CSCI-GA.2565-001 Machine Learning: Homework 0

Due 5pm EST, February 8, 2022 on Gradescope

We encourage LaTeX-typeset submissions but will accept quality scans of hand-written pages.

1 Joint Distributions and Independence

Let X, Y, Z have joint density $p(X, Y, Z) = p(X)p(Y|X)p(Z|X)$.

(A) Use law of total probability to write down the marginal distribution of Y in terms of $p(X), p(Y|X), p(Z|X)$

Given,

$$p(X, Y, Z) = p(X)p(Y/X)p(Z/X)$$

Now,

$$P(Y) = \sum_x \sum_z p(X)p(Y/X)p(Z/X)$$

$$P(Y) = \sum_x p(X)p(Y/X) \sum_z p(Z/X)$$

For the conditional of Z given X,

$$\sum_z p(Z/X) = 1$$

Therefore,

$$P(Y) = \sum_x p(X)p(Y/X)$$

(B) Use Bayes rule to write down the conditional distribution $Z|Y$ in terms of $p(X), p(Y|X), p(Z|X)$

$$p(Z/Y) = \frac{p(Y, Z)}{p(Y)}$$

$$P(Y, Z) = \sum_x p(X, Y, Z) = \sum_x p(X)p(Y/X)p(Z/X)$$

From 1.A,

$$P(Y) = \sum_x p(X)p(Y/X)$$

Therefore,

$$p(Z/Y) = \frac{\sum_x p(X)p(Y/X)p(Z/X)}{\sum_x p(X)p(Y/X)}$$

(C) Without further assumptions, which variables are independent? Which are conditionally independent?

$$p(Y, Z/X) = p(X, Y, Z)/p(X)$$

$$p(Y, Z/X) = \frac{p(X)p(Y/X)p(Z/X)}{p(X)}$$

$$p(Y, Z/X) = p(Y/X)p(Z/X)$$

Therefore, Y and Z are conditionally independent given X.

X,Y and Z are not independent.

(D) Conditional Specification.
We study two jointly distributed variables X, Y with conditional densities $p(Y = y|X = x) = g(x, y)$ and $p(X = x|Y = y) = h(x, y)$. What conditions if any do g, h need to satisfy (i.e. to specify a well-defined joint density)?
Hint: How can you relate $p(X = x|Y = y)$ and $p(Y = y|X = x)$ using an equality, possibly involving the marginals of X, Y?

From Bayes rule,

$$p(X/Y)p(Y) = p(Y/X)p(X)$$

Given,

$$p(Y = y|X = x) = g(x, y)$$
$$p(X = x|Y = y) = h(x, y)$$

Let $f(x, y)$ denote the joint distribution. Let $f_1(x)$ and $f_2(y)$ be the marginals for X and Y respectively.

$$h(x, y) * f_2(y) = g(x, y) * f_1(x)$$

$$\frac{h(x, y)}{g(x, y)} = \frac{f_1(x)}{g_1(y)}$$

As there exists a conditional PDF of X, given Y = y, that is given by $fX|Y(x|y) = f(x, y)/f_2(y)$ we have $f_2(y) > 0$. And similarly for p(Y=y/X=x).

If our conditional distributions are defined, we can assume finite marginals. The ratio of conditional probabilities is the ratio of marginal probabilities. If P(X=x/Y=y) is non zero, then for finite marginals, P(Y=y/X=x) is also non zero and vice versa. If P(X=x/Y=y) is zero, then P(Y=y/X=x) will also be zero and vice versa.

$$\int_y P(Y/X = x)dy = 1$$

$$\int_x P(X/Y = y)dx = 1$$

(E) Construct two continuous random variables X, Y and a non-constant function f such that $f(X, Y)$ is independent of X and $f(X, Y)$ is independent of Y. If impossible, explain why.

Consider the joint pdf as follows.

R.V. X such that x ranges from 0 to 1. R.V. Y such that y ranges from 0 to 1.

Case 1: $y \leq 1/4$
newline

$f(x,y) = 1$ if $x \in [0, 1/4]$

$f(x,y) = 0$ if $x \in (1/4, 1/2]$

$f(x,y) = 1$ if $x \in (1/2, 3/4]$

$f(x,y) = 0$ if $x \in [3/4, 1]$

Case 2: $1/4 < y \leq 1/2$

$f(x,y) = 0$ if $x \in [0, 1/4]$

$f(x,y) = 1$ if $x \in (1/4, 1/2]$

$f(x,y) = 0$ if $x \in (1/2, 3/4]$

$f(x,y) = 1$ if $x \in [3/4, 1]$

Case 3: $1/2 < y \leq 3/4$

$f(x,y) = 1$ if $x \in [0, 1/4]$

$f(x,y) = 0$ if $x \in (1/4, 1/2]$

$f(x,y) = 1$ if $x \in (1/2, 3/4]$

$f(x,y) = 0$ if $x \in [3/4, 1]$

Case 4: $3/4 < y \leq 1$

$f(x,y) = 0$ if $x \in [0, 1/4]$

$f(x,y) = 1$ if $x \in (1/4, 1/2]$

$f(x,y) = 0$ if $x \in (1/2, 3/4]$

$f(x,y) = 1$ if $x \in [3/4, 1]$

f(X,Y) is either 0 or 1.

For any X=x', there will be two intervals of Y of length 1/4 each where f(x,y) is 1. Else f(x',y) is 0. We use the geometrical interpretation of probability. $p(f(X,Y) = 0/X = x')$ will be 1/2 regardless of the X value chosen. Similarly for $p(f(X,Y) = 1/X = x')$ So f(X,Y) is independent of X.

For any Y=y', there will be two intervals of Y of length 1/4 each where f(x,y) is 1. Else f(x,y') is 0. $p(f(X,Y) = 0/X = x')$ will be 1/2 regardless of the Y value chosen. Similarly for $p(f(X,Y) = 1/Y = y')$ So f(X,Y) is independent of Y.

2 Moments

(A) Construct a random variable X, such that $\mathbb{P}(X < \infty) = 1$, but $\mathbb{E}[X] = \infty$. Show both properties.

Consider a sequence of random variables $X_1, X_2, X_3.....X_n$

$$P(X_n = n^2) = \frac{1}{n}$$

$$P(X_n = 0) = 1 - \frac{1}{n}$$

$$E[X_n] = \frac{1}{n}.n^2 + 0\{1 - \frac{1}{n}\}$$

$$\lim_{n\to\infty} E[X_n] = n = \infty$$

If n tends to infinity.

$$P(X < \infty) = P(X_n = 0) = 1 - \frac{1}{n}$$

$$P(X < \infty) = 1$$

Since 1/n tends to 0.

If n does not tend to infinity,

$$P(X_n < \infty) = P(X_n = 0) + P(X_n = n^2)$$
$$P(X_n < \infty) = 1 - \frac{1}{n} + \frac{1}{n}$$
$$P(X_n < \infty) = 1$$

X_n is our desired R.V.

(B) Prove $\mathbb{E}[Y] = \mathbb{E}\Big[\mathbb{E}[Y|X]\Big]$.

$$E[Y/X] = \int_y p(y/x).y.dy$$

$$E[E[Y/X]] = \int_x p(x) \left(\int_y p(y/x).y.dy \right) dx$$

$$E[E[Y/X]] = \int_x \int_y p(y/x)p(x)dydx$$

Adjusting the order of integration,

$$E[E[Y/X]] = \int_y \left(\int_x p(y/x)p(x)dx \right) dy$$

$$E[E[Y/X]] = \int_y \int_x p(y,x)dxdy$$

$$E[E[Y/X]] = \int_y p(y)dy$$

$$E[E[Y/X]] = E[Y]$$

Note: Since the inner expectation is a function of only X, we will generally omit the subscript on the outer expectation since it has to correspond to X.

3 Some Normal Math

In the section below, we use "Gaussian" and "Normal" interchangeably. The univariate Gaussian $\mathcal{N}(\mu, \sigma^2)$ with mean μ and variance $\sigma^2 > 0$ has PDF

$$p(X = x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big[-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big]$$

(A) Given a sample $X \sim \mathcal{N}(0, 1)$, specify a function f (not relying on any other random variables) such that $f(X) \sim \mathcal{N}(3, 2)$

For X $X \sim \mathcal{N}(0, 1)$ if we want to get Y such that $Y \sim \mathcal{N}(3, 2)$, we use the property of Linear transformations of a Gaussian distribution.

If R.V. X is a Gaussian parameterized by μ, σ^2, Let,

$$Y = aX + b$$

then, Y is given by $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$,

Therefore,

$$Y = f(X) = \sqrt{2}X + 3$$

(B) Given a sample $X \sim \mathcal{N}(0, 1)$, name a random variable Y such that $X + Y \sim \mathcal{N}(3, 2)$.

Let

$$X + Y = Z$$

where,

$$Z \sim \mathcal{N}(3, 2)$$

$$Y = Z - X$$

Linearity of expectation,

$$\mu_Y = \mu_Z - \mu_X$$

Using independence,

$$Var[Z - X] = Var(Z) - Var(X)$$

$$Y \sim \mathcal{N}(3, 1)$$

From part A,

(C) Let μ be a D dimensional real vector. Let Σ be a $D \times D$ positive semi-definite matrix. The multivariate Gaussian PDF in D dimensions with mean μ and covariance Σ is:

$$p(X = x) = det(2\pi\Sigma)^{-\frac{1}{2}} \exp\Big[-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big]$$

The marginals of each dimension are normal with $X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$. The 2D case is called the Bivariate Normal. Let $X = [X_1, X_2]$ be Bivariate Normal $\mathcal{N}(\mu, \Sigma)$ with

$$\mu = [\mu_1, \mu_2], \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & c \\ c & \sigma_2^2 \end{bmatrix}$$

such that Σ is positive semi-definite. Letting $\rho = \frac{c}{\sigma_1\sigma_2}$, the 2D case can be written as $p(X_1 = x_1, X_2 = x_2) =$

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\Big[\Big(\frac{x_1-\mu_1}{\sigma_1}\Big)^2 - 2\rho\Big(\frac{x_1-\mu_1}{\sigma_1}\Big)\Big(\frac{x_2-\mu_2}{\sigma_2}\Big) + \Big(\frac{x_2-\mu_2}{\sigma_2}\Big)^2\Big]\right]$$

Compute the conditional density $p(X_1 = x_1|X_2 = x_2)$.

Hint: Using either form for the 2D Normal PDF, start with Bayes rule and remember that the marginals are Gaussian with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. You may also use the fact that conditionals of Gaussians are Gaussian. Since Gaussians are fully specified by their mean and variance, this means you only need to identify the mean and variance of $p(X_1 = x_1|X_2 = x_2)$.

$p(X_1, X_2) =$

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right]$$

Let,

$$\delta = \left[-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right]$$

$$p(X_1/X_2) = \frac{p(X_1, X_2)}{\int_{x_1} p(X_1, X_2)dx_1}$$

The normalization of the joint with respect to X_1 will give us the conditional for X_1. The denominator will be a function of X_2. After the normalization of the joint, with respect to X_1, we still get a exponential containing a quadratic form in X_1 which is analogous to the form of the Gaussian distribution.

The term containing x_1, x_2 is delta.

Therefore if we look at the quadratic form in delta, make it a quadratic in X_1, and adjust for the constants, we can get the conditional distribution of X_1/X_2.

$$\delta = \left[-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1\mu_1}{\sigma_1^2}\right) + \frac{\mu_1^2}{\sigma_1^2} - \left(\frac{2\rho}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2}\right)x_1 + constant\right)\right]\right]$$

$$\delta = \frac{1}{2(1-\rho^2)}\left[-\frac{x_1^2}{\sigma_1^2} + \frac{2x_1}{\sigma_1^2}\left(u_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2-u_2)\right) + constant\right]$$

$$\delta = -\frac{1}{2}\left[\frac{1}{(1-\rho^2)}\frac{x_1^2}{\sigma_1^2} + \frac{2x_1}{\sigma_1^2(1-\rho^2)}\left(u_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2-u_2)\right) + constant\right]$$

We recollect the equation for the Gaussian distribution,

$$p(X = x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Therefore,

$$\delta_{standard} = -\frac{1}{2}\left[\frac{x^2}{\sigma^2} - \frac{2x\mu}{\sigma} + constant\right]$$

Where μ is the mean and σ^2 is the variance.

Comparing with the form of the standard Gaussian, as discussed,

$$mean_{X_1/X_2} = \left(u_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2-u_2)\right)$$

$$variance_{X_1/X_2} = \sigma_1^2(1-\rho^2)$$

$$p(X_1/X_2) = \mathcal{N}\left(\left(u_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2-u_2)\right), \sigma_1^2(1-\rho^2)\right)$$

(D) Construct a pair of variables X, Y that have $Cov(X, Y) = 0$ but X is not independent of Y. Is this possible if X, Y are jointly Gaussian?

Let X be a R.V. such that

$$p(X = 3) = 1/3$$
$$p(X = 0) = 1/3$$
$$p(X = -3) = 1/3$$
$$E(X) = \frac{1}{3}3 + \frac{1}{3}0 + \frac{1}{3}(-3) = 0$$

Let Y be a R.V. such that $Y = X^2$

Y is dependent on X because Y is a direct function of X.

$$E(Y) = \frac{1}{3}9 + \frac{1}{3}0 + \frac{1}{3}9 = 6$$

$$Cov(X, Y) = E[(X(Y - 6)]$$
$$Cov(X, Y) = E[XY] - 6E[X]$$

$$E[XY] = \sum_x \sum_y p(X = x, Y = y)xy$$

$$E[XY] = 1/3(-3.9 + 0.0 + 3.9) = 0$$

$$-6E[X] = -6.0 = 0$$

Hence,

$$Cov(X, Y) = 0$$

Thus we get our desired example.

Consider $X = [X_1, X_2]$ be Bivariate Normal $\mathcal{N}(\mu, \Sigma)$ with

$$\mu = [\mu_1, \mu_2], \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & c \\ c & \sigma_2^2 \end{bmatrix}$$

such that Σ is positive semi-definite. Having $\rho = \frac{c}{\sigma_1 \sigma_2}$.

If X_1, X_2 are jointly Gaussian and co-variance of X,Y is zero, This would imply,

$$\rho \sigma_1 \sigma_2 = 0$$

For non zero variances, rho is zero.

$p(X_1 = x_1, X_2 = x_2) =$

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1 - \mu_1}{\sigma_1}\right)\left(\frac{x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 \right] \right]$$

If rho is zero, we will get the following.
$p(X_1 = x_1, X_2 = x_2) =$

$$\frac{1}{2\pi\sigma_1\sigma_2}\exp\left[-\frac{1}{2}\Big[\Big(\frac{x_1-\mu_1}{\sigma_1}\Big)^2+\Big(\frac{x_2-\mu_2}{\sigma_2}\Big)^2\Big]\right]$$

This can be factorized as,

$$p(X_1 = x_1, X_2 = x_2) = \frac{1}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right)\frac{1}{\sigma_2\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right)$$

Therefore,

$$p(X_1, X_2) = p(X_1)p(X_2)$$

Which would mean that X_1, X_2 are independent.

Therefore, in the jointly Gaussian case, for non zero variances, if Cov(X,Y) is zero, then, X, Y are independent.

4 Monte Carlo

Suppose you are given N samples independently drawn from some distribution D. Let's call these samples $\{X_i\}_{i=1}^N$. You are given that the variance is finite, that is $\mathbf{Var}_{X\sim D}[X] = \sigma^2 < \infty$.

(A) Is the mean of the random variable $X \sim D$ finite, i.e. $\mu = \mathbb{E}_{X\sim D}[X] < \infty$? If yes, why? If not, construct an example.

For the distribution D,

The variance exists and is finite. Using a transformation of the coordinates, The second order moments exist and are finite. If a moment of order t exists for any point, moments of lesser order exist for every other point in the probability space.

$$E[X] = \int_x xp(x)dx$$

$$E[X] = \int_{x\leq 1} xp(x)dx + \int_{x>1} xp(x)dx$$

We can see that,

$$\int_{x\leq 1} xp(x)dx \leq \int_{x\leq 1} p(x)dx \leq p(X \leq 1) : -finite$$

$$\int_{x>1} xp(x)dx \leq \int_{x>1} x^2p(x)dx \leq EX^2 : -finite$$

Also, the expression for Variance $E[X-\mu]^2$ is contingent on a μ value which is finite. Therefore given a finite variance, we have a finite mean.

(B) One estimate of the mean of $\mu = \mathbb{E}(X)$ is $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^N X_i$. Find the mean and variance of $\hat{\mu}_N$.

$$\overline{x} = \frac{\sum_{i=1}^N X_i}{N}$$

$$E[\overline{x}] = E\left[\frac{\sum_{i=1}^N X_i}{N}\right]$$

$$E[\overline{x}] = \frac{\sum_{i=1}^N E[X_i]}{N}$$

For X_i coming from the distribution,

$$E[X_i] = \mu$$

Therefore,

$$E[\overline{x}] = \frac{N\mu}{N}$$

$$E[\overline{x}] = \mu$$

$$Var(\overline{x}) = Var\left(\frac{\sum_{i=1}^N X_i}{N}\right)$$

$$Var(\overline{x}) = \frac{1}{N^2} Var\left(\sum_{i=1}^N X_i\right)$$

$X_1, X_2, ..X_i$ are independently drawn.

$$Var(\overline{x}) = \frac{1}{N^2}\left(\sum_{i=1}^{N} Var(X_i)\right)$$

Now,

$$Var(X_i) = \sigma^2$$

$$Var(\overline{x}) = \frac{N\sigma^2}{N^2}$$

$$Var(\overline{x}) = \frac{\sigma^2}{N}$$

(C) Suppose that $\mathbf{Var}_{X\sim D}[X] = \sigma^2 < \infty$ and $\mu = \mathbb{E}[X] < \infty$, then prove that for any $k > 0$ we have the following inequality:

$$\mathbb{P}(|X - \mu| > k) \leq \frac{\sigma^2}{k^2}.$$

Now,

$$E[X] = \int_{x\leq t} xp(x)dx + \int_{x>t} xp(x)dx$$

$$\int_{x>t} xp(x)dx \geq tP(X > t)$$

$$E[X] \geq tP(X > t)$$

Therefore we get,

$$P(X > t) \leq \frac{E[X]}{t}$$

Let $Y = [X - \mu]^2$ Substituting for Y in the above equation,

$$P(Y > k^2) \leq \frac{E[X - \mu]^2}{k^2}$$

$$P([X - \mu]^2 > k^2) \leq \frac{\sigma^2}{k^2}$$

$$P(|X - \mu| > k) \leq \frac{\sigma^2}{k^2}$$

Hence proved.

(D) Let $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $\mathbb{E}[X_i] < \infty, \mathbf{Var}(X) < \infty$. Using parts (b) and (c), prove that for any $k > 0$:

$$\lim_{N\to\infty} \mathbb{P}(|\hat{\mu}_N - \mu| > k) = 0.$$

From part C

$$P(|\hat{\mu}_N - \mu| > k) \leq \frac{\hat{\sigma_N}^2}{k^2}$$

From part B

$$\hat{\sigma_N}^2 = Var(\overline{x}) = \frac{\sigma^2}{N}$$

$$P(|\hat{\mu}_N - \mu| > k) \leq \frac{\sigma^2}{k^2 N}$$

As N tends to infinity, $\frac{\sigma^2}{k^2 N}$ tends to zero.

Therefore,

$$\lim_{N \to \infty} \mathbb{P}(|\hat{\mu}_N - \mu| > k) = 0.$$

5 KL Divergence

Suppose we have a strictly-convex function f and X is a (non-constant) random variable. Jensen's inequality states that:

$$f\left(\mathbb{E}(X)\right) < \mathbb{E}\left(f(X)\right).$$

One way to measure the similarity between two distributions P, Q is the KL divergence, which is defined using their densities p, q as:

$$KL(P||Q) = \int_{x\in\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} dx.$$

The KL is non-negative and is 0 if and only if the two distributions are equal. These properties also hold when P, Q are discrete.

Assume that the densities $p(x), q(x) > 0$ for all $x \in \mathbb{R}$. Prove the following two statements:

- when $P = Q$, $KL(P||Q) = 0$.
- when $P \neq Q$, $KL(P||Q) > 0$ (strict inequality). **Hint:** Use Jensen's inequality.

$$KL(P||Q) = \int_{x\in\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} dx.$$

$$KL(P||Q) = -\int_{x\in\mathbb{R}} p(x) \log \frac{q(x)}{p(x)} dx.$$

$$KL(P||Q) = E\left[-ln\frac{q(x)}{p(x)}\right]$$

Since -lnx is a convex function, Using Jensens inequality where q(x)/p(x) is non constant,

$$E\left[-ln\frac{q(x)}{p(x)}\right] > -ln\left[E\left(\frac{q(x)}{p(x)}\right)\right]$$

Therefore,

$$KL(P||Q) > -ln\left(\int p(x)\frac{q(x)}{p(x)}dx\right)$$

$$KL(P||Q) > -ln\left(\int q(x)dx\right) > -ln1$$

$$KL(P||Q) > 0$$

when P=Q,

$$KL(P||Q) = \int_{x\in\mathbb{R}} p(x) \log 1 dx.$$

$$KL(P||Q) = 0$$

6 Calculus

(A) Let $Y \sim \text{Exp}(\lambda)$. That is, Y is distributed according to the Exponential distribution with parameter λ. Let $\text{Exp}(y; \lambda)$ denote the evaluation of the exponential density at the value $Y = y$. Use (univariate) calculus to maximize $\text{Exp}(2; \lambda)$ with respect to λ. We suggest maximizing the log of the density.

$\text{Exp}(y; \lambda)$ denotes the evaluation of the exponential density at the value $Y = y$

$$f(y; \lambda) = \lambda \exp(-\lambda y)$$

$$f(2; \lambda) = \lambda \exp(-\lambda 2)$$

looking at the log likelihood,

$$L(\lambda) = \log f(2; \lambda) = -2\lambda + \log \lambda$$

Taking the derivative and equating to zero,

$$\frac{dL(\lambda)}{d\lambda} = 0$$

$$-2 + \frac{1}{\lambda} = 0$$

$$\lambda = 1/2$$

$$L''(\lambda) = \frac{-1}{\lambda^2}$$

for $\lambda = 1/2$

$$L''(\lambda) < 0$$

Therefore we have a maxima at $\lambda = 1/2$. Correspondingly,

$$f(2; 1/2) = \frac{1}{2e}$$

(B) Let $x \in \mathbb{R}^2$ be a 2 dimensional real vector where $x = [x_1, x_2]$. Define the scalar-valued function $f : \mathbb{R}^2 \to \mathbb{R}$ by:

$$f(x) = \exp[\log(x_1^2) + x_1 x_2]$$

Compute $\nabla_x f$.

$$f(x) = \exp[\log(x_1^2) + x_1 x_2]$$

$$\nabla_x f = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right]$$

$$\nabla_x f = \left[\exp[\log(x_1^2) + x_1 x_2] * \left(\frac{1}{x_1^2}(2x_1) + x_2\right), \exp[\log(x_1^2) + x_1 x_2] * x_1\right]$$

$$\nabla_x f = \left[f * \left(\frac{2}{x_1} + x_2\right), f * x_1\right]$$

(C) The CDF of the Exponential distribution is

$$F(y;\lambda) = 1 - \exp[-\lambda y]$$

Derive the Exponential PDF $f(y;\lambda)$.

Given CDF,

$$F(y;\lambda) = 1 - \exp[-\lambda y]$$

PDF for this distribution,

$$\frac{\partial F}{\partial y} = \lambda exp[-\lambda y]$$

$$f(y;\lambda) = \lambda exp[-\lambda y]$$

7 PyTorch

This question is mostly to get you to install PyTorch, one of the two popular machine learning libraries for python (the other being Tensorflow), and to start writing a few lines of sampling code. It should be easy to get started by choosing your system settings on this page https://pytorch.org/get-started/locally/. The non-GPU version for your regular laptop is fine for our purposes.

Assuming you have installed the library you should be able to `import torch`. We expect you are familiar with basic usage of Numpy, where `np.array`'s are the main data structure. In Torch, the equivalent is a `torch.tensor`:

- `x=torch.tensor([[1.0,2.0],[3.0,4.0]])` is a 2×2 matrix. You can verify the shape by using `x.shape`.
- `tensor`s have lots of convenient methods. Try `x.sum(), x.sum(0), x.sum(1), x.mean(0), x.std(), x.abs(), x.pow(2)` etc... See https://pytorch.org/docs/stable/index.html for more.

For this homework question, we want you to teach yourself how to do the following in PyTorch:

(A) Draw N univariate normal samples $x_i \sim \mathcal{N}(0, \sigma^2)$ for some value of σ^2. For this you will need

`torch.distributions.Normal`

Be sure to give the right arguments (e.g. standard deviation and not variance). Compute the square of each sample and record the average of these squares $\hat{\mu}_N = \frac{1}{N} \sum_i x_i^2$.

(B) Let's call the previous estimate "one trial". Now perform T trials for a fixed choice of N. Denote the mean produced by trial t with $\hat{\mu}_{N,t}$ for $t \in \{1, \ldots .T\}$. Now, report the mean and standard deviation across trials of $\hat{\mu}_{N,t}$. For example, for the mean, you would compute $\frac{1}{T} \sum_t \hat{\mu}_{N,t}$.

(C) For $T = 100$ and $\sigma^2 = 10$, try this for $N \in \{1, 10, 50, 100, 200, 500, 1000\}$ and plot the means and variances (e.g. with `matplotlib`). Make sure the mean and variances follow the expected trend as N increases (see the previous Monte Carlo question on this homework)

(D) Try to find another function (instead of squaring) that has higher variance across trials for a given value of N

Please See Notebook.

ML Assignment 0, Question 7

In [1]:
```
import torch
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:
```
x=torch.tensor([[1.0,2.0],[3.0,4.0]])
```

In [3]:
```
x
```

Out[3]:
```
tensor([[1., 2.],
        [3., 4.]])
```

In [4]:
```
x.shape
```

Out[4]:
```
torch.Size([2, 2])
```

Part A
We define a function to draw a sample of a given size and a given variance. torch.normal takes the std as the argument. We also define a function to obtain the square average of an array.

In [5]:
```
def sample(variance, size):
    deviation=torch.sqrt(torch.tensor(variance))
    return torch.normal(mean=0, std= deviation* torch.ones(size))
```

In [6]:
```
def square_avg(arr):
    return arr.dot(arr)/(arr.size()[0])
```

In [7]:
```
x=sample(3,10)
```

In [8]:
```
x.size()[0]
```

Out[8]:
```
10
```

In [9]:
```
x
```

Out[9]:
```
tensor([-0.4594, -0.6307,  0.8052, -0.5855, -1.4998,  2.3501, -1.4504, -0.3247,
         1.0625,  4.2317])
```

In [10]:
```
torch.mean(x)
```

Out[10]:
```
tensor(0.3499)
```

In [11]:
```
square_avg(x)
```

Out[11]:
```
tensor(3.0618)
```

Part B
We conduct this experiment across trails and record the square averages. We can then find the mean and the variance of square averages for that trial. We then define a function in part C to find the mean and variance of the recorded square average value across trails. We do this for varying values of N.

In [12]:
```
def trails(T,N,var):
    u=torch.empty(T)

    for i in range(T):
        arr=sample(var,N)
        x=square_avg(arr)
        u[i]=x
    return u
```

In [13]:
```
t=trails(10,20,3)
t
```

Out[13]:
```
tensor([4.6852, 2.5278, 2.9221, 5.0253, 1.8385, 1.6065, 2.4826, 4.0149, 4.5221,
        2.6558])
```

In [14]:
```
torch.mean(t)
```

Out[14]:
```
tensor(3.2281)
```

In [15]:
```
torch.std(t)
```

Out[15]:
```
tensor(1.2326)
```

In [16]:
```
torch.var(t)
```

Out[16]:
```
tensor(1.5194)
```

Part C
We now define a function to find the mean and variance of the recorded square average value across trails. We do this for variying values of N.

In [17]:
```
N_array=[1, 10, 50, 100, 200, 500, 1000]
T=100
variance=10

def find_means_and_variances(variance,T,sizearr):
    umean=torch.empty(len(sizearr))
    uvar=torch.empty(len(sizearr))
    for i in range(len(sizearr)):
        x=trails(T,sizearr[i],variance)
        umean[i] = torch.mean(x)
        uvar[i] = torch.var(x)


    return umean,uvar
```

In [18]:
```
find_means_and_variances(10,100,N_array)
```

Out[18]:
```
(tensor([10.7412, 11.3290, 10.2798,  9.9835, 10.0363,  9.9454,  9.9684]),
 tensor([2.5471e+02, 1.8586e+01, 3.8177e+00, 2.4367e+00, 9.3616e-01, 4.0279e-01,
         1.7402e-01]))
```

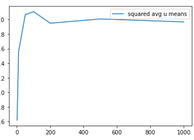
In [19]:
```
(umean, uvar) = find_means_and_variances(10,100,N_array)
```

Part C
We do the plotting

In [20]:
```
sizes=np.array(N_array)
```

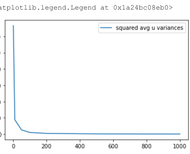
In [21]:
```
plt.plot(sizes,umean.numpy(),label='squared avg u means')
plt.legend()
```

Out[21]:
```
<matplotlib.legend.Legend at 0x1a24bb78c70>
```



In [22]:
```
plt.plot(sizes,uvar.numpy(),label='squared avg u variances')
plt.legend()
```

Out[22]:
```
<matplotlib.legend.Legend at 0x1a24bc08eb0>
```



Explanation

Let $Y = \sum_{i=1}^{N} \frac{x_i^2}{N}$

The random variable Y is representing the square average means and is our variable of interest.

$E[Y] = \sum_{i=1}^{N} \frac{1}{N} E[x_i^2]$ Therefore, $E[Y] = E[X^2]$ We know, $E[X^2]$ is $\mu^2 + \sigma^2$ which is 10. We can see from the monte carlo properties, that the sample mean converges to the true mean.We also know that the sample variance will tend to zero as N keeps on increasing. This is what we observe in our graph.

Part D We take the squared sum function instead of the squared avg. We can see that there is no N in our denominator and this affects our monte carlo calculations.

In [42]:
```
def square_sum(arr):
    return arr.dot(arr)
```

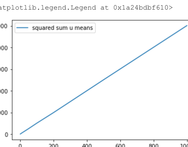
In [43]:
```
def trails(T,N,var):
    u=torch.empty(T)

    for i in range(T):
        arr=sample(var,N)
        x=square_sum(arr)
        u[i]=x
    return u
```

In [68]:
```
N_array=[1, 10, 50, 100, 200, 500, 1000]
T=100
variance=10

def find_means_and_variances(variance,T,sizearr):
    umean=torch.empty(len(sizearr))
    uvar=torch.empty(len(sizearr))
    for i in range(len(sizearr)):
        x=trails(T,sizearr[i],variance)
        umean[i] = torch.mean(x)
        uvar[i] = torch.var(x)


    return umean,uvar
```

In [69]:
```
(umean, uvar) = find_means_and_variances(10,100,N_array)
```

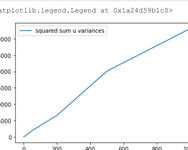
In [70]:
```
sizes=np.array(N_array)
plt.plot(sizes,umean.numpy(),label='squared sum u means')
plt.legend()
```

Out[70]:
```
<matplotlib.legend.Legend at 0x1a24bdbf610>
```



In [71]:
```
plt.plot(sizes,uvar.numpy(),label='squared sum u variances')
plt.legend()
```

Out[71]:
```
<matplotlib.legend.Legend at 0x1a24d59b1c0>
```



We can see that the variances are increasing and greater than the earlier sub-part for our given N.

In [ ]:
```
```