CS 15-759: Homework 2

Due: 2/7/2025, 11:59 PM on Canvas

Bonus points: If you find typos in my homework or lecture notes, please email me. You will earn +1 bonus points per typo found, and potentially more for especially egregious typos.

Hints: Hints are on the last page. It is recommended to think about the problem without hints for a while, and then look at the hints when stuck. The problems are meant to be difficult, so there is no shame in looking at the hints. If you make partial progress on problems (e.g., by following the hints) you will get partial points.

Problem 1: Continuity of Convex Functions (15 Points)

This is a problem that is similar in spirit to problem 2.

Problem: Prove that if $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then f is continuous. Use the following definitions of *convex* and *continuous.*

Convex: A function f is convex if $f(tx + (1-t)y) \le t \cdot f(x) + (1-t) \cdot f(y)$ for all $x, y \in \mathbb{R}^n$.

Continuous: A function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous if for all $x \in \mathbb{R}^n$ and $\varepsilon > 0$, there is some $\delta > 0$ depending on x and ε so that $|f(y) - f(x)| \le \varepsilon$ for all y satisfying $\|y - x\|_2 \le \delta$.

Problem 2: Smoothness and Strong Convexity (30 Points)

(a) Solve Exercise 6 in the Lecture Notes (equivalent notions of smoothness).

(b) Formally prove Theorem 3.5 in the Lecture Notes (gradient descent for smooth and strongly convexity functions with respect to a PSD matrix B). Take the definition of strong convexity and smoothness with respect to B to mean that $\mu B \preceq \nabla^2 f(x) \preceq LB$.

Problem 3: Practice with Cauchy-Schwarz (20 Points)

The Cauchy-Schwarz inequality states that for any real numbers $x_1, \dots, x_n$ and $y_1, \dots, y_n$ that

$$(x_1y_1 + \cdots + x_ny_n)^2 \le (x_1^2 + \cdots + x_n^2)(y_1^2 + \cdots + y_n^2).$$

Prove the following inequalities using the Cauchy-Schwarz inequality.

(a) For any positive real numbers $a_1, \ldots, a_n > 0$ and $b_1, \ldots, b_n > 0$ it holds that:

$$\sum_{i=1}^{n} \frac{a_i^2}{b_i} \geq \frac{\left(\sum_{i=1}^{n} a_i\right)^2}{\sum_{i=1}^{n} b_i}.$$

(b) Prove for any real numbers $x_0, x_1, \ldots, x_n$ that

$$(x_n - x_0)^2 \leq n \sum_{i=0}^{n-1} (x_{i+1} - x_i)^2.$$

Problem 4: Coordinate Descent (35 Points)

In this problem you will analyze an algorithm called *coordinate descent* on functions that have different smoothness parameters in different coordinates.

Definitions: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function with minimizer x^*. Let $e_i \in \mathbb{R}^n$ denote the standard basis vectors (the vector with a 1 in coordinate i, and the rest 0). For $i = 1, \ldots, n$ let $L_i \geq 0$ be *coordinate smoothness* parameters satisfying the following inequality:

$$|(\nabla f(x + \delta e_i))_i - (\nabla f(x))_i| \leq L_i |\delta| \quad \text{for all} \quad x \in \mathbb{R}^n, \delta \in \mathbb{R}.$$

Here $(\cdot)_i$ denotes the i-th coordinate of the vector.

Finally let $L = \sum_{i=1}^{n} L_i$ and for $i = 1, \ldots, n$ define $p_i = \frac{L_i}{L}$.

Consider the following algorithm, for an input point $x^{(0)} \in \mathbb{R}^n$. For $t = 0, 1, \ldots, T-1$:

1. Sample a coordinate $i \in [n]$ such that i is sampled with probability p_i.
2. Set $x^{(t+1)} = x^{(t)} - \frac{1}{2L_i}(\nabla f(x^{(t)}))_i$.

Prove that $\mathbb{E}[f(x^{(T)})] - \mathbb{E}[f(x^*)] \leq \frac{2LR^2}{T}$ where $R = \sup_{y : f(y) \leq f(x^{(0)})} \|y - x\|_2$. Note that this definition of R is a bit different from its definition in the case of smooth gradient descent than in the Lecture Notes.

Benefits of coordinate descent. While the number of steps of coordinate descent is often larer than that of gradient descent, the benefit is that in each iteration, only a single coordinate changes. Thus, even though there are more total iterations, the total time to do the iterations may be faster. This is useful when eg. parallel complexity is not a consideration.

1 Hints

Problem 1: First, solve the problem when $n = 1$. For the n-dimensional version for $n > 1$, the only new claim you need is that there is some ball around x on which the function is uniformly upper bounded. This can be proven by applying convexity again.

Problem 2(b): If you are having trouble, look at the analysis of Richardson iteration in Section 2.1.1 of the Lecture Notes.

Problem 4: First establish the following analogue of the quadratic upper bound for smooth functions:

$$f(x + \delta e_i) \leq f(x) + \delta(\nabla f(x))_i + \frac{L_i \delta^2}{2}.$$

Use this to upper bound $\mathbb{E}[f(x^{(t+1)})]$ in terms of $f(x^{(t)})$ and $\|\nabla f(x^{(t)})\|_2^2$. Now conclude by copying the analysis of smooth gradient descent in Lecture 4.

The following inequality may be useful: for a random variable X it holds that $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$. This is yet another consequence of the Cauchy-Schwarz inequality.

References