Nonlinear Optimization 1, Fall 2023 - Homework 2
Due at 6PM on Friday 9/29 (Gradescope)

Your submitted solutions to assignments should be your own work. You are allowed to discuss homework problems with other students, but should carry out the execution of any thoughts/directions discussed independently, on your own. Acknowledge any source you consult.

Problem 1 - Smoothness

We saw in class that Lipschitz continuity of the gradient is key to proving convergence rates. In this exercise, you will learn more examples of functions with Lipschitz continuous gradients. Prove that the following functions have Lipschitz continuous gradients and provide a Lipschitz constant.

(a) (Composition with linear). Suppose that $f\colon \mathbb{R}^d \to \mathbb{R}$ has an L-Lipschitz gradient and let $A \in \mathbb{R}^{n\times d}$ be a matrix and let $b \in \mathbb{R}^n$ be a vector. Show that the function $x \mapsto f(Ax+b)$ has a Lipschitz continuous gradient, and find its Lipschitz constant. (Hint: use the chain rule to compute the gradient.)

(b) (Sums) Assume that $f_1, f_2\colon \mathbb{R}^d \to \mathbb{R}$ are functions with L_1 and L_2-Lipschitz gradients, respectively. Prove that the gradient of $f_1 + f_2$ is Lipschitz continuous, and determine its Lipschitz constant.

(c) (Squared norm). $f(x) = \frac{1}{2}\|x\|^2$.

(d) (Logistic loss). $f(x) = \sum_{i=1}^d \log(1+\exp(x_i))$.

(e) (Shifted norm) $f(x) = \sum_{i=1}^d \sqrt{1+x_i^2}$.

Problem 2 - Strong convexity and quadratic growth

Let $h\colon \mathbb{R}^d \to \mathbb{R}$ be a continuous μ-strongly convex, i.e., $x \mapsto h(x) - \frac{\mu}{2}\|x\|^2$ is convex. For this problem you will show that strongly convex functions have a unique minimizer and exhibit quadratic growth.

(a) Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a function that has a closed epigraph and bounded sublevel sets, i.e., for any $\alpha \in \mathbb{R}$ the set $\{x \in \mathbb{R}^d \mid f(x) \le \alpha\}$ is bounded. Show that f has at least one minimizer.

(b) Use what you proved in (a), to show that h has at least one minimizer $x^\star$.

(c) Show that the following inequality holds for all $x \in \mathbb{R}^d$

$$h(x^\star) + \frac{\mu}{2}\|x - x^\star\|^2 \le h(x),$$

and deduce that $x^\star$ is unique.

Problem 3 - Harder, Better, Faster, Stronger

For this problem you will analyze the convergence of a variant of Nesterov's accelerated gradient descent for strongly convex smooth functions. Let $h\colon \mathbb{R}^d \to \mathbb{R}$ be a continuous μ strongly convex function with L Lipschitz gradient.

Algorithm 1 `AcceleratedGradientDescent`

Input: Start point x_0, number of iterations T

Initialize $y_0 \leftarrow x_0$, $\lambda_0 \leftarrow 0$

Iterate (k $\leq T-1$) :

$$\lambda_{k+1} \leftarrow \frac{1+\sqrt{1+4\lambda_k^2}}{2}, \quad y_{k+1} \leftarrow x_k - \frac{1}{L}\nabla f(x_k), \quad x_{k+1} \leftarrow y_{k+1} + \left(\frac{\lambda_k - 1}{\lambda_{k+1}}\right)(y_{k+1} - y_k)$$

Output y_T

(a) In class we saw that Algorithm 1 achieves a rate of convergence of $f(y_k) - \min f \leq \frac{2L\|x_0 - x^\star\|^2}{k^2}$. Find an integer value $T^\star > 0$ (in terms of μ and L) such that if we set $k \geq T^\star$ we have

$$f(y_k) - \min f \leq (f(x_0) - \min f)/2.$$

(b) Consider the method that at each iteration updates:

$$z_{i+1} \leftarrow \texttt{AcceleratedGradientDescent}(z_i, T^\star)$$

starting with $z_0 = x_0$. Here $T^\star$ is the value you found in the previous question. That is, at each iteration of the algorithm we run an instance of Accelerated Gradient Descent for $T^\star$ iterations and then restart the method. This is known as "Restarted accelerated gradient descent." The total number of iterations of this algorithm is the sum of all the iterations of Accelerated Gradient Descent that the algorithm performs. Show that for any $\varepsilon > 0$, this algorithm finds a point $z \in \mathbb{R}^d$ such that $f(z) - \min f \leq \varepsilon$ after a total number of iterations of at most

$$\lceil 4\sqrt{L/\mu} \rceil \cdot \left\lceil \log_2 \left(\frac{f(x_0) - \min f}{\varepsilon} \right) \right\rceil \quad \text{iterations.}$$

Problem 4 - Fun with least squares

Consider the following least squares optimization problem

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2$$

for a given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

(a) Show that $x \mapsto \frac{1}{2}\|Ax - b\|^2$ has $\lambda_{\max}(A^\top A)$-Lipschitz gradient and is $\lambda_{\min}(A^\top A)$-strongly convex.

(b) Write a program that generates random A and b with i.i.d. normally distibuted entries $N(0,1)$ with $n = 1000$ and $m = 2000$.

(c) Implement and run 100 steps of Gradient Descent on you random problem using $x_0 = 0$ and the theoretically justified stepsize $\alpha_k = 1/\lambda_{\max}(A^\top A)$. Print out $\|\nabla f(x_k)\|$ at each iteration or generate a plot.

(d) Implement and run 100 steps of Accelerated Gradient Descent and the same stepsize $\alpha_k = 1/\lambda_{\max}(A^\top A)$. Print out $\|\nabla f(x_k)\|$ at each iteration or generate a plot. How does it compare with (b)?

(e) Implement and run 100 steps of Restarted Accelerated Gradient Descent from P3 (every 25 iterations, restart the above accelerated method, initialized at its last iterate). Print out $\|\nabla f(x_k)\|$ at each iteration or generate a plot. How does it compare with (c)?