ECE 3803, Fall 2021

Homework #4

Due Thursday, October 7, at 11:59pm

1. Prepare a one paragraph summary of what we talked about in class in the last week. I do not want just a bulleted list of topics, I want you to use complete sentences and establish context (Why is what we have learned relevant? How does it connect with other classes?). The more insight you give, the better.

2. Show that the following functions are convex. For each, indicate if it is strictly convex or not. [Hint: recall the second-order conditions of convexity from the notes.]
   (a) $f(x) = x^2$
   (b) $f(x) = e^{x^2}$
   (c) $f(x) = \log(1 + e^x)$

3. Consider $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$, where $\|\boldsymbol{x}\|$ denotes any valid norm.
   (a) Prove that $f(\boldsymbol{x})$ is convex using the basic properties of a norm.
   (b) Prove that no norm can be strictly convex.

4. (a) Consider the so-called "rectified linear unit" or ReLU activation function that is commonly used in neural networks:
   $$r(x) = \max(0, x).$$
   Show that $r(x)$ is convex.
   (b) Let $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ be convex functions on $\mathbb{R}^N$. Generalize the previous result by showing that
   $$f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\}$$
   is convex.
   (c) If $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ are convex, can you say anything about the convexity or concavity of
   $$f(\boldsymbol{x}) = \min\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\}?$$
   Sketch a one-dimensional example that supports your argument.

5. In class we showed that or a differentiable function f being convex is equivalent to the statement that
   $$f(\boldsymbol{x}) \geq f(\boldsymbol{y}) + \langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{y}) \rangle, \tag{1}$$
   for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$. Here we will provide another equivalent characterization of convexity for differentiable f. Specifically, the first-order condition in (1) is equivalent to the statement that
   $$\langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) \rangle \geq 0 \tag{2}$$
   for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$. This is often called the *monotone gradient* condition.

(a) Prove that (1) implies (2). (In fact, these are equivalent, but proving the other direction is a bit harder and I will not ask you to do this.)

(b) Consider a one-dimensional differentiable convex function $f(x)$. Assume that $f(x)$ has a unique global minimum $x^\star$. What does the above condition say about $f'(x)$ for $x > x^\star$? What about $f'(x)$ for $x < x^\star$?

6. A central focus when considering different optimization algorithms is the *rate of convergence*. It is often not enough to merely argue that the algorithm converges – we want to know how quickly it will do so, and the rate of convergence allows us to quantify this. In the problems below, we assume that $\{x_k\}$ is a sequence that converges to $x^\star$. We say that $\{x_k\}$ converges *linearly* with a rate of convergence of β if

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^\star|}{|x_k - x^\star|} = \beta$$

for some $\beta \in (0, 1)$. If $\beta = 1$ we say that $\{x_k\}$ converges *sublinearly*, and if $\beta = 0$ we say that $\{x_k\}$ converges *superlinearly*.

(a) Suppose that $\{x_k\}_{k=0}^\infty$ satisfies $|x_{k+1} - x^\star| \le \beta |x_k - x^\star|$ for some $0 < \beta < 1$ (and hence converges linearly with rate β). Prove that $|x_k - x^\star| \le \epsilon$ for all

$$k \ge \frac{\log\left(\frac{|x_0 - x^\star|}{\epsilon}\right)}{\log\left(\frac{1}{\beta}\right)}.$$

(b) Now suppose that $\{x_k\}_{k=0}^\infty$ satisfies $|x_k - x^\star| = \frac{k}{k+1}|x_{k-1} - x^\star|$. Is this convergence linear, sublinear, or superlinear? How large must k be to ensure that $|x_k - x^\star| \le \epsilon$

(c) A finer-grained distinction among different kinds of linear/superlinear convergence is the *order of convergence*. We say that $\{x_k\}$ converges with order q if

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^\star|}{|x_k - x^\star|^q} = \gamma$$

for some $\gamma > 0$ (not necessarily less than 1). For $q = 1$ this reverts to linear convergence, but $q = 2$ is called *quadratic* convergence, $q = 3$ *cubic* convergence, and so on. Note that q need be an integer. It might not be initially obvious, but quadratic convergence is *much* faster than linear convergence. Consider the two sequences defined by

$$x_k = \frac{1}{2^k} \qquad z_k = \frac{1}{2^{2^k}}.$$

Both converge to zero. Show that $\{x_k\}$ converges linearly and compute the rate. Show that $\{z_k\}$ converges quadratically. Submit a plot (on a log scale) that illustrates the difference in how quickly these converge to zero.

7. The bisection method is a strategy for one-dimensional problems of the form

$$\underset{x}{\text{minimize}}\ f(x) \quad \text{subject to} \quad x_l \le x \le x_u, \tag{3}$$

where $x_l, x_u \in \mathbb{R}$ satisfy $x_l < x_u$ and $f : \mathbb{R} \to \mathbb{R}$ is convex. If f is also differentiable, the bisection algorithm can be expressed as follows:

Set initial bounds: $a = x_l, b = x_u$
Initialize $k = 0$
while not converged **do**
    $x_k = (a + b)/2$
    if $f'(x) > 0$ **then**
        $b = x$
    else
        $a = x$
    end if
    $k = k + 1$
end while

(a) Provide an intuitive explanation for why this algorithm makes sense in light of the monotone gradient property of convex functions.

(b) Assume that f is strictly convex, and hence (3) has a unique solution, denoted $x^\star$. Let x_k denote the estimate provided by the bisection method after k iterations. Argue that x_k converges to $x^\star$.

(c) Determine whether x_k converges linearly, sublinearly, or superlinearly. If linear, compute the rate of convergence β.