Stochastic Gradient Descent

Cong Ma

University of Chicago, Winter 2026

Outline

- Stochastic gradient descent (stochastic approximation)
- Convergence analysis

Stochastic Programming

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

- $\boldsymbol{\Xi}$ denotes the random source in the objective.
- Assume $f(\cdot, \boldsymbol{\Xi})$ is convex for every realization of $\boldsymbol{\Xi}$; then F 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 $j \sim \text{Unif}(1, \ldots, n)$ 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 regularity conditions, we can interchange gradient and expectation:

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

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 Approximation / Stochastic Gradient Descent

— Robbins, Monro '51

stochastic approximation / stochastic gradient descent (SGD)

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

where $\boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\Xi}^t)$ is an *unbiased* estimate of $\nabla F(\boldsymbol{x}^t)$:

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

- 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

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla_{\boldsymbol{x}} f\left(\boldsymbol{x}^t; \{\boldsymbol{a}_{i_t}, y_{i_t}\}\right).$$

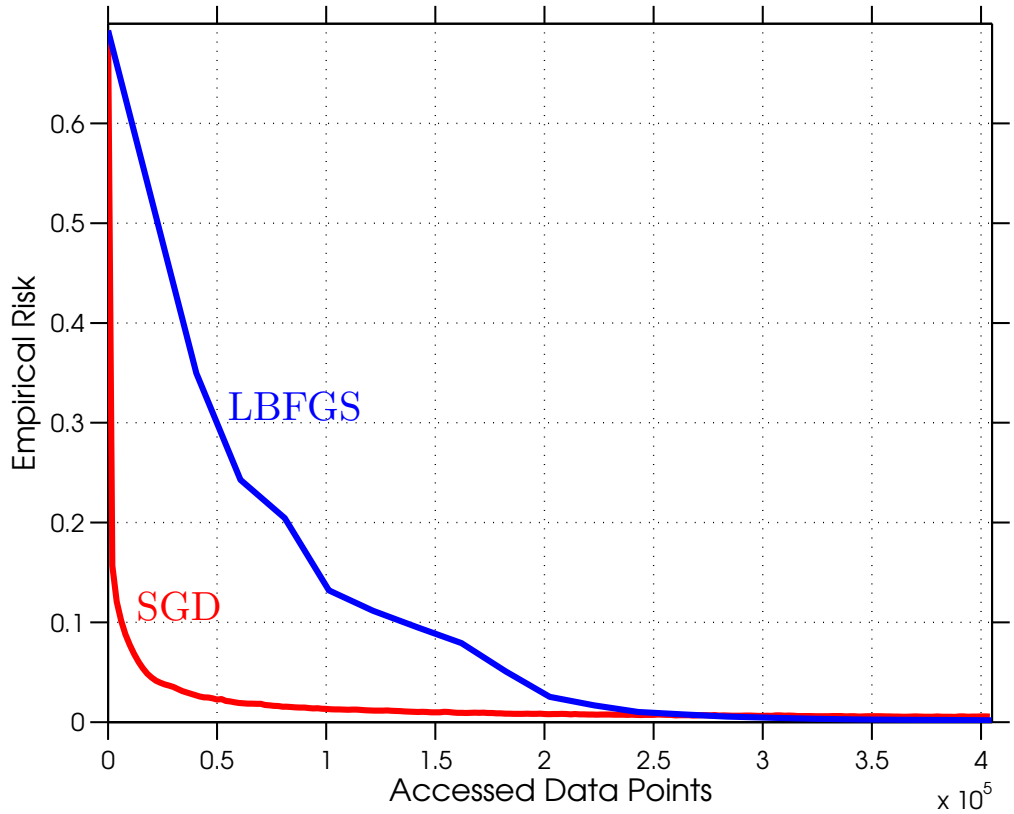
Example: SGD for Empirical Risk Minimization

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 ($\eta_t \equiv 4$)

Convergence analysis

Convergence Roadmap

- We first study strongly convex and smooth objectives under a second-moment bound on stochastic gradients.
- With fixed stepsizes, SGD converges linearly to a noise-dominated neighborhood.
- With diminishing stepsizes, SGD converges to $\boldsymbol{x}^*$ at rate $O(1/t)$.

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
- $\boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\Xi}^t)$ is conditionally unbiased:

$$\mathbb{E}\big[\boldsymbol{g}(\boldsymbol{x}^t; \boldsymbol{\Xi}^t) \,\big|\, \boldsymbol{\Xi}^0, \ldots, \boldsymbol{\Xi}^{t-1}\big] = \nabla F(\boldsymbol{x}^t)$$

- for all $\boldsymbol{x}$,

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

Convergence: Fixed Stepsizes

Theorem 1 (Convergence of SGD for strongly convex problems; fixed stepsizes)

Under the assumptions in Frame 13, if $\eta_t \equiv \eta \leq \frac{1}{Lc_{\mathrm{g}}}$, *then SGD* (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

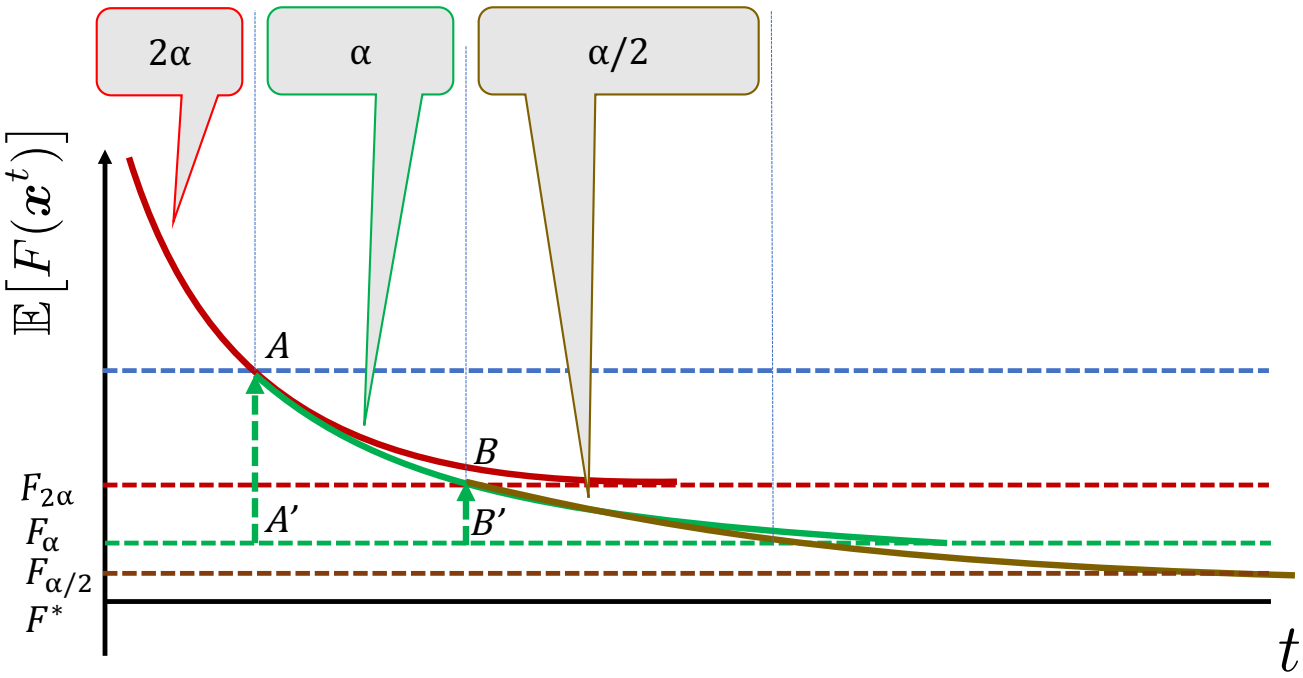
$$\mathbb{E}\big[F(\boldsymbol{x}^t) - F(\boldsymbol{x}^*)\big] \leq \frac{\eta L \sigma_{\mathrm{g}}^2}{2\mu} + (1-\eta\mu)^t \Big(F(\boldsymbol{x}^0) - F(\boldsymbol{x}^*)\Big)$$

- 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 fixed stepsizes η yield better terminal accuracy

One Practical Strategy

Run SGD with a fixed stepsize; whenever progress stalls, reduce the stepsize and continue.

— *Bottou, Curtis, Nocedal '18*



Whenever progress stalls, halve the stepsize and repeat.

Convergence with Diminishing Stepsizes

Theorem 2 (Convergence of SGD for strongly convex problems; diminishing stepsizes)

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

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

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 $\eta_t \asymp 1/t$