EE 248 HW#1 - Convexity and Gradient Methods

Due date: Sunday October 17, 2021 midnight over iLearn.

- You only need to submit solutions for 2 out of the 3 problems.
- If you submit all 3 solutions you get 10% bonus. If your effort is not sufficient, you may not get it (at TA's discretion).
- Only 1 problem of TA's choice will be graded (out of the ones you submitted).
- You are free to discuss with others but you are not allowed to look at other's HW solutions. It has to be 100% your writing, your work.
- Please LaTeX the solutions.

Problem 1

Part a. Let $(y_i, \boldsymbol{x}_i)_{i=1}^n$ be a dataset with $y \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^d$. Fix $\lambda \geq 0$ and consider the ridge regression problem

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2.$$

- Derive $\nabla\mathcal{L}(\boldsymbol{\beta})$ and $\nabla^2\mathcal{L}(\boldsymbol{\beta})$. Make sure to use matrix/vector notation where $\boldsymbol{X} = [\boldsymbol{x}_1 \ \dots \ \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} = [y_1 \ \dots \ y_n] \in \mathbb{R}^n$.
- Determine the precise conditions under which $\mathcal{L}(\boldsymbol{\beta})$ is a strongly-convex function.

Part b. Now consider the loss function

$$\bar{\mathcal{L}}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^n (y_i - |\boldsymbol{x}_i^\top \boldsymbol{\beta}|)^2.$$

- Given a choice of $(\boldsymbol{x}_i)_{i=1}^n$, construct $(y_i)_{i=1}^n$ such that the loss function $\bar{\mathcal{L}}(\boldsymbol{\beta})$ is convex. What is the most general set (to which $(y_i)_{i=1}^n$ belongs) that guarantees convexity?
- Given a choice of $(\boldsymbol{x}_i)_{i=1}^n$, construct $(y_i)_{i=1}^n$ such that $\bar{\mathcal{L}}(\boldsymbol{\beta})$ is nonconvex.

Solution a. $\nabla\mathcal{L}(\boldsymbol{\beta}) = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}) + \lambda\boldsymbol{\beta}$. $\nabla^2\mathcal{L}(\boldsymbol{\beta}) = \boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}$.

The Problem is strongly convex iff $\nabla^2\mathcal{L}(\boldsymbol{\beta}) \succ 0$ i.e. Hessian is positive definite. For this to happen, we either need

- $\lambda > 0$: positive ridge regression or
- $\boldsymbol{X}^\top\boldsymbol{X} \succ 0$: $\boldsymbol{X}$ is rank d. This is same as $n \geq d$ and $\boldsymbol{X}$ is full rank.

Solution b: Q1. Expanding the function we find $(y_i - |\boldsymbol{x}_i^\top \boldsymbol{\beta}|)^2 = y_i^2 - 2y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}| + (\boldsymbol{x}_i^\top \boldsymbol{\beta})^2$. Observe that y_i^2 is a constant, $(\boldsymbol{x}_i^\top \boldsymbol{\beta})^2$ is a convex function and $|\boldsymbol{x}_i^\top \boldsymbol{\beta}|$ is a convex function as well. However, $-2y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}|$ can be convex or concave depending on y_i.

Main observation is that if we have $y_i \le 0$ for all $1 \le i \le n$, then the problem is convex because each $-2y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}|$ term becomes convex.

Solution b: Q2. If all of $(\boldsymbol{x}_i)_{i=1}^n$ are zero problem is convex. Suppose one of the examples $(\boldsymbol{x}_i)_{i=1}^n$ are nonzero. Then, we can choose some $\boldsymbol{\beta}$ so that $\boldsymbol{X}\boldsymbol{\beta} \ne 0$ and set $\boldsymbol{y} = |\boldsymbol{X}\boldsymbol{\beta}|$. This $\boldsymbol{y}$ choice leads to nonconvexity. The reason is that $\bar{\mathcal{L}}(\boldsymbol{\beta}) = \bar{\mathcal{L}}(-\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - |\boldsymbol{X}\beta|\|_{\ell_2}^2 = 0$ so $\pm\boldsymbol{\beta}$ are global minima. However $\bar{\mathcal{L}}(0) = \frac{1}{2}\|\boldsymbol{y}\|_{\ell_2}^2 > 0$ which contradicts with the convexity since 0 is linear combination of $\pm\boldsymbol{\beta}$.

Solution b: Q1: Most general set. Here, the cleanest answer is all choices of y_i that makes $f(\boldsymbol{\beta}) = -\sum_{i=1}^n y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}|$ convex.

- This automatically covers the case y_i are nonpositive.
- Conversely, if y_i's are all non-negative and there is at least one pair $(y, \boldsymbol{x})$ obeying $y > 0$ and $\boldsymbol{x} \ne 0$, then $f(\boldsymbol{\beta})$ is nonconvex because

$$f(\boldsymbol{x}) = f(-\boldsymbol{x}) \le y|\boldsymbol{x}^\top \boldsymbol{x}| = y\|\boldsymbol{x}\|_{\ell_2}^2 < f(0) = 0.$$

I say the convexity of $f(\boldsymbol{\beta})$ is the cleanest answer because $\bar{\mathcal{L}}(\boldsymbol{\beta}) = 0.5\|\boldsymbol{y}\|_{\ell_2}^2 + f(\boldsymbol{\beta}) + f'(\boldsymbol{\beta})$ where $f'(\boldsymbol{\beta}) = 0.5\sum_{i=1}^n (\boldsymbol{x}_i^\top \boldsymbol{\beta})^2$. $0.5\|\boldsymbol{y}\|_{\ell_2}^2$ term is constant and $f'(\boldsymbol{\beta})$ has no importance. To see the latter observe that $\bar{\mathcal{L}}$ is convex if and only if f is convex.

- If $f(\boldsymbol{\beta})$ is convex, $f'(\boldsymbol{\beta})$ is also convex so $\bar{\mathcal{L}}(\boldsymbol{\beta})$ is convex.
- If $f(\boldsymbol{\beta})$ is not convex, there is $\boldsymbol{\beta}, \boldsymbol{\beta}'$ such that $\alpha f(\boldsymbol{\beta}) + (1-\alpha)f(\boldsymbol{\beta}') < f(\boldsymbol{\beta}_\alpha)$ where $\boldsymbol{\beta}_\alpha = \alpha\boldsymbol{\beta} + (1-\alpha)\boldsymbol{\beta}'$. Let ε be a small scalar of our choice. Observe that

  $$\bar{\mathcal{L}}(\varepsilon\boldsymbol{\beta}) = 0.5\|\boldsymbol{y}\|_{\ell_2}^2 + \varepsilon f(\boldsymbol{\beta}) + \varepsilon^2 f'(\boldsymbol{\beta}).$$

  Thus, we find

  $$(1/\varepsilon)[\alpha\bar{\mathcal{L}}(\varepsilon\boldsymbol{\beta}) + (1-\alpha)\bar{\mathcal{L}}(\varepsilon\boldsymbol{\beta}') - \bar{\mathcal{L}}(\varepsilon\boldsymbol{\beta}')] = \underbrace{[\alpha f(\boldsymbol{\beta}) + (1-\alpha)f(\boldsymbol{\beta}') - f(\boldsymbol{\beta}_\alpha)]}_{<0} + \underbrace{\varepsilon[\alpha f'(\boldsymbol{\beta}) + (1-\alpha)f'(\boldsymbol{\beta}') - f'(\boldsymbol{\beta}_\alpha)]}_{\to 0 \quad \text{as} \quad \varepsilon \to 0}$$

  Thus, letting $\varepsilon \to 0$, we find that $\bar{\mathcal{L}}(\boldsymbol{\beta})$ is not convex.

Problem 2

Part a. Recall that a convex function f may not be differentiable but it admits subgradients. The set of all subgradients of a function f is called the subdifferential and is denoted by $\partial f(\boldsymbol{a})$. Consider the ℓ_1 norm of a vector $\boldsymbol{a} \in \mathbb{R}^d$ which is defined as $\|\boldsymbol{a}\|_1 = \sum_{i=1}^d |a_i|$. Determine the set $\partial\|\boldsymbol{a}\|_1$.

Part b. Suppose you wish to pick a subgradient $\boldsymbol{g} \in \partial\|\boldsymbol{a}\|_1$ such that ℓ_1 norm reduces as much as possible when you take an ε small step in the $-\boldsymbol{g}$ direction. Determine this optimal $\boldsymbol{g}$ at a given $\boldsymbol{a}$.

Part c. Consider the lasso problem which is ℓ_1-penalized least-squares regression

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

Suppose you wish to solve $\arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta})$ using an iterative algorithm with constant learning rate η. Based on Parts a. & b., derive how to take a step from the current iterate $\boldsymbol{\beta}$ to obtain the next iterate $\hat{\boldsymbol{\beta}}$.

Hint: Part of the loss is not differentiable so you need to choose the best sub-gradient.

Part d. People use ℓ_1 norm for a reason right? It turns out, it is because ℓ_1 norm is the smallest p for which ℓ_p quasi-norm is convex. Let us define the ℓ_p quasi-norm of a vector $\|\boldsymbol{a}\|_p = (\sum_{i=1}^d |a_i|^p)^{1/p}$ for all $p > 0$. Prove that ℓ_p quasi-norm is a nonconvex function for all $0 < p < 1$.

Solution a. Here a key observation is that ℓ_1 norm is decomposable into individual coordinates since $\|\boldsymbol{x}\|_1 = \sum_{i=1}^d |\boldsymbol{x}_i|$. Thus subgradient is the concatenation of the subgradients of the individual absolute value functions because $\partial\|\boldsymbol{x}\|_1/\partial\boldsymbol{x}_i = \partial|\boldsymbol{x}_i|/\partial\boldsymbol{x}_i$. Let $\mathrm{sgn}(\boldsymbol{a})$ be the sign vector which is 1 if $\boldsymbol{a}_i > 0$, -1 if $\boldsymbol{a}_i < 0$ and 0 else. Subdifferential is given by

$$\partial\|\boldsymbol{a}\|_1 = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_\infty \le 1,\ \boldsymbol{x}_i = \mathrm{sgn}(\boldsymbol{a}_i),\ \text{whenever } \boldsymbol{a}_i \ne 0\}.$$

In words, subgradient is equal to $\mathrm{sgn}(\boldsymbol{a})$ over the nonzero entries and the remaining entries are arbitrary with absolute value bounded by 1.

Solution b. $\boldsymbol{g} = \mathrm{sgn}(\boldsymbol{a})$. If $\boldsymbol{g}$ had a nonzero entry in a coordinate $\boldsymbol{a}_i = 0$, then we would have $\|\boldsymbol{a} - \varepsilon\boldsymbol{g}\|_1 > \|\boldsymbol{a} - \varepsilon\mathrm{sgn}(\boldsymbol{a})\|_1$ because $|\boldsymbol{g}_i + \boldsymbol{a}_i| = |\boldsymbol{g}_i| > 0$ would increase the ℓ_1 norm. Thus, optimal $\boldsymbol{g}$ has to be $\mathrm{sgn}(\boldsymbol{a})$.

Solution c. Any subgradient can be written as $g_{\mathcal{L}}(\boldsymbol{\beta}) = 0.5\nabla\|\boldsymbol{y} - \boldsymbol{X\beta}\|_{\ell_2}^2 + \lambda g_{\|\cdot\|_1}(\boldsymbol{\beta})$. Following part (b), we choose $g_{\mathcal{L}}(\boldsymbol{\beta}) = \boldsymbol{X}^\top(\boldsymbol{X\beta} - \boldsymbol{y}) + \lambda\mathrm{sgn}(\boldsymbol{\beta})$.

Solution d. If $d = 1$, all norms are equal to absolute value and convex. So assume $d \ge 2$. Choose $\boldsymbol{x} = [1\ 0]^\top$ and $\boldsymbol{y} = [0\ 1]^\top$ and $\boldsymbol{z} = [0.5\ 0.5]^\top$. For $p < 1$, convexity is violated via

$$\frac{\|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p}{2} = 1 < 2^{1/p-1} = 2^{1/p}0.5 = \|\boldsymbol{z}\|_p.$$

Problem 3

Part a. Let $\mathcal{C} \in \mathbb{R}^d$ be a closed convex set. Let $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ two points. Define the projection operator to be the unique point $\Pi_{\mathcal{C}}(\boldsymbol{a}) = \arg\min_{\boldsymbol{v}\in\mathcal{C}} \|\boldsymbol{a} - \boldsymbol{v}\|_2$. Prove that projection is contractive, that is,

$$\|\Pi_{\mathcal{C}}(\boldsymbol{a}) - \Pi_{\mathcal{C}}(\boldsymbol{b})\|_2 \le \|\boldsymbol{a} - \boldsymbol{b}\|_2. \tag{CONTRACTION}$$

Hint: Use the fact that projection satisfies $\langle \boldsymbol{v} - \Pi_{\mathcal{C}}(\boldsymbol{a}), \boldsymbol{a} - \Pi_{\mathcal{C}}(\boldsymbol{a})\rangle \le 0$ for any $\boldsymbol{v} \in \mathcal{C}$.

Significance: In the special case of $\boldsymbol{b} \in \mathcal{C}$, this result implies: *"Projecting $\boldsymbol{a}$ onto $\mathcal{C}$ is guaranteed to return a closer point to $\boldsymbol{b}$"* and forms the basis of iterative algos that use gradient-descent followed by projection.

Part b. Set $\mathcal{C}$ to be the set of s-sparse signals, that is, $\boldsymbol{v} \in \mathcal{C}$ iff $\boldsymbol{v}$ has at most s-nonzero entries for some $s \le d$. Construct an example where (CONTRACTION) is violated for $\mathcal{C}$.

Part c. Consider the ℓ_∞ norm-ball defined as $\mathcal{C} = \{\boldsymbol{v} \in \mathbb{R}^d \mid |v_i| \le 1 \quad \text{for all} \quad 1 \le i \le d\}$. Derive (the analytical expression for) the projection onto this ball i.e. $\Pi_{\mathcal{C}}(\boldsymbol{x})$.

Part d. In Problem 2a. you calculated the subdifferential of the ℓ_1 norm. Now, let us set $\mathcal{C}$ to be this subdifferential i.e. $\mathcal{C} = \partial\|\boldsymbol{a}\|_1$. Derive the projection onto this subdifferential i.e. $\Pi_{\mathcal{C}}(\boldsymbol{x})$. Comment on any similarities to the norm-ball projection in Problem 3c. (**Hint:** What happens when $\boldsymbol{a} = 0$?)

Solution a (easiest to understand is by drawing on a paper). Let $\boldsymbol{a}' = \Pi_{\mathcal{C}}(\boldsymbol{a}), \boldsymbol{b}' = \Pi_{\mathcal{C}}(\boldsymbol{b})$. Let P denote all points between $\boldsymbol{a}', \boldsymbol{b}'$. Note that P is in the convex set $\mathcal{C}$ and $\boldsymbol{b}' = \Pi_P(\boldsymbol{b})$ is the projection of $\boldsymbol{b}$ onto $P \subset \mathcal{C}$. Consider the two dimensional plane $\mathcal{S}$ induced by the vectors $\boldsymbol{a} - \boldsymbol{a}'$ and $\boldsymbol{a} - \boldsymbol{b}$. Obtain $\boldsymbol{b}_2 = \Pi_{\mathcal{S}}(\boldsymbol{b})$ which projects $\boldsymbol{b}$ onto the plane (now you can draw $\boldsymbol{a}, \boldsymbol{b}_2, \boldsymbol{a}', \boldsymbol{b}'$ on a paper). Since $\boldsymbol{b}_2$ is a subspace projection and $P \subset \mathcal{S}$, we still have that $\boldsymbol{b}'$ is the closest point i.e. $\boldsymbol{b}' = \Pi_P(\boldsymbol{b}_2) = \Pi_P(\boldsymbol{b})$.

Let $\mathcal{S}'$ be the 1 dimensional subspace P lies on. Using the hint observe that

$$\|\boldsymbol{a} - \boldsymbol{b}\|_{\ell_2} \ge \|\boldsymbol{a} - \boldsymbol{b}_2\|_{\ell_2} \ge \|\Pi_{\mathcal{S}'}(\boldsymbol{a}) - \Pi_{\mathcal{S}'}(\boldsymbol{b}_2)\|_{\ell_2} \ge \|\boldsymbol{a}' - \boldsymbol{b}'\|_{\ell_2}$$

Solution b. Let us set $d = 2$ and $s = 1$. Consider $\boldsymbol{a} = [1\ 0.9]$ and $\boldsymbol{b} = [0.9\ 1]$. We have that $\|\boldsymbol{a} - \boldsymbol{b}\|_{\ell_2}^2 = 0.02$. On the other hand $\boldsymbol{a}' = \Pi_{\mathcal{C}}(\boldsymbol{a}) = [1\ 0]$ and $\boldsymbol{b}' = \Pi_{\mathcal{C}}(\boldsymbol{b}) = [0\ 1]$ thus $\|\boldsymbol{a}' - \boldsymbol{b}'\|_{\ell_2}^2 = 2 > \|\boldsymbol{a} - \boldsymbol{b}\|_{\ell_2}^2$. Violation for general $s < d$ can be done with similar construction (except for $s = d$ where sparse set $\mathcal{C}$ is the whole $\mathbb{R}^d$ and convex).

Solution c. The key observation again is that this projection is decomposable i.e. it can be done entrywise. The ℓ_∞ projection is given by

$$\Pi_\infty(\boldsymbol{x}_i) = \min(|\boldsymbol{x}_i|, 1) \cdot \mathrm{sgn}(\boldsymbol{x}_i) \quad \text{for} \quad 1 \le i \le d.$$

Solution d. At the special case of $\boldsymbol{a} = 0$, ℓ_1 subdifferential is exactly the ℓ_∞ norm ball. This connection arises from the fact that ℓ_∞ and ℓ_1 are so-called *dual norms* of each other. Let $\mathcal{S}$ be the set of nonzero coordinates of $\boldsymbol{a}$. Noticing entrywise decomposability again, the projection is same as Π_∞ over $\mathcal{S}^c$ and it is simply $\mathrm{sgn}(\boldsymbol{a}_i)$ over $\mathcal{S}$ as follows

$$\Pi_{\mathcal{C}}(\boldsymbol{x}_i) = \begin{cases} \min(|\boldsymbol{x}_i|, 1) \cdot \mathrm{sgn}(\boldsymbol{x}_i) & i \in \mathcal{S}^c \\ \mathrm{sgn}(\boldsymbol{a}_i) & i \in \mathcal{S} \end{cases}$$

Vector notation is $\Pi_{\mathcal{C}}(\boldsymbol{x}) = \mathrm{sgn}(\boldsymbol{a}) + \Pi_\infty(\boldsymbol{x}_{\mathcal{S}^c})$ where $\boldsymbol{x}_{\mathcal{S}^c}$ returns the entries over $\mathcal{S}^c$ and zeros over $\mathcal{S}$.