DSAI5104: Gradient Descent Tutorial

Problems

Problem 1: (LASSO and Regularization) Consider the LASSO problem:

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2}\|X\beta - b\|^2 + \lambda\|\beta\|_1,$$

where $X \in \mathbb{R}^{N\times n}$, $b \in \mathbb{R}^N$, and $\lambda > 0$.

(a) Prove that the objective function is convex but not differentiable everywhere. Specifically, identify which part causes non-differentiability and why.

(b) Calculate the gradient of the loss term $\frac{1}{2}\|X\beta - b\|^2$ with respect to β.

(c) Discuss why the non-differentiability of the ℓ_1-norm makes the optimization problem more challenging for standard gradient descent, and why ℓ_1-regularization encourages sparsity in solutions compared to ℓ_2-regularization.

Problem 2: (Convexity and Gradient Descent Directions) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable convex function. Suppose we are at a point x^k with $\nabla f(x^k) \neq 0$. Define the direction $d^k = -\nabla f(x^k)$.

(a) Using the mean-value theorem (as in lecture), prove that for sufficiently small step size $\alpha > 0$, we have $f(x^k + \alpha d^k) < f(x^k)$.

(b) Show that if f is also μ-strongly convex (i.e., $f(y) \geq f(x) + \nabla f(x)^\top(y - x) + \frac{\mu}{2}\|y - x\|^2$ for all x, y), then the decrease can be quantified as:

$$f(x^{k+1}) - f(x^*) \leq (1 - \mu\alpha)(f(x^k) - f(x^*))$$

for a suitable choice of α, where x^* is the unique minimizer. (Hint: Use the optimality condition $\nabla f(x^*) = 0$ and strong convexity.)

Problem 3: (Lipschitz Continuity and Gradient Bounds) A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be L-Lipschitz if $|f(x) - f(y)| \leq L\|x - y\|$ for all x, y.

(a) Let f be differentiable and L-Lipschitz. Prove that $\|\nabla f(x)\| \leq L$ for all x.

(b) Consider the Huber loss function with parameter $\delta > 0$:

$$h_\delta(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq \delta, \\ \delta(|x| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases}$$

Show that h_δ is L-Lipschitz for some L (find the smallest such L).

Problem 4: (Smoothness and Iteration Compute)

Let $n = 2$ and define

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ -2 \end{pmatrix},$$

and the quadratic function

$$f(x) = \frac{1}{2}x^\top Ax + b^\top x, \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2.$$

(a) Compute the gradient of f and show that ∇f is L-Lipschitz continuous, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^2.$$

By computing the eigenvalues of A, determine the smallest possible Lipschitz constant L.

(b) Let $x^k = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and choose the step size

$$\alpha = \frac{1}{L}, \qquad L = \lambda_{\max}(A).$$

Calculate one gradient descent step

$$x^{k+1} = x^k - \alpha \nabla f(x^k).$$

(c) You are encouraged to perform the gradient descent method in Python on the same function.

Problem 5: (Convergence Rate Comparison) Consider the two gradient descent setups from the lecture:

(a) For a convex, L-Lipschitz function, we use step size $\alpha = \epsilon/L^2$ and run for $T = \frac{D^2 L}{\epsilon^2}$ iterations, where $D = \|x^1 - x^*\|$. The average iterate $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x^t$ satisfies $f(\bar{x}) \leq f(x^*) + \epsilon$.

(b) For a convex, L-smooth function, we use step size $\alpha = 1/L$ and run for $T = \frac{LD^2}{2\epsilon}$ iterations. Then the last iterate x^T satisfies $f(x^T) \leq f(x^*) + \epsilon$.

Suppose we want an accuracy of $\epsilon = 10^{-4}$, with $L = 10$ and $D = 1$. Calculate the number of iterations required in each case. Discuss which case requires fewer iterations. What does this tell us about the importance of smoothness assumptions in optimization?

Problem 6: (Finite Termination for a Specific Quadratic Function)

Let $n = 2$ and define

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

and the quadratic function

$$f(x) = \frac{1}{2}x^\top A x + b^\top x, \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2.$$

(a) Compute the eigenvalues of A and determine the largest eigenvalue $L = \lambda_{\max}(A)$. Then write the gradient descent update with step size $\alpha = 1/L$ in the form:

$$x^{k+1} = Mx^k + c$$

for some matrix M and vector c (which you should specify explicitly).

(b) Starting from $x^0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, perform two gradient descent steps (compute x^1 and x^2 explicitly). Then verify that x^2 is exactly the minimizer $x^* = -A^{-1}b$ by computing x^* directly.

(c) Using the result from (b), explain why gradient descent with step size $\alpha = 1/L$ terminates in exactly two steps for this particular quadratic function, despite A not being a multiple of the identity matrix. (Hint: Consider the relationship between the eigenvalues of A and the matrix M in the gradient descent update.)