AMS 553.761: Nonlinear Optimization I
Midterm, Fall 2020

- **There are 4 questions on this test.**
- **You have to upload your answers on Blackboard by 4:30 pm on Tuesday, October 13, 2020. If something goes wrong with the upload you may email it to me before 4:30pm on Tuesday, October 13, 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 we have mentioned in class or from the HWs without proof. If you cite a result (e.g., from a book) that was NOT mentioned in class, 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)** Give complete proofs of Lemmas 2.2, 2.3 and 2.6 from the lecture notes on "Conjugate Gradient".

2. **(25 pts)** Prove that the bisection method for Wolfe linesearch, i.e., Algorithm 13 from slide 76 in the "Line Search Methods" slides on the course webpage, terminates with a steplength satisfying the weak Wolfe conditions. Assume that the function f is bounded from below and has a gradient that is Lipschitz continuous with a global Lipschitz constant γ.

3. **(20 pts)** Prove Theorem 4.8 from slide 103 in the "Line Search Methods" slides on the course webpage, that establishes a global convergence rate of $O\big((\frac{1}{\epsilon})^2\big)$ for a modified or quasi Newton method with *Wolfe linesearch*. You may use Theorem 3.4 on slide 71 (due to Zoutendijk) without proof. [Recall that the condition number of B_k is given by $\frac{\lambda_{max}(B_k)}{\lambda_{min}(B_k)}$ which is the ratio of the largest eigenvalue of B_k to the smallest eigenvalue of B_k.]

4. **(10 pts)** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, not necessarily differentiable. Show that any local minimum is also a global minimum.

1. **(25 pts)** Give complete proofs of Lemmas 2.2, 2.3 and 2.6 from the lecture notes on "Conjugate Gradient".

Lemmas 2.2 9/10

Consider L is span of a set of orthorgonal basis $\{b_1, b_2, \cdots b_k\}$ and $f(\bar{x})$ is the minimizer of $f(x)$ restricted to $x_0 + L$, while $\nabla f(\bar{x}) = g(\bar{x})$ is not orthgonal to L w.r.t standard inner product. Thus

$$\langle g(\bar{x}), \sum_{i=1}^{k} \lambda_i b_i \rangle \neq 0$$

Then we can compute the projection of $g(\bar{x})$ into L:

$$\text{projection of } g(\bar{x}) = s' = \sum_{i=1}^{k} \frac{\langle g(\bar{x}), b_i \rangle}{\langle b_i, b_i \rangle} b_i$$

Thus, we can compute direction derivate of $f(\bar{x})$ w.r.t direction s'. $f(p) = \frac{1}{2} p^T A p - b^T p$ is quadratic function, which is differential continuously. Thus

$$f'(\bar{x}; s') = \nabla f(\bar{x}) \cdot s' = g(\bar{x})^T \cdot s' \neq 0$$

if $g(\bar{x})^T \cdot s' > 0$, then we set $s = -s'$ and conside $\phi(\lambda) = f(\bar{x} + \lambda s)$

$$\phi'(0) = g(\bar{x})^T s = -g(\bar{x})^T \cdot s' < 0$$

since f is continuas differentiable, so is ϕ, Thus there exists $0 < \delta < \varepsilon$ such that $\phi'(c) < 0$ for all $c \in (-\delta, \delta)$

By the Mean Value theorem:

$$\phi(\delta) = \phi(0) + \phi'(\xi)(\delta - 0)$$

for some $\xi \in (0, \delta)$ since $\phi'(\xi) < 0$ we have

$$f(\bar{x} + \delta s) = \phi(\delta) < \phi(0) = f(\bar{x})$$

If $g(\bar{x})^T s' < 0$, then we can set $s = s'$, by doing same proof, we can show that there must exist a point that make $f(x) < f(\bar{x})$, which is a contradition to our hypothesis that $f(\bar{x})$ is the minimizer of $f(p)$ restricted to $x_0 + L$. check $x \in x_0 + L$

Also. [if $\bar{x} \in x_0 + L$ and $s \in x_0 + L$, then $x_0 + \delta s \in x_0 + L$] (Due to the linearity of space). Thus. if $f(\bar{x})$ is the minimizer of $f(x)$ restricted to L, then $\nabla f(\bar{x})$ is orthogonal to L w.r.t the standard inner product.

Lemma 2.3 9/10

we assume lemma 2.6 is correct, which will be proved later, but we will ues it to prove lemma 2.3

The minimizer of $\phi(\alpha) = f(\bar{x} + \alpha s)$ is $\alpha = \frac{s^T(b - A\bar{x})}{\langle s, s \rangle_A}$

we consider $\bar{x}$ is the minimizer of $f(p)$ restricted to $x_0 + L$

Then $\bar{x}$ can be consider as a linear combination of basises of L which is $s_1, \cdots, s_k$. Thus we can represent $\bar{x} = \sum_{i=1}^{k} \lambda_k s_k$ where s_k are mutually A-conjugate. Thus

$$\alpha = \frac{s^T(b - A\bar{x})}{\langle s, s \rangle_A} = \frac{s^T b - s^T A\bar{x}}{\langle s, s \rangle_A} = \frac{s^T b - \langle s, \bar{x} \rangle_A}{\langle s, s \rangle_A}$$

$$\langle s, \bar{x} \rangle_A = \langle s, \sum_{i=1}^{k} \lambda_i s_i \rangle_A = \sum_{i=1}^{k} \lambda_i \langle s, s_i \rangle_A$$

$\because$ s is any vector that is A-conjugate to L

$\therefore$ $\langle s, s_i \rangle_A = 0 \quad \forall s_i$ for $i = 1, \cdots k$

$\therefore$ $\alpha_{\text{minimizer}} = \frac{s^T b}{\langle s, s \rangle_A}$

And when we want to find the minimizer of $f(p)$ restricted to $\text{span}(L \cup \{s\})$, and s is A-conjugate to L, then every point can be represented as linear combination of $\{s_1, \cdots, s_k, s\}$

Thus we consider a global minimizer of $f(p)$ restricted to $x^0 + \text{span}(L \cup \{s\})$ is $x' = \sum_{i=1}^{k} \lambda_i s_i + \alpha' s + x^0$

According to the lecture note, For each coefficient of basis, [it can be computed independtly due to the covexity of the quadratic function:]

$$\lambda_i = \frac{b^T s_i}{\langle s_i, s_i \rangle_A} \quad \text{and} \quad \lambda' = \frac{b^T s}{\langle s, s \rangle_A}$$

And it's see to see that the first k coordinate of x' should be same as $\bar{x}$ and

$$\lambda' = \frac{b^T s}{\langle s, s \rangle_A} = \frac{s^T b}{\langle s, s \rangle_A}$$

which is as same as λ minimizer and $\hat{x} = \bar{x} + \frac{s^T b}{\langle s, s \rangle_A} s$

Thus, it's clear to show that

$$x' \overset{\neq}{\equiv} \hat{x} \qquad x' \in \text{span}(L \cup \{s\}) \quad -1$$

$$\hat{x} \in x_0 + \text{span}(L \cup \{s\})$$

By having such proof, we can say that $\hat{x}$ is also the minimizer of $f(p)$ restricted to $\text{span}\{L \cup \{s\}\}$

Lemma 2.6 415

$$\phi(\alpha) = f(x+\alpha s) = \frac{1}{2}(x+\alpha s)^T A(x+\alpha s) - b^T(x+\alpha s)$$

$$= \frac{1}{2}(x+\alpha s)^T(Ax+\alpha As) - b^T(x+\alpha s)$$

$$= \frac{1}{2}(x^T+\alpha s^T)(Ax+\alpha As) - b^Tx - \alpha b^Ts$$

$$= \frac{1}{2}x^TAx + \frac{1}{2}\alpha x^TAs + \frac{1}{2}\alpha s^TAx + \frac{1}{2}\alpha^2 s^TAs$$

$$- b^Tx - \alpha b^Ts$$

$$= \frac{1}{2}x^TAx + \alpha x^TAs + \frac{1}{2}\alpha^2 s^TAs - b^Tx - \alpha b^Ts$$

$$\frac{\partial \phi(\alpha)}{\partial \alpha} = x^TAs + \alpha s^TAs - b^Ts$$

$$\frac{\partial^2 \phi(\alpha)}{\partial \alpha^2} = s^TAs$$

To find minimum of $\phi(\alpha)$ w.r.t α, we set $\frac{\partial \phi(\alpha)}{\partial \alpha} = 0$

$$\therefore \quad x^TAs + \alpha s^TAs - b^Ts = 0$$

$$\Rightarrow \quad \alpha \cdot \langle s, s\rangle_A = b^Ts - x^TAs = s^Tb - s^TAx$$

$$\Rightarrow \quad \alpha = \frac{s^T(b-Ax)}{\langle s, s\rangle_A}$$

also $\frac{\partial^2 \phi(\alpha)}{\partial \alpha^2} = s^TAs$, and A is a positive definite matrix

$\therefore s^TAs \geq 0$ for $\forall s \in \mathbb{R}^n$ $\therefore \frac{\partial^2 \phi(\alpha)}{\partial \alpha^2} \geq 0$

$\therefore \alpha = \frac{s^T(b-Ax)}{\langle s, s\rangle_A}$ is the minimizer of $\phi(\alpha)$ (*)

-1

25/25

2. **(25 pts)** Prove that the bisection method for Wolfe linesearch, i.e., Algorithm 13 from slide 76 in the "Line Search Methods" slides on the course webpage, terminates with a steplength satisfying the weak Wolfe conditions. Assume that the function f is bounded from below and has a gradient that is Lipschitz continuous with a global Lipschitz constant γ.

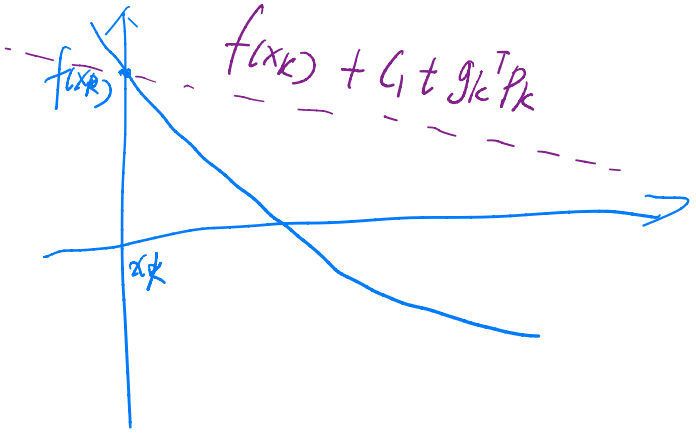
Answer:

The only situation that the Algorithm won't terminate is that

$f(x_k + t p_k) \le f(x_k) + c_1 t g_k^T p_k$ but $g(x_k + t p_k)^T p_k < c_2 (g_k^T p_k)$

In each iteration, α is doubled and β is always $= \infty$

which mean the function $f(x)$ is unbounded below and $g(x)^T p_k < 0$ for all x, which can be shown on following graph



But as said in the question, $f(x)$ is bounded below, this situation is impossible, $f(x_k) + c_1 t g_k^T p_k$ must intersect with $f(x)$ at least once.

If we assume $\alpha_k = 0$ for all k, which mean

$$f(x_k + t_k p_k) > f(x_k) + c_1 t_k g_k^T p_k \quad \forall k$$

$$\Rightarrow \frac{f(x_k + t_k p_k) - f(x_k)}{t_k} - c_1 g_k^T p_k > 0$$

When k become sufficiently large, due to each iteration

$t_k = \frac{\alpha_k + \beta_k}{2} = \frac{\beta_k}{2}$ and $\beta_k = t_{k-1}$

Thus $\lim_{k\to\infty} t_k = 0$

$$\therefore \lim_{k\to\infty} \frac{f(x_k + t_k p_k) - f(x_k)}{t_k} = g_k^T p_k$$

Then $\lim_{k\to\infty} g_k^T p_k \geq c_1 g_k^T p_k$, which is contradiction. Thus we now can assume that for specific k_0 when $k \geq k_0$, $\alpha_k > 0$ and $0 < \alpha_k < t_k < \beta_k < \infty$, while α_k, t_k, β_k are approching a specific value $\bar{t}$. Then we can compute following inequality:

$$f(x + \alpha_k p_k) \leq f(x) + c_1 \alpha_k g_k^T p_k \quad \text{—— ①}$$

$$f(x + \beta_k p_k) > f(x) + c_1 \beta_k g_k^T p_k \quad \text{—— ②}$$

$$\nabla f(x + \alpha_k p_k)^T p_k < c_2 \nabla f(x)^T p_k$$

$$\lim_{k\to\infty} \nabla f(x + \alpha_k p_k)^T p_k < c_2 \nabla f(x)^T p_k = \nabla f(x + \bar{t} p_k)^T p_k \leq c_2 \nabla f(x)^T p_k$$

By adding ① and ② together:

$$f(x + \alpha_k p_k) + f(x) + c_1 \beta_k g_k^T p_k \leq f(x) + c_1 \alpha_k g_k^T p_k + f(x + \beta_k p_k)$$

$$f(x + \beta_k p_k) - f(x + \alpha_k p_k) \geq c_1 (\beta_k - \alpha_k) g_k^T p_k \quad \text{— ③}$$

By using meen value theorem:

$$f(x + \beta_k p_k) - f(x + \alpha_k p_k) = (\beta_k - \alpha_k) \cdot \nabla f(x + \hat{t} p_k)^T p_k \quad \text{— ④}$$

when $\alpha_k \leq \hat{t} \leq \beta_k$

Combining ③ and ④, we can get:

$$(\beta_k - \alpha_k)\nabla f(x+\hat{t}p_k)^T p_k \geq c_1(\beta_k - \alpha_k)\nabla f(x)^T p_k$$

$$\Rightarrow \quad \nabla f(x+\hat{t}p_k)^T p_k \geq c_1 \nabla f(x)^T p_k$$

When $k \to \infty$, $\hat{t} \approx \bar{t}$, which yield

$$\nabla f(x+\bar{t}p_k)^T p_k \geq c_1 \nabla f(x)^T p_k$$

However, $\nabla f(x)^T p_k$ is supposed to be less then 0 for the reason that p_k is the descent direction of $f(x)$.

Thus with $0 < c_1 < c_2 < 1$:

$c_1 \nabla f(x)^T p_k$ is supposed to be larger than $c_2 \nabla f(x)^T p_k$

but in here:

$$\nabla f(x+\bar{t}p_k)^T p_k \geq c_1 \nabla f(x)^T p_k$$

$$\nabla f(x+\bar{t}p_k)^T p_k \leq c_2 \nabla f(x)^T p_k$$

which is a contradiction to our hypothesis. Thus we can say bisection method can find a step length to satisfy wolfe condition

3. **(20 pts)** Prove Theorem 4.8 from slide 103 in the "Line Search Methods" slides on the course webpage, that establishes a global convergence rate of $O((\frac{1}{\epsilon})^2)$ for a modified or quasi Newton method with *Wolfe linesearch*. You may use Theorem 3.4 on slide 71 (due to Zoutendijk) without proof. [Recall that the condition number of B_k is given by $\frac{\lambda_{max}(B_k)}{\lambda_{min}(B_k)}$ which is the ratio of the largest eigenvalue of B_k to the smallest eigenvalue of B_k.]

Answer:

Based on Wolfe linesearch, each iterat satify the following conditions

$$f(x_{k+1}) \leq f(x_k) - c_1 \alpha_k |g_k^T p_k| \quad -(1)$$

$$g_{k+1}^T p_k \geq c_2 g_k^T p_k \quad -(2)$$

while "∇f is lipschitz continuous on $\mathbb{R}^n$" implies that

$$(g_{k+1} - g_k)^T p_k \leq \alpha_k \gamma^2 \|p_k\|_2^2 \quad -(3)$$

From wolfe condition (2), we can get

$$(g_{k+1} - g_k)^T p_k \geq (c_2 - 1) g_k^T p_k \quad -(4)$$

Combining (3) and (4), we can get following relation

$$\alpha_k \geq \frac{c_2 - 1}{\gamma^2} \cdot \frac{g_k^T p_k}{\|p_k\|_2^2} \quad -(5)$$

By substituting (5) back to wolfe condition (1), we can obtain:

$$f(x_{k+1}) \leq f(x_k) - c_1 \frac{c_2 - 1}{\gamma^2} \frac{(g_k^T p_k)^2}{\|p_k\|_2^2} \quad -(6)$$

By summing iteration from 0 to T and set $c = c_1 \frac{c_2 - 1}{\gamma^2}$, we can obtain following inequality:

$$\left.\begin{aligned} f(x_{T+1}) &\le f(x_T) - c\frac{(g_T^T p_T)^2}{\|p_T\|_2^2} \\ &\vdots \\ f(x_1) &\le f(x_0) - c\frac{(g_0^T p_0)^2}{\|p_0\|_2^2} \end{aligned}\right\} \xrightarrow{\text{sum}} f(x_{T+1}) \le f(x_0) - c\sum_{k=0}^{T}\frac{(g_k^T p_k)^2}{\|p_k\|_2^2} \quad -(7)$$

While $f(x)$ is bounded below, we can rewrite (7) as

$$c\sum_{k=0}^{T}\frac{(g_k^T p_k)^2}{\|p_k\|_2^2} \le f(x_0) - f(x_{T+1}) \le M' \quad -(8)$$

Due to the modified Newton method ($B_k p_k = -g_k$), and B_k is a positive definite matrix, Thus

$$\lambda_{min}(B_k) \le \frac{s^T B_k s}{\|s\|^2} \le \lambda_{max}(B_k) \Leftrightarrow \lambda_{max}^{-1}(B_k) \le \frac{s^T B_k^{-1} s}{\|s\|^2} \le \lambda_{min}^{-1}(B_k)$$

Where $\lambda_{min}(B_k)$ and $\lambda_{max}(B_k)$ are the smallest and largest eigenvalue respectively of B_k, Thus:

$$|g_k^T p_k| = |g_k^T B_k^{-1} g_k| \ge \lambda_{max}^{-1}(B_k)\|g_k\|_2^2 \quad -(9)$$

$$\|p_k\|_2^2 = g_k^T B_k^{-2} g_k \le \lambda_{min}^{-2}(B_k)\|g_k\|_2^2$$

$$\hookrightarrow \|p_k\|_2 \le \lambda_{min}^{-1}(B_k)\|g_k\|_2 \quad -(10)$$

Combining (9) and (10), we can obtain:

$$\frac{|g_k^T p_k|}{\|p_k\|_2} \ge \frac{\lambda_{max}^{-1}(B_k)\|g_k\|_2^2}{\lambda_{min}^{-1}(B_k)\|g_k\|_2} = \frac{\lambda_{min}(B_k)}{\lambda_{max}(B_k)}\|g_k\|_2 = \frac{1}{cond(B_k)}\|g_k\|_2$$

While $cond(B_k)$ is bounded by β, Thus:

$$\frac{|g_k^T p_k|}{\|p_k\|_2} \geq \frac{1}{cond(B_k)} \|g_k\|_2 \geq \frac{1}{\beta} \cdot \|g_k\|_2$$

Thus

$$\frac{(g_k^T p_k)^2}{\|p_k\|_2^2} \geq \frac{1}{\beta^2} \cdot \|g_k\|_2^2 \quad \text{——} (11)$$

Combining (11) and (8), we can obtain following inequality:

$$c \sum_{k=1}^{T} \frac{1}{\beta^2} \|g_k\|_2^2 \leq c \sum_{k=0}^{T} \frac{(g_k^T p_k)^2}{\|p_k\|_2^2} \leq M'$$

Thus

$$\sum_{k=0}^{T} \|g_k\|_2^2 \leq \frac{\beta^2 M'}{c}, \quad \text{where } c = c_1 \frac{c_2 - 1}{\gamma^2} \text{ —— (12)}$$

while it's intuitively to prove that

$$\sum_{k=1}^{T} \|g_k\|_2^2 \geq (T+1) \cdot \min_{k=0,\cdots,T} \|g_k\|_2^2 \text{ —— (13)}$$

With (12) and (13), then

$$\min_{k=0,1,\cdots T} \|g_k\| \leq \sqrt{\frac{\beta^2 M'}{c(T+1)}} = \frac{M}{\sqrt{T+1}}$$

where $M = \sqrt{\frac{\beta^2 M'}{c}} = \beta\gamma^2 \sqrt{\frac{M'}{c_1(c_2-1)}}$

4. **(10 pts)** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, not necessarily differentiable. Show that any local minimum is also a global minimum.

Answer:

Consider $f(x^*)$ is a local minimum for $f: \mathbb{R}^n \to \mathbb{R}$ in x^* neighbor region $B(x, \varepsilon)$ where $\varepsilon \in \mathbb{R}^n$, Thus

$$f(x^*) < f(x), \quad \forall x \in B(x, \varepsilon)$$

and we consider a point $x' = (1-t)x^* + ty$ where y can be any point within domain $(f(x))$, $y \in$ domain (f), if we make t sufficiently small to make x' land in x^* neighbor, Thus

$$f(x^*) \leq f(x') = f((1-t)x^* + ty) \quad \text{— (1)}$$

While $f(x)$ is a convex function, it implies that

$$f((1-t)x^* + ty) \leq (1-t)f(x^*) + tf(y) \quad \text{— (2)}$$

Combining (1) and (2), we can get: 10/10

$$f(x^*) \leq (1-t)f(x^*) + tf(y)$$

$$\Rightarrow \quad tf(x^*) \leq tf(y)$$

$$\Rightarrow \quad f(x^*) \leq f(y) \qquad (0 < t \leq 1)$$

With the fact that y can be any point within the domain (f) we can conclude that $f(x^*)$ is the global minimizer of f

Thus, it can be proved that local minimizer of convex function is also a global minimizer.