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

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

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) $\alpha$ and $\beta$, such that

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

# Cauchy-Schwarz Inequality

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

# 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

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

- $l_2$ 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, $\sigma_i$ 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
- $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, and $\lambda_1, \ldots, \lambda_d$ are eigenvalues

Some property:

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

## 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 property:

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

## 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 $S_m = \sum_{i=1}^m X_i$,

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

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

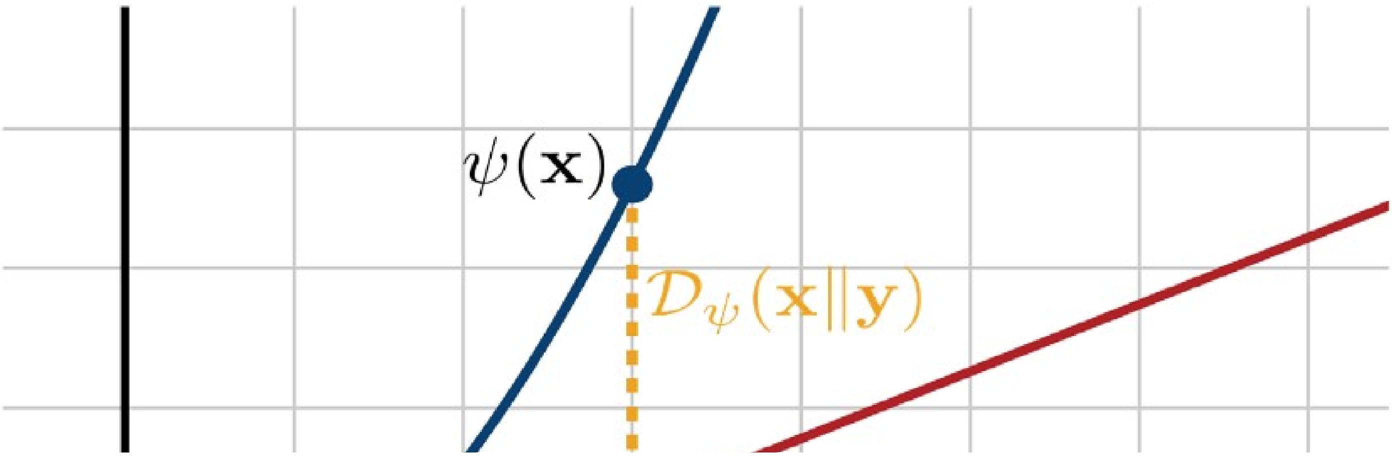
Some property:

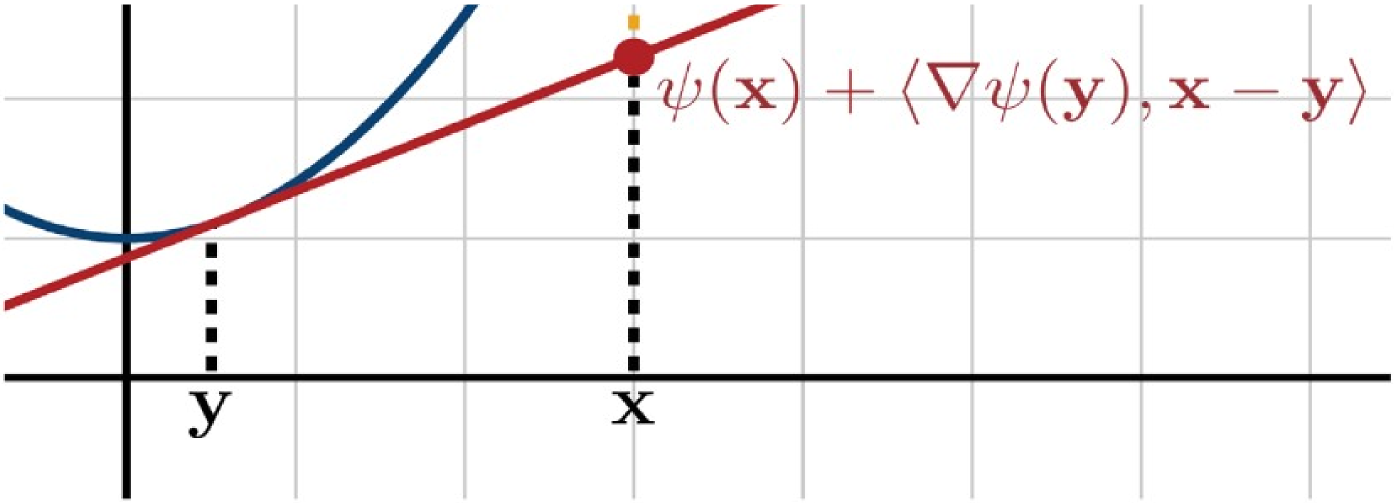
- KL divergence is always non-negative
- Pinsker's Inequality: $KL(\mathbf{p}\|\mathbf{q}) \geq \frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_1^2$
- $KL(\mathbf{p}\|\mathbf{q})$ doesn't always equal to $KL(\mathbf{q}\|\mathbf{p})$

# Bregman Divergence

**Definition** (Bregman Divergence). Let $\psi$ 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 $\psi$ 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.