CS7545, Spring 2023: Machine Learning Theory - Solutions #1

Jacob Abernethy, Zihao Hu, Guanghui Wang, Yeojoon Youn Due: Tuesday, January 31 at 11:59 p.m.

1) **Norm.** We will prove a generic statement which implies (a)-(d).

Let $p > q \geq 1$, and r be a number such that $\frac{1}{p} + \frac{1}{r} = \frac{1}{q}$. Then, $q < p, r$ and $(\frac{p}{q}, \frac{r}{q})$ is a conjugate norm pair. Let $\mathbf{a} \in \mathbb{R}^N$ such that $a_i = |x_i|^q$, and let $\mathbf{y} = (1, \ldots, 1) \in \mathbb{R}^N$. Now we use Holder's inequality:

$$\mathbf{a}^\top \mathbf{y} = \sum_{i=1}^N |x_i|^q \leq \|\mathbf{a}\|_{\frac{p}{q}} \|\mathbf{y}\|_{\frac{r}{q}} = \left(\sum_{i=1}^N |x_i|^p\right)^{\frac{q}{p}} n^{\frac{q}{r}}.$$

By exponentiating each side with $1/q$, we get

$$\|\mathbf{x}\|_q \leq \|\mathbf{x}\|_p n^{\frac{1}{r}} = \|\mathbf{x}\|_p n^{\frac{1}{q} - \frac{1}{p}}$$

Also note that

$$\|\mathbf{x}\|_q^p = \left(\sum_{i=1}^N |x_i|^q\right)^{\frac{p}{q}} \geq \sum_{i=1}^N |x_i|^p = \|\mathbf{x}\|_p^p$$

which implies $\|\mathbf{x}\|_q \geq \|\mathbf{x}\|_p$. The inequality follows since $(\sum |x_i|)^\alpha \geq \sum |x_i|^\alpha$ whenever $\alpha \geq 1$.

2) **Hölder.**

(a) Let $p > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$.Consider the following two vectors:

$$\mathbf{x}, \mathbf{y} \in \mathbb{R}^N : x_i = p_i^{\frac{1}{q} - 1}, y_i = p_i^{\frac{1}{p}},$$

then by Hölder's Inequality,

$$\|\mathbf{x}\|_q \|\mathbf{y}\|_p \geq \mathbf{x}^T \mathbf{y} = \sum_i p_i^{\frac{1}{p} + \frac{1}{q} - 1} = N.$$

where $\|\mathbf{x}\|_q = (\sum_i p_i^{1-q})^{\frac{1}{q}}$ and $\|\mathbf{y}\|_p = 1$. Therefore,

$$\sum_i \left(\sum_i \frac{1}{p_i^{q-1}}\right)^{\frac{1}{q}} \geq N \Rightarrow \sum_i \frac{1}{p_i^{q-1}} \geq N^q.$$

Remark. You can also use Jensen's inequality. Consider the function $f(p) = \frac{1}{p^q}$ and note that $f(\sum_{i=1}^N p_i p_i) \leq \sum_{i=1}^N p_i f(p_i)$.

(b) By Jensen's Inequality,

$$\sum_i p_i^2 \geq \sum_i \frac{p_i}{N} = \frac{1}{N}.$$

Therefore, we have

$$\sum_i \left(\frac{1}{p_i} + p_i\right)^2 = \sum_i p_i^2 + \sum_i 2 + \sum_i \frac{1}{p_i^2} \geq \frac{1}{N} + 2N + N^3.$$

Remark. You can use 1(a) to show that $\sum_i p_i^2 = \|\mathbf{p}\|_2^2 \geq \frac{\|\mathbf{p}\|_1^2}{N}$.

3) **Projection.**

(a) If $x \in \mathcal{K}$, the projection of x is obviously x. Suppose that for some point $x \notin \mathcal{K}$, the projection $\Pi_{\mathcal{K}}(x) > 1$ (we know $\Pi_{\mathcal{K}} > 0$ because the set is non-empty and closed). Let y_1 and y_2 be two points in $\Pi_{\mathcal{K}}(x)$ and let $||y_1 - x||_2 = ||y_2 - x||_2 = d$. Let $z = \alpha y_1 + (1-\alpha) y_2$ for $\alpha \in (0,1)$, so z is on the line between y_1 and y_2. Then by the triangle inequality:

$$\begin{aligned} ||x - z||_2 &= ||x - \alpha y_1 - (1-\alpha) y_2||_2 && (1)\\ &\leq \alpha \, ||x - y_1||_2 + (1-\alpha) \, ||x - y_2||_2 && (2)\\ &= d && (3) \end{aligned}$$

Since y_1 and y_2 are projections of x, we know $||x - z||_2 \geq d$, so the triangle inequality holds with equality. This can only occur when $x - y_1$ is collinear with $x - y_2$, which means either $y_1 = y_2$ or x is on the line between y_1 and y_2, which would imply $x \in \mathcal{K}$ by convexity. Both of these would cause a contradiction.

Alternately, consider a ball of radius d around x, which must intersect $\mathcal{K}$ at y_1 and y_2. The line between y_1 and y_2 lies inside this ball, which means every point between y_1 and y_2 is strictly closer to x than y_1 and y_2.

(b) We need to show that if $\mathcal{K}$ is non-convex, there is a point x that has a non-unique projection. Assume for contradiction that all points have unique projections. For non-convex $\mathcal{K}$, we know there are two points $y_1, y_2 \in \mathcal{K}$ such that for some $\alpha \in (0,1)$, $z = \alpha y_1 + (1-\alpha) y_2$ is not in $\mathcal{K}$. If z has a non-unique projection, we are done. Else, let x_0 be the unique projection of z. Then every point on the ray from x_0 in the direction of z must also uniquely project to x_0 (otherwise there'd be some point that projected to two points in $\mathcal{K}$). Take a point z_t at distance t from x_0 on the ray from x_0 towards z. Then by definition of projection of z_t:

$$t = ||z_t - x_0||_2 < ||z_t - y_1||_2 \tag{4}$$

However, for sufficiently large t, $z_t - x_0$ and $z_t - y_0$ will be approximately collinear, and we will have $||z_t - y_1||_2 < ||z_t - x_0||_2$ or $||z_t - y_2||_2 < ||z_t - x_0||_2$, which contradicts the fact that z_t projects to x_0.

4) **Fenchel.**

(a) The conjugate of f_α is defined as

$$f_\alpha^*(\theta) = \sup_{\mathbf{x}} \mathbf{x}^T \theta - f_\alpha(\mathbf{x}) = \alpha \left(\sup_{\mathbf{x}} \mathbf{x}^T \frac{\theta}{\alpha} - f(\mathbf{x}) \right) = \alpha g \left(\frac{1}{\alpha} \theta \right).$$

(b) The conjugate of f is defined as

$$f^*(\theta) = \sup_x x\theta - \sqrt{1+x^2}.$$

Let $h(x,\theta) = x\theta - \sqrt{1+x^2}$. As h is strictly concave in x, $\frac{\partial h(x,\theta)}{\partial x}$ has at most one zero for a fixed θ. We have

$$\frac{\partial h(x,\theta)}{\partial x} = \theta - \frac{x}{\sqrt{1+x^2}}.$$

As $|\frac{x}{\sqrt{1+x^2}}| < 1$ for all $x \in \mathbb{R}$, consider the three cases:

- $|\theta| > 1$, then $h(x,\theta)$ is monotonic in x since $|\frac{\partial h(x,\theta)}{\partial x}| > |\theta| - 1 > 0$. Therefore $f^*(\theta)$ is not defined.

- $|\theta| < 1$, then the supremum is achieved where the gradient is zero, i.e., $x = \frac{\theta}{\sqrt{1-\theta^2}}$. Therefore we have $f^*(\theta) = -\sqrt{1-\theta^2}$.
- $|\theta| = 1$. For $\theta = 1$ the gradient approaches 0 as x goes to infinity, and hence

$$f^*(\theta) = \lim_{x\to\infty} x - \sqrt{1+x^2} = 0.$$

  Similarly, we have $f^*(-1) = 0$.

To summerize, we have $f^*(\theta) = -\sqrt{1-\theta^2}, \theta \in [0, 1]$.

5) **Hoeffding.**

(a) By Markov's Inequality, $\Pr[X \geq t] \leq \Pr[e^{\lambda X} \geq e^{\lambda t}] \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}}$ for all $\lambda \geq 0$. Therefore

$$\Pr[X \geq t] \leq \inf_{\lambda \geq 0} e^{-\lambda t + \log \mathbb{E}[e^{\lambda X}]} = \exp\{-\sup_{\lambda \geq 0} \lambda t - \log \mathbb{E}[e^{\lambda X}]\}$$

It suffices to prove that $\sup_{\lambda \geq 0} \lambda t - \log \mathbb{E}[e^{\lambda X}] = \sup_{\lambda} \lambda t - \log \mathbb{E}[e^{\lambda X}] = f^*(t)$ given that $t > \mathbb{E}[X]$.

We first prove that $\lambda t - \log \mathbb{E}[e^{\lambda X}] \leq 0 \cdot t - \log \mathbb{E}[e^{0 \cdot X}] = 0$ for all $\lambda \leq 0$. We have

$$\lambda t - \log \mathbb{E}[e^{\lambda X}] \leq \lambda \mathbb{E}[X] - \log \mathbb{E}[e^{\lambda X}] = \log \frac{e^{\lambda \mathbb{E}[X]}}{\mathbb{E}[e^{\lambda X}]}.$$

As $f(x) = e^{\lambda x}$ is convex, by Jensen's inequality we have $0 \leq e^{\lambda \mathbb{E}[X]} \leq \mathbb{E}[e^{\lambda X}]$, therefore $\log \frac{e^{\lambda \mathbb{E}[X]}}{\mathbb{E}[e^{\lambda X}]} \leq 0$, which means that $\sup_{\lambda \leq 0} \lambda t - \log \mathbb{E}[e^{\lambda X}] \leq 0 \leq \sup_{\lambda \geq 0} \lambda t - \log \mathbb{E}[e^{\lambda X}] \Rightarrow \sup_{\lambda \geq 0} \lambda t - \log \mathbb{E}[e^{\lambda X}] = \sup_{\lambda} \lambda t - \log \mathbb{E}[e^{\lambda X}]$.

Remark. You want to prove for every t, not every λ. Nowhere in the statement does λ actually appear; it appears as a *dummy variable* inside the definition of convex conjugate. Showing the statement separetely for positive and negative λ doesn't work here.

(b) Let $X_i^{(j)}$ be the random variable denoting the result of the j-th toss of the i-th coin (1 for a head and 0 for a tail). For notational convenience, we assume the first coin is the special one.

Define $Y^{(j)} = \frac{X_1^{(j)} - X_2^{(j)} + 1}{2}$, which is a random variable taking values in $[0, 1]$. Note that

$$\mathbb{E}[Y^{(j)}] = \frac{1+\rho}{2}$$

and that

$$\sum_{j=1}^{n} X_1^{(j)} \geq \sum_{j=1}^{n} X_2^{(j)} \text{ if and only if } \frac{1}{n}\sum_{j=1}^{n} Y^{(j)} \geq \frac{1}{2}.$$

So, we can bound the probability that the coin 2 came up heads more frequently than the coin 1 did, using Hoeffding's lemma:

$$\Pr\left[\frac{1}{n}\sum_{j=1}^{n} Y^{(j)} \leq \frac{1}{2}\right] = \Pr\left[\frac{1}{n}\sum_{j=1}^{n} Y^{(j)} - \frac{1+\rho}{2} \leq -\frac{\rho}{2}\right] \leq \exp\left(-\frac{n\rho^2}{2}\right).$$

Since all the coins except the speical one is i.i.d., we can apply the union bound and conclude that

$$\Pr\left[\exists i \neq 1 : \sum_{j=1}^{n}(X_i^{(j)} - X_1^{(j)}) \geq 0\right] \leq (m-1)\exp\left(-\frac{n\rho^2}{2}\right).$$

So, the speical coin has the most heads with probability at least $1 - m\exp\left(-\frac{n\rho^2}{2}\right)$.

Remark. You can also show that the probability of the biased coin coming up with most heads is lower bounded by the probabililty that the biased coin gets at least $\frac{1+\rho}{2}n$ heads and the rest gets at most $\frac{1+\rho}{2}n$ heads. If you do this, you get a bound that looks like $\left(1-\exp\left(-\frac{n\rho^2}{2}\right)\right)^m$, which for a small enough ρ is eseentially the same as the above bound.

A common mistake is to consider Z_i, the difference between the number of heads between the biased coin and the non-biased coin i, and assume independence among Z_i's. Although you get the same $\left(1-\exp\left(-\frac{n\rho^2}{2}\right)\right)^m$ bound, this argument is wrong.

(c) It suffices to choose n such that $m\exp(-\frac{1}{2}n\rho^2) \leq \delta$, or equivalently, $n \geq 2\rho^{-2}\log\frac{m}{\delta}$ to ensure that the probability of failure is upper bounded by δ.

6) **Shannon.**

(a) Let $h(\boldsymbol{\theta}) = \langle \mathbf{x}, \boldsymbol{\theta}\rangle - g(\boldsymbol{\theta})$. The supremum of h is achieved when $\nabla g(\boldsymbol{\theta}) = \mathbf{x}$. The i-th coordinate of the gradient of g is $\frac{\exp(\theta_i)}{\sum_{j=1}^{n}\exp(\theta_j)}$. If $\theta_i = \log x_i$, $\nabla g(\theta) = \frac{x_i}{\sum_{j=1}^{n}x_j} = x_i$. So, $g^*(\mathbf{x}) = \langle \mathbf{x}, \log\mathbf{x}\rangle - g(\log\mathbf{x}) = \langle \mathbf{x}, \log\mathbf{x}\rangle$ because $g(\log\mathbf{x}) = \log(\sum x_i) = 0$ for any $\mathbf{x} \in \Delta^N$.

It remains to show that the domain of g is Δ^N. Suppose $x_k < 0$. Then, h grows unboundedly by setting $\theta_k = -t$ and $\theta_i = 0$ for all $i \neq k$, as t goes to infinity. Now suppose $x_i \geq 0$ for every i. Consider $\boldsymbol{\theta} = (t, \ldots, t)$. Then,

$$h(\boldsymbol{\theta}) = t\sum x_i - t - \log n.$$

Setting $t = +\infty$ if $\sum x_i > 1$ and $t = -\infty$ if $\sum x_i < 1$ shows h is unbounded.

Remark. You can skip the part that shows the domain of g^* is indeed Δ^N, if you prove the other duality, i.e., $f^* = g$. Note that generally speaking, for any convex function h, $h(x) = \infty$ for all $x \notin \mathrm{dom}(h)$. (which makes sense considering the definition of convex conjugate.)

(b) The bregman diveregence of the negative entropy function is exactly the KL divergence. Let $\mathbf{x} = (p, 1-p)$ and $\mathbf{y} = (q, 1-q)$. Without loss of generality, we assume that $p \geq q$. Note $\|\mathbf{x}-\mathbf{y}\|_1 = 2(p-q)$ Define $h_p(q) = D_f(\mathbf{x}, \mathbf{y}) - \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_1^2$, and differentiate it:

$$\begin{aligned}
\frac{\partial h_p}{\partial q} &= \frac{\partial\left(p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} - 2(p-q)^2\right)}{\partial q} \\
&= -\frac{p}{q} + \frac{1-p}{1-q} + 4(p-q) \\
&= \frac{q-p}{q(1-q)} + 4(p-q) \\
&= (p-q)\left(4 - \frac{1}{q(1-q)}\right).
\end{aligned}$$

Since $p \geq q$ and $\frac{1}{q(1-q)} \geq 4$, the derivative is non-positive. Therefore, h_p is minimized at the boundary $p = q$, where h_p is exactly 0. This proves the 1-strong convexity of the negative entropy function.

(c) For general Δ_n, note that $\|\mathbf{x}-\mathbf{y}\|_1 = \sum_{i:x_i \leq y_i}(y_i - x_i) + \sum_{i:x_i > y_i}(x_i - y_i)$. Consider the following two vectors in Δ_2:

$$\mathbf{x}_A = \left(\sum_{i\in A} x_i, \sum_{i\notin A} x_i\right) \text{ and } \mathbf{y}_A = \left(\sum_{i\in A} y_i, \sum_{i\notin A} y_i\right)$$

where $A = \{i : x_i > y_i\}$. Then, we have $\|\mathbf{x}-\mathbf{y}\|_1 = \|\mathbf{x}_A - \mathbf{y}_A\|_1$. We certainly have $D(\mathbf{x}_A\|\mathbf{y}_A) \geq \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_1$ by (b) and $D(\mathbf{x}\|\mathbf{y}) \geq D(\mathbf{x}_A\|\mathbf{y}_A)$ by data-processing inequality; therefore the 1-strong convexity hold for general Δ_n.

7) **Bregman.**

(a) By definition,

$$\begin{aligned}&D_f(\mathbf{x},\mathbf{y}) + D_f(\mathbf{y},\mathbf{z}) - D_f(\mathbf{x},\mathbf{z})\\ &= f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + f(\mathbf{y}) - f(\mathbf{z}) - \langle \nabla f(\mathbf{z}), \mathbf{y}-\mathbf{z}\rangle - f(\mathbf{x}) + f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x}-\mathbf{z}\rangle\\ &= \langle \nabla f(\mathbf{z}) - \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle.\end{aligned}$$

(b) Take $\mathbf{z} = \mathbf{x}$ in the equation above, we have

$$D_f(\mathbf{y},\mathbf{x}) + D_f(\mathbf{x},\mathbf{y}) = \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle.$$

Strong convexity indicates that

$$D_f(\mathbf{y},\mathbf{x}) + D_f(\mathbf{x},\mathbf{y}) \geq \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2 + \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2 = \|\mathbf{x}-\mathbf{y}\|^2;$$

on the other hand, by Hölder's inequality we have $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \leq \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*\|\mathbf{x}-\mathbf{y}\|$. Concatenating these two inequalities we have

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \geq \|\mathbf{x}-\mathbf{y}\|,$$

which still holds for the special case $\|\mathbf{x}-\mathbf{y}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$.

(c) Denote f^* the Fenchel duality of f. By the properties of Fenchel duality, we have

$$D_f(\mathbf{x},\mathbf{y}) = D_{f^*}(\nabla f(\mathbf{y}), \nabla f(\mathbf{x})).$$

Also, f is 1-strongly convex with respect to $\|\cdot\|$ if and only if f^* is 1-strongly smooth with respect to $\|\cdot\|_*$. Therefore we only need to prove the latter proposition. Fenchel duality tells us that

$$\nabla f^*(\nabla f(\mathbf{x})) = \mathbf{x}.$$

As $\nabla f^*(\theta) = \nabla f^{-1}(\theta)$, by $\forall \mathbf{x},\mathbf{y}, \|\mathbf{x}-\mathbf{y}\| \leq \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*$ we have

$$\|\nabla f^*(\psi) - \nabla f^*(\varphi)\| \leq \|\psi - \varphi\|_*.$$

For arbitrarily fixed φ, θ, consider the function $g(\alpha) = D_{f^*}(\varphi+\alpha\theta, \varphi) = f^*(\varphi+\alpha\theta) - f^*(\varphi) - \alpha\nabla f^*(\varphi)^T\theta$. It suffices to prove that $g(\alpha) \leq \frac{1}{2}\alpha^2\|\theta\|_*^2$. Apparently,

$$g(0) = 0, g'(\alpha) = \langle \nabla f^*(\varphi+\alpha\theta) - \nabla f^*(\varphi), \theta\rangle \leq \alpha\|\theta\|_*^2,$$

therefore $g(\alpha) = \int_0^\alpha g'(\beta)d\beta \le \|\theta\|_*^2 \int_0^\alpha \beta d\beta = \frac{1}{2}\alpha^2\|\theta\|_*^2$. Therefore f^* is 1-strongly smooth and f is 1-strongly convex.

With the stronger condition $\forall \mathbf{x}, \mathbf{y}, \langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \ge \|\mathbf{x} - \mathbf{y}\|_2^2$, however, the proof can be much easier as we can get rid of Cauchy Schwarz. Use the method above, for fixed $\mathbf{y}, \mathbf{z}$ let $h(\alpha) = D_f(\mathbf{y}+\alpha\mathbf{z}, \mathbf{y})$. We only need to prove that $h(\alpha) \ge \frac{1}{2}\alpha^2\|\mathbf{z}\|^2$ for all α. Apparently,

$$h'(\alpha) = \langle \nabla f(\mathbf{y} + \alpha\mathbf{z}) - \nabla f(\mathbf{y}), \mathbf{z}\rangle \ge \alpha\|\mathbf{z}\|^2.$$

Also by $h(0) = 0$ we conclude $h(\alpha) \ge \int_0^\beta \alpha\|\mathbf{z}\|^2 d\beta = \frac{1}{2}\alpha^2\|\mathbf{z}\|^2$, finishing the proof.