CIS 4190/5190: Applied Machine Learning — Spring 2026

## Homework 1

*Handed Out: Feb 2* — *Due: Feb 16, 8 p.m.*

- You are encouraged to format your solutions using LaTeX. Handwritten solutions are permitted, but remember that you bear the risk that we may not be able to read your work and grade it properly — we will not accept post hoc explanations for illegible work. You will submit your solution manuscript for written HW 1 as a single PDF file.

- The homework is **due at 8 PM** on the due date. We will be using Gradescope for collecting the homework assignments. Please submit your solution manuscript as a PDF file via Gradescope. Post on Ed Discussion and contact the TAs if you are having technical difficulties in submitting the assignment.

- Make sure to assign pages to each question when submitting homework to Gradescope. The TA may deduct 0.2 points per sub-question if a page is not assigned to a question.

- Items marked [**5190 Only**] are mandatory for students enrolled in CIS 5190, and optional for CIS 4190.

# 1 Written Questions

Note: You do not need to show work for multiple choice questions. If formatting your answer in LaTeX, use our LaTeX template `hw_template.tex` (This is a read-only link. You'll need to make a copy before you can edit. Make sure you make only private copies.).

1. [Bias-Variance Tradeoff] (15 pts) Suppose we have an $L_2$-regularized linear regression model running until convergence, which has loss $L(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\big(f_{\boldsymbol{\beta}}(\mathbf{x}_i) - y_i\big)^2 + \lambda\|\boldsymbol{\beta}\|_2^2$. For each of the following, assuming it is the only variable changing, indicate whether it tends to increase **bias**, decrease bias, or keep bias the same, and similarly for **variance**. Consider bias and variance as the quality of model fit and sensitivity to data changes, rather than the magnitude of error. You will get full points if you get the answer right even without any explanation. If you get it wrong though, your explanations may be assigned partial points if they demonstrate partially correct reasoning.

   (a) Increase the number of training examples $n$

   (b) Decrease the regularization parameter $\lambda$

   (c) Decrease the dimension $d$ of the features $\phi(x) \in \mathbb{R}^d$

   (d) Increase $c$, where we replace the features $\phi(x)$ with $c \cdot \phi(x)$, for some $c \in \mathbb{R}_{>0}$ (for this part, assume no regularization, i.e., $\lambda = 0$)

   (e) Increase the gradient descent step size $\alpha$ (but not so much that gradient descent diverges)

(f) Suppose you fit a model and find that it has low loss on the training data but high loss on the test data; for each of the above five values $n$, $\lambda$, $d$, $c$, and $\alpha$, indicate whether you should increase or decrease it to reduce the test loss, or it has no impact on the test loss.

2. [Regularization/Sparsity] (7 pts) In class, we demonstrated the intuition behind $\ell_1$ and $\ell_2$ regularization. In this problem we will try to see why $\ell_1$ regularization creates sparsity (i.e. reduces some elements of $\beta$ to zero) from the perspective of gradient descent. As a reminder, here's the $\ell_1$ regularized linear regression objective.

$$\mathcal{L}_{\ell_1}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta^T x_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

(a) (3 pts) Write down the partial derivative of the $\ell_1$ and $\ell_2$ regularization term (i.e. second term) with respect to an individual weight $\beta_j$. You can ignore the case $\beta_j = 0$ where the gradient may be undefined. [Hint: Recall that the $l_1$ regularization term is $\mathcal{L}_{\ell_1}(\beta) = \lambda \sum_{j=1}^{d} |\beta_j|$, and the $\ell_2$ regularization term is $\mathcal{L}_{\ell_2}(\beta) = \lambda \sum_{j=1}^{d} \beta_j^2$.]

(b) (2 pts) Does $\ell_2$ regularization encourage sparsity (i.e., push some coefficients $\beta_j$ to exactly zero)? Briefly justify your answer using the gradient in the previous question.

(c) (2 pts) Does $\ell_1$ regularization encourage sparsity? (Analyze how the gradient of the second term in Equation 1 affects the gradient of the first term)

3. [Linear Regression] (17 pts) We are interested here in a linear regression problem with $n$ samples and $d$ features. The dataset has its features stored in a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with corresponding labels in the vector $\mathbf{y} \in \mathbb{R}^n$. The objective function for this problem can be written as:

$$J(\mathbf{w}) = \frac{1}{n}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}),$$

(a) (5 pts) Derive the closed-form solution for $\mathbf{w}^*$ that minimizes $J(\mathbf{w})$.

(b) (5 pts) Is there a closed-form solution for linear regression with L2 regularization? If yes, can you derive the formula of the closed-form solution for same? If not, please explain your answer.

(c) (2 pts) Is linear regression with no regularization guaranteed to have a unique solution for any dataset?

(d) **[5190 ONLY]** (5 pts) Show that in the special case of simple linear regression with an intercept, where $\hat{y}_i = w_0 + w_1 x_i$, the optimal parameters $(w_0^*, w_1^*)$ satisfy

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)(x_i - \bar{x}) = 0$$

4. [Linear Regression] (7 pts)

   Suppose we have a weight vector $\boldsymbol{w} = [w_1, w_2]^T$ with input vectors $\boldsymbol{x_n} \in \mathbb{R}^2$ and $y_n \in \{0, 1\}$ ($y = \boldsymbol{wx} = w_1 x_1 + w_2 x_2$). Let us initialize all the weights to be 0. Also, suppose we have $N = 2$ examples in our dataset: $(\boldsymbol{x_1} = [1, -1]^T, y_1 = 0), (\boldsymbol{x_2} = [-1, -1]^T, y_2 = 1)$. Show what happens in each iteration of the training process for an $l_2$ regularized ($\lambda = 1$) linear regression model with batch gradient descent (learning rate = 1) on the above dataset for two epochs (steps).

   (a) (2 pts) What is the value of the loss function at the beginning?

   (b) (5 pts) What is the final state of the trained weight vector after 2 steps, and the corresponding value of the loss function? (Hint: derive the partial derivative of the loss function with respect to weights, and calculate their values after each step)

$$L = \frac{1}{n} \Sigma_{i=1}^{n} (y_i - \boldsymbol{w x_i})^2 + \lambda ||\boldsymbol{w}||^2$$

5. [Gradient Descent] (4 pts) Suppose you are training a linear regression model on a small dataset using gradient descent. Consider the following two separate scenarios:

   (a) (2 pts) The training loss is oscillating and even increasing. What is the likely cause, and how would you address it?

   (b) (2 pts) After 1,000 iterations, the training loss is still decreasing at a steady rate. What does this suggest, and how would you address it?