3.1. **Operator norms** $\boxed{1}$

Let E and F be normed spaces. Prove the following statements about linear functions $A, B \in L(E, F)$.

(a) $\|Ax\| \leq \|A\|\|x\|$.

(b) $\|AB\| \leq \|A\|\|B\|$.

(c) $\|A + B\| \leq \|A\| + \|B\|$.

Notation. For linear functions, it is common to omit the brackets around the argument. Thus Ax stands for $A(x)$. Moreover, composition of linear functions is commonly denoted by juxtaposition. Thus, AB stands for $A \circ B$. This notation stems from linear algebra and has been widely adopted in functional analysis and elsewhere.

3.2. **Convexity** $\boxed{1}$

Are the following functions $f : \mathbb{R}^2 \to \mathbb{R}$ convex, strictly convex, or uniformly convex?

(a) $f(x, y) = x + y$

(b) $f(x, y) = xy$

(c) $f(x, y) = x^2 + y^2$

(d) $f(x, y) = x^2 - y^2$

3.3. **Coercivity** $\boxed{1}$

Are the following functions $f : \mathbb{R}^2 \to \mathbb{R}$ coercive?

(a) $f(x, y) = x + y$

(b) $f(x, y) = xy$

(c) $f(x, y) = x^2 + y^2$

(d) $f(x, y) = x^2 - y^2$

3.4. **On the choice of learning rate** $\boxed{1}$

Compute and draw the first couple of steps of the gradient descent started at $x_0 = 1$ for the objective function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \frac{1}{2}x^2$ for the following learning rates α:

(a) $\alpha = 1$

(b) $\alpha = 2$

(c) $\alpha = \frac{1}{2}$

3.5. **Gradient of quadratic functions** $\boxed{1}$

Let A be a 2×2 matrix, and let

$$f : \mathbb{R}^2 \to \mathbb{R}, \qquad f(x) = x^\top A x.$$

Show that $\nabla f(x) = (A + A^\top)x$.

3.6. **Gradient descent for quadratic functions** $\boxed{1}$

Follow the instructions in the Jupyter notebook.

You can achieve a total of $\boxed{6}$ points for this problem set.

Author: Zhang Tianyu

Collaborator: Song Ruiqi

3.1. **Operator norms** $\boxed{1}$

Let E and F be normed spaces. Prove the following statements about linear functions $A, B \in L(E, F)$.

(a) $\|Ax\| \leq \|A\|\|x\|$.

(b) $\|AB\| \leq \|A\|\|B\|$.

(c) $\|A + B\| \leq \|A\| + \|B\|$.

Notation. For linear functions, it is common to omit the brackets around the argument. Thus Ax stands for $A(x)$. Moreover, composition of linear functions is commonly denoted by juxtaposition. Thus, AB stands for $A \circ B$. This notation stems from linear algebra and has been widely adopted in functional analysis and elsewhere.

(a) When $x=0$, $\|Ax\| = 0 \leq 0 = \|A\|\|x\|$.

when $x \neq 0$, $\|x\| \neq 0$, let $x' = \frac{1}{\|x\|} x$, s.t. $\|x'\| = 1$

By definition, $\|A\| \geq \|Ax'\| = \|A \frac{1}{\|x\|} x\| = \frac{1}{\|x\|} \|Ax\|$.

Hence, we have $\|A\| \geq \frac{1}{\|x\|} \|Ax\|$.

$\Leftrightarrow \|Ax\| \leq \|A\| \cdot \|x\|$.

(b). $\|AB\| = \sup\{ \|ABx\|, x \in E, \|x\| \leq 1 \}$

from (a), $\|ABx\| \leq \|A\| \cdot \|Bx\|$

Hence, $\|AB\| \leq \sup\{ \|A\| \cdot \|Bx\|, x \in E, \|x\| \leq 1 \}$

$= \|A\| \cdot \|B\|$

(c) $\|A+B\| = \sup\{ \|(A+B)x\|, x \in E, \|x\| \leq 1 \}$, by triangle inequality.

$\leq \sup\{ \|Ax\| + \|Bx\|, x \in E, \|x\| \leq 1 \}$

$= \|A\| + \|B\|$.

3.2. **Convexity**

Are the following functions $f : \mathbb{R}^2 \to \mathbb{R}$ convex, strictly convex, or uniformly convex?

(a) $f(x, y) = x + y$

(b) $f(x, y) = xy$

(c) $f(x, y) = x^2 + y^2$

(d) $f(x, y) = x^2 - y^2$

for $\lambda \in [0,1]$.

(a) $f(\lambda(x_1,y_1)+(1-\lambda)(x_2,y_2)) = \lambda(x_1+y_1)+(1-\lambda)(x_2+y_2) = \lambda f(x_1,y_1)+(1-\lambda)f(x_2,y_2)$

Convex, not strictly convex, not uniformly convex.

(b). $f(\lambda(x_1,y_1)+(1-\lambda)(x_2,y_2)) = f(\lambda x_1+(1-\lambda)x_2,\ \lambda y_1+(1-\lambda)y_2)$

$= \lambda^2 x_1 y_1 + \lambda(1-\lambda)(x_1y_2+x_2y_1)+(1-\lambda)^2 x_2y_2 \quad (*)$.

$\lambda f(x_1,y_1)+(1-\lambda)f(x_2,y_2) = \lambda x_1y_1+(1-\lambda)x_2y_2. \quad (**)$.

$** - *$: $\lambda(1-\lambda)x_1y_1 - \lambda(1-\lambda)(x_1y_2+x_2y_1)+\lambda(1-\lambda)x_2y_2$

$= \lambda(1-\lambda)(x_1y_1-x_1y_2-x_2y_1+x_2y_2) = \lambda(1-\lambda)(x_1-x_2)(y_1-y_2)$.

$**-*$ is negative when $x_1 > x_2$ and $y_1 < y_2$. Hence, not convex

(c). $f(\lambda(x_1,y_1)+(1-\lambda)(x_2,y_2)) = (\lambda x_1+(1-\lambda)x_2)^2 + (\lambda y_1+(1-\lambda)y_2)^2$

$= \lambda^2x_1^2+\lambda^2y_1^2+(1-\lambda)^2x_2^2+(1-\lambda)^2y_2^2+2\lambda(1-\lambda)(x_1x_2+y_1y_2)$ $(*)$

$\lambda f(x_1,y_1)+(1-\lambda)f(x_2,y_2) = \lambda x_1^2+\lambda y_1^2+(1-\lambda)x_2^2+(1-\lambda)y_2^2$. $(**)$

$**-*$: $\lambda(1-\lambda)[x_1^2+y_1^2+x_2^2+y_2^2-2x_1x_2-2y_1y_2]$

$= \lambda(1-\lambda)[(x_1-x_2)^2+(y_1-y_2)^2]$

$= \lambda(1-\lambda)\cdot \|(x_1,y_1)-(x_2,y_2)\|^2 \geq 0$. convex

when $(x_1,y_1) \neq (x_2,y_2)$. $**-* > 0$. Hence, strictly convex

let $\phi(t) = t^2$, $\mathbb{R} \to \mathbb{R}$

$f(\lambda(x_1,y_1)+(1-\lambda)(x_2,y_2)) \leq \lambda f(x_1,y_1)+(1-\lambda)f(x_2,y_2) - \lambda(1-\lambda)\phi(\|(x_1,y_1)-(x_2,y_2)\|$

Thus, f is convex, strictly convex. and uniformly convex.

(d) $f(\lambda(x_1,y_1)+(1-\lambda)(x_2,y_2))=(\lambda x_1+(1-\lambda)x_2)^2-(\lambda y_1+(1-\lambda)y_2)^2$

$$=\lambda^2x_1^2-\lambda^2y_1^2+(1-\lambda)^2x_2^2-(1-\lambda)^2y_2^2$$
$$+2\lambda(1-\lambda)(x_1x_2-y_1y_2) \quad (*)$$

$$\lambda f(x_1,y_1)+(1-\lambda)f(x_2,y_2)=\lambda x_1^2-\lambda y_1^2+(1-\lambda)x_2^2-(1-\lambda)y_2^2 \quad (**)$$

$** - *$: $\lambda(1-\lambda)[x_1^2-y_1^2+x_2^2-y_2^2-2x_1x_2+2y_1y_2]$

$= \lambda(1-\lambda)[(x_1-x_2)^2-(y_1-y_2)^2] < 0$ when $x_1=x_2=y_1=0,\ y_2>0$

Hence, not any kind of convex

3.3. **Coercivity** $\boxed{1}$

Are the following functions $f:\mathbb{R}^2\to\mathbb{R}$ coercive?

(a) $f(x,y)=x+y$

(b) $f(x,y)=xy$

(c) $f(x,y)=x^2+y^2$

(d) $f(x,y)=x^2-y^2$

(a). $f(0,y)=y$. when $y\to-\infty$, $\|(0,y)\|\to\infty$, $f(0,y)\to-\infty$

Not coercive.

(b) $f(1,y)=y$, when $y\to-\infty$, $\|(1,y)\|\to\infty$, $f(1,y)\to-\infty$

Not coercive.

(c). $f(x,y)=x^2+y^2=\|(x,y)\|^2$, $\lim_{\|(x,y)\|\to\infty} f(x,y)=\infty$

Yes, it is coercive.

(d). $f(0,y)=-y^2$, when $y\to\infty$. $\|(0,y)\|\to\infty$. $f(0,y)\to-\infty$

Not coercive.

3.4. **On the choice of learning rate** $\boxed{1}$

Compute and draw the first couple of steps of the gradient descent started at $x_0 = 1$ for the objective function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \frac{1}{2}x^2$ for the following learning rates α:

(a) $\alpha = 1$

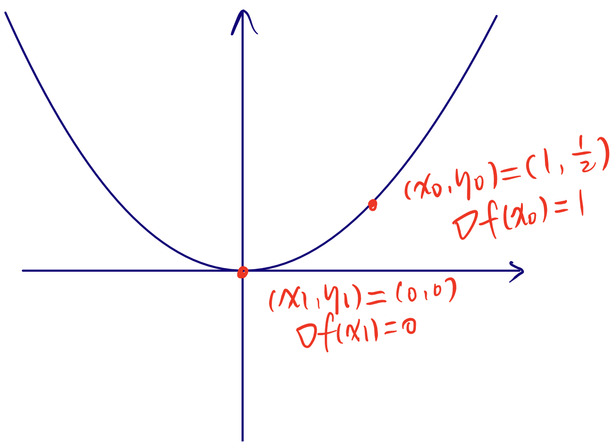
(b) $\alpha = 2$

(c) $\alpha = \frac{1}{2}$

$\nabla f(x) = x$. $\nabla f(x_0) = 1$.

(a). $\alpha = 1$, $x_1 = x_0 - \alpha \nabla f(x_0) = 0$.

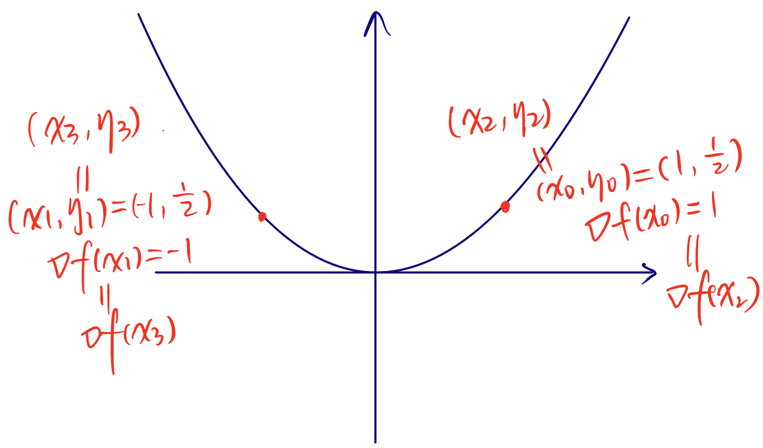
$\nabla f(x_1) = 0$, converges.



(b). $\alpha = 2$, $x_1 = x_0 - \alpha \nabla f(x_0) = -1$

$\nabla f(x_1) = -1$, $x_2 = x_1 - \alpha \nabla f(x_1) = 1 = x_0$

$\nabla f(x_2) = 1$, $x_3 = x_1 = -1$ loop

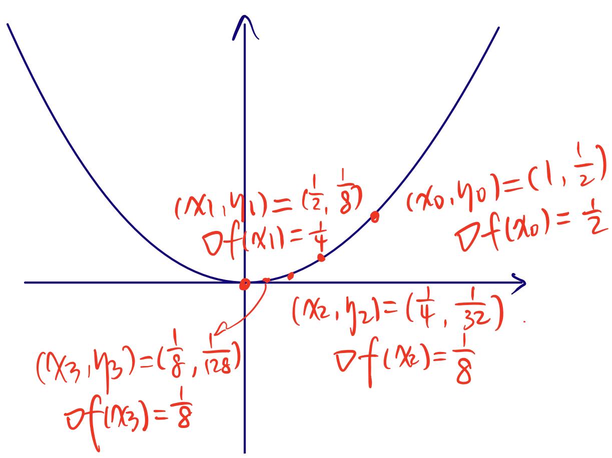


(c). $\alpha = \frac{1}{2}$, $x_1 = x_0 - \alpha \nabla f(x_0) = \frac{1}{2}$. $\nabla f(x_1) = \frac{1}{2}$.

$x_2 = x_1 - \alpha \nabla f(x_1) = \frac{1}{2}^2$ $\nabla f(x_2) = \frac{1}{2}^2$

$x_3 = x_2 - \alpha \nabla f(x_2) = \frac{1}{2}^3$, $\nabla f(x_3) = \frac{1}{2}^3$

.... infinite steps



3.5. **Gradient of quadratic functions** $\boxed{1}$

Let A be a 2×2 matrix, and let

$$f : \mathbb{R}^2 \to \mathbb{R}, \qquad f(x) = x^\top A x.$$

Show that $\nabla f(x) = (A + A^\top)x$.

let $x = (x_1, x_2)^T$. $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$.

$$f(x) = (x_1, x_2)\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (a_{11}x_1 + a_{21}x_2, a_{12}x_1 + a_{22}x_2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

$$= a_{11}x_1^2 + a_{21}x_1x_2 + a_{12}x_1x_2 + a_{22}x_2^2$$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{x_1} \\ \frac{\partial f(x)}{x_2} \end{bmatrix} = \begin{bmatrix} 2a_{11}x_1 + a_{21}x_2 + a_{12}x_2. \\ 2a_{22}x_2 + a_{21}x_1 + a_{12}x_1. \end{bmatrix} = (A + A^T)x$$

3.6. **Gradient descent for quadratic functions** $\boxed{\mathbf{1}}$

Follow the instructions in the Jupyter notebook.

You can achieve a total of $\boxed{\mathbf{6}}$ points for this problem set.

https://github.com/shahuiz/MH3520_Mathematics_of_Deep_Learning/blob/main/03_problemset.ipynb