CIS 3333: Mathematics of Machine Learning Fall 2025

Homework 6

Release Date: October 29, 2025 *Due Date: November 5, 2025*

Submission Instructions: Submit a single PDF (clear scans or photos compiled) to Gradescope and assign pages for each problem. Show key steps and justify answers.

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.

Problem 1: Representer Theorem (12 points)

In lecture, we proved the Representer Theorem for a regularizer of the form $\lambda\|f\|^2_{\mathcal{H}_K}$, which penalizes solutions with large norms and effectively shrinks the optimal function f^* towards the zero function.

In this exercise, we will prove a more general version where the regularizer incorporates a **prior** or **reference function** $h \in \mathcal{H}_K$. This function h represents a baseline belief about the solution, and the new regularizer, $\lambda\|f - h\|^2_{\mathcal{H}_K}$, penalizes functions that deviate from this prior.

The goal is to solve the following optimization problem:

$$f^* = \arg\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{N} \ell\left(f(x_i), y_i\right) + \lambda\|f - h\|^2_{\mathcal{H}_K} \right]$$

where $h \in \mathcal{H}_K$ is a fixed function.

Let $U = \operatorname{span}\{K_{x_1}, \ldots, K_{x_N}\}$. Then we can project both f and h onto this subspace and its orthogonal complement where $f = f_U + f_{U^\perp}$ and $h = h_U + h_{U^\perp}$.

Your task is to prove the theorem by following these steps, which closely mirror the proof from the lecture.

Step 1: Decompose the regularizer (4 points). Show that the norm decomposes along the orthogonal subspaces, i.e.,

$$\|f - h\|^2_{\mathcal{H}_K} = \|f_U - h_U\|^2_{\mathcal{H}_K} + \|f_{U^\perp} - h_{U^\perp}\|^2_{\mathcal{H}_K} \tag{1}$$

Step 2: Analyze the loss term (2 points). Using the reproducing property of $\mathcal{H}_K$, show that the predictions at the data points only depend on the projection onto U:

$$f(x_i) = f_U(x_i) \tag{2}$$

Conclude that the loss term $\sum_{i=1}^{N} \ell(f(x_i), y_i)$ is independent of the orthogonal component $f_{U^\perp}$.

Step 3: Combine and conclude (6 points). Use the results from the previous steps to argue that the optimal solution f^* must satisfy $f^*_{U^\perp} = h_{U^\perp}$. Then, show that this implies the solution must have the form:

$$f^* = \sum_{i=1}^{N} \alpha_i K_{x_i} + h \tag{3}$$

for some coefficients $\alpha_i \in \mathbb{R}$.

Problem 2: Application to Kernel Ridge Regression (13 points)

This problem will guide you through the derivation of the Kernel Ridge Regression (KRR) algorithm, which is a direct application of the Representer Theorem. The KRR objective is:

$$\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{N} (f(x_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}_K} \right]$$

Let $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N$ be the vector of labels and $\mathbf{X} = \begin{bmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{bmatrix} \in \mathbb{R}^{N \times D}$ be the data matrix. Let $\mathbf{K} \in \mathbb{R}^{N \times N}$ be the kernel matrix where $\mathbf{K}_{ij} = K(x_i, x_j)$ for all $i, j \in \{1, \dots, N\}$.

Step 1: Applying the Representer Theorem. By Representer Theorem, we know that the optimal solution f^* must have the form

$$f^*(x) = \sum_{i=1}^{N} \alpha_i^* K(x_i, x)$$

Let $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} \in \mathbb{R}^N$ be the vector of coefficients.

Step 2: Vectorize the Predictions (2 points). Show that the vector of predictions on the training data,

$$f(\mathbf{X}) = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{bmatrix} = \mathbf{K}\boldsymbol{\alpha}.$$

Note that we are overloading the notation f to mean both a function and a vector of function values.

Step 3: Vectorize the Regularizer (3 points). Show that the regularization term can be written in matrix form:

$$\|f\|^2_{\mathcal{H}_K} = \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}.$$

Step 4: Rewrite the Objective (3 points). Using the results above, show that the KRR objective can be rewritten as:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} \left[\|\mathbf{K}\boldsymbol{\alpha} - \mathbf{y}\|_2^2 + \lambda\boldsymbol{\alpha}^\top\mathbf{K}\boldsymbol{\alpha}\right]$$

Step 5: Solve for the Optimal Coefficients. By taking the gradient of your new objective with respect to $\boldsymbol{\alpha}$[1] and setting it to zero, we get the optimal coefficients:

$$\boldsymbol{\alpha}^* = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}.$$

Application: Linear Kernel (5 points). Now, consider the specific case of a linear kernel, $K(x_i, x_j) = x_i^T x_j$. First show that

$$\mathbf{K} = \mathbf{X}\mathbf{X}^\top$$

Now show that the KRR prediction function,

$$f^*(x) = \sum_{i=1}^{N} \alpha_i^* K(x_i, x)$$

is equivalent to a linear model

$$f^*(x) = \mathbf{x}^\top\mathbf{w}^* \text{ with } \mathbf{w}^* = \mathbf{X}^\top\boldsymbol{\alpha}^* = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{y}.$$

Connection to Least-Squares. If we solved the least-squares problem with regularization:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2,$$

we would get the solution:

$$\mathbf{w}^* = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Turns out that these two expressions are mathematically identical! Just two different ways of solving the same problem. *Bonus question: Why are these two expressions mathematically identical?*

[1] We will deal with matrix gradients in the next module.