CIS3333: Mathematics of Machine Learning Fall 2025

Homework 8

Due: November 19, 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.

Problem 1: First-Order Condition and Convexity (6 points)

Recall that a differentiable function f is convex if and only if it satisfies the first-order condition (i.e., it lies above all of its tangents):

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) \quad \text{for all } x, y$$

In lecture, we showed this condition follows from the definition of convexity.

(a) (6 points) Prove the other direction: that the first-order condition implies the definition of convexity. That is, if f satisfies the first-order condition, then for any x, y and $\alpha \in [0, 1]$:

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y).$$

Hint: For any x, y and $\alpha \in [0, 1]$, define an intermediate point $z = \alpha x + (1 - \alpha)y$. First try to get bounds on $f(x)$ and $f(y)$ relative to z. Then, combine these bounds to get the desired inequality.

Problem 2: Operations Preserving Convexity (12 points)

In lecture, we stated that certain operations preserve convexity. In this problem, you will prove these results. You should only need to use the definition of convexity to prove the results.

(a) (4 points) **Non-negative weighted sum.** Prove that if $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ are convex functions and $w_1, w_2 \geq 0$, then $f(x) = w_1 f_1(x) + w_2 f_2(x)$ is convex.

(b) (4 points) **Composition with affine function.** Prove that if $f : \mathbb{R}^m \to \mathbb{R}$ is convex, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$, then $g(x) = f(Ax + b)$ is convex.

Application: Ridge Regression. Consider the ridge regression objective function:

$$L(\theta) = \frac{1}{2}\|X\theta - y\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2$$

where $X \in \mathbb{R}^{N \times d}$, $y \in \mathbb{R}^N$, $\theta \in \mathbb{R}^d$, and $\lambda > 0$.

(c) (4 points) Prove that $L(\theta)$ is convex. *You may find it helpful to use the properties you proved in the parts above.*

Problem 3: Gradient Descent for Least Squares (12 points)

In this problem, we will show that the gradient descent algorithm converges to the optimal solution at a linear rate for the least squares problem.

Consider the ordinary least squares (OLS) objective

$$f(\theta) = \frac{1}{2}\|X\theta - y\|_2^2,$$

where $X \in \mathbb{R}^{N \times D}$ has full column rank and $y \in \mathbb{R}^N$. Let $H := X^\top X$ and note that $H \succ 0$. The unique minimizer is $\theta^* = H^{-1}X^\top y$.

The gradient and Hessian are

$$\nabla f(\theta) = X^\top(X\theta - y) = H\theta - X^\top y, \qquad \nabla^2 f(\theta) = H.$$

We set $\mu := \lambda_{\min}(H)$ and $L := \lambda_{\max}(H)$ such that $0 < \mu \le L$. Note that this implies that $f(\theta)$ is μ-strongly convex and L-smooth.

We will analyze the convergence of gradient descent with step size $\eta = 1/L$. Let θ_t be the iterate at iteration t with initialization being θ_0.

(a) (3 points) Recall that the optimal solution is $\theta^* = H^{-1}X^\top y$. Using the gradient descent update, show that for all $t \ge 1$,

$$\theta_t - \theta^* = (I - \eta H)(\theta_{t-1} - \theta^*).$$

Applying the above result iteratively, we get

$$\theta_t - \theta^* = (I - \eta H)^t(\theta_0 - \theta^*).$$

Now our goal is to bound the error $\|\theta_t - \theta^*\|_2^2$ after t iterations.

(b) (2 points) Show that $(I - \eta H)^t$ is symmetric. *Recall that a matrix A is symmetric if $A = A^\top$.*

Using the symmetry of $(I - \eta H)^t$, we can bound $\|\theta_t - \theta^*\|_2^2$ as follows:

$$\begin{aligned}
\|\theta_t - \theta^*\|_2^2 &= (\theta_t - \theta^*)^\top(\theta_t - \theta^*) \\
&= (\theta_0 - \theta^*)^\top \left((I - \eta H)^t\right)^\top \left((I - \eta H)^t\right)(\theta_0 - \theta^*) \\
&= (\theta_0 - \theta^*)^\top (I - \eta H)^{2t}(\theta_0 - \theta^*) \\
&\le \lambda_{\max}\left((I - \eta H)^{2t}\right)\|\theta_0 - \theta^*\|_2^2.
\end{aligned}$$

Here in the last inequality, we use the fact that for any symmetric matrix A, we have $v^\top Av \leq \lambda_{\max}(A)\|v\|_2^2$. Now let's bound the maximum eigenvalue of $(I - \eta H)^{2t}$.

(c) (4 points) Show that if λ is an eigenvalue of H, then $1 - \eta\lambda$ is an eigenvalue of $I - \eta H$. Using this show that $(1 - \eta\lambda)^{2t}$ is an eigenvalue of $(I - \eta H)^{2t}$.

(d) (2 points) Using the above result, show that

$$\lambda_{\max}\left((I - \eta H)^{2t}\right) \leq \left(1 - \frac{\mu}{L}\right)^{2t}.$$

Substituting this into the inequality befor and taking the square root, we get

$$\|\theta_t - \theta^*\|_2 \leq \left(1 - \frac{\mu}{L}\right)^t \|\theta_0 - \theta^*\|_2.$$

This shows that the error converges to zero exponentially fast and the rate depends on the ratio $\kappa := L/\mu$, also known as the condition number of the problem.

(e) (1 point) As κ gets larger, does the rate of convergence increase or decrease?