Nonlinear Optimization I, Fall 2021
Final
Due by midnight EST on 12/15
Please only in file formats .pdf, .zip, standard image formats (.jpg, .png, etc.).

Your submitted solutions to the final should be entirely your own work. Do not consult with anyone (other than the professor) or any online source about exam questions. Feel free to cite any result presented in class or in homeworks without proof and may access any material posted on the course blackboard.

You can write solutions by hand or type them up (the LaTeX code for this pdf is on blackboard).

Q1. (True/False and Why, Your Turn, 15pts) Invent three "True-False and Why" questions that illustrate a variety of topics from this course that you might put on this exam (yes, they must be different from any "True-False and Why" questions on the midterm). For each statement, clearly state whether it is true or false and then give your insightful reason why.

Your questions can vary in difficulty but should require more than citing a single definition or result from lecture. Tricky edge cases or unexpected counterexamples you ran into throughout the semester would be perfect.

Q2. (Smoothness, Strong Convexity and Preconditioning, 20pts) Consider minimizing a continuously differentiable $f\colon \mathbb{R}^d \to \mathbb{R}$ that has L-Lipschitz gradient and is μ-strongly convex.

(a) (5pts) Prove $L \geq \mu$.

(b) (5pts) If $L = \mu > 0$, show that there exists a minimizer x^* of f such that

$$f(x) = f(x^*) + \frac{L}{2}\|x - x^*\|_2^2 .$$

(c) (5pts) If $L = \mu > 0$, show that Gradient Descent reaches the optimal solution in one step. (Namely, for any point x, prove $x - \nabla f(x)/L$ globally minimizes f.)

(d) (5pts) If $L = \mu > 0$, show that Cyclic Coordinate Descent reaches optimal in d steps. (Namely, for any initialization x_1, prove that $x_{k+1} = x_k - (\nabla f(x_k)^T e_k)e_k/L$ has x_{d+1} globally minimizes f, where e_k is the kth standard basis vector[1].)

[1] $\nabla f(x_k)^T e_k$ is the kth partial derivative of f at x_k, denoted (confusingly) in the initial exam posting as $\frac{\partial}{\partial x_k} f(x_k)$.

Q3. (Limitations of our Convergence Theory, 30pts) Consider the following continuously differentiable minimization problem, attaining its minimum value of zero at the origin

$$\min_{x \in \mathbb{R}^d} f(x) = \|x\|_3^3 \tag{1}$$

where $\|x\|_3 = (\sum_{i=1}^d |x_i|^3)^{1/3}$.

(a) (10pts) Consider solving (1) using gradient descent with a constant stepsize $\alpha > 0$ from some initialization $x_0 \in \mathbb{R}^d$,

$$x_{k+1} = x_k - \alpha \nabla f(x_k) \ .$$

For what values of x_0 and α does this iteration converge and at what kind of rate[2]?
Explain the cause of any differences from the theory we developed in lecture.

(b) (10pts) Consider solving (1) using Newton's method from some initialization $x_0 \in \mathbb{R}^d$,

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k) \ .$$

For what values of x_0 does this iteration converge and at what kind of rate?
Explain the cause of any differences from the theory we developed in lecture.

(c) (5pts) Conjecture how you think an Accelerated Gradient method would perform here?
No rigorous proof is needed, but do justify your expectations.

(d) (5pts) Conjecture how you think a Quasi-Newton method would perform here?
No rigorous proof is needed, but do justify your expectations.

[2] State and justify whether it is sublinear, linear, superlinear, or quadratic.

Q4. (Trust Region Subproblem Optimization, 15pts) Consider the following constrained quadratic minimization

$$\begin{aligned} &\min_{s\in\mathbb{R}^d} g^T s + \frac{1}{2}s^T Bs \\ &\text{subject to } \|s\|_2 \le \Delta \end{aligned} \tag{2}$$

for some vector $g \in \mathbb{R}^d$, symmetric matrix $B \in \mathbb{R}^{d\times d}$ (not necessarily positive definite), and $\Delta > 0$. Recall a vector s^* is a minimizer of (2) if and only if $\|s^*\| \le \Delta$ and some $\lambda \ge 0$ satisfies the conditions

$$(B + \lambda I)s^* = -g \tag{3}$$

$$\lambda(\|s^*\|_2 - \Delta) = 0 \tag{4}$$

$$B + \lambda I \succeq 0\ . \tag{5}$$

(a) (5pts) Propose g, B, Δ such that exactly two different optimal solutions s_1^* and s_2^* to (2) exist.

(b) (5pts) Suppose some vector s^* has $\|s^*\|_2 \le \Delta$ and some $\lambda \ge 0$ exists satisfying (3), (4), and $B + \lambda I \succ 0$ (*strictly!*). Prove or disprove s^* must be the unique minimizer of (2).

(c) (5pts) Suppose some vector s^* has $\|s^*\|_2 \le \Delta$ and some $\lambda \ge 0$ exists satisfying (3), (4), and (5). Prove or disprove λ must be the unique number satisfying (3), (4), and (5).

Q5. (Trust Region Subproblem Optimization Continued, 20pts) Define the indicator function of the trust region constraint $S = \{s \mid \|s\|_2 \leq \Delta\}$ as

$$\delta_S(s) = \begin{cases} 0 & \text{if } s \in S \\ \infty & \text{if } s \notin S \ . \end{cases}$$

Then the trust region subproblem (2) can be equivalently viewed as the following minimization

$$\min_{s \in \mathbb{R}^d} g^T s + \frac{1}{2} s^T B s + \delta_S(s) \ . \tag{6}$$

(a) (5pts) Derive the following formula for the set of subgradients $\partial \delta_S(s)$ at any $s \in S$

$$\partial \delta_S(s) = \begin{cases} \{\lambda s \mid \lambda \geq 0\} & \text{if } \|s\|_2 = \Delta \\ \{0\} & \text{if } \|s\|_2 < \Delta. \end{cases}$$

(b) (5pts) Derive the following formula for the proximal operator $\text{prox}_{\alpha \delta_S}(s)$

$$\text{prox}_{\alpha \delta_S}(s) = \begin{cases} \Delta s / \|s\|_2 & \text{if } \|s\|_2 \geq \Delta \\ s & \text{if } \|s\|_2 < \Delta. \end{cases}$$

for any $\alpha > 0$ and $s \in \mathbb{R}^d$.

(c) (5pts) Based on these, state a necessary condition for a point s^* to be a local minimizer of (6). How does your condition relate to the necessary and sufficient conditions (3), (4), and (5)?

(d) (5pts) Consider applying proximal gradient descent to (6) with $\alpha = 1/\|B\|$ and $s_0 \in S$

$$s_{k+1} = \text{prox}_{\alpha \delta_S}(s_k - \alpha(g + B s_k)) \ .$$

Prove or disprove: The iterates s_k must converge to a global minimizer of (6)