MATH 173A: Homework #3

Due on Oct 29, 2024 at 23:59pm

Professor Cloninger

Ray Tsai

A16848188

Problem 1

Determine whether each function is Lipschitz, and if so find the smallest possible Lipschitz constant for the function. For all problems, $\|\cdot\|$ represents the Euclidean norm (2-norm).

(a) $f : \mathbb{R}^n \to \mathbb{R}$ for $f(x) = \|x\|$

Proof. By the triangle inequality,

$$|f(x) - f(y)| = |\|x\| - \|y\|| \leq \|x - y\|.$$

Since the equality holds when $y = 2x \neq 0$, f is Lipschitz, with smallest possible Lipschitz constant being $L = 1$. □

(b) $f : \mathbb{R}^n \to \mathbb{R}$ for $f(x) = \|x\|^2$

Proof. Note that f is convex and differentiable. Since $\|\nabla f(x)\| = 2\|x\|$ is unbounded, f is not Lipschitz. □

(c) $\rho : \mathbb{R} \to \mathbb{R}$ for $\rho(x) = \frac{1}{1+e^{-x}}$.

Proof. Since $0 < \rho(x) < 1$ for all $x \in \mathbb{R}$,

$$|\rho'(x)| = \left|\frac{e^{-x}}{(1 + e^{-x})^2}\right| = |\rho(x)||(1 - \rho(x))| \leq \frac{1}{4},$$

where the equality holds when $x = 0$. Thus, ρ is Lipschitz with the smallest Lipschitz constant being $L = \frac{1}{4}$. □

(d) $f : \mathbb{R}^n \to \mathbb{R}$ for $f(x) = \rho(w^T x + b)$ for some weight vector $w \in \mathbb{R}^n$, $b \in \mathbb{R}$, and ρ from part (c).

Proof. Since $0 < \rho(x) < 1$ for all $x \in \mathbb{R}$,

$$\|\nabla f(x)\| = \rho'(x)\|w\| \leq \frac{1}{4}\|w\|,$$

where the equality holds when $x = 0$. Thus, f is Lipschitz with the smallest Lipschitz constant being $L = \frac{1}{4}\|w\|$. □

Problem 2

Let f be a convex and differentiable. Let x^* be the global minimum and suppose $x^{(0)}$ is the initialization such that $\|x^* - x^{(0)}\| \leq 5$.

(a) Let f be $L-$Lipschitz function where $L = 3$. Determine the step size μ and number of steps needed to satisfy

$$\left\| f\left(\frac{1}{t}\sum_{s=0}^{t-1} x^{(s)}\right) - f\left(x^*\right)\right\| \leq 10^{-4}.$$

Proof. By the rate of convergence theorem, putting $\mu = \frac{5}{3\sqrt{t}}$ yields

$$\left\| f\left(\frac{1}{t}\sum_{s=0}^{t-1} x^{(s)}\right) - f\left(x^*\right)\right\| \leq \frac{15}{\sqrt{t}},$$

and thus $t \geq 2.25 \times 10^{10}$ to satisfy the requirement. This makes the step size $\mu \leq \frac{1}{90000}$. □

(b) Let f be $L-$smooth where $L = 3$. Determine the step size μ and number of steps needed to satisfy

$$\left\| f\left(x^{(t)}\right) - f\left(x^*\right)\right\| \leq 10^{-4}.$$

Proof. Pick $\mu = \frac{1}{3}$. Then the gradient descent equation satisfies

$$\left\| f\left(x^{(t)}\right) - f\left(x^*\right)\right\| \leq \frac{5}{2t\mu} = \frac{15}{2t}.$$

Thus, $t \geq 7.5 \times 10^4$ to satisfy the requirement. □

Problem 3

Consider the function $f(x_1, x_2) = (2x_1 - 1)^4 + (x_1 + x_2 - 1)^2$.

(a) Find the global minimum of f, and justify your answer.

Proof. Note that $f(x_1, x_2) \geq 0$ as it is a sum of squares. Hence $(x_1, x_2) = (\frac{1}{2}, \frac{1}{2})$ is the global minimum of f as it achieves the minimum value of 0. $\square$

(b) Starting at $x^{(0)} = (0, 0)$, perform gradient descent with backtracking line-search.

i. Starting at $x^{(0)} = (0, 0)$ with learning rate $\mu^{(0)}$, write down the gradient descent equation for $x^{(1)}$.

Proof. Since $\nabla f(x) = \begin{bmatrix} 8(2x_1 - 1)^3 + 2(x_1 + x_2 - 1) \\ 2(x_1 + x_2 - 1) \end{bmatrix}$, we have

$$x^{(1)} = x^{(0)} - \mu^{(0)} \nabla f(x^{(0)}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \mu^{(0)} \begin{bmatrix} -2 \\ -10 \end{bmatrix} = \mu^{(0)} \begin{bmatrix} 2 \\ 10 \end{bmatrix}.$$

$\square$

ii. Suppose we want to set $\mu^{(0)}$ using backtracking line search with $\gamma = 0.2$ and Armijo's condition $f(x^{(1)}) \leq f(x^{(0)}) - \mu^{(0)} \gamma \|\nabla f(x^{(0)})\|_2^2$. Find a value of $\mu^{(0)}$ that satisfies this.

Proof. We already know $f(x^{(0)}) = 2$ and $\|\nabla f(x^{(0)})\|_2^2 = 104$. Computing $f(x^{(1)})$, we have

$$f(x^{(1)}) = \left(4\mu^{(0)} - 1\right)^4 + \left(2\mu^{(0)} + 10\mu^{(0)} - 1\right)^2.$$

By the Armijo's condition, we get

$$\left(4\mu^{(0)} - 1\right)^4 + \left(2\mu^{(0)} + 10\mu^{(0)} - 1\right)^2 \leq 2 - 20.8\mu^{(0)}.$$

Putting $\mu^{(0)} = 0.01$ satisfies the condition. $\square$

iii. Suppose instead you started with $\mu^{(0)} = 1$ and an update of $\mu^{(0)} \leftarrow \frac{1}{2}\mu^{(0)}$ (i.e. $\beta = \frac{1}{2}$). In the worst case, how many steps of back-tracking would you have to take before accepting $x^{(1)}$?

Proof. Define $g(\mu) = 2 - 20.8\mu - (4\mu - 1)^4 - (2\mu + 10\mu - 1)^2$. We then have

$$g(1) = -220.8, \quad g\left(\frac{1}{2}\right) = -34.4, \quad g\left(\frac{1}{4}\right) = -7.2, \quad g\left(\frac{1}{8}\right) = -0.9125, g\left(\frac{1}{16}\right) = 0.32109375.$$

Thus, we would have to take at most 4 steps of backtracking before accepting $x^{(1)}$. $\square$