Lecture 6. Subgradient Methods

Bao Wang
Department of Mathematics
Scientific Computing and Imaging Institute
University of Utah
Math 5750/6880, Fall 2021

Loss function

So far, we have formulated training machine learning models as

$$\min f(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_i(\boldsymbol{x}) + R(\boldsymbol{x})$$

where $\boldsymbol{x}$ is the parameter of the machine learning model, $\mathcal{L}_i(\boldsymbol{x})$ is the loss of the ith training instance, and $R(\boldsymbol{x})$ is the regularization term.

How to find the optimal $\boldsymbol{x}^*$ if $R(x)$ is not differentiable everywhere, e.g. ℓ_1-regularization?

Subgradient methods or proximal gradient methods.

We will focus on the "subgradient-based methods" in this lecture, i.e.,

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

where $\boldsymbol{g}^t$ is any subgradient of f at $\boldsymbol{x}^t$.

Subgradients

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

$$f(\boldsymbol{z}) \geq \underbrace{f(\boldsymbol{x}) + \boldsymbol{g}^\top(\boldsymbol{z} - \boldsymbol{x})}_{\text{a linear under-estimate of } f}, \quad \forall \boldsymbol{z}. \tag{2}$$

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

Example

Let $f(x) = |x|$, then

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

Example (a subgradient of norms at 0)

Let $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ for any norm $\|\cdot\|$, then for any $\boldsymbol{g}$ obeying $\|\boldsymbol{g}\|_* \leq 1$, then

$$\boldsymbol{g} \in \partial f(0),$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$ (i.e. $\|\boldsymbol{x}\|_* := \sup_{\boldsymbol{z}:\|\boldsymbol{z}\|\leq 1}\langle \boldsymbol{z}, \boldsymbol{x}\rangle$).

Proof. To see this, it suffices to prove that

$$f(\boldsymbol{z}) \geq f(0) + \langle \boldsymbol{g}, \boldsymbol{z} - 0\rangle, \ \forall \boldsymbol{z} \ \Leftrightarrow \langle \boldsymbol{g}, \boldsymbol{z}\rangle \leq \|\boldsymbol{z}\|, \ \forall \boldsymbol{z}.$$

This follows from generalized Cauchy-Schwarz, i.e.

$$\langle \boldsymbol{g}, \boldsymbol{z}\rangle \leq \|\boldsymbol{g}\|_* \|\boldsymbol{z}\| \leq \|\boldsymbol{z}\|.$$

Basic rules of subgradient methods

Scaling: $\partial(\alpha f) = \alpha \partial f$ for $\alpha > 0$.

Basic rules of subgradient methods

Summation: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$.

Basic rules of subgradient methods

Affine transformation: if $h(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$, then $\partial h(\boldsymbol{x}) = \boldsymbol{A}^\top \partial f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$.

Chain rule: suppose f is convex, and g is differentiable, nondecreasing, and convex. Let $h = g \circ f$, then

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

Composition: suppose $f(\boldsymbol{x}) = h(f_1(\boldsymbol{x}), \cdots, f_n(\boldsymbol{x}))$, where f_i's are convex, and h is differentiable, nondecreasing, and convex. Let $\boldsymbol{q} = \nabla h(\boldsymbol{y})|_{\boldsymbol{y}=[f_1(\boldsymbol{x}),\cdots,f_n(\boldsymbol{x})]}$, and $\boldsymbol{g}_i \in \partial f_i(\boldsymbol{x})$. Then

$$q_i \boldsymbol{g}_1 + \cdots + q_n \boldsymbol{g}_n \in \partial f(\boldsymbol{x})$$

Basic rules of subgradient methods

Pointwise maximum: if $f(\boldsymbol{x}) = \max_{1\le i\le k} f_i(\boldsymbol{x})$, then

$$\partial f(\boldsymbol{x}) = \underbrace{conv\Big\{ \cup \{\partial f_i(\boldsymbol{x}) | f_i(\boldsymbol{x}) = f(\boldsymbol{x})\}\Big\}}_{\text{convex hull of subdifferentials of all active functions}}$$

Pointwise supremum: if $f(\boldsymbol{x}) = \sup_{\alpha \in \mathcal{F}} f_\alpha(\boldsymbol{x})$, then

$$\partial f(\boldsymbol{x}) = closure\Big(conv\Big\{ \cup \{\partial f_\alpha(\boldsymbol{x}) | f_\alpha(\boldsymbol{x}) = f(x)\}\Big\}\Big)$$

Example

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

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

Example (ℓ_1 norm)

$f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 = \sum_{i=1}^n \underbrace{|x_i|}_{:=f_i(\boldsymbol{x})}$ since

$$\partial f_i(\boldsymbol{x}) = \begin{cases} sgn(x_i)\boldsymbol{e}_i & \text{if } x_i \neq 0 \\ [-1,1]\cdot \boldsymbol{e}_i & \text{if } x_i = 0 \end{cases}$$

Note that $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 = \sum_{i,|x_i|\neq 0} |x_i|$, thus we have

$$\sum_{i,x_i\neq 0} sgn(x_i)\boldsymbol{e}_i \in \partial f(\boldsymbol{x}).$$

How about the subgradient at $\boldsymbol{x} = 0$?

Example

Let $h(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\|_1$, and denote $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ and $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m]^\top$, we have

$$\boldsymbol{g} = \sum_{i:\boldsymbol{a}_i^\top \boldsymbol{x} + b_i \neq 0} sgn(\boldsymbol{a}_i^\top \boldsymbol{x} + b_i)\boldsymbol{e}_i \in \partial f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$$

$$\Rightarrow \boldsymbol{A}^\top \boldsymbol{g} = \sum_{i:\boldsymbol{a}_i^\top \boldsymbol{x} + b_i \neq 0} sgn(\boldsymbol{a}_i^\top \boldsymbol{x} + b_i)\boldsymbol{a}_i \in \partial h(\boldsymbol{x}).$$

Example

Consider the piecewise linear function

$$f(\boldsymbol{x}) = \max_{1 \le i \le m} \{\boldsymbol{a}_i^\top \boldsymbol{x} + b_i\},$$

pick any $\boldsymbol{a}_j$ s.t. $\boldsymbol{a}_j^\top \boldsymbol{x} + b_j = \max_i \{\boldsymbol{a}_i^\top \boldsymbol{x} + b_i\}$, then

$$\boldsymbol{a}_j \in \partial f(\boldsymbol{x}).$$

Example (ℓ_∞ norm)

Let $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty = \max_{1\le i\le n} |x_i|$, if $\boldsymbol{x} \neq 0$, then pick any x_j obeying $|x_j| = \max_i |x_i|$ to obtain

$$sgn(x_j)\boldsymbol{e}_j \in \partial f(\boldsymbol{x}).$$

Negative subgradients are not necessarily descent directions

Consider $f(\boldsymbol{x}) = |x_1| + 3|x_2|$, at $\boldsymbol{x} = (1, 0)$: $\boldsymbol{g}_1 = (1, 0) \in \partial f(\boldsymbol{x})$ and $-\boldsymbol{g}_1$ is a descent direction; $\boldsymbol{g}_2 = (1, 3) \in \partial f(\boldsymbol{x})$ while $-\boldsymbol{g}_2$ is not a descent direction. This is because f is not continuous at $\boldsymbol{x}$, one can change directions significantly without violating the validity of subgradients.

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

$$f^{best,t} := \min_{1 \leq i \leq t} f(\boldsymbol{x}^i).$$

We also denote by $f^{opt} := \min_{\boldsymbol{x}} f(\boldsymbol{x})$ the optimal objective value.

Clearly, we cannot analyze all nonsmooth functions. A nice class to start with is Lipschitz functions, i.e. the set of all f obeying

$$|f(\boldsymbol{x}) - f(\boldsymbol{z})| \leq L_f \|\boldsymbol{x} - \boldsymbol{z}\|_2, \quad \forall \boldsymbol{x} \textit{ and } \boldsymbol{z}.$$

We'd like to optimize $\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2$, but do not have access to $\boldsymbol{x}^*$. The key idea is **majorization-minimization**: find another function that majorizes $\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2$, and optimize the majorizing function

Lemma 1. Subgradient update rule (1) obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \leq \underbrace{\underbrace{\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2}_{\text{fixed}} - 2\eta_t\left(f(\boldsymbol{x}^t) - f^{opt}\right) + \eta_t^2\|\boldsymbol{g}^t\|_2^2}_{\text{majorizing function}} \tag{3}$$

Proof of Lemma 1

$$\begin{aligned}
\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 &= \|\boldsymbol{x}^t - \eta_t \boldsymbol{g}^t - \boldsymbol{x}^*\|_2^2 \\
&= \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - 2\eta_t \langle \boldsymbol{x}^t - \boldsymbol{x}^*, \boldsymbol{g}^t \rangle + \eta_t^2 \|\boldsymbol{g}^t\|_2^2 \\
&\leq \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - 2\eta_t (f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)) + \eta_t^2 \|\boldsymbol{g}^t\|_2^2
\end{aligned}$$

where the last line uses the subgradient inequality

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}^t) \geq \langle \boldsymbol{x}^* - \boldsymbol{x}^t, \boldsymbol{g}^t \rangle.$$

The majorizing function in (3) suggests a step size (Polyak's stepsize rule)

$$\eta_t = \frac{f(\boldsymbol{x}^t) - f^{opt}}{\|\boldsymbol{g}_t\|_2^2}, \tag{4}$$

which leads to error reduction

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \le \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2}{\|\boldsymbol{g}^t\|_2^2}. \tag{5}$$

The algorithm is useful if f^{opt} is known, and the estimation error is monotonically decreasing with Polyak's stepsize.

Convergence of projected subgradient method with Polyak's stepsize

Theorem 1. Suppose f is convex and L_f-Lipschitz continuous. Then the projected subgradient method (1) with Polyak's stepsize rule obeys

$$f^{best,t} - f^{opt} \le \frac{L_f \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2}{\sqrt{t+1}}.$$

The rate $O(1/\sqrt{t})$ is called sublinear convergence rate.

Proof. We have seen from (5) that

$$\begin{aligned}(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2 &\le \{\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2\}\|\boldsymbol{g}^t\|_2^2 \\ &\le \{\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2\} L_f^2.\end{aligned}$$

Applying it recursively for all iterations (from 0th to tth) and summing them up yield

$$\sum_{k=0}^{t}(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*))^2 \le \left\{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2\right\} L_f^2.$$

therefore

$$(t+1)(f^{best,t} - f^{opt})^2 \le \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 L_f^2$$

which concludes the proof.

Polyak's stepsize rule requires knowledge of f^{opt}, which is often unknown a *priori*. We might often need simpler rules for setting stepsizes.

How about the other stepsize rules?

Theorem 2. [Subgradient methods for convex and Lipschitz functions] Suppose f is convex and L_f-Lipschitz continuous. Then the projected subgradient update rule (1) obeys

$$f^{best,t} - f^{opt} \leq \frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 + L_f^2 \sum_{i=0}^{t} \eta_i^2}{2 \sum_{i=0}^{t} \eta_i}.$$

Proof. Applying Lemma 1 recursively gives

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \le \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 - 2\sum_{i=0}^{t} \eta_i (f(\boldsymbol{x}^i) - f^{opt}) + \sum_{i=0}^{t} \eta_i^2 \|\boldsymbol{g}^i\|_2^2.$$

Rearranging terms, we are left with

$$\begin{aligned} 2\sum_{i=0}^{t} \eta_i (f(\boldsymbol{x}^i) - f^{opt}) &\le \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 + \sum_{i=0}^{t} \eta_i^2 \|\boldsymbol{g}^i\|_2^2 \\ &\le \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 + L_f^2 \sum_{i=0}^{t} \eta_i^2. \end{aligned}$$

Thus

$$f^{best,t} - f^{opt} \le \frac{\sum_{i=0}^{t} \eta_i (f(\boldsymbol{x}^i) - f^{opt})}{\sum_{i=0}^{t} \eta_i} \le \frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 + L_f^2 \sum_{i=0}^{t} \eta_i^2}{2\sum_{i=0}^{t} \eta_i}.$$

Other stepsize rules

Constant step size $\eta_t \equiv \eta$:

$$\lim_{t\to\infty} f^{best,t} - f^{opt} \le \frac{L_f^2 \eta}{2},$$

i.e. may converge to non-optimal points. (Note that $2\sum_{i=0}^{t} \eta_i = \infty$)

Diminishing step size obeying

$$\sum_t \eta_t^2 \le \infty \quad \text{and} \quad \sum_t \eta_t \to \infty : \quad \lim_{t\to\infty} f^{best,t} - f^{opt} = 0,$$

i.e. converges to optimal points.

Optimal choice?

$$\eta_t = \frac{1}{\sqrt{t}}, \quad f^{best,t} - f^{opt} \lesssim \frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 + L_f^2 \log t}{\sqrt{t}},$$

i.e. attains ϵ-accuracy within about $O(1/\epsilon^2)$ iterations (ignoring the log factor).

Strongly convex and Lipschitz problems

Theorem 3. [Subgradient methods for strongly convex and Lipschitz functions] Let f be μ-strongly convex and L_f-Lipschitz continuous over $\mathcal{C}$. If $\eta_t \equiv \eta = \frac{2}{\mu(t+1)}$, then

$$f^{best,t} - f^{opt} \leq \frac{2L_f^2}{\mu} \cdot \frac{1}{t+1}.$$

Proof of Theorem 3. When f is μ-strongly convex, we can improve Lemma 1 to (exercise)

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \leq (1 - \mu\eta_t)\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - 2\eta_t\Big(f(\boldsymbol{x}^t) - f^{opt}\Big) + \eta_t^2\|\boldsymbol{g}^t\|_2^2$$

$$\Rightarrow f(\boldsymbol{x}^t) - f^{opt} \leq \frac{1 - \mu\eta_t}{2\eta_t}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{1}{2\eta_t}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 + \frac{\eta_t}{2}\|\boldsymbol{g}^t\|_2^2$$

Strongly convex and Lipschitz problems (Proof Cont'd)

Since $\eta_t = 2/(\mu(t+1))$, we have

$$f(\boldsymbol{x}^t) - f^{opt} \le \frac{\mu(t-1)}{4}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{\mu(t+1)}{4}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 + \frac{1}{\mu(t+1)}\|\boldsymbol{g}^t\|_2^2$$

and hence

$$t(f(\boldsymbol{x}^t) - f^{opt}) \le \frac{\mu t(t-1)}{4}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{\mu t(t+1)}{4}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 + \frac{1}{\mu}\|\boldsymbol{g}^t\|_2^2.$$

Summing over all iterations before t, we get

$$\sum_{k=0}^{t} k\Big(f(\boldsymbol{x}^k) - f^{opt}\Big) \le 0 - \frac{\mu t(t+1)}{4}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 + \frac{1}{\mu}\sum_{k=0}^{t}\|\boldsymbol{g}^k\|_2^2 \le \frac{t}{\mu}L_f^2.$$

$$\Rightarrow f^{best,k} - f^{opt} \le \frac{L_f^2}{\mu}\frac{t}{\sum_{k=0}^{t} k} \le \frac{2L_f^2}{\mu}\frac{1}{t+1}.$$

Subgradient method summary

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

In contrast, gradient descent is much faster!