# Subgradient Method

**SIE 449/549: Optimization for Machine Learning**

Afrooz Jalilzadeh

Department of Systems and Industrial Engineering
University of Arizona

# Subgradient

- For a convex and differentiable function $f$ for all $x, y$:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

- Subgradients are motivated for the case when $f$ is **non-differentiable**, and are used to define the tightest affine function that underestimates $f$

### Definition 1 (Subgradient)

$g$ is a subgradient of a convex function $f$ at $x$ if

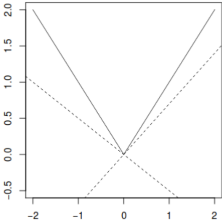
$$f(y) \geq f(x) + g^T (y - x), \quad \forall y$$

- If $f$ is indeed differentiable at $x$, then $g = \nabla f(x)$ uniquely
- The definition can hold for non-convex functions too. However, it could be possible that $g$ doesn't exist

# Examples

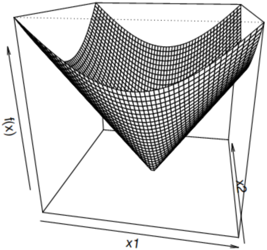
**$\mathbf{f(x) = |x|}$**

- Consider $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) = |x|$. It has one point of non-differentiability, namely at $x = 0$
- For $x \neq 0$, the subgradient is unique and is $g = \text{sign}(x)$
- For $x = 0$, the subgradient is any element of $[-1, 1]$



# Examples

$\mathbf{f(x)} = \|\mathbf{x}\|_2$



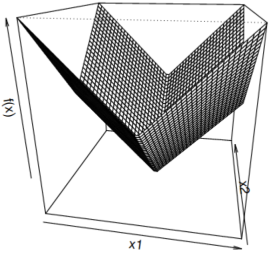

- Consider $f : \mathbb{R}^n \to \mathbb{R}$ defined as $f(x) = \|x\|_2$. It has one point of non-differentiability, namely at $x = \mathbf{0}$
- For $x \neq \mathbf{0}$, the subgradient is unique and is $g = \dfrac{x}{\|x\|_2}$
- For $x = \mathbf{0}$, the subgradient is any element of

$$\{v : \|v\|_2 \leq 1\}$$

# Examples

$\mathbf{f(x)} = \|\mathbf{x}\|_1$

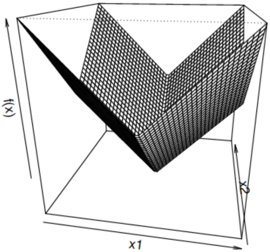
- Consider $f : \mathbb{R}^n \to \mathbb{R}$ defined as $f(x) = \|x\|_1$. It has more than one point of non-differentiabilty that is when any one of the components equal 0.
- For $x_i \neq 0$, the $i^{th}$ component of the subgradient is unique and is $g_i = \text{sign}(x_i)$
- For $x_i = 0$, the $i^{th}$ component the subgradient is any element of $[-1, 1]$
- Note that this coincides with the first example ($f(x) = |x|$) when $n = 1$



# Examples

$\mathbf{f(x)} = \max\{\mathbf{f_1(x)}, \mathbf{f_2(x)}\}$

- Consider $f : \mathbb{R}^n \to \mathbb{R}$ defined as $f(x) = \max\{f_1(x), f_2(x)\}$, where $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ are convex and differentiable
- if $f(x) = f_1(x)$ i.e., $f_1(x) > f_2(x)$, then $g$ is unique and is given by $\nabla f_1(x)$
- if $f(x) = f_2(x)$ i.e., $f_2(x) > f_1(x)$, then $g$ is unique and is given by $\nabla f_2(x)$
- if $f_1(x) = f_2(x)$, then $g$ is any point on the line segment between $\nabla f_1(x)$ and $\nabla f(x_2)$



# Subdifferentials

**Definition 2 (Subdifferential)**

The subdifferential of a convex function $f$ at $x \in \mathrm{dom}(f)$ is the collection of all subgradients of $f$ at $x$

$$\partial f(x) = \{g : f(y) \geq f(x) + g^T(y - x)\}$$

Some properties of the subdifferential:

- For convex $f$, $\partial f(x) \neq \emptyset$. However, for concave $f$, $\partial f(x) = \emptyset$
- $\partial f(x)$ is closed and convex for any $f$
- $\partial f(x) = \{\nabla f(x)\}$ when $f$ is differentiable at $x$

## Subgradient calculus

- Positive scaling: $\partial(\alpha f) = \alpha \partial f$ if $\alpha > 0$
- Addition: $\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$ for $x \in \text{dom}(f_1) \bigcap \text{dom}(f_2)$
- Affine composition: Let $g(x) = f(Ax + b)$, then $\partial g(x) = A^T \partial f(Ax + b)$
- Norms: To each norm $\|\cdot\|$, there is a dual norm $\|\cdot\|_*$ such that:

$$\|x\| = \max_{\|z\|_* \leq 1} z^T x$$

if $f(x) = \|x\|_p$, consider $q$ satisfying the relation $\dfrac{1}{p} + \dfrac{1}{q} = 1$, then:

$$\|x\|_p = \max_{\|z\|_q \leq 1} z^T x$$

Also, $\partial f(x) = \underset{\|z\|_q \leq 1}{\text{argmax}}\ z^T x$

# Gradient Method

- Consider the following optimization problem

$$\min_{x} \; f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and **differentiable**

- Gradient descent: choose initial $x_0 \in \mathbb{R}^n$, repeat:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

If $\nabla f(x)$ is Lipschitz, gradient descent has convergence rate $\mathcal{O}(1/\epsilon)$

- What if $f$ is not differentiable?

## Subgradient Method

- Replacing gradients with subgradients: choose initial $x_0 \in \mathbb{R}^n$, repeat:

$$x_{k+1} = x_k - \alpha_k g_k$$

  where $g_k \in \partial f(x_k)$, any subgradient of $f$ at $x_k$

- **Constant stepsize**: $\alpha_k = \alpha > 0$
- **Diminishing stepsize**: $\alpha_k$ satisfies the following two conditions

$$\lim_{k\to\infty} \alpha_k = 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

- Subgradient method is not necessarily a descent method
- Thus, the best solution among all of the iterations is used as the final solution:

$$f(x_k^{best}) = \min_{i=0,\ldots,k} f(x_i)$$

## Subgradient is not necessarily descent

**Example.** $f(x) = \max\left\{x_1^2 + (x_2+1)^2, x_1^2 + (x_2-1)^2\right\}$ at point $x = (1,0)$

### Lemma 3

*Assume $f : \mathbb{R}^n \to \mathbb{R}$ is convex, and L-Lipschitz, i.e., $|f(x) - f(y)| \leq L\|x - y\|$, then $\|g\| \leq L$, for any $g \in \partial f(x)$ and $x \in \mathbb{R}^n$.*

# Convergence Result

## Lemma 4

*Suppose $f$ is convex and Lipschitz continuous with constant $L$, and $\{x_k\}$ be the sequence generated by subgradient method, then:*

$$f(x_T^{best}) - f(x^*) \leq \frac{\|x_0 - x^*\|^2 + L^2 \sum_{k=0}^{T-1} \alpha_k^2}{2 \sum_{k=0}^{T-1} \alpha_k}$$

## Convergence Result

### Theorem 5

*For fixed step size $\alpha$:*

$$\lim_{T\to\infty} f(x_T^{best}) = f(x^*) + L^2\alpha/2$$

- Note that with fixed step size, the optimal value is not achieved in the limit. Smaller fixed step sizes will be reduce the gap between $f(x_T^{best})$ and $f(x^*)$

### Theorem 6

*For diminishing step size:*

$$\lim_{T\to\infty} f(x_T^{best}) = f(x^*)$$

## Convergence Rate

### Theorem 7

*Suppose $f$ is convex and Lipschitz continuous with constant $L$, and $\{x_k\}$ be the sequence generated by subgradient method. Choose stepsize $\alpha = \epsilon/L^2$, then*

$$f(x_T^{best}) - f(x^*) \leq \mathcal{O}(1/\epsilon^2).$$

# Projected Subgradient Method

- Subgradient method has convergence rate $\mathcal{O}(1/\epsilon^2)$, compare this to $\mathcal{O}(1/\epsilon)$ rate of gradient descent

- Consider $\min_{x\in C} \ f(x)$ where $C$ is a closed convex subset of $\mathbb{R}^n$:

$$x_{k+1} = P_C\left(x_k - \alpha_k g_k\right)$$

  where $g_k \in \partial f(x_k)$, any subgradient of $f$ at $x_k$

- Same convergence guarantees as the usual subgradient method, with the same step size choices