CIS3333: Mathematics of Machine Learning Fall 2025

Homework 10

Due: December 10, 2025, 11:59 PM ET

Submission Instructions: Submit a single PDF to Gradescope. Show key steps and justify your answers conceptually.

Collaboration & AI Policy: You may discuss approaches with classmates, but write up your own solutions and list collaborators. If you use computational tools (including LLMs) for checking, cite them and ensure the reasoning is your own.

Note: This is the final homework of the course.

Problem 1: Projected Gradient Descent (9 points)

Recall that Projected Gradient Descent (PGD) solves constrained optimization problems $\min_{x\in\mathcal{C}} f(x)$ via:

$$y_{t+1} = x_t - \eta \nabla f(x_t)$$
$$x_{t+1} = \Pi_{\mathcal{C}}(y_{t+1})$$

where $\Pi_{\mathcal{C}}(y) = \arg\min_{x\in\mathcal{C}} \|x - y\|_2$ is the Euclidean projection onto $\mathcal{C}$.

In lecture, we proved that projections onto *convex* sets are non-expansive. In this problem, we will compute projections for the box constraint set:

$$\mathcal{C}_{\ell,u} = \{x \in \mathbb{R}^n : \ell_i \leq x_i \leq u_i \text{ for all } i\},$$

for some vectors $\ell, u \in \mathbb{R}^n$ with $\ell_i \leq u_i$ for all i.

First let's prove that the box constraint set is convex. Recall that a set $\mathcal{C}$ is convex if for all $x, y \in \mathcal{C}$ and $t \in [0, 1]$, we have $tx + (1 - t)y \in \mathcal{C}$.

(a) (3 points) Show that $\mathcal{C}_{\ell,u}$ is convex.

Now let's compute the projections for these sets.

(b) (6 points) Show that the projection onto the box constraint is given coordinate-wise for $i = 1, \ldots, n$ is

$$\Pi_{\mathcal{C}_{\ell,u}}(y)_i = \min\big(u_i, \max(\ell_i, y_i)\big).$$

Hint: Separate the problem into n independent problems for each coordinate i.

Problem 2: Why Convexity Matters (6 points)

Consider the unit circle $\mathcal{C} = \{x \in \mathbb{R}^2 : \|x\|_2 = 1\}$ (a non-convex set).

(a) (2 points) Show that $\mathcal{C}$ is not convex.

Let $y = (0.1, 0)$ and $x = (-1, 0) \in \mathcal{C}$.

(b) (4 points) Compute $\Pi_{\mathcal{C}}(y)$, then compute $\|y - x\|_2$ and $\|\Pi_{\mathcal{C}}(y) - x\|_2$. Does the non-expansiveness property $\|\Pi_{\mathcal{C}}(y) - x\|_2 \leq \|y - x\|_2$ hold?

Problem 3: Affine Projections (9 points)

Recall from the linear algebra module that if U is a subspace of $\mathbb{R}^n$ then the orthogonal projection of y onto U is the same as the projection onto the constraint set U:

$$\Pi_U(y) = \min_{u \in U} \|y - u\|_2$$

In this problem, we will look at projecting onto **affine subspaces**. Recall that an affine subspace is a shifted subspace: $\mathcal{A} = \{x_0 + u : u \in U\}$ for some subspace U and point x_0.

(a) (3 points) Show that the projection of y onto $\mathcal{A}$ is:

$$\Pi_{\mathcal{A}}(y) = x_0 + \Pi_U(y - x_0)$$

Now let's see what kind of constraints we can model as affine subspaces. Consider the constraint set formed by a linear equation:

$$\mathcal{A}_{\text{linear}} = \{x \in \mathbb{R}^n : a^\top x = b\} \quad \text{where } a \in \mathbb{R}^n \text{ and } b \in \mathbb{R}.$$

(b) (6 points) Show that $\mathcal{A}_{\text{linear}}$ is an affine subspace by finding a point $x_0 \in \mathcal{A}_{\text{linear}}$ and a subspace U such that $\mathcal{A}_{\text{linear}} = x_0 + U$.

Hint: What happens when you subtract two points that both lie in $\mathcal{A}_{linear}$? This should help you find U. Then use that to find x_0.

Problem 4: Generalization Meets Optimization (6 points)

Consider training a model by minimizing empirical risk $\hat{R}(\theta)$ using gradient descent. There are two sources of error:

Generalization gap: The difference between true risk and empirical risk. With high probability,

$$|R(\theta) - \hat{R}(\theta)| \leq \frac{C_1}{\sqrt{N}} \quad \text{for all } \theta$$

where N is the number of training samples.

Optimization gap: GD doesn't find the exact minimizer. After T steps,

$$\hat{R}(\theta_T) - \hat{R}(\theta^*) \leq \frac{C_2}{T}$$

where $\theta^* = \arg\min_\theta \hat{R}(\theta)$.

(a) (2 points) Combine these to bound the excess true risk $R(\theta_T) - R(\theta^*)$.

(b) (4 points) If you have a total budget of B operations (where collecting one sample costs the same as one GD step), how should you split between N and T to minimize the bound from part (a)?

Bonus Problem: Teach Something

Create a short explainer (max 2 pages, or a Jupyter notebook) demonstrating one concept from this course that initially confused you or took time to understand. Email me your explainer by December 12, 2025, 11:59 PM ET.

Grade: This is worth up to 3% on your final grade to help you bump to the next grade. This will not be taken into account when deciding the threshold for the grades.

Note: I am looking for genuine effort and creativity. An explanation of what you struggled with and how you overcame it is more insightful than a textbook explanation.