

CSE 203B W21 Homework 3

Due Time : 11:50pm, Wednesday Feb. 1, 2023 Submit to Gradescope
Gradescope: <https://gradescope.com/>

In this homework, we work on exercises from text book including level sets of convex, concave, quasi-convex, quasi-concave functions (3.1, 3.2), second-order conditions for convexity on affine sets (3.9), Kullback-Leibler divergence (3.13), saddle points of convex-concave functions (3.14) determination of convex, concave, quasi-convex, quasi-concave functions (3.16), conjugate functions (3.36), and gradient and Hessian of conjugate functions (3.40). Extra assignments are given on softmax functions, dual norms, and a conjugate function.

Exercises are graded by completion, assignments are graded by content. We may just grade a subset of the problems.

I. Exercises from textbook chapter 3 (8 pts)

3.1, 3.2, 3.9, 3.13, 3.14, 3.16, 3.36, 3.40.

II. Assignments (42 pts)

II. 1. Softmax Functions.

Given a function $f(x) = \max_i x_i - \min_i x_i$, where $x = [x_i] \in \mathbb{R}^n$, we use a softmax expression $\tilde{f}(x) = \log \sum_i e^{x_i} + \log \sum_i e^{-x_i}$ to approximate the function $f(x)$.

II.1.1. Prove or disprove that function $f(x)$ is convex.

Solution: Prove $g(x) = \max(x_1, x_2, \dots, x_n)$ is convex:
for $0 \leq \theta \leq 1$

$$\begin{aligned} g(\theta x + (1 - \theta)y) &= \max_i (\theta x + (1 - \theta)y) \\ &\leq \theta \max_i x_i + (1 - \theta) \max_i y_i \\ &= \theta g(x) + (1 - \theta)g(y) \end{aligned}$$

Prove $h(x) = \min(x_1, x_2, \dots, x_n)$ is concave:
for $0 \leq \theta \leq 1$

$$\begin{aligned} h(\theta x + (1 - \theta)y) &= \min_i (\theta x + (1 - \theta)y) \\ &\geq \theta \min_i x_i + (1 - \theta) \min_i y_i \end{aligned}$$

$$= \theta h(x) + (1 - \theta)h(y)$$

Sum of two convex functions is convex

II.1.2. Prove or disprove that the approximation function $\tilde{f}(x)$ is convex.

Solution:

Let $g(x) = \log \sum_{k=1}^n \exp x_k$, $z_k = \exp x_k$ and $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise

$$\frac{\partial}{\partial x_j} g(x) = \frac{\partial}{\partial z_j} \log 1^t z. \quad \frac{\partial z_j}{\partial x_j} = \frac{1}{1^t z} z_j$$

Second derivative:

$$\begin{aligned} \frac{\partial^2 g}{\partial x_i \partial x_j} &= \frac{\partial}{\partial z_i} \left(\frac{z_j}{1^t z} \right) \cdot \frac{\partial z_i}{\partial x_i} \\ &= \frac{\delta_{ij} 1^t z - z_j}{(1^t z)^2} \cdot \exp x_i \\ &= \frac{\delta_{ij} z_i \cdot 1^t z - z_j z_i}{(1^t z)^2} \\ &= \frac{\delta_{ij} z_i}{1^t z} - \frac{z_i z_j}{(1^t z)^2} \\ &= \left(\frac{1}{1^t z} \text{diag}(z) - \frac{1}{(1^t z)^2} z z^t \right)_{ij} \end{aligned}$$

to show $\nabla^2 g(x) \geq 0$, we must verify that $v^T \nabla^2 g(x) v \geq 0$ for all v :

$$v^T \nabla^2 g(x) v = \frac{(\sum_k z_k v_k^2) (\sum_k z_k) - (\sum_k v_k z_k)^2}{(\sum_k z_k)^2} \geq 0$$

since, $(\sum_k v_k z_k)^2 \leq (\sum_k z_k v_k^2) (\sum_k z_k)$, from Cauchy-Schwartz inequality.

We can similarly prove $h(x) = \log \sum_{k=1}^n e^{-x_k}$ is convex by take second derivative. Therefore, the sum of convex functions is a convex function.

II.1.3. Show and prove the worst error of the approximation.

Solution:

Consider the case when $(n - 1)$ entries in x are equal to x_{max} and one entry is 0.

$$\begin{aligned} \tilde{f}(x) &= \log((n - 1)e^{x_{max}} + 1) + \log((n - 1)e^{-x_{max}} + 1) \\ &\approx \log(n - 1) + x_{max} + \log((n - 1)e^{-x_{max}} + 1) \\ f(x) &= x_{max} \end{aligned}$$

$$\tilde{f}(x) - f(x) \approx \log(n - 1) + \log((n - 1)e^{-x_{max}} + 1)$$

Now, consider the case when all (n) entries in x are equal to x_{max} .

$$\tilde{f}(x) = \log((n)e^{x_{max}}) + \log((n)e^{-x_{max}})$$

$$f(x) = 0$$

$$\tilde{f}(x) - f(x) = 2\log(n)$$

Maximum approximation error can be up to $2\log(n)$

II.1.4. Design an improved approximation function that improves the worst error. Show and prove the worst ratio of the new function.

Solution: First approximation:

$$\max_{x_i} \approx \text{LogSumExp}_\alpha(x_1, \dots, x_n) = \frac{1}{\alpha} \log \sum_{i=1}^n \exp \alpha x_i$$

$$\min_i \approx \frac{1}{\alpha} \log \sum_{i=1}^n \exp -\alpha x_i$$

The worst error, in this case, will be $\frac{1}{\alpha}$ times than in the last case.

Second approximation: Using Mellowmax Operator: [Link](#)

You are free to use other approximations as well.

II.1.5. Prove or disprove that your approximation function is convex.

Solution: Proof is similar to proof in 1.2 by taking the second derivative of subproblems. Both the subproblems need to be convex so that their sum is convex.

II. 2. Dual Norm.

Given a dual norm $f(x) = \max_{\|y\|_p} y^T x$, where $x, y \in R^n$.

II.2.1. Prove that the function can be expressed as $f(x) = \|x\|_q$, where $1/p + 1/q = 1$, when $p \geq 1$.

Proof 1:

Solve $\nabla_y \frac{x^T y}{\|y\|_p} = 0$. Using the fractions differentiation, we can simplify the

equation and arrive at the expression $x_i \|y\|_p = (x^T y) \frac{y_i^{p-1} \|y\|_p}{\sum_{k=1}^n y_k}$. Next, we raise

each x_i to the exponential of $\frac{p}{p-1}$ and sum up the n elements. This will lead

us to $\sum x_i^{\frac{p}{p-1}} = (x^T y)^{\frac{p}{p-1}}$. Last, we let $q = \frac{p}{p-1}$, and we can express $x^T y$

as $(\sum x_i^q)^{\frac{1}{q}}$, which is equivalent to $\|x\|_q$

Proof 2: User Holder's Inequality to find the maximum value that the dual norm can achieve. Then for specific value of y , prove that it attains that maximum value. Follow this link for complete proof: [Link](#)

II.2.2. Does the dual norm $f(x) = \max_{\|y\|_p \leq 1} y^T x$ remain to be convex when $0 < p < 1$? Explain your answer.

Solution: It is a convex function based on the property that pointwise supremum preserves the convexity of a function.

II.2.3. Derive the formula of the dual norm $f(x)$ when $0 < p < 1$.

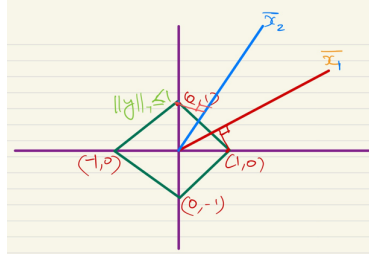


Figure 1: Dual norm of L1 norm

In the above figure 1, let us consider an L1 norm with $\|y\|_1 \leq 1$. The dual norm is given by $\max_{\|y\|_1 \leq 1} y^T x$. Consider a vector x_1 the supremum is projection of all the point in the norm (green color). So, the maximum projection length is of corner point $(1, 0)$, all the other points will have smaller projection. Similar case for vector x_2 , the maximum projection will be of point $(0, 1)$. Therefore the dual norm of L_1 norm is L_∞ norm.

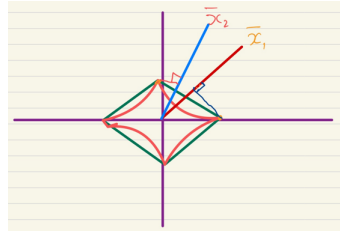


Figure 2: Dual norm of L_p for $0 < p < 1$ norm

Now for L_p for $0 < p < 1$ norms the argument is similar to that in L_1 norm. For any vector x such as x_1 or x_2 . The dual norm is the maximum projection of all the points in $\|y_p \leq 1\|$. It turns out that corner points have the maximum projection similar to L_1 norm. Therefore dual norm of all $0 < p < 1$ is L_∞ .

II. 3. Conjugate Functions.

Find the conjugate function of the following function.

$$f(x) = \begin{cases} \|Ax + b\|_2^2, & \|Ax + b\|_2 \leq \alpha, \\ \alpha(2\|Ax + b\|_2 - \alpha), & \|Ax + b\|_2 > \alpha. \end{cases}$$

where $A \in R^{mn}$, $x \in R^n$, $b \in R^m$, $\alpha \in R_{++}$.

Solution:

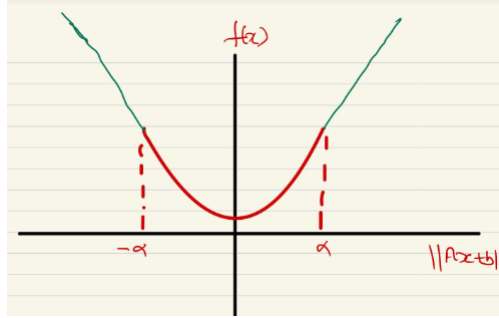


Figure 3: Visualize Conjugate as supporting hyperplane of $f(x)$

By definition, the conjugate function $f^*(y) = \sup_x t^T x - f(x)$, let

$$g(y) = \sup_x (y^T x - \|Ax + b\|_2^2), \quad \|Ax + b\|_2 \leq \alpha,$$

$$h(y) = \sup_x (y^T x - \alpha(2\|Ax + b\|_2 - \alpha)), \quad \|Ax + b\|_2 > \alpha.$$

we have, $f^*(y) = \max(g(y), h(y))$.

Note that A is not guaranteed to be invertible. Therefore we are going to use pseudo inverse, A^\dagger , for A in the following part.

Case 1: $\|Ax + b\| \leq \alpha$

$$\nabla_x g(y) = \nabla_x (y^T x - (Ax + b)^T (Ax + b))$$

By setting $\nabla_x g(y) = 0$, we get

$$y - 2A^T(Ax + b) = 0$$

$$\frac{1}{2}y = A^T(Ax + b) = A^T Ax + A^T b$$

$$\frac{1}{2}y - A^T b = A^T Ax$$

$$(A^T A)^\dagger (\frac{1}{2}y - A^T b) = x$$

Therefore, we have $x = \hat{x} = \frac{1}{2}(A^T A)^\dagger y - A^\dagger (A^T)^\dagger A^T b$ to reach the maximum. Note that, $A\hat{x} + b = \frac{1}{2}A(A^T A)^\dagger y - AA^\dagger (A^T)^\dagger A^T b + b = \frac{1}{2}(A^T)^\dagger y$

Plugging $A\hat{x} + b = \frac{1}{2}(A^T)^\dagger y$ back into $g(y)$, we have

$$\begin{aligned} g(y) &= \frac{1}{2}y^T(A^T A)^\dagger y - y^T(A^T A)^\dagger A^T b - \left(\frac{1}{2}(A^T)^\dagger y\right) \left(\frac{1}{2}(A^T)^\dagger y\right) \\ &= \frac{1}{2}y^T(A^T A)^\dagger y - y^T(A^T A)^\dagger A^T b - \frac{1}{4}y^T(A^T A)^\dagger y \\ &= \frac{1}{4}y^T(A^T A)^\dagger y - y^T A^\dagger b \end{aligned}$$

Case 2: $\|Ax + b\| \geq \alpha$

In this region supporting hyperplane will be linear and parallel to $f(x)$.

$$\nabla_x h(y) = \nabla_x (y^T x - \alpha(2\|Ax + b\|_2) - \alpha)$$

By setting it to zero, we get¹:

$$\begin{aligned} y &= \frac{2\alpha A^T(Ax + b)}{\|Ax + b\|_2} \\ (A^T)^\dagger y &= \frac{2\alpha}{Ax + b} \|Ax + b\|_2 \end{aligned}$$

by square both side, we have:

$$\begin{aligned} ((A^T)^\dagger y)^T ((A^T)^\dagger y) &= 4\alpha^2 \frac{(Ax + b)^T(Ax + b)}{\|Ax + b\|_2^2} \\ &= 4\alpha^2 \end{aligned}$$

Therefore, $\|(A^T)^\dagger y\|_2 = 2\alpha$.

Case 2.1: When $\|(A^T)^\dagger y\|_2 \leq 2\alpha$, we have

$$\begin{aligned} \left\| \frac{1}{2}(A^T)^\dagger y \right\|_2 &\leq \alpha < \|Ax + b\|_2 \\ \frac{1}{2}((A^T)^\dagger y)^T(Ax + b) &\leq \|(A^T)^\dagger y\|_2 \|Ax + b\|_2 \\ \implies y^T(x + A^\dagger b) &\leq \|(A^T)^\dagger y\|_2 \|Ax + b\|_2 \\ y^T x - \alpha(2\|Ax + b\|_2 - \alpha) &\leq (\|(A^T)^\dagger y\|_2 - 2\alpha) \|Ax + b\|_2 + \alpha^2 - y^T A^\dagger b \end{aligned}$$

We get $(\|(A^T)^\dagger y\|_2 - 2\alpha) \|Ax + b\|_2$ is negative. To maximize the bound, we can only minimize $\|Ax + b\|_2$, which may only approach α as the smallest value here. We now have:

¹<https://math.stackexchange.com/questions/2849868/derivative-of-norm-2> shows how to solve derivative of norm 2.

$$\sup (y^T x - \alpha(2\|Ax + b\|_2 - \alpha)) = \alpha\|(A^T)^\dagger y\|_2 - \alpha^2 - y^T A^\dagger b, \text{ s.t. } \|Ax + b\|_2 > \alpha \text{ and } \|(A^T)^\dagger y\|_2 \leq 2\alpha$$

Case 2.2: When $\|(A^T)^\dagger y\|_2 > 2\alpha$, consider the dual norm of it:

$$\|(A^T)^\dagger y\|_2 = \sup_z \frac{((A^T)^\dagger y)^T z}{\|z\|_2}$$

Let z^* be the argument that gives supremum in this case:

$$\begin{aligned} \|(A^T)^\dagger y\|_2 &> 2\alpha \\ \frac{((A^T)^\dagger y)^T z^*}{\|z^*\|_2} &> 2\alpha \\ \implies ((A^T)^\dagger y)^T z^* - 2\alpha\|z^*\|_2 &> 0 \end{aligned}$$

Let $Ax + b = tz^*$, $t \in \mathbb{R}_\infty$, $x = A^\dagger(tz^* - b)$

$$\begin{aligned} &\lim_{t \rightarrow \infty} (y^T x - \alpha(2\|Ax + b\|_2 - \alpha)) \\ &= \lim_{t \rightarrow \infty} (y^T x - 2\alpha((tz^*)^T(tz^*)^{1/2} + \alpha^2)) \rightarrow \infty \end{aligned}$$

Summarizing the cases:

$$f^*(y) = \begin{cases} \max(\frac{1}{4}y^T(A^T A)^\dagger y - y^T A^\dagger b, \alpha\|(A^T)^\dagger y\|_2 - \alpha^2 - y^T A^\dagger b), & \|(A^T)^\dagger y\|_2 \leq 2\alpha, \\ \infty, & \|(A^T)^\dagger y\|_2 > 2\alpha. \end{cases}$$

Note that $\frac{1}{4}y^T(A^T A)^\dagger y - y^T A^\dagger b$ is larger than $\alpha\|(A^T)^\dagger y\|_2 - \alpha^2 - y^T A^\dagger b$, finally we have

$$f^*(y) = \begin{cases} \frac{1}{4}y^T(A^T A)^\dagger y - y^T A^\dagger b, & \|(A^T)^\dagger y\|_2 \leq 2\alpha, \\ \infty, & \|(A^T)^\dagger y\|_2 > 2\alpha. \end{cases}$$