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

Review of Gradient Descent

For unconstrained convex optimization, the **gradient descent** method starts with an initial point $\mathbf{x}_0$, and iteratively computes

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

For constrained convex optimization with constraint $\mathcal{C}$, the **projected gradient descent** method starts with an initial point $\mathbf{x}_0$, and iteratively computes

$$\mathbf{x}_{t+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)).$$

Review of Convergence Rate

condition	constrained	convergence rate	iteration complexity
strongly convex & smooth	no	$O\left(\left(1-\frac{1}{\kappa}\right)^t\right)$	$O(\kappa \log \frac{1}{\varepsilon})$
strongly convex & smooth	yes	$O\left(\left(1-\frac{1}{\kappa}\right)^t\right)$	$O(\kappa \log \frac{1}{\varepsilon})$
convex & smooth	no	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$
convex & smooth	yes	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$

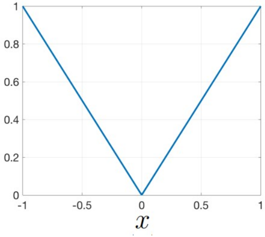
Table: Convergence Properties of GD & PGD

Outline

Nondifferentiable Problems

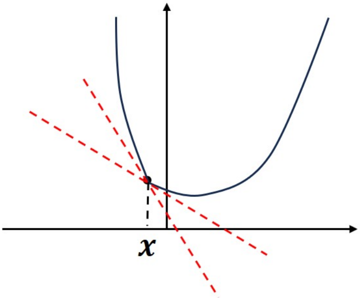
Consider the objection function $f(x) = |x|$. If we perform GD with initial point $x_0 = \frac{\eta}{2}$ and constant stepsize η, it will generate the sequence

$$\frac{\eta}{2}, -\frac{\eta}{2}, \frac{\eta}{2}, -\frac{\eta}{2}, \ldots$$



The descent directions may undergo large / discontinuous changes

Subgradient (次梯度)

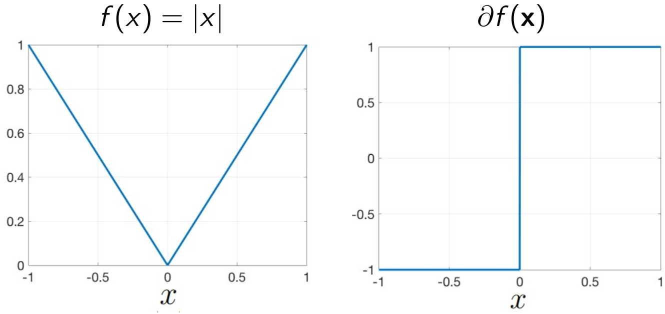


We say $\mathbf{g}$ is a subgradient of f at the point $\mathbf{x}$ if

$$f(\mathbf{y}) \geq \underbrace{f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle}_{\text{a linear under-estimate of } f}, \qquad \forall \mathbf{y} \in \mathrm{dom}\ f$$

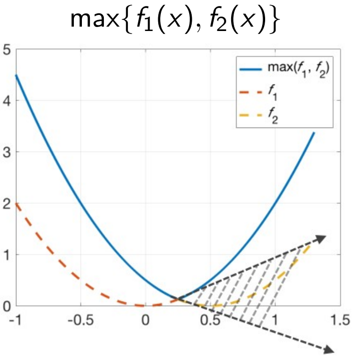
The set of all subgradients of f at $\mathbf{x}$ is called the subdifferential of f at $\mathbf{x}$, denoted by $\partial f(\mathbf{x})$.

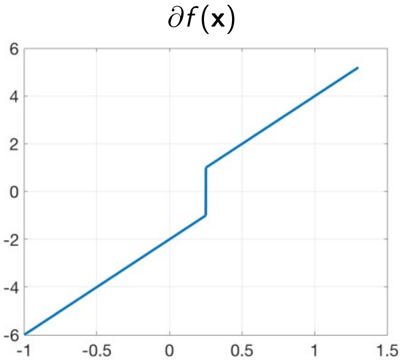
Example: $f(x) = |x|$



$$f(x) = |x| \qquad \partial f(\mathbf{x}) = \begin{cases} \{-1\}, & \text{if } x < 0 \\ [-1, 1], & \text{if } x = 0 \\ \{1\}, & \text{if } x > 0 \end{cases}$$

Example: $\max\{f_1(x), f_2(x)\}$





$f(x) = \max\{f_1(x), f_2(x)\}$ where $f_1(x)$ and $f_2(x)$ are differentiable.

$$\partial f(\mathbf{x}) = \begin{cases} \{f_1'(x)\}, & \text{if } f_1'(x) > f_2'(x) \\ [f_1'(x), f_2'(x)], & \text{if } f_1'(x) = f_2'(x) \\ \{f_2'(x)\}, & \text{if } f_1'(x) < f_2'(x) \end{cases}$$

Subgradient of Differentiable Functions

If a function is differentiable, the only subgradient at each point is the gradient.

Optimality Condition for Nondifferentiable Functions

$\mathbf{x}$ is a minimum of f if and only if the zero vector is a subgradient of f at $\mathbf{x}$.

Under strict convexity the minimum is unique.

Basic Rules of Subgradient

- **scaling:** $\partial(\alpha f) = \alpha \partial f$, for $\alpha > 0$
- **summation:** $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$

Example: Compute the subdifferential of ℓ_1 norm

$$f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_i|$$

Basic Rules of Subgradient (cont.)

- **chain rule:** suppose f is convex, and g is differentiable, nondecreasing, and convex. Let $h(\mathbf{x}) = g(f(\mathbf{x}))$, then

$$\partial h(\mathbf{x}) = g'(f(\mathbf{x}))\partial f(\mathbf{x})$$

- Suppose f is convex, and let $h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$. Then

$$\partial h(\mathbf{x}) = \mathbf{A}^\top \partial f(\mathbf{A}\mathbf{x} + \mathbf{b})$$

Example: Find a subgradient of $\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_1$.

Basic Rules of Subgradient (cont.)

- **pointwise maximum:** if $f(\mathbf{x}) = \max_{1 \le i \le k} f_i(\mathbf{x})$, then

$$\partial f(\mathbf{x}) = \mathrm{conv}\left\{ \bigcup \{ \partial f_i(\mathbf{x}) | f_i(\mathbf{x}) = f(\mathbf{x}) \} \right\}$$

- **pointwise supremum:** if $f(\mathbf{x}) = \sup_{\alpha \in \mathcal{F}} f_\alpha(\mathbf{x})$, then

$$\partial f(\mathbf{x}) = \mathrm{closure}\left(\mathrm{conv}\left\{ \bigcup \{ \partial f_\alpha(\mathbf{x}) | f_\alpha(\mathbf{x}) = f(\mathbf{x}) \} \right\} \right)$$

Example:

$$f(\mathbf{x}) = \max_{1 \le i \le k} \{ \mathbf{a}_i^\top \mathbf{x} + b_i \}$$

$$f(\mathbf{x}) = \|\mathbf{x}\|_\infty = \max_{1 \le i \le d} |x_i|$$

Subgradient Characterization of Convexity

A function f is convex if and only if $\mathrm{dom}\ f$ is convex and $\partial f(\mathbf{x}) \neq \emptyset$ for all $\mathbf{x} \in \mathrm{dom}\ f$.

Outline

Subgradient Descent Method (次梯度下降法)

In each iteration, the (projected) subgradient descent method computes

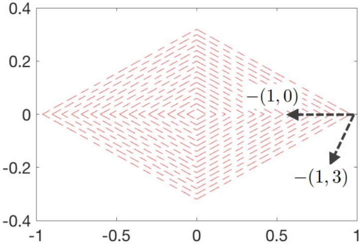
$$\mathbf{x}_{t+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{x}_t - \eta_t \mathbf{g}_t),$$

where $\mathbf{g}_t$ is any subgradient of f at $\mathbf{x}_t$.

Note: this update rule does not necessarily yield reduction w.r.t. the objective values.

Negative subgradients are not necessarily descent directions

Example: $f(\mathbf{x}) = |x_1| + 3|x_2|$



at $\mathbf{x} = (1,0)$:

- $\mathbf{g}_1 = (1,0) \in \partial f(\mathbf{x})$, $-\mathbf{g}_1$ is a descent direction;
- $\mathbf{g}_2 = (1,3) \in \partial f(\mathbf{x})$, $-\mathbf{g}_2$ is not a descent direction.

Negative subgradients are not necessarily descent directions

Since $f(\mathbf{x}_t)$ is not necessarily monotone, we will keep track of the best point

$$f_{best,t} \triangleq \min_{1 \leq i \leq t} f(\mathbf{x}_i)$$

We denote $f^* = \min_{\mathbf{x}} f(\mathbf{x})$ the optimal objective value.

Convex and Lipschitz Problems

Clearly, we cannot analyze all nonsmooth functions. Thus we start with Lipschitz continuous functions.

Remember that a function $f : \mathbb{R}^d \to \mathbb{R}$ is G-Lipschitz continuous if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq G \|\mathbf{x} - \mathbf{y}\|_2 .$$

f is G-Lipschitz continuous implies that all its subgradients $\mathbf{g}$ is bounded, i.e., $\|\mathbf{g}\|_2 \leq G$.

Polyak's Stepsize

We'd like to optimize $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2$, but don't have access to $\mathbf{x}^*$

Key idea (majorization-minimization): find another function that majorizes $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2$, and optimize the majorizing function

Lemma

Projected subgradient update rule obeys

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \le \underbrace{\underbrace{\|\mathbf{x}_t - \mathbf{x}^*\|_2^2}_{\textit{fixed}} - 2\eta_t(f(\mathbf{x}_t) - f^*) + \eta_t^2 \|\mathbf{g}_t\|_2^2}_{\textit{majorizing function}}$$

Polyak's Stepsize

The majorizing function in (4.3) suggests a stepsize (Polyak '87)

$$\eta_t = \frac{f(\mathbf{x}_t) - f^*}{\|\mathbf{g}_t\|_2^2}$$

which leads to error reduction

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \frac{(f(\mathbf{x}_t) - f^*)^2}{\|\mathbf{g}_t\|_2^2}$$

- require to know f^*
- the estimation error is monotonically decreasing with Polyak's stepsize

Convergence Rate with Polyak's Stepsize

Suppose f is convex and G-Lipschitz continuous over $\mathcal{C}$. The projected subgradient descent with Polyak's stepsize obeys

$$f_{best,t} - f^* \leq \frac{G \left\| \mathbf{x}_0 - \mathbf{x}^* \right\|_2}{\sqrt{t+1}}$$

Other Stepsize

Suppose f is convex and G-Lipschitz continuous over $\mathcal{C}$. The projected subgradient descent obeys

$$f_{best,t} - f^* \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + G^2 \sum_{k=0}^{t} \eta_k^2}{2 \sum_{k=0}^{t} \eta_k}.$$

If we choose $\eta_t = \frac{1}{\sqrt{t+1}}$, we get

$$f_{best,t} - f^* \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + G^2 \log(t)}{4\sqrt{t+1}}.$$

Strongly Convex and Lipschitz Problems

Let f be μ-strongly convex and G-Lipschitz continuous over $\mathcal{C}$. If $\eta_t = \frac{2}{\mu(t+1)}$, then the projected subgradient descent obeys

$$f_{best,t} - f^* \leq \frac{2G^2}{\mu(t+1)}.$$

Summary

condition	stepsize	convergence rate	iteration complexity
convex & smooth	$\eta_t = \frac{1}{L}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$
strongly convex & smooth	$\eta_t = \frac{1}{L}$	$O\left(\left(1-\frac{1}{\kappa}\right)^t\right)$	$O(\kappa \log \frac{1}{\varepsilon})$

Table: Convergence Properties of GD & PGD

	stepsize	convergence rate	iteration complexity
convex & smooth	$\eta_t \approx \frac{1}{\sqrt{t}}$	$O\left(\frac{1}{\sqrt{t}}\right)$	$O(\frac{1}{\varepsilon^2})$
strongly convex & smooth	$\eta_t \approx \frac{1}{t}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$

Table: Convergence Properties of Subgradient Descent

Questions