1.

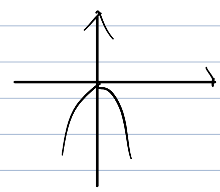
Statement: nonlinear optimization problems can have only one local optimal solution.

False, for example, consider $f(x) = e^x$, there is no local optimal solution at all

2.

Statement For a nonlinear optimization problem, if Newton's method converges, then it converges to a local minimum.

False, let $f(x) = -x^2$



and $x_0 = 1$

it converges,

then, Newton's method will stop at $(0,0)$ which is not a local minimum

3. Statement: there is no function could be both convex and concave.

False, affine function: $f(x) = ax + b$ could be both convex and concave.

Q_2.

a) from L-lipschitz, we have:

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\| \quad \forall, x, y$$

from u-strongly convex:

$$(\nabla f(y) - \nabla f(x))^T (y-x) \geqslant \mu \|y-x\|_2^2$$

so $L\|y-x\|_2^2 \geqslant \mu \|y-x\|_2^2$

so $L \geqslant \mu$

b) if $L = \mu$

From Hw3 Q_2 we know for $u>0$, u-strongly convex f have only one minimizer

we take the minimizer as x^*

since $f(x)$ is differentiable, $\nabla f(x^*) = 0$

$$f(x) \le f(x^*) + \nabla f(x^*)^T (x-x^*) + \frac{L}{2}\|x-x^*\|_2^2$$

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

for u-strongly convex,

$$f(x) \geq f(x^*) + \nabla f(x^*)^T (x-x^*) + \frac{\mu}{2}\|x-x^*\|_2^2$$

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

and $L = \mu$

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

c) from b) x^* is the unique minimizer,

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

so $\nabla f(x) = L(x-x^*)$

so $x - \nabla f(x)/L = x^*$, so it only takes one step

d) $\nabla f(x_1) = L(x_1 - x^*)$ $\qquad$ x_{ij} means jth element in x_i

$$x_2 = x_1 - \frac{\partial f}{\partial x_1}(x_1)\, e_1 / L = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1d} \end{bmatrix} - \begin{bmatrix} x_{11} - x_1^* \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} x_1^* \\ x_{12} \\ \vdots \\ x_{1d} \end{bmatrix}$$

Similarly $\quad x_3 = \begin{bmatrix} x_1^* \\ x_2^* \\ x_{13} \\ \vdots \\ x_{1d} \end{bmatrix}$ $\quad$ in each step, then ith element in x_i matches x_i^*

so it takes d steps to reach optimal

Q3.

(a) $f(x) = \sum_{i=1}^{d} |x_i|^3$, $\frac{\partial}{\partial x_j} f(x) = \frac{\partial}{\partial x_j} \sum_{i=1}^{n} |x_i|^3 = 3x_j^2$ if $x_j > 0$

$-3x_j^2$ if $x_j < 0$

from gradient descent, we have: $x_{k+1} = x_k - \alpha \nabla f(x_k)$

x_{k_i} = i-th element in x_k

and we know that $x^* = \vec{0}$,

so $\|x_{k+1} - x^*\|_2^2 = \sum_{i=1}^{d} (x_{k+1,i} - 0)^2 = \sum_{i=1}^{d} (x_{k,i} - 3\alpha x_{k,i}^2 \cdot \text{sign}(x_{ki}))^2$

$\le \max_{i=1\dots d}(1 - 3\alpha |x_{k_i}|)^2 \cdot \sum_{j=1}^{d} x_{k_j}^2$

we need $(1 - 3\alpha |x_{k_i}|)^2 < 1$ $i \in 1\dots d$ so $\alpha \in \left(0, \min_{i \in 1\dots d} \left(\frac{2}{3|x_{k_i}|}\right)\right)$

iteration coverge rate should be sublinear Since $\max_{i \in 1\dots d}(1 - 3\alpha |x_{k_i}|)^2$ will be closer and closer to 1, so it is sublinear.

$$\nabla^2 f(x) = 6 \begin{pmatrix} |x_1| & 0 & 0 & 0 & \cdots & 0 \\ 0 & |x_2| & 0 & 0 & \cdots & 0 \\ 0 & 0 & |x_3| & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & & \\ 0 & 0 & \cdots & & & |x_n| \end{pmatrix} \succeq 0$$

for all x, so it is a convex but unlike what we had in class, $f(x)$ here doesn't have L-Lipschitz gradient

b)

from (a) we can get $\nabla^2 f(x) = 6 \begin{pmatrix} |x_1| & 0 & 0 & 0 & \cdots & 0 \\ 0 & |x_2| & 0 & 0 & \cdots & 0 \\ 0 & 0 & |x_3| & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & & \\ 0 & 0 & \cdots & & & |x_n| \end{pmatrix}$

so $\nabla^2 f(x)$ is positive-semi-definite.

when $x_i \neq 0$ for all $i < d$ $\nabla^2 f(x)$ is positive definite, and will converge to $\nabla f(x') = 0$

Since only when $x = \vec{0}$, $\nabla f(x) = 0$, so when $x_{0_i} \neq 0$ for all $i \in 1\dots d$

and it converges at a quadratic rate

unlike what we develped in class, in this problem:

the Hessian matrix is singular at the optimal point

c) $\nabla^2 f(x) = 6\begin{pmatrix} |x_1| & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & |x_2| & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & |x_3| & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & & \cdots & & & \\ 0 & 0 & \cdots & \cdots & & & |x_n| \end{pmatrix} \succeq \mu I \quad \mu \le \min\{6|x_i|\}\ i \in 1 \cdots n$

when $\min|x_i|$ is much greater than 0, we can use accelerated gradient method.
but when x is close to $\vec{0}$, μ becomes very small, it will be similar to gradiene descent

So the converge rate is sublinear

d) Quasi-Newton method should converge linearly in this case
For Quasi-Newton method, when x_k is not close to the origin,
we can pick a μ mentioned in (c), it will converge fast
but when x_k is close to the origin, $\det(B_k)$ will be close to ∞

Q4 $B=-1$ $g=0$ $\Delta=1$ $\lambda=1$

a) so the problem turns into : $\min_{s\in R} -\frac{1}{2}s^2$ $|s|\le 1$

then, $s_1^*=1$ $s_2^*=-1$ so two optimal solutions exist

b) for a symmetric matrix we can have: $N=QMQ^T$

N is symmetric, Q is orthognal M is diagnal,

if $N>0 \Rightarrow M>0$ and all eigenvalues>0 so N is invertible. and its inverse is unique.

then, $B+\lambda I$ is invertible

so $s^* = -(B+\lambda I)^{-1}g$ is unique.

c) if λ is not unique, we can have λ' satisfies (3) (4) (5) with the same s^*

since $\lambda\neq\lambda' \Rightarrow$ one of $\lambda, \lambda' \neq 0$ $\Rightarrow \|s^*\|_2-\Delta=0$ assume $\lambda\neq 0$

if $B+\lambda I \geqslant 0$ and $\lambda\neq 0$, then, $B+2\lambda I>0$,

then, $g^Ts^*+\frac{1}{2}s^{*T}Bs^* = -\lambda s^{*T}s^* - \frac{1}{2}s^{*T}Bs^* = -\lambda'\Delta - \frac{1}{2}s^{*T}Bs^*$

So $\lambda=\lambda'$ so λ is unique

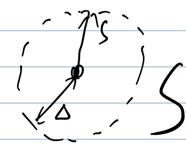
5. $\delta(s') \geq \delta(s) + g^T(s'-s)$ g is a subgradient

a) when $\|s\|_2 < \Delta$, $\delta_S(s) = 0$

if $\|s'\|_2 < \Delta$, then, $\delta(s') = \delta(s)$ so $g = \vec{0}$

if $\|s'\|_2 > 0$, then. $\delta(s) = \infty$, g could be any vector $\in R^d$

$\Rightarrow g = \vec{0}$



when $\|s\|_2 = \Delta$, $\delta_S(s) = 0$

if $\|s'\| > \Delta$ $\delta_S(s') = \infty$, g could be any vector $\in R^d$

goal: $\delta(s') \geq \delta(s) + g^T(s'-s)$

if $\|s'\| \leq \Delta$ we want $g^T(s'-s) \leq 0$ since $\delta(s') = \delta(s)$

for all the vectors $\{(s'-s)\}$ only vector in $\vec{s}$ direction satisfy $g^T(s'-s) \leq 0$

so $g = \lambda \cdot s$. $\lambda \geq 0$

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

(b) $\text{prox}_{\alpha \delta_S}(s) = \underset{s'}{\text{argmin}} \left\{ \delta_S(s') + \frac{1}{2\alpha}\|s'-s\|_2^2 \right\}$

$\delta_S(s') \geq 0$, $\frac{1}{2\alpha}\|s'-s\|_2^2 \geq 0$, so $\delta_S(s') + \frac{1}{2\alpha}\|s'-s\|_2^2 \geq 0$

when $\|s\|_2 < \Delta$, $\text{Prox}_{\alpha\delta_S}(s) = s$ since $\delta_S(s) + \frac{1}{2\alpha}\|s-s\|_2^2 = 0$

when $\|s\|_2 \geq \Delta$ $\underset{s'}{\text{argmin}} \left\{ \delta_S(s') + \frac{1}{2\alpha}\|s'-s\|_2^2 \right\}$ is the point in S which is closest to s.

so $s' = \Delta \cdot \frac{s}{\|s\|_2}$

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

c) let $J(s) = g^T s + \frac{1}{2} s^T B s + \delta(s)$

when s is a local minimizer,
subgradient of $J(s)$ should be parallel to s or $\vec{0}$
when $\|s\| < \Delta$, subgradint of $J(s)$ is $g + Bs = 0 \Rightarrow \lambda = 0 \Rightarrow (3)\ (4)$

when $\|s\| = \Delta$ subgradient of $J(s)$ is : $g + Bs + ks$ $\quad k \geq 0$
$= \lambda s + ks \Rightarrow$ parallel to s

$\uparrow$
used (3) (4)

d) We know that proximal gradient descent will go to a local minimizer
proximal operator always force $s^* \in S$, we can ignore $\delta(s)$ term
if B is positive definite or p.s.d $\quad$ (6) becomes a convex
$\Downarrow$
then, the local minimizer will be a global minimizer
if B is not pd or p.s.d,
(6)'s shape will be a saddle,
and global minimal will be at the boundary
when it does, (4) is satisfied,

from 4(c) we know λ is unique,

from 4(b) we know if $B+\lambda I > 0$ s^* is unique $\Rightarrow$ global minimizer

if $B+\lambda I \geq 0$, then all s^* have $(B+\lambda I)s^* = -g$

$$g^T s + \frac{1}{2} s^T B s = -\lambda s^{*T} s^* - \frac{1}{2} s^{*T} B s^*$$ $\frac{1}{2}B+\lambda Z$ could be deformed

$$= - s^{*T}(\frac{1}{2}B+\lambda I)s^*$$ as QMQ^T ← orthogonal

↑ diagnal

$$-s^{*T}(\frac{1}{2}B+\lambda Z)s^* = - s^{*T} Q M Q^T s^*$$

$\|s^*\|_2 < \Delta$ $M = diag(a_1, a_2 \cdots a_k, 0 \cdots 0)$ $a_1, a_2, \cdots a_k > 0$

on each dimension, it is a $a_i x_i^2$, so s^* might not be unique, but there value should be the same $\Rightarrow$ global minimizer