Optimization for Machine Learning
机器学习中的优化方法

陈 程

华东师范大学 软件工程学院

chchen@sei.ecnu.edu.cn

Outline

Empirical risk minimization

Let $\{\mathbf{a}_i, b_i\}_{i=1}^n$ be n random samples. In machine learning, we usually learn model parameters $\mathbf{x}$ by optimizing

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}; \{\mathbf{a}_i, b_i\}).$$

- hinge loss (support vector machine):

$$f(\mathbf{x}; \{\mathbf{a}_i, b_i\}) = \max\{1 - b_i\mathbf{a}_i^\top \mathbf{x}, 0\}$$

- logistic loss (logistic regression):

$$f(\mathbf{x}; \{\mathbf{a}_i, b_i\}) = \log(1 + \exp(-b_i\mathbf{a}_i^\top \mathbf{x}))$$

- neural network

Stochastic optimization

More generally, we consider the stochastic optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \underbrace{\mathbb{E}_\xi[f(\mathbf{x};\xi)]}_{\text{expectation setting}},$$

where the random variable $\xi \sim \mathcal{D}$.

- ξ is the randomness in problem.
- In this lecture, we suppose $F(\mathbf{x})$ is differentiable and convex.

Finite-sum setting

The finite-sum setting is a special case of the expectation setting:

$$F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}).$$

If one draws index i from $\{1, 2, \ldots, n\}$ uniformly at random, then

$$F(\mathbf{x}) = \mathbb{E}_i[f_i(\mathbf{x})].$$

A natural solution

Under "mild" assumptions, we have

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{x}_t - \eta_t \nabla F(\mathbf{x}_t) \\ &= \mathbf{x}_t - \eta_t \nabla \mathbb{E}[f(\mathbf{x}_t, \xi)] \\ &= \mathbf{x}_t - \eta_t \mathbb{E}[\nabla_{\mathbf{x}} f(\mathbf{x}_t, \xi)] \end{aligned}$$

issues:

- For the expectation setting, distribution of ξ may be unknown.
- For the finite-sum setting, computing full gradient is very expensive when n is very large.

A natural solution

Under "mild" assumptions, we have

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{x}_t - \eta_t \nabla F(\mathbf{x}_t) \\ &= \mathbf{x}_t - \eta_t \nabla \mathbb{E}[f(\mathbf{x}_t, \xi)] \\ &= \mathbf{x}_t - \eta_t \mathbb{E}[\nabla_{\mathbf{x}} f(\mathbf{x}_t, \xi)] \end{aligned}$$

issues:

- For the expectation setting, distribution of ξ may be unknown.
- For the finite-sum setting, computing full gradient is very expensive when n is very large.

Outline

Stochastic gradient descent (SGD)

Stochastic gradient descent:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t g(\mathbf{x}_t, \xi),$$

where $g(\mathbf{x}_t, \xi)$ is an unbiased estimator of $\nabla F(\mathbf{x}_t)$, i.e.,

$$\mathbb{E}[g(\mathbf{x}_t, \xi)] = \nabla F(\mathbf{x}_t).$$

For the finite-sum setting, we can choose index i_t from $\{1, 2, \ldots, n\}$ uniformly at random. Then $\nabla f_{i_t}(\mathbf{x}_t)$ is an unbiased estimator of $\nabla F(\mathbf{x}_t)$:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f_{i_t}(\mathbf{x}_t)$$

Stochastic gradient descent (SGD)

Stochastic gradient descent:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t g(\mathbf{x}_t, \xi),$$

where $g(\mathbf{x}_t, \xi)$ is an unbiased estimator of $\nabla F(\mathbf{x}_t)$, i.e.,

$$\mathbb{E}[g(\mathbf{x}_t, \xi)] = \nabla F(\mathbf{x}_t).$$

For the finite-sum setting, we can choose index i_t from $\{1, 2, \ldots, n\}$ uniformly at random. Then $\nabla f_{i_t}(\mathbf{x}_t)$ is an unbiased estimator of $\nabla F(\mathbf{x}_t)$:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f_{i_t}(\mathbf{x}_t)$$

Outline

Strongly convex and smooth problems

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \mathbb{E}_\xi[f(\mathbf{x};\xi)]$$

Assumptions:

- $F(\mathbf{x})$ is L-smooth and μ-strongly convex (we do not require assumptions on $f(\mathbf{x};\xi)$);
- Given $\xi_0,\ldots,\xi_{t-1}$, $g(\mathbf{x}_t,\xi_t)$ is an unbiased estimator of $\nabla F(\mathbf{x}_t)$, i.e.,

$$\mathbb{E}\left[g(\mathbf{x}_t,\xi_t)\middle|\xi_0,\ldots,\xi_{t-1}\right] = \nabla F(\mathbf{x}_t);$$

- For all $\mathbf{x}$, we have $\underbrace{\mathbb{E}\left[\|g(\mathbf{x},\xi)\|_2^2\right] \le \sigma^2}_{\text{bounded variance}}$.

Strongly convex and smooth problems

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) \triangleq \mathbb{E}_\xi[f(\mathbf{x}; \xi)]$$

Assumptions:

- $F(\mathbf{x})$ is L-smooth and μ-strongly convex (we do not require assumptions on $f(\mathbf{x}; \xi)$);
- Given $\xi_0, \ldots, \xi_{t-1}$, $g(\mathbf{x}_t, \xi_t)$ is an unbiased estimator of $\nabla F(\mathbf{x}_t)$, i.e.,

$$\mathbb{E}\left[g(\mathbf{x}_t, \xi_t) \middle| \xi_0, \ldots, \xi_{t-1}\right] = \nabla F(\mathbf{x}_t);$$

- For all $\mathbf{x}$, we have $\underbrace{\mathbb{E}\left[\|g(\mathbf{x}, \xi)\|_2^2\right] \leq \sigma^2}_{\text{bounded variance}}$.

Strongly convex and smooth problems

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \mathbb{E}_\xi[f(\mathbf{x};\xi)]$$

Assumptions:

- $F(\mathbf{x})$ is L-smooth and μ-strongly convex (we do not require assumptions on $f(\mathbf{x};\xi)$);
- Given $\xi_0,\ldots,\xi_{t-1}$, $g(\mathbf{x}_t,\xi_t)$ is an unbiased estimator of $\nabla F(\mathbf{x}_t)$, i.e.,

$$\mathbb{E}\left[g(\mathbf{x}_t,\xi_t)\middle|\xi_0,\ldots,\xi_{t-1}\right] = \nabla F(\mathbf{x}_t);$$

- For all $\mathbf{x}$, we have $\underbrace{\mathbb{E}\left[\|g(\mathbf{x},\xi)\|_2^2\right] \le \sigma^2}_{\text{bounded variance}}$.

Convergence with fixed stepsizes

Under the assumptions in page 7, if $\eta_t = \eta \leq 1/(2L)$, then SGD achieves

$$\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] \leq (1 - 2\eta\mu)^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \frac{\eta\sigma^2}{2\mu};$$

- fast (linear) convergence at the very beginning
- converges to some neighborhood of $\mathbf{x}^*$
- smaller stepsize η yield better converging points

Convergence with fixed stepsizes

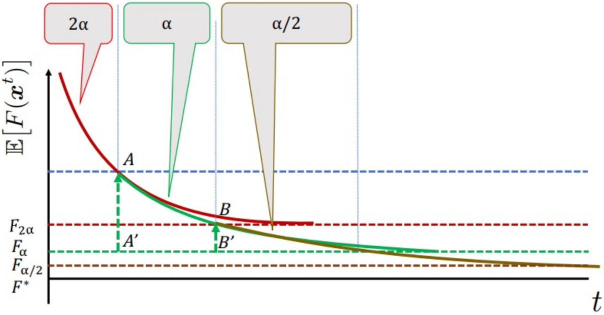
Under the assumptions in page 7, if $\eta_t = \eta \le 1/(2L)$, then SGD achieves

$$\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] \le (1 - 2\eta\mu)^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \frac{\eta\sigma^2}{2\mu};$$

- fast (linear) convergence at the very beginning
- converges to some neighborhood of $\mathbf{x}^*$
- smaller stepsize η yield better converging points

One practical strategy

Run SGD with fixed stepsizes; whenever progress stalls, half the stepsize and continue SGD.



Convergence with diminishing stepsizes

Under the assumptions in page 7, if $\eta_t = \frac{\theta}{t+1}$ for some $\theta > \frac{1}{2\mu}$, then SGD achieves

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

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

Convex and smooth problems

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \mathbb{E}_\xi[f(\mathbf{x};\xi)]$$

Assumptions:

- F is L-smooth and convex;
- Given $\xi_0, \ldots, \xi_{t-1}$, $g(\mathbf{x}_t, \xi_t)$ is an unbiased estimator of $\nabla F(\mathbf{x}_t)$;
- For all $\mathbf{x}$, we have $\underbrace{\mathbb{E}[\|g(\mathbf{x},\xi)\|_2^2] \le \sigma^2}_{\text{bounded variance}}$.

Convex and smooth problems

Suppose we return a weighted average

$$\tilde{\mathbf{x}}_t = \sum_{k=0}^{t} \frac{\eta_k}{\sum_{j=0}^{t} \eta_j} \mathbf{x}_k$$

If F is convex, we have

$$\mathbb{E}[F(\tilde{\mathbf{x}}_t) - F(\mathbf{x}^*)] \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \sum_{k=0}^{t} \sigma^2 \eta_k^2}{2\sum_{k=0}^{t} \eta_k}.$$

If we choose $\eta_t = \Theta(1/\sqrt{t})$, we can get

$$\mathbb{E}[F(\tilde{\mathbf{x}}_t) - F(\mathbf{x}^*)] \leq O\left(\frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \sigma^2 \log t}{\sqrt{t}}\right).$$

Convex and smooth problems

If we return

$$\tilde{\mathbf{x}}_t = \sum_{k=\lceil \frac{t}{2} \rceil}^{t} \frac{\eta_k \mathbf{x}_k}{\sum_{j=\lceil \frac{t}{2} \rceil}^{t} \eta_j}$$

Then we have

$$\mathbb{E}[F(\tilde{\mathbf{x}}_t) - F(\mathbf{x}^*)] \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \sum_{k=\lceil \frac{t}{2} \rceil}^{t} \sigma^2 \eta_k^2}{2 \sum_{k=\lceil \frac{t}{2} \rceil}^{t} \eta_k} = O\left(\frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \sigma^2}{\sqrt{t}} \right).$$

Comparisons with batch GD

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}).$$

	iteration complexity	per-iteration	total
batch GD	$\kappa \log(1/\epsilon)$	n	$n\kappa \log(1/\epsilon)$
SGD	$1/\epsilon$	1	$1/\epsilon$

Table: Convergence rate for the strongly convex case

	iteration complexity	per-iteration	total
batch GD	$1/\epsilon$	n	n/ϵ
SGD	$1/\epsilon^2$	1	$1/\epsilon^2$

Table: Convergence rate for the convex case