Homework 5

This homework is due at 11 PM on October 4th, 2023.

Submission Format: Your homework submission should consist of a single PDF file that contains all of your answers (any handwritten answers should be scanned).

NOTE: While the problems on this homework can be completed in any order, they are likely to be most cohesive in the order in which they appear.

1. **Practice with Convexity**

   (a) Let $f_1, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ be convex functions. **Prove that**

$$\textbf{the set} \qquad S \doteq \{\vec{x} \in \mathbb{R}^n \mid f_i(\vec{x}) \leq 0 \quad \forall i = 1, \ldots, k\} \qquad \textbf{is convex.} \tag{1}$$

   (b) Let $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$. **Prove that** $f : \mathbb{R}^n \to \mathbb{R}$ **given by** $f(\vec{x}) \doteq \max_{1 \leq i \leq n} |x_i|$ **is convex.**

2. Ridge Regression for Bounded Output Perturbation

We will first solve the ridge regression problem in the case where our output measurements $\vec{y}$ are perturbed and we have some bounds on this perturbation, as well as some specific knowledge about data matrix A.

Let square matrix $A \in \mathbb{R}^{n \times n}$ have the singular value decomposition $A = U\Sigma V^\top$, and suppose its singular values satisfy $\sigma_1(A) \geq \cdots \geq \sigma_n(A) > 0$.

(a) Is A invertible? If so, write the singular value decomposition of A^{-1}.

(b) Consider the linear equation $A\vec{x} = \vec{y}_p$, where $\vec{y}_p \in \mathbb{R}^n$ is a perturbed measurement satisfying

$$\|\vec{y}_p - \vec{y}\|_2 \leq r \tag{2}$$

for some vector $\vec{y} \in \mathbb{R}^n$ and $r > 0$. Let $\vec{x}^\star(\vec{y})$ denote the solution of $A\vec{x} = \vec{y}$.

Show that

$$\max_{\vec{y}_p : \|\vec{y}_p - \vec{y}\|_2 \leq r} \|\vec{x}^\star(\vec{y}_p) - \vec{x}^\star(\vec{y})\|_2 = \frac{r}{\sigma_{\min}\{A\}} \tag{3}$$

(c) What happens if the smallest singular value of A is very close to zero? Why is this problematic for finding our solution vector $\vec{x}^\star$?

(d) Now assume that we find optimal value $\vec{x}^\star$ via ridge regression, i.e., we compute

$$\vec{x}^\star_\lambda(\vec{y}_p) = \operatorname*{argmin}_{\vec{x}} \left\{ \|A\vec{x} - \vec{y}_p\|_2^2 + \lambda \|\vec{x}\|_2^2 \right\} \tag{4}$$

for some chosen value $\lambda \geq 0$. Compute $\vec{x}^\star_\lambda(\vec{y}_p)$, our optimal solution vector (now parameterized by λ), by solving this optimization problem. You may use the solution from class for this part.

(e) Show that for all $\lambda > 0$,

$$\max_{\vec{y}_p : \|\vec{y}_p - \vec{y}\|_2 \leq r} \|\vec{x}^\star_\lambda(\vec{y}_p) - \vec{x}^\star_\lambda(\vec{y})\|_2 \leq \frac{r}{2\sqrt{\lambda}}. \tag{5}$$

How does the value of λ affect the sensitivity of your solution $\vec{x}^\star_\lambda(\vec{y})$ to the perturbation level in $\vec{y}$? *HINT: For every $\lambda > 0$, we have*

$$\max_{\sigma > 0} \frac{\sigma}{\sigma^2 + \lambda} = \frac{1}{2\sqrt{\lambda}}. \tag{6}$$

You need not show this; this optimization can be solved by setting the derivative of the objective function to 0 and solving for σ.

3. (OPTIONAL) Ridge Regression for Data Matrix Noise

Next, we will solve the ridge regression problem in the case where our data matrix A is noisy and we know some properties of this noise.

Consider the standard least-squares problem

$$\min_{\vec{x}} \|A\vec{x} - \vec{y}\|_2^2, \tag{7}$$

in which the data matrix $A \in \mathbb{R}^{m\times n}$ is noisy. We model this noise by assuming that each row $\vec{a}_i^\top$, where $\vec{a}_i^\top \in \mathbb{R}^n$, has the form $\vec{a}_i = \vec{p}_i + \vec{u}_i$, where the independent noise vectors $\vec{u}_i \in \mathbb{R}^n$ have zero mean (i.e., $\mathbb{E}_U[\vec{u}_i] = \vec{0}$) and covariance matrices $\sigma^2 I_n$ (i.e., $\mathbb{E}_U\left[\vec{u}_i\vec{u}_i^\top\right] = \sigma^2 I_n$ and $\mathbb{E}_U\left[\vec{u}_i\vec{u}_j^\top\right] = O_n$ whenever $i \neq j$), with $\sigma \in \mathbb{R}$ a measure of the size of the noise, and the vector $\vec{p}_i \in \mathbb{R}^n$ is the "true" data point. Therefore, now the matrix A is a function of the random matrix U whose rows are $\vec{u}_1^\top, \ldots, \vec{u}_m^\top$; to make this dependence clear we use the notation $A = A_U$. We will use P to denote the "true" data matrix with rows $\vec{p}_i^\top$, $i = 1, \ldots, m$. To account for this noise, we replace the standard least squares formulation above with

$$\min_{\vec{x}} \ \mathbb{E}_U\left[\|A_U\vec{x} - \vec{y}\|_2^2\right], \tag{8}$$

where $\mathbb{E}_U$ denotes the expected value with respect to the random variable U. Show that this problem can be written as

$$\min_{\vec{x}} \left\{\|P\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_2^2\right\}, \tag{9}$$

where $\lambda \geq 0$ is some regularization parameter, which you will determine. In other words, show that regularized least-squares can be interpreted as a way to take into account uncertainties in the matrix A, in the expected value sense.

HINT: Compute the expected value of $((\vec{p}_i + \vec{u}_i)^\top\vec{x} - y_i)^2$, for a specific row index i. What is the sum of squares of these expected values over all i?

4. Linear Regression with Weights

In this problem, we discuss multiple interpretations of weighted linear regression.

Let $A \in \mathbb{R}^{m\times n}$ be a data matrix whose data points are the m rows $\vec{a}_1^\top, \ldots, \vec{a}_m^\top \in \mathbb{R}^{1\times n}$. Suppose $m \geq n$ and A has full column rank. Let $\vec{y} \in \mathbb{R}^m$ be a vector of outputs, each corresponding to a data point. Let $\vec{w} \in \mathbb{R}^m_{++}$ be a vector of *positive* real numbers, also called weights, each corresponding to a data point—output pair. We are interested in the following least-squares type optimization problem:

$$\min_{\vec{x}\in\mathbb{R}^n} \sum_{i=1}^{m} w_i(\vec{a}_i^\top \vec{x} - y_i)^2. \tag{10}$$

In general, assigning a high weight w_i means that we want our learned linear predictor $\vec{a}_i^\top \vec{x}$ to achieve a close value to y_i; that is, we believe this data point is significant or important to get right.

(a) Show that the problem in Equation (10) is equivalent to the problem:

$$\min_{\vec{x}\in\mathbb{R}^n} \left\| W^{1/2}(A\vec{x} - \vec{y}) \right\|_2^2 \tag{11}$$

where $W \doteq \operatorname{diag}(\vec{w}) \in \mathbb{R}^{m\times m}$ is a diagonal matrix whose diagonal entries are the entries of $\vec{w}$.

(b) Using Equation (11), compute the gradient (with respect to $\vec{x}$) of the objective function

$$f(\vec{x}) \doteq \left\| W^{1/2}(A\vec{x} - \vec{y}) \right\|_2^2. \tag{12}$$

(c) Show that the optimal solution to Equation (11) is given by

$$\vec{x}^\star_{\mathrm{WLR}} \doteq (A^\top W A)^{-1} A^\top W \vec{y}. \tag{13}$$

HINT: There are multiple ways to do this problem; one uses the gradient that you just computed, and another finds the least squares solution of a particular linear system.

HINT: If using the gradient method, you may assume that f is minimized at any $\vec{x}^\star$ such that $\nabla_{\vec{x}} f(\vec{x}^\star) = \vec{0}$; this is because f is convex, as we will see a little later in the course.

(d) Now we will look at this problem from a probabilistic interpretation. Suppose our output value $\vec{y}$ is noisy, and in particular there is some $\vec{x}_0$ such that for every i we have $y_i = \vec{a}_i^\top \vec{x}_0 + u_i$, where u_i is a random variable. Here we assume the u_i are independent but *not* identically distributed. In particular, we assume that for each i we have that u_i is distributed according to a Gaussian $\mathcal{N}(0, \sigma_i^2)$ where $\sigma_i > 0$ is a known noise parameter for each data point i.

To recover $\vec{x}_0$ given data A and $\vec{y}$, as well as the σ_i^2, we want to compute the maximum likelihood estimator (MLE). Show that the maximum likelihood problem

$$\operatorname*{argmax}_{\vec{x}\in\mathbb{R}^n} p(\vec{y} \mid A, \vec{x}) \tag{14}$$

is equivalent to the weighted linear regression problem:

$$\operatorname*{argmin}_{\vec{x}\in\mathbb{R}^n} \sum_{i=1}^{m} w_i(\vec{a}_i^\top \vec{x} - y_i)^2. \tag{15}$$

for some choice of $\vec{w}$. What choice of $\vec{w}$ makes them equivalent?

(e) In addition to assigning weights to data points to place higher importance on getting those points right, sometimes we want to make sure that our learned $\vec{x}$ is close to some value $\vec{z} \in \mathbb{R}^n$. This could be true, for example, if we had prior information that said $\vec{x}$ is close to $\vec{z}$.

In particular, we want to make sure that the quantity

$$\sum_{i=1}^{n} s_i(x_i - z_i)^2 \tag{16}$$

is small, where $\vec{s} \in \mathbb{R}^n_{++}$ is a vector of *positive* weights. We may add this term to the weighted least squares objective function, with a regularization parameter $\lambda \geq 0$, to create a modified objective function

$$g(\vec{x}) \doteq \left\| W^{1/2}(A\vec{x} - \vec{y}) \right\|_2^2 + \lambda \left\| S^{1/2}(\vec{x} - \vec{z}) \right\|_2^2 \tag{17}$$

where $S \doteq \operatorname{diag}(\vec{s}) \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose entries are the entries of $\vec{s}$. This formulation is called *Tikhonov regression*.

Compute the gradient (with respect to $\vec{x}$) of g, and show that the optimal solution is

$$\vec{x}^\star_{\mathrm{TR}} \doteq (A^\top W A + \lambda S)^{-1}(A^\top W \vec{y} + \lambda S \vec{z}). \tag{18}$$

HINT: You may assume that g is minimized at any $\vec{x}^\star$ such that $\nabla_{\vec{x}} g(\vec{x}^\star) = \vec{0}$; this is because g is convex, as we will see a little later in the course.

5. Visualizing Rank 1 Matrices

In this problem, we explore the effect of rank constraints on the convexity of matrix sets.

First, consider the set of all 2×2 matrices with diagonal elements $(1, 2)$, which we can write explicitly as

$$\mathcal{S} = \left\{ \begin{bmatrix} 1 & x \\ y & 2 \end{bmatrix} \;\middle|\; x, y \in \mathbb{R} \right\}. \tag{19}$$

(a) Is set $\mathcal{S}$ convex? If so, provide a proof, and if not, provide a counterexample.

(b) Suppose we now wish to define $\mathcal{S}_1 \subset \mathcal{S}$, the set of all rank-1 matrices in $\mathcal{S}$. Write out conditions on x and y (i.e. equation constraints that x and y must satisfy) to define $\mathcal{S}_1$ explicitly.

(c) Is set $\mathcal{S}_1$ convex? If so, provide a proof, and if not, provide a counterexample. Plot the x–y curve described by constraints you found in the earlier part and observe its shape. *HINT: Any linear function applied to a convex set generates another convex set, and the function that maps set $\mathcal{S}_1$ to variables (x, y) is linear.* [1]

(d) In this class, we will sometimes pose optimization problems in which we optimize over sets of matrices. Since low-dimensional models are often easier to interpret, it would be nice to impose rank constraints on these solution matrices. Suppose we wish to solve the optimization problem

$$\min_{A \in \mathcal{S}_1} \|A\|_F^2 \tag{20}$$

which is equivalent to

$$\min_{A \in \mathcal{S}} \quad \|A\|_F^2 \tag{21}$$

$$\text{s.t.} \quad \operatorname{rank}(A) = 1. \tag{22}$$

Is this optimization problem convex? (That is, are the objective function $\|A\|_F^2$ and feasible set $\mathcal{S}_1$ convex?)

[1] We can show this directly from the definition of linearity: define function $f : \mathcal{S}_1 \to \mathbb{R}^2$ that maps each set element s to its corresponding off-diagonal values (x, y). Then for any two elements $s_1, s_2 \in \mathcal{S}_1$ and $\alpha_1, \alpha_2 \in \mathbb{R}$, we have $f(\alpha_1 s_1 + \alpha_2 s_2) = \alpha_1 f(s_1) + \alpha_2 f(s_2)$.

6. Quadratics and Least Squares

In this question, we will see that every least squares problem can be considered as the minimization of a quadratic cost function; whereas not every quadratic minimization problem corresponds to a least-squares problem. To begin with, consider the quadratic function, $f : \mathbb{R}^2 \to \mathbb{R}$ given by:

$$f(\vec{w}) = \vec{w}^\top A\vec{w} - 2\vec{b}^\top \vec{w} + c \tag{23}$$

where $A \in \mathbb{S}^2_+$ (set of symmetric positive semidefinite matrices in $\mathbb{R}^{2\times 2}$), $\vec{b} \in \mathbb{R}^2$ and $c \in \mathbb{R}$.

(a) Assume $c = 0$, and assume that setting $\nabla f(\vec{w}) = 0$ allows us to find the unique minimizer. Give a concrete example of a matrix $A \succ 0$ and a vector $\vec{b}$ such that the point $\vec{w}^\star = \begin{bmatrix} -1 & 1 \end{bmatrix}^\top$ is the unique minimizer of the quadratic function $f(\vec{w})$.

(b) Assume $c = 0$. Give a concrete example of a matrix $A \succeq 0$, and a vector $\vec{b}$ such that the quadratic function $f(\vec{w})$ has infinitely many minimizers and all of them lie on the line $w_1 + w_2 = 0$.

HINT: Take the gradient of the expression and set it to zero. What needs to be true for there to be infinitely many solutions to the equation?

(c) Assume $c = 0$. Let $\vec{w} = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$. Give a concrete example of a **non-zero** matrix $A \succeq 0$ and a vector $\vec{b}$ such that the quadratic function $f(\alpha\vec{w})$ tends to $-\infty$ as $\alpha \to \infty$. *HINT: Use the eigenvalue decomposition to write $A = \sigma_1\vec{u}_1\vec{u}_1^\top + \sigma_2\vec{u}_2\vec{u}_2^\top$ and express $\vec{w}$ in the basis formed by $\vec{u}_1, \vec{u}_2$.*

(d) Say that we have the data set $\{(\vec{x}_i, y_i)\}_{i=1,\dots,n}$ of data points $\vec{x}_i \in \mathbb{R}^d$ and values $y_i \in \mathbb{R}$. Define $X = \begin{bmatrix} \vec{x}_1 & \dots & \vec{x}_n \end{bmatrix}^\top$ and $\vec{y} = \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix}^\top$. In terms of X and $\vec{y}$, find a matrix A, a vector $\vec{b} \in \mathbb{R}^d$ and a scalar c, so that we can express the sum of the square losses $\displaystyle\sum_{i=1}^{n}(\vec{w}^\top\vec{x}_i - y_i)^2$ as the quadratic function $f(\vec{w}) = \vec{w}^\top A\vec{w} - 2\vec{b}^\top\vec{w} + c$.

(e) Here are three statements with regards to the minimization of a quadratic loss function:

i. It can have a unique minimizer.

ii. It can have infinitely many minimizers.

iii. It can be unbounded from below, i.e. there is some direction, $\vec{w}$ so that $f(\alpha\vec{w})$ goes to $-\infty$ as $\alpha \to \infty$.

All three statements apply to general minimization of a quadratic cost function. Parts (a), (b) and (c) give concrete examples of quadratic cost functions where (i), (ii) and (iii) apply respectively. However, notice that statement (iii) cannot apply to the least squares problem as the objective is always positive. The least-squares problem can have infinitely many minimizers though. How? Consider the gradient of the least squares problem in part (d) at an optimal solution $\vec{w}^\star$:

$$\nabla f(\vec{w}^\star) = 2X^\top X\vec{w}^\star - 2\vec{b} = 0. \tag{24}$$

Therefore, the least squares problem only has multiple solutions if $X^\top X$ is not full rank. This means that $\operatorname{rank}(X^\top X) = \operatorname{rank}(X) < d$. Finally, the rank of X is less than d when the data points $\{\vec{x}_i\}_{i=1}^n$ do not span $\mathbb{R}^d$. This can happen when the number of data points n is less than d or when $\{\vec{f}_i\}_{i=1}^d$ are linearly dependent where $\vec{f}_i$ are the columns of X, i.e., the features.

We will see soon that these cases correspond to the *convexity* of the function: if the function is strictly convex, then it has a unique minimizer; and if it is just convex, then it can have multiple minimizers; and in both cases, it can have no minimizers. We will see soon how to prove that the quadratic objective functions we discuss in this problem are convex, strictly convex, or even non-convex.

Indicate below that you have read and understood the discussion above.

7. Homework Process

With whom did you work on this homework? List the names and SIDs of your group members.

NOTE: If you didn't work with anyone, you can put "none" as your answer.

