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

$$\partial f(x) = \nabla f(x) = \frac{1}{2\sqrt{x^T x}} \cdot 2x$$

$$= \frac{x}{\|x\|}$$

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

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

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

$$f(y) = \|y\| \geq \underbrace{f(0)}_{=0} + g^T(y-0) \qquad -(1)$$

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 $\forall\, y \in \mathbb{R}^n$

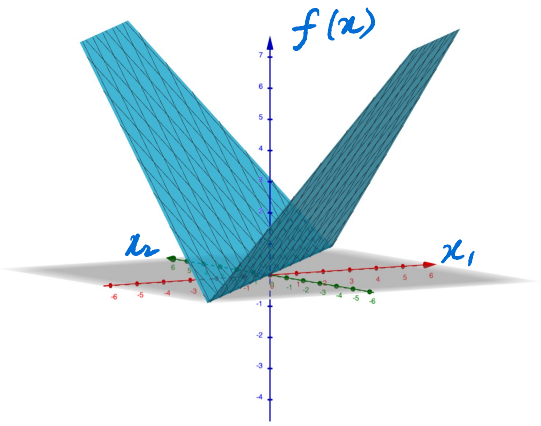
To establish the converse, suppose $\|g\| > 1$.

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

$\Rightarrow \|y\| = 1$ but $g^T y = \|g\| > 1$

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

Example $f(x) = |x_1 - x_2|$ ← convex
(by Triangle Ineq)



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 $x_1 = x_2$, $\partial f(x) = \{(a,b) : a = -b, |a| \le 1\}$.

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

$$|y_1 - y_2| \ge \underbrace{f(c,c)}_{=0} + \overbrace{\begin{bmatrix} a & b \end{bmatrix}}^{g^T} \overbrace{\begin{bmatrix} y_1 - c \\ y_2 - c \end{bmatrix}}^{y - x}$$

i.e., $|y_1 - y_2| \ge a y_1 + b y_2 - c(a+b)$ — (2)

With $a = -b$ and $|a| \le 1$

$$a y_1 + b y_2 - c(a+b) = a(y_1 - y_2) \le |y_1 - y_2|$$

$\Rightarrow$ (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 $y_1 = y_2 = 0$

$\Rightarrow |y_1 - y_2| = 0$, 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 $y_1 = y_2 = 2c$

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

i.e., (2) fails to hold

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

$\Rightarrow |y_1 - y_2| = 0$ 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

$f(x_k - \alpha g_k)$ may be $\geq f(x_k)$ $\forall \alpha > 0$.

for some g_k.

Example $f(x) = |x_1| + \frac{1}{2} x_2^2$. ← convex, min $x^* = (0,0)$

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

$$\partial f(0,1) = (\boxed{-1,1}, 1)$$

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 $\alpha > 0$.

ie, $-g_k$ 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 $g \in \partial f(x)$, $\|g\| \leq a$ ← (may not be known)

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

Analysis

$$\|x_{k+1} - x^*\|^2 = \|x_k - \alpha_k g_k - x^*\|^2$$

$$= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k g_k^T (x_k - x^*)$$

$$\leq \|x_k - x^*\|^2 + \alpha_k^2 a^2 - 2\alpha_k g_k^T (x_k - x^*)$$

By definition of g_k,

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

$$\Rightarrow \|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + \alpha_k^2 a^2 + 2\alpha_k (f^* - f(x_k))$$

$$\leq \|x_{k-1} - x^*\|^2 + \alpha_k^2 a^2 + \alpha_{k-1}^2 a^2$$

$$+ 2\alpha_k (f^* - f(x_k)) + 2\alpha_{k-1} (f^* - f(x_{k-1}))$$

Iterating starting from N and down to 0,

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

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

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

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

$$\leq \frac{\|x_0 - x^*\|^2 + a^2 \sum_{k=0}^{N-1} \alpha_k^2}{2 \sum_{k=0}^{N-1} \alpha_k}$$

Diminishing step-size

Suppose $\{\alpha_k\}$ is such that $\sum_{k=0}^{N-1} \alpha_k \to \infty$ as $N \to \infty$,

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

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

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

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

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

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

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

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

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

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

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

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

Still worse than GD