Optimization for Machine Learning (Spring 2026): Homework 2

- LO4ML stands for the textbook *Linear Algebra and Optimization for Machine Learning*.
- You **must email your submission** as a **PDF file** to kbala@wsu.edu. You are welcome to write answers by hand and scan the pages. Put all the images on a PDF file, though.
- Your file name should identify you in the following manner. If you are Napoleon Dynamite, you should **name your submission NapoleonDynamite_Hw2.pdf**. If you want to add more bits to the title, e.g., Math565, you could name it NapoleonDynamite_Math565_Hw2.pdf, for instance. But you should **start the file name with NapoleonDynamite. And please avoid white spaces in the file name**.
- **Begin the SUBJECT of your email submission with the same `FirstnameLastname`, expression, e.g., "NapoleonDynamite Hw2 submission".**
- **This homework is due by 10:00 PM on Tuesday, February 17**.

1. In this problem, you will explore the basic gradient descent (GD) algorithm and how its convergence could depend on the learning rate (step size) α. Recall the basic GD update step: $\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla J$. Consider the loss function $J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T A \mathbf{w}$ for $\mathbf{w} \in \mathbb{R}^2$ and a symmetric positive semidefinite matrix $A \in \mathbb{R}^{2\times 2}$.

   (a) (15) For a given starting point $\mathbf{w}_0 = [1 \;\; 1]^T$ and learning rate $\alpha = 0.1$, specify two different instances A_s and A_f of A on which GD runs *slow* and *fast*, respectively. Define on your own the exact characteristics of slow and fast modes of convergence. Also create visualizations of the two loss function surfaces J_s and J_f using A_s and A_f, respectively, for the A matrix.

   (b) (15) Computationally demonstrate the convergence behavior of GD for both the slow and fast instances of J. Include any code you want to share in the your homework PDF file itself (no need to send them separately).

   (c) (10) How does the rate of convergence of GD depend on the eigenvalues of A, or of the related matrix in the update step of GD?

   (d) (10) Explain why the same learning rate α that works well for A_f causes issues for A_s.

2. (20) LO4ML Problem 4.8.2 from Page 187. This is the same result presented as Lemma 8 in Lecture 8. For this as well as other problems, you are encouraged to use the notation we have seen in class: $\mathbf{w}$ for the variable (parameter) vector and $\mathbf{x}_i$ for the data vector of the i-th instance, etc., rather than the notation used by LO4ML, e.g., W, X_i, etc.

3. (20) LO4ML Problem 6 from Page 200.

4. (25) LO4ML Problem 12 from Page 201.