
- "*Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization*," A. Agarwal, P. Bartlett, P. Ravikumar, M. Wainwright, *IEEE Transactions on Information Theory*, 2011.
- "*Information-based complexity, feedback and dynamics in convex programming*," M. Raginsky, A. Rakhlin, *IEEE Transactions on Information Theory*, 2011.
- "*First-order methods in optimization*," A. Beck, Vol. 25, SIAM, 2017.
- "*Acceleration of stochastic approximation by averaging*," B. Polyak, A. Juditsky, *SIAM Journal on Control and Optimization*, 1992.
- "*A convergence theorem for nonnegative almost supermartingales and some applications*," H. Robbins, D. Siegmund, *Optimizing methods in statistics*, 1971.