

Homework 3

*Release Date: March 19, 2023**Due Date: April 1, 2023*

- HW3 will count for 10% of the grade. This grade will be split between the written (35 points) and programming (32 points) parts.
- All written homework solutions are required to be formatted using L^AT_EX. Please use the template [here](#). Do not modify the template. **This** is a good resource to get yourself more familiar with L^AT_EX, if you are still not comfortable.
- You will submit your solution for the written part of HW3 as a single PDF file via Gradescope. The deadline is **11:59 PM ET**. Contact TAs on Ed if you face any issues uploading your homeworks.
- Collaboration is permitted and encouraged for this homework, though each student must understand, write, and hand in their own submission. In particular, it is acceptable for students to discuss problems with each other; it is not acceptable for students to look at another student's written Solutions when writing their own. It is also not acceptable to publicly post your (partial) solution on Ed, but you are encouraged to ask public questions on Ed. If you choose to collaborate, you must indicate on each homework with whom you collaborated.
- **Bonus Questions:** We have added two bonus questions in this homework for extra credit. These are intended to be more challenging than the non-bonus homework questions.

Please refer to the notes and slides posted on the website if you need to recall the material discussed in the lectures.

1 Written Questions (35 points + 7 bonus points)

Problem 1: Kernels (13 points + 4 bonus points)

In this problem we will show that several algebraic operations preserve validity of a kernel.

1.1 (9 points) Consider two kernel functions k_1 and k_2 and their feature maps ϕ_1 and ϕ_2 respectively. In particular,

$$k_1(x, x') = \phi_1(x)^\top \phi_1(x') \text{ and } k_2(x, x') = \phi_2(x)^\top \phi_2(x').$$

For each of the following kernel functions, show that it is a valid kernel. Specifically, design a feature map ϕ using ϕ_1, ϕ_2 such that $k(x, x') = \phi(x)^\top \phi(x')$.

- (a) (2 point) $k(x, x') = c \cdot k_1(x, x')$ for any $c \geq 0$
- (b) (2 points) $k(x, x') = k_1(x, x') + k_2(x, x')$
- (c) (5 points) $k(x, x') = k_1(x, x') \cdot k_2(x, x')$

1.2 (4 points) Using the above properties, show that $k'(x, x') = \sum_{i=1}^d \alpha_i k(x, x')^i$ is a valid kernel for any valid kernel k if $\alpha_i \geq 0$ for all $i \in [d]$.

Bonus (4 points) Show that $k'(x, x') = \exp(\min(x, x'))$ is a valid kernel over input space $\mathcal{X} = [0, 1]$.

Problem 2: Gradient Descent (12 points)

Consider a training dataset $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ where for all $i \in [m]$, $\|x_i\|_2 \leq 1$ and $y_i \in \{-1, 1\}$. Suppose we want to run regularized logistic regression, that is, solve the following optimization problem: for regularization term $\mathcal{R}(w)$,

$$\min_w \frac{1}{m} \sum_{i=1}^m \log \left(1 + \exp \left(-y_i w^\top x_i \right) \right) + \mathcal{R}(w)$$

Recall: For showing that a twice differentiable function f is μ -strongly convex, it suffices to show that the hessian satisfies: $\nabla^2 f \succeq \mu I$. Similarly to show that a twice differentiable function f is L -smooth, it suffices to show that the hessian satisfies: $LI \succeq \nabla^2 f$. Here I is the identity matrix of the appropriate dimension.

2.1 (2 points) In the case where $\mathcal{R}(w) = 0$, we know that the objective is convex. Is it strongly convex? Explain your answer.

2.2 (3 points) In the case where $\mathcal{R}(w) = 0$, show that the objective is 1-smooth.

2.3 (1 point) What is the convergence rate of gradient descent on this problem with $\mathcal{R}(w) = 0$? In other words, what is the asymptotic number of iterations needed to get within ϵ of the minimum? *Note: Do not derive the convergence rate, just provide the rate in terms of number of iterations T .*

2.4 (5 points) Consider the following variation of the ℓ_2 norm regularizer called the weighted ℓ_2 norm regularizer: for $\lambda_1, \dots, \lambda_d \geq 0$,

$$\mathcal{R}(w) = \sum_{j=1}^d \lambda_j w_j^2.$$

Show that the objective with $\mathcal{R}(w)$ as defined above is μ -strongly convex and L -smooth for $\mu = 2 \min_{j \in [d]} \lambda_j$ and $L = 1 + 2 \max_{j \in [d]} \lambda_j$.

2.5 (1 point) What is the convergence rate of gradient descent on the regularized logistic regression problem with the weighted ℓ_2 norm penalty? In other words, what is the asymptotic number of iterations needed to get within ϵ of the minimum?

Note: Do not derive the convergence rate, just provide the rate in terms of number of iterations T .

Problem 3: SVM (10 points + 3 bonus points)

Consider running hard-margin kernel-SVM with the following kernel:

$$k(x, x') = \begin{cases} 1 & \text{if } x = x' \\ 0 & \text{otherwise.} \end{cases}$$

Assume that the training dataset is $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ such that all x_i are distinct (that is, for all $i \neq j$, $x_i \neq x_j$) and $y_i \in \{-1, 1\}$. Further assume that S is *balanced*, that is, the number of training examples with label 1 is the same as the number of training examples with label -1 (mathematically this means $|\{i : i \in [m], y_i = 1\}| = |\{i : i \in [m], y_i = -1\}| = m/2$ for even m).

3.1 (5 points) Recall that the dual objective is:

$$\begin{aligned} \text{maximize over } \alpha \quad & -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j k(x_i, x_j) + \sum_{i=1}^m \alpha_i \\ \text{such that} \quad & \sum_{i=1}^m \alpha_i y_i = 0 \\ & \forall i \in [m], \alpha_i \geq 0. \end{aligned}$$

Find the optimizer α_* of the above optimization problem with the given kernel and dataset.

3.2 (2 points) Using α_* from above show that the classifier learned is:

$$f(x) = \text{sign} \left(\sum_{i=1}^m y_i \mathbb{1}[x = x_i] \right).$$

3.3 (1 point) What is the error of the classifier on the training dataset?

3.4 (2 points) What is the prediction on any x not seen in the training dataset? What does this say about the generalization error of this classifier?

Bonus (3 points) Find the classifier learned when we remove the balanced assumption. Express it in terms of m_+ (the number of training examples with label 1) and m_- (the number of training examples with label -1).

2 Programming Questions (32 points + 2 bonus points)

Use the link [here](#) to access the Google Colaboratory (Colab) file for this homework. Be sure to make a copy by going to “File”, and “Save a copy in Drive”. As with the previous homeworks, this assignment uses the PennGrader system for students to receive immediate feedback. As noted on the notebook, please be sure to change the student ID from the default ‘99999999’ to your 8-digit PennID.

Instructions for how to submit the programming component of HW 3 to Gradescope are included in the Colab notebook. You may find this [PyTorch linear algebra reference](#) and this [general PyTorch reference](#) to be helpful in perusing the documentation and finding useful functions for your implementation.