Exercise 05

Tuesday, November 14, 2023 11:43 AM

Exercise 16 (10 points)

(a) (3 point) Compute the differential and the derivative (or its transpose) of the map $\mathbb{R}^{n\times m} \to \mathbb{R}$ $X \mapsto \mathrm{tr}(AX^TB)$ by the method given in the lecture (here A should also be an $n \times m$-matrix and B an $n \times k$-matrix).

(b) (4 points) Compute the differential and the derivative of the map $\mathbb{R}^{n\times n} \to \mathbb{R}$ $X \mapsto \mathrm{tr}(X^n)$ by the method given in the lecture.

[Remark: You can use without proof that $\mathrm{tr}(Y\epsilon Z\epsilon W)$ is $o(\|\epsilon\|)$ for any matrices Y, Z, W.]

(c) (3 points) The matrix exponential function is the function $\mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, $X \mapsto e^X$ defined by $e^X := \sum_{k=0}^{\infty} \frac{1}{k!}X^k$. One can show that this sum of matrices converges to a matrix. Compute the differential of the map $\mathbb{R}^{n\times n} \to \mathbb{R}$ $X \mapsto \mathrm{tr}(e^X)$ by the method given in the lecture.
[Remark: You can use without proof that you can pull the above infinite sum out of the trace. Reason: trace is a linear map (can pull out sums), and therefore continuous (can pull out the limit).]

(a) $\mathbb{R}^{n\times m} \to \mathbb{R}$

$X \xmapsto{f} \mathrm{tr}(AX^TB)$

$f(X+\varepsilon) = f(X) + Df(X)(\varepsilon) + O(\|\varepsilon\|)$

$\mathrm{tr}(AX^TB) = \mathrm{tr}(A(X+\varepsilon)^TB)$

$= \mathrm{tr}(A(X^T+\varepsilon^T)B)$

$= \mathrm{tr}(AX^TB) + \mathrm{tr}(A\varepsilon^TB)$ ($= \mathrm{tr}((A\varepsilon^TB)^T) = \mathrm{tr}(B^T\varepsilon A^T) = \mathrm{tr}((BA)^T\varepsilon)$)

$= \langle BA, X\rangle_F + \langle BA, \varepsilon\rangle_F + 0$

$f(X)$ $\quad Df(X)(\varepsilon)$ $\quad O(\|\varepsilon\|)$

So the differential $Df(X)(\varepsilon) = \langle BA, \varepsilon\rangle_F$

derivative is BA (a "constant" matrix).

tensor calculus

$X_{ij} \xmapsto{f} A_{ij}X^T_{jk}B_{ki} = A_{ij}X_{kj}B_{ki}$

$\frac{\partial f}{\partial X_{mn}} = A_{ij}B_{ki}\frac{\partial X_{kj}}{\partial X_{mn}}$

$= A_{ij}B_{ki}\delta_{km}\delta_{jn}$

$= A_{in}B_{mi}$

$= B_{mi}A_{in}$

$\Rightarrow \frac{\partial f}{\partial \underline{X}} = \underline{B}\cdot\underline{A}$

(b) $\mathbb{R}^{n\times n} \to \mathbb{R}$

$X \xmapsto{f} \mathrm{tr}(X^n)$

$f(X+\varepsilon) = \mathrm{tr}((X+\varepsilon)^n)$

$\overset{n\times n}{=\!=} \mathrm{tr}\big(\binom{n}{n}X^n + \underbrace{X^{n-1}\varepsilon + X^{n-2}\varepsilon X + \cdots + \varepsilon X^{n-1}}_{\binom{n}{1}}$

$+ \underbrace{X^{n-2}\varepsilon^2 + X^{n-3}\varepsilon X\varepsilon + \cdots}_{\binom{n}{2}+\cdots+\binom{n}{n}}\big)$

$= \mathrm{tr}(X^n) + n\,\mathrm{tr}(X^{n-1}\varepsilon) + o(\|\varepsilon\|)$

$= \mathrm{tr}(X^n) + \langle n(X^{n-1})^T, \varepsilon\rangle_F + o(\|\varepsilon\|)$

$= f(X) + \langle n(X^T)^{n-1}, \varepsilon\rangle_F + o(\|\varepsilon\|)$

So the differential $Df(X)(\varepsilon) = \langle n(X^T)^{n-1}, \varepsilon\rangle_F$

derivative is $n(X^T)^{n-1}$

Tensor calculus

$\frac{\partial X^2}{\partial \underline{X}} =$

$\frac{\partial X_{ij}X_{jm}}{\partial X_{kl}} = \delta_{ik}\delta_{jl}X_{jm} + \delta_{jk}\delta_{ml}X_{ij}$

$= \delta_{ik}X_{lm} + X_{ik}\delta_{ml}$

$= \delta_{ik}X_{ml} + X_{ik}\delta_{ml}$

$= \underline{I}\otimes\underline{X} + \underline{X}\otimes\underline{I}$

$\frac{\partial \mathrm{tr}(\underline{X}^n)}{\partial \underline{X}} = \frac{\partial \underline{I}:\underline{X}^n}{\partial \underline{X}} = \underline{I}:\frac{\partial \underline{X}^n}{\partial \underline{X}}$

$\delta_{im}\cdot\frac{\partial X_{ij}X_{jm}}{\partial X_{kl}} = \delta_{ik}X_{il} + X_{ik}\delta_{il}$

$= X_{kl} + X_{lk} = \underline{X} + \underline{X}^T$

(c) (c) (3 points) The matrix exponential function is the function $\mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, $X \mapsto e^X$ defined by $e^X := \sum_{k=0}^{\infty} \frac{1}{k!}X^k$. One can show that this sum of matrices converges to a matrix. Compute the differential of the map $\mathbb{R}^{n\times n} \to \mathbb{R}$ $X \mapsto \mathrm{tr}(e^X)$ by the method given in the lecture.
[Remark: You can use without proof that you can pull the above infinite sum out of the trace. Reason: trace is a linear map (can pull out sums), and therefore continuous (can pull out the limit).]

(c)

(c) (3 points) The matrix exponential function is the function $\mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, $X \mapsto e^X$ defined by $e^X := \sum_{k=0}^{\infty} \frac{1}{k!} X^k$. One can show that this sum of matrices converges to a matrix. Compute the differential of the map $\mathbb{R}^{n\times n} \to \mathbb{R}$ $X \mapsto \operatorname{tr}(e^X)$ by the method given in the lecture.
[Remark: You can use without proof that you can pull the above infinite sum out of the trace. Reason: trace is a linear map (can pull out sums), and therefore continuous (can pull out the limit).]

$\mathbb{R}^{n\times n} \to \mathbb{R}$

$X \xrightarrow{f} \operatorname{tr}(e^X)$

Q: $e^{X+\varepsilon} = e^X \cdot e^{\varepsilon}$?

$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{1}{i!}\frac{1}{j!} X^i \varepsilon^j$

$i+j=k$ => seems like it

$$f(X+\varepsilon) = \operatorname{tr}(e^{X+\varepsilon})$$

$$= \operatorname{tr}\left(\sum_{k=0}^{\infty} \frac{1}{k!}(X+\varepsilon)^k\right)$$

$$= \sum_{k=0}^{\infty} \frac{1}{k!} \operatorname{tr}\left((X+\varepsilon)^k\right) = \operatorname{tr}(I) + \sum_{k=1}^{\infty} \frac{1}{k!} \operatorname{tr}\left((X+\varepsilon)^k\right)$$

$$\overset{\text{by (b)}}{=\!=} \underbrace{\operatorname{tr}(I) + \sum_{k=1}^{\infty} \frac{1}{k!}\left(\operatorname{tr}(X^k) + k\operatorname{tr}(X^{k-1}\varepsilon) + o(\|\varepsilon\|)\right)}$$

$$= \operatorname{tr}(e^X) + \operatorname{tr}\left(\sum_{k=1}^{\infty} \frac{1}{(k-1)!} X^{k-1}\varepsilon\right) + o\|\varepsilon\|$$

$$\overset{k-1=j}{=\!=} \operatorname{tr}(e^X) + \operatorname{tr}\left(\sum_{j=0}^{\infty} \frac{1}{j!} X^j \varepsilon\right) + o\|\varepsilon\| = f(X) + \operatorname{tr}(e^X\varepsilon) + o\|\varepsilon\|$$

$$= f(X) + \langle (e^X)^T, \varepsilon\rangle_F + o\|\varepsilon\|$$

So the differential $Df(X)(\varepsilon) = \langle (e^X)^T, \varepsilon\rangle_F$

derivative is $(e^X)^T = \left(\sum_{k=0}^{\infty} \frac{1}{k!} X^k\right)^T = \sum_{k=0}^{\infty} \frac{1}{k!} (X^T)^k = e^{X^T}$

Exercise 17 (10 points)

A casino owner wants to offer a betting game with n possible outcomes, which give her revenues of $a_1, \ldots, a_n$ with $a_i \in \mathbb{R}$. She can freely set up the game choosing the the probability p_i of outcome i. She wants to base her choice on various factors, e.g. the expected revenue should be high enough, it should also be reliable enough (she wants not too big variance), but also not too predictable (the variance should not be low). The following questions are part of checking which of the requirements can be achieved by solving *convex* optimization problems.

(a) (2 points) Consider the set of possible probability distributions $P = \{p = (p_1, \ldots, p_n) \mid p_i \geq 0, \Sigma_i p_i = 1\}$. Show that this set is a convex subset of $\mathbb{R}^n$.

a: revenue

p: outcome prob.

$E(a)\uparrow$ $\quad Var(a)\downarrow$ $\quad Var(p)\uparrow$

(a) Suppose $p, q \in P \subseteq \mathbb{R}^n$ is arbitrary

p can be written as $(p_1, \cdots, p_n)$ $\quad p_i \geq 0,\ \sum_i p_i = 1$

q can be written as $(q_1, \cdots, q_n)$ $\quad q_i \geq 0,\ \sum_i q_i = 1$

For $\forall \theta \in [0,1]$

$$\theta p + (1-\theta) q = (\theta p_1 + (1-\theta) q_1,\ \theta p_2 + (1-\theta) q_1,\ \cdots,\ \theta p_n + (1-\theta) q_n) \in \mathbb{R}^n$$

- Since $\theta \in [0,1]$, $\theta \geq 0$, $1-\theta \geq 0$

  and since $p_i \geq 0$, $q_i \geq 0$ $\Rightarrow$ $\boxed{\theta p_i + (1-\theta) q_i \geq 0}$

- $\boxed{\sum_{i=1}^{n} [\theta p_i + (1-\theta) q_i]} = \theta \sum_{i=1}^{n} p_i + (1-\theta) \sum_{i=1}^{n} q_i = \theta + 1 - \theta \boxed{= 1}$

Therefore $\theta p + (1-\theta) q \in P$, for $\forall \theta \in [0,1]$, $p, q \in P$.

$\Rightarrow$ $P \subseteq \mathbb{R}^n$ is a convex set

$\mathbb{R}^n$ is certainly convex $\Longrightarrow$ P is a convex subset of $\mathbb{R}^n$

(b) (2 points) The expected revenue is $E(p) := \sum_{i=1}^{n} p_i a_i$. Is $E\colon P \to \mathbb{R}$ a convex function? Is it a concave function?

Let $A_{1\times n} = (a_1\ a_2\ \cdots\ a_n)$

$E(p) = Ap$ is an affine function (even linear)

Since P is convex, $E(p): P \to \mathbb{R}$ is a convex function.

(c) (2 points, tricky) The variance of her revenue is $Var(p) := \sum_{i=1}^{n} p_i(a_i - E(p))^2$. Is $Var\colon P \to \mathbb{R}$ a convex function? Is it a concave function?

[Hint: You can use without proof that $Var(p) = \sum_{i=1}^{n} p_i a_i^2 - E(p)^2$ – this is the formula for random variables $Var(X) = E(X^2) - (EX)^2$ You might also want to prove and use that $x \mapsto x^2$ is a convex function and then apply Example 3.3.9.7 of the notes]

$Var(p) = \sum_i^n p_i a_i^2 - E(p)^2$

$f_2: \mathbb{R} \to \mathbb{R}: = f_2(x) = x^2$

It's second derivative $Hf = \frac{\partial^2 f_2}{\partial x^2} = 2 \geq 0 \Rightarrow f_2$ is convex

by 3.3.9.5

$f_2(E(p)) = f_2(Ap + 0)$

since $E(p)$ is affine, f_2 is convex

$f_2(E(p)) = (E(p))^2 = E(p)^2$ is convex

Suppose $p, q \in P$ is arbitrary, $\forall \theta \in [0,1]$:

$r = \theta p + (1-\theta) q$

$-Var(r) = -\sum_{i=1}^{n} \theta p_i a_i^2 - \sum_{i=1}^{n} (1-\theta) q_i a_i^2 + \underbrace{E(r)^2}_{\text{convex}}$

$\leq -\theta \sum_{i=1}^{n} p_i a_i^2 - (1-\theta) \sum_{i=1}^{n} q_i a_i^2 + \theta E(p)^2 + (1-\theta) E(q)^2$

$= \theta(-Var(p)) + (1-\theta)(-Var(q))$

$\Rightarrow -Var(p)$ is a convex function

$\Rightarrow Var(p)$ is concave

(d) (2 points) Let $\alpha \in \mathbb{R}$. Is $\{p \in P \mid Var(p) \geq \alpha\}$ a convex set?

(e) (2 points) Let $\alpha \in \mathbb{R}$. Is $\{p \in P \mid Var(p) \leq \alpha\}$ a convex set?

(d) Suppose $p, q \in \{p \in P \mid Var(p) \geq \alpha\} \subseteq P$ are arbitrary. $-Var(p) \leq -\alpha$

$\forall \theta \in [0,1]$, since $Var(p)$ is concave:

$-Var(\theta p + (1-\theta) q) \leq -\theta\, Var(p) + (1-\theta)[-Var(q)]$

$\leq -\theta\alpha - (1-\theta)\alpha = -\alpha$

$\Rightarrow Var(\theta p + (1-\theta) q) \geq \alpha$

So $\forall p, q, \theta \in [0,1]$: $\theta p + (1-\theta) q \in \{p \in P \mid Var(p) \geq \alpha\}$

$\{p \in P \mid Var(p) \geq \alpha\}$ is then a convex set.

(e) let $n = 2$, $\alpha = 1$, so the set is $\{p \in P \mid Var(p) \leq 0\}$

we choose $p = (1, 0)^T$, $q = (0, 1)^T$, $a = (-1, 1)^T$

(e) let $n=2$, $\alpha=1$, so the set is $\{p \in P \mid \mathrm{Var}(p) \le 0\}$

we choose $p=(1,0)^T$, $q=(0,1)^T$, $a=(-1,1)^T$

$\mathrm{Var}(p) = 1-(-1)^2 = 0$, $\mathrm{Var}(q) = 1-1^2 = 0$

let $\theta = 0.5$ $\quad \theta p + (1-\theta)q = (0.5, 0.5)^T$

$\mathrm{Var}(\theta p + (1-\theta)q) = 1 \not\le 0$

$\mathrm{Var}((0.5, 0.5)^T) \notin \{p \in P \mid \mathrm{Var}(p) \le 0\}$

Therefore $\{p \in P \mid \mathrm{Var}(p) \le 0\}$ is not convex.

Exercise 18 (12 points)

(a) (2 points) Is the function $f: \mathbb{R}^2 \to \mathbb{R}$, $(x,y) \mapsto \frac{x}{y}$ convex? Prove your answer.

$\mathbb{R}^2$ is a convex set.

$$Hf = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial f}{\partial x \partial y} \\ \frac{\partial f}{\partial y \partial x} & \frac{\partial f}{\partial y^2} \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{y^2} \\ -\frac{1}{y^2} & \frac{2x}{y^3} \end{pmatrix}$$

at $(0,1)$, $Hf = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$ $\quad \chi_{Hf}(\lambda) = \det\begin{pmatrix} \lambda & 1 \\ 1 & \lambda \end{pmatrix}$

$= \lambda^2 - 1 = 0$

$\lambda_2 = -1 < 0$

Hf is not positive semi-definite in $\mathbb{R}^2 \Rightarrow f$ is not convex

(b) (2 points) A quadratic function $f: \mathbb{R}^n \to \mathbb{R}$ is a function of the form $x \mapsto x^T A x + b^T x + c$, for some $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Give conditions on A, b and b which characterize when exactly such a function is convex.

$f(x) = x^T A x + b^T x + c$

$\frac{\partial f}{\partial x} = Ax + A^T x + b$

$Hf(x) = \frac{\partial^2 f}{\partial x^2} = A + A^T$

Since $f(x)$ is convex iff. $Hf(x)$ is positive semi-definite.

$A + A^T$ must be positive semi-definite. There is no limitation on b.

Further more, $A + A^T$ is symmetric therefore diagonalizable.

$A + A^T$ have non-negative eigenvalues $\Leftrightarrow$ $f(x)$ is convex

(c) (4 points) Consider the set of probability distributions on an n-element set, $P := \{p = (p_1, \dots, p_n) \mid p_i \geq 0, \Sigma_i p_i = 1\}$. Show that the following function is strictly convex on P:

$$f\colon P \to \mathbb{R}, \qquad f(p) := \sum_{i=1}^{n} p_i \log p_i$$

Here we adopt the convention that $0 \cdot \log 0 = 0$.

[Remark: The function f is called *negative entropy*. The *entropy* of a probability distribution (i.e. $-f(p)$) is a measure of how uncertain is the result of a random experiment with the probabilities given by p. It is widely used in Machine Learning, and we will cover it in the part on Information Theory. By (c) entropy is a concave function, and that is good, because its maximization is a standard task in Bayesian Statistics.]

$f(p) = \sum_{i=1}^{n} p_i \log p_i$

$\nabla f = \frac{\partial f(p)}{\partial p} = \frac{\partial (p \log(p))_i}{\partial p_j} e_j = (\log(p_j) + 1) e_j$

$\nabla f = \frac{\partial}{\partial p_k} e_k (\log(p_j) + 1) e_j$

$= (\delta_{jk} \frac{1}{p_j}) e_j e_k = \frac{1}{p} \cdot I$

For $1 \leq i \leq n$

$\frac{\partial f(p)}{\partial p_i} = \log p_i + p_i \cdot \frac{1}{p_i}$

$= \log p_i + 1$

$\frac{\partial f(p)}{\partial p_i \partial p_j} = \begin{cases} \frac{1}{p_i} & i = j \\ 0 & i \neq j \end{cases}$ $\Rightarrow Hf(p) = \operatorname{diag}(\frac{1}{p_1}, \frac{1}{p_2}, \dots, \frac{1}{p_n})$

since $p_i \geq 0$ $Hf(p)$ is positive semi-definite

$\sum_{i=1}^{n} p_i = 1$, $\exists\, p_i \in \{p_1, \dots, p_n\}$ s.t. $p_i \neq 0 \Rightarrow Hf(p)$ is positive definite

$\Rightarrow f$ is strictly convex on P $\square$

(d) (4 points) The *Kullback-Leibler divergence* is a distance measure between probability distributions. It is the map $D_{KL}(-\|-)\colon P \to \mathbb{R}$ defined by $D_{KL}(p\|q) := \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right)$. It does not satisfy all the properties of a metric, but at least the following two of them:

Show that $D_{KL}(p\|q) \geq 0$ for all $p, q \in P$ and that $D_{KL}(p\|q) = 0$ if and only if $p = q$.

[Hint: Use the function f of part (c): Show that $D_{KL}(p\|q) = f(p) - f(q) - \nabla f(q)^T (p - q)$. Then use the first order characterization of strict convexity from Prop. 3.3.3(a) of the manuscript, applied to f: This says that a differentiable function f is convex if and only if its domain is convex and $f(y) \geq f(x) + (gradf)(x)(y - x)$ for all x, y.]

Suppose $p, q \in P$ are arbitrary

$\nabla f(q) = \frac{\partial f(q)}{\partial q} = \begin{pmatrix} \log(q_1) + 1 \\ \vdots \\ \log(q_n) + 1 \end{pmatrix}$

$\nabla f(q)^T (p - q) = \sum_{i=1}^{n} (\log(q_i)(p_i - q_i) + p_i - q_i)$

$f(p) - f(q) - \nabla f(q)^T (p - q) = \sum_{i=1}^{n} (p_i \log p_i - q_i \log q_i - p_i \log q_i + q_i \log q_i + p_i - q_i)$

$= \sum_{i=1}^{n} p_i (\log p_i - \log q_i) + \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} q_i$

$= \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} = D_{KL}(p\|q)$

Since f is strictly convex:

$i=1$ q_i $D_{KL}(p \| q)$

Since f is strictly convex:

$f(p) \geq f(q) + \nabla f(q)^T(p-q)$ and $f(p) = f(q) + \nabla f(q)^T(p-q)$ iff $p=q$

$\Rightarrow$ for $\forall p, q \in P$, $D_{KL}(p \| q) \geq 0$ and $D_{KL}(p \| q) = 0$ iff $p=q$