All methods we have seen so far work under the assumption that the objective function f is smooth and in particular differentiable. In this lecture, we consider nonsmooth functions.

Examples include the absolute value $f(x) = |x|$ and more generally the ℓ_1 norm $f(x) = \|x\|_1 = \sum_{i=1}^{d} |x(i)| = \sum_{i=1}^{d} \max\{x(i), -x(i)\}$,[1] as well as the so-called Rectified Linear Unit (ReLU) $f(x) = \max\{x, 0\}$. In general, the maximum of (finitely many) smooth functions is a nonsmooth function.

1 Nonsmooth optimization

Consider the problem

$$\min_{x \in \mathcal{X}} f(x). \tag{P}$$

Assumptions:

- f is M-Lipschitz continuous for some $M \in (0, \infty)$, i.e.,

$$|f(x) - f(y)| \leq M \|x - y\|, \quad \forall x, y \in \mathrm{dom}(f),$$

  under some norm $\|\cdot\|$, whose dual norm is $\|\cdot\|_*$. Here, $\|\cdot\|$ can be an arbitrary norm. Later when we discuss the projected subgradient descent method, we will restrict to the ℓ_2 norm.

- f is convex and minimized by some $x^* \in \mathrm{argmin}_{x \in \mathcal{X}} f(x)$.

- $\mathcal{X} \subseteq \mathbb{R}^d$ is closed, convex and non-empty, and we can efficiently compute projection onto $\mathcal{X}$.

In this setting, f is not necessarily differentiable. But, it is *subdifferentiable*.

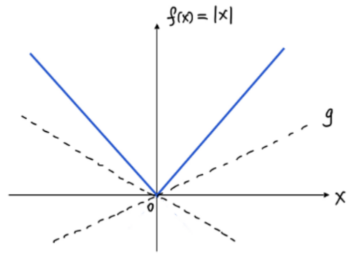
2 Subdifferentiability

Definition 1. We say that a convex function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is subdifferentiable at $x \in \mathrm{dom}(f)$ if there exists $g_x \in \mathbb{R}^d$ such that

$$\forall y \in \mathbb{R}^d : \quad f(y) \geq f(x) + \langle g_x, y - x \rangle .$$

Such a vector g_x is called a *subgradient* of f at x. The set of all subgradients of f at x is called the *subdifferential* of f at x and denoted by $\partial f(x)$.

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



$\forall y : f(y) \geq f(x) + \langle g, y - x \rangle$

$f(x) = \mathrm{ReLu}(x) = \max\{x, 0\}$. $\quad \partial f(x) = \begin{cases} \{1\} & x > 0. \\ [0, 1]. & x = 0. \\ \{0\}. & x < 0 \end{cases}$

Fact: f is cvx + differentiable $\Rightarrow \forall x,\ \partial f(x) = \{\nabla f(x)\}$.

Pf: $f'(x;d) = \langle \nabla f(x), d \rangle$. Let $g \in \partial f(x)$, we'll show $g = \nabla f(x)$.

By max formula, $\langle \nabla f(x), d \rangle = f'(x;d) \geq \langle g, d \rangle$. $\Rightarrow \langle g - \nabla f(x), d \rangle \leq 0$, $\boxed{\forall d.}$

$\|x\| = \max_{\|d\|=1} \langle x, d \rangle$. Hence $\|g - \nabla f(x)\| = \max_{\|d\|=1} \langle g - \nabla f(x), d \rangle \leq 0$.

$\Rightarrow g = \nabla f(x)$.

2.1 Optimality condition

For a differentiable convex function f, we know from previous lectures that x^* is a minimizer if and only if $\nabla f(x^*) = 0$. The following theorem provides a generalization to potentially non-differentiable functions.

Theorem 1. *For a convex function f, a point x^* is a minimizer if and only if $0 \in \partial f(x^*)$.*

Pf: $0 \in \partial f(x^*) \Leftrightarrow \forall y,\ f(y) \geq f(x^*) + \langle 0, y - x^* \rangle = f(x^*) \Leftrightarrow x^*$ is a minimizer.

2.2 Properties of subdifferential (optional)

The subdifferential has many important properties. We discuss a few of them below; see Wright-Recht Sections 8.2–8.4 for more.

Fact 1. *Every convex lower semicontinuous function is subdifferentiable everywhere on the (interior its domain.)*

Example 2. Let $I_{\mathcal{X}}(x) = \begin{cases} 0, & x \in \mathcal{X}, \\ \infty, & x \notin \mathcal{X}, \end{cases}$ be the indicator function of a closed convex nonempty set $\mathcal{X}$. Then for each $x \in \mathcal{X}$, $\boxed{\partial I_{\mathcal{X}}(x) = N_{\mathcal{X}}(x)}$, where $N_{\mathcal{X}}(x)$ is the normal cone at x.

For smooth functions, the gradient has a linearity property: $\nabla(af + bh)(x) = a\nabla f(x) + b\nabla h(x)$. A similar property holds for the subdifferential.

Fact 2 (Linearity). *For any two convex functions f, h and any positive constants a, b, we have*

$$\partial(af + bh)(x) = a\partial f(x) + b\partial(x) = \{ag + bg' : g \in \partial f(x), g' \in \partial h(x)\}$$

for x in the interior of $\mathrm{dom}(f) \cap \mathrm{dom}(g)$.

$\partial I_{\mathcal{X}}(x) = N_{\mathcal{X}}(x)$.

Pf: $\forall x \in \mathcal{X}$, $y \in \partial I_{\mathcal{X}}(x) \Leftrightarrow I_{\mathcal{X}}(z) \geq I_{\mathcal{X}}(x) + \langle y, z - x \rangle$, $\forall z \in \mathcal{X}$.

$\Leftrightarrow \langle y, z - x \rangle \leq 0$. $\forall z \in S$. $\Leftrightarrow y \in N_{\mathcal{X}}(x)$.

Exercise: $\partial f(x)$ for l_1-norm. $f(x) = \|x\|_1 = \sum_{i=1}^{d} |x_i|$

Example 3.3 (subdifferential of norms at 0). Let $f : \mathbb{E} \to \mathbb{R}$ be given by $f(\mathbf{x}) = \|\mathbf{x}\|$, where $\|\cdot\|$ is the endowed norm on $\mathbb{E}$. We will show that the subdifferential of f at $\mathbf{x} = \mathbf{0}$ is the dual norm unit ball:

$$\boxed{\partial f(\mathbf{0}) = B_{\|\cdot\|_*}[\mathbf{0}, 1] = \{\mathbf{g} \in \mathbb{E}^* : \|\mathbf{g}\|_* \le 1\}. \qquad (3.2)}$$

To show (3.2), note that $\mathbf{g} \in \partial f(\mathbf{0})$ if and only if

$$f(\mathbf{y}) \ge f(\mathbf{0}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{0} \rangle \text{ for all } \mathbf{y} \in \mathbb{E},$$

which is the same as

$$\|\mathbf{y}\| \ge \langle \mathbf{g}, \mathbf{y} \rangle \text{ for all } \mathbf{y} \in \mathbb{E}. \qquad (3.3)$$

We will prove that the latter holds true if and only if $\|\mathbf{g}\|_* \le 1$. Indeed, if $\|\mathbf{g}\|_* \le 1$, then by the generalized Cauchy–Schwarz inequality (Lemma 1.4),

$$\langle \mathbf{g}, \mathbf{y} \rangle \le \|\mathbf{g}\|_* \|\mathbf{y}\| \le \|\mathbf{y}\| \text{ for any } \mathbf{y} \in \mathbb{E},$$

implying (3.3). In the reverse direction, assume that (3.3) holds. Taking the maximum of both sides of (3.3) over all $\mathbf{y}$ satisfying $\|\mathbf{y}\| \le 1$, we get

$$\|\mathbf{g}\|_* = \max_{\mathbf{y}:\|\mathbf{y}\| \le 1} \langle \mathbf{g}, \mathbf{y} \rangle \le \max_{\mathbf{y}:\|\mathbf{y}\| \le 1} \|\mathbf{y}\| = 1.$$

We have thus established the equivalence between (3.3) and the inequality $\|\mathbf{g}\|_* \le 1$, which is the same as the result (3.2). ∎

Example 3.4 (subdifferential of the l_1-norm at 0). Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(\mathbf{x}) = \|\mathbf{x}\|_1$. Then, since this is a special case of Example 3.3 with $\|\cdot\| = \|\cdot\|_1$, and since the l_∞-norm is the dual of the l_1-norm, it follows that

$$\partial f(\mathbf{0}) = B_{\|\cdot\|_\infty}[\mathbf{0}, 1] = [-1, 1]^n.$$

In particular, when $n = 1$, then $f(x) = |x|$, and we have

$$\partial f(0) = [-1, 1].$$

The linear underestimators that correspond to -0.8, -0.3, and $0.7 \in \partial f(0)$, meaning $-0.8x$, $-0.3x$, and $0.7x$, are described in Figure 3.1. ∎

★ Unify optimality cond for $\begin{cases} \text{unconstrained} \\ \text{constrained} \end{cases}$ opt.

$-\nabla f(x) \in N_{\mathcal{X}}(x)$

$\Leftrightarrow -\nabla f(x) \in \partial I_{\mathcal{X}}(x).$

$\Leftrightarrow 0 \in \nabla f(x) + \partial I_{\mathcal{X}}(x)$

$\Leftrightarrow 0 \in \partial(f + I_{\mathcal{X}})(x)$

Subgrad & Lipschitz: Connection.

Theorem 2. *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be a convex function. f is M-Lipschitz-continuous w.r.t a norm $\|\cdot\|$ if and only if*

$$(\forall x \in \operatorname{dom}(f))\,(\forall g_x \in \partial f(x)): \quad \|g_x\|_* \le M.$$

Pf: $(\Rightarrow)$ $M\|y-x\| \ge |f(y)-f(x)| \ge f(y)-f(x) \ge \langle g_x, y-x\rangle$, $\forall x, y$.

$$M \ge \frac{\langle g_x, y-x\rangle}{\|y-x\|} = \frac{\langle g_x, u\rangle}{\|u\|}$$

set $u = y-x$, u is arbitrary. maximum for u on both sides.

$$\|g_x\|_* = \max_u \frac{\langle g_x, u\rangle}{\|u\|} \le M.$$

$(\Leftarrow)$ $\forall g_x \in \partial f(x)$, $\|g_x\|_* \le M$. $\forall x$, $g_x \in \partial f(x) \Rightarrow \forall y,\ f(y) \ge f(x) + \langle g_x, y-x\rangle$

$f(x) - f(y) \le \langle g_x, x-y\rangle \le \|g_x\|_* \|x-y\| \le M\|x-y\|$.

Switching role of x, y, $f(y) - f(x) \le M\|y-x\|$.

$\Rightarrow |f(y)-f(x)| \le M\|y-x\|$.

f is M-Lipschitz.

3 Projected subgradient descent

For the rest of the lecture, we assume f is M-Lipschitz w.r.t. the *Euclidean ℓ_2 norm* $\|\cdot\|_2$.

We consider the following projected subgradient descent (PSubGD) method:

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{y \in \mathcal{X}} \left\{ a_k \langle g_{x_k}, y - x_k\rangle + \frac{1}{2}\|y - x_k\|_2^2 \right\} = \operatorname{argmin}\left\{\frac{1}{2}\|y - x_k + g_{x_k}\|_2^2\right\}. \\ &= P_{\mathcal{X}}(x_k - a_k g_{x_k}), \end{aligned}$$

$y - (x_k - g_{x_k})$

where one may take any subgradient g_{x_k} from the set $\partial f(x_k)$, and $a_k > 0$ is the stepsize.

randomly ?

Without smoothness, we cannot get a descent lemma. In particular, it is not necessarily true that $f(x_{k+1}) \le f(x_k)$. Nevertheless, we can still argue about convergence for the (weighted) *average of the iterates*, defined as

$$x_k^{\text{out}} := \frac{1}{A_k}\sum_{i=0}^{k} a_i x_i,$$

where $A_k := \sum_{i=0}^{k} a_i$.

3.1 Convergence rate

Assumption: f is cvx + M-Lipschitz.

We follow the proof strategy introduced in the Frank-Wolfe lecture and restated below.

General strategy:

1. Maintain an upper bound $U_k \ge f(x_k^{\text{out}})$ and a lower bound $L_k \le f(x^*)$.
2. With $G_k := U_k - L_k \ge f(x_k^{\text{out}}) - f(x^*)$, show that
$$A_k G_k - A_{k-1} G_{k-1} \le E_k \implies G_k \le \frac{A_0 G_0 + \sum_{i=1}^k E_i}{A_k}.$$
3. Choose $\{a_k\}$ so that the above right hand decays to 0 fast.

why this strategy works?

lower bd: By subdifferentiable, $\forall i$, $f(x^*) \ge f(x_i) + \langle g_{x_i}, x^* - x_i \rangle$.

$$\Rightarrow f(x^*) \ge \frac{1}{A_k} \sum_{i=0}^k a_i \left(f(x_i) + \langle g_{x_i}, x^* - x_i \rangle \right) := L_k.$$

upbd: $f(x_k^{\text{out}}) = f\left(\frac{1}{A_k}\sum_{i=0}^k a_i x_i\right) \le \frac{1}{A_k}\sum_{i=0}^k a_i f(x_i) := U_k$.

$$f(x_k^{\text{out}}) - f(x^*) \le U_k - L_k := G_k. \qquad G_k = \frac{1}{A_k}\sum_{i=0}^k a_i \langle x_i - x^*, g_{x_i} \rangle$$

$$A_k G_k - A_{k-1} G_{k-1} = a_k \langle x_k - x^*, g_{x_k} \rangle$$
$$= a_k \langle g_{x_k}, x_k - x_{k+1} \rangle + a_k \langle g_{x_k}, x_{k+1} - x^* \rangle$$

$x_{k+1} = P_{\mathcal{X}}(x_k - a_k g_k)$. By minimum priciple, $\langle x_{k+1} - (x_k - a_k g_k), y - x_{k+1} \rangle \ge 0$, $\forall y \in \mathcal{X}$.

Set $y = x^*$. $\quad a_k \langle g_{x_k}, x_{k+1} - x^* \rangle \le \langle x_{k+1} - x_k, x^* - x_{k+1} \rangle$

$$= \frac{1}{2}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - \|x_{k+1} - x_k\|^2 \right)$$

minimum principle
$\Downarrow$
3 pt identity.

It follows that

$$A_k G_k - A_{k-1} G_{k-1} \le \frac{1}{2}\|x_k - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x_k\|^2 + M a_k \|x_k - x_{k+1}\|_2$$
$$\le \frac{1}{2}\|x_k - x^*\|^2 - \frac{1}{2}\|x_{k+1} - x^*\|^2 + \frac{1}{2} M^2 a_k^2$$

$$A_0 G_0 = a_0 \langle g(x_0), x^* - x_0 \rangle \le \frac{1}{2}\|x_0 - x^*\|^2 - \frac{1}{2}\|x_1 - x^*\|^2 + \frac{1}{2} M^2 a_0^2. \quad \text{(Similarly).}$$

$$\Rightarrow \quad A_K G_K = A_0 G_0 + \sum_{i=1}^{K} (A_i G_i - A_{i-1} G_{i-1}) \le \frac{1}{2}\|x_0 - x^*\|^2 - \frac{1}{2}\|x_{K+1} - x^*\|^2 + \frac{M^2}{2}\sum_{i=0}^{K} a_i^2$$

$$\le \frac{1}{2}\|x_0 - x^*\|^2 + \frac{M^2}{2}\sum_{i=0}^{K} a_i^2.$$

$$G_K \le \frac{\|x_0 - x^*\|^2 + M^2 \sum_{i=0}^{K} a_i^2}{2 A_K}$$

Set $a_i = C$. $A_K = C(K+1)$. $\sum_{i=1}^{K} a_i^2 = KC^2$. $\frac{M^2 \sum_{i=1}^{K} a_i^2}{2A_K} = \frac{M^2 K C^2}{2C(K+1)} = \frac{M^2 C}{2} \frac{K}{K+1} \le \frac{M^2 C}{2}$

const step size.

$$f(x_K^{out}) - f(x^*) \le G_K \le \frac{\|x_0 - x^*\|^2}{2C(K+1)} + \frac{M^2 C}{2}$$

RHS can be minimized when $C^2 (K+1) M^2 = \|x_0 - x^*\|^2$ $\quad C = \frac{\|x_0 - x^*\|}{\sqrt{K+1}\, M}$

$$f(x_K^{out}) - f(x^*) \le \frac{M\|x_0 - x^*\|_2}{\sqrt{K+1}}$$

$O(\frac{1}{\sqrt{K}})$ rate, slower than smooth cvx function ($O(\frac{1}{K})$)

Moreover, it analyses x_K^{out} rather than x_{K+1}.

3.2 Other considerations

The above choice of $\{a_k\}$ and the final bound require:

(i) knowing $\|x_0 - x^*\|_2$;

(ii) fixing the total number of iterations K before setting $\{a_k\}$.

(i). Define $D = \max_{x,y \in X} \|x - y\|_2$. Choose $C = \frac{D}{M\sqrt{K+1}}$. $\quad f(x_K^{out}) - f(x^*) \le \frac{MD}{\sqrt{K+1}}$

Total # iters.

(ii) Choose $a_K = \frac{D}{M\sqrt{K+1}}$. $\Rightarrow f(x_K^{out}) - f(x^*) = O\left(\frac{DM \log K}{\sqrt{K+1}}\right)$

Finally, if D is unknown or unbounded, then we can use $\boxed{a_k = \frac{1}{\sqrt{k+1}}}$. Note that this choice does not require knowledge of the Lipschitz M either. In this case we have

$$f(x_K^{\text{out}}) - f(x^*) = O\left(\frac{\left(\|x_0 - x^*\|_2^2 + M^2\right)\log K}{\sqrt{K+1}}\right).$$

4 Lower bounds (optional)

The $\boxed{O\left(\frac{1}{\sqrt{K}}\right)}$ rate above is order-wise optimal for first-order methods in a sense similar to the optimality of AGD. Consider a first-order method that generates iterates $x_1, x_2, x_3 ...$ satisfying $x_1 = 0$ and

$$x_{k+1} \in \text{Lin}\{g_1, \dots g_k\}, \quad \forall k \geq 1,$$

where $g_k \in \partial f(x_k)$ is an arbitrary subgradient at x_k. Note that the iterates x_k and x_k^{out} of PSubGD both satisfy this assumption. We have the following lower bound.

Theorem 3. *There exists a convex and M-Lipschitz function f such that for any first-order method satisfying the above assumption, we have*

$$\min_{1\leq k\leq K} f(x_k) - f(x^*) \geq \frac{M\|x^* - x_1\|_2}{2(1+\sqrt{K})}.$$

Proof. Consider a function $f : \mathbb{R}^K \to \mathbb{R}$ defined as

$$f(x) = \gamma \max_{1\leq i\leq K} x(i) + \frac{1}{2}\|x\|_2^2,$$

where $\gamma = \frac{M\sqrt{K}}{1+\sqrt{K}}$. Then

$$\partial f(x) = x + \gamma\, \text{conv}\left\{e_i : i \in \operatorname*{argmax}_{1\leq j\leq K} x(j)\right\},$$

where $e_i \in \mathbb{R}^K$ is the ith standard basis vector and conv$\{\cdot\}$ denotes the convex hull.

A minimizer of f is x^* with $x^*(i) = -\frac{\gamma}{K}, \forall i$, because $0 \in \partial f(x^*)$ (Theorem 1). Hence

$$\|x^* - x_1\|_2 = \|x^*\|_2 = \frac{\gamma}{\sqrt{K}} = \frac{M}{1+\sqrt{K}} \tag{3}$$

and the optimal value is

$$f(x^*) = -\frac{\gamma^2}{K} + \frac{1}{2}\frac{\gamma^2}{K} = -\frac{M^2}{2(1+\sqrt{K})^2}.$$

Note that if $\|x\|_2 \leq \frac{\gamma}{\sqrt{K}}$, then $\|g\|_2 \leq \frac{\gamma}{\sqrt{K}} + \gamma = M, \forall g \in \partial f(x)$. By Theorem 2 we know that f is M-Lipschitz on the ball $\left\{x : \|x\|_2 \leq \frac{\gamma}{\sqrt{K}}\right\}$.

Under our assumption for first-order methods, it is easy to see that

$$x_k \in \text{Lin}\{g_1, \dots g_{k-1}\} \subseteq \text{Lin}\{e_1, \dots, e_{k-1}\}.$$

Therefore, for all $k \leq K$, we have $x_k(K) = 0$ and thus $f(x_k) \geq 0$. It follows that the optimality gap is lower bounded as

$$f(x_k) - f(x^*) \geq 0 - \frac{M^2}{2(1+\sqrt{K})^2} = \frac{M\|x^* - x_1\|_2}{2(1+\sqrt{K})},$$

where the last step follows from (3). □