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

Review: 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: 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	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$

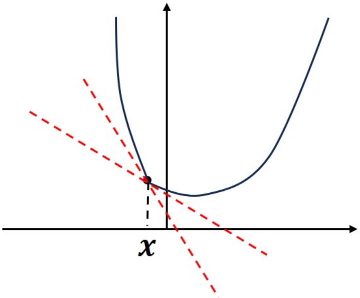
Table: Convergence Properties of GD & PGD

Can we drop the smoothness condition?

Outline

1. Subgradient
2. Subgradient descent method

Subgradient (次梯度)

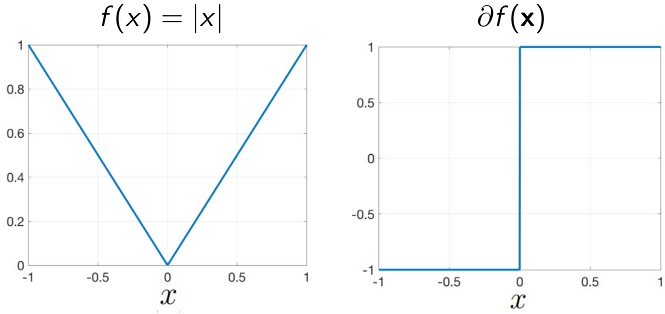


We say $\mathbf{g}$ is a subgradient of a function f at the point $\mathbf{x}$ if

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

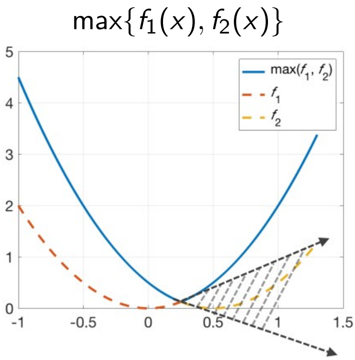
The set of all subgradients of f at $\mathbf{x}$ is called the subdifferential of f at $\mathbf{x}$, denoted by $\partial f(\mathbf{x})$.

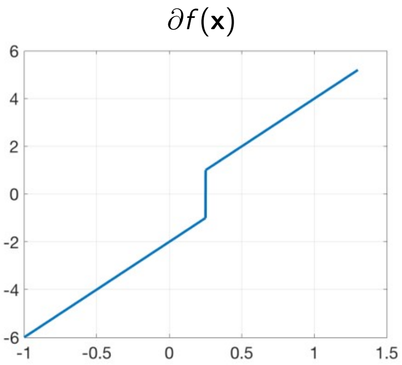
Example: $f(x) = |x|$



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

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, i.e.,

$$\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}.$$

Proof. Let $\mathbf{g}$ be a subgradient of $\mathbf{x}$ and $\mathbf{y} = \mathbf{x} + t\mathbf{h}$, we can get

$$f(\mathbf{x} + t\mathbf{h}) \geq f(\mathbf{x}) + \langle \mathbf{g}, t\mathbf{h} \rangle$$

Let $t \to 0$, we obtain

$$\lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{h}) - f(\mathbf{x})}{t} = \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle \geq \langle \mathbf{g}, \mathbf{h} \rangle.$$

Thus we have $\langle \nabla f(\mathbf{x}) - \mathbf{g}, \mathbf{h} \rangle \geq 0$.
Since this inequality holds for all $\mathbf{h}$, we must have $\nabla f(\mathbf{x}) = \mathbf{g}$.

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

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

- **affine transformation:** suppose f is convex and $h(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$. 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: Find subgradients of following functions:

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

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

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 $\mathbf{x} \in (\mathrm{dom}\ f)^{\circ}$.

Proof. "$\Leftarrow$": Suppose $\partial f(\mathbf{x}) \neq \emptyset$ and set $\mathbf{g} \in \partial f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y})$.

$$f(\mathbf{x}) \geq f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) + (1-\alpha)\langle \mathbf{g}, \mathbf{x} - \mathbf{y} \rangle \tag{1}$$

$$f(\mathbf{y}) \geq f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) - \alpha \langle \mathbf{g}, \mathbf{x} - \mathbf{y} \rangle \tag{2}$$

Then we have

$$\alpha \cdot (1) + (1-\alpha) \cdot (2) : \quad \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) \geq f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}),$$

which means f is convex.

"$\Rightarrow$" is left for homework.

Outline

1. Subgradient
2. Subgradient descent method

Subgradient descent method (次梯度下降法)

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

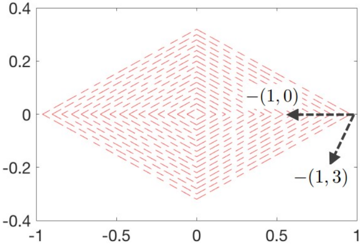
$$\mathbf{x}_{t+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{x}_t - \eta_t \mathbf{g}_t),$$

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

Remark: 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

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \leq \underbrace{\underbrace{\|\mathbf{x}_t - \mathbf{x}^*\|_2^2}_{\textit{fixed}} - 2\eta_t(f(\mathbf{x}_t) - f^*) + \eta_t^2 \|\mathbf{g}_t\|_2^2}_{\textit{majorizing function}} \tag{3}$$

Polyak's Stepsize

The majorizing function in equation (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

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

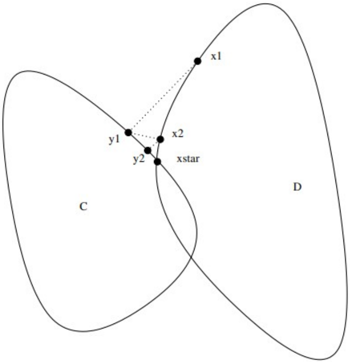
Example: projection onto intersection of convex sets

Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be closed convex sets and suppose $\mathcal{C}_1 \cap \mathcal{C}_2 \neq \emptyset$. We want to find $\mathbf{x} \in \mathcal{C}_1 \cap \mathcal{C}_2$ which is the solution of

$$\min_{\mathbf{x}} \max\{dist_{\mathcal{C}_1}(\mathbf{x}), dist_{\mathcal{C}_2}(\mathbf{x})\},$$

where $dist_{\mathcal{C}}(\mathbf{x}) \triangleq \min_{\mathbf{y} \in \mathcal{C}} \|\mathbf{x} - \mathbf{y}\|_2$.

Example: projection onto intersection of convex sets



For this problem, the subgradient method with Polyak's stepsize rule is equivalent to alternating projection

$$\mathbf{x}_{t+1} = \mathcal{P}_{\mathcal{C}_1}(\mathbf{x}_t), \qquad \mathbf{x}_{t+2} = \mathcal{P}_{\mathcal{C}_2}(\mathbf{x}_{t+1})$$

Other Stepsize

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

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

Diminishing step size: $\frac{\sum_{t=0}^{T} \eta_t^2}{\sum_{t=0}^{T} \eta_t} \to 0$ as $T \to \infty$

Other Stepsize

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

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

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

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

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

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

Without knowing $f_{best,t}$

Now we consider $\bar{\mathbf{x}}_t = \sum_{k=0}^{t} \frac{\eta_k \mathbf{x}_k}{\sum_{j=0}^{t} \eta_j}$. By Jensen's inequality, we have

$$
\begin{aligned}
\sum_{k=0}^{t} \eta_k (f(\mathbf{x}_k) - f^*) &= \left(\sum_{k=0}^{t} \eta_k \right) \left(\sum_{k=0}^{t} \frac{\eta_k}{\sum_{j=0}^{t} \eta_j} (f(\mathbf{x}_k) - f^*) \right) \\
&\geq \left(\sum_{k=0}^{t} \eta_k \right) \left(f \left(\sum_{k=0}^{t} \frac{\eta_k \mathbf{x}_k}{\sum_{j=0}^{t} \eta_j} \right) - f^* \right) \\
&= \left(\sum_{k=0}^{t} \eta_k \right) (f(\bar{\mathbf{x}}_t) - f^*)
\end{aligned}
$$

Optimal result

Suppose f is convex and G-Lipschitz continuous over $\mathcal{C}$. Suppose $\mathcal{C}$ is bounded and convex with diameter $D > 0$, i.e., $\|\mathbf{x} - \mathbf{y}\|_2 \geq D$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$. If we choose $\eta_t = \frac{D}{G\sqrt{t+1}}$, we get

$$f(\bar{\mathbf{x}}_t) - f^* \leq \frac{DG}{\sqrt{t+1}},$$

where $\bar{\mathbf{x}}_t = \sum_{k=\lceil \frac{t}{2} \rceil}^{t} \frac{\eta_k \mathbf{x}_k}{\sum_{j=\lceil \frac{t}{2} \rceil}^{t} \eta_j}$ or $\bar{\mathbf{x}}_t = \min_{\lceil \frac{t}{2} \rceil \leq i \leq t} f(\mathbf{x}_i)$.

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	$\eta_t \approx \frac{1}{\sqrt{t}}$	$O\left(\frac{1}{\sqrt{t}}\right)$	$O(\frac{1}{\varepsilon^2})$
strongly convex	$\eta_t \approx \frac{1}{t}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$

Table: Convergence Properties of Subgradient Descent