Homework 9

This homework is due at 11 PM on November 1, 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), as well as a printout of your completed Jupyter notebook(s).

1. Diagonally Dominant Matrices

(a) Given a symmetric matrix $A \in \mathbb{R}^{n \times n}$ we say that A is diagonally dominant if

$$\forall i \in \{1, \ldots, n\} \text{ we have } A_{i,i} \geq \sum_{j \neq i} |A_{i,j}|$$

That is, the diagonal entry is greater than the sum of the absolute values of the off-diagonal entries of that row (equivalently column, as it's symmetric). Prove that A is positive semi-definite.

(b) Show that $f(\vec{x}) = \log\left(\sum_{i=1}^{n} e^{x_i}\right)$ is a convex function with domain $\mathbb{R}^n$. You might find the previous part helpful.

2. Convergence of Gradient Descent for Ridge Regression

Let $A \in \mathbb{R}^{m \times n}$, $\vec{y} \in \mathbb{R}^m$, and $\lambda > 0$. Consider a slight variation of the ridge regression problem where the least squares loss is normalized by the number of data points:

$$\min_{\vec{x} \in \mathbb{R}^n} f_\lambda(\vec{x}) \qquad \text{where} \qquad f_\lambda(\vec{x}) \doteq \frac{1}{2}\left\{\frac{1}{m}\|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_2^2\right\}. \tag{1}$$

In this problem, we will examine the behavior of gradient descent (GD) on this problem, and in particular the interplay between the step size $\eta > 0$ and regularization parameter $\lambda > 0$ in determining the convergence of gradient descent.

(a) Show that the unique solution to the problem in Equation (1) is

$$\vec{x}_\lambda^\star = \left(A^\top A + \lambda m I\right)^{-1} A^\top \vec{y}. \tag{2}$$

(b) Show that the GD update

$$\vec{x}_{t+1} = \vec{x}_t - \eta\left(\frac{1}{m}A^\top(A\vec{x}_t - \vec{y}) + \lambda \vec{x}_t\right) \tag{3}$$

can be rearranged into the form

$$\vec{x}_{t+1} - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)(\vec{x}_t - \vec{x}_\lambda^\star). \tag{4}$$

Use this to show that

$$\vec{x}_t - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star). \tag{5}$$

for every positive integer t.

(c) We now discuss the insight that the SVD can give us regarding the convergence of GD. Let $A = U\Sigma V^\top$ be a full SVD of A. Let $\vec{z}_t = V^\top \vec{x}_t$ and $\vec{z}_\lambda^\star = V^\top \vec{x}_\lambda^\star$. Show that

$$\vec{z}_t - \vec{z}_\lambda^\star = \left(I - \eta\left(\frac{\Sigma^\top \Sigma}{m} + \lambda I\right)\right)^t (\vec{z}_0 - \vec{z}_\lambda^\star), \tag{6}$$

and, moreover, show that for each $i \in \{1, \ldots, n\}$, we have

$$(\vec{z}_t)_i - (\vec{z}_\lambda^\star)_i = \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right)^t ((\vec{z}_0)_i - (\vec{z}_\lambda^\star)_i) \tag{7}$$

where $\sigma_i\{A\}$ is the i^{th} largest singular value of A. This shows that the *rate of convergence* of $\vec{z}_t$ to $\vec{z}_\lambda^\star$ along the i^{th} component is influenced by the interaction between $\sigma_i\{A\}$, λ, and η, but, critically, no other singular values. Thus, one considers the V basis to be the "natural" basis for thinking about GD for ridge regression.

(d) Show that $\lim_{t\to\infty} \vec{z}_t = \vec{z}_\lambda^\star$ for all initializations $\vec{x}_0 = V\vec{z}_0$ if and only if

$$\max_{i \in \{1,\ldots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1. \tag{8}$$

Use this to show that GD converges for all initializations $\vec{x}_0$ if and only if

$$\eta \in \left(0, \frac{2m}{\sigma_{\max}\{A\}^2 + \lambda m}\right) \tag{9}$$

where $\sigma_{\max}\{A\} = \sigma_1\{A\}$ is the largest singular value of A.

(e) The attached notebook, `gd_convergence.ipynb`, will examine the computational aspects of GD on ridge regression. Implement the GD and stochastic gradient descent (SGD) functions at the top of the notebook, which are marked with `TODO`s.

(f) Click through the notebook and run the sections $n = 1$, $n = 2$, and $n \gg 2$. Change the values of λ and η and re-run the cells a few times. Write down your observations about how the convergence of GD works under different values of λ and η.

(g) In the sections $n = 1$, $n = 2$, and $n \gg 2$, change the calls to GD to instead call SGD. Write down your observations about how the convergence of SGD works under different values of λ and η. Compare the behavior of GD and SGD.

(h) You might have noticed that if we think of convergence in the "last iterate" sense, i.e., $\lim_{T\to\infty} \vec{x}_T = \vec{x}^\star_\lambda$, then *SGD rarely converges*. This is because even if we reach the global optimum, the gradient estimate used by SGD is in general nonzero, and so the iterates end up bouncing around near the optimum. Another different, weaker, notion of convergence under which one might show that SGD actually does converge is convergence "in time average", i.e., $\lim_{T\to\infty} \overline{\vec{x}}_T = \vec{x}^\star_\lambda$ where $\overline{\vec{x}}_T \doteq \frac{1}{T}\sum_{t=1}^{T} \vec{x}_t$. Visualize this by adding the argument `time_avg=True` to each plotting function; the plot will now visualize the sequence of $\overline{\vec{x}}_t$. Re-run the notebook. Write down your observations, especially regarding the stability of SGD and convergence in the last-iterate sense versus the time-average sense.

(i) **(OPTIONAL)** One way we can derive an "optimal" step size $\eta^\star$ to minimize the largest rate of convergence:

$$\eta^\star \in \operatorname*{argmin}_{\eta\in\mathbb{R}} \max_{i\in\{1,\ldots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right|. \tag{10}$$

One important property of $\eta^\star$ is that it makes the rates of convergence $\left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right|$ associated with the largest and smallest singular values of A equal. Use this property to show that

$$\eta^\star = \frac{2m}{\sigma_{\max}\{A\}^2 + \sigma_{\min}\{A\}^2 + 2\lambda m} \tag{11}$$

where $\sigma_{\min}\{A\} = \sigma_n\{A\}$ is the n^{th} largest singular value of A.

NOTE: There are several useful notions of optimal step size in this context; this is just one of them.

3. Trust region

In optimization a trust region refers to the region where a certain model (usually quadratic) can be used to approximate the original objective function.

Consider the problem

$$p^* = \min_{\vec{x}} \ \vec{x}^\top Q\vec{x} + 2\vec{c}^\top \vec{x} \ : \ \|\vec{x}\|_2 = 1.$$

where $Q \in \mathbb{S}^n$ is symmetric (not necessarily positive semi-definite) and $\vec{c} \in \mathbb{R}^n$.

(a) Is the problem, as stated convex? What if Q is positive semi-definite?

(b) Show that the problem can be reduced to

$$p^* = \min_{\vec{y}} \ \sum_{i=1}^n \left(\lambda_i y_i^2 + 2d_i y_i\right) \ : \ \sum_{i=1}^n y_i^2 = 1,$$

for appropriate vectors $\vec{\lambda}, \vec{d} \in \mathbb{R}^n$, which you will determine as functions of the problem data.

(c) Show that the problem can be further reduced to the convex problem

$$p^* = \min_{\vec{z}} \ \sum_{i=1}^n \left(\lambda_i z_i - 2|d_i|\sqrt{z_i}\right) \ : \ \sum_{i=1}^n z_i = 1, \ \ z \geq 0.$$

4. Formulating Optimization problems

(a) **Linear Separability.** Let $(\vec{x}_i, y_i)$ be given data points with $\vec{x}_i \in \mathbb{R}^n$ and binary labels $y_i \in \{-1, 1\}$. We want to know if it is possible to find a hyperplane $\mathcal{L} = \{\vec{x} \in \mathbb{R}^n : \vec{h}^\top \vec{x} + b = 0\}$ that separates all the points with labels $y_i = -1$ from all the points with labels $y_i = 1$. In other words, can we find a vector $\vec{h} \in \mathbb{R}^n$ and a scalar $b \in \mathbb{R}$ such that $\vec{h}^\top \vec{x}_i + b \leq 0$ for all i satisfying $y_i = 1$, and $\vec{h}^\top \vec{x}_i + b > 0$ for all i satisfying $y_i = -1$? We want to cast this task as the following LP

$$p^\star = \min_{\vec{h}, b, z} \quad f_0(\vec{h}, b, z) \tag{12}$$

$$s.t. \quad \vec{h}^\top \vec{x}_i + b \leq 0 \qquad \forall i : y_i = 1, \tag{13}$$

$$\vec{h}^\top \vec{x}_i + b \geq z \qquad \forall i : y_i = -1, \tag{14}$$

where z is some scalar. Complete this formulation by specifying a linear objective function f_0. What does the solution $p^\star$ say about the existence of the separating hyperplane?

(b) **Chebyshev Center.** Let $\mathcal{P} \subset \mathbb{R}^n$ be a non-empty polyhedron defined as the intersection of m hyperplanes $\mathcal{P} = \{\vec{x} : \vec{a}_i^\top \vec{x} \leq b_i \ \forall i = 1, 2, \ldots, m\}$. We define the closed Euclidean ball in $\mathbb{R}^n$ with radius R and center $\vec{x}_0$ as the set $\mathcal{B}(\vec{x}_0, R) = \{\vec{x} \in \mathbb{R}^n : \|\vec{x} - \vec{x}_0\|_2 \leq R\}$. We want to find a point $\vec{x}_0 \in \mathcal{P}$ that is the center of the largest closed Euclidean ball contained in $\mathcal{P}$. Cast this problem as an LP.

5. L-smooth functions

Let $L > 0$ be a fixed constant. Consider the following three definitions of L-smooth functions.

- **Class definition**—In class we used the following definition. Given a function $f : \mathbb{R}^n \to \mathbb{R}$ with domain $\mathbb{R}^n$, consider the function
$$h(\vec{x}) := \frac{L}{2}\|\vec{x}\|_2^2 - f(\vec{x}),$$
with domain $\mathbb{R}^n$. Then f will be called L-smooth if h is a convex function. Note that this definition does not even require f to be differentiable. Let us call this the *class definition*.
- **Course reader definition**—In the course reader you will see the following definition. Given a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ with domain $\mathbb{R}^n$, it is called L-smooth iff it satisfies
$$f(\vec{y}) \leq f(\vec{x}) + \nabla f(\vec{x})^\top(\vec{y} - \vec{x}) + \frac{L}{2}\|\vec{y} - \vec{x}\|_2^2, \tag{15}$$
for all $\vec{x}, \vec{y} \in \mathbb{R}^n$. This definition, as stated, works for differentiable functions. Let us call this the *course reader definition*.
- **Natural definition**—There is a third definition. Given a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ with domain $\mathbb{R}^n$, it is called L-smooth iff it satisfies
$$\|\nabla f(\vec{y}) - \nabla f(\vec{x})\|_2 \leq L\|\vec{x} - \vec{y}\|_2, \tag{16}$$
for all $\vec{x}, \vec{y} \in \mathbb{R}^n$. This definition, as stated, works for differentiable functions. Let us call this the *natural definition*.

Note: The natural definition of L-smoothness is in general *not* equivalent to the class and course reader definitions of L-smoothness, as this problem will illustrate.

Remark: Let $g : \mathbb{R}^n \to \mathbb{R}^m$ with domain $\mathbb{R}^n$. Then g is said to be *Lipschitz with Lipschitz constant* L if we have
$$\|g(\vec{y}) - g(\vec{x})\|_2 \leq L\|\vec{y} - \vec{x}\|_2,$$
for all $\vec{x}, \vec{y} \in \text{dom}(g)$. Thus, a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, with domain $\mathbb{R}^n$, satisfies (16) precisely when $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, with domain $\mathbb{R}^n$, is Lipschitz with Lipschitz constant L. This is why we call this third definition the "natural" definition - it captures smoothness of the way in which the gradient changes as we move around in the domain of the function.

Example: Let $A \in \mathbb{S}_+^n$. Consider the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(\vec{x}) := \frac{1}{2}\vec{x}^\top A\vec{x}$, with $\text{dom}(f) = \mathbb{R}^n$. Then $\nabla f(\vec{x}) = A\vec{x}$ and $\nabla^2 f(\vec{x}) = A$. For $\vec{x}, \vec{y} \in \mathbb{R}^n$, we have
$$\|\nabla f(\vec{y}) - \nabla f(\vec{x})\|_2 = \|A(\vec{x} - \vec{y})\|_2.$$
Thus f is L-smooth according to the natural definition if and only if $\lambda_{\max}(A) \leq L$. Note that
$$\frac{L}{2}\|\vec{x}\|_2^2 - \frac{1}{2}\vec{x}^\top A\vec{x} = \frac{1}{2}\vec{x}^\top(LI - A)\vec{x}$$
defines a convex function if and only if $L \geq \lambda_{\max}(A)$. Thus, the same condition illustrates that f is L-smooth according to the class definition.

(a) Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function with domain $\mathbb{R}^n$. Show that f is L-smooth in the sense of the class definition if and only if it is L-smooth in the sense of the course reader definition.

Hint: Use the first order condition for characterizing convexity of differentiable functions, applied to $h(\vec{x}) := \frac{L}{2}\|\vec{x}\|_2^2 - f(\vec{x})$.

(b) Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable function with domain $\mathbb{R}^n$. Show that f is L-smooth in the sense of the class definition if and only if $LI - \nabla^2 f(\vec{x})$ is a positive semidefinite matrix for all $\vec{x} \in \mathbb{R}^n$.

Hint: Use the second order condition for characterizing convexity of twice differentiable functions, applied to $h(\vec{x}) := \frac{L}{2}\|\vec{x}\|_2^2 - f(\vec{x})$.

(c) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with domain $\mathbb{R}^n$ that is L-smooth in the sense of the natural definition. Show that we have

$$(\nabla f(\vec{y}) - \nabla f(\vec{x}))^\top (\vec{y} - \vec{x}) \leq L\|\vec{y} - \vec{x}\|_2^2, \tag{17}$$

for all $\vec{x}, \vec{y} \in \mathbb{R}^n$.

Hint: Use the Cauchy-Schwarz inequality.

(d) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function with domain $\mathbb{R}^n$ that is L-smooth in the sense of the natural definition. Show that it is L-smooth in the sense of the class definition.

Hint: Use the result of part (a) of problem 4 of Homework 6 for the function h.

So at this point, for differentiable f, we know that the class definition and the course reader definition of L-smoothness are equivalent, and we also know that if f is L-smooth in the sense of the natural definition, then it is L-smooth in the sense of the class definition (and hence also in the sense of the class reader definition).

Remark: L-smoothness is understood in a much broader context, without differentiability assumptions and without assuming that the domain of the function is all of $\mathbb{R}^n$. Also, there is a general theory involving norms other than L^2 norms, where the dual norm of the norm in question plays a role.

We will now show that all three definitions are equivalent for twice differentiable *convex* functions. Reading what follows is **optional**.

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable convex function with domain $\mathbb{R}^n$ that is L-smooth in the sense of the class definition. From part (b) we then know that $LI - \nabla^2 f(\vec{x})$ is a positive semidefinite matrix for all $\vec{x} \in \mathbb{R}^n$ and so, in particular, we have $\|\nabla^2 f(\vec{x})\|_2 \leq L$ for all $\vec{x} \in \mathbb{R}^n$, where $\|\nabla^2 f(\vec{x})\|_2$ denotes the induced L^2 norm. Another ingredient in the proof is the following integration formula, which you should make sure that you understand

$$\nabla f(\vec{y}) - \nabla f(\vec{x}) = \int_0^1 \nabla^2 f(\vec{x} + t(\vec{y} - \vec{x}))(\vec{y} - \vec{x})dt,$$

for all $\vec{x}, \vec{y} \in \mathbb{R}^n$. Finally, the third ingredient of the proof is the following characterization of the L^2 norm in $\mathbb{R}^n$

$$\|\vec{x}\|_2 = \max_{\vec{u} \in \mathbb{R}^n, \|\vec{u}\|_2 = 1} \vec{x}^\top \vec{u}$$

for all $\vec{x} \in \mathbb{R}^n$. This is true when $\vec{x} = \vec{0}$, since both sides of the above equality evaluate to 0, while for $\vec{x} \neq \vec{0}$ it can be easily proved using the Cauchy-Schwarz inequality (the maximum on the RHS occurs at $\frac{\vec{x}}{\|\vec{x}\|_2}$).

Putting these ingredients together, we have, for all $\vec{x}, \vec{y} \in \mathbb{R}^n$,

$$\begin{aligned}
\|\nabla f(\vec{y}) - \nabla f(\vec{x})\|_2 &= \max_{\vec{u} \in \mathbb{R}^n, \|\vec{u}\|_2 = 1} \vec{u}^\top (\nabla f(\vec{y}) - \nabla f(\vec{x})) \\
&= \max_{\vec{u} \in \mathbb{R}^n, \|\vec{u}\|_2 = 1} \int_0^1 \vec{u}^\top \nabla^2 f(\vec{x} + t(\vec{y} - \vec{x}))(\vec{y} - \vec{x})dt \\
&\leq \int_0^1 \left(\max_{\vec{u} \in \mathbb{R}^n, \|\vec{u}\|_2 = 1} \vec{u}^\top \nabla^2 f(\vec{x} + t(\vec{y} - \vec{x}))(\vec{y} - \vec{x}) \right) dt \\
&= \int_0^1 \|\nabla^2 f(\vec{x} + t(\vec{y} - \vec{x}))(\vec{y} - \vec{x})\|_2 dt
\end{aligned}$$

$$\begin{aligned}
&\leq \int_0^1 \|\nabla^2 f(\vec{x}+t(\vec{y}-\vec{x}))\|_2 \|(\vec{y}-\vec{x})\|_2 dt \\
&= \|(\vec{y}-\vec{x})\|_2 \int_0^1 \|\nabla^2 f(\vec{x}+t(\vec{y}-\vec{x}))\|_2 dt \\
&\leq \|(\vec{y}-\vec{x})\|_2 \int_0^1 L\, dt \\
&= L\|(\vec{y}-\vec{x})\|_2,
\end{aligned}$$

which establishes the claim. Here, in the last inequality, we have used the fact that $LI - \nabla^2 f(\vec{x}+t(\vec{y}-\vec{x}))$ is symmetric positive semi-definite, which is a consequence of L-smoothness in the sense of the class definition for twice differentiable functions. From this we can conclude that, for each $\vec{x} \in \mathbb{R}^n$, all the eigenvalues of $\nabla^2 f(\vec{x})$ are bounded above by L, and since they are nonnegative by the assumption of convexity, we can conclude that $\|\nabla^2 f(\vec{x})\|_2$ is bounded above by L.