ESE 5460: Principles of Deep Learning Fall 2025

Homework 0

Release Date: August 27, 2025 *Due Date: September 03, 2025*

Name: Ravi Raghavan

PennKey: rr1133

PennID: 31109579

1 Written Questions

Problem 1.

Solution
(a) We define the following: $g(x) = \min_{y \in B} f(x, y)$ and $h(y) = \max_{x \in A} f(x, y)$. We see that

$$\forall x \in A, g(x) \leq f(x, y), \forall y \in B \tag{1}$$

and

$$\forall y \in B, f(x, y) \leq h(y), \forall x \in A \tag{2}$$

Overall, we come to the conclusion that

$$\forall x \in A \; \forall y \in B, \; g(x) \leq f(x, y) \leq h(y) \tag{3}$$

Therefore,

$$\max_{x \in A} \min_{y \in B} f(x, y) = \max_{x \in A} g(x) \leq \min_{y \in B} h(y) = \min_{y \in B} \max_{x \in A} f(x, y)$$

(b) We define the following: $g(x) = \inf_{y \in B} f(x, y)$ and $h(y) = \sup_{x \in A} f(x, y)$

We see that

$$\forall x \in A, \; g(x) \leq f(x, y), \; \forall y \in B \tag{4}$$

and

$$\forall y \in B, f(x, y) \leq h(y), \forall x \in A \tag{5}$$

Overall, we come to the conclusion that

$$\forall x \in A \; \forall y \in B, \; g(x) \leq f(x, y) \leq h(y) \tag{6}$$

This implies that

$$\sup_{x \in A} \inf_{y \in B} f(x, y) = \sup_{x \in A} g(x) \leq \inf_{y \in B} h(y) = \inf_{y \in B} \sup_{x \in A} f(x, y)$$

Problem 2.

Solution
The function given is $f(x_1, x_2) = 2x_1^2 - 1.05x_1^4 + \frac{1}{6}x_1^6 - x_1x_2 + x_2^2$.

(a)
To find the global minima, we first find the stationary points by setting the partial derivatives to zero.

$$\nabla f(x_1, x_2) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 4x_1 - 4.2x_1^3 + x_1^5 - x_2 \\ -x_1 + 2x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Using all the solutions, we see that $(0, 0)$ is global minima

(b) yes, $(\pm 1.74, \pm 0.87)$ and $(\pm 1.08, \pm 0.54)$

(c)

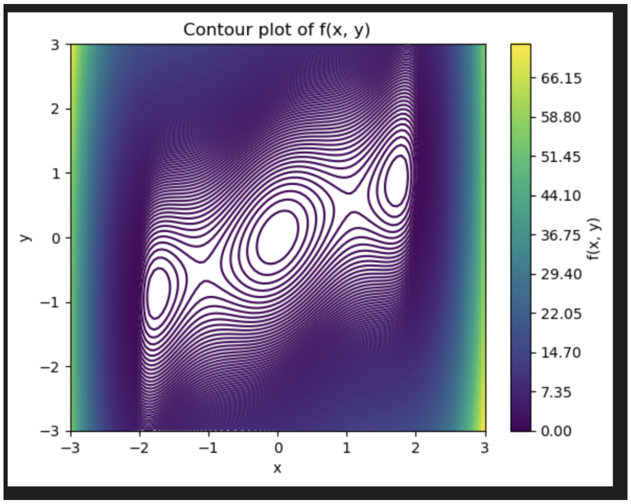

Figure 1: Contour

Problem 3.

Solution

We minimize

$$f(x, y) = x^2 + y^2 - 6xy - 4x - 5y$$

subject to the two constraints

$$(\text{C1}): \; y \leq -(x-2)^2 + 4, \qquad (\text{C2}): \; y \geq -x + 1.$$

We write (C1) in the standard form for optimization problems

$$g_1(x, y) := y + (x-2)^2 - 4 \leq 0,$$

and (C2) as

$$g_2(x, y) := -y - x + 1 \leq 0.$$

To solve this problem, we will use KKT conditions

(a)

We start by constructing the Lagrangian and its gradient

$$L(x, y, \lambda) = x^2 + y^2 - 6xy - 4x - 5y + \lambda_1 \left(y + (x-2)^2 - 4\right) + \lambda_2 \left(-y - x + 1\right) \tag{7}$$

$$\nabla L(x, y, \lambda) = \begin{bmatrix} 2x - 6y - 4 + 2\lambda_1 x - 4\lambda_1 - \lambda_2 \\ 2y - 6x - 5 + \lambda_1 - \lambda_2 \end{bmatrix} \tag{8}$$

Case # 1: Both inequality constraints are inactive
In this case, due to complementary slackness, $\lambda_1 = \lambda_2 = 0$, which makes the gradient

$$\nabla L(x, y, \lambda) = \begin{bmatrix} 2x - 6y - 4 \\ 2y - 6x - 5 \end{bmatrix} \tag{9}$$

Setting this to 0 gives us $x = \frac{-19}{16}$, $y = \frac{-17}{16}$, but this violates both constriants.

Case # 2: 1st inequality constraints is active
In this case, the gradient boils down to

$$\nabla L(x, y, \lambda) = \begin{bmatrix} 2x - 6y - 4 + 2\lambda_1 x - 4\lambda_1 \\ 2y - 6x - 5 + \lambda_1 \end{bmatrix} \tag{10}$$

We also have that $y + (x-2)^2 = 4$. Using this and setting the gradient to 0 gives us $\approx (2.7, 3.5)$ with a positive $\lambda_1 \approx 14.15$, which satisfies the other KKT constraints making it a feasible KKT point! The optimal loss value is around -65.6

Case # 3: 2nd inequality constraints is active In this case, the gradient boils down to

$$\nabla L(x, y, \lambda) = \begin{bmatrix} 2x - 6y - 4 - \lambda_2 \\ 2y - 6x - 5 - \lambda_2 \end{bmatrix} \tag{11}$$

We also have that $-y - x + 1$. Using this, and setting the gradient to 0, gives us that $(7/16, 9/16)$ but this gives a negative λ_2, which makes it infeasible.

(b) By ideas from local sensitivity analysis, we pose the problem as follows:

$$\lambda_1^* = \frac{-\partial f^\star(0,0)}{\partial u} \tag{12}$$

where $f^*(u, v)$ is the optimal value of f when perturbing the constraints $C1$ and $C2$ by u and v respectively.

Since $\lambda_1^* \approx 14.15$, we see that $\Delta f = -\lambda_1^* \cdot \Delta u = -14.15 * 0.1 = -1.415$

We expect the change in optimal value to be -1.415. Hence, we expect the optimal value to be around -67.015.

(c) Code uploaded and our result are verified!

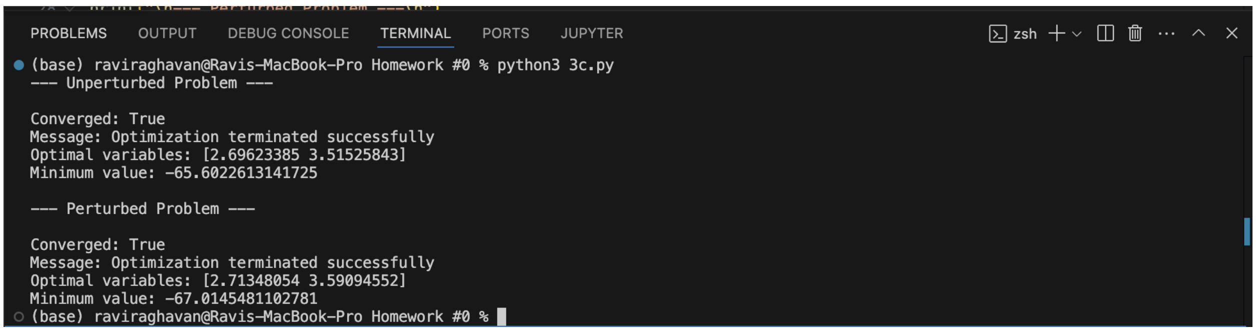

Figure 2: Verification of Results

Problem 4.

Solution

As per the problem statement, we know that $X, Y \in \{-1, +1\}$, with X and Y independent, $\mathbb{P}(X = 1) = q$ ($\mathbb{P}(X = -1) = 1 - q$), and Y is uniform on $\{-1, +1\}$. Subsequently, we know that $Z = XY \in \{-1, +1\}$.

(a) Conditional distribution $\mathbb{P}(Y \mid Z)$

By Bayes Rule, we proceed as follows

$$\mathbb{P}(Y = y \mid Z = z) = \frac{\mathbb{P}(Y = y, Z = z)}{\mathbb{P}(Z = z)} = \frac{\mathbb{P}(Y = y, X = z/y)}{\mathbb{P}(Z = z)} = \frac{\mathbb{P}(Y = y)\,\mathbb{P}(X = z/y)}{\mathbb{P}(Z = z)},$$

Note: Independence of X and Y is used in the last equality.

First note that Y is uniform, so $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = \frac{1}{2}$. Hence, $\mathbb{P}(Y = y) = \frac{1}{2}$. Also,

$$\mathbb{P}(Z = z) = \sum_{y \in \{\pm 1\}} \mathbb{P}(Y = y)\mathbb{P}(X = z/y) = \mathbb{P}(Y = 1)\mathbb{P}(X = z) + \mathbb{P}(Y = -1)\mathbb{P}(X = -z) = \tag{13}$$

$$\tfrac{1}{2}\big(\mathbb{P}(X = z) + \mathbb{P}(X = -z)\big). \tag{14}$$

We see that

$$P(Z = 1) = \tfrac{1}{2}\big(\mathbb{P}(X = 1) + \mathbb{P}(X = -1)\big) = \tfrac{1}{2}\big(q + 1 - q\big) = \tfrac{1}{2} \tag{15}$$

$$P(Z = -1) = \tfrac{1}{2}\big(\mathbb{P}(X = -1) + \mathbb{P}(X = 1)\big) = \tfrac{1}{2}\big(1 - q + 1\big) = \tfrac{1}{2} \tag{16}$$

Therefore

$$\boxed{\mathbb{P}(Y = y \mid Z = z) = \mathbb{P}(X = z/y) = \begin{cases} q, & z/y = +1, \\ 1 - q, & z/y = -1. \end{cases}}$$

Equivalently,

$$\mathbb{P}(Y = 1 \mid Z = z) = \mathbb{P}(X = z) = \begin{cases} q, & z = +1, \\ 1 - q, & z = -1, \end{cases} \qquad \mathbb{P}(Y = -1 \mid Z = z) = \mathbb{P}(X = -z) = \begin{cases} 1 - q, & z = +1, \\ q, & z = -1. \end{cases}$$

(b) Conditional mean $\mathbb{E}[Y \mid Z = z]$

Using part (a),

$$\mathbb{E}[Y \mid Z = z] = \sum_{y \in \{\pm 1\}} y\,\mathbb{P}(Y = y \mid Z = z) = \mathbb{P}(Y = 1 \mid Z = z) - \mathbb{P}(Y = -1 \mid Z = z).$$

Hence

$$\mathbb{E}[Y \mid Z = +1] = q - (1 - q) = 2q - 1, \qquad \mathbb{E}[Y \mid Z = -1] = (1 - q) - q = 1 - 2q.$$

Thus, compactly,

$$\boxed{\mathbb{E}[Y \mid Z = z] = (2q - 1)\, z.}$$

(c) Distribution of the random variable $\mu_{Y|Z} = \mathbb{E}[Y \mid Z]$

From (b),

$$\mu_{Y|Z} = (2q - 1)Z.$$

Since Z is uniform on $\{\pm 1\}$ (computed above in part (a)), $\mu_{Y|Z}$ takes two values,

$$\boxed{\mu_{Y|Z} = \begin{cases} +(2q - 1), & \text{with probability } \frac{1}{2}, \\ -(2q - 1), & \text{with probability } \frac{1}{2}. \end{cases}}$$

Problem 5.

Solution
By the triangle inequality theorem, we state the following:
Three positive lengths x_1, x_2, x_3 can form the sides of a triangle iff each side is less than or equal to than the sum of the other two. In other words, we need the following 3 conditions to hold

$$X_1 \leq X_2 + X_3 \tag{17}$$
$$X_2 \leq X_1 + X_3 \tag{18}$$
$$X_3 \leq X_1 + X_2 \tag{19}$$

The complement, that we cannot form a triangle, requires at least one of the above conditions to be FALSE. For convenience, we proceed to compute the probability of the complement event: that some X_i is at least the sum of the other two. Define

$$A_i = \{X_i \geq X_j + X_k\}, \qquad \{i, j, k\} = \{1, 2, 3\}.$$

These events are pairwise disjoint because two different coordinates cannot each be at least the sum of the other two simultaneously. Hence, by symmetry, we see that

$$\Pr(\text{no triangle}) = \Pr\big(A_1 \cup A_2 \cup A_3\big) = \sum_{i=1}^{3} \Pr(A_i) = 3\Pr(A_1),$$

We compute $\Pr(A_1) = \Pr(X_1 \geq X_2 + X_3)$. Using independence and the uniform density on $[0, 1]^3$,

$$\Pr(X_1 \geq X_2 + X_3) = \int_0^1 \Pr(X_2 + X_3 \leq x_1)\, dx_1.$$

For a fixed $x_1 \in [0, 1]$, the set $\{(x_2, x_3) \in [0, 1]^2 : x_2 + x_3 \leq x_1\}$ is a right triangle of area $\frac{1}{2}x_1^2$. Therefore

$$\Pr(X_1 \geq X_2 + X_3) = \int_0^1 \frac{1}{2}x_1^2\, dx_1 = \frac{1}{2} \cdot \frac{x_1^3}{3}\Big|_0^1 = \frac{1}{6}.$$

Thus

$$\Pr(\text{no triangle}) = 3 \cdot \frac{1}{6} = \frac{1}{2},$$

and therefore the probability that the three sticks *do* form a triangle is

$$\boxed{\Pr(\text{triangle}) = 1 - \Pr(\text{no triangle}) = 1 - \tfrac{1}{2} = \tfrac{1}{2}.}$$

Problem 6.

Solution

(a) The analytical expression for the weights $\mathbf{w}$ and the bias b that minimize the average of the squared residuals is derived from the loss function:

$$L(\mathbf{w}, b) = \frac{1}{2n} \sum_{i=1}^{n} (y^i - \mathbf{w}^T \mathbf{x}^i - b)^2$$

To simplify the derivation, we augment the feature vectors $\mathbf{x}^i$ with a constant term of 1 and the weight vector $\mathbf{w}$ with the bias b. This is something I learned in my ML course. Let

$$\hat{\mathbf{x}}^i = \begin{bmatrix} 1 \\ \mathbf{x}^i \end{bmatrix} \quad \text{and} \quad \hat{\mathbf{w}} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$$

The loss function can now be written in a more compact form using matrix notation. Let $\mathbf{X}$ be the design matrix where each row is an augmented feature vector $\hat{\mathbf{x}}^i$, and $\mathbf{Y}$ be the vector of target values.

$$L(\hat{\mathbf{w}}) = \frac{1}{2n} ||\mathbf{Y} - \mathbf{X}\hat{\mathbf{w}}||^2 = \frac{1}{2n} (\mathbf{Y} - \mathbf{X}\hat{\mathbf{w}})^T (\mathbf{Y} - \mathbf{X}\hat{\mathbf{w}})$$

To find the optimal $\hat{\mathbf{w}}$, we take the derivative of the loss function with respect to $\hat{\mathbf{w}}$ and set it to zero:

$$\frac{\partial L}{\partial \hat{\mathbf{w}}} = \frac{1}{2n} \frac{\partial}{\partial \hat{\mathbf{w}}} (\mathbf{Y}^T\mathbf{Y} - 2\hat{\mathbf{w}}^T\mathbf{X}^T\mathbf{Y} + \hat{\mathbf{w}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}) = 0$$

$$\frac{1}{2n} (-2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}) = 0$$

This leads to the normal equation:

$$\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^T\mathbf{Y}$$

Assuming the matrix $\mathbf{X}^T\mathbf{X}$ is invertible, we can solve for the optimal $\hat{\mathbf{w}}$:

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

This is the analytical expression for the weights and bias that minimize the mean squared error.

(b)

```
--- Experiment 1 ---
Training MSE: 0.5179
Validation MSE: 0.5559

--- Experiment 2 ---
Training MSE: 0.5281
Validation MSE: 0.5098

--- Experiment 3 ---
Training MSE: 0.5232
Validation MSE: 0.5348

--- Summary ---
Mean Training Error: 0.5231
Standard Deviation of Training Error: 0.0042
Mean Validation Error: 0.5335
Standard Deviation of Validation Error: 0.0188
```

Figure 3: Metrics