DS-GA 1003 Machine Learning: Homework 0

Due 11.59 p.m. EST, February 13, 2024 on Gradescope

(Jiasheng Ni)

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

1 Probability Distributions

For parts (A) to (C), suppose that 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 $p(X), p(Y|X), p(Z|X)$.

Solution. By the definition of marginal probability:

$$\begin{aligned} P(Y = y) &= \sum_{x \in \mathcal{X}, z \in \mathcal{Z}} P(X = x, Y = y, Z = z) \\ &= \sum_{z \in \mathcal{Z}} \sum_{x \in \mathcal{X}} P(X = x) P(Y = y \mid X = x) P(Z = z \mid X = x) \end{aligned}$$

Thus we know that $P(Y) = \sum_{X,Z} P(X)P(Y \mid X)P(Z \mid X)$ □

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

Solution.

$$\begin{aligned} P(Z \mid Y) &= \frac{P(Y, Z)}{P(Y)} \\ &= \frac{\sum_{x \in \mathcal{X}} P(X = x, Y, Z)}{\sum_{x \in \mathcal{X}, z \in \mathcal{Z}} P(X, Y = x, Z = z)} \\ &= \frac{\sum_{x \in \mathcal{X}} P(X = x) P(Y \mid X = x) P(Z \mid X = x)}{\sum_{x \in \mathcal{X}, z \in \mathcal{Z}} P(X) P(Y \mid X = x) P(Z = z \mid X = x)} \\ &= \frac{\sum_{X} P(X) P(Y \mid X) P(Z \mid X)}{\sum_{X,Z} P(X) P(Y \mid X) P(Z)} \end{aligned}$$

□

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

Solution. From total law of probability:

$$\begin{aligned} P(X, Y, Z) &= P(Y, Z \mid X) P(X) \\ \because P(X, Y, Z) &= P(Y|X|P(Z \mid X) P(X) \\ \therefore P(Y \mid X) P(Z \mid X) &= P(Y, Z \mid X) \end{aligned}$$

$\therefore Y, Z$ are conditionally independent given X. We cannot make any conclusions about independence. □

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

Solution.

$$\begin{aligned}
E[E[Y \mid X]] &= E\left[\int_{y\in\mathcal{Y}} y \cdot f_{Y|X}(y \mid X)dy\right] \\
&= \int_{X\in\mathcal{X}} \cdot \int_{Y\in\mathcal{Y}} y f_{Y|X}(y \mid x)dy f_X(x)dx \\
&= \int_{X\in\mathcal{X}} \int_{y\in\mathcal{Y}} y f_{X,Y}(x,y)dydx \\
&= \int_{y\in\mathcal{Y}} y \cdot \left(\int_{X\in\mathcal{X}} f_{X,Y}(x,y)dx\right) dy \\
&= \int_{y\in\mathcal{Y}} y \cdot f_Y(y)dy \\
&= E[Y]
\end{aligned}$$

□

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

Solution. Suppose we have a continuous R.V that have the following Pdf:

$$f_X(x) = \frac{1}{\pi\,(1+x^2)}$$

$$\begin{aligned}
P(X < \infty) &= \int_{-\infty}^{\infty} \frac{1}{\pi\,(1+x^2)}dx \\
&= \frac{1}{\pi}\arctan x\Big|_{-\infty}^{\infty} \\
&= \frac{1}{\pi}\cdot\pi \\
&= 1
\end{aligned}$$

$$\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} \frac{x}{\pi\,(1+x^2)}dx \\
&= \frac{1}{2\pi}\log\left(x^2+1\right)\Big|_{-\infty}^{\infty} \\
&= \infty
\end{aligned}$$

□

(F) 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.

Solution. It's impossible. If $f(X,Y) \perp X$ and $f(X,Y) \perp Y$, then. knowing the information from X and Y wont give us any information about $f(X,Y)$ However, $f(X,Y)$ is a non-constant function of X and Y, if we know the value of X and Y, we definitely know the value of $f(X,Y)$ $\therefore$ $f(X,Y)$ is dependent on X and Y, which is a contradiction. □

2 Gradients

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

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

Compute $\nabla_x f$.

Solution.

$$\begin{aligned}\nabla_x f &= \begin{bmatrix} \frac{\partial f(\vec{x})}{\partial x_1} & \frac{\partial f(\vec{x})}{\partial x_2} \end{bmatrix}^\top \\ &= \begin{bmatrix} f(\vec{x}) \cdot \left(\frac{2}{x_1} + x_2\right) & x_1 \end{bmatrix}^\top \end{aligned} \tag{1}$$

□

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

Solution.

$$f(y; \lambda) = \lambda e^{-\lambda y}$$
$$f(2; \lambda) = \lambda e^{-2\lambda}$$

Fist, since taking a ln at $f(\lambda) = \lambda e^{-2\lambda}$ wont change our optimizing direction i- We are optimizing;

$$\max_\lambda \ln \lambda - 2\lambda$$

It's equivalent to optimize $g(\lambda) = -\ln \lambda + 2\lambda$:

$$\min_\lambda -\ln \lambda + 2\lambda$$

We take second-denivative of $g(\lambda)$ and get:

$$H_y = \frac{1}{\lambda^2} \geqslant 0 \quad \forall \lambda$$

$\therefore g(\lambda)$ is convex in λ $\therefore$ We can set derivative to zero and get:

$$\begin{aligned} -\tfrac{1}{\lambda} + 2 &= 0 \\ \lambda^* &= \tfrac{1}{2} \\ \therefore \min_\lambda g(\lambda) &= \ln 2 + 1 \\ \therefore \max_\lambda f(\lambda) &= -(\ln 2 + 1) \end{aligned}$$

□

(C) The CDF of the Exponential distribution is

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

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

Solution. We tale the derivative of $F(y; \lambda)$ and get:

$$f(y; \lambda) = \lambda e^{-\lambda y}$$

□

3 The Gaussian Distribution

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 $f(X) \sim \mathcal{N}(3, 2)$.

Solution. We first prove that linear transformation on X wont change its family of distribution: Claim: $ax + b$ where $a \neq 0, b$ are constant is still normal distribution.

Proof: $x \sim N(0, 1)$

$$\begin{aligned} \therefore f_X(x) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} \\ F_Y(y) &= P(ax + b \leq y) \\ &= \int_{-\infty}^{\frac{y-b}{a}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} dx \end{aligned}$$

By fundamental theorem of calculus we have:

$$\begin{aligned} f_Y(y) &= \frac{\partial F_Y(y)}{\partial y} \\ &= \frac{1}{\sqrt{2\pi}a} \exp\left\{-\frac{1}{2}\left(\frac{y-b}{a}\right)^2\right\} \\ \therefore Y &\sim N\left(b, a^2\right) \end{aligned}$$

$\therefore$ We let $f(X) = \sqrt{2}X + 3$ and thus $f(X) \sim N(3, 2)$ □

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

Solution.

$$\begin{aligned} \because M_{X+Y}(s) &= E\left[e^{s(x+Y)}\right] \\ &= e^{3s+s^2} \\ &= e^{\frac{1}{2}s^2} \cdot e^{3s+\frac{1}{2}s^2} \\ &= M_X(s)M_Y(s) \end{aligned}$$

$\therefore$ We could let $Y \sim N(3, 1)$ and $X \perp Y$. □

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

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

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 $p(X_1|X_2 = x_2)$.

Solution. Claim: Proof: We rewrite the joint pdf as follows:

$$\begin{aligned}
f_{X_1,X_2}(x_1,x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{s(x_1)^2 - 2\rho s(x_1)s(x_2) + s(x_2)^2}{2(1-\rho^2)}\right\} \\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{(s(x_1) - \rho s(x_2))^2 + s(x_2)^2 - \rho^2 s(x_2)^2}{2(1-\rho^2)}\right\} \\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}s(x_2)^2\right\} \cdot \exp\left\{-\frac{(s(x_1) - \rho s(x_2))^2}{2(1-\rho^2)}\right\} \\
f_{X_2}(x_2) &= \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{1}{2}s(x_2)^2\right\} \\
\therefore f_{X_1|X_2}(x_1 \mid x_2) &= \frac{f_{x_1,x_2}(x_1,x_2)}{f_{x_2}(x_2)} \\
&= \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}} \exp\left\{-\frac{(s(x_1) - \rho s(x_2))^2}{2(1-\rho^2)}\right\} \\
&= \frac{1}{\sqrt{2\pi{\sigma_1}^2(1-\rho^2)}} \exp\left\{-\frac{(x_1 - \mu_1 - \rho\sigma_1 s(x_2))^2}{2{\sigma_1}^2(1-\rho^2)}\right\}
\end{aligned}$$

$\therefore$ We know that $X_1 \,\middle|\, X_2 = x_2 \sim N\left(\mu_1 + \rho\sigma_1\frac{x_2-\mu_2}{\sigma_2}, \sigma_1^2(1-\rho^2)\right)$ □

(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? Why or why not?

Solution. Suppose we have two variable that satisfies the following condition:

(a) X, Y can only take values from $\{-1, 0, 1\}$
(b) If $X = 0, Y$ take 1 or -1 with probability $\frac{1}{2}$
(c) If $Y = 0, X$ take 1 or -1 with probability $\frac{1}{2}$
(d) Either $X = 0$ with probability $\frac{1}{2}$ or $Y = 0$ w.p $\frac{1}{2}$

We then claim that $\text{Cov}(X, Y) = 0$, but X is dependent on Y.

Proof: Notice that since there can only be one zero, so all the possible x, y) pair could be

$$\begin{aligned}
\Omega &= \{(0,1),(0,-1),(1,0),(-1,0)\} \\
P_{X,Y}(0,1) &= P_{Y|X}(1 \mid 0)P_X(0) \\
&= \frac{1}{2} \times \frac{1}{2} \\
&= \frac{1}{4} \\
P_{X,Y}(1,0) &= P_{X|Y}(1 \mid 0)P_Y(0) \\
&= \frac{1}{2} \times \frac{1}{2} \\
&= \frac{1}{4}
\end{aligned}$$

$P_{X,Y}(0,-1) = P_{X,Y}(-1,0) = \frac{1}{4}$ with the same reasoning.

$$\begin{aligned}
\therefore E_{X,Y}[XY] &= \sum_{x,y \in \Omega} 0 \cdot P_{X,Y}(x,y) \\
&= 0
\end{aligned}$$

$$\begin{aligned}
E_X[x] &= P_X(0) \cdot 0 + P_X(1) \cdot 1 + P_X(-1) \cdot (-1) \\
&= \frac{1}{2} \times 0 + \sum_y P_{x,Y}(1,y) \cdot 1 + \sum_y P_{x,y}(-1,y) \cdot (-1) \\
&= 0 + \frac{1}{4} - \frac{1}{4} \\
&= 0
\end{aligned}$$

$E_Y[Y] = 0$ with the same reasoning

$$\begin{aligned}
\therefore \mathrm{Cov}(X,Y) &= E[XY] - E[X]E[Y] \\
&= 0
\end{aligned}$$

However, since by construction:

$$P_{X|Y}(1 \mid 0) = \frac{1}{2} \quad P_X(1) = \frac{1}{4}$$

$\therefore X$ and Y aren't independent. For X, Y that's jointly gaussian, if $\mathrm{cov}(X,Y) = 0$. then $\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} u_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right)$

$$\begin{aligned}
\therefore f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{s(x)^2 + s(y)^2}{2}\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{1}{2}s(x)^2\right\} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{1}{2}s(y)^2\right\} \\
&= f_X(x) f_Y(y) \\
\therefore X &\perp Y
\end{aligned}$$

□

4 Monte Carlo Estimators

Let $X \sim D$ be a random variable and denote $\mu = \mathbb{E}_{X\sim D}[X]$ and $\sigma^2 = \mathbf{Var}_{X\sim D}[X]$ as its mean and variance respectively. Assume that X has finite variance, i.e. $\sigma^2 < \infty$. While you do not know μ or σ^2, you can collect N independent samples of X, which we denote as $\{X_i\}_{i=1}^N$.

(A) Is the mean μ finite? If yes, why? If not, construct an example of such a random variable X.

Solution. Yes. For the sake of contradiction, suppose M is not finite. Then $E[x] = \int_{x\in X} xf(x)dx = \infty$

$$\begin{aligned} \because \quad \text{Var}[X] &= \int_{x\in X} (X - E(X))^2 f(x)dx \\ &= \int_{x\in X} x^2 f(x)dx + E[x]^2 f(x)dx - 2xE(x)f(x)dx \end{aligned}$$

If $E(X) = \infty$, the $\text{Var}[X]$ cannot be finite. $\therefore$ Contradiction!

□

(B) From your N samples, you can construct a **Monte Carlo estimator** of μ as:

$$\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Find the mean and variance of $\hat{\mu}_N$.

Solution.

$$\begin{aligned} E\left[\hat{\mu}_N\right] &= \frac{1}{N}\sum_{i=1}^{N} E\left[X_i\right] \\ &= \frac{1}{N}\cdot N\mu \\ &= \mu \\ \text{Var}\left[\hat{\mu}_N\right] &= \frac{1}{N^2}\left[\sum_{i=1}^{N}\text{Var}\left(X_i\right) + 2\sum_{i<j}\text{cov}\left(X_i, X_j\right)\right] \\ &= \frac{1}{N^2}\left[N\cdot\sigma^2 + 0\right] \\ &= \frac{\sigma^2}{N} \end{aligned}$$

□

(C) Based on your answer in (B), name a potential advantage and disadvantage of using $\hat{\mu}_N$ to estimate μ.

Solution. Since $E\left[\hat{\mu}_N\right] = \mu, \therefore\ \hat{\mu}_N$ is unbiased estimator of μ When N is small $\text{Var}\left[\hat{\mu}_N\right]$ could be large and thus $\widehat{d_N}$ may be unstable. □

(D) Assuming that $\sigma^2 < \infty$ and $\mu < \infty$, then prove for any $k > 0$ the following inequality:

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

Solution. Since $(X-\mu)^2 \geq 0$, thus by markov inequality we have:

$$P\left[(X-\mu)^2 > k^2\right] \leq \frac{E\left[(X-\mu)^2\right]}{k^2}$$
$$P[|X-\mu| > k] \leqslant \frac{E\left[(X-\mu)^2\right]}{k^2} = \frac{\sigma^2}{k^2}$$

□

(E) Using parts (B) and (D), prove for any $k > 0$ that:

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

Solution. Since $E[\hat{\mu_N}] = \mu$ and $Var(\hat{\mu_N}) < \infty$, by chebyshev's inequality we have:

$$P\left(|\hat{\mu}_N - \mu| > k\right) \leqslant \frac{\mathrm{Var}\left(\hat{\mu}_N\right)}{k^2}$$
$$= \frac{\sigma^2}{Nk^2}$$
$$\therefore 0 \leq \lim_{N\to\infty} P\left(|\hat{\mu}_N - \mu| > k\right) \leq \lim_{N\to\infty} \frac{\sigma^2}{Nk^2} = 0$$

By squeeze theorem we know:

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

□

(F) In your own words, why is the result in (E) useful?

Solution. The result in (E) is what's called weak law of large number. It says $\hat{\mu}_N \xrightarrow{p} \mu$, ie. $\hat{\mu}_N$ converges in probability to μ. If we have large number of samples, $\hat{\mu}_N$ would be both an unbiased and consistent estimator of μ By saying unbiased, we say that the estimator wont over/under estimate the true parameter. By saying consistent, we say that the estimator improves with more information. Combined, the estimator ensures that even with limited sample size, the estimate is as accurate as possible. □

5 Kullback-Liebler Divergence

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, which states that given a strictly-convex function f and a (non-constant) random variable X:

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

Solution. (1):

$$\begin{aligned} kL(P\|P) &= \int_{x \in R} p(x) \log \frac{p(x)}{p(x)} dx \\ &= \int_{x \in R} p(x) \cdot 0 dx \\ &= 0 \end{aligned}$$

(2): We could chase $f(x) = -\log x$, which is strizfy convex over R. $\therefore$ By Jensen's inequality we have:

$$\begin{aligned} KL(P\|Q) = \int_{x \in R} p(x) \log \frac{p(x)}{q(x)} dx &= \int_{x \in R} -p(x) \log \frac{q(x)}{p(x)} dx \\ &= E_{x \sim p}\left(-\log \frac{q(x)}{p(x)}\right] \\ &> -\log E_{x \sim p}\left[\frac{q(x)}{p(x)}\right] \quad \text{(Jensen's inequality)} \\ &= -\log \int_{x \in \mathcal{X}} p(x) \cdot \frac{q(x)}{p(x)} dx \\ &= -\log \int_{x \in \mathcal{X}} q(x) dx \\ &= -\log \cdot 1 \\ &= 0 \end{aligned}$$

□

6 Setting Up 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` is 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:

1. 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 $\hat{\mu}_N = \frac{1}{N} \sum_i x_i^2$.

2. Let's call the estimate $\hat{\mu}_N$ we obtain in Step 1 as a single "trial". Now perform T trials for a fixed choice of N. Denote the mean produced by trial t as $\hat{\mu}_{N,t}$ for $t \in \{1, \ldots .T\}$. Now, compute the mean and standard deviation across trials of $\hat{\mu}_{N,t}$. For example, for the mean, you would compute $\frac{1}{T} \sum_t \hat{\mu}_{N,t}$.

Now that you can code these two steps:

(A) Set $T = 100$ and $\sigma^2 = 10$. Perform Steps 1 and 2 for each value of $N \in \{1, 10, 50, 100, 200, 500, 1000\}$. Plot the means and variances on a single graph each, i.e. you should have two graphs, one for the means and one for the variances, where the x-axis is $\log N$.

(B) What do you observe about the mean and variances as N increases? How do these trends relate to your answers in Question 4?

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

Problem 6.1

In [ ]:
```
# Problem 6.1
mean = 0.0
std = 1.0
N = 50
sample_generator = torch.distributions.Normal(mean, std)
samples = sample_generator.sample((N,))

# P6.1 Compute the square of each sample and record the average
samples.square().mean()
```

Out[ ]:
```
tensor(0.7462)
```

Problem 6.2

In [ ]:
```
# P6.2 Compute the estimator T times
T = 50
estimators = [(sample_generator.sample((N,))).square().mean().item() for i i
mean = np.mean(estimators)
std = np.std(estimators)
std
```

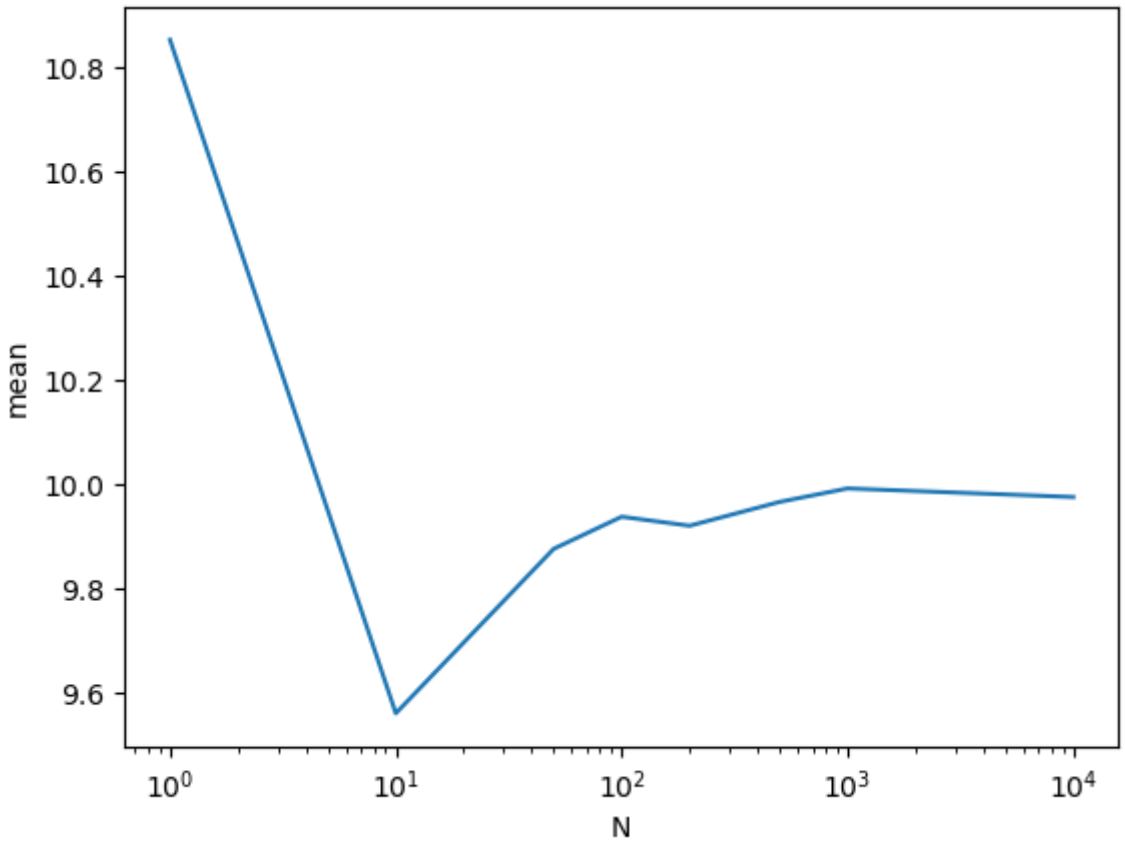
Out[ ]:
```
0.18128779864147673
```

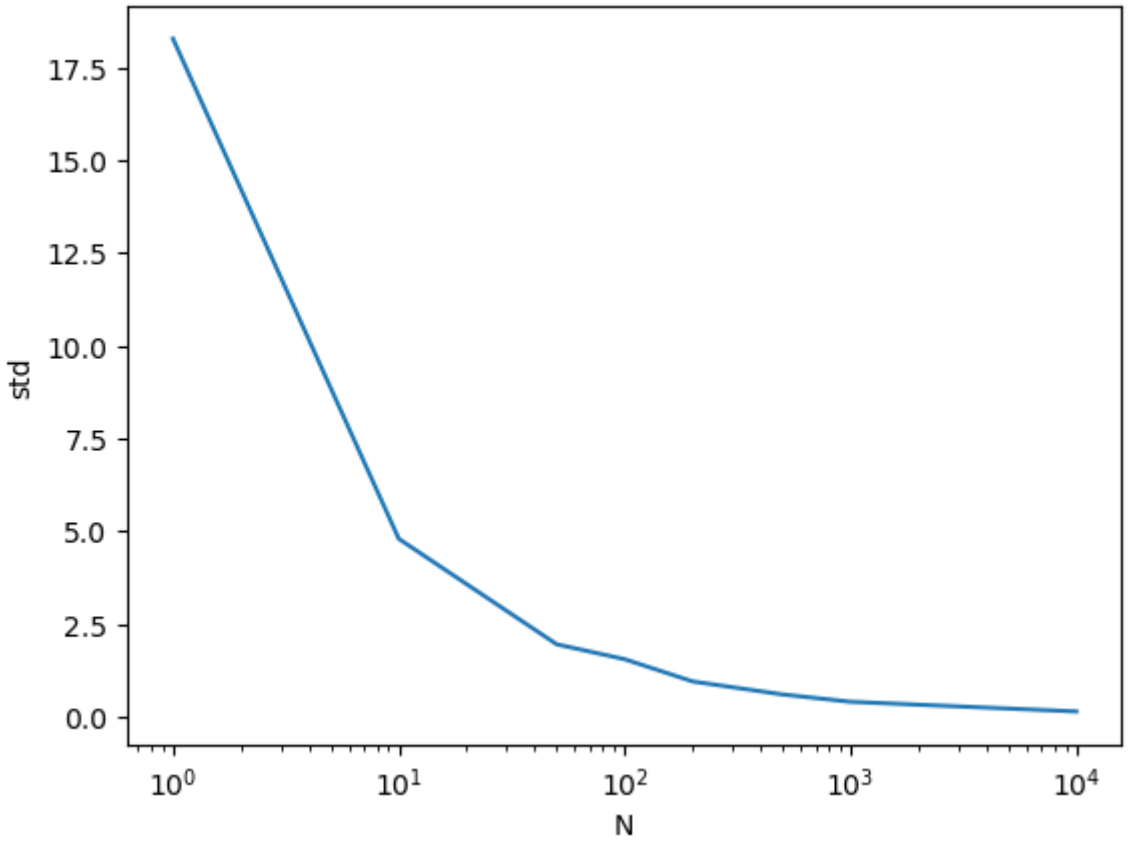
Problem 6.3(A)

In [ ]:
```
# P6.3
def generate_T_sample(mean, std, N):
    sample_generator = torch.distributions.Normal(mean, std)
    samples = sample_generator.sample((N,))
    return samples

def compute_estimator_desc(mean, std, N, T):
    T_samples = [generate_T_sample(mean , std, N).square().mean().item() for
    estimator_mean = np.mean(T_samples)
    estimator_std = np.std(T_samples)
    return [estimator_mean, estimator_std]

def collect_estimator(mean, std, N_list, T):
    return [compute_estimator_desc(mean, std, N, T) for N in N_list]

def plot_estimator(estimators_desc, N_list):
```

```
    print(estimators_desc)
    mean_list = list(map(lambda x: x[0], estimators_desc))
    std_list = list(map(lambda x: x[1], estimators_desc))
    plt.plot()
    plt.semilogx(N_list, mean_list)
    plt.xlabel("N")
    plt.ylabel("mean")
    plt.show()
    plt.semilogx(N_list, std_list)
    plt.xlabel("N")
    plt.ylabel("std")
    plt.show()

N_list = [1,10,50,100,200,500,1000, 10000]
mean = 0
std = np.sqrt(10)
T = 100
estimators_desc = collect_estimator(mean, std, N_list, T)
plot_estimator(estimators_desc, N_list)
```

```
[[10.854389454252013, 18.267171864183172], [9.561431233882905, 4.799733233449349], [9.877097930908203, 1.9590385411132993], [9.938786916732788, 1.5573714609819649], [9.921476850509643, 0.9582612686629485], [9.96716851234436, 0.6086864385648383], [9.993071165084839, 0.41086452832110637], [9.976598720550538, 0.15048812253635147]]
```





Problem 6.3(B)

Observation:

- $\hat{\mu_N} = \frac{1}{N}\sum_i x_i^2$ is an unbiased estimator for σ^2 since it is always close to the true variance, no matter what value N takes.
- It is also consistent since we observe that the variance of this estimator goes to zero as $N \to \infty$, and the plot verifies this.

Claim, $\hat{\mu_N}$ is an unbiased and consistent estimator of σ^2.

Proof:

- MGF of X_i: $M_X(t) = e^{\frac{\sigma^2 t^2}{2}}$.
- Second Moment: $\nabla M_X(t)\big|_{t=0} = \sigma^2$
- Fourth Moment: $\nabla^{(4)} M_X(t)\big|_{t=0} = 5\sigma^4$
- Unbiasedness: $E[\hat{\mu_N}] = \frac{1}{N}\sum_{i=1}^{N} E[X_i^2] = \frac{\sigma^2}{N} \cdot N = \sigma^2$
- Consistency: $\lim_{N\to\infty} P(|\widehat{\mu_N} - \sigma^2| > \epsilon) \leq \frac{Var(\widehat{\mu_N})}{k^2} = \frac{5\sigma^4}{Nk^2}$(using theorem from problem 4E), by squeeze theorem we have consistency.

In [ ]: