写一些不会的东西。

Notational Convention

- $[n] = \{1, 2, \ldots, n\}$
- $\mathbf{x}, \mathbf{y}, \mathbf{v}$: vectors
- A, B: matrices
- $\mathcal{X}, \mathcal{Y}, \mathcal{K}$: domains
- d, m, n: dimensions
- I: identity matrix
- X, Y: random variables
- $\mathbf{p}, \mathbf{q}$: probability distributions

Calculus

Hessian

$$\nabla^2 f(\mathbf{x}) = \left[\frac{\partial^2 f}{\partial x_i, x_j}(\mathbf{x})\right]_{1 \le i, j \le d}$$

Reference: The Matrix Cookbook

> link <

Linear Algebra

Positive (Semi-)Definite Matrix

Positive Definite matrix => PD, $\forall \mathbf{x} \in \mathbb{R}^d, \mathbf{x}^T A \mathbf{x} > 0 \Leftrightarrow A \succ 0$

Positive Semi-Definite matrix => PSD, $\forall \mathbf{x} \in \mathbb{R}^d, \mathbf{x}^T A \mathbf{x} \ge 0 \Leftrightarrow A \succeq 0$

Inner Product

- Vector Space:

$$\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$
$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^{d} x_i y_i$$

- Matrix Space:

$$A, B \in \mathbb{R}^{m \times n}$$
$$\langle A, B \rangle = \mathrm{Tr}(A^T B) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}$$

Norm

- Quadratic Norm:

$$\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^T A \mathbf{x}}, A \text{ is positive semi-definite}$$

Dual Norm

$$\|\mathbf{y}\|_* = \sup\{\mathbf{y}^T \mathbf{x} \mid \|\mathbf{x}\| \leq 1\}$$

The dual norm of l_p-norm is the l_q-norm with $\frac{1}{p} + \frac{1}{q} = 1$

Hölder's Inequality: $\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|_*$

Norm Relationship

Lemma (Mathematical Equivalence of Norms). Suppose that $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbb{R}^d$, there exist positive "constants"(**depend on dimension**) α and β, such that

$$\alpha \|\mathbf{x}\|_a \leq \|\mathbf{x}\|_b \leq \beta \|\mathbf{x}\|_a$$

Cauchy-Schwarz Inequality

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|_*$$
$$\left(\sum_{i=1}^{n} a_i b_i \right)^2 \leq \left(\sum_{i=1}^{n} a_i^2 \right) \cdot \left(\sum_{i=1}^{n} b_i^2 \right)$$

$$\left(\int_a^b f(x)g(x)\mathrm{d}x\right)^2 \le \left(\int_a^b f^2(x)\mathrm{d}x\right)\cdot\left(\int_a^b g^2(x)\mathrm{d}x\right)$$

Matrix Operator Norm

Definition (Matrix Operator Norm). The operator norm (or called induced norm) of a matrix $A \in \mathbb{R}^{m\times n}$ is defined by

$$\|A\|_{\mathrm{op},p} \triangleq \max\left\{\frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p} \mid \mathbf{x}\in\mathbb{R}^d, \mathbf{x}\neq\mathbf{0}\right\}$$

- l_1 norm:

$$\|A\|_{\mathrm{op},1} = \max_{j\in[n]}\sum_{i=1}^m |A_{ij}|$$

- l_∞ norm:

$$\|A\|_{\mathrm{op},\infty} = \max_{i\in[m]}\sum_{j=1}^n |A_{ij}|$$

- l_2 norm (Spectral Norm):

$$\|A\|_{\mathrm{op},2} = \max_{i\in[r]}|\sigma_i|$$

Where $A=\sum_{i=1}^r \sigma_i\mathbf{u}_i\mathbf{v}_i^T$, namely, σ_i is the i-th singular value.

Matrix Entrywise Norm

Definition (Matrix Entrywise Norm). The entrywise norm of a matrix $A\in\mathbb{R}^{m\times n}$ is defined by

$$\|A\|_{\mathrm{en},p} \triangleq \left(\sum_{i=1}^m\sum_{j=1}^n |A_{i,j}|^p\right)^{1/p}$$

$$\|A\|_{\mathrm{F}} = \|A\|_{\mathrm{en},2}$$

Eigen Value Decomposition

Let A be an $d\times d$ PSD matrix, then it can be factored as

$$A = Q\Lambda Q^T$$

where

- $Q = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d) \in \mathbb{R}^{d\times d}$ is orthogonal, and $\mathbf{v}_1, \ldots, \mathbf{v}_d$ are eigenvectors
- $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$, and $\lambda_1, \ldots, \lambda_d$ are eigenvalues

Some properties:

- $A = \sum_{i=1}^d \lambda_i \mathbf{v}_i \mathbf{v}_i^T$
- $\det(A) = \prod_{i=1}^d \lambda_i$
- $\operatorname{Tr}(A) = \sum_{i=1}^d \lambda_i$
- $\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^d \lambda_i^2}$

Singular Value Decomposition

Suppose $A \in \mathbb{R}^{m\times n}$ has a rank r, then it can be factored as

$$A = U\Sigma V^T$$

where

- $U = (\mathbf{u}_1, \ldots, \mathbf{u}_r) \in \mathbb{R}^{m\times r}$ satisfies $U^T U = I$; $V = (\mathbf{v}_1, \ldots, \mathbf{v}_r) \in \mathbb{R}^{n\times r}$ satisfies $V^T V = I$
- $\Sigma = (\sigma_1, \ldots, \sigma_r)$ and $\sigma_1, \ldots, \sigma_r$ are singular valuess.

Some properties:

- $A = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$
- $\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^r \sigma_i^2}$

Schatten Norm

Definition (Matrix Schatten Norm). The Schatten norm of a matrix $A \in \mathbb{R}^{m\times n}$ with rank r is defined by

$$\|A\|_{\mathrm{Sc},p} \triangleq \left(\sum_{i=1}^r \sigma_i^p\right)^{1/p}$$

Probability and Statistics

Cauchy-Schwarz Inequality in Probability

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]$$

Concentration Inequalities

Theorem (Markov's Inequality). Let X be a non-negative random variable with $\mathbb{E}[X] < \infty$, then for all $t > 0$,

$$\Pr[X \geq t\mathbb{E}[X]] \leq \frac{1}{t}$$

Theorem (Chebyshev's Inequality). Let X be a non-negative random variable with $\mathbb{E}[X], \mathrm{Var}[X] < \infty$, then for all $\epsilon > 0$,

$$\Pr[|X - \mathbb{E}[X]| \geq \epsilon] \leq \frac{\mathrm{Var}[X]}{\epsilon^2}$$

Theorem (Hoeffding's Inequality). Let $X_1, \ldots, X_m$ be independent random variables with X_i taking values in $[a_i, b_i]$ for all $i \in [m]$. Then, for any $\epsilon > 0$, the following inequalities hold for $S_m = \sum_{i=1}^m X_i$,

$$\begin{aligned} \Pr[S_m - \mathbb{E}[S_m] \geq \epsilon] &\leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m (b_i - a_i)^2}\right) \\ \Pr[S_m - \mathbb{E}[S_m] \leq -\epsilon] &\leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m (b_i - a_i)^2}\right) \end{aligned}$$

Entropy

Definition (Entropy). The enotropy of a discrete random variable X with probability mass function $\mathbf{p}(x) = \Pr[X = x]$ is denoted by $H(X)$:

$$H(X) = -\sum_{x \in \mathcal{X}} \mathbf{p}(x) \log \mathbf{p}(x)$$

The entropy is a lower bound on lossless data compression.

A explanation of entropy: $\log_2(1/\mathbf{p}(x))$ is the code length needed to encode the information, and $H(X)$ measures the expected code length to encode a distribution $\mathbf{p}$.

Definition (Condition Entropy).

$$
\begin{aligned}
H(Y|X) &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \mathbf{p}(x,y)\log\left[\frac{\mathbf{p}(x,y)}{\mathbf{p}(x)}\right] \\
&= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \mathbf{p}(x,y)\log\mathbf{p}(x,y) + \sum_{x\in\mathcal{X}}\mathbf{p}(x)\log\mathbf{p}(x) \\
&= H(X,Y) - H(X)
\end{aligned}
$$

Definition (Mutual Information).

$$
\begin{aligned}
I(X,Y) &= KL(\mathbf{p}(x,y)\|\mathbf{p}(x)\mathbf{p}(y)) \\
&= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \mathbf{p}(x,y)\log\left[\frac{\mathbf{p}(x,y)}{\mathbf{p}(x)\mathbf{p}(y)}\right] \\
&= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \mathbf{p}(x,y)\log\mathbf{p}(x,y) - \sum_{x\in\mathcal{X}}\mathbf{p}(x)\log\mathbf{p}(x) - \sum_{y\in\mathcal{Y}}\mathbf{p}(y)\log\mathbf{p}(y) \\
&= H(X) + H(Y) - H(X,Y)
\end{aligned}
$$

with the conventions: $0\log 0 = 0, 0\log\frac{0}{0} = 0, \text{ and } a\log\frac{a}{0} = +\infty \text{ for } a > 0$

KL Divergence (Relative Entropy)

Definition (KL Divergence). The KL divergence of two distributions p and q is defined by $KL(\mathbf{p}\|\mathbf{q})$:

$$
KL(\mathbf{p}\|\mathbf{q}) = \sum_{x\in\mathcal{X}}\mathbf{p}(x)\log\left[\frac{\mathbf{p}(x)}{\mathbf{q}(x)}\right]
$$

with the conventions: $0\log 0 = 0, 0\log\frac{0}{0} = 0, \text{ and } a\log\frac{a}{0} = +\infty \text{ for } a > 0$

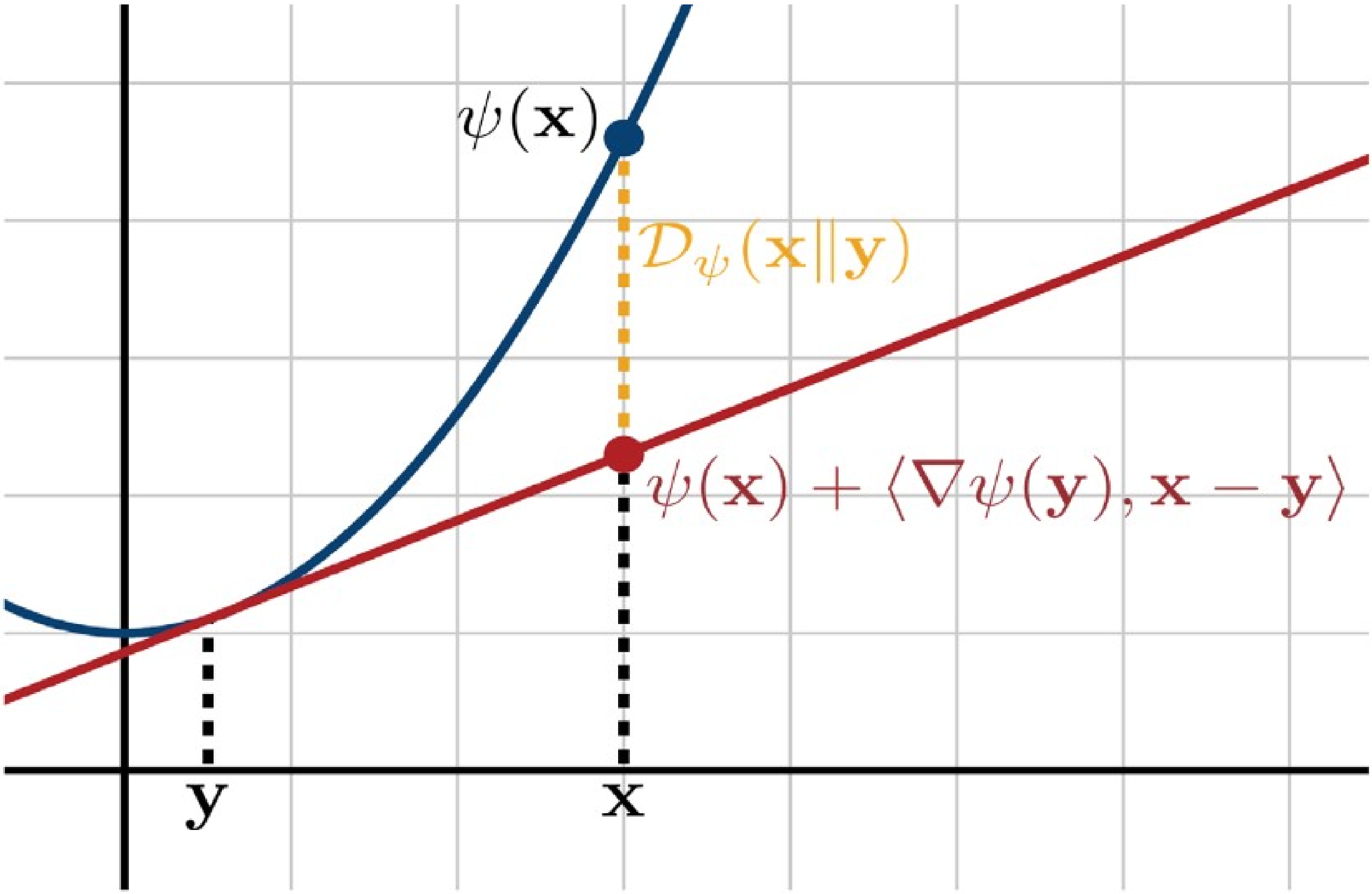
Some properties:

- KL divergence is always non-negative
- Pinsker's Inequality: $KL(\mathbf{p}\|\mathbf{q}) \geq \frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_1^2$
- $KL(\mathbf{p}\|\mathbf{q})$ doesn't always equal to $KL(\mathbf{q}\|\mathbf{p})$

Bregman Divergence

Definition (Bregman Divergence). Let ψ be a convex and differentiable function over a convex set $\mathcal{K}$, and then for any $\mathbf{x},\mathbf{y}\in\mathcal{K}$, the bregman divergence $\mathcal{D}_\psi$ associated to ψ is defined as

$$
\mathcal{D}_\psi(\mathbf{x}\|\mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle\nabla\psi(\mathbf{y}),\mathbf{x}-\mathbf{y}\rangle
$$



Bregman divergence measures the difference of a function and its linear approximation.

KL divergence is a special case when $\mathbf{p}(x)$ is defined as negative entropy: $\mathbf{p}(x) = \sum_i x_i \log x_i$
.

Asymptotic Notations

Definition

- $\Theta(g(n)) = \{f(n) \mid$ there exist positive constants c_1, c_2, and n_0 such that $0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n)$ for all $n \geq n_0\}$.
- $\mathcal{O}(g(n)) = \{f(n) \mid$ there exist positive constants c and n_0 such that $0 \leq f(n) \leq cg(n)$ for all $n \geq n_0\}$.
- $\Omega(g(n)) = \{f(n) \mid$ there exist positive constants c and n_0 such that $0 \leq cg(n) \leq f(n)$ for all $n \geq n_0\}$.
- $o(g(n)) = \{f(n) \mid$ for any positive constant $c > 0$, there exists a constant $n_0 > 0$ such that $0 \leq f(n) < cg(n)$ for all $n \geq n_0\}$.
- $\omega(g(n)) = \{f(n) \mid$ for any positive constant $c > 0$, there exists a constant

$n_0 > 0$ such that $0 \leq cg(n) < f(n)$ for all $n \geq n_0\}$.

Optimization in Machine Learning

Learning by Optimization

The fundamental goal of (supervised) learning: **Risk Minimization**.

$$\min_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{x}, y \in \mathcal{D}}[f(h(\mathbf{x}), y)]$$

where:

- h denotes the hypothesis (model) from the hypothesis space $\mathcal{H}$
- $(\mathbf{x}, y)$ is an instance chosen from a unknown distribution $\mathcal{D}$
- $f(h(\mathbf{x}), y)$ denotes the loss of using hypothesis h on the instance $(\mathbf{x}, y)$

Empirical Risk Minimization

The distribution of the data is unavailable, and the risk can't be computed.

In practice, the learner instead tries to optmize empirical risk.

$$\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} f(h(\mathbf{x}_i), y_i)$$

- IID assumption: **independent** and **identically distributed** random variables.
- ERM approximates RM: All instance are i.i.d. sampled from the same distribution.

Structured ERM

In practice, we often explicitly control the complexity of the learner by adding a **regularization term** in the optimization objective.

$$\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} f(h(\mathbf{x}_i), y_i) + \lambda \mathcal{R}(h)$$

(Constrained) Optimization Problem

$$\min f(\mathbf{x}), \quad \text{s.t. } \mathbf{x} \in \mathcal{X}$$

Unconstrained Optimization

Add a barrier/indicator function.

$$\min h(\mathbf{x}) \triangleq f(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x}), \quad \text{s.t. } \mathbf{x} \in \mathbb{R}^d$$

$$\delta_{\mathcal{X}}(\mathbf{x}) = \begin{cases} 0 & , \quad \mathbf{x} \in \mathcal{X} \\ +\infty & , \quad \text{otherwise} \end{cases}$$

Convex Optimization

Convex Set

已经会了

Projection onto Convex Sets

Definition(Projection). The projection a given point $\mathbf{y}$ onto a convex set $\mathcal{X}$ is defined as the closet point inside the convex set. Formally,

$$\mathbf{x}^* = \Pi_{\mathcal{X}}[\mathbf{y}] \triangleq \arg\min_{x \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$$

Theorem(Pythagoras Theorem). Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex, $\mathbf{y} \in \mathbb{R}^d$. Then for any $\mathbf{z} \in \mathcal{X}$ we have

$$\|\mathbf{y} - \mathbf{z}\| \geq \|\Pi_{\mathcal{X}}[\mathbf{y}] - \mathbf{z}\|$$

Convex Function

Definition(Convex Function). A function $f : \mathcal{X} \to \mathbb{R}$ is called *convex* if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\forall \alpha \in [0,1], f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})$$

Definition(Concave Function). A function $f : \mathcal{X} \to \mathbb{R}$ is called *concave* if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\forall \alpha \in [0,1], f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \geq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})$$

Theorem. A function f is convex iff $\operatorname{dom} f$ is convex and one one of the following properties hold, for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ and $\alpha \in [0,1]$,

1. $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})$
2. $f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq f(\mathbf{y})$
3. $\nabla^2 f(\mathbf{x}) \succeq 0$

Jensen's Inequality

Theorem(Jesen Inequality). If X is a random variable such that $X \in \operatorname{dom} f$ with probability 1, and f is convex, then we have

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Convex Optimization Problem

- minimization language:

$$\begin{aligned} \min f(\mathbf{x}) \\ \text{s.t.} \quad g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m \end{aligned}$$

$\operatorname{dom} f$ should be convex or half-plane.

Subgradient

Definition(Subgradient). Let $f : \mathcal{X} \to \mathbb{R}$ be a proper function and let $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$. A vector $\mathbf{g} \in \mathbb{R}^d$ is called *subgradient* of f at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \forall \mathbf{y} \in \mathbb{R}^d$$

- If $\forall x \in \mathcal{X}$, its subgradients exist, then f is convex.
- f is convex doesn't imply that if $\forall x \in \mathcal{X}$, its subgradients exist. (e.g. $f = -\sqrt{x},\ x \geq 0$. When $x = 0$, the subgradient doesn't exist). $\Rightarrow$ Only consider interial point of feasible domain of f.

Subdifferential

Definition(Subdifferential). The set of all subgradients of f at $\mathbf{x}$ is called *subdifferential* of f at $\mathbf{x}$ and is denoted as by $\partial f(\mathbf{x})$,

$$\partial f(\mathbf{x}) \triangleq \{\mathbf{g} \in \mathbb{R}^d | f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \forall \mathbf{y} \in \mathbb{R}^d\}$$

Optimality Condition

Fermat’s Optimality Condition

- Unconstrained case

Theorem(Fermat's Optimality Condition). Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper convex function. Then

$$\mathbf{x}^* \in \arg\min\{f(\mathbf{x})|\mathbf{x} \in \mathbb{R}^d\}$$

iff $\mathbf{0} \in \partial f(\mathbf{x}^*)$.

First-order Optimality Condition

- Constrained case

Theorem(First-order Optimality Condition). Let f be convex and $\mathcal{X}$ be a closed convex set on which f is differentiable. Then $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ iff there exists $\mathbf{g} \in \partial f(\mathbf{x})$ such that

$$\langle \mathbf{g}, \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \forall \mathbf{x} \in \mathcal{X}$$

KKT Conditions

Theorem(KKT Conditions). Consider the minimization problem

$$\min f(\mathbf{x}), \quad \text{s.t.} \quad g_i(\mathbf{x}) \leq 0, \quad i \in [m] \tag{1}$$

where $f, g_1, \ldots, g_m$ are real-valued convex functions.

1. Let $\mathbf{x}^*$ be optimal solution of (1), and assume that Slater's condition is satisfied. Then there exist $\lambda_1, \ldots, \lambda_m \geq 0$ for which

$$\mathbf{0} \in \partial f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \partial g_i(\mathbf{x}^*) \tag{2}$$

$$\lambda_i \partial g_i(\mathbf{x}^*) = 0, \quad i \in [m] \tag{3}$$

2. If $\mathbf{x}^*$ satisfies conditions (2) and (3) for some $\lambda_1, \ldots, \lambda_m \geq 0$, then it is an optimal solution of problem (1).

Function Properties

Lipschitz Continuity

已经会了

Lipschitzness and Subgradient

Theorem Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function. Consider the following two claims:

1. Lipschitzness: $|f(\mathbf{x}) - f(\mathbf{y})| \leq G\|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$
2. Bounded subgradient: $\|\mathbf{g}\|_* \leq G, \forall \mathbf{g} \in \partial f(\mathbf{x}), x \in \mathcal{X}$

Then

- 2 $\Rightarrow$ 1
  - Proof:

$$\begin{gathered}
|f(\mathbf{y}) - f(\mathbf{x})| \leq |\langle \mathbf{g}, \mathbf{y} - \mathbf{x}\rangle| \leq \|\mathbf{y} - \mathbf{x}\| \cdot \|\mathbf{g}\|_* \\
\because \|\mathbf{g}\|_* \leq G \\
\therefore |f(\mathbf{x}) - f(\mathbf{y})| \leq G\|\mathbf{x} - \mathbf{y}\|
\end{gathered}$$

- If $\mathcal{X}$ is open, then 1 $\Leftrightarrow$ 2

Smoothness

Definition(Smoothness). A function f is L-smooth with repect to the $\|\cdot\|$ norm if, for any $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\|$$

- Why use dual norm?

$$\begin{gathered}
|\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle| \leq \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \cdot \|\mathbf{x} - \mathbf{y}\| \\
|\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle| \leq L\|\mathbf{x} - \mathbf{y}\|^2
\end{gathered}$$

Lemma(Descent Lemma). Let f be L-smooth function over a given convex set $\mathcal{X}$. Then for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

Theorem(First-order Characterizations of L-smoothness). Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function, differentiable over $\mathcal{X}$. Then the following claims are equivalent:

1. f is L-smoothness
2. $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$
3. $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2L}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_*^2$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$
4. $\langle \nabla f(\mathbf{x}) - f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \geq \frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$
5. $f(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) \geq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y}) - \frac{L}{2}\lambda(1 - \lambda)\|\mathbf{x} - \mathbf{y}\|^2$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}, \lambda \in [0, 1]$

Theorem(Second-order Characterizations of L-smoothness). Let f be a twice continuously differentiable function over $\mathbb{R}^d$. L-smoothness w.r.t. the l_p-norm($p \in [0, +\infty]$) is equivalent to

$$\|\nabla^2 f(\mathbf{x})\|_{\mathrm{op},p} \leq L$$

for any $x \in \mathbb{R}^d$.

Strong Convexity

Definition(Strong Convexity). A function f is σ-strongly convex with respect to norm $\|\cdot\|$ if, for any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}\ f$ and $\lambda \in [0, 1]$,

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \frac{\sigma}{2}\lambda(1-\lambda)\|\mathbf{x} - \mathbf{y}\|^2$$

Theorem(First-order Characterizations of Strong Convexity). Let f be a proper closed and convex function. The followings equal:

1. f is σ-strongly convex.
2. For any $\mathbf{x} \in \mathrm{dom}(\partial f), \mathbf{y} \in \mathrm{dom}(f)$ and $\mathbf{g} \in \partial f(\mathbf{x})$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle + \frac{\sigma}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

3. For any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(\partial f), \mathbf{g}_\mathbf{x} \in \partial f(\mathbf{x}), \mathbf{g}_\mathbf{y} \in \partial f(\mathbf{y})$,

$$\langle \mathbf{g}_\mathbf{x} - \mathbf{g}_\mathbf{y}, \mathbf{x} - \mathbf{y} \rangle \geq \sigma \|\mathbf{x} - \mathbf{y}\|^2$$

4. $f(\cdot) - \frac{\sigma}{2}\|\cdot\|^2$ is convex.

Theorem(Second-order Characterizations of Strong Convexity). Let $\mathcal{X}$ be a Eucildean space. Then f is σ-strongly convex iff for $\mathbf{x}, \mathbf{w} \in \mathcal{X}$,

$$\mathbf{w}^T \nabla^2 f(\mathbf{x}) \mathbf{w} = \|\mathbf{w}\|_{\nabla^2 f(\mathbf{x})} \geq \sigma \|\mathbf{w}\|^2$$

When using l_2-norm, $\nabla^2 f(\mathbf{x}) \succeq \sigma I$.

Theorem Let f be a proper closed and σ-strongly convex function. Then

1. f has a unique minimizer, denoted by $\mathbf{x}^*$.
2. $f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^*\|^2, \forall \mathbf{x} \in \mathrm{dom}\ f$.

Strongly Convex and Smooth

If function f is both σ-strongly convex and L-smooth w.r.t. l_2-norm, then

1. $\sigma I \preceq \nabla^2 f(\mathbf{x}) \preceq LI$

2. f is γ-well-conditioned where $\gamma \triangleq \sigma/L \leq 1$ is called the condition number.

Theorem(Conjugate Correspondence). Consider the conjugate function:

$$f^*(\mathbf{y}) \triangleq \max_{\mathbf{x}\in\mathcal{X}}\{\langle\mathbf{y},\mathbf{x}\rangle - f(\mathbf{x})\}$$

1. If the function f is convex and $1/\sigma$-smooth w.r.t. the norm $\|\cdot\|$, then its conjugate f^* is σ-strongly convex w.r.t. the dual norm $\|\cdot\|_*$.
2. If the function f is convex and σ-strongly convex w.r.t. the norm $\|\cdot\|$, then its conjugate f^* is $1/\sigma$-smooth w.r.t. the dual norm $\|\cdot\|_*$.

Some understanding from Kimi.