CS 6190: Probabilistic Machine Learning Spring 2023

Homework 0

Handed out: 10 Jan, 2023
Due: 11:59pm, 20 Jan, 2023

- 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

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

solution

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$
$$\text{So, } p(A) + p(B) - p(A \cap B) \leq p(A) - p(B)$$

$$p(A \cap B) \leq p(A)$$
$$p(A \cap B) = p(A) - p(A \cap \overline{B})$$
$$\text{So, } p(A) - p(A \cap \overline{B}) \leq p(A)$$

$$p(A \cap B) \leq p(B)$$
$$p(A \cap B) = p(B) - p(B \cap \overline{A})$$
$$\text{So, } p(B) - p(B \cap \overline{A}) \leq p(B)$$

When does the equality hold?

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)

solution

Base Case:
Assume: n = 2

$$p(A_1 \cup A_2) \leq p(A_1) + p(A_2)$$

$$--> p(A_1 \cup A_2) = p(A_1) + p(A_2) - p(A_1 \cap A_2)$$

$$--> p(A_1) + p(A_2) - p(A_1 \cap A_2) \leq p(A_1) + p(A_2)$$

$$--> -p(A_1 \cap A_2) \leq 0$$

Base case proved

Inductive Hypothesis:

Assume $p(\cup_{i=1}^{k} A_i) \leq \sum_{i=1}^{k} p(A_i)$ is true.

Inductive step:

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

$$p(A_1 \cup A_2 \cup ...A_{k+1}) \leq p(A_1) + p(A_2) + ...p(A_{k+1})$$

$$p(A_1 \cup A_2 \cup ...A_{k+1}) = p(A_1) + p(A_2) + ...p(A_{k+1}) - p(A_1 \cap A_2) - p(A_2 \cap A_3) - ...p(A_1 \cap A_2 \cap ...A_{k+1})$$

$$-p(A_1 \cap A_2) - p(A_2 \cap A_3) - ...p(A_1 \cap A_2 \cap ...A_{k+1}) \leq 0$$

Inductive step proved

This inequality holds for all cases. Both when A and B are not mutually exclusive and when A and B are mutually exclusive.

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 $p(Y)$

$$x = 1, p(x) = 2/10 + 4/10 ==> p(x = 1) = \mathbf{6/10}$$
$$x = 0, p(x) = 3/10 + 1/10 ==> p(x = 0) = \mathbf{4/10}$$
$$y = 1, p(y) = 1/10 + 4/10 ==> p(y = 1) = \mathbf{5/10}$$
$$y = 0, p(y) = 3/10 + 2/10 ==> p(y = 0) = \mathbf{5/10}$$

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

$p(X|Y)$:

$$x = 0, y = 1 : (1/10)/(5/10) ==> p(x|y) = \mathbf{1/5}$$
$$x = 1, y = 1 : (4/10)/(5/10) ==> p(x|y) = \mathbf{4/5}$$
$$x = 0, y = 0 : (3/10)/(5/10) ==> p(x|y) = \mathbf{3/5}$$
$$x = 1, y = 0 : (2/10)/(5/10) ==> p(x|y) = \mathbf{2/5}$$

$p(Y|X)$:

$$x = 0, y = 1 : (1/10)/(4/10) ==> p(y|x) = \mathbf{1/4}$$
$$x = 1, y = 1 : (4/10)/(6/10) ==> p(y|x) = \mathbf{2/3}$$
$$x = 0, y = 0 : (3/10)/(4/10) ==> p(y|x) = \mathbf{3/4}$$
$$x = 1, y = 0 : (2/10)/(6/10) ==> p(y|x) = \mathbf{1/3}$$

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

$$\mathbb{E}(X) = 0(4/10) + 1(6/10) = \mathbf{6/10}$$
$$\mathbb{E}(Y) = 0(5/10) + 1(5/10) = \mathbf{5/10}$$
$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$
$$\mathbb{V}(X) = \int p(X)(X - \mathbb{E}(X))^2$$
$$\mathbb{V}(X) = (4/10)(0 - (6/10))^2 + (6/10)(1 - (6/10))^2 = \mathbf{0.168}$$
$$\mathbb{V}(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^2]$$
$$\mathbb{V}(Y) = \int p(Y)(Y - \mathbb{E}(Y))^2$$
$$\mathbb{V}(Y) = (5/10)(0 - (5/10))^2 + (5/10)(1 - (5/10))^2 = \mathbf{0.25}$$

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

$$\mathbb{E}(Y|X=0) = \sum y * p(Y=y|X=0)$$
$$= 0 * (p(Y=0|X=0)) + 1 * (p(Y=1|X=0))$$
$$= 0 * (3/4) + 1 * (1/4)$$
$$\mathbb{E}(Y|X=0) = \mathbf{1/4}$$
$$\mathbb{E}(Y|X=1) = 0 * (p(Y=0|X=1)) + 1 * (p(Y=1|X=1))$$
$$= 0 * (1/3) + 1 * (1/4)$$
$$\mathbb{E}(Y|X=1) = \mathbf{1/4}$$
$$\mathbb{V}(Y|X=0) = \sum p(Y=y|X=0) * (y - \mathbb{E}(Y=y|X=0))^2$$
$$= p(Y=0|X=0) * (0 - \mathbb{E}(Y|X=0))^2 + p(Y=1|X=0) * (1 - \mathbb{E}(Y|X=0))^2$$
$$= 3/4 * (0 - 1/4)^2 + 1/4 * (1 - 1/4)^2$$
$$\mathbb{V}(Y|X=0) = \mathbf{3/16}$$
$$\mathbb{V}(Y|X=1) = \sum p(Y=y|X=1) * (y - \mathbb{E}(Y=y|X=1))^2$$
$$= p(Y=0|X=1) * (0 - \mathbb{E}(Y|X=1))^2 + p(Y=1|X=1) * (1 - \mathbb{E}(Y|X=1))^2$$
$$= 1/3 * (0 - 1/4)^2 + 2/3 * (1 - 1/4)^2$$
$$\mathbb{V}(Y|X=1) = \mathbf{19/48}$$

v. the covariance between X and Y

$$\mathrm{Cov}[X,Y] = \sum \mathbb{E}[(X - \mathbb{E}(X)) * (Y - E(Y))$$
$$= 3/10 * (0 - 6/10) * (0 - 5/10) + 1/10 * (0 - 6/10) * (1 - 5/10) + 2/10 * (1 - 6/10) * (0 - 5/10) + 4/10 * (1 - 6/10) * (1 - 5/10)$$
$$\mathrm{Cov}[X,Y] = \mathbf{0.1}$$

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

$$p(X=0 \cap Y=0) = 3/10 \mathrel{!}= 20/10 = p(X=0)p(Y=0)$$
$$p(X=0 \cap Y=1) = 1/10 \mathrel{!}= 20/10 = p(X=0)p(Y=1)$$
$$p(X=1 \cap Y=0) = 2/10 \mathrel{!}= 30/10 = p(X=1)p(Y=0)$$
$$p(X=1 \cap Y=1) = 4/10 \mathrel{!}= 30/10 = p(X=1)p(Y=1)$$

No, because $p(A)! = p(A)P(B)$ as shown above, and for independence, you need $p(A \cap B) = p(A)p(B)$ for all cases.

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

$$\mathbb{E}(Y|X=0) = 1/4 == \mathbb{E}(Y|X=1) = 1/4$$
$$\mathbb{V}(Y|X=0) = 3/16 \mathrel{!}= \mathbb{V}(Y|X=1) = 19/48$$

$\mathbb{E}(Y|X)$ is constant, but $\mathbb{V}(Y|X)$ is not constant as shown above.

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) $\mathbb{E}(Y)$

$$\mathbb{E}(Y) = \int g(x)p(x)dx \text{ for functions with continuous random variables}$$
$$\mathbb{E}(Y) = \int e^{-x^2} * (e^{-(x^2)/2}/(\sqrt{2\pi}))dx$$
$$= 1/\sqrt{2\pi} \int e^{-3x^2/2} dx$$

$$\text{we will say that } u = \sqrt{3x^2/2} sou^2 = 3x^2/2$$
$$\mathbb{E}(Y) = 1/(\sqrt{3\pi}) \int e^{-t^2} dt$$
$$= 1/(\sqrt{3\pi}) * \sqrt{\pi}$$
$$= 1/(\sqrt{3})$$

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

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$$
$$= \int (e^{-x^2} * (e^{-(x^2)/2}/(\sqrt{2\pi}))dx)^2 - (1/\sqrt{3})^2$$

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

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

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

   (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 $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1/2 \\ -1/3 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$.

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

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

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}[\sum y * p(Y = y|X = x])$$
$$= \sum_x [\sum_y y * p(Y = y|X = x)] * p(X = x)$$
$$= \sum_x \sum_y y * p(Y = y, X = x)$$
$$= \sum_y \sum_x y * p(Y = y, X = x)$$
$$= \sum_y y \sum_x p(Y = y, X = x)$$
$$= \sum_y y * p(Y = y)$$
$$= \mathbb{E}(X)$$

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

$$V(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$$
$$\mathbb{E}(Y^2) = \mathbb{E}(\mathbb{E}(Y^2|X)) \text{ - law of total expectation}$$
$$\mathbb{E}(Y^2) = \mathbb{E}(\mathbb{V}(Y|X) + (\mathbb{E}(Y|X))^2 \text{ - definition of variance}$$
$$\text{So, } \mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X) + (\mathbb{E}(Y|X))^2) - (\mathbb{E}(\mathbb{E}(Y|X)))^2 \text{ - substitution}$$
$$\mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = (\mathbb{E}(\mathbb{V}(Y|X)) + (\mathbb{E}(\mathbb{E}(Y|X)^2) - (\mathbb{E}(\mathbb{E}(Y|X)))^2) \text{ - rearranging}$$
$$\mathbb{E}(\mathbb{E}(Y|X)^2) - (\mathbb{E}(\mathbb{E}(Y|X)))^2 = \mathbb{V}(\mathbb{E}(Y|X)) \text{ - definition of variance}$$
$$\mathbf{V(Y) = E(V(Y—X)) + V(E(Y—X))} \text{ - substitution}$$

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

$$\sigma = 1/(1 + e^{(-a^\top x)})$$
$$df(x)/dx = (-1 + e^{-a^\top x})^{-2} d/dx(1 + e^{-a^\top x}))$$
$$= (-1 + e^{-a^\top x})^{-2} * (d/dx[1] + d/dx[e^{-a^\top x}])$$
$$= (-1 + e^{-a^\top x})^{-2} * (0 + (-a * e^{-a^\top x}))$$
$$df(x)/dx = (a^\top e^{-a^\top x})/(1 + e^{-a^\top x})^{-2}$$

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

$$\frac{\mathrm{d}^2 f(\mathbf{x})}{\mathrm{d}\mathbf{x}^2} = df(x)/dx = d/dx((a^\top * e^{-a^\top x})/(1 + e^{-a^\top x})^2$$
$$= ((1 + e^{-a^\top x})^2 * (-a^\top a e^{-a^\top x}) - (a^\top e^{-a^\top x}) * (-2a^\top e^{-a^\top x}) * (e^{-a^\top x} + 1))/(1 + e^{-a^\top x})^4$$
$$= ((1 + e^{-a^\top x}) * (-a^\top a e^{-a^\top x}) - (a^\top e^{-a^\top x}) * (-2a^\top e^{-a^\top x}))/(1 + e^{-a^\top x})^3$$
$$= ((1 + e^{-a^\top x}) * (-a^\top a e^{-a^\top x}) - (2a^\top a e^{-a^\top x})/(1 + e^{-a^\top x})^3$$

$$\text{Hessian matrix of } \frac{\mathrm{d}^2 f(\mathbf{x})}{\mathrm{d}\mathbf{x}^2} \text{ in indicial notation:}$$
$$= ((1 + e^{-a_j x}) * (-a_j a_i e^{-a_j x}) - (2 a_j a_i e^{-a_j x})/(1 + e^{-a_j x})^3$$

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

$$\nabla(-log(f(\mathbf{x}))) = d/dx(-log(f(x))) = -1/f(x)$$
$$\nabla^2(-log(f(\mathbf{x}))) = d/dx(-1/f(x)) = 1/(f(x)^2)$$
Since $0 \leq f(x) \leq 1, \nabla^2(-log(f(\mathbf{x}))) = 1/(f(x)^2)$ is always ≥ 0, which means it is convex.

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

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

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

f is differentiable ($f'(x) = -1/x$) and the max gap occurs at $f'(x) = y$ so the max is at $-1/x = y$(slope) or $x = -1/y$(x-point) which makes the y-point $=$ $f(-1/y) = -ln(-1/y) = ln(-y)$.
which gives us in point slope form: $y - (ln(-y)) = y(x - (-1/y))$
@ x = 0, y=ln(-y) + 1
f*(y) = -ln(y) - 1

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

f*(y) = max($y^T x - x^T A^{-1} x$)
We differentiate and set equal to 0 to get the max
$$y - 2A^{-1}x = 0$$
solving for x and substituting into f*(y) we get
$$f*(y) = y^T 1/2Ay - 1/2A^T y^T A^{-1} 1/2Ay$$
$$f*(y) = y^T 1/2Ay - 1/4A^T y^T y$$

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 $\frac{\partial f}{\partial \mathbf{x}}$.

$$df = d(x^T A x)$$
lets make $w = Ax$
so $f = x^T w$
$$df = x^T d(w) + wd(x)^T$$
$$df = 2Ax^T dx$$
$$df/dx = 2Ax^T$$

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

define $B = I + xx^T$
$$dy = d(B^{-1}x)$$
$$dy = d(B^{-1})x + B^{-1}d(x)$$
$$dy = -B^{-1}(d(B))B_B^{-1-1}d(x)$$
we can define $d(B) = dx * x^T + x * dx^T$
so, $dy = (-B^{-1}B^{-1}dx * x^T - B^{-1}B^{-1}x * dx^T) + B^{-1}dx$
since B is symmetric, the transpose is the same
$$dy = -2(x^T B^{-1} B^{-1} dx) + B^{-1} dx$$
$$dy/dx = -2x^T B^{-1} B^{-1} + B^{-1}$$

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

$$df = d(log(det(K + \alpha I)))$$
$$B = K + \alpha I$$
$$df = Tr(B^{-1}d(B))$$
$$d(B) = d(K) + d(\alpha I)$$
$$d(B) = dK + \alpha d(I)$$
assuming K and I are constants
$$d(B) = 0$$
assuming K and I are both symmetric, the inverse is same as the non-inverse
$$df = Tr(Bd(B))$$
$$df = Tr((k + \alpha I)(0))$$
$$df/d\alpha = 0$$

(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 $\frac{\partial f}{\partial \boldsymbol{\Sigma}}$ (Hint: check Minka's notes about the definition of gradient w.r.t a matrix).

We will say that $w = \mathbf{A}\boldsymbol{\mu}$ and $C = \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top$
$$f(w, C) = log(1/\sqrt{(2\pi)^{n/2}}) - log(det(C))/2 + log(e^{-1/2(x-w)^\top C^{-1}(x-w)})$$
$$f(w, C) = log(1/\sqrt{(2\pi)^{n/2}}) - 1/2 * (x - w)^\top C^{-1}(x - w)$$
$$df/dw = -1/2d((x - w)^\top C^{-1}(x - w))/dw$$
$$= (-1/2)d((x - w)^\top C^{-1}(x - w))/d(x - w) * d(x - w)/dw$$
$$= -(x - w)^\top (C^{-1} + C^{-1\top})$$
$$df/dw = 1/2(x - w)^\top (C^{-1} + (C^{-1})^\top)$$
assuming C is symmetric, C^{-1} and its transpose are the same
$$= 1/2(x - w)^\top (2C^{-1})$$
$$= (x - w)^\top C^{-1}$$
$$df/d\boldsymbol{\mu} = (x - \mathbf{A}\boldsymbol{\mu})^\top (\mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top)^{-1}$$
-
$$df/dC = (-1/2)(d((x - w)^\top C^{-1}(x - w)))/(dC) - d(log(det(C)))/2dC$$
$$= -1/2(d(tr((x - w)^\top C^{-1}(x - w))))/(dC) - (1/2det(C))(d(det(C)))/dC$$
$$= -1/2tr((d((x - w)^\top C^{-1}(x - w))))/(dC) - tr(C^{-1})/2$$
$$= 1/2tr(((x - w)(x - w)^\top C^{-1}C^{-1}))) - tr(C^{-1})/2$$
$$= 1/2tr(((x - w)(x - w)^\top 2C^{-1}))) - tr((C^{-1})/2)\mathbf{S}\mathbf{S}^\top$$

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

The derivative of the kronecker product is
$$d(\mathbf{K} \otimes \boldsymbol{\Sigma})/d(\boldsymbol{\Sigma}) = d(\mathbf{K})/d(\boldsymbol{\Sigma}) \otimes \boldsymbol{\Sigma} + \mathbf{K} \otimes d/d(\boldsymbol{\Sigma}) * (\boldsymbol{\Sigma})$$
$$= \mathbf{K} \otimes \mathbf{I}$$
subsituting this into the derivative we got from the second part of part d and we get...
$$df/d\boldsymbol{\Sigma} = (1/2tr(((x - w)(x - w)^\top 2C^{-1}))) - tr(C^{-1})/2)(\mathbf{K} \otimes \mathbf{I})$$

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 $\mathbf{x} = \boldsymbol{\mu}$.

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.

say that $z = \int \exp(-\frac{1}{2\sigma^2}x^2)\mathrm{d}x$

squaring this and using a double integral, we get

$$z^2 = \int_0^\infty \int_0^\infty \exp(-\frac{x^2+y^2}{2\sigma^2})\mathrm{d}x\mathrm{d}y$$

turning x,y coordinates into polar coordinates, we can get

$$z^2 = \int_0^{2\pi} \int_0^\infty r\exp(-\frac{r^2}{2\sigma^2})\mathrm{d}r\mathrm{d}\theta$$

Evaluating this double integral gives us $z = \sigma\sqrt{(2\pi)} = \sqrt{2\pi\sigma^2}$

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.

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

$$\Gamma(x+1) = \int_0^\infty u^x e^{-u}\mathrm{d}u.$$
$$\Gamma(x+1) = \int_0^\infty u^x e^{-u}\mathrm{d}u.$$
$$\Gamma(x+1) = [-u^x e^{-u}]_0^\infty + \int_0^\infty xu^{x-1}e^{-u}\mathrm{d}u.$$
$$\Gamma(x+1) = \lim_{u\infty}(-u^x e^{-u}) - (-0^x e^{-0}) + x\int_0^\infty u^{x-1}e^{-u}\mathrm{d}u$$
We can see that $-u^x e^{-u} --> 0$ as $u --> \infty$
$$\Gamma(x+1) = x\int_0^\infty u^{x-1}e^{-u}\mathrm{d}u$$
$$\Gamma(x+1) = x\Gamma(x)$$

13. [2 points] By using Jensen's inequality with $f(x) = \log(x)$, show that for any collection of positive numbers $\{x_1, \ldots, x_N\}$,

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

14. [2 points] Given two probability density functions $p(\mathbf{x})$ and $q(\mathbf{x})$, show that

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

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

$$= -\mathbb{E}[\log\tfrac{q(\mathbf{x})}{p(\mathbf{x})}]$$
(jensens inequality for concave function)

$$\int p(\mathbf{x})\log\frac{p(\mathbf{x})}{q(\mathbf{x})}\mathrm{d}\mathbf{x} \geq -\log(\mathbb{E}[\frac{q(\mathbf{x})}{p(\mathbf{x})}])$$

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

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

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

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