EECS 127/227AT Optimization Models in Engineering UC Berkeley Fall 2023

Homework 9

Self grades are due at 11 PM on November 9, 2023.

1. Diagonally Dominant Matrices

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

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

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.

Solution: To prove that A is PSD we have to show that for any $\vec{x} \in \mathbb{R}^n$ we have $\vec{x}^\top A \vec{x} \geq 0$.

$$\begin{aligned}
\vec{x}^\top A \vec{x} &= \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \\
&= \sum_{i=1}^{n} A_{ii} x_i^2 + \sum_{i=1}^{n} \sum_{j \neq i} A_{ij} x_i x_j \\
&\geq \sum_{i=1}^{n} \sum_{j \neq i} |A_{ij}| x_i^2 - \sum_{i=1}^{n} \sum_{j \neq i} |A_{ij}| |x_i| |x_j| \\
&= \sum_{i=1}^{n} \sum_{j>i} |A_{ij}| x_i^2 + \sum_{i=1}^{n} \sum_{j<i} |A_{ij}| x_i^2 - \sum_{i=1}^{n} \sum_{j>i} |A_{ij}| |x_i| |x_j| - \sum_{i=1}^{n} \sum_{j<i} |A_{ij}| |x_i| |x_j| && (1) \\
&= \sum_{i=1}^{n} \sum_{j>i} |A_{ij}| x_i^2 + \sum_{j=1}^{n} \sum_{i>j} |A_{ij}| x_i^2 - \sum_{i=1}^{n} \sum_{j>i} |A_{ij}| |x_i| |x_j| - \sum_{j=1}^{n} \sum_{i>j} |A_{ij}| |x_i| |x_j| && (2) \\
&= \sum_{i=1}^{n} \sum_{j>i} |A_{ij}| x_i^2 + \sum_{i=1}^{n} \sum_{j>i} |A_{ji}| x_j^2 - \sum_{i=1}^{n} \sum_{j>i} |A_{ij}| |x_i| |x_j| - \sum_{i=1}^{n} \sum_{j>i} |A_{ji}| |x_j| |x_i| && (3) \\
&= \sum_{i=1}^{n} \sum_{j>i} |A_{ij}| x_i^2 + \sum_{i=1}^{n} \sum_{j>i} |A_{ij}| x_j^2 - \sum_{i=1}^{n} \sum_{j>i} |A_{ij}| |x_i| |x_j| - \sum_{i=1}^{n} \sum_{j>i} |A_{ij}| |x_i| |x_j| && (4) \\
&= \sum_{i=1}^{n} \sum_{j>i} |A_{ij}| (x_i^2 + x_j^2 - 2|x_i||x_j|) \\
&= \sum_{i=1}^{n} \sum_{j>i} |A_{ij}| (|x_i| - |x_j|)^2 \\
&\geq 0.
\end{aligned}$$

In the above equations, (2) follows from (1) by exchanging the order of summation over the indices i and j in the second and fourth terms, (3) follows from (2) by exchanging the indices i and j in the second and fourth terms, and (4) follows from (3) because A is symmetric, and thus $A_{ij} = A_{ji}$ for all $i, j \in \{1, \cdots, n\}$.

(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.

Solution: Computing the Hessian of f,

$$\frac{\partial f(\vec{x})}{\partial x_i} = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}},$$

$$
\begin{aligned}
\frac{\partial^2 f(\vec{x})}{\partial x_i^2} &= \frac{e^{x_i}\sum\limits_{j=1}^{n} e^{x_j} - e^{2x_i}}{\left(\sum\limits_{j=1}^{n} e^{x_j}\right)^2}, \\
\frac{\partial^2 f(\vec{x})}{\partial x_i \partial x_j} &= \frac{-e^{x_i+x_j}}{\left(\sum\limits_{j=1}^{n} e^{x_j}\right)^2}, \text{ if } j \neq i.
\end{aligned}
$$

Thus, the Hessian is symmetric and diagonally dominant and hence is PSD, which implies that f is convex.

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:

$$\min_{\vec{x} \in \mathbb{R}^n} f_\lambda(\vec{x}) \qquad \text{where} \qquad f_\lambda(\vec{x}) \doteq \frac{1}{2}\left\{\frac{1}{m}\|A\vec{x} - \vec{y}\|_2^2 + \lambda\|\vec{x}\|_2^2\right\}. \tag{5}$$

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 (5) is

$$\vec{x}_\lambda^\star = \left(A^\top A + \lambda m I\right)^{-1} A^\top \vec{y}. \tag{6}$$

Solution: The function $\lambda\|\vec{x}\|_2^2$ in Equation (5) is strictly convex, so f_λ is strictly convex. Thus the problem in Equation (5) has strictly convex objective and convex feasible set $\mathbb{R}^n$, so it has at most one solution. We can find a solution by setting the gradient to $\vec{0}$:

$$\nabla f_\lambda(\vec{x}_\lambda^\star) = \frac{1}{m}A^\top(A\vec{x}_\lambda^\star - \vec{y}) + \lambda\vec{x}_\lambda^\star \tag{7}$$

$$= \left(\frac{A^\top A}{m} + \lambda I\right)\vec{x}_\lambda^\star - \frac{1}{m}A^\top\vec{y} \tag{8}$$

$$\implies \left(\frac{A^\top A}{m} + \lambda I\right)\vec{x}_\lambda^\star = \frac{1}{m}A^\top\vec{y} \tag{9}$$

$$\implies \left(A^\top A + \lambda m I\right)\vec{x}_\lambda^\star = A^\top\vec{y} \tag{10}$$

$$\implies \vec{x}_\lambda^\star = \left(A^\top A + \lambda m I\right)^{-1} A^\top\vec{y}. \tag{11}$$

(b) Show that the GD update

$$\vec{x}_{t+1} = \vec{x}_t - \eta\left(\frac{1}{m}A^\top(A\vec{x}_t - \vec{y}) + \lambda\vec{x}_t\right) \tag{12}$$

can be rearranged into the form

$$\vec{x}_{t+1} - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)(\vec{x}_t - \vec{x}_\lambda^\star). \tag{13}$$

Use this to show that

$$\vec{x}_t - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)^t(\vec{x}_0 - \vec{x}_\lambda^\star). \tag{14}$$

for every positive integer t.

Solution: We have

$$\vec{x}_{t+1} = \vec{x}_t - \eta\left(\frac{1}{m}A^\top(A\vec{x}_t - \vec{y}) + \lambda\vec{x}_t\right) \tag{15}$$

$$= \vec{x}_t - \eta \cdot \frac{A^\top A}{m}\vec{x}_t + \eta \cdot \frac{1}{m}A^\top\vec{y} + \eta\lambda\vec{x}_t \tag{16}$$

$$= \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)\vec{x}_t + \eta \cdot \frac{1}{m}A^\top\vec{y} \tag{17}$$

$$\implies \vec{x}_{t+1} - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)\vec{x}_t + \eta \cdot \left(\frac{A^\top A}{m} + \lambda I\right)\vec{x}_\lambda^\star - \vec{x}_\lambda^\star \tag{18}$$

$$= \left(I - \eta\left(\frac{A^{\top}A}{m} + \lambda I\right)\right)(\vec{x}_t - \vec{x}_{\lambda}^{\star}). \tag{19}$$

Iterating this relation obtains the second equality.

(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

$$\vec{z}_t - \vec{z}_{\lambda}^{\star} = \left(I - \eta\left(\frac{\Sigma^{\top}\Sigma}{m} + \lambda I\right)\right)^t (\vec{z}_0 - \vec{z}_{\lambda}^{\star}), \tag{20}$$

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

$$(\vec{z}_t)_i - (\vec{z}_{\lambda}^{\star})_i = \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right)^t ((\vec{z}_0)_i - (\vec{z}_{\lambda}^{\star})_i) \tag{21}$$

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.

Solution: If $A = U\Sigma V^{\top}$ then

$$A^{\top}A = V\Sigma^{\top}U^{\top}U\Sigma V^{\top} = V\Sigma^{\top}\Sigma V^{\top}. \tag{22}$$

Thus we have

$$\begin{aligned}
\vec{x}_t - \vec{x}_{\lambda}^{\star} &= \left(I - \eta\left(\frac{A^{\top}A}{m} + \lambda I\right)\right)^t (\vec{x}_0 - \vec{x}_{\lambda}^{\star}) && (23)\\
&= \left(I - \eta\left(\frac{V\Sigma^{\top}\Sigma V^{\top}}{m} + \lambda I\right)\right)^t (\vec{x}_0 - \vec{x}_{\lambda}^{\star}) && (24)\\
&= \left(I - \eta\left(V\left(\frac{\Sigma^{\top}\Sigma}{m} + \lambda I\right)V^{\top}\right)\right)^t (\vec{x}_0 - \vec{x}_{\lambda}^{\star}) && (25)\\
&= \left(I - \eta V\left(\frac{\Sigma^{\top}\Sigma}{m} + \lambda I\right)V^{\top}\right)^t (\vec{x}_0 - \vec{x}_{\lambda}^{\star}) && (26)\\
&= \left(V\left(I - \eta\left(\frac{\Sigma^{\top}\Sigma}{m} + \lambda I\right)\right)V^{\top}\right)^t (\vec{x}_0 - \vec{x}_{\lambda}^{\star}) && (27)\\
&= V\left(I - \eta\left(\frac{\Sigma^{\top}\Sigma}{m} + \lambda I\right)\right)^t V^{\top}(\vec{x}_0 - \vec{x}_{\lambda}^{\star}) && (28)\\
\implies V^{\top}(\vec{x}_t - \vec{x}_{\lambda}^{\star}) &= \left(I - \eta\left(\frac{\Sigma^{\top}\Sigma}{m} + \lambda I\right)\right)^t V^{\top}(\vec{x}_0 - \vec{x}_{\lambda}^{\star}) && (29)\\
\implies \vec{z}_t - \vec{z}_{\lambda}^{\star} &= \left(I - \eta\left(\frac{\Sigma^{\top}\Sigma}{m} + \lambda I\right)\right)^t (\vec{z}_0 - \vec{z}_{\lambda}^{\star}). && (30)
\end{aligned}$$

Now note that $I - \eta(\frac{\Sigma^{\top}\Sigma}{m} + \lambda I)$ is a diagonal matrix. Thus we have

$$\begin{aligned}
\vec{z}_t - \vec{z}_{\lambda}^{\star} &= \left(I - \eta\left(\frac{\Sigma^{\top}\Sigma}{m} + \lambda I\right)\right)^t (\vec{z}_0 - \vec{z}_{\lambda}^{\star}) && (31)\\
\implies (\vec{z}_t - \vec{z}_{\lambda}^{\star})_i &= \left(I - \eta\left(\frac{\Sigma^{\top}\Sigma}{m} + \lambda I\right)\right)_i^t (\vec{z}_0 - \vec{z}_{\lambda}^{\star})_i && (32)\\
\implies (\vec{z}_t)_i - (\vec{z}_{\lambda}^{\star})_i &= \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right)^t ((\vec{z}_0)_i - (\vec{z}_{\lambda}^{\star})_i) && (33)
\end{aligned}$$

(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

$$\max_{i\in\{1,\dots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1. \tag{34}$$

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

$$\eta \in \left(0, \frac{2m}{\sigma_{\max}\{A\}^2 + \lambda m}\right) \tag{35}$$

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

Solution: We have

$$\begin{aligned}
& \lim_{t\to\infty} \vec{z}_t = \vec{z}_\lambda^\star, \qquad \forall \vec{x}_0 && (36)\\
\iff & \lim_{t\to\infty} (\vec{z}_t)_i = (\vec{z}_\lambda^\star)_i, \qquad \forall i \quad \forall \vec{x}_0 && (37)\\
\iff & \lim_{t\to\infty} \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right)^t = 0, \qquad \forall i && (38)\\
\iff & \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1, \qquad \forall i && (39)\\
\iff & \max_{i\in\{1,\dots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1. && (40)
\end{aligned}$$

This proves what is required in the first part of the question. The solution to the second part of the question follows by noting that

$$\begin{aligned}
& \max_{i\in\{1,\dots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1 && (41)\\
\iff & \max_{i\in\{1,\dots,n\}} \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right) < 1 && (42)\\
& \text{and} \quad \min_{i\in\{1,\dots,n\}} \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right) > -1 && (43)\\
\iff & 1 - \eta\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right) < 1 && (44)\\
& \text{and} \quad 1 - \eta\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right) > -1. && (45)
\end{aligned}$$

Now the first equation is always satisfied for $\eta > 0$ and $\lambda > 0$ because $\frac{\sigma_{\min}\{A\}^2}{m} + \lambda > 0$, so $1 - \eta(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda) < 1$. The second equation is satisfied when $\eta < \frac{2m}{\sigma_{\max}\{A\}^2 + \lambda m}$. Since $\lim_{t\to\infty} \vec{x}_t = \vec{x}_\lambda^\star$ if and only if $\lim_{t\to\infty} \vec{z}_t = \vec{z}_\lambda^\star$, we have that gradient descent converges for all initializations $\vec{x}_0$ if and only if $0 < \eta < \frac{2m}{\sigma_{\max}\{A\}^2 + \lambda m}$.

(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., $\lim_{T\to\infty} \vec{x}_T = \vec{x}_\lambda^\star$, 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:

$$\eta^\star \in \operatorname*{argmin}_{\eta\in\mathbb{R}} \max_{i\in\{1,\ldots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right|. \tag{46}$$

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

$$\eta^\star = \frac{2m}{\sigma_{\max}\{A\}^2 + \sigma_{\min}\{A\}^2 + 2\lambda m} \tag{47}$$

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.

Solution: We have

$$\begin{aligned}
\left|1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right)\right| &= \left|1 - \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right)\right| && (48)\\
1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right) &= -\left(1 - \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right)\right) && (49)\\
1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right) &= \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right) - 1 && (50)\\
2 &= \eta^\star\left(\frac{\sigma_{\max}\{A\}^2 + \sigma_{\min}\{A\}^2}{m} + 2\lambda\right) && (51)\\
\eta^\star &= \frac{2m}{\sigma_{\max}\{A\}^2 + \sigma_{\min}\{A\}^2 + 2m\lambda}. && (52)
\end{aligned}$$

Here the second equality is the most challenging to derive. It follows from the first inequality by the following reasoning:

- If $1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right)$ and $1 - \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right)$ have the same sign, then by the first equality, they must be equal. This means that $\sigma_{\max}\{A\} = \sigma_1\{A\} = \sigma_2\{A\} = \cdots = \sigma_n\{A\} = \sigma_{\min}\{A\}$ and the optimal step size $\eta^\star$ sets each rate $1 - \eta^\star\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)$ to 0 simultaneously, ensuring convergence in one step. If both sides are 0 then the second equality holds (because $0 = -0$).
- Otherwise, $1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right)$ and $1 - \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right)$ have opposite signs. Since $\sigma_{\max}\{A\} > \sigma_{\min}\{A\}$ (since if they were equal we would be in the first case), we have $1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right) > 1 - \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right)$. Thus $1 - \eta^\star\left(\frac{\sigma_{\min}\{A\}^2}{m} + \lambda\right)$ must be positive and $1 - \eta^\star\left(\frac{\sigma_{\max}\{A\}^2}{m} + \lambda\right)$ must be negative. The absolute value of a negative number is its negative, so the second equality follows directly from the first equality.

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?

Solution: The problem, as stated, is not convex, due to the non-convex constraint. This is true even when Q is positive semi-definite; in that case only the objective function is convex.

(b) Show that the problem can be reduced to

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

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

Solution: Let $Q = U \Lambda U^\top$ be the eigenvalue decomposition of Q, with U orthonormal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ the diagonal matrix containing the eigenvalues in decreasing order. With the change of variable $\vec{y} = U^\top \vec{x}$, and with $\vec{d} = U^\top \vec{c}$, we have $\vec{c}^\top \vec{x} = \vec{d}^\top \vec{y}$, and $\vec{x}^\top Q \vec{x} = \vec{y}^\top \Lambda \vec{y}$. Since U is orthonormal the constraint on $\vec{x}$ becomes $\|\vec{y}\|_2 = 1$. This proves the result.

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

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

Solution: Note that given a solution $\vec{y}^*$, if y_i is the same sign as d_i, you can decrease the objective value by changing the sign of y_i while still remaining feasible. Therefore, at optimum, y_i must have the opposite sign of d_i. Then the problem can be written as

$$\min_{\vec{\xi}} \sum_{i=1}^{n} \left(\lambda_i \xi_i^2 - 2|d_i|\xi_i\right) \ : \ \sum_{i=1}^{n} \xi_i^2 = 1, \ \ \xi \geq 0,$$

with $\xi_i = -\mathbf{sign}(d_i) y_i$, $i = 1, \ldots, n$.

The new formulation results from the change of variable $z_i = \xi_i^2$, $i = 1, \ldots, n$. The optimal $\vec{y}$ is then obtained as $y_i = -\mathbf{sign}(d_i)\sqrt{z_i}$, $i = 1, \ldots, n$.

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

$$\begin{aligned} p^\star = \min_{\vec{h}, b, z} \quad & f_0(\vec{h}, b, z) && && (53) \\ s.t. \quad & \vec{h}^\top \vec{x}_i + b \leq 0 && \forall i : y_i = 1, && (54) \\ & \vec{h}^\top \vec{x}_i + b \geq z && \forall i : y_i = -1, && (55) \end{aligned}$$

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?

Solution: With the choice of the objective function

$$f_0(\vec{h}, b, z) = -z, \tag{56}$$

the separating hyperplane exists if $p^\star < 0$.

To see this note that $p^\star < 0$ if and only if the optimal solution to the problem $(\vec{h}^\star, b^\star, z^\star)$ is such that $z^\star > 0$. Now consider the hyperplane $\mathcal{L}^\star = \{\vec{x} \in \mathbb{R}^n : \vec{h}^{\star\top} \vec{x} + b^\star = 0\}$. From feasibility of the optimal solution, we can see that

$$\begin{aligned} & \vec{h}^{\star\top} \vec{x}_i + b^\star \leq 0 && \forall i : y_i = 1, && (57) \\ & \vec{h}^{\star\top} \vec{x}_i + b^\star \geq z^\star > 0 && \forall i : y_i = -1, && (58) \end{aligned}$$

which is the definition of linear separability. Thus $\mathcal{L}^\star$ is indeed a separating hyperplane for this data.

(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.

Solution: Any point $\vec{x} \in \mathcal{B}(\vec{x}_0, R)$ can be expressed as $\vec{x} = \vec{x}_0 + \vec{u}$ where $\|\vec{u}\|_2 \leq R$. To satisfy the condition that $\mathcal{B}(\vec{x}_0, R) \subset \mathcal{P}$ we need for all $\vec{u} \in \mathbb{R}^n$ with norm $\|\vec{u}\|_2 \leq R$:

$$\vec{a}_i^\top (\vec{x}_0 + \vec{u}) \leq b_i \qquad \forall i = 1, 2, \ldots, m \tag{59}$$

We take the maximum over $\vec{u}$ of both sides to get the equivalent condition

$$\begin{aligned} \max_{\|\vec{u}\|_2 \leq R} \left(\vec{a}_i^\top (\vec{x}_0 + \vec{u}) \right) \leq b_i & && (60) \\ \vec{a}_i^\top \vec{x}_0 + \max_{\|\vec{u}\|_2 \leq R} \left(\vec{a}_i^\top \vec{u} \right) \leq b_i & \qquad \forall i = 1, 2, \ldots, m && (61) \end{aligned}$$

The inner product $\vec{a}_i^\top \vec{u}$ is maximized when $\vec{u}$ is the longest possible vector along the direction of $\vec{a}_i$, thus

$$\begin{aligned} \max_{\|\vec{u}\|_2 \leq R} \left(\vec{a}_i^\top \vec{u} \right) &= \vec{a}_i^\top \left(\frac{R}{\|\vec{a}_i\|_2} \vec{a}_i \right) && (62) \\ &= R \|\vec{a}_i\|_2 && (63) \end{aligned}$$

This gives the following conditions

$$\vec{a}_i^\top \vec{x}_0 + R \left\|\vec{a}_i\right\|_2 \leq b_i \qquad \forall i = 1, 2, \ldots, m. \tag{64}$$

Now we can write the problem of finding the largest ball enclosed in $\mathcal{P}$ as

$$\min_{\vec{x}_0, R} - R \tag{65}$$

$$s.t. \quad \vec{a}_i^\top \vec{x}_0 + R \left\|\vec{a}_i\right\|_2 \leq b_i \qquad \forall i = 1, 2, \ldots, m. \tag{66}$$

This is a linear program, the variables being the coordinates of $\vec{x}_0$ and the scalar R. One can add the constraint $R \geq 0$ if one wishes, but it is not necessary because we are told that $\mathcal{P}$ is non-empty, so we have a feasible point with $\vec{x}_0 \in \mathcal{P}$ and $R = 0$, which guarantees that we must have $R \geq 0$ at an optimal point.

Note: If we had not been told that $\mathcal{P}$ is nonempty, then it would have been necessary to impose the constraint $R \geq 0$ to obtain a correct formulation. For example, suppose the polytope $\mathcal{P}$ were defined by the constraints $x \leq 0$ and $x \geq 3$, where $x \in \mathbb{R}$. Then $\mathcal{P}$ is the empty set. However, you can check that $x_0 = \frac{3}{2}$ and $R = -\frac{3}{2}$ is feasible for the LP above, and is indeed the optimal solution to this LP, so the LP does not represent the original problem in this case.

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
$$f(\vec{y}) \leq f(\vec{x}) + \nabla f(\vec{x})^\top(\vec{y} - \vec{x}) + \frac{L}{2}\|\vec{y} - \vec{x}\|_2^2, \tag{67}$$
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
$$\|\nabla f(\vec{y}) - \nabla f(\vec{x})\|_2 \leq L\|\vec{x} - \vec{y}\|_2, \tag{68}$$
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 (68) 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 $h(\vec{x}) := \frac{L}{2}\|\vec{x}\|_2^2 - f(\vec{x})$.

Solution: Let

$$h(\vec{x}) := \frac{L}{2}\|\vec{x}\|_2^2 - f(\vec{x}),$$

with domain $\mathbb{R}^n$, and note that

$$\nabla h(\vec{x}) = L\vec{x} - \nabla f(\vec{x}).$$

Suppose h is convex. Then, for all $\vec{x}, \vec{y} \in \mathbb{R}^n$, we have

$$h(\vec{y}) \geq h(\vec{x}) + \nabla h(\vec{x})^\top (\vec{y} - \vec{x}). \tag{69}$$

Substituting for h and ∇h in terms of f and ∇f respectively, this gives

$$\frac{L}{2}\|\vec{y}\|_2^2 - f(\vec{y}) \geq \frac{L}{2}\|\vec{x}\|_2^2 - f(\vec{x}) + \left(L\vec{x} - \nabla f(\vec{x})\right)^\top (\vec{y} - \vec{x}). \tag{70}$$

Rearranging this gives (67). This shows that for a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ with domain $\mathbb{R}^n$, being L-smooth in the sense of the class definition implies that it is L-smooth in the sense of the course reader definition.

Conversely, suppose (67) holds for all $\vec{x}, \vec{y} \in \mathbb{R}^n$. Rearranging this gives (70), for all $\vec{x}, \vec{y} \in \mathbb{R}^n$, which is just (69), which implies that h is convex. This shows that for a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ with domain $\mathbb{R}^n$, being L-smooth in the sense of the course reader definition implies that it is L-smooth in the sense of the class definition.

(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 $h(\vec{x}) := \frac{L}{2}\|\vec{x}\|_2^2 - f(\vec{x})$.

Solution: Let $h(\vec{x}) := \frac{L}{2}\|\vec{x}\|_2^2 - f(\vec{x})$, with domain $\mathbb{R}^n$. Note that h is twice differentiable. Note that $\nabla h(\vec{x}) = L\vec{x} - \nabla f(\vec{x})$ and

$$\nabla^2 h(\vec{x}) = LI - \nabla^2 f(\vec{x}).$$

By the second order condition for characterizing convexity of twice differentiable functions we see that h is convex if and only if $LI - \nabla^2 f(\vec{x})$ is a positive semidefinite matrix for all $\vec{x} \in \mathbb{R}^n$. This establishes the claim.

(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

$$\left(\nabla f(\vec{y}) - \nabla f(\vec{x})\right)^\top (\vec{y} - \vec{x}) \leq L\|\vec{y} - \vec{x}\|_2^2, \tag{71}$$

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

Hint: Use the Cauchy-Schwarz inequality.

Solution: We have

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

where the first inequality is an application of the Cauchy-Schwarz inequality and the second inequality is from the definition of L-smoothness in the sense of the natural definition. This establishes the claim.

(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.

Solution: Note that

$$\nabla h(\vec{x}) = L\vec{x} - \nabla f(\vec{x}).$$

For any $\vec{x}, \vec{y} \in \mathbb{R}^n$, since

$$\begin{aligned}(\nabla h(\vec{x}) - \nabla h(\vec{y}))^\top (\vec{y} - \vec{x}) &= (L\vec{x} - \nabla f(\vec{x}) - L\vec{y} + \nabla f(\vec{y}))^\top (\vec{y} - \vec{x}) \\ &= (\nabla f(\vec{y}) - \nabla f(\vec{x}))^\top (\vec{y} - \vec{x}) - L\|\vec{y} - \vec{x}\|_2^2,\end{aligned}$$

we see that (71) is equivalent to

$$(\nabla h(\vec{y}) - \nabla h(\vec{x}))^\top (\vec{y} - \vec{x}) \geq 0,$$

which we know, by the result of part (a) of problem 4 of Homework 6, is equivalent to h being convex. Since (71) is implied by the assumption of being L-smooth in the sense of the natural definition, this proves what was desired.

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

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

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

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.