CS861: Theoretical Foundations of Machine Learning

Chapter 0: A Toolkit for CS861

Kirthevasan Kandasamy

UW-Madison

Contents

These slides give a quick overview of some background topics for the course. We will cover some in class, but you are encouraged to review the rest on your own.

Ch 0.1: Some common inequalities

Basic Logarithmic and Exponential Inequalities.

- For all $x > -1$,
$$\frac{x}{1+x} \;\le\; \log(1+x) \;\le\; x.$$

- For all real x and all positive integers n,
$$e^x \;\ge\; \left(1+\frac{x}{n}\right)^n \;\ge\; 1+x.$$

- For $x \ge -1$ and $r \ge 1$,
$$(1+x)^r \ge 1 + rx.$$
When $r \in [0,1]$, the reverse holds.

Classical Inequalities: Hölder, Cauchy–Schwarz, Minkowski

Let $x, y \in \mathbb{R}^d$. Define $\|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$, for $p \in [1, \infty)$ and $\|x\|_\infty = \max_i |x_i|$.

- Hölder's inequality: for p, q such that $p, q \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$,

$$\|xy\|_1 \ \leq \ \|x\|_p \cdot \|y\|_q.$$

- Cauchy–Schwarz inequality: Hölder when $p = q = 2$:

$$\langle x, y \rangle \ \leq \ \|x\|_2 \, \|y\|_2.$$

- Minkowski's inequality (triangle inequality for ℓ^p-spaces). Let $1 \leq p \leq \infty$. Then,

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

- Let $1 \leq p \leq q \leq \infty$. Then, (Proof in HW0)

$$\|x\|_q \leq \|x\|_p, \qquad \text{and} \qquad \frac{1}{d^{1/p}}\|x\|_p \leq \frac{1}{d^{1/q}}\|x\|_q.$$

Jensen's and Harris' Inequalities

Jensen's inequality. Let f be a convex function and X be an integrable random variable. Then $f\big(\mathbb{E}[X]\big) \le \mathbb{E}\big[f(X)\big]$. Similarly, if f is a concave function, $f\big(\mathbb{E}[X]\big) \ge \mathbb{E}\big[f(X)\big]$.

Harris' covariance inequality. Let X be a random variable, and let f and g be real-valued functions that are both non-decreasing. Then,

$$\mathbb{E}\big[f(X)\,g(X)\big] \ge \mathbb{E}\big[f(X)\big]\,\mathbb{E}\big[g(X)\big].$$

Ch 0.2: Bayesian Inference

Frequentist approach. In the frequentist paradigm, a parameter of interest θ is fixed and unknown. Statistical inference targets objective estimation: ask how well data can estimate θ via sampling distributions, confidence intervals, hypothesis tests, etc. Example: Problem 1 in HW0 for normal mean estimation.

Bayesian approach.

- Here, θ is treated as a random variable taking values in Θ.
- A learner's prior knowledge/beliefs about θ, before seeing data, are captured by a *prior distribution* $\pi(\theta)$.
- Once the data $X = x$ are observed, the prior is updated with the observed data via the likelihood $p_{X|\theta}(x \mid \theta)$, yielding a *posterior distribution* $\pi(\theta \mid x)$.

In this class, we will mostly focus on the frequentist paradigm but will use Bayesian inference for proving lower bounds (Chapter 1).

Bayesian Setup: General Formulation

Given data $X = x$, we have:

Prior: $\pi(\theta)$, `Learner's prior beliefs about` θ.

Likelihood: $p_{X|\theta}(x \mid \theta)$,

`Probability of observing data` x `given that the parameter is` θ.

$$\text{Posterior: } \pi(\theta \mid x) = \frac{p_{X|\theta}(x \mid \theta)\,\pi(\theta)}{\int_{\Theta} p_{X|\theta}(x \mid \theta)\,\pi(\theta)\,d\theta} \;\propto\; p_{X|\theta}(x \mid \theta)\,\pi(\theta).$$

`Learner's beliefs about` θ `after observing data.`

From the posterior, one can derive point estimates (e.g., posterior mean, mode) or credible intervals.

Example 1: Beta prior, Binomial likelihood

Prior $\pi(\theta)$.

$$\theta \in (0,1), \qquad \pi(\theta) = \mathrm{Beta}(\alpha,\beta): \ \pi(\theta) = \frac{1}{B(\alpha,\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Likelihood $p_{x|\theta}(x \mid \theta)$.

$$X \mid \theta \sim \mathrm{Binomial}(n,\theta), \quad x \in \{0,\ldots,n\}, \qquad p_{x|\theta}(x \mid \theta) = \binom{n}{x}\,\theta^{x}(1-\theta)^{n-x}.$$

Posterior $\pi(\theta \mid x)$.

$$\pi(\theta \mid x) \ \propto \ p_{x|\theta}(x \mid \theta)\,\pi(\theta) \ \propto \ \left[\theta^{x}(1-\theta)^{n-x}\right] \left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right] = \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}.$$

Thus, $\pi(\theta \mid x) = \mathrm{Beta}\big(\alpha + x,\ \beta + n - x\big)$.

Example 2: Normal prior and likelihood

Prior $\pi(\mu)$**.**

$$\mu \sim \mathcal{N}(\mu_0, \tau^2), \quad \pi(\mu) \propto \exp\left(-\frac{(\mu - \mu_0)^2}{2\tau^2}\right).$$

Likelihood $p_{\mathbf{x}|\mu}(\mathbf{x} \mid \mu)$**.**

$$X_i \mid \mu \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad \text{so} \quad p(\mathbf{x} \mid \mu) \propto \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right).$$

Posterior $\pi(\mu \mid \mathbf{x})$ is a normal distribution (Proof in HW0)

$$\pi(\mu \mid \mathbf{x}) = \mathcal{N}\left(\frac{(\tau^{-2})\mu_0 + (n/\sigma^2)\bar{x}}{\tau^{-2} + n/\sigma^2}, \ \frac{1}{\tau^{-2} + n/\sigma^2}\right).$$

Bayes' risk and Bayes' estimator

Definition. Given loss $\ell(\theta, a)$ for estimating θ with a, the risk (expected loss) of an estimator $\widehat{\theta}$ is

$$R(\pi, \widehat{\theta}) = \mathbb{E}\big[\ell(\theta, \widehat{\theta}(X))\big] = \mathbb{E}_\theta\left[\mathbb{E}_X\left[\ell(\theta, \widehat{\theta}(X)) \mid \theta\right]\right].$$

An estimator $\widehat{\theta}$ which minimizes $R(\pi, \widehat{\theta})$, if it exists, is called the *Bayes' estimator*.

The minimum value $R(\pi, \widehat{\theta})$ is called the *Bayes' risk*.

Computing the Bayes' estimator. Let us write,

$$R(\pi, \widehat{\theta}) = \mathbb{E}_X\Big[\underbrace{\mathbb{E}_\theta\left[\ell(\theta, \widehat{\theta}(X)) \mid X\right]}_{(*)}\Big].$$

If we can find an estimator $\widehat{\theta}$ which minimizes $(*)$ for all X, it is the Bayes' estimator.

Bayes' Estimator (cont'd)

Squared loss. When $\ell(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$ is the squared loss, the Bayes' estimator is the posterior mean $\widehat{\theta}_\Lambda(X) = \mathbb{E}[\theta|X]$.

Proof. Define $\widehat{\theta}(X) = \mathbb{E}[\theta|X]$. Consider any other estimator $\widehat{\theta}'$.

$$
\begin{aligned}
\mathbb{E}_P[(\widehat{\theta}'(X) - \theta(P))^2|X] &= \mathbb{E}_P[(\widehat{\theta}' - \theta)^2|X] \quad \texttt{here, } \widehat{\theta}'(X) \to \widehat{\theta}'\texttt{, } \theta(P) \to \theta\texttt{.} \\
&= \mathbb{E}_P[(\widehat{\theta}' - \widehat{\theta})^2 + (\widehat{\theta} - \theta)^2 + 2(\widehat{\theta}' - \widehat{\theta})(\widehat{\theta} - \theta)|X] \\
&= \mathbb{E}_P[(\widehat{\theta}' - \widehat{\theta})^2|X] + \mathbb{E}_P[(\widehat{\theta} - \theta)^2|X] + 2(\widehat{\theta}' - \widehat{\theta})\mathbb{E}[(\widehat{\theta} - \theta)|X] \\
&\qquad \widehat{\theta}\texttt{, } \widehat{\theta}' \texttt{ are functions of } X\texttt{.} \\
&= \mathbb{E}_P[(\widehat{\theta}' - \widehat{\theta})^2|X] + \mathbb{E}_P[(\widehat{\theta} - \theta)^2|X] \quad \texttt{as } \widehat{\theta} = \mathbb{E}_P[\theta|X]\texttt{.} \\
&\geq \mathbb{E}_P[(\widehat{\theta} - \theta)^2|X].
\end{aligned}
$$

Absolute loss. When $\ell(\theta_1, \theta_2) = |\theta_1 - \theta_2|$ is the absolute loss, the Bayes' estimator is the posterior median. (Proof: try at home)

Ch 0.3: Sub-Gaussian concentration

We will start with some common probability inequalities

Union bound. For any events $A_1, A_2, \ldots, A_n$,

$$\Pr\Big(\bigcup_{i=1}^{n} A_i\Big) \;\leq\; \sum_{i=1}^{n} \Pr(A_i).$$

Markov's inequality. Let X be a non-negative random variable and let $a > 0$. Then

$$\Pr(X \geq a) \;\leq\; \frac{\mathbb{E}[X]}{a}.$$

Chebyshev's inequality. Let X be any random variable with finite mean $\mu = \mathbb{E}[X]$ and finite variance $\sigma^2 = \mathrm{Var}(X)$. For any $b > 0$,

$$\Pr\big(|X - \mu| \geq b\big) \;\leq\; \frac{\sigma^2}{b^2}.$$

Sub-Gaussian Random Variables (cont'd)

Sub-Gaussian random variables. If a random variable X satisfies

$$\mathbb{E}\left[e^{t(X-\mathbb{E}[X])}\right] \leq e^{\frac{\sigma^2 t^2}{2}},$$

for all $t \in \mathbb{R}$, then it is said to be σ-sub-Gaussian.

Intuitively, the tail decays at least as fast as a $\mathcal{N}(0, \sigma^2)$ RV.

Sub-Gaussian Random Variables (cont'd)

Sub-Gaussian concentration. If X is σ-sub-Gaussian, then

$$\mathbb{P}(X - \mathbb{E}[X] > \epsilon) \leq e^{\frac{-\epsilon^2}{2\sigma^2}}, \qquad \mathbb{P}(X - \mathbb{E}[X] < -\epsilon) \leq e^{\frac{-\epsilon^2}{2\sigma^2}}.$$

Proof. Assume $\mathbb{E}[X] = 0$, for simplicity. Then, for all t,

$$\begin{aligned} \mathbb{P}(X > \epsilon) &= \mathbb{P}\left(e^{tX} > e^{t\epsilon}\right) \\ &\leq \frac{\mathbb{E}[e^{tX}]}{e^{t\epsilon}} && \texttt{by Markov inequality, } \mathbb{P}(Z > a) \leq \frac{\mathbb{E}[Z]}{a} \texttt{ for } Z > 0 \\ &\leq e^{\frac{1}{2}t^2\sigma^2 - t\epsilon} && \texttt{sub-Gaussianity definition, true for all } t. \end{aligned}$$

By choosing $t = \epsilon/\sigma^2$ to minimize the exponent, we have,

$$\mathbb{P}(X > \epsilon) \leq e^{\frac{-\epsilon^2}{2\sigma^2}}.$$

Sub-Gaussian Random Variables (cont'd)

Some examples:

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then X is σ-sub-Gaussian.
- If $\textbf{supp}\,(X) \subset [a, b]$, then X is $\frac{1}{2}(b-a)$-sub-Gaussian. Sub-Gaussian concentration is Hoeffding's inequality.

Some properties of sub-Gaussian RVs: (Proofs: try at home)

1) Let $a \in \mathbb{R}$. If X is σ-sub-Gaussian, then aX is $|a|\sigma$-sub-Gaussian.

2) If $X_1, \ldots, X_n$ are independent sub-Gaussian RVs with constants $\sigma_1, \ldots, \sigma_n$, then $\sum_{i=1}^n X_i$ is $\sqrt{\sigma_1^2 + \cdots + \sigma_n^2}$ sub-Gaussian.

3) If $X_1, \ldots, X_n$ are independent σ-sub-Gaussian RVs, then

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_i (X_i - \mathbb{E}[X_i])\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-n\epsilon^2}{2\sigma^2}\right).$$

Sub-Gaussian Random Variables (cont'd)

The maximal inequality. Let $Z_1, \ldots, Z_n$ be zero mean σ-sub-Gaussian random variables (not necessarily independent). Then,

$$\mathbb{E}\left[\max_{i \in [n]} Z_i\right] \leq \sigma\sqrt{2\log(n)}.$$

Proof. The following holds for all t:

$$\begin{aligned}
\exp\left(t\mathbb{E}\left[\max_{i \in [n]} Z_i\right]\right) &\leq \mathbb{E}\left[\exp\left(t \max_{i \in [n]} Z_i\right)\right] && \texttt{By Jensen's and concavity of exp} \\
&= \mathbb{E}\left[\max_{i \in [n]} e^{tZ_i}\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{n} e^{tZ_i}\right] \\
&\leq n e^{\frac{t^2\sigma^2}{2}} && \texttt{Sub-Gaussianity}
\end{aligned}$$

Proof of maximal inequality (cont'd)

We just showed,

$$\exp\left(t\mathbb{E}\left[\max_{i\in[n]} Z_i\right]\right) \leq ne^{\frac{t^2\sigma^2}{2}}$$

Taking log on both sides, we have, for all t,

$$\mathbb{E}\left[\max_{i\in[n]} Z_i\right] \leq \frac{\log(n)}{t} + \frac{t\sigma^2}{2}.$$

Choosing $t = \frac{1}{\sigma}\sqrt{2\log(n)}$ to minimize the RHS, gives the bound.

McDiarmid's inequality

Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $X_i \in \mathcal{X}_i$ for all i. Let $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ satisfy the bounded differences condition: for each i, there exists c_i such that for all $x_1, x_2, \ldots, x_i, x_i', \ldots, x_n$, we have,

$$\left| f(x_1, \ldots, x_i, \ldots, x_n) \; - \; f(x_1, \ldots, x_i', \ldots, x_n) \right| \; \leq \; c_i .$$

Then for any $t > 0$,

$$\Pr\left(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \geq t\right) \; \leq \; \exp\left(-\frac{2\,t^2}{\sum_{i=1}^n c_i^2}\right),$$

$$\Pr\left(\, f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \leq -t\right) \; \leq \; \exp\left(-\frac{2\,t^2}{\sum_{i=1}^n c_i^2}\right).$$

Ch 0.4: Covering and Packing Numbers

Metric. Let $\mathcal{X}$ be a set and let $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a function. We say ρ is a metric if it satisfies the following four properties for all $x, y, z \in \mathcal{X}$:
(i) $\rho(x, x) = 0$.
(ii) $\rho(x, y) = 0 \implies x = y$.
(iii) $\rho(x, y) = \rho(y, x)$.
(iv) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ (triangle inequality).

If ρ satisfies (i), (iii), and (iv) it is called a **pseudo-metric**.

We say that $(\mathcal{X}, \rho)$ is a **(pseudo-)metric space**.

Norm balls. Let $\mathcal{B}(x, \epsilon, \rho) = \{x'; \rho(x, x') \leq \epsilon\}$. .

Some special cases

Let $1 \leq p \leq \infty$.

Euclidean spaces. When $\mathcal{X} = \mathbb{R}^d$, we can define the ℓ_p norms as follows:

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p}, \qquad \|x\|_\infty = \max_{i\in[d]} |x_i|.$$

$\rho(x, y) = \|x - y\|_p$ is a metric.

Function spaces. When $\mathcal{X} = \mathbb{R}^{\mathcal{S}} = \{f : f : \mathcal{S} \to \mathbb{R}\}$ for some set $\mathcal{S}$, we define the L_p norm for $f \in \mathcal{X}$ as follows:

$$\|f\|_p = \left(\int_{\mathcal{S}} |f(s)|^p \mathrm{d}s\right)^{1/p} \qquad \|f\|_\infty = \max_{s\in\mathcal{S}} |f(s)|.$$

$\rho(f, y) = \|f - g\|_p$ is a metric.

Covering numbers

Let $(\mathcal{X}, \rho)$ be a pseudo-metric space and let $A \subset \mathcal{X}$. Let $\epsilon > 0$. A set $C \subset A$ is called an ϵ-**cover** of A if, for all $x \in A$, there exists $c \in C$ such that $\rho(x, c) \leq \epsilon$.

The ϵ-**covering number** is the size of the smallest ϵ-cover of A,

$$N(\epsilon, A, \rho) = \min\left\{|C|;\ C \text{ is an } \epsilon\text{-cover of A. }\right\}$$

The function $\epsilon \mapsto \log(N(\epsilon, A, \rho))$ is called the **metric entropy** of A.

Packing numbers

Let $(\mathcal{X}, \rho)$ be a pseudo-metric space and let $A \subset \mathcal{X}$. Let $\epsilon > 0$. A set $P \subset A$ is called an ϵ-**packing** of A if, $\rho(x, x') > \epsilon$ (note strict inequality) for all $x, x' \in P$ such that $x \neq x'$.

The ϵ-**packing number** is the size of the largest ϵ-packing of A,

$$M(\epsilon, A, \rho) = \max\{|P|;\ P \text{ is an } \epsilon\text{-packing of A. }\}$$

Some useful results

Property. Let $A \subset A'$. Then, for any $\epsilon > 0$, we have

$$N(\epsilon, A, \rho) \le N(\epsilon, A', \rho), \qquad M(\epsilon, A, \rho) \le M(\epsilon, A', \rho).$$

Proof. Any cover of A' also covers A and any packing in A is also a packing in A'.

Theorem 1 (Packing-covering sandwich). For any $\epsilon > 0$ and $A \subset \mathcal{X}$, we have

$$M(2\epsilon, A, \rho) \le N(\epsilon, A, \rho) \le M(\epsilon, A, \rho).$$

Proof. In HW0.

Some useful results (cont'd)

For any $A, A' \subset \mathbb{R}^d$ and $\alpha \in \mathbb{R}$, denote $A + \alpha A' = \{a + \alpha a'; a \in A, a' \in A'\}$.

Theorem 2 (Bounds in Euclidean spaces). Let $\mathcal{X} = \mathbb{R}^d$ and let $\|\cdot\|$ be any norm. Let $B = \{x \in \mathbb{R}^d; \|x\| \leq 1\}$ be the unit ball. Then,

$$\left(\frac{1}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)} \leq N(\epsilon, A, \|\cdot\|) \leq M(\epsilon, A, \|\cdot\|) \leq \frac{\mathrm{vol}\left(A + \frac{\epsilon}{2}B\right)}{\mathrm{vol}\left(\frac{\epsilon}{2}B\right)}. \tag{1}$$

Moreover if A is a convex set and contains ϵB, then,

$$M(\epsilon, A, \|\cdot\|) \leq \left(\frac{3}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)}. \tag{2}$$

Proof of Theorem 2

Denote $\mathcal{B}(x, \epsilon) = \mathcal{B}(x, \epsilon, \|\cdot\|)$.

Middle inequality in (1): follows from the packing-covering sandwich.

Left inequality: Let C be any covering of A. We have $A \subset \bigcup_{c \in C} \mathcal{B}(c, \epsilon)$. Therefore

$$\mathrm{vol}(A) \leq \sum_{c \in C} \mathrm{vol}(\mathcal{B}(c, \epsilon)) = |C| \mathrm{vol}(\epsilon B) = |C| \epsilon^d \mathrm{vol}(B).$$

Here, we have used the fact that $\mathrm{vol}(\alpha A') = \alpha^d \mathrm{vol}(A')$ for any $A' \subset \mathbb{R}^d$.

Therefore, $|C| \geq \frac{1}{\epsilon^d} \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)}$. Taking the minimum over all coverings C yields the result.

Proof of Theorem 2 (cont'd)

Right inequality in (1): Let P be any packing of A.
We have $A + \frac{\epsilon}{2}B \supset \bigcup_{p \in P} \mathcal{B}(p, \epsilon/2)$. Therefore,

$$\mathrm{vol}\left(A + \frac{\epsilon}{2}B\right) \geq \sum_{p \in P} \mathrm{vol}\left(\mathcal{B}(p, \epsilon/2)\right) = |P| \mathrm{vol}\left(\frac{\epsilon}{2}B\right).$$

Therefore, $|P| \leq \frac{\mathrm{vol}\left(A + \frac{\epsilon}{2}B\right)}{\mathrm{vol}\left(\frac{\epsilon}{2}B\right)}$. Now take maximum over all packings.

Result (2): We will show: *(i)* when $\epsilon B \subset A$, then $A + \frac{\epsilon}{2}B \subset A + \frac{1}{2}A$, *(ii)* when A is convex, $A + \frac{1}{2}A \subset \frac{3}{2}A$. Therefore, $A + \frac{\epsilon}{2}B \subset \frac{3}{2}A$. Hence,

$$M(\epsilon, A, \|\cdot\|) \underbrace{\leq}_{\text{from (1)}} \frac{\mathrm{vol}\left(A + \frac{\epsilon}{2}B\right)}{\mathrm{vol}\left(\frac{\epsilon}{2}B\right)} \leq \frac{\mathrm{vol}\left(\frac{3}{2}A\right)}{\mathrm{vol}\left(\frac{\epsilon}{2}B\right)} = \frac{(3/2)^d \mathrm{vol}\left(A\right)}{(\epsilon/2)^d \mathrm{vol}\left(B\right)} = \left(\frac{3}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)}.$$

Proof of Theorem 2 (cont'd)

Read at home.

Proof of (i) when $\epsilon B \subset A$, then $A + \frac{\epsilon}{2}B \subset A + \frac{1}{2}A$.

Suppose $x \in A + \frac{\epsilon}{2}B$. Then, we can write $x = a + \frac{\epsilon}{2}b$ where $a \in A$ and $b \in B$. Hence, $\frac{\epsilon}{2}b \in \frac{1}{2}\epsilon B \subset \frac{1}{2}A$. Therefore, $x = a + \frac{\epsilon}{2}b \in A + \frac{1}{2}A$.

Proof of (ii) when A is convex, $A + \frac{1}{2}A \subset \frac{3}{2}A$.

Suppose $x \in A + \frac{1}{2}A$. Then, we can write $x = a + \frac{1}{2}a'$ where $a, a' \in A$.
As A is convex, $A \ni \frac{2}{3}a + \frac{1}{3}a' = \frac{2}{3}x$. Therefore, $x \in \frac{3}{2}A$.

N.B. You can also check that $A + \frac{1}{2}A \supset \frac{3}{2}A$ for all A so $A + \frac{1}{2}A = \frac{3}{2}A$ when A is convex.

Some useful results (cont'd)

Recall: Theorem 2. Let $\mathcal{X} = \mathbb{R}^d$ and let $\|\cdot\|$ be any norm. Let $B = \{x \in \mathbb{R}^d; \|x\| \leq 1\}$ be the unit ball. Suppose A is a convex set and contains ϵB. Then,

$$\left(\frac{1}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)} \leq N(\epsilon, A, \|\cdot\|) \leq M(\epsilon, A, \|\cdot\|) \leq \frac{\mathrm{vol}\left(A + \frac{\epsilon}{2}B\right)}{\mathrm{vol}\left(\frac{\epsilon}{2}B\right)} \leq \left(\frac{3}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)}.$$

A useful corollary of Theorem 2. Let $\mathcal{X} = \mathbb{R}^d$ and let $\epsilon \leq 1$. For any norm $\|\cdot\|$ and corresponding unit ball $B = \{x \in \mathbb{R}^d; \|x\| \leq 1\}$, we have

$$\left(\frac{1}{\epsilon}\right)^d \leq N(\epsilon, B, \|\cdot\|) \leq M(\epsilon, B, \|\cdot\|) \leq \left(1 + \frac{2}{\epsilon}\right)^d \leq \left(\frac{3}{\epsilon}\right)^d.$$

Some useful results (cont'd)

Theorem 3. Let $\mathcal{X} = \mathbb{R}^d$ and $\epsilon \leq 1$. Let $p, q \in [1, \infty]$. Denote $\mathcal{B}_r = \{x; \|x\|_r \leq 1\}$. Then,

$$\left(\frac{1}{\epsilon}\right)^d \frac{\mathrm{vol}(\mathcal{B}_q)}{\mathrm{vol}(\mathcal{B}_p)} \leq N(\epsilon, \mathcal{B}_q, \|\cdot\|p). \tag{3}$$

Moreover, when $q \geq p$, then

$$N(\epsilon, \mathcal{B}_q, \|\cdot\|p) \leq \left(\frac{3d^{1/p-1/q}}{\epsilon}\right)^d. \tag{4}$$

Some notes.

1. If you need a lower bound on the packing number, you can use (3) along with the packing-covering sandwich.
2. Explicit expression for $\mathrm{vol}(\mathcal{B}_q)$ in $\mathbb{R}^d$:

$$\mathrm{vol}(\mathcal{B}_q) = \frac{\big(2\,\Gamma(1 + 1/q)\big)^d}{\Gamma\big(1 + d/q\big)}.$$

Proof of Theorem 3

Recall: Theorem 2. Let $\mathcal{X} = \mathbb{R}^d$ and let $\|\cdot\|$ be any norm. Let $B = \{x \in \mathbb{R}^d; \|x\| \leq 1\}$ be the unit ball. Suppose A is a convex set and contains ϵB. Then,

$$\left(\frac{1}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)} \leq N(\epsilon, A, \|\cdot\|) \leq M(\epsilon, A, \|\cdot\|) \leq \frac{\mathrm{vol}\left(A + \frac{\epsilon}{2} B\right)}{\mathrm{vol}\left(\frac{\epsilon}{2} B\right)} \leq \left(\frac{3}{\epsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)}.$$

Proof. Inequality (3) follows directly from Theorem 2:

$$N(\epsilon, \mathcal{B}_q, \|\cdot\|_p) \geq \left(\frac{1}{\epsilon}\right)^d \frac{\mathrm{vol}(\mathcal{B}_q)}{\mathrm{vol}(\mathcal{B}_p)}.$$

For inequality (4), we will use Theorem 2 again as follows,

$$N\left(\frac{\epsilon}{d^{1/p-1/q}}, \mathcal{B}_p, \|\cdot\|_p\right) \leq \left(\frac{3d^{1/p-1/q}}{\epsilon}\right)^d \frac{\mathrm{vol}(\mathcal{B}_p)}{\mathrm{vol}(\mathcal{B}_p)}.$$

To complete the proof, we will show $N(\epsilon, \mathcal{B}_q, \|\cdot\|p) \leq N\left(\frac{\epsilon}{d^{1/p-1/q}}, \mathcal{B}_p, \|\cdot\|_p\right)$.

Proof of Theorem 3 (cont'd)

Recall the following:

Property. Let $A \subset A'$. Then, $\forall\, \epsilon > 0$, $N(\epsilon, A, \rho) \leq N(\epsilon, A', \rho)$, and $M(\epsilon, A, \rho) \leq M(\epsilon, A', \rho)$.

Inequality. Let $1 \leq p \leq q$. Then, in $\mathbb{R}^d$, we have $\frac{1}{d^{1/p}}\|\cdot\|_p \leq \frac{1}{d^{1/q}}\|\cdot\|_q$.

From the above inequality we have $\mathcal{B}_q \subset d^{1/p-1/q}\mathcal{B}_p$ via the following argument:

$$
\begin{aligned}
d^{1/q}\mathcal{B}_q = \left\{ d^{1/q}x;\ \|x\|_q \leq 1 \right\} &= \left\{ x;\ \frac{1}{d^{1/q}}\|x\|_q \leq 1 \right\} \\
&\subset \left\{ x;\ \frac{1}{d^{1/p}}\|x\|_p \leq 1 \right\} = d^{1/p}\mathcal{B}_p.
\end{aligned}
$$

The claim follows from the above property:

$$
N(\epsilon, \mathcal{B}_q, \|\cdot\|p) \leq N\left(\epsilon, d^{1/p-1/q}\mathcal{B}_p, \|\cdot\|_p\right) = N\left(\frac{\epsilon}{d^{1/p-1/q}}, \mathcal{B}_p, \|\cdot\|_p\right).
$$

Some useful results

Theorem 4. Denote $N_p(\epsilon, A) = N\left(\epsilon, A, \frac{1}{d^{1/p}}\|\cdot\|_p\right)$, and $N_\infty(\epsilon, A) = N(\epsilon, A, \|\cdot\|_\infty)$. Then,

$$N_1(\epsilon, A) \leq N_2(\epsilon, A) \leq \cdots \leq N_\infty(\epsilon, A).$$

Proof. In HW0.

Ch 0.5: Distances/Divergences Between Distributions

Let P, Q be probability distributions with densities or pmfs p, q.

1. Kullback-Leibler divergence:

$$\mathrm{KL}(P, Q) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\, dx$$

2. Total variation distance:

$$\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$$

3. L_1 distance:

$$\|P - Q\|_1 = \int |p(x) - q(x)| dx$$

Distances/Divergences (cont'd)

4. Chi-squared divergence:

$$\chi^2(P,Q) = \int \frac{(p(x)-q(x))^2}{q(x)}\,dx = \mathbb{E}_Q\left[\left(\frac{p(X)}{q(X)} - 1\right)^2\right]$$

5. Hellinger distance:

$$\mathrm{H}^2(P,Q) = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = 2 - 2\int \sqrt{p(x)q(x)}dx$$

6. Affinity (measure of similarity):

$$\|P \wedge Q\| = \int \min(p(x), q(x))dx$$

Relations Between Divergences

1. For n i.i.d samples:

$$\mathrm{KL}(P^n, Q^n) = n\mathrm{KL}(P, Q), \quad \mathrm{H}^2(P^n, Q^n) = 2 - 2\left(1 - \frac{1}{2}\mathrm{H}^2(P, Q)\right)^n.$$

2. $\mathrm{TV}(P, Q) = \frac{1}{2}\|P - Q\|_1 = 1 - \|P \wedge Q\|$.
3. $\mathrm{H}^2(P, Q) \leq \|P - Q\|_1$.
4. Pinsker's inequality: $\mathrm{TV}(P, Q) \leq \sqrt{\frac{1}{2}\mathrm{KL}(P, Q)}$.
5. $\mathrm{KL}(P, Q) \leq \chi^2(P, Q)$.
6. Affinity bound: $\|P \wedge Q\| \geq \frac{1}{2}e^{-\mathrm{KL}(P,Q)}$.

We will prove 6. You will prove 1–5 in HW1.

Proof: Affinity Bound via KL Divergence

Affinity bound: $\|P \wedge Q\| \geq \frac{1}{2} e^{-\mathrm{KL}(P,Q)}$

$$
\begin{aligned}
2\|P \wedge Q\| &= 2 \int \min(p,q) \\
&\geq 2 \int \min(p,q) - \left(\int \min(p,q) \right)^2 \\
&= \int \min(p,q) \times \left(2 - \int \min(p,q) \right) \\
&= \int \min(p,q) \times \int \max(p,q) \qquad \texttt{As } \int \min(p,q) + \int \max(p,q) = \int p + \int q = 2 \\
&\geq \left(\int \sqrt{\min(p,q) \cdot \max(p,q)} \right)^2 \qquad \texttt{CS ineq: } \left(\int uv \right)^2 \leq \int u^2 \int v^2 \\
&= \left(\int \sqrt{p \cdot q} \right)^2 = \exp\left(2 \log \left(\int \sqrt{p \cdot q} \right) \right)
\end{aligned}
$$

Proof: Affinity Bound via KL Divergence (cont'd)

Affinity bound: $\|P \wedge Q\| \geq \frac{1}{2} e^{-\mathrm{KL}(P,Q)}$.

$$2\|P \wedge Q\| \geq \exp\left(2\log\left(\int \sqrt{p \cdot q}\right)\right)$$

$$= \exp\left(2\log\left(\int p\sqrt{\frac{q}{p}}\right)\right)$$

$$\geq \exp\left(2\int p\log\left(\sqrt{\frac{q}{p}}\right)\right)$$

`by Jensen's ineq:` $\log\left(\mathbb{E}_{X\sim P}\left[\sqrt{\frac{q(X)}{p(X)}}\right]\right) \geq \mathbb{E}_{X\sim P}\left[\log\left(\sqrt{\frac{q(X)}{p(X)}}\right)\right]$

$$= \exp\left(-\int p\log\left(\frac{p}{q}\right)\right) = e^{-\mathrm{KL}(P,Q)}.$$

Examples of KL Divergence

1) If $P = \mathcal{N}(\mu_1, \sigma^2)$ and $Q = \mathcal{N}(\mu_2, \sigma^2)$, then $\mathrm{KL}(P, Q) = \frac{1}{2\sigma^2}(\mu_1 - \mu_2)^2$.

2) If $P = \mathrm{Bern}\,(p)$ and $Q = \mathrm{Bern}\,(q)$, for $p, q \in [0, 1]$, then

$$\mathrm{KL}(P, Q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q))$$

Moreover, it satisfies:

$$2(p - q)^2 \underbrace{\leq}_{\text{Pinsker's}} \mathrm{KL}(P, Q) \underbrace{\leq}_{\mathrm{KL} \leq \chi^2} \frac{(p - q)^2}{q(1 - q)}.$$

If q is bounded away from 0 and 1, we have $\mathrm{KL}(P, Q) \in \Theta((p - q)^2)$.

Ch 0.6: Information Theory

Let P be a distribution with density (pdf or pmf) p, and let $\textbf{supp}\,(P) = \mathcal{X}$.

Entropy[1]**.** Let X have distribution P. Then,

$$H(X) = H(P) = \mathbb{E}_{X\sim P}[-\log(p(X))].$$

N.B. It is customary to write this as a function of a random variable $H(X)$, although it is really a function of the distribution $H(P)$.

For discrete and continuous RVs, we have respectively,

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log(p(x)), \qquad H(X) = -\int_{x\in\mathcal{X}} p(x)\log(p(x))udx$$

[1]Not to be confused with metric entropy.

Entropy (cont'd)

Some examples:

- $\mathrm{Bern}\,(p)$: $H(X) = -p\log(p) - (1-p)\log(1-p)$.
- $\mathcal{N}(\mu, \sigma^2)$: $H(X) = -\frac{1}{2}\log(2e\pi\sigma^2)$.

Property. For discrete random variables, taking values in $\mathcal{X}$, we have, (Proof: try at home),

$$0 \leq H(X) \leq \log(|\mathcal{X}|).$$

Some interpretations, when X is discrete:

- The measure of the spread/uncertainty in a random variable.
- The amount of information in a random variable.

Joint and Conditional Entropy

Joint Entropy. Let random variables X, Y have joint distribution P. The joint entropy is defined as:

$$H(X, Y) \;=\; \mathbb{E}_{X,Y \sim P}[-\log p(X, Y)]\,.$$

Conditional Entropy. Let random variables X, Y have joint distribution P. The conditional entropy of X given Y is:

$$H(X \mid Y) \;=\; \mathbb{E}_{X,Y \sim P}[-\log p(X \mid Y)]\,.$$

Conditional Entropy (cont'd)

To understand the conditional entropy further, let us consider discrete X, Y and define:

$$H(X \mid Y = y) = -\sum_{x \in \mathcal{X}} p(x \mid y) \log p(x \mid y).$$

Measures how much information/uncertainty is left in X after observing $Y = y$.

How much information does Y reveal about X *on average*?

$$\begin{aligned}
\mathbb{E}_{Y' \sim P_Y}\left[H(X \mid Y = Y')\right] &= \sum_{y \in \mathcal{Y}} p(y)\, H(X \mid Y = y) \\
&= \sum_{x,y} -p(x, y) \log p(x \mid y) \\
&= \mathbb{E}_{X,Y \sim P}\left[-\log p(X \mid Y)\right] \\
&= H(X \mid Y).
\end{aligned}$$

Conditional Entropy (cont'd)

Chain rule for conditional entropy. Let $X_1, \ldots, X_n, Y$ be random variables. Then,

1. $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_1, \ldots, X_{i-1})$.
2. $H(X_1, \ldots, X_n \mid Y) = \sum_{i=1}^{n} H(X_i \mid X_1, \ldots, X_{i-1}, Y)$.

Proof Sketch of (1). For two random variables, we have:

$$p(x_1, x_2) = p(x_1)\, p(x_2 \mid x_1),$$

$$\text{Therefore,} \quad \log p(x_1, x_2) = \log p(x_1) + \log p(x_2 \mid x_1).$$

Taking expectation on both sides yields:

$$H(X_1, X_2) = H(X_1) + H(X_2 \mid X_1).$$

The general result follows by induction.

KL Divergence and Mutual Information

KL Divergence (a.k.a. Relative Entropy). Let P and Q be two distributions. Then, the KL divergence is defined as:

$$\mathrm{KL}(P, Q) = \mathbb{E}_{X \sim P}\left[\log\left(\frac{p(X)}{q(X)}\right)\right].$$

Mutual Information. Let X, Y have joint distribution P_{XY} with marginals P_X, P_Y. The MI is the KL between the joint distribution and the product of its marginals:

$$I(X, Y) = \mathrm{KL}(P_{XY}, P_X \times P_Y) = \mathbb{E}\left[\log\left(\frac{p_{X,Y}(X, Y)}{p_X(X)\, p_Y(Y)}\right)\right].$$

$I(X, Y)$ measures how much information X has about Y, and vice versa.

Properties of KL Divergence and Mutual Information (1/3)

1. **Symmetry of Mutual Information:** $I(X, Y) = I(Y, X)$.

2. **Non-negativity of KL divergence:** $\mathrm{KL}(P, Q) \geq 0$, with equality iff $P = Q$.

   Proof: Write $\mathrm{KL}(P, Q) = \mathbb{E}_P\left[-\log\left(\frac{q(X)}{p(X)}\right)\right]$ and apply Jensen's inequality.

3. **Non-negativity of Mutual Information:** $I(X, Y) \geq 0$, with equality iff $X \perp Y$.

   Proof: From property (2), $I(X, Y) = \mathrm{KL}(P_{XY}, P_X P_Y) \geq 0$. Equality holds when $P_{XY} = P_X P_Y$, i.e. $X \perp Y$.

Properties of KL Divergence and Mutual Information (2/3)

4. **MI as Reduction in Entropy:**

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

Proof: From the definition:

$$\begin{aligned} I(X, Y) &= \mathbb{E}_{X,Y}\left[\log\left(\frac{p(X, Y)}{p_X(X)p_Y(Y)}\right)\right] \\ &= \mathbb{E}_{X,Y}\left[\log\left(\frac{p(X|Y)p_Y(Y)}{p_X(X)p_Y(Y)}\right)\right] \\ &= H(X) - H(X|Y). \end{aligned}$$

5. **Conditioning reduces entropy:** $H(X|Y) \leq H(X)$.

   Proof: From (4), we have $H(X) - H(X|Y) = I(X, Y) \geq 0$.

Properties of KL Divergence and Mutual Information (3/3)

6. **MI via Joint Entropy:**

$$I(X, Y) = H(X) + H(Y) - H(X, Y).$$

Proof: By the chain rule, $H(X, Y) = H(Y) + H(X|Y)$. Substituting into property (4): $I(X, Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y)$.

7. **Self-information property:**

$$I(X, X) = H(X).$$

Proof: Using property (4), $I(X, X) = H(X) - H(X|X)$. Since $H(X|X) = 0$, we get $I(X, X) = H(X)$.

Conditional Mutual Information and Chain Rule

Conditional Mutual Information. Let X, Y, Z have joint distribution P. Then,

$$I(X, Y|Z) = H(X|Z) - H(X|Y, Z) \ = \mathbb{E}_{X,Y,Z\sim P}\left[\log\left(\frac{p(X, Y|Z)}{p(X|Z)\, p(Y|Z)}\right)\right].$$

Interpretation: Measures how much information X and Y share *given* Z.

Chain Rule for Mutual Information. For random variables $X_1, \ldots, X_n$ and Y:

$$I((X_1, \ldots, X_n), Y) = \sum_{i=1}^{n} I(X_i, Y \mid X_1, \ldots, X_{i-1}).$$

Chain Rule for MI

Chain Rule for Mutual Information. For random variables $X_1, \ldots, X_n$ and Y:

$$I((X_1, \ldots, X_n), Y) = \sum_{i=1}^{n} I(X_i, Y \mid X_1, \ldots, X_{i-1}) .$$

Proof (for $n = 2$).

$$\begin{aligned} I((X_1, X_2), Y) &= H(X_1, X_2) - H(X_1, X_2|Y) \qquad \texttt{By property 4 above.} \\ &= H(X_1) + H(X_2|X_1) \; - \; (H(X_1|Y) + H(X_2|X_1, Y)) , \\ & \qquad \texttt{Chain rule for entropy.} \\ &= \underbrace{H(X_1) - H(X_1|Y)}_{I(X_1,Y)} + \underbrace{H(X_2|X_1) - H(X_2|X_1, Y)}_{I(X_2,Y|X_1)} . \end{aligned}$$

Data Processing Inequality

Data Processing Inequality. Let X, Y, Z be random quantities such that $X \perp Z \mid Y$. Then, $I(X, Y) \geq I(X, Z)$. And hence, $H(X|Y) \leq H(X|Z)$.

Proof. We will apply the chain rule for MI in two ways:

$$I(X, (Y, Z)) = I(X, Z) + I(X, Y|Z) = I(X, Y) + I(X, Z|Y).$$

Since $X \perp Z \mid Y$, we have $I(X, Z|Y) = 0$. Also, $I(X, Y|Z) \geq 0$. Hence,

$$I(X, Y) \geq I(X, Z).$$

For the conditional entropy statement, note that by property 4:

$$H(X) - H(X|Z) = I(X, Z) \leq I(X, Y) = H(X) - H(X|Y),$$

which implies $H(X|Y) \leq H(X|Z)$.

Ch 0.7: Convex Analysis

Definition (Convex set). A set $\Omega \subset \mathbb{R}^d$ is called *convex* if, for every two points $\omega, \omega' \in \Omega$ and every $\alpha \in [0,1]$, we have $\alpha\omega + (1-\alpha)\omega' \in \Omega$.

Convex functions

Definition (convex function)

- A function $f : \Omega \to \mathbb{R}$ is convex if Ω is a convex set and $\forall \alpha \in [0, 1]$ and all $u, v \in \Omega$ we have, $f(\alpha u + (1 - \alpha)v) \leq \alpha f(u) + (1 - \alpha)f(v)$.

- Equivalently, f is convex if, for all $\omega \in \Omega$, there exists $g \in \mathbb{R}^n$ such that $\forall \omega' \in \Omega$, we have $f(\omega') \geq f(\omega) + g^\top(\omega' - \omega)$.

Subgradients and Subdifferentials

Convex function: A function f is convex if, for all $\omega \in \Omega$, there exists $g \in \mathbb{R}^n$ such that $\forall \omega' \in \Omega$, we have $f(\omega') \geq f(\omega) + g^\top(\omega' - \omega)$.

Any g which satisfies the theorem above is called a subgradient of f at ω.

The set of all subgradients of ω are called the subdifferential, and denoted $\partial f(\omega)$.

Some useful facts about subgradients: (Proofs: Try at home)

- If f is differentiable at ω, then $\partial f(\omega) = \{\nabla f(\omega)\}$.

- $\mathbf{0} \in \partial f(\omega) \iff \omega \in \operatorname{argmin}_{\omega' \in \Omega} f(\omega')$.

- If $g_1 \in \partial f_1(\omega)$ and $g_2 \in \partial f_2(\omega)$, then

$$\alpha g_1 + \beta g_2 \in \partial(\alpha f_1 + \beta f_2) \qquad \text{for all } \alpha, \beta \in \mathbb{R}$$

Strong convexity

Convex function: A function f is convex if, for all $\omega \in \Omega$, there exists $g \in \mathbb{R}^n$ such that $\forall \omega' \in \Omega$, we have $f(\omega') \geq f(\omega) + g^\top(\omega' - \omega)$.

Definition (strong convexity) A convex function $f : \Omega \to \mathbb{R}$ is α–strongly convex in some norm $\| \cdot \|$ if, $f(\omega') \geq f(\omega) + g^\top(\omega' - \omega) + \frac{\alpha}{2}\|\omega' - \omega\|^2 \quad \forall g \in \partial f(\omega)$.

Remark. If f is strongly convex in $\| \cdot \|_2$, this is equivalent to saying that $f(\omega) - \frac{\alpha}{2}\|\omega\|_2^2$ is convex, i.e f is at least as convex as a quadratic function.

Define $h(\omega) = f(\omega) - \frac{\alpha}{2}\|\omega\|_2^2$. Then, $g \in \partial f(\omega) \iff g - \alpha\omega \in \partial h(\omega)$.

$$
\begin{aligned}
&h(\omega') \geq h(\omega) + (g - \alpha\omega)^\top(\omega' - \omega) \iff \\
&\quad f(\omega') - \frac{\alpha}{2}\|\omega'\|_2^2 \geq f(\omega) - \frac{\alpha}{2}\|\omega\|_2^2 + (g - \alpha\omega)^\top(\omega' - \omega) \iff \\
&\quad f(\omega') \geq f(\omega) + g^\top(\omega' - \omega) + \frac{\alpha}{2}\|\omega - \omega'\|_2^2.
\end{aligned}
$$

Strong convexity examples

Example 1. $f(\omega) = \frac{1}{2}\|\omega\|_2^2$ is 1–strongly convex in $\|\cdot\|_2$.
Proof. As $\frac{1}{2}\|\omega\|_2^2 - \frac{1}{2}\|\omega\|_2^2 = 0$ is convex.

Example 2. The negative entropy $f(\omega) = \sum_{i=1}^{K} \omega(i)\log(\omega(i))$ is 1–strongly-convex in $\|\cdot\|_1$, when $\Omega = \Delta([K]) = \{\omega \in \mathbb{R}_+^K; \mathbf{1}^\top \omega = 1\}$.

Proof. As f is differentiable, $\partial f(\omega) = \{\nabla f(\omega)\}$. Therefore, we need to show, for all $\omega, \omega' \in \Delta([K])$, we have

$$f(\omega') \geq f(\omega) + \nabla f(\omega)^\top(\omega' - \omega) + \frac{1}{2}\|\omega' - \omega\|_1^2.$$

$$\iff \quad f(\omega') - f(\omega) - \nabla f(\omega)^\top(\omega' - \omega) \geq \frac{1}{2}\|\omega' - \omega\|_1^2.$$

Note that

$$\frac{\partial f(\omega)}{\partial \omega(i)} = 1 + \log(\omega(i)).$$

Strong convexity examples (cont'd)

We need to show, for all $\omega, \omega' \in \Delta([K])$, we have

$$f(\omega') - f(\omega) - \nabla f(\omega)^\top (\omega' - \omega) \geq \frac{1}{2}\|\omega' - \omega\|_1^2.$$

We just showed $\frac{\partial f \omega(i))}{\partial \omega(i)} = 1 + \log(\omega(i))$.

Therefore,

$$\begin{aligned}
\mathrm{LHS} &= \sum_{i=1}^{K} \omega'(i) \log(\omega'(i)) - \sum_{i=1}^{K} \omega(i) \log(\omega(i)) - \sum_{i=1}^{K} (1 + \log(\omega(i)))(\omega'(i) - \omega(i)) \\
&= \sum_{i=1}^{K} \omega'(i) \log\left(\frac{\omega'(i)}{\omega(i)}\right) \\
&= \mathrm{KL}(\omega', \omega) \geq \frac{1}{2}\|\omega' - \omega\|_1^2.
\end{aligned}$$

The last step follows by Pinsker's, $\mathrm{KL}(P, Q) \geq 2\mathrm{TV}(P, Q)^2 = 2\left(\frac{1}{2}\|P - Q\|_1\right)^2$.

Some useful properties about strongly convex functions

1. If f_1 is α-strongly convex and f_2 is convex then $\beta f_1 + f_2$ is $(\beta\alpha)$–strongly convex.

Proof. *Try at home*

2. If $\omega_\star = \operatorname{argmin}_{\omega \in \Omega} f(\omega)$, where f is α–strongly convex with respect to $\|\cdot\|$, then $f(\omega) \geq f(\omega_\star) + \frac{\alpha}{2}\|\omega - \omega_\star\|^2$.

Proof. By definition of strong convexity,

$$f(\omega) \geq f(\omega_\star) + g^\top(\omega - \omega_\star) + \frac{\alpha}{2}\|\omega - \omega_\star\|^2, \qquad \text{for all } g \in \partial f(\omega_\star).$$

Claim follows by noting $\mathbf{0} \in \partial f(\omega_\star)$.

□

Dual norm

Definition (dual norm) Given a norm $\|\cdot\|$, the dual norm $\|\cdot\|_\star$ is defined as

$$\|\omega\|_\star = \max_{\|u\|\le 1} u^\top \omega.$$

Examples of dual norm pairs: $(\|\cdot\|_2, \|\cdot\|_2)$, $(\|\cdot\|_1, \|\cdot\|_\infty)$.

For the 2-norm,

$$\|\omega\|_\star = \max_{\|u\|_2\le 1} u^\top \omega = \left(\frac{\omega}{\|\omega\|_2}\right)^\top \omega = \|\omega\|_2.$$

For the ∞-norm,

$$\|\omega\|_\star = \max_{\|u\|_\infty\le 1} u^\top \omega = \mathbf{1}^\top |\omega| = \|\omega\|_1.$$

Dual norm (cont'd)

$$\|\omega\|_\star = \max_{\|u\| \leq 1} u^\top \omega.$$

Hölder's inequality For all $a, b \in \mathbb{R}^d$, we have $a^\top b \leq \|a\| \cdot \|b\|_\star$.

Proof.

$$\begin{aligned} a^\top b &= b^\top \left(\frac{a}{\|a\|} \right) \|a\| \\ &\leq \|a\| \max_{\omega, \|\omega\| \leq 1} b^\top \omega \\ &= \|a\| \cdot \|b\|_\star. \end{aligned}$$