

Subgradient Methods

Example $f(x) = \|x\| = \sqrt{x^T x}$

- f is convex (by Triangle Inequality)
- For $x \neq 0$, $\nabla f(x)$ exists and

$$\begin{aligned}\nabla f(x) &= \nabla f(x) = \frac{1}{2\sqrt{x^T x}} \cdot 2x \\ &= \frac{x}{\|x\|}\end{aligned}$$

- If $x=0$, $\nabla f(x)$ does not exist

Claim $\nabla f(0) = \{g \in \mathbb{R}^n : \|g\| \leq 1\}$

Proof Need to show that for $\|g\| \leq 1$ and $\forall y \in \mathbb{R}^n$

$$f(y) = \|y\| \geq f(0) + g^T(y - 0) \quad -(1)$$

But by Cauchy-Schwarz Inequality, for $\|g\| \leq 1$

$$g^T y \leq \|g\| \|y\| \leq \|y\|, \quad \forall y \in \mathbb{R}^n$$

i.e., (1) holds $\forall y \in \mathbb{R}^n$

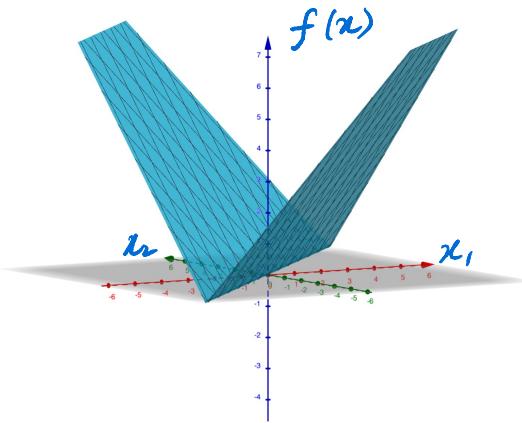
To establish the converse, suppose $\|g\| > 1$.

Then, setting $y = \frac{g}{\|g\|}$

$$\Rightarrow \|y\| = 1 \quad \text{but} \quad g^T y = \|g\| > 1$$

i.e., (1) fails to hold.

Example $f(x) = |x_1 - x_2| \leftarrow \text{convex}$



If $x_1 > x_2$ then $|x_1 - x_2| = x_1 - x_2$, and

∇f exists and $= (1, -1)$

If $x_1 < x_2$, then $|x_1 - x_2| = x_2 - x_1$, and

∇f exists and $= (-1, 1)$

Claim If $x_1 = x_2$, $\nabla f(x) = \{(a, b) : a = -b, |a| \leq 1\}$.

Proof Suppose $x_1 = x_2 = c$. Then we need to show $\forall y \in \mathbb{R}^2$, $(a, b) \text{ s.t. } a = -b, |a| \leq 1$

$$|y_1 - y_2| \geq f(c, c) + \underbrace{[a \ b]}_{\stackrel{g^T}{\rightarrow}} \underbrace{\begin{bmatrix} y_1 - c \\ y_2 - c \end{bmatrix}}_{y - c}$$

$$\therefore |y_1 - y_2| \geq ay_1 + by_2 - c(a+b) \quad -(2)$$

with $a = -b$ and $|a| \leq 1$

$$ay_1 + by_2 - c(a+b) = a(y_1 - y_2) \leq |y_1 - y_2|$$

$\Rightarrow (2)$ holds $\forall y \in \mathbb{R}^2$.

To show the converse, suppose $a \neq -b$, i.e. $a+b \neq 0$.

If $c(a+b) < 0$, setting $y_1 = y_2 = 0$

$$\Rightarrow |y_1 - y_2| = 0, \text{ and } ay_1 + by_2 - c(a+b) = -c(a+b) > 0$$

i.e., (2) fails to hold.

If $c(a+b) > 0$, then setting $y_1 = y_2 = 2c$

$$\Rightarrow |y_1 - y_2| = 0 \text{ and } ay_1 + by_2 - c(a+b) = c(a+b) > 0$$

i.e., (2) fails to hold

If $c = 0$, then setting $y_1 = y_2 = (a+b)$

$$\Rightarrow |y_1 - y_2| = 0 \text{ and } ay_1 + by_2 - c(a+b) = (a+b)^2 > 0$$

(2) fails

Finally, if $a = -b$, with $|a| > 1$, first suppose $a > 1$, then setting $y_1 = y_2 + 1$ makes (2) fail

If $a < -1$, setting $y_1 = y_2 - 1$ makes (2) fail

$$\partial f(x) = \begin{cases} (1, -1) & \text{if } x_1 > x_2 \\ (-1, 1) & \text{if } x_1 < x_2 \\ a(1, -1) \text{ with } |a| < 1 & \text{if } x_1 = x_2 \end{cases}$$

Subgradient "Descent" = Subgradient is not necessarily a descent direction, i.e. if g_k is a subgradient of f at x_k , then

$$f(x_k - \alpha g_k) \geq f(x_k) + \alpha > 0,$$

for some g_k .

Example $f(x) = |x_1| + \frac{1}{2} x_2^2$.

Suppose $x_k = (0, 1)$. Then it is easy to show:

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

Consider $g_k = (-1, 1) \in \partial f(0, 1)$

$$f(x_k - \alpha g_k) = f(0 + \alpha, 1 - \alpha) = f(\alpha, 1 - \alpha)$$

$$= |\alpha| + \frac{1}{2} (1 - \alpha)^2$$

$$= \alpha + \frac{1}{2} (1 - 2\alpha + \alpha^2)$$

$$= \frac{1}{2} (1 + \alpha^2) > \frac{1}{2} = f(x_k)$$

for all $\alpha > 0$.

i.e., $-g_k$ is not a descent direction!

If f is convex, there is some $g_k \in \partial f(x_k)$ for which $-g_k$ is a descent direction, but finding such a g_k may be difficult in high-dimensional settings.

This means, we cannot use back-tracking algorithms (Armijo's Rule) for adapting step-size.

Sub-gradient "descent" with Diminishing Step size

Assumptions

(i) f is convex on \mathbb{R}^n .

(ii) $f^* = \inf_{x \in \mathbb{R}^n} f(x)$ exists and

there exists an x^* s.t. $f(x^*) = f^*$

(iii) For all $x \in \mathbb{R}^n$ and for all $g \in \partial f(x)$,

$$\|g\| \leq a$$

$$x_{k+1} = x_k - \alpha_k g_k, \quad g_k \in \partial f(x_k)$$

Analysis

$$\begin{aligned}
 \|x_{k+1} - x^*\|^2 &= \|x_k - \alpha_k g_k - x^*\|^2 \\
 &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k g_k^T (x_k - x^*) \\
 &\leq \|x_k - x^*\|^2 + \alpha_k^2 a^2 - 2\alpha_k g_k^T (x_k - x^*)
 \end{aligned}$$

By definition of g_k ,

$$f(x_k) + g_k^T (x^* - x_k) \leq f(x^*) = f^*.$$

$$\begin{aligned}
 \Rightarrow \|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 + \alpha_k^2 a^2 + 2\alpha_k (f^* - f(x_k)) \\
 &\leq \|x_{k+1} - x^*\|^2 + \alpha_k^2 a^2 + \alpha_{k-1}^2 a^2 \\
 &\quad + 2\alpha_k (f^* - f(x_k)) + 2\alpha_{k-1} (f^* - f(x_{k-1}))
 \end{aligned}$$

Iterating starting from N and down to 0,

$$\|x_N - x^*\|^2 \leq \|x_0 - x^*\|^2 + a^2 \sum_{k=0}^{N-1} \alpha_k^2 + 2 \sum_{k=0}^{N-1} \alpha_k (f^* - f(x_k))$$

$$f_N^* = \min \{f(x_0), \dots, f(x_{N-1})\}$$

$$\|x_N - x^*\|^2 \leq \|x_0 - x^*\|^2 + a^2 \sum_{k=0}^{N-1} \alpha_k^2 + 2(f^* - f_N^*) \sum_{k=0}^{N-1} \alpha_k$$

$$\begin{aligned} \Rightarrow f_N^* - f^* &\leq \frac{\|x_0 - x^*\|^2 - \|x_N - x^*\|^2 + a^2 \sum_{k=0}^{N-1} \alpha_k^2}{2 \sum_{k=0}^{N-1} \alpha_k} \\ &\leq \frac{\|x_0 - x^*\|^2 + a^2 \sum_{k=0}^{N-1} \alpha_k^2}{2 \sum_{k=0}^{N-1} \alpha_k} \end{aligned}$$

Diminishing step-size

Suppose $\{\alpha_k\}$ is such that

$$\lim_{N \rightarrow \infty} \frac{\sum_{k=0}^{N-1} \alpha_k^2}{\sum_{k=0}^{N-1} \alpha_k} = 0$$

Then $\lim_{N \rightarrow \infty} f_N^* = f^*$.

Examples of $\{\alpha_k\}$ and Convergence Rates

1) $\alpha_k = \frac{1}{k+1}, k=0, 1, \dots$

$$\sum_{k=0}^{N-1} \alpha_k = \sum_{k=0}^{N-1} \frac{1}{k+1} = \sum_{k=1}^N \frac{1}{k} > \log N.$$

$$\sum_{k=0}^{N-1} \alpha_k^2 = \sum_{k=1}^N \frac{1}{k^2} \xrightarrow{N \rightarrow \infty} \frac{\pi^2}{6}$$

$$\Rightarrow (f_N^* - f^*) \sim O\left(\frac{1}{\log N}\right)$$

Compare with $O\left(\frac{1}{N}\right)$ for gradient descent (lec 7)

2) $\alpha_k = \frac{1}{\sqrt{k+1}}, k=0, 1, \dots$

$$\sum_{k=0}^{N-1} \alpha_k^2 = \sum_{k=1}^N \frac{1}{k} < \log N + 1$$

$$\sum_{k=0}^{N-1} \alpha_k = \sum_{k=1}^N \frac{1}{\sqrt{k}} > 2\sqrt{N} - 2$$

$$\Rightarrow f_N^* - f^* \sim O\left(\frac{\log N}{\sqrt{N}}\right)$$

Still worse than GD