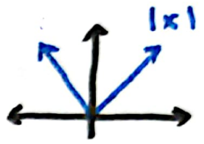
Nonlinear Optimization I

Fall 2021

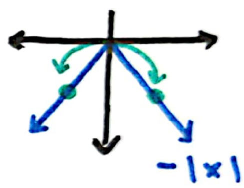
Midterm Sample Solutions



Q1 (a) $\boxed{\text{False}}$, for example, consider $f(x) = |x|$.
This minimizes at 0, but $\nabla f(0)$ does not exist.

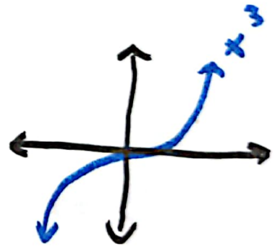
(True by necessary condition if we assume C^1)

(b) $\boxed{\text{False}}$, for example, consider $f(x) = -|x|$.
At $\bar{x} = 0$, prox can descend moving left or right $\text{prox}_f(0) = \{\pm 1\}$.



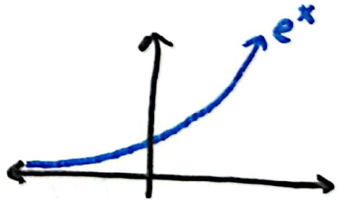
(True by strong convexity giving unique min existing if f is convex)

(c) $\boxed{\text{False}}$, for example, consider $f(x) = x^3$.
This is not locally minimized at 0, but has $f'(0) = f''(0) = 0$.



(True by 2^{nd}-order sufficient condition if we assume strict positive def).

(d) $\boxed{\text{False}}$, for example, consider $f(x) = e^x$.
$\inf e^x = 0$, but every $x \in \mathbb{R}$ has $e^x > 0$ (strictly).



(True by HW3, Q2 if we assume strong convexity)

(e) $\boxed{\text{True}}$, We are given $0 \in \partial f(\bar{x}^*) + \rho(\bar{x}^* - \bar{x})$

$$= \partial f(\bar{x}^*) + \nabla\left(\tfrac{\rho}{2}\|\cdot - \bar{x}\|_2^2\right)(\bar{x}^*)$$

(by Sum Rule) $$= \partial\left(f + \tfrac{\rho}{2}\|\cdot - \bar{x}\|_2^2\right)(\bar{x}^*)$$

$$\Rightarrow \forall y \quad f(y) + \tfrac{\rho}{2}\|y - \bar{x}\|_2^2 \geq f(\bar{x}^*) + \tfrac{\rho}{2}\|\bar{x}^* - \bar{x}\|_2^2 + 0^T(y - x).$$

$\Rightarrow \bar{x}^*$ globally minimizes.

(I did not specify $\rho > 0$. False would be a perfect answer if you caught my mistake omitting that. Full credit either way.)

$\boxed{\text{Q2}}$ (a) Note that $\nabla\left(f + \frac{1}{2\alpha}\|\cdot\|^2\right)(x) = \nabla f(x) + \frac{x}{\alpha}$.

Then for any $x, y \in \mathbb{R}^d$,

$$\left\|\left(\nabla f(x) + \tfrac{x}{\alpha}\right) - \left(\nabla f(y) + \tfrac{y}{\alpha}\right)\right\|_2$$

(by Triangle Inequality) $$\leq \|\nabla f(x) - \nabla f(y)\|_2 + \left\|\tfrac{x}{\alpha} - \tfrac{y}{\alpha}\right\|_2$$

(by Lipschitz assumption) $$\leq L\|x - y\|_2 + \tfrac{1}{\alpha}\|x - y\|_2$$

$$= (L + 1/\alpha)\|x - y\|_2. \qquad \square$$

(b) We prove this by showing the gradient at any $x \in \mathbb{R}^d$ of $f + \frac{1}{2\alpha}\|\cdot\|^2$ gives a quadratic lower bound:

By our Taylor-Approximation Theorem lower bound

$$\forall y \quad f(y) \geq f(x) + \nabla f(x)^T(y - x) - \tfrac{L}{2}\|y - x\|_2^2$$

(Adding & Subtracting $\frac{1}{2\alpha}\|\cdot\|^2$ at x and y)

$$\Rightarrow \forall y \quad f(y) + \tfrac{1}{2\alpha}\|y\|_2^2 \geq f(x) + \tfrac{1}{2\alpha}\|x\|_2^2 + \nabla f(x)^T(y - x) - \tfrac{L}{2}\|y - x\|_2^2 - \tfrac{1}{2\alpha}\|x\|_2^2 + \tfrac{1}{2\alpha}\|y\|_2^2$$

Lets simplify these last ~~four~~ three terms to get the result...

$$-\frac{L}{2}\|y-x\|_2^2-\frac{1}{2\alpha}\|x\|_2^2+\frac{1}{2\alpha}\|y\|_2^2$$

$$= \frac{(1/\alpha - L)}{2}\|y-x\|_2^2 + \frac{1}{2\alpha}\left(\|y\|_2^2 - \|x\|_2^2 - \|y-x\|_2^2\right)$$

$$= \frac{(1/\alpha - L)}{2}\|y-x\|_2^2 + \frac{1}{2\alpha}\left(-2\|x\|_2^2 + 2y^Tx\right)$$

$$= \frac{(1/\alpha - L)}{2}\|y-x\|_2^2 + \frac{1}{\alpha}x^T(y-x).$$

Plugging this back in, we have one of our equivalent characterizations of strong convexity:

$$\forall y \left(f(y)+\frac{1}{2\alpha}\|y\|_2^2\right) \geq \left(f(x)+\frac{1}{2\alpha}\|x\|_2^2\right) + \overbrace{\left(\nabla f(x)+\frac{x}{\alpha}\right)}^{=\nabla\left(f+\frac{1}{2\alpha}\|\cdot\|^2\right)(x)}{}^T(y-x)$$

$$+ \underbrace{\frac{(1/\alpha - L)}{2}}_{\text{claimed strong convexity constant.}}\|y-x\|_2^2. \quad \square$$

(c) Computing $\text{prox}_{\alpha f}(\bar{x})$ amounts to minimizing

$$f(x)+\frac{1}{2\alpha}\|x-\bar{x}\|_2^2$$

$$= \underbrace{f(x)+\frac{1}{2\alpha}\|x\|_2^2}_{\left(\frac{1}{\alpha}-L\right)\text{-strongly convex by (b)}} \underbrace{-\frac{1}{\alpha}\bar{x}^Tx+\frac{1}{2\alpha}\|\bar{x}\|_2^2}_{\text{An affine function of } x \Rightarrow 0\text{-strongly convex.}}$$

By Lemma in lecture, the sum of strongly convex functions is strongly convex (adding up their constants).

$\Rightarrow f+\frac{1}{2\alpha}\|\cdot-\bar{x}\|_2^2$ is $\frac{1}{\alpha}-L>0$ -strongly convex.

By HW3, Q2, this must have existence and uniqueness for its minimizer.

$\Rightarrow \text{prox}_{\alpha f}(\bar{x})$ is a singleton. $\square$

(d) Similar to the reasoning in (c), we also know from part (a), that the proximal problem's objective

$$f(x) + \frac{1}{2\alpha}\|x - \bar{x}\|_2^2$$

is $\left(\frac{1}{\alpha} + L\right)$-smooth (it has $\frac{1}{\alpha} + L$-Lipschitz gradient).

Our best algorithm for smooth, strongly convex minimization is our Restarted Accelerated Method from HW2.

(alternatively the modified accelerated method from Nesterov's book gives the same rate.)

Applying this method to minimize $h(x) := \left(f + \frac{1}{2\alpha}\|\cdot - \bar{x}\|_2^2\right)(x)$ with restarts every $\left\lceil 4\sqrt{\frac{1/\alpha + L}{1/\alpha - L}}\right\rceil$ steps will find a point y with

$$\overbrace{f(y) + \frac{1}{2\alpha}\|y - \bar{x}\|_2^2}^{h(y)} - \overbrace{\min_x f(x) + \frac{1}{2\alpha}\|x - \bar{x}\|_2^2}^{\min h(x)} \le \varepsilon$$

after at most $\left\lceil 4\sqrt{\frac{1/\alpha + L}{1/\alpha - L}}\right\rceil \left\lceil \log_2\left(\frac{h(x_0) - \min h(x)}{\varepsilon}\right)\right\rceil$ steps. □

Q3 Recall from lecture we have $\partial(|x|)(x) = \begin{cases} +1 & \text{if } x>0 \\ [-1,+1] & \text{if } x=0 \\ -1 & \text{if } x<0. \end{cases}$

for any scalar x.

First lets show the generalized version of this for the one-norm:

$$\partial(\|x\|_1)(x) = \left\{ g = \begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix} \;\middle|\; g_i \in \partial|\cdot|(x_i) \right\}$$

Proof. "$\subseteq$" consider $g \in \partial\|\cdot\|_1(x)$, then for any i, lets look at $y = x + \lambda e_i$ where $\lambda \in \mathbb{R}$ and e_i is the i^{th} basis vector.

$$\Rightarrow f(x+\lambda e_i) \geq f(x) + g^T(\lambda e_i), \text{ where } f = \|\cdot\|_1$$

$$\Leftrightarrow |x_i + \lambda| \geq |x_i| + \lambda g_i$$

$$\Rightarrow g_i \in \partial|\cdot|(x_i). \checkmark$$

"$\supseteq$" Suppose for each i, we have $g_i \in \partial|\cdot|(x_i)$.

$$\Rightarrow |y_i| \geq |x_i| + g_i^T(y_i - x_i) \quad \text{for any } y_i \in \mathbb{R}$$

$$\Rightarrow \sum |y_i| \geq \sum |x_i| + \sum g_i^T(y_i - x_i)$$

$$\Leftrightarrow \|y\|_1 \geq \|x_i\|_1 + g^T(y-x). \checkmark$$

$\square$

Then our Sum Rule ensures the LASSO objectives subgrad's are given by

$$\partial f(x) = A^T(Ax-b) + \gamma \cdot \partial\|\cdot\|_1(x).$$

$$= \left\{ A^T(Ax-b) + \gamma g \;\middle|\; g_i \in \partial|\cdot|(x_i) \right\}$$

$$= \left\{ \begin{bmatrix} A_1^T(Ax-b) + \gamma g_1 \\ \vdots \\ A_n^T(Ax-b) + \gamma g_n \end{bmatrix} \;\middle|\; g_i \in \partial|\cdot|(x_i) \right\}.$$

So the question is asking us to show x^* is a global minimizer of $f(x) = \frac{1}{2}\|Ax-b\|_2^2 + \gamma\|x\|_1$ if and only if

$$0 \in \partial f(x).$$

This is true for generic f as

$$\begin{aligned}
&x^* \text{ is a global minimizer} \\
&\Leftrightarrow \forall y \;\; f(y) \geq f(x^*) \\
&\Leftrightarrow \forall y \;\; f(y) \geq f(x^*) + 0^T(y - x^*) \\
&\Leftrightarrow 0 \in \partial f(x^*).
\end{aligned}$$

$\square$

$\boxed{\text{Q4}}$ (a) (Note: This is very similar to HW 3 Q2 (a))

For any x, strong convexity gives the lower bound

$$\begin{aligned}
f(y) &\geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2 \\
&= f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2 + \underbrace{\frac{\mu}{2}\left\|y - \left(x - \frac{1}{\mu}\nabla f(x)\right)\right\|_2^2}_{\geq 0} \quad \text{(completing the square)} \\
&\geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2
\end{aligned}$$

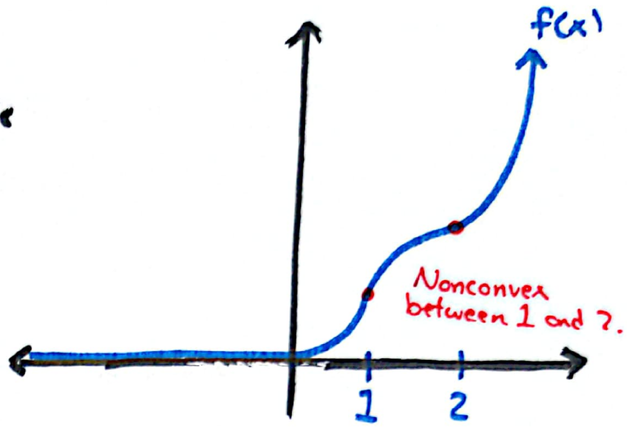
Minimizing both sides over y,

$$\min f(y) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2.$$

Rearranging this gives the claim.

$\square$

(b) First, a picture. We would like to have the derivative growing large as we move away from optimal. Something like...



We can build a piecewise quadratic function that looks like this.

Some simple polynomials also work here:
x^4+x^3+x works,
$x^6-x^4+\frac{1}{2}x^2$ works

$$f(x)=\begin{cases} 0 & \text{if } x\in[-\infty,0] \\ \frac{1}{2}x^2 & \text{if } x\in[0,1] \\ \frac{1}{2}+1\cdot(x-1)-\frac{1}{4}(x-1)^2 & \text{if } x\in[1,2] \\ \frac{5}{4}+\frac{1}{2}(x-2)+\frac{1}{2}(x-2)^2 & \text{if } x\in[2,\infty]. \end{cases}$$

This is differentiable with derivative squared of

$$(f'(x))^2=\begin{cases} 0 & \text{if } x\in[-\infty,0] \\ x^2 & \text{if } x\in[0,1] \\ \left(\frac{-1}{2}x+\frac{3}{2}\right)^2 & \text{if } x\in[1,2] \\ \left(-\frac{3}{2}+x\right)^2 & \text{if } x\in[2,\infty]. \end{cases}$$

by hand since they are quadratics or using wolframalpha.

Noting $\min f(x)=0$, we can check in each case $\mu=1/5$ has $(f'(x))^2\geq\mu(f(x)-0)$. (these are equal at $x=2$, so no larger μ would work).

(c) Our Descent Lemma ensures that we decrease

$$f(x_{k+1})\leq f(x_k)-\frac{1}{2L}\|\nabla f(x_k)\|_2^2 \qquad \text{by Descent Lemma}$$

$$\leq f(x_k)-\frac{\mu}{L}\left(f(x_k)-\min f(x')\right) \qquad \text{by Error Bound Assumption.}$$

Subtracting $\min f(x')$ from both sides gives the geometric recurrence relation

$$\delta_{k+1} \le \delta_k - \frac{\mu}{L}\delta_k = \left(1-\frac{\mu}{L}\right)\delta_k$$

where $\delta_k = f(x_k) - \min f(x')$.

Iteratively applying this gives the claim as

$$\delta_k \le \left(1-\frac{\mu}{L}\right)^k \delta_0 . \qquad \square$$

(d) We do still have $f(x_k) \to \min f(x')$. Similar to our previous proof, we have

$$f(x_{k+1}) \le f(x_k) - \frac{\mu^2}{2L}\left(f(x_k) - \min f(x')\right)^2$$

($\Rightarrow$ When $f(x_k) - \min f$ is large, we have a constant decrease)

In particular, we have the recurrence relation

$$\delta_{k+1} \le \delta_k - \frac{\mu^2}{2L}\delta_k^2 \le \underbrace{\delta_k - \frac{\mu^2}{2L}\delta_k \cdot \delta_{k+1}}_{\text{using that } \delta_k \text{ is decreasing}}.$$

We solved a recurrence of this form when we analyzed convex, smooth gradient descent. Namely, looking at this recurrence divided by $\delta_k \cdot \delta_{k+1}$ gives

$$\frac{1}{\delta_k} \le \frac{1}{\delta_{k+1}} - \frac{\mu^2}{2L}.$$

$\Rightarrow \frac{1}{\delta_k}$ increases by at least $\frac{\mu^2}{2L}$ per step.

$$\Rightarrow \frac{1}{\delta_K} \geq \frac{1}{\delta_0} + K \cdot \frac{\mu^2}{2L} \quad \text{by induction.}$$

$$\geq K \cdot \frac{\mu^2}{2L}$$

$$\Rightarrow f(x_K) - \min f(x') \leq \frac{2L}{\mu^2 K}. \qquad \square$$

$\boxed{Q5}$ (a) By the descent lemma, we know

$$\min f(x) \leq f(y_K - \tfrac{1}{L}\nabla f(y_K))$$

$$\leq f(y_K) - \frac{1}{2L}\|\nabla f(y_K)\|_2^2 . \quad \text{(by Descent Lemma)}$$

$$\Rightarrow \|\nabla f(y_K)\|_2^2 \leq 2L \cdot (f(y_K) - \min f(x))$$

$$\leq 2L \cdot \frac{2L\|x_0 - x^*\|^2}{K^2}. \quad \text{(by accelerated rate)}$$

Taking the squareroot of both sides gives the claim.

(b) Our nonconvex gradient descent guarantee established that

$$\frac{1}{K}\sum_{i=0}^{K-1} \|\nabla f(\bar{x}_i)\|_2^2 \leq \frac{2L(f(\bar{x}_0) - \min f(x))}{K}$$

$$\leq \frac{(2L)^2\|x_0 - x^*\|^2}{K \cdot K^2} \quad \text{(by accelerated rate } \bar{x}_0 = y_K\text{)}$$

Observing that the smallest $\|\nabla f(\bar{x}_i)\|^2$ is at most the average and taking a squareroot gives the claim.