CORNELL TECH | HOME OF THE JACOBS TECHNION-CORNELL INSTITUTE

[Fall 2025] ECE 5290/7290 and ORIE 5290
Distributed Optimization for Machine Learning and AI
Homework 1

Gradescope Due: September 12th at 5PM

Objective of This Assignment

The objective is to bridge the gap between the mathematical foundations of optimization and their practical application in machine learning. The first two problems are designed to reinforce the essential prerequisite knowledge from linear algebra and multivariate calculus, focusing on the concepts of matrix properties and gradient calculations that form the bedrock of continuous optimization. The subsequent problems will build on this foundation, guiding you to implement a gradient descent algorithm on real loss functions.

Question 1: Linear Algebra Review (15 points)

This problem reviews the concept of positive semidefinite (PSD) matrices, which is fundamental to convex optimization.

(a) (5 points) Consider the following three matrices:

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 4 & -2 \\ -2 & 1 \end{pmatrix}$$

For each matrix, determine if it is positive definite (PD), positive semidefinite (PSD), or indefinite.

A is PD **(2 Points)**. B is indefinite **(2 Points)**. C is PSD **(1 Point)**.

(b) (5 points) Briefly explain your reasoning for each matrix. You can justify your answer by computing eigenvalues, checking principal minors, or using the definition $\boldsymbol{\theta}^\top M\boldsymbol{\theta}$.

- For A: The leading principal minor is $2 > 0$ and $\det(A) = 3 > 0$. Hence A is positive definite. **(2 Points)**
- For B: The determinant is $\det(B) = -3 < 0$, so eigenvalue have opposite signs (since $\det(B) = \lambda_1\lambda_2$). Hence B is indefinite. **(2 Points)**
- For C: Observe that $C = vv^\top$ with $v = (2, -1)$, so for any x, $x^\top Cx = (v^\top x)^2 \geq 0$. We have $v^\top x = 0$ when $x = (-1, 2)$. **(1 Point)**

(c) (5 points) Consider the quadratic function

$$f(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top M\boldsymbol{\theta} - \mathbf{b}^\top\boldsymbol{\theta}.$$

Explain why the convexity of this function depends on the properties of the matrix M. What property must M have for $f(\boldsymbol{\theta})$ to be convex?

We use the fact that if the Hessian of a function is positive semidefinite (PSD) everywhere, then the function is convex.

(2 points): Gradient. The gradient is

$$\nabla f(\boldsymbol{\theta}) = \tfrac{1}{2}(M + M^\top)\boldsymbol{\theta} - \mathbf{b}.$$

(2 points): Hessian. Taking another derivative gives the Hessian:

$$\nabla^2 f(\boldsymbol{\theta}) = \tfrac{1}{2}(M + M^\top).$$

(1 point): Convexity condition. Since a function is convex if and only if its Hessian is positive semidefinite (PSD), f is convex precisely when $M + M^\top$ is PSD. In the special case where M is symmetric, this condition reduces to requiring that M itself be PSD.

Question 2: Calculus Review (40 points)

This problem reviews multivariate calculus, which is essential for gradient-based optimization.

(a) (5 points) Find the first derivative, $f'(\theta)$, for the following function, which requires the chain rule:

$$f(\theta) = \exp(-(\theta - 2)^2)$$

$$f'(\theta) = -2(\theta - 2)\exp(-(\theta - 2)^2)$$

(b) (5 points) Consider the function $h(\theta) = (\theta - 5)^2 + 3$. Find the value of θ that minimizes this function. Explain how you can use the derivative to find this minimum.

We first compute its gradient and hessian.

(3 points) Find Gradient and Hessian.

$$h'(\theta) = 2\theta - 10, \quad h''(\theta) = 2$$

(2 points) Final minimum. Since $h''(\theta) > 0$, the function is convex. The minimum θ^* is attained when $h'(\theta^*) = 0$, which gives $\theta^* = 5$.

(c) (5 points) Let $\boldsymbol{\theta} \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, and $\mathbf{b} \in \mathbb{R}^m$. Consider the standard least-squares objective function:

$$f(\boldsymbol{\theta}) = \|A\boldsymbol{\theta} - \mathbf{b}\|_2^2$$

Derive the gradient of this function with respect to $\boldsymbol{\theta}$, which is the vector $\nabla f(\boldsymbol{\theta}) \in \mathbb{R}^n$.

(1 Points) Notice that we can write $f(\theta)$ in the following way:

$$\begin{aligned} f(\theta) = \|A\boldsymbol{\theta} - \mathbf{b}\|_2^2 &= (A\boldsymbol{\theta} - \mathbf{b})^\top (A\boldsymbol{\theta} - \mathbf{b}) \\ &= \big((A\boldsymbol{\theta})^\top - \mathbf{b}^\top\big)(A\boldsymbol{\theta} - \mathbf{b}) \\ &= \theta^\top A^\top A\theta - (A\theta)^\top b - b^\top (A\boldsymbol{\theta}) - b^\top b \\ &= \theta^\top A^\top A\theta - 2b^\top A\theta - b^\top b \end{aligned}$$

(4 Points) Notice that $A^\top A$ is a symetric matrix. Using the result we derive in Question 1, the gradient is:

$$\nabla f(\theta) = (A^\top A + (A^\top A)^\top)\theta - 2(b^\top A)^\top = 2A^\top A\theta - 2A^\top b$$

(d) (5 points) Derive the Hessian of $f(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, which is the matrix $\nabla^2 f(\boldsymbol{\theta}) \in \mathbb{R}^{n\times n}$.

Notice that we have $\nabla^2_{i,j} f(\boldsymbol{\theta}) = (2A^\top A)_{(i,j)}$, so $\nabla^2 f(\boldsymbol{\theta}) = 2A^\top A$.

(e) (5 points) Using your result from part (d), what can you say about the convexity of $f(\boldsymbol{\theta})$? (Hint: Think about the properties of the matrix $A^\top A$ and your conclusions from Problem 1).

For any vector $x \in \mathbb{R}^n$, we have $x^\top A^\top Ax = (Ax)^\top (Ax) \geq 0$ so $f(\theta)$ is convex.

(f) (5 points) The sigmoid function, $\sigma(z) = (1 + e^{-z})^{-1}$, is a core component of logistic regression. Find its derivative, $\sigma'(z)$, with respect to z.
Hint: Show that the derivative can be simplified to the well-known form: $\sigma(z)(1 - \sigma(z))$.

(2 Points) Notice that $(1 - \sigma(z)) = 1 - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}}$.

(3 Points) We then compute the gradient:

$$\sigma'(z) = (1 + e^{-z})^{-2} e^{-z} = (1 + e^{-z})^{-1} \frac{e^{-z}}{1 + e^{-z}} = \sigma(z)(1 - \sigma(z))$$

(g) (5 points) The Binary Cross-Entropy (BCE) loss for a single example $(\mathbf{x}, y)$ with $y \in \{0, 1\}$ is:

$$L_{BCE}(\boldsymbol{\theta}) = -y \log(\sigma(\boldsymbol{\theta}^\top \mathbf{x})) - (1 - y)\log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}))$$

A more compact form, often used in optimization literature, is the log-sum-exp form:

$$L_{\text{comp}}(\boldsymbol{\theta}) = -y\boldsymbol{\theta}^\top \mathbf{x} + \log(1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}))$$

Prove that these two forms are mathematically equivalent.
Hint: Let $z = \boldsymbol{\theta}^\top \mathbf{x}$*. Start with the BCE form and substitute the definition of* $\sigma(z)$*. You may find the identity* $\log(1 + e^{-z}) = \log(1 + e^z) - z$ *useful.*

Let $z = \boldsymbol{\theta}^\top \mathbf{x}$, we can derive the following chain of equality:

$$\begin{aligned}
L_{BCE}(\boldsymbol{\theta}) &= -y\log(\sigma(z)) - (1-y)\log(1-\sigma(z)) \\
&= y\log(1+e^{-z}) - (1-y)\log(\frac{e^{-z}}{1+e^{-z}}) \\
&= y\log(1+e^{-z}) - (1-y)(-z - \log(1+e^{-z})) \\
&= \log(1+e^{-z}) + (1-y)z \\
&= \log(1+e^{z}) - z + (1-y)z \\
&= \log(1+e^{z}) - yz \\
&= -y\boldsymbol{\theta}^\top\mathbf{x} + \log(1+\exp(\boldsymbol{\theta}^\top\mathbf{x})) = L_{\text{comp}}(\boldsymbol{\theta})
\end{aligned}$$

(h) (5 points) Using the **compact form**, $L_{\text{comp}}(\boldsymbol{\theta})$, derive its gradient with respect to the parameter vector $\boldsymbol{\theta}$, which is $\nabla L_{\text{comp}}(\boldsymbol{\theta})$.
Hint: Your answer should be in a simple form involving the prediction $\sigma(\boldsymbol{\theta}^\top\mathbf{x})$ and the true label y.

Let $z = \boldsymbol{\theta}^\top \mathbf{x}$. Differentiating the compact form gives

$$\frac{d}{dz}L_{\text{comp}}(z) = -y + \frac{e^z}{1+e^z} = \sigma(z) - y.$$

By the chain rule, since $\nabla_{\boldsymbol{\theta}} z = \mathbf{x}$, we obtain

$$\nabla L_{\text{comp}}(\boldsymbol{\theta}) = \big(\sigma(\boldsymbol{\theta}^\top\mathbf{x}) - y\big)\,\mathbf{x}.$$

Question 3: Linear Regression (20 points)

Consider the linear regression problem of finding $\boldsymbol{\theta}$ that minimizes the following least square loss function

$$L(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{\theta}^\top\mathbf{x}_i - y_i)^2 \tag{1}$$

where $\mathcal{D} = \left\{\mathbf{x}_i, y_i\right\}_{i=1}^N$ is the training data set with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, $\boldsymbol{\theta} \in \mathbb{R}^d$ is the parameter to be determined, and $\lambda \in \mathbb{R}$ is the regularization coefficient. We plan to implement the (batch) gradient descent algorithms to solve this problem, with a constant learning rate of α.

(a) (5 points) Write the batch gradient descent update at the t^{th} iteration, for solving this problem.

$$\boldsymbol{\theta}^k = \boldsymbol{\theta}^{k-1} - \alpha\sum_{i=1}^{N}(\mathbf{x}_i^\top\boldsymbol{\theta}^{k-1} - y_i)\mathbf{x}_i$$

(b) (5 points) Now consider you are provided with the data set given in Table 1. For this part, let $\alpha = 1$. By taking the initial parameter $\boldsymbol{\theta}^0 = (0,\ 0)^\top$, find the parameter $\boldsymbol{\theta}^1$ obtained after updating for one iteration using batch gradient descent.

i	$\mathbf{x}_i$	y_i
1	$(1,\ 1)^\top$	-1
2	$(-1,\ 1)^\top$	-1
3	$(\ -1,\ 0)^\top$	1

Table 1: Dataset for Problem 3

$$\nabla L(\boldsymbol{\theta}^0) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 - \alpha \nabla L(\boldsymbol{\theta}^0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 1 \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -1 \\ -2 \end{pmatrix}$$

(c) (10 points) For the batch gradient descent in (b), after updating $\boldsymbol{\theta}^0$ via two batch gradient descent iterations, does the loss $L(\boldsymbol{\theta})$ decrease at $\boldsymbol{\theta}^2$ compared with $\boldsymbol{\theta}^0$? If yes, please show the decrements $L(\boldsymbol{\theta}^2) - L(\boldsymbol{\theta}^0)$; if no, please suggest a way to address the stepsize α that decreases in the loss $L(\boldsymbol{\theta})$.

(4 Points) No.

(6 Points) Use $\alpha = 1/2$.

$$\nabla L(\boldsymbol{\theta}^0) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 - \alpha \nabla L(\boldsymbol{\theta}^0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} \\ -1 \end{pmatrix}$$

$$\nabla L(\boldsymbol{\theta}^1) = \begin{pmatrix} -\frac{1}{2} \\ 0 \end{pmatrix}$$

$$\boldsymbol{\theta}^2 = \boldsymbol{\theta}^1 - \alpha \nabla L(\boldsymbol{\theta}^1) = \begin{pmatrix} -\frac{1}{2} \\ -1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} -\frac{1}{2} \\ 0 \end{pmatrix} = \begin{pmatrix} -\frac{1}{4} \\ -1 \end{pmatrix}$$

Checking for change in $L(\boldsymbol{\theta})$, we get

$$L(\boldsymbol{\theta}^2) - L(\boldsymbol{\theta}^0) \approx -1.1562 < 0$$

$\Longrightarrow$ Decrease in $L(\boldsymbol{\theta})$.

Question 4: Logistic Regression (30 points)

We consider using logistic regression for a 2-class classification setting. Let $D = \{\mathbf{x}_i, y_i\}_{i=1}^N$ be a dataset for 2-class classification, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$.

(a) (10 points) The assumption for logistic regression is that the log-odds over the label class is affine i.e.

$$\log\left(\frac{P(\mathbf{x}|y=1)}{P(\mathbf{x}|y=0)}\right) = \boldsymbol{\theta}^\top \mathbf{x}, \tag{2}$$

Assuming additionally the label distribution is even (i.e. $P(y=0) = P(y=1) = 0.5$), derive the posterior $P(y|\mathbf{x})$ for both $y = 0$ and $y = 1$. Hint: use the Bayes rule.

By assumption,

$$\frac{P(\mathbf{x}|y=1)}{P(\mathbf{x}|y=0)} = \frac{P(\mathbf{x}|y=1)}{1 - P(\mathbf{x}|y=1)} = \exp\left(\boldsymbol{\theta}^\top \mathbf{x}\right)$$

(4 Points) Therefore,

$$\begin{aligned} P(\mathbf{x}|y=1) &= \frac{\exp\left(\boldsymbol{\theta}^\top \mathbf{x}\right)}{1 + \exp\left(\boldsymbol{\theta}^\top \mathbf{x}\right)} \\ P(\mathbf{x}|y=0) &= \frac{1}{1 + \exp\left(\boldsymbol{\theta}^\top \mathbf{x}\right)} \end{aligned}$$

(6 Points) Using Bayes' theorem,

$$\begin{aligned} P(y=1|\mathbf{x}) &= \frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=1)P(y=1) + P(\mathbf{x}|y=0)P(y=0)} = \frac{\exp\left(\boldsymbol{\theta}^\top \mathbf{x}\right)}{1 + \exp\left(\boldsymbol{\theta}^\top \mathbf{x}\right)} \\ P(y=0|\mathbf{x}) &= \frac{P(\mathbf{x}|y=0)P(y=0)}{P(\mathbf{x}|y=1)P(y=1) + P(\mathbf{x}|y=0)P(y=0)} = \frac{1}{1 + \exp\left(\boldsymbol{\theta}^\top \mathbf{x}\right)} \end{aligned}$$

(b) (10 points) For binary classification with $y_i \in \{0, 1\}$, the prediction rule using logistic regression is:

$$\begin{aligned} &\hat{y}_i = 1 \text{ if } P(y=1|\mathbf{x}) > 0.5, \\ &\hat{y}_i = 0 \text{ otherwise.} \end{aligned}$$

The loss function of the logistic regression is given by

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \left[-y_i \boldsymbol{\theta}^\top \mathbf{x}_i + \log(1 + \exp(\boldsymbol{\theta}^\top \mathbf{x}_i))\right]. \tag{3}$$

We would like to conduct the Gradient Descent on $\boldsymbol{\theta}$ to minimize the loss function. Can you perform 1-step Gradient Descent update with step size $\eta > 0$ to find $\boldsymbol{\theta}^{t+1}$ from $\boldsymbol{\theta}^t$?

(4 Points) To conduct Gradient Descent, we first calculate the gradient

$$\begin{aligned}\nabla L(\boldsymbol{\theta}) &= \sum_{i=1}^{n} -y_i \mathbf{x}_i + \frac{\exp(\boldsymbol{\theta}^T \mathbf{x}_i)}{1+\exp(\boldsymbol{\theta}^T \mathbf{x}_i)} \mathbf{x}_i \\ &= \sum_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\theta}^T \mathbf{x}_i)}{1+\exp(\boldsymbol{\theta}^T \mathbf{x}_i)} - y_i \right) \mathbf{x}_i\end{aligned}$$

(6 Points) The 1-step Gradient Descent update with step size $\eta > 0$ is

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \eta \nabla L(\boldsymbol{\theta}_t) \\ &= \boldsymbol{\theta}_t - \eta \sum_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\theta}^T x_i)}{1+\exp(\boldsymbol{\theta}^T x_i)} - y_i \right) \mathbf{x}_i\end{aligned}$$

(c) (10 points) Following all the settings in question (b), additionally assume that at the current iteration t we have $\boldsymbol{\theta}^t$ that successfully classifies all data points, and the data points with both labels exist in the dataset. Compare the norm of $\boldsymbol{\theta}^{t+1}$ to the norm of $\boldsymbol{\theta}^t$. Which is greater?

(2 Points) To compare the norms, it is equivalent to compare the square norm:

$$\begin{aligned}\|\boldsymbol{\theta}_{t+1}\|^2 &= \langle \boldsymbol{\theta}_t - \eta \nabla L(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \eta \nabla L(\boldsymbol{\theta}_t) \rangle \\ &= \|\boldsymbol{\theta}_t\|^2 + \eta^2 \|\nabla L(\boldsymbol{\theta}_t)\|^2 + 2\eta \langle \boldsymbol{\theta}_t, -\nabla L(\boldsymbol{\theta}_t) \rangle\end{aligned}$$

We know that $\eta^2 \|\nabla L(\boldsymbol{\theta}_t)\|^2 \geq 0$. Furthermore,

$$\langle \boldsymbol{\theta}_t, -\nabla L(\boldsymbol{\theta}_t) \rangle = \sum_{i=1}^{n} \left(y_i - \frac{\exp(\boldsymbol{\theta}^T \mathbf{x}_i)}{1+\exp(\boldsymbol{\theta}^T \mathbf{x}_i)} \right) \langle \boldsymbol{\theta}_t, \mathbf{x}_i \rangle$$

(2 Points) As $\boldsymbol{\theta}_t$ successfully classifies all data points, for $i \in \{i \in [n] : y_i = 1\}$,

$$\begin{aligned}&P(y=1|\mathbf{x}) = \frac{\exp\left(\boldsymbol{\theta}^\top \mathbf{x}\right)}{1+\exp\left(\boldsymbol{\theta}^\top \mathbf{x}\right)} > 0.5 \\ &\Rightarrow \boldsymbol{\theta}^\top \mathbf{x} = \langle \boldsymbol{\theta}_t, \mathbf{x}_i \rangle > 0\end{aligned}$$

(2 Points) For $i \in \{i \in [n] : y_i = 0\}$,

$$\begin{aligned}&P(y=1|\mathbf{x}) = \frac{\exp\left(\boldsymbol{\theta}^\top \mathbf{x}\right)}{1+\exp\left(\boldsymbol{\theta}^\top \mathbf{x}\right)} \leq 0.5 \\ &\Rightarrow \boldsymbol{\theta}^\top \mathbf{x} = \langle \boldsymbol{\theta}_t, \mathbf{x}_i \rangle \leq 0\end{aligned}$$

(2 Points) Therefore,

$$\begin{aligned}\langle \boldsymbol{\theta}_t, -\nabla L(\boldsymbol{\theta}_t)\rangle = &\sum_{i\in\{i\in[n]:y_i=0\}} \left(-\frac{\exp(\boldsymbol{\theta}^T\mathbf{x}_i)}{1+\exp(\boldsymbol{\theta}^T\mathbf{x}_i)}\right)\langle \boldsymbol{\theta}_t, \mathbf{x}_i\rangle \\ &+ \sum_{i\in\{i\in[n]:y_i=1\}} \left(1-\frac{\exp(\boldsymbol{\theta}^T\mathbf{x}_i)}{1+\exp(\boldsymbol{\theta}^T\mathbf{x}_i)}\right)\langle \boldsymbol{\theta}_t, \mathbf{x}_i\rangle \\ &> 0 \text{ as } \frac{\exp(\boldsymbol{\theta}^T\mathbf{x}_i)}{1+\exp(\boldsymbol{\theta}^T\mathbf{x}_i)} \in (0,1).\end{aligned}$$

(2 Points) In this way,

$$\begin{aligned}\|\boldsymbol{\theta}_{t+1}\|^2 &= \|\boldsymbol{\theta}_t\|^2 + \eta^2\|\nabla L(\boldsymbol{\theta}_t)\|^2 + 2\eta\langle \boldsymbol{\theta}_t, -\nabla L(\boldsymbol{\theta}_t)\rangle \\ &\geq \|\boldsymbol{\theta}_t\|^2\end{aligned}$$

We can conclude that the norm of $\boldsymbol{\theta}_{t+1}$ is greater than the norm of $\boldsymbol{\theta}_t$.