CSCI-GA 2565 Machine Learning Homework 1

Chuanyang Jin

September 22, 2022

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.
$p(Y) = \int_x \int_z p(X, Y, Z)dzdx = \int_x \int_z p(X)p(Y|X)p(Z|X)dzdx$ □

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

Solution.
$p(Z|Y) = \frac{p(Y,Z)}{p(Y)} = \frac{\int_x p(X,Y,Z)dx}{\int_x \int_z p(X,Y,Z)dzdx} = \frac{\int_x p(X)p(Y|X)p(Z|X)dx}{\int_x \int_z p(X)p(Y|X)p(Z|X)dzdx}$ □

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

Solution.
Without further assumptions, we can not conclude that any variable is independent.
Y and Z are conditionally independent given X, since

$$\begin{aligned} p(Y|X, Z) &= \frac{p(X, Y, Z)}{p(X, Z)} \\ &= \frac{p(X)p(Y|X)p(Z|X)}{p(X)p(Z|X)} \\ &= p(Y|X) \end{aligned}$$

□

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

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.

Solution.

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{E}[Y|X]\Big] &= \mathbb{E}\Big[\int_{-\infty}^{\infty} yP(Y=y|X=x)dy\Big] \\
&= \int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty} yP(Y=y|X=x)dy\Big)P(X=x)dx \\
&= \int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty} y\frac{P(Y=y,X=x)}{P(X=x)}dy\Big)P(X=x)dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\frac{P(Y=y,X=x)}{P(X=x)}P(X=x)dydx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yP(Y=y,X=x)dydx \\
&= \int_{-\infty}^{\infty} y\Big(\int_{-\infty}^{\infty} P(Y=y,X=x)dx\Big)dy \\
&= \int_{-\infty}^{\infty} yP(Y=y)dy \\
&= \mathbb{E}[Y]
\end{aligned}$$

□

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

Solution.

$P(X=x)=\frac{1}{x}$ for $x=2^k, k\in\mathbb{Z}^+$ and $P(X=x)=0$ otherwise.
$P(x<\infty)=\Sigma_{k=1}^{\infty}\frac{1}{2^k}=1$
$E(X)=\Sigma_{k=1}^{\infty}x\cdot P(X=x)=\Sigma_{k=1}^{\infty}x\cdot\frac{1}{x}=\Sigma_{k=1}^{\infty}1=\infty$ □

(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.

Construct X, Y to be random variables with probability density functions

$$f_X(x)=\begin{cases}1 & 0<x<1\\ 0 & otherwise\end{cases},$$

$$f_Y(y)=\begin{cases}1 & 0<y<1\\ 0 & otherwise\end{cases}.$$

And define $f(X,y)$ as

$$f(X,Y)=\begin{cases}1 & (0<x<\frac{1}{2} \text{ and } \frac{1}{2}<y<1) \text{ or } (\frac{1}{2}<x<1 \text{ and } 0<y<\frac{1}{2})\\ 0 & otherwise\end{cases}.$$

Then $f(X,Y)$ is independent of X and $f(X,Y)$ is independent of Y. □

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.

$$\begin{aligned}\frac{\partial f(x)}{\partial x_1} &= \exp[\log(x_1^2) + x_1 x_2] \cdot \frac{\partial(log(x_1^2 + x_1 x_2))}{\partial x_1} \\ &= \exp[\log(x_1^2) + x_1 x_2] \cdot (\frac{2}{x_1} + x_2)\end{aligned}$$

$$\begin{aligned}\frac{\partial f(x)}{\partial x_2} &= \exp[\log(x_1^2) + x_1 x_2] \cdot \frac{\partial(log(x_1^2 + x_1 x_2))}{\partial x_2} \\ &= \exp[\log(x_1^2) + x_1 x_2] \cdot x_1\end{aligned}$$

Therefore,

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

□

(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.

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

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

To maximize $f(2; \lambda)$, it's equivalent to maximize $\ln f(2; \lambda)$. Note that $\ln f(2; \lambda)$ is a convex function, and its derivative with respect to λ is $\frac{1}{\lambda} - 2$. We set $\frac{1}{\lambda} - 2$ equal to 0, and get $\lambda = \frac{1}{2}$, and $f(2; \lambda) = \frac{1}{2e}$. Therefore, when $\lambda = \frac{1}{2}$, $f(2; \lambda)$ has its maximum value $\frac{1}{2e}$. □

(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.

For $y \geq 0$,

$$f(y : \lambda) = \frac{\partial F(y : \lambda)}{\partial y} = \lambda \exp[-\lambda y]$$

Therefore, for $y \geq 0$, $f(y : \lambda) = \lambda \exp[-\lambda y]$. □

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.

$f(X) = aX + b \sim \mathcal{N}(a\mu_X + b, a^2\sigma_X^2) = \mathcal{N}(b, a^2) = \mathcal{N}(3, 2)$.
Therefore, $f(X) = \sqrt{2}X + 3$. □

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

Solution.

$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) = \mathcal{N}(\mu_Y, 1 + \sigma_Y^2) = \mathcal{N}(3, 2)$.
Therefore, $Y \sim \mathcal{N}(3, 1)$. □

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

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

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 $p(X_1 = x_1 | X_2 = x_2)$.

Solution.

$X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

By Bayes rule,

$$
\begin{aligned}
p(X_1 = x_1 | X_2 = x_2) &= \frac{p(X = x_1, X_2 = x_2)}{p(X_2 = x_2)} \\
&= \frac{\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right] \right]}{\frac{1}{\sigma_2\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right]} \\
&= \frac{1}{\sigma_1\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right) - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \rho^2\left(\frac{x_2-\mu_2}{\sigma_2}\right) \right]^2 \right] \\
&= \frac{1}{\sigma_1\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right) - \rho\left(\frac{x_2-\mu_2}{\sigma_2}\right) \right]^2 \right]
\end{aligned}
$$

□

(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.

Let X be a random variable that is 1 or +1 with probability 0.5. Let Y be a random variable such that $Y = 0$ if $X = 1$, and Y is randomly 1 or +1 with probability 0.5 if $X = 1$.
$Cov(X, Y) = E[XY] - E[X]E[Y] = 0 - 0 = 0$ but clearly X is not independent of Y.
This is not possible if X, Y are jointly Gaussian. In part (C), if X, Y are jointly Gaussian, then $c = 0$, and $\rho = \frac{c}{\sigma_1\sigma_2} = 0$.

$$
p(X = x | Y = y) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left[\left(\frac{x-\mu_x}{\sigma_1}\right) \right]^2 \right]
$$

So X must be independent of Y. □

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, the mean μ is finite.
The variance $var(X) = E[X^2] - E[X]^2$ is finite, so $E[X^2] - E[X]^2 < \infty$.
By Jensen's inequality, $E[X]^2 \leq E[X^2]$, so $E[X]^2$ is finite, so the mean $\mu = E[X]$ is finite..

□

(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.

Since X include N independent samples,

$$\begin{aligned} E(\hat{\mu}_N) &= E(\frac{1}{N}\sum_{i=1}^{N} X_i) \\ &= \frac{1}{N}E(\sum_{i=1}^{N} X_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} E(X_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} \mu \\ &= \mu \end{aligned}$$

$$\begin{aligned} Var(\hat{\mu}_N) &= Var(\frac{1}{N}\sum_{i=1}^{N} X_i) \\ &= \frac{1}{N^2}Var(\sum_{i=1}^{N} X_i) \\ &= \frac{1}{N^2}\sum_{i=1}^{N} Var(X_i) \\ &= \frac{1}{N^2}\sum_{i=1}^{N} \sigma^2 \\ &= \frac{\sigma^2}{N} \end{aligned}$$

□

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

Solution.

Advantage: the mean of $\hat{\mu}_N$ is exactly μ.
Disadvantage: the estimation is not accurate when N is small. □

(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.

For any nonnegative random variable Y and $a > 0$,

$$\begin{aligned} aP(Y \geq a) &= a \int_a^\infty P(Y = y)dy \\ &\leq \int_a^\infty yP(Y = y)dy \\ &\leq \int_{-\infty}^\infty yP(Y = y)dy \\ &= E(Y) \end{aligned}$$

So we have the Markov's inequality $P(Y \geq a) \leq \frac{E(Y)}{a}$.
Let $Y = (X - \mu)^2, a = k^2$, we get

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

and equivalently,

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

. □

(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.

Using parts (B), $\hat{\mu}_N$ has mean μ and variance $\frac{\sigma^2}{N}$.
Using the inequality from parts (D) and let X be $\hat{\mu}_N$, we get

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

Then

$$\lim_{N\to\infty} P(|\hat{\mu}_N - \mu| > k) \leq \lim_{N\to\infty} \frac{\sigma^2}{Nk^2} = 0.$$

Since the probability must be nonnegative, we get

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

□

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

Solution.

From the result we can know that: if we have enough samples, the Monte Carlo estimator $\hat{\mu}_N$ will be arbitrarily close to μ. □

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.

When $P = Q$,

$$\begin{aligned} KL(P||Q) &= \int_{x \in \mathbb{R}} p(x) \log \frac{p(x)}{q(x)} dx \\ &= \int_{x \in \mathbb{R}} p(x) \log 1 dx \\ &= \int_{x \in \mathbb{R}} p(x) \cdot 0 dx \\ &= 0. \end{aligned}$$

When $P \neq Q$,
let $f(x) = -\log x$, and we have $f''(x) = \frac{1}{x^2} > 0$ for $x > 0$, so $f(x)$ is a strictly convex function.
By Jensen's inequality, $E(-\log \frac{q(x)}{p(x)}) > -\log E(\frac{q(x)}{p(x)})$, so

$$\begin{aligned} KL(P||Q) &= \int_{x \in \mathbb{R}} p(x) \log \frac{p(x)}{q(x)} dx \\ &= \int_{x \in \mathbb{R}} -p(x) \log \frac{q(x)}{p(x)} dx \\ &= E(-\log \frac{q(x)}{p(x)}) \\ &> -\log E(\frac{q(x)}{p(x)}) \\ &= -\log \int_{x \in \mathbb{R}} p(x) \frac{q(x)}{p(x)} dx \\ &= -\log \int_{x \in \mathbb{R}} q(x) dx \\ &= \log 1 \\ &= 0. \end{aligned}$$

□

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`.
- `tensors` 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

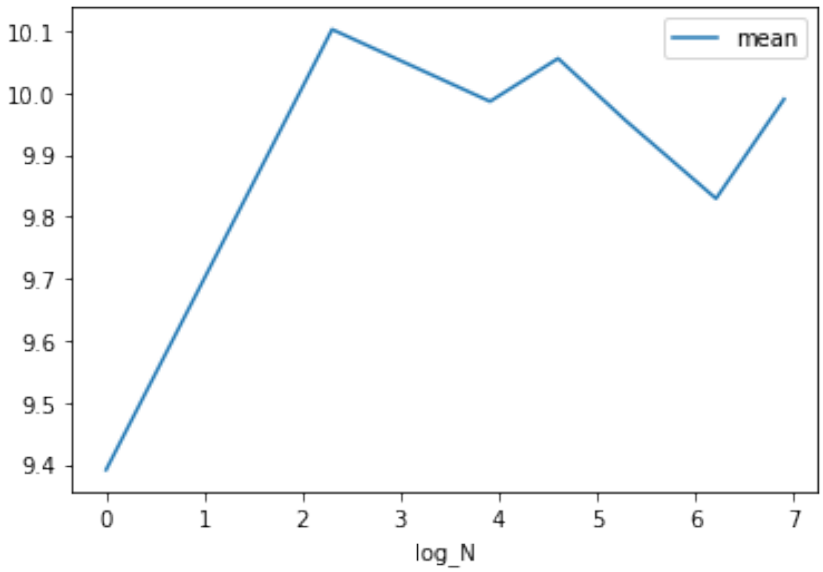
   `torch.distributions.Normal`

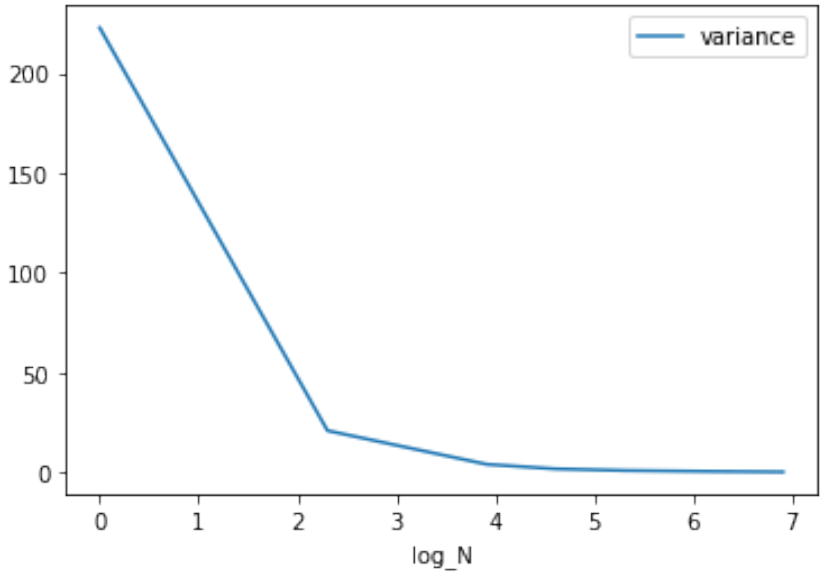
   Be sure to give the right arguments (e.g. standard deviation and not variance). Compute the square of each sample and record the average of these squares $\hat{\mu}_N = \frac{1}{N} \sum_i x_i^2$.

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?

Solution.

As N increases, the mean is more and more stable and the variance keeps decreasing.
According to my answers in Question 4, $\hat{\mu}_N$ has mean μ, so $\hat{\mu}_N$ will be initially around μ and becomes more and more stable.
$\hat{\mu}_N$ has variance $\frac{\sigma^2}{N}$, which decreases as N increases, and eventually approaches negative infinity. □