# LOG RG on Convex Optimization

## Week 1

Jinseok Chung, Donghyun Oh

January 11, 2023

# Table of Contents

# Table of Contents

# Linear (in)dependence

A set of vectors $\{v_1, v_2, \cdots, v_n\}$ in a vector space $\mathbb{V}$ is called **linearly independent** if the linear combination

$$\sum_{i=1}^{n} \alpha_i v_i = 0$$

implies that $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$. If not, is called **linearly dependent**.

# Span and Basis

For a set of vectors $\{v_1, v_2, \cdots, v_n\}$ in a vector space $\mathbb{V}$, the set of linear combinations of vectors in $\mathbb{V}$ is called the **span** of $\{v_1, v_2, \cdots, v_n\}$. i.e,

$$\textbf{span}\{v_1, v_2, \cdots, v_n\} = \{\sum_{i=1}^{n} \alpha_i v_i | \forall \alpha_i \in \mathbb{R}, i = 1, 2, \cdots, n\}$$

A **basis** of a vector space $\mathbb{V}$ is an independent set of vectors that spans $\mathbb{V}$.

# Matrix rank

Let $A \in \mathbb{R}^{d \times m}$ be a matrix, the **column rank** of $A$ is defined as the number of linearly independent columns, similar the **row rank** of $A$ is defined as the number of linear independent rows

- Row and column ranks are always the same for given matrix
- So we call it simply rank

# (vector)Subspace

Let $\mathbb{V} = (V, +, \cdot)$ be a vector space and $\emptyset \neq U \subseteq V$. Then $\mathbb{U} = (U, +, \cdot)$ is called vector **subspace** of $\mathbb{V}$ if $\mathbb{U}$ is closed under $(+, \cdot)$ operations.

- $0 \in \mathbb{V}$ always belongs to any subspaces
- Lines and planes through the $0 \in \mathbb{R}^3$ are subspaces in $\mathbb{R}^3$
- The intersection of arbitrarily many subspaces is a subspace itself

# Affine subspace

Let $\mathbb{V}$ be a vector space, $x \in \mathbb{V}$ and $\mathbb{U} \subseteq \mathbb{V}$ a subspace. Then the subset

$$L = x + \mathbb{U} := \{x + u |\ u \in \mathbb{U}\}$$

is called **affine subspace** $\mathbb{V}$.

- An affine subspace excludes 0 if $x \notin \mathbb{U}$
- Points, lines and planes are affine subspaces in $\mathbb{R}^3$
- In $\mathbb{R}^d$, the (d-1)-dimensional affine subspaces are called hyperplanes.

# Dot / Inner product

A **dot product** of $x = (x_1, x_2, \cdots, x_d), y = (y_1, y_2, \cdots, y_d) \in \mathbb{R}^d$ is defined as

$$x \cdot y = \sum_{i=1}^{d} x_i y_i$$

# Dot / Inner product

A **dot product** of $x = (x_1, x_2, \cdots, x_d), y = (y_1, y_2, \cdots, y_d) \in \mathbb{R}^d$ is defined as

$$x \cdot y = \sum_{i=1}^{d} x_i y_i$$

An **inner product** of a real scalar vector space $\mathbb{V}$ is a function of vector pairs $x, y \in \mathbb{V}$, which is denoted by $\langle x, y \rangle$ and satisfies the following three properties:

- (commutativity) $\langle x, y \rangle = \langle y, x \rangle$ for any $x, y \in \mathbb{V}$.
- (linearity) $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$ for any $\alpha, \beta \in \mathbb{R}$ and $x, y, z \in \mathbb{V}$.
- (positive definiteness) $\langle x, x \rangle \geq 0$ for any $x \in \mathbb{V}$ and $\langle x, x \rangle = 0$ if and only if $x = 0$.

# Dot / Inner product

A **dot product** of $x = (x_1, x_2, \cdots, x_d), y = (y_1, y_2, \cdots, y_d) \in \mathbb{R}^d$ is defined as

$$x \cdot y = \sum_{i=1}^{d} x_i y_i$$

An **inner product** of a real scalar vector space $\mathbb{V}$ is a function of vector pairs $x, y \in \mathbb{V}$, which is denoted by $\langle x, y \rangle$ and satisfies the following three properties:

- (commutativity) $\langle x, y \rangle = \langle y, x \rangle$ for any $x, y \in \mathbb{V}$.
- (linearity) $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$ for any $\alpha, \beta \in \mathbb{R}$ and $x, y, z \in \mathbb{V}$.
- (positive definiteness) $\langle x, x \rangle \geq 0$ for any $x \in \mathbb{V}$ and $\langle x, x \rangle = 0$ if and only if $x = 0$.

- A dot product is an inner product but the reverse is not true
- For $x = (x_1, x_2), y = (y_1, y_2)$, operation $\langle x, y \rangle = x_1 y_1 + 2 x_2 y_2$ is also an inner product

# Outer product

The outer product is operation between two vectors $x \in \mathbb{R}^d, y \in \mathbb{R}^m$ defined as

$$x \otimes y = xy^T \in \mathbb{R}^{d \times m}$$

- Rank of outer product of two vectors is 1

# Norms

A norm $||\cdot||$ on a vector space $\mathbb{V}$ is a function $||\cdot|| : \mathbb{V} \to \mathbb{R}$ satisfying the following properties:

- (nonnegativity) $||x|| \geq 0$ for any $x \in \mathbb{V}$ and $||x|| = 0$ if and only if $x = 0$
- (positive homogeneity) $||\alpha x|| = |\alpha| \cdot ||x||$ for any $x \in \mathbb{V}$ and $\alpha \in \mathbb{R}$.
- (triangle inequality) $||x + y|| \leq ||x|| + ||y||$ for any $x, y \in \mathbb{V}$

## $\ell_p$ Norms

For a $p \geq 1$, the $\ell_p$-norm on $\mathbb{R}^d$ is given by the formula

$$||x||_p = \sqrt[p]{\sum_{i=1}^{d} |x_i|^p}$$

For a $p = \infty$, the $\ell_\infty$-norm on $\mathbb{R}^d$ is given by

$$||x||_\infty = \max_{i=1,2,\cdots,d} |x_i|$$

# $\ell_p$ Norms

For a $p = 0$, the $\ell_0$-norm on $\mathbb{R}^d$ is given by

$$\|x\|_0 = \# \text{ of non zero components}$$

- e.g) $x = (2, 0, 3), \|x\|_0 = 2$
- This assumes $0^0 = 1$
- In fact, $\ell_0$ norm is not a norm. Because it doesn't satisfy the positive homogeneity

# Induced norm

Any inner product induces a norm, defined as

$$||x|| := \sqrt{\langle x, x \rangle}$$

which is called **induced norm** by the given inner product $\langle \cdot, \cdot \rangle$

- Induced norms satisfy the properties of norms
- $\ell_2$ norm is induced norm by dot product
- Not all the norms are induced norm, e.g, $\ell_1$-norm

# Angle

An **angle** $\omega$ of two vectors $x, y \in \mathbb{V}$ equipped with $\langle \cdot, \cdot \rangle$ is defined by

$$\omega = \frac{\langle x, y \rangle}{||x|| \; ||y||}$$

where the $|| \cdot ||$ is induced norm

- $-1 \leq \omega \leq 1$
- The two vectors have different angles depending on which inner product is used
- $arccos(w)$ is an angle of two vectors in radian

## Cauchy–Schwarz inequality

For vector space $\mathbb{V}$, an inner product $\langle \cdot, \cdot \rangle$ and its induced norm $|| \cdot ||$ satisfies the **Cauchy-Schwarz inequality**

$$|\langle x, y \rangle| \leq ||x|| \; ||y||$$

# Hölder inequality

For vector space $\mathbb{V}$ and $p, q \in [1, \infty]$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$, the $||\cdot||_p, ||\cdot||_q$ satisfy the **Hölder inequality**

$$x \cdot y \leq ||x||_p \, ||y||_q$$

- The pair $(p, q)$ are called Hölder conjugates of each other
- Cauchy–Schwarz inequality is the case of $p = q = 2$

# Minkowski inequality

For vector space $\mathbb{V}$ and $p \in [1, \infty]$, the $||\cdot||_p$ satisfies the **Minkowski inequality**

$$||x + y||_p \leq ||x||_p + ||y||_p$$

- From this inequality, $\ell_p$-norms satisfy the triangle inequality property

# Young's inequality

For $a, b \geq 0$ and $p, q > 1$ s.t $\frac{1}{p} + \frac{1}{q} = 1$, the **Young's inequality** is as follows

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

Equality holds if and only if $a^p = b^q$

# Eigenvalues and eigenvectors

Let $A \in \mathbb{R}^{d \times d}$ be a square matrix. Then $\lambda \in \mathbb{R}$ is an **eigenvalue** of $A$ and $x \in \mathbb{R}^d/\{0\}$ is the corresponding **eigenvector** of $A$ if

$$Ax = \lambda x$$

- There are at most d eigenvalues(and corresponding eigenvectors)
- For any symmetric (real)matrix $A$, all its eigenvalues are real
- All eigenvectors of a symmetric (real)matrix are orthogonal to each other

# Eigenspace and Eigenspectrum

The set of all eigenvectors of $A$ associated with an eigenvalue $\lambda$ spans a subspace of $\mathbb{R}^d$, which is called the **eigenspace** of $A$ with respect to $\lambda$ and is denoted by $E_\lambda$

The span of all the eigenvectors of $A$ is called the **eigenspectrum** of $A$

# Positive (semi)definite matrices

The given square matrix $A \in \mathbb{R}^{d \times d}$ is called **positive semidefinite**, if for any $x \in \mathbb{R}^d$

$$x^T A x \geq 0$$

when the inequality holds strictly, the matrix called **positve definite**

- For any matrix $A$, $A^T A$ is symmetric and positive semidefinite
- For any square matrix $A$,
  positive definiteness(respectively, semidefiniteness) $\iff$
  every eigenvalues of $A$ are positive (respectively, nonnegative)

# Matrix Norms

Frobenius norm is norm of any $m \times n$ matrix A defined as square root of component wise squared sum

$$\|A\|_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

- $\|A\|_F = \sqrt{Tr(AA^T)}$

# Matrix Norms

Vector $\ell_p$ norm induced matrix $\ell_p$ norm is norm of any $m \times n$ matrix A defined as

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

- Matrix $\ell_1$ norm is the maximum absolute column sum of the matrix
- Matrix $\ell_\infty$ norm is the maximum absolute row sum of the matrix

# Matrix Norms

- For example, for $A = \begin{bmatrix} -3 & 5 & 7 \\ 2 & 6 & 4 \\ 0 & 2 & 8 \end{bmatrix}$ we have that

$$\|A\|_1 = \max(|-3| + 2 + 0; 5 + 6 + 2; 7 + 4 + 8) = \max(5, 13, 19) = 19$$
$$\|A\|_\infty = \max(|-3| + 5 + 7; 2 + 6 + 4; 0 + 2 + 8) = \max(15, 12, 10) = 15$$

- Matrix $\ell_2$ norm, also called spectral norm, is the largest singular value of $A$

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sigma_{\max}(A)$$

- $\|A\|_2^2 = \|A^T A\|_2 = \|AA^T\|_2$
- $\|A\|_2 \leq \|A\|_F$

# Table of Contents

# Gradient

For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, $x \in \mathbb{R}^d$ and $x = (x_1, x_2, \cdots, x_d)$, the collection of partial derivatives is called the **gradient** of $f$ defined as

$$\nabla_x f = grad\ f = \frac{df}{dx} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix} \in \mathbb{R}^d$$

- It has the steepest ascending direction infinitesimally, similarly the opposite is the steepest descending direction.

# Hessian

For a twice continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, $x \in \mathbb{R}^d$ and $x = (x_1, x_2, \cdots, x_d)$, the collection of second-order partial derivatives is called the **Hessian** of $f$ defined as

$$\nabla_x^2 f = H\ f = \frac{d^2 f}{dx^2} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_d} & \frac{\partial^2 f}{\partial x_2 \partial x_d} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \in \mathbb{R}^{d \times d}$$

- If the above function is twice continuously differentiable, the Hessian matrix is always a real symmetric matrix
- The eigenvectors and eigenvalues of Hessian is directly related to curvature of the given function

# Table of Contents

# Convex Sets: Definition

## Definition

A set $C \subseteq \mathbb{R}^n$ is **convex** if

$$\forall x_1, x_2 \in C, \forall \lambda \in [0,1] \Rightarrow \lambda x_1 + (1-\lambda) x_2 \in C$$

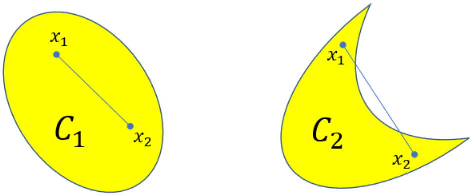


Figure: Examples of convex and non-convex sets

# Examples of Convex Sets

- **Simplex**: A simplex $\{x \in \mathbb{R}^n \mid \sum_{i=1}^n x_i = 1 \text{ and for all } i = 1, 2, ..., n, x_i \geq 0\}$ in $\mathbb{R}^n$ is a convex set. For any $x, y$ in the simplex, and $\lambda \in [0, 1]$,

$$\sum_{i=1}^n (\lambda x + (1-\lambda)y)_i = \lambda \sum_{i=1}^n x_i + (1-\lambda) \sum_{i=1}^n y_i = 1$$

  and each element of $\lambda x + (1-\lambda)y$ is still non-negative.

- **Set of psd matrices**: A set of $n \times n$ positive semidefinite (psd) matrices, denoted by $S_+^n$, is convex. Take $M_1, M_2 \in S_+^n$. Then for all $x \in \mathbb{R}^n, \lambda \in [0, 1]$,

$$x^T(\lambda M_1 + (1-\lambda)M_2)x = \lambda(x^T M_1 x) + (1-\lambda)(x^T M_2 x) \geq 0$$

  So $\lambda M_1 + (1-\lambda)M_2 \in S_+^n$.

- **Set of copositive matrices**: An $n \times n$ matrix $M$ is copositive if $x^T M x \geq 0$ for any $x \in \mathbb{R}_+^n$. We can show in the same way as above that the set of copositive matrices is a convex set. Note that since a psd matrix is always copositive, the set of psd matrices is included in the set of copositive matrices.

# Hyperplane and Half-spaces

## Definition

In $\mathbb{R}^n$, given some $s \in \mathbb{R}^n$ and $b \in \mathbb{R}$, we define a **hyperplane** as

$$H_{s,b} = \{x \in \mathbb{R}^n \mid s^T x = b\}$$

Here, $s$ is called the **normal vector** of $H_{s,b}$.
Moreover, a hyperplane $H_{s,b}$ divides $\mathbb{R}^n$ into two **half-spaces**

$$H^-_{s,b} = \{x \in \mathbb{R}^n \mid s^T x \le b\}, \;\; H^+_{s,b} = \{x \in \mathbb{R}^n \mid s^T x \ge b\}$$

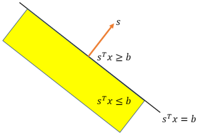


Figure: Hyperplane with normal vector, two half-spaces

# Hyperplane and Convexity

A convex set can be "carved out" from half-spaces. Formally, a closed convex set is the intersection of every closed half-spaces that contain the set. This property is equivalent to the separating hyperplane theorem.

## Theorem

**Separating hyperplane theorem**: *Let $\mathcal{X} \subset \mathbb{R}^n$ be a closed convex set, and $x_0 \in \mathbb{R}^n \setminus \mathcal{X}$. Then, there exists $w \in \mathbb{R}^n$ and $t \in \mathbb{R}$ such that*

$$\langle w, x_0 \rangle < t, \text{ and } \forall x \in \mathcal{X}, \langle w, x \rangle \geq t$$

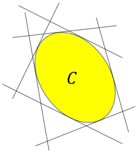


Figure: Convex Set can be seen as intersection of half-spaces

# Example of Separating Hyperplanes

Separating hyperplane between the set of psd matrices and a symmetric matrix $\hat{M}$ that is not psd.

Since $\hat{M}$ is symmetric, it has the eigenvalue decomposition $\hat{M} = \sum_i \hat{\lambda}_i \hat{v}_i \hat{v}_i^T$, where the eigenvectors are orthonormal. Since it is not psd, there is $i \in \{1, 2, ..., n\}$ such that $\hat{\lambda}_i < 0$. For simplicity, assume that $i = 1$. Now let $s = \hat{v}_1 \hat{v}_1^T$ and $b = 0$. Then, we have

$$
\begin{aligned}
\langle s, \hat{M} \rangle &= \langle \hat{v}_1 \hat{v}_1^T, \hat{M} \rangle = tr(\hat{M} \hat{v}_1 \hat{v}_1^T) \\
&= tr(\hat{v}_1^T M \hat{v}_1) \\
&= \hat{v}_1^T \hat{M} \hat{v}_1 \\
&= \hat{v}_1^T (\sum_i \hat{\lambda}_i \hat{v}_i \hat{v}_i^T) \hat{v}_1 = \hat{v}_1^T (\hat{\lambda}_1 \hat{v}_1) = \hat{\lambda}_1 < 0
\end{aligned}
$$

This implies that $\hat{M} \in H^-_{s,b}$.

For any $M \in S^n_+$, $\langle s, M \rangle = \langle \hat{v}_1 \hat{v}_1^T, M \rangle = \hat{v}_1^T M \hat{v}_1 \geq 0$ since $M$ is psd. So, $S^n_+ \subseteq H^+_{s,b}$, and $H_{s,b}$ is the separating hyperplane.

# Cones and Polar Cones

## Definition

A set $K$ is called a **cone** if

$$\forall x_1, x_2 \in K, \forall \alpha_1, \alpha_2 \geq 0 \Rightarrow \alpha_1 x_1 + \alpha_2 x_2 \in K$$

Given a cone $K$, a **polar cone** of $K$, $K^o$ is also a cone defined as

$$K^o = \{z \mid \langle z, x \rangle \leq 0, \forall x \in K\}$$

**Note**. A cone is always convex. Any subspace is a cone, but not vice versa.

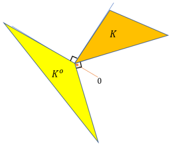


Figure: A cone and its polar cone

# Tangent Cones and Normal cones

**Definition**

Given a set $\mathfrak{X}$ and a point $x \in \mathfrak{X}$, a **tangent cone** of $\mathfrak{X}$ at $x$, denoted as $T_{\mathfrak{X}}(x)$, is informally the set of directions $x$ can move inside $\mathfrak{X}$.

**Definition**

A **normal cone** of $\mathfrak{X}$ at $x$, denoted as $N_{\mathfrak{X}}(x)$, is a polar cone of the tangent cone of $\mathfrak{X}$ at $x$.

# Tangent Cones and Normal Cones

- If $x$ is an interior point of $\mathfrak{X}$, then $T_{\mathfrak{X}}(x) = \mathbb{R}^n$ and $N_{\mathfrak{X}}(x) = \{0\}$.
- If $x$ is a "smooth" boundary point, then $T_{\mathfrak{X}}(x)$ is the half-space including $\mathfrak{X}$ and $N_{\mathfrak{X}}(x) = \{s\}$ where the half-space and normal vector $s$ are from the supporting hyperplane $H_{s,b}$ of $\mathfrak{X}$ at $x$.
- If $x$ is a "non-smooth" boundary point of $\mathfrak{X}$, then $T_{\mathfrak{X}}(x)$ and $N_{\mathfrak{X}}(x)$ are shown below.

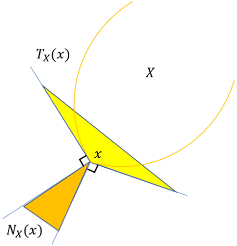


Figure: A tangent cone and its normal cone at non-smooth boundary point

# Table of Contents

# Convex Functions: Definitions

**Definition**

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if

$$\forall x, y \in \mathbb{R}^n, \forall \lambda \in [0,1] \Rightarrow f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

**Definition**

Suppose a function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then it is convex if given any $x \in \mathbb{R}^n$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \text{ for all } y \in \mathbb{R}^n$$

**Definition**

Suppose a function $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable. Then it is convex if for any $x \in \mathbb{R}^n$, the Hessian $\nabla f^2(x)$ is positive semidefinite.

# Example: Quadratic Function

We will show that the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = x^T Qx$ where $Q \succcurlyeq 0$, is convex using the three definitions.

**Definition 1**:

$$\{\lambda f(x) + (1-\lambda)f(y)\} - f(\lambda x + (1-\lambda)y) = (\lambda - \lambda^2)(y-x)^T Q(y-x) \geq 0$$

**Definition 2**: $\nabla f(x) = 2Qx$. Then

$$f(y) - \{f(x) + \langle \nabla f(x), y - x\rangle\} = (y-x)^T Q(y-x) \geq 0$$

for all $y \in \mathbb{R}^n$.

**Definition 3**: $\nabla^2 f(x) = 2Q \succcurlyeq 0$.

# Example: Maximum of Convex Function

The maximum function of convex functions is convex. This can be shown from Definition 1.

- A function returning the largest element: A function $f : \mathbb{R}^n \to \mathbb{R}$ defined as $f(x) = f(x_1, x_2, ..., x_n) = max(x_1, x_2, ..., x_n) = max(e_1^T x, e_2^T x, ..., e_n^T x)$ is convex since each $e_i^T x$ is linear and hence convex.
- Maximum eigenvalue of symmetric matrix: For a symmetric matrix $Q$, $f(Q) = \lambda_{max}(Q)$. We show that $f$ is convex. Recall that if $Q$ is symmetric, then $x^T Qx \leq \lambda_{max}||x||_2^2$ and the equality holds when $x$ is the eigenvector corresponding to $\lambda_{max}$. So $\lambda_{max} = \sup x^T Qx$ subject to $||x||_2 = 1$, and since each $x^T Qx$ is a linear function of $Q$, (observe that $x^T Qx = \langle xx^T, Q\rangle$), $f$ is a convex function.

# Subgradients and Subdifferentials

The second definition of convex functions can be extended to non-differentiable convex functions.

**Definition**

Given a function $f : \mathbb{R}^n \to \mathbb{R}$, it is convex if given any $x \in \mathbb{R}^n$, there exists $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + \langle g, y - x \rangle \text{ for all } y \in \mathbb{R}^n$$

Note that if $f$ is convex and differentiable, $g = \nabla f(x)$ is unique $g$ that satisfy the above inequality.

**Definition**

For a convex function $f : \mathbb{R}^n \to \mathbb{R}$ and a point $x \in \mathbb{R}^n$, a vector $g \in \mathbb{R}^n$ such that $f(y) \geq f(x) + \langle g, y - x \rangle$ for all $y \in \mathbb{R}^n$ is called a **subgradient** at $x$. A set of subgradients at $x$ is called the **subdifferential** of $f$ at $x$ and is denoted as $\partial f(x)$.

# Monotone Property of Gradient

**Theorem**

*$f$ is convex according to Definition 2 if and only if it's gradient has the monotone property, that is,* $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$.

**Proof.**

First, assume that $f$ is convex according to Definition 2.
(i) Then, by Definition 2, we have two inequalities

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \;\; f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$$

for all $x, y \in \mathbb{R}^n$. Adding them gives $0 \geq \langle \nabla f(x) - \nabla f(y), y - x \rangle$.

$\square$

# Monotone Property of Gradient

Proof.

Now assume that $\nabla f$ has the monotone property.

(ii) Define a function $g : [0,1] \to \mathbb{R}$ as $g(t) = f(tx + (1-t)y) = f(y + t(x-y))$. Its gradient is given as $g'(t) = [\nabla f(y + t(x-y))]^T (x-y) = \langle \nabla f(y + t(x-y)), x-y \rangle$. By the Fundamental Theorem of Calculus, we have

$$\int_0^1 g'(t)dt = g(1) - g(0) = f(x) - f(y) \text{ or } f(x) = f(y) + \int_0^1 g'(t)dt$$

(iii) We claim $g'(t)$ is minimized at $t = 0$. By the monotone property,

$$\begin{aligned}
&\langle \nabla f(y + t(x-y)) - \nabla f(y), y + t(x-y) - y \rangle \\
&= \langle \nabla f(y + t(x-y)), t(x-y) \rangle - \langle \nabla f(y), t(x-y) \rangle \\
&= t(g'(t) - g'(0)) \geq 0
\end{aligned}$$

Proof.

Therefore, $g'(t)$ has its minimum at $t = 0$. (iv) From the results of (ii) and (iii),

$$\begin{aligned} f(x) &= f(y) + \int_0^1 g'(t)dt \\ &\geq f(y) + \int_0^1 g'(0)dt \\ &= f(y) + g'(0) \\ &= f(y) + \langle \nabla f(y), x - y \rangle \end{aligned}$$

and $f$ is convex according to Definition 2. $\square$

# Table of Contents

# Optimality Condition for Smooth, Unconstrained Problem

Consider the problem **min** $f(x)$ where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable.

## Theorem

*$\hat{x}$ is an optimal solution to the above problem if and only if $\nabla f(\hat{x}) = 0$.*

## Proof.

Only prove the "if" part. Apply Definition 2 for convex functions to point $\hat{x}$. Then for all $y \in \mathbb{R}^n$, $f(y) \geq f(\hat{x}) + \langle \nabla f(\hat{x}), y - \hat{x} \rangle = f(\hat{x})$ for all $y \in \mathbb{R}^n$. So, $\hat{x}$ minimizes $f$. $\square$

## Optimality Condition for Unconstrained Problem

Now consider the problem **min** $f(x)$ where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and not necessarily differentiable.

Theorem

*$\hat{x}$ is an optimal solution to the above problem if and only if $0 \in \partial f(\hat{x})$.*

Proof.

Only prove the "if" part. Apply definition of subgradients to function $f$ at point $\hat{x}$. Then for all $y \in \mathbb{R}^n$, $f(y) \geq f(\hat{x}) + \langle 0, y - \hat{x} \rangle = f(\hat{x})$ for all $y \in \mathbb{R}^n$. So, $\hat{x}$ minimizes $f$. $\square$

# Examples

- Sum of Squares: Given $a_1, a_2, ..., a_n \in \mathbb{R}$, find $\hat{x}$ that minimizes $\frac{1}{n}\sum_{i=1}^{n}(a_i - x)^2$ for $x$. We can check that the objective, the sum of convex functions, is convex. Taking derivative w.r.t $x$ and setting it to 0 gives $-\frac{2}{n}\sum_{i=1}^{n}(a_i - \hat{x}) = 0$ or $\hat{x} = \frac{1}{n}\sum_{i=1}^{n} a_i$.
- Sum of absolute values: Given $a_1, a_2, ..., a_n \in \mathbb{R}$, find $x \in \mathbb{R}$ that minimizes $\frac{1}{n}\sum_{i=1}^{n} |a_i - x|$. $\hat{x}$ is the optimal solution if $0 \in \frac{1}{n}\sum_{i=1}^{n} \partial(|a_i - \hat{x}|)$. Recall that the subdifferential of $|x|$ is given by $\partial|x| = \begin{cases} -1, & \text{if } x < 0 \\ 1, & \text{if } x > 0 \\ [-1, 1], & \text{if } x = 0 \end{cases}$
  Now, assume WLOG that $a_1 \leq a_2 \leq ... \leq a_n$. If $n = 2k$ is even, any $\hat{x} \in [a_k, a_{k+1}]$ is optimal. If $n = 2k + 1$ is odd, $\hat{x} = a_{k+1}$ is optimal.

# Optimality Condition for Constrained Problem

We now consider the constrained optimization problem **min** $f(x)$ **subject to** $x \in \mathfrak{X}$, where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and $\mathfrak{X}$ is convex.

**Theorem**

*$\hat{x}$ is an optimal solution to the above problem if and only if* $0 \in \partial f(\hat{x}) + N_{\mathfrak{X}}(\hat{x})$.

**Proof.**

Only prove the "if" part. Then for $g \in \partial f(\hat{x})$, we have $-g \in N_{\mathfrak{X}}(\hat{x})$. Also, for any $y \in \mathfrak{X}$, we have $y - \hat{x} \in T_{\mathfrak{X}}(\hat{x})$. Combining two results, we have $\langle -g, y - \hat{x} \rangle \leq 0$ for all $y \in \mathfrak{X}$. Now from the definition of subgradient, we have for all $y \in \mathfrak{X}$, $f(y) \geq f(\hat{x}) + \langle g, y - \hat{x} \rangle \geq f(\hat{x})$. Therefore, $\hat{x}$ is the optimal solution. $\square$

# Projection

## Definition

Given a convex set $\mathfrak{X}$ and point $y \notin \mathfrak{X}$, the **projection** of $y$ onto $\mathfrak{X}$ is a point $\hat{x} = Pr_{\mathfrak{X}}(y)$ that solves the problem $\min ||x - y||_2^2$ subject to $x \in \mathfrak{X}$.

From the previous theorem, $\hat{x}$ is optimal iff $0 \in (\hat{x} - y) + N_{\mathfrak{X}}(\hat{x})$ or $y - \hat{x} \in N_{\mathfrak{X}}(\hat{x})$. Since for any $x \in \mathfrak{X}$, $x - \hat{x} \in T_{\mathfrak{X}}(\hat{x})$, we have $\langle y - \hat{x}, x - \hat{x} \rangle \leq 0$ if $\hat{x}$ is optimal.

**Theorem**

*Given a convex set $\mathfrak{X}$ and two points $y_1, y_2$ outside $\mathfrak{X}$, let $x_1, x_2$ be projections of $y_1, y_2$ onto $\mathfrak{X}$. Then $||y_1 - y_2||_2 \geq ||x_1 - x_2||_2$.*

**Proof.**

Previous discussion gives two inequalities

$$\langle y_1 - x_1, x_2 - x_1 \rangle \leq 0 \text{ and } \langle y_2 - x_2, x_1 - x_2 \rangle \leq 0$$

Summing them gives

$$\langle (y_1 - y_2) - (x_1 - x_2), x_2 - x_1 \rangle \leq 0 \text{ or } ||x_1 - x_2||_2^2 \leq \langle y_1 - y_2, x_1 - x_2 \rangle$$

Also, by the Cauchy-Schwarz Inequality, $\langle y_1 - y_2, x_1 - x_2 \rangle \leq ||y_1 - y_2||_2 ||x_1 - x_2||_2$. Combining the two inequalities give the result. $\square$

# Acknowledgement

This work is based on Bubeck, 2014; Deisenroth, Faisal, and Ong, 2020; Beck, 2017; Bubeck, 2014

# References I


- Beck, Amir (2017). *First-Order Methods in Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611974997. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611974997. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611974997.
- Bubeck, Sébastien (2014). "Convex Optimization: Algorithms and Complexity". In: DOI: 10.48550/ARXIV.1405.4980. URL: https://arxiv.org/abs/1405.4980.
- Deisenroth, Marc Peter, A Aldo Faisal, and Cheng Soon Ong (2020). *Mathematics for machine learning*. Cambridge University Press.