AMS 553.761: Nonlinear Optimization I
Final, Fall 2020

- **There are 6 questions on this test.**
- **You have to upload your answers on Blackboard by 5:00 pm on Thursday, December 17, 2020. If something goes wrong with the upload you may email it to me before 5:00pm on Thursday, December 17, 2020. Please use the email option only if the upload does not work. No answers will be accepted after this deadline.**

  Please hand in ONE submission - multiple submissions will not be tolerated.
- **You are not allowed to discuss any problem with any other human being, except the instructor.**
- **You can use a computer only as a word processor; in particular, you cannot consult the internet in regards to this midterm. You CAN use the slides from class and books from the library.**
- **You CAN cite any result/inequality we have proved in class or from the HWs, without proof. If you cite a result (e.g., from a book) that was NOT mentioned in class or the HWs, you should include a complete proof of this fact.**
- **The level of rigor expected is the same as the HW solutions. Make sure you justify all your answers.**

1. **(25 pts)** Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a *convex*, L-smooth function (i.e., ∇f is a Lipschitz continuous with constant L) that has a minimizer $x^\star$. We wish to use a modified/quasi Newton linesearch method with Armijo backtracking to solve

$$\min_{x \in \mathbb{R}^n} f(x).$$

Recall that this means at every iteration, a positive definite matrix B_k is chosen and the search direction is set to $p_k := -B_k^{-1} g_k$, where g_k is the gradient at iterate x_k. The step size α_k is chosen by an Armijo backtracking search, with initialization α_{init} (see Algorithm 9 on slide 58 of the "Line Search Methods" slides). The other relevant parameters are $0 < \eta < 1$, which is used to define the Armijo condition (see condition (7) on slide 59 of the "Line Search Methods" slides), and $0 < \tau < 1$ as the backtracking parameter (see Algorithm 9 on slide 58 of the "Line Search Methods" slides). Assume further that $x^\star$ and all iterates x_k lie in a bounded set, i.e., their norm is at most some constant $R > 0$. Finally, we assume that the eigenvalues of B_k are uniformly bounded for all k: lower bounded by $\lambda_{\min}$ and upper bounded by $\lambda_{\max}$.

Prove that there exists a constant C such that for any $T \geq 1$, we have $f(x_T) - f(x^\star) \leq \frac{C}{T}$.

2. **(10 pts)** Recall the BFGS formula used in quasi-Newton methods. Given the current approximation $B_k \in \mathbb{R}^{n \times n}$ to the Hessian, and vectors $s, y \in \mathbb{R}^n$, define

$$B_{k+1} = B_k - \frac{1}{s^T B_k s}(B_k s)(B_k s)^T + \frac{1}{y^T s} y y^T.$$

Show that if B_k is positive definite, and $y^T s > 0$, then B_{k+1} is positive definite.

3. **(25 pts)** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function with a Lipschitz continuous Hessian map. Show that f is μ-strongly convex (for some $\mu > 0$), i.e., show that

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \leq f(y) \qquad \forall x, y \in \mathbb{R}^n \tag{1}$$

if and only if the smallest eigenvalue of $\nabla^2 f(x)$ is greater than or equal to μ for all $x \in \mathbb{R}^n$. [Hint: You might need the third order Taylor approximation bound that uses the Lipschitz constant of the Hessian map; see the "Background and basics" slides.]

4. Let $L \geq \mu > 0$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and μ-strongly convex, i.e., f satisfies (1). Let $x^\star$ be a minimizer for f.

   (a) **(10 pts)** Show that $f(x_k) - f(x^\star) \geq \frac{1}{2L}\|\nabla f(x_k)\|^2$. [Note: This inequality follows just from the L-smoothness property.]

   (b) **(15 pts)** Use part (a) and the μ-strongly convex property (1) to show that if we consider the iterates generated by standard steepest/gradient descent starting from an initial iterate x_0, i.e., generate iterates $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$, then the following holds

$$\|x_{k+1} - x^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)\|x_k - x^\star\|^2.$$

   Thus, conclude that after $T \geq 0$ iterations, $\|x_T - x^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)^T \|x_0 - x^\star\|^2$.

5. **(20 pts)** Let $L \geq \mu > 0$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and μ-strongly convex, i.e., f satisfies (1). Suppose we use random coordinate choice as a stochastic gradient oracle, with step lengths $\alpha_k = \frac{1}{nL}$. Thus, $x_{k+1} = x_k - \alpha_k a_k$ where a_k is equal to $n\frac{\partial f}{\partial x_i}e^i$ for $i \in \{1, \ldots, n\}$, chosen uniformly at random (as usual, e^i is the i-th standard unit vector, i.e., $e^i_j = 0$ if $j \neq i$ and $e^i_i = 1$). Show that for any $T \geq 1$,

$$\mathbb{E}[f(x_T) - f^\star] \leq \left(1 - \frac{\mu}{nL}\right)^T (f(x_0) - f^\star).$$

6. **(5 pts)** Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function and suppose $x_0, x_1, x_2 \ldots$ is a sequence of iterates for some algorithm. Suppose this sequence converges to a point $x^\star = \lim_{k\to\infty} x_k$. Suppose further that $\nabla f(x_k)$ converges to zero and the Hessian $\nabla^2 f(x_k)$ is positive definite for all k. Show that the limit $x^\star$ satisfies second order necessary conditions for optimality.

   Is it necessarily true that $x^\star$ also satisfies the second order *sufficient* conditions? Why or why not?