CS 6190: Probabilistic Machine Learning Spring 2022

Homework 0

Handed out: 10 Jan, 2022
Due: 11:59pm, 21 Jan, 2022

- You are welcome to talk to other members of the class about the homework. I am more concerned that you understand the underlying concepts. However, you should write down your own solution. Please keep the class collaboration policy in mind.
- Feel free discuss the homework with the instructor or the TAs.
- Your written solutions should be brief and clear. You need to show your work, not just the final answer, but you do *not* need to write it in gory detail. Your assignment should be **no more than 10 pages**. Every extra page will cost a point.
- Handwritten solutions will not be accepted.
- The homework is due by **midnight of the due date**. Please submit the homework on Canvas.

Warm up[100 points + 5 bonus]

1. [2 points] Given two events A and B, prove that

$$p(A \cup B) \leq p(A) + p(B)$$
$$p(A \cap B) \leq p(A), p(A \cap B) \leq p(B)$$

When does the equality hold?

Answer

a) Let $P(A)$ and $P(B)$ be probability of two events A and B respectively. From the definition of probability we know that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. As

$$0 \leq P(A \cap B) \leq 1$$

$$P(A \cup B) \leq P(A) + P(B)$$

The equality holds when A and B are independent.

b) Let $P(A)$ and $P(B)$ be probability of two events A and B respectively. From the definition of probability we know that $P(A \cap B) = P(A|B)P(B)$ and $P(A \cap B) = P(B|A)P(A)$. As

$$0 \leq P(A|B) \leq 1 \text{ and } 0 \leq P(B|A) \leq 1$$

$$P(A \cap B) \leq P(B) \text{ and } P(A \cap B) \leq P(A)$$

The equality holds when $P(A|B) = 1$ and $P(B|A) = 1$ respectively i.e. $P(A) = p(B)$ (A and B cover the same area in Venn diagram).

2. [2 points] Let $\{A_1, \ldots, A_n\}$ be a collection of events. Show that

$$p(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n p(A_i).$$

When does the equality hold? (Hint: induction)

Answer

From Inclusion Exclusion principle, for events $A_1, A_2, \cdots, A_n$ in probability space we know that

$$p(\cup_{i=1}^n A_i) = \sum_{i=1}^n p(A_i) - \sum_{1\leq i<j\leq n} p(A_i \cap A_j) + \sum_{1\leq i<j<k\leq n} p(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} P(\cap_{i=1}^n A_i)$$

In the above equation, when we subtract multiple double integrations, we're also subtraction triple integrations twice. To remedy this, we add the small amount back. This goes on and on until we reach $P(\cap_{i=1}^n A_i)$. Note that the amount added is always smaller than the one subtracted before hence we can say that

$$p(\cup_{i=1}^n A_i) = \sum_{i=1}^n p(A_i) - Z; Z \geq 0$$

Z in the above equation is the sum of all the terms after the first one. So we can finally say that

$$p(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n p(A_i)$$

The equality holds when all events are independent.

3. [14 points] We use $\mathbb{E}(\cdot)$ and $\mathbb{V}(\cdot)$ to denote a random variable's mean (or expectation) and variance, respectively. Given two discrete random variables X and Y, where $X \in \{0, 1\}$ and $Y \in \{0, 1\}$. The joint probability $p(X, Y)$ is given in as follows:

	$Y = 0$	$Y = 1$
$X = 0$	$3/10$	$1/10$
$X = 1$	$2/10$	$4/10$

(a) [10 points] Calculate the following distributions and statistics.

i. the the marginal distributions $p(X)$ and $p(Y)$

Answer

$P(X = 0) = P(X = 0 \cap Y = 0) + P(X = 0 \cap Y = 1) = 0.3 + 0.1 = 0.4$
$P(X = 1) = P(X = 1 \cap Y = 0) + P(X = 1 \cap Y = 1) = 0.2 + 0.4 = 0.6$
$P(Y = 0) = P(Y = 0 \cap X = 0) + P(Y = 0 \cap X = 1) = 0.3 + 0.2 = 0.5$
$P(Y = 1) = P(Y = 1 \cap X = 0) + P(Y = 1 \cap X = 1) = 0.1 + 0.4 = 0.5$

ii. the conditional distributions $p(X|Y)$ and $p(Y|X)$

Answer

We know that $p(A|B) = \frac{p(A,B)}{p(B)}$, Hence using this definition

$$p(X = 0|Y = 0) = \frac{0.3}{0.5} = 0.6$$

$$p(X = 0|Y = 1) = \frac{0.1}{0.5} = 0.2$$

$$p(X=1|Y=0) = \frac{0.2}{0.5} = 0.4$$

$$p(X=1|Y=1) = \frac{0.4}{0.5} = 0.8$$

$$p(Y=0|X=0) = \frac{0.3}{0.4} = 0.75$$

$$p(Y=0|X=1) = \frac{0.2}{0.6} = 0.33$$

$$p(Y=1|X=0) = \frac{0.1}{0.4} = 0.25$$

$$p(Y=1|X=1) = \frac{0.4}{0.6} = 0.66$$

iii. $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathbb{V}(X)$, $\mathbb{V}(Y)$

Answer

From definition we know that $\mathbb{E}(A) = \sum ap(a)$, hence using this definition

$$\mathbb{E}(X) = 0.6$$

$$\mathbb{E}(Y) = 0.5$$

From definition we know that $\mathbb{V}(A) = \sum(a - \mathbb{E}(a))^2 p(a)$, hence using this definition

$$\mathbb{V}(X) = (0-0.6)^2 \times 0.4 + (1-0.6)^2 \times 0.6 = 0.24$$

$$\mathbb{V}(Y) = (0-0.5)^2 \times 0.5 + (1-0.5)^2 \times 0.5 = 0.25$$

iv. $\mathbb{E}(Y|X=0)$, $\mathbb{E}(Y|X=1)$, $\mathbb{V}(Y|X=0)$, $\mathbb{V}(Y|X=1)$

Answer

We know that $\mathbb{E}(A|B=0) = \sum ap(a|b=0)$ hence

$$\mathbb{E}(Y|X=0) = 0.25$$

$$\mathbb{E}(Y|X=1) = 0.66$$

We know that $\mathbb{V}(A|B=0) = \sum(a - \mathbb{E}(a|b=0))^2 p(a|b=0)$ hence

$$\mathbb{V}(Y|X=0) = (0-0.25)^2 \times 0.75 + (1-0.25)^2 \times 0.25 = 0.18$$

$$\mathbb{V}(Y|X=1) = (0-0.66)^2 \times 0.33 + (1-0.66)^2 \times 0.66 = 0.22$$

v. the covariance between X and Y

Answer

We know that $cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$, hence $cov(X,Y) = 0.4 - 0.3 = 0.1$

(b) [2 points] Are X and Y independent? Why?

Answer

No they are not independent as $p(x|y) \neq p(x)$ for all x, and also $cov(X,Y) \neq 0$.

(c) [2 points] When X is not assigned a specific value, are $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ still constant? Why?

Answer

As $p(Y|X)$ is a function of X and Y, $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ will not be constant if a specific value is not assigned to X.

4. [9 points] Assume a random variable X follows a standard normal distribution, i.e., $X \sim \mathcal{N}(X|0,1)$. Let $Y = e^{-X^2}$. Calculate the mean and variance of Y.

   (a) $\mathbb{E}(Y)$

   Answer

   As X follows the normal distribution, $p(X) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$

   Then expectation of Y is given as

$$\mathbb{E}(Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{3x^2}{2}\right) dx$$

   Let $\sqrt{3}x = t$, then $\sqrt{3}dx = dt$

$$\mathbb{E}(Y) = \frac{1}{\sqrt{6\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt$$

$$\mathbb{E}(Y) = \sqrt{\frac{2\pi}{6\pi}} = \frac{1}{\sqrt{3}}$$

   (b) $\mathbb{V}(Y)$

   Answer

$$\mathbb{V}(Y) = \int_{-\infty}^{\infty} (\exp(-x^2) - \frac{1}{\sqrt{3}}) \frac{\exp(-\frac{x^2}{2})}{\sqrt{2\pi}}$$

$$\mathbb{V}(Y) = \int_{-\infty}^{\infty} \frac{\exp(-\frac{3x^2}{2})}{\sqrt{2\pi}} dx - \int_{-\infty}^{\infty} \frac{\exp(-\frac{x^2}{2})}{\sqrt{6\pi}} dx$$

   Solving the above equation we get

$$\mathbb{V}(Y) = \frac{1}{\sqrt{3}} - \frac{1}{\sqrt{3}} = 0$$

   (c) $\text{cov}(X, Y)$

   Answer

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

   We know that $\mathbb{E}(X) = 0$, hence $\text{cov}(X, Y) = \mathbb{E}(XY)$.

$$\mathbb{E}(XY) = \mathbb{E}(X \exp(-X^2)) = \frac{1}{\sqrt{2\pi}} \int x \exp(-3x^2/2) dx = \frac{1}{3} e^{-\frac{3}{2}x^2}$$

5. [8 points] Derive the probability density functions of the following transformed random variables.

   (a) $X \sim \mathcal{N}(X|0,1)$ and $Y = X^3$.

   Answer

   PDF is given by $f_X(g^{-1}(y)) \cdot \left|\frac{d}{dy} g^{-1}(y)\right|$, where $f_X = \mathcal{N}(X|0,1)$ and $g^{-1}(y) = y^{\frac{1}{3}}$. Plugging these into the formula gives

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^{2/3}}{2}) \cdot \left|\frac{1}{3} y^{-2/3}\right|$$

$$f_Y(y) = \frac{1}{3\sqrt{2\pi}} \exp(-\frac{y^{2/3}}{2}) \cdot y^{-2/3}$$

(b) $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} | \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix})$ and $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1/2 \\ -1/3 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$.

Answer

If $\mathbf{x}$ follows multivariate normal distribution such that $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ and the linear transformation of $\mathbf{x}$ is also multivariately distributed such that $\mathbf{y} = \mathbf{A}\mathbf{x}$, then PDF is given as

$$\mathbf{F}_Y(\mathbf{y}) = \exp\left(\mathbf{y}^T\mathbf{A}\mu + \frac{1}{2}\mathbf{y}^T\mathbf{A}\Sigma\mathbf{A}^T\mathbf{y}\right)$$

Plugging the given matrices and vectors in above equation we get

$$\mathbf{F}_Y(\mathbf{y}) = \exp\left(\frac{1}{2}\mathbf{y}^T\mathbf{K}\mathbf{y}\right)$$

where $\mathbf{K} = \begin{bmatrix} 0.75 & -0.25 \\ -0.25 & 1.44 \end{bmatrix}$.

6. [10 points] Given two random variables X and Y, show that

   (a) $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$

   Answer

$$\mathbb{E}(\mathbb{E}(Y|X)) = \int_{-\infty}^{\infty} \mathbb{E}(Y|X=x)p_X(x)dx$$
$$\mathbb{E}(\mathbb{E}(Y|X)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yp_{Y|X}(y|X=x)p_X(x)dydx$$
$$\mathbb{E}(\mathbb{E}(Y|X)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yp_{Y,X}(y,x)dydx$$
$$\mathbb{E}(\mathbb{E}(Y|X)) = \int_{-\infty}^{\infty} y\int_{-\infty}^{\infty} p_{Y,X}(y,x)dxdy$$
$$\mathbb{E}(\mathbb{E}(Y|X)) = \int_{-\infty}^{\infty} yp_Y(y)dy$$
$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$$

   (b) $\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X))$

   Answer

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$$
$$\mathbb{V}(Y) = \mathbb{E}(\mathbb{E}(Y^2|X)) - \mathbb{E}(\mathbb{E}(Y|X))^2$$
$$\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X) + \mathbb{E}(Y|X)^2) - \mathbb{E}(\mathbb{E}(Y|X))^2$$
$$\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X) + \mathbb{E}(\mathbb{E}(Y|X)^2)) - \mathbb{E}(\mathbb{E}(Y|X))^2$$
$$\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X))$$

   (Hints: using definition.)

7. [9 points] Given a logistic function, $f(\mathbf{x}) = 1/(1+\exp(-\mathbf{a}^\top\mathbf{x}))$ ($\mathbf{x}$ is a vector),

(a) derive $\frac{\mathrm{d}f(\mathbf{x})}{\mathrm{d}\mathbf{x}}$

Answer
We know that

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

and

$$\frac{\partial f}{\partial x_k} = (1 + e^{-\sum_{i=1}^{n} a_i x_i})^{-2} \cdot e^{-\sum_{i=1}^{n} a_i x_i} \cdot a_k = f(1-f)a_k$$

Hence combined, the whole vector can be described by the equation

$$\nabla f(\mathbf{x}) = f(\mathbf{x})(1 - f(\mathbf{x})) \cdot a$$

(b) derive $\frac{\mathrm{d}^2 f(\mathbf{x})}{\mathrm{d}\mathbf{x}^2}$, i.e., the Hessian matrix

Answer
We know that the Hessian is defined as

$$\nabla^2 f(\mathbf{x}) = \nabla(\nabla f(\mathbf{x})) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \frac{\partial^2 f}{\partial x_1 x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \vdots & & & \\ \frac{\partial^2 f}{\partial x_1 x_n} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

and

$$\frac{\partial^2 f}{\partial x_i x_j} = (f'(1-f) + f(1-f)')a_i$$

$$\frac{\partial^2 f}{\partial x_i x_j} = (f(1-f)^2 a_j + f(-f(1-f)a_j))a_i$$

$$\frac{\partial^2 f}{\partial x_i x_j} = f(1-f)(1-2f)a_i a_j$$

which can be written as

$$\nabla^2 f(\mathbf{x}) == f(1-f)(1-2f)aa^T$$

(c) show that $-\log(f(\mathbf{x}))$ is convex

Answer
For $g(x) = -\log(f(\mathbf{x}))$ to be convex means $g''(x) > 0$. Differentiating $g(x)$ once we get

$$g'(x) = -\frac{1}{f(x)} \cdot f'(x) = (f-1) \cdot a = \begin{bmatrix} (f-1)a_1 \\ (f-1)a_2 \\ \vdots \\ (f-1)a_n \end{bmatrix}$$

Differentiating the gradient, we get Hessian as follows

$$g''(x) = \begin{bmatrix} f(1-f)a_1 a_1 & f(1-f)a_1 a_2 & \cdots & f(1-f)a_1 a_n \\ f(1-f)a_2 a_1 & f(1-f)a_2 a_2 & \cdots & f(1-f)a_2 a_n \\ \vdots & & & \\ f(1-f)a_n a_1 & f(1-f)a_n a_2 & \cdots & f(1-f)a_n a_n \end{bmatrix}$$

The above matrix can be written as

$$g''(x) = f(1-f)aa^T$$

and we can see that product $f(1-f) > 0$, and aa^T is a positive definite matrix (diagonals will always be positive) hence $g(x)$ is a convex function.

Note that $0 \leq f(\mathbf{x}) \leq 1$.

8. [10 points] Derive the convex conjugate for the following functions

   (a) $f(x) = -\log(x)$

   Answer

   Convex conjugate for $f(x)$ is defined as $\sup_x(yx - f(x))$ hence for the problem at hand $g(y) = \sup_x(yx + log(x))$. Differentiating the above function to find the max distance and then substituting x in terms of y we get convex conjugate as $f^*(y) = -1 - log(-y)$.

   (b) $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}^{-1}\mathbf{x}$ where $\mathbf{A} \succ 0$

   Answer

   Convex conjugate for $f(\mathbf{x})$ is defined as

   $$\sup_x(\mathbf{y}^T\mathbf{x} - f(\mathbf{x}))$$

   hence for the problem at hand

   $$g(\mathbf{y}) = \sup_{\mathbf{x}}(\mathbf{y}^T\mathbf{x} - \mathbf{x}^T\mathbf{A}^{-1}\mathbf{x}).$$

   Differentiating the above function to find the max distance we get

   $$dg(\mathbf{x}) = \mathbf{y}^T d\mathbf{x} - \mathbf{x}^T(\mathbf{A}^{-1} + (\mathbf{A}^{-1})^T)d\mathbf{x}$$

   and setting this to zero gives

   $$\mathbf{x} = (\mathbf{A}^{-1} + (\mathbf{A}^{-1})^T)^{-1}\mathbf{y}.$$

   Now to find convex conjugate of the function, we just plug this result in and get

   $$f^*(\mathbf{y}) = \mathbf{y}^T\big(\mathbf{A}^{-1} + (\mathbf{A}^{-1})^T\big)^{-1}\mathbf{y} - \mathbf{y}^T\big(\mathbf{A}^{-1} + (\mathbf{A}^{-1})^T\big)^{-1}\mathbf{A}^{-1}\big(\mathbf{A}^{-1} + (\mathbf{A}^{-1})^T\big)^{-1}\mathbf{y}$$

   .

9. [20 points] Derive the (partial) gradient of the following functions. Note that bold small letters represent vectors, bold captial letters matrices, and non-bold letters just scalars.

   (a) $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$, derive $\frac{\partial f}{\partial \mathbf{x}}$

   Answer

   $$\partial f = \partial \mathbf{x}^T \mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{A}\partial\mathbf{x}$$

   $$\partial f = \mathbf{x}^T\mathbf{A}^T\partial\mathbf{x} + \mathbf{x}^T\mathbf{A}\partial\mathbf{x}$$

   $$\partial f = \mathbf{x}^T(\mathbf{A}^T + \mathbf{A})\partial\mathbf{x}$$

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

(b) $f(\mathbf{x}) = \left(\mathbf{I} + \mathbf{x}\mathbf{x}^\top\right)^{-1}\mathbf{x}$, derive $\frac{\partial f}{\partial \mathbf{x}}$
Answer

$$\partial f = \partial(\mathbf{I} + \mathbf{x}\mathbf{x}^T)^{-1}\mathbf{x} + (\mathbf{I} + \mathbf{x}\mathbf{x}^T)^{-1}\partial\mathbf{x}$$
$$\partial f = -(\mathbf{I} + \mathbf{x}\mathbf{x}^T)^{-1}\partial(\mathbf{x}\mathbf{x}^T)(\mathbf{I} + \mathbf{x}\mathbf{x}^T)^{-1}\mathbf{x} + (\mathbf{I} + \mathbf{x}\mathbf{x}^T)^{-1}\partial\mathbf{x}$$

For the sake of simplifying the answer let's call $(\mathbf{I} + \mathbf{x}\mathbf{x}^T)^{-1} = \mathbf{P}$. The final answer is

$$\mathbf{P} - (\mathbf{x}^T\mathbf{P}\mathbf{x}\mathbf{P} + \mathbf{P}\mathbf{x}(\mathbf{x}^T(\mathbf{I}^T + \mathbf{x}\mathbf{x}^T)))$$

(c) $f(\alpha) = \log|\mathbf{K} + \alpha\mathbf{I}|$, where $|\cdot|$ means the determinant. Derive $\frac{\partial f}{\partial \alpha}$
Answer

$$\partial f = Tr((\mathbf{K} + \mathbf{x}\mathbf{I})^{-1}\partial\alpha)$$

$$\partial f = Tr((\mathbf{K} + \mathbf{x}\mathbf{I})^{-1})\partial\alpha$$

$$\frac{\partial f}{\partial \alpha} = Tr((\mathbf{K} + \mathbf{x}\mathbf{I})^{-1})$$

(d) $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log\left(\mathcal{N}(\mathbf{a}|\mathbf{A}\boldsymbol{\mu}, \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top)\right)$, derive $\frac{\partial f}{\partial \boldsymbol{\mu}}$ and $\frac{\partial f}{\partial \boldsymbol{\Sigma}}$,
Answer
log of a PDF is defined as

$$f(\mathbf{a}) = \frac{1}{\mathbf{a}\mathbf{S}\Sigma\mathbf{S}^T\sqrt{2\pi}}\exp(\frac{-(\log(a - \mathbf{A}\mu)^2}{2(\mathbf{S}\Sigma\mathbf{S}^T)^2})$$

(e) $f(\boldsymbol{\Sigma}) = \log\left(\mathcal{N}(\mathbf{a}|\mathbf{b}, \mathbf{K} \otimes \boldsymbol{\Sigma})\right)$ where $\otimes$ is the Kronecker product (Hint: check Minka's notes).
Answer
log of a PDF is defined as

$$f(\mathbf{a}) = \frac{1}{\mathbf{a}\mathbf{K} \otimes \boldsymbol{\Sigma}\sqrt{2\pi}}\exp(\frac{-(\log(a - \mathbf{b})^2}{2(\mathbf{K} \otimes \boldsymbol{\Sigma})^2})$$

10. [2 points] Given the multivariate Gaussian probability density,

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-(\mathbf{x} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right).$$

Show that the density function achieves the maximum when $\mathbf{x} = \boldsymbol{\mu}$.
Answer
Maximum is found out by differentiating the function and setting it to zero.

$$\partial p = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\partial\exp\left(-(\mathbf{x} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

$$\partial p = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-(\mathbf{x} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)(-\partial(\left((\mathbf{x} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right))$$

$$\frac{\partial p}{\partial \mathbf{x}} = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-(\mathbf{x} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)(\left(\left((\mathbf{x} - \boldsymbol{\mu})^\top(\boldsymbol{\Sigma}^{-1T} + \boldsymbol{\Sigma}^{-1})\right)\right))$$

From the equation above we can see that it can be zero only when $\mathbf{x} = \mu$.

11. [5 points] Show that

$$\int \exp(-\frac{1}{2\sigma^2}x^2)\mathrm{d}x = \sqrt{2\pi\sigma^2}.$$

Note that this is about how the normalization constant of the Gaussian density is obtained. Hint: consider its square and use double integral.

Answer

Let

$$Q = \int \exp(-\frac{x^2}{2\sigma^2})dx$$

Replacing $\frac{x}{\sigma} = u$, we get $dx = \sigma du$

$$Q = \sigma \int \exp(-\frac{u^2}{2})du$$

Squaring the above equation we get

$$Q^2 = \sigma^2 \int\int \exp(-\frac{u^2+v^2}{2})dudv$$

Using polar coordinates $u = r\cos\theta; 0 \leq \theta \leq 2\pi$ and $v = r\sin\theta; 0 \leq r \leq \infty$

$$Q^2 = \sigma^2 \int_0^{2\pi}\int_0^{\infty} \exp(-\frac{r^2}{2})|\mathbf{J}|drd\theta$$

$$Q^2 = \sigma^2 \int_0^{2\pi}\int_0^{\infty} \exp(-\frac{r^2}{2})rdrd\theta$$

Let $r^2 = w$, then $dr = \frac{dw}{2r}$

$$Q^2 = \frac{\sigma^2}{2} \int_0^{2\pi}\int_0^{\infty} \exp(-\frac{w}{2})dwd\theta$$

$$Q^2 = -\sigma^2 \int_0^{2\pi} \left[\exp(-\frac{w}{2})\right]_0^{\infty} d\theta$$

$$Q^2 = \sigma^2 \int_0^{2\pi} d\theta$$

$$Q^2 = 2\sigma^2\pi$$

$$Q = \sqrt{2\sigma^2\pi}$$

12. [5 points] The gamma function is defined as

$$\Gamma(x) = \int_0^{\infty} u^{x-1}e^{-u}\mathrm{d}u.$$

Show that $\Gamma(1) = 1$ and $\Gamma(x+1) = x\Gamma(x)$. Hint: using integral by parts.

Answer

$$\Gamma(1) = \int_0^{\infty} \exp(-u)du$$

$$\Gamma(1) = \Big[-\exp(-u) \Big]_0^\infty$$
$$\Gamma(1) = 1$$

$$\Gamma(x+1) = \int_0^\infty u^x e^{-u} du$$
$$\Gamma(x+1) = -u^x e^{-u} + \int_0^\infty x u^{x-1} e^{-u} du$$
$$\Gamma(x+1) = 0 + x\Gamma(x)$$

13. [2 points] By using Jensen's inequality with $f(x) = \log(x)$, show that for any collection of positive numbers $\{x_1, \ldots, x_N\}$,
$$\frac{1}{N} \sum_{n=1}^{N} x_n \geq \left(\prod_{n=1}^{N} x_n \right)^{\frac{1}{N}}.$$

Answer

Taking expectation on both sides

$$\mathbb{E}(f(x)) = \mathbb{E}(\log(x))$$

From Jensen's equality we know that

$$\mathbb{E}(\log(x)) \geq \log(\mathbb{E}(x)))$$

$$\sum_{n=1}^{N} (\log(x_n)) \geq \log(\sum_{n=1}^{N}(x)))$$

$$\log \prod_{n=1}^{N} x_n \geq \log(\sum_{n=1}^{N}(x)))$$

$$\frac{1}{N} \log \prod_{n=1}^{N} x_n \leq \log(\frac{1}{N} \sum_{n=1}^{N}(x)))$$

$$\log(\prod_{n=1}^{N} x_n)^{\frac{1}{N}} \leq \log(\frac{1}{N} \sum_{n=1}^{N}(x)))$$

$$\frac{1}{N} \sum_{n=1}^{N} x_n \geq \left(\prod_{n=1}^{N} x_n \right)^{\frac{1}{N}}$$

14. [2 points] Given two probability density functions $p(\mathbf{x})$ and $q(\mathbf{x})$, show that
$$\int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \mathrm{d}\mathbf{x} \geq 0.$$

Answer

Using Jensen's inequality, we can say that

$$\int p(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x} \leq \log \int p(\mathbf{x}) \frac{q(\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x}$$

$$\int p(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x} \leq \log \int q(\mathbf{x}) \mathrm{d}\mathbf{x}$$

as $\log \int q(\mathbf{x}) \mathrm{d}\mathbf{x} \leq 0$, we can write

$$\int p(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x} \leq 0$$

Now inverting the values in the log, we get

$$-\int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \mathrm{d}\mathbf{x} \leq 0$$

$$\int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \mathrm{d}\mathbf{x} \geq 0$$

15. [**Bonus**][5 points] Show that for any square matrix $\mathbf{X} \succ 0$, $\log |\mathbf{X}|$ is concave to $\mathbf{X}$.

Answer

Let $A = \log |\mathbf{X}|$. To find concavity, we need to find the second order derivative of this equation and show that it's always negative.

$$dA = d \log |\mathbf{X}|$$
$$dA = Tr(\mathbf{X}^{-1}) d\mathbf{X}$$
$$d^2 A = Tr(d\mathbf{X}^{-1}) d\mathbf{X}$$
$$d^2 A = Tr(-\mathbf{X}^{-1} d\mathbf{X} \mathbf{X}^{-1}) d\mathbf{X}$$
$$d^2 A = -Tr(\mathbf{X}^{-1} d\mathbf{X} \mathbf{X}^{-1}) d\mathbf{X}$$

As the matrix at hand is square and positive definite, we can argue that the value returned from the Tr will always be positive. hence the overall 2nd order derivative is always negative, hence it's concave.