Problem 1:

1. [4p]

Fill in the remaining part of the function `problem1_rejection` in order to produce samples from the below density using rejection sampling:

$$f[x] = Cx^{0.2}(1-x)^{1.3}$$

for $0 \leq x \leq 1$, where C is a value such that f above is a density (i.e. integrates to one).

Hint: you do not need to know the value of C to perform rejection sampling.

We want to sample from the density

$$f(x) = Cx^{0.2}(1-x)^{1.3}, \quad 0 \leq x \leq 1,$$

where C is an unknown constant. Rejection sampling lets us ignore C.

Step 1: Identify unnormalized target density

We only know the density up to a constant factor:

- True density:

$$f(x) = C \cdot x^{0.2}(1-x)^{1.3}, \quad 0 \leq x \leq 1$$

- Unnormalised version (ignoring C):

$$\tilde{f} = x^{0.2}(1-x)^{1.3}$$

Rejection sampling uses $\tilde{f}(x)$: we never need to know C.

Step 2: Choose a proposal distribution $g(x)$:

The density is supported on $[0,1]$, so the obvious simple proposal is:

$$X \sim \text{Uniform}(0,1)$$

The density of a uniform(0,1) random variable is $g(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$

For a uniform distribution on an interval of length 1, the probability density is constant everywhere on that interval and zero outside it.

A probability density must integrate to 1, so we require: $\int_0^1 g(x)\,dx = 1$

If $g(x)$ is constant, call it k:

$$\int_0^{\pi} k\,dx = k \cdot 1 = 1$$

$$\Rightarrow k \cdot 1 = 1 \Rightarrow k = \frac{1}{1} \Rightarrow g(x) = 1$$

$$\Rightarrow g(x) = 1 \quad , (0 \le x \le 1)$$

Step 3: Find a suitable constant M such that: $\tilde{f}(x) \le M \cdot g(x)$

we need a constant M such that:

$$\hat{f}(x) \le M \cdot g(x) \quad \text{for all } x.$$

Since $g(x) = 1$ on $[0,1]$, this requirement becomes

$$x^{0.2}(1-x)^{1.3} \le M \quad \text{for all } x \in [0,1].$$

Now, note:

- For $x \in [0,1]$, we always have $0 \le x^{0.2} \le 1$
- Also, $0 \le (1-x)^{1.3} \le 1$ on $[0,1]$

So their product satisfies:

$$0 \le x^{0.2}(1-x)^{1.3} \le 1$$

This means a perfectly valid choise is

$$M = 1$$

we don't care about the exact maximum, we only need an upper bound, and 1 is simple and safe upper bound.

Step 4: Compute acceptance probability:

The rejection sampling acceptance probability is

$$\alpha(x) = \frac{\hat{f}(x)}{M \cdot g(x)} = \frac{x^{0.2}(1-x)^{1.3}}{1 \cdot 1} = x^{0.2}(1-x)^{1.3}$$

So: The accept/reject rule becomes:

- Draw $X \sim \text{Uniform}(0,1)$
- Draw $U \sim \text{Uniform}(0,1)$ independently
- Accept X if:

$$U \le x^{0.2}(1-x)^{1.3}$$

(If rejected, we draw a new X and a new U

and try again).

Problem 1:

② Just a simple plotting.

Problem 1:

③ 3. [2p]

Define X as a random variable with the density given in part 1. Denote $Y = \sin(10X)$ and use the above 100000 samples to estimate

$$E[Y]$$

and store the result in `problem1_expectation`.

we have:

- A random variable X with the density from part ①:

$$f(x) = C \cdot x^{0.2}(1-x)^{1.3}, \quad 0 \le x \le 1$$

- We also have $n = 100\,000$ samples from the distribution stored in problem1_samples.

- We define a new random variable:

$$Y = \sin(10X)$$

- The goal is to estimate the expectation $E[Y]$ and store it in problem1_expectation.

Step 1: What is $E[Y]$ in theory?

By definition:

$$E[Y] = E[\sin(10X)]$$

Since X has density f, that means

$$E[Y] = \int_0^1 \sin(10x) f(x) dx = \int_0^1 \sin(10x) C x^{0.2} (1-x)^{1.3} dx$$

This integral is ugly to compute by hand, and we don't even know C explicitly. That's why we use Monte Carlo.

Step 2: Monte Carlo intuition:

Monte Carlo estimation says:

- If you can generate i.i.d samples $X_1, \dots, X_n$ from a distribution of X,
- and you want $E[g(X)]$ for some function g,

then a natural estimator is the sample mean:

$$\hat{E}[g(X)] = \frac{1}{n} \cdot \sum_{i=1}^{n} g(X_i)$$

Here;

- $g(x) = \sin(10x)$
- the samples are $X_1, \dots, X_n$ stored in problem1_samples

So: The estimator for $E[Y]$ is:

$$\hat{E}[Y] = \frac{1}{n} \cdot \sum_{i=1}^{n} \sin(10 X_i),$$

where X_i are the values in problem1_samples.

Problem 4:

4. [2p]

Use Hoeffdings inequality to produce a 95% confidence interval of the expectation above and store the result as a tuple in the variable `problem1_interval`.

we already estimated $\hat{\mu} = \hat{E}[Y] = \frac{1}{n} \cdot \sum_{i=1}^{n} \sin(10 x_i)$

Now we want a 95% confidence interval for the true expectation:

$$\mu = E[\sin(10X)]$$

using Hoeffding's inequality.

Step 1: What hoeffding inequality says:

If a random variable Y is bounded:

$$a \le Y \le b$$

then with probability at least $1-\alpha$:

$$|\bar{Y} - E[Y]| \le (b-a)\sqrt{\frac{\ln(2/\alpha)}{2n}}$$

① Hoeffding inequality says:

$$\mathbb{P}(|\bar{Y} - E[Y]| \ge \varepsilon) \le 2e^{\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)}$$

② we want a bound of the form "with probability at least 95%."

we want a 95% confidence statement. That means we want:

$$\mathbb{P}(|\bar{Y} - E[Y]| \leq \varepsilon) \geq 0.95$$

<u>Equivalently:</u>

$$\mathbb{P}(|\bar{Y} - E[Y]| \geq \varepsilon) \leq 0.05$$

So we set

- Confidence level: $1 - \alpha = 0.95$
- So $\alpha = 0.05$

We want to choose ε so that the hoeffding bound is at most α:

$$2e^{\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)} \leq \alpha$$

③ <u>Solve this inequality for ε:</u>

$$2e^{\left(\frac{-2n\varepsilon^2}{(b-a)}\right)} = \alpha$$

$$\Rightarrow e^{\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)} = \frac{\alpha}{2}$$

$$\Rightarrow \frac{-2n\varepsilon^2}{(b-a)^2} = \ln\left(\frac{\alpha}{2}\right)$$

$$\Rightarrow \varepsilon^2 = \frac{(b-a)^2 \cdot \ln\left(\frac{\alpha}{2}\right)}{-2n}$$

$$\Rightarrow \varepsilon = (b-a) \cdot \sqrt{\frac{\ln\left(\frac{\alpha}{2}\right)}{-2n}}$$

What are a and b?

Our Y is: $Y = \sin(10x)$,

and regardless of X, the sine function always satisfies:

$$-1 \leq \sin(\cdot) \leq 1$$

Therefore:

- $a = -1$ (smallest possible value)
- $b = 1$ (largest possible value)

The 95% confidence interval:

$$\Rightarrow (\hat{\mu} - \varepsilon, \hat{\mu} + \varepsilon)$$

Problem 1:

5

5. [4p]

Can you calculate an approximation of the value of C from part 1 using random samples? Provide a plot of the histogram from part 2 together with the true density as a curve, recall that this requires the value of C. Explain what method you used and what answer you got.

key idea:

- Because f is a density, its integral must be 1, and that condition determines C. We then approximate that integral using random samples (Monte Carlo integration), and from that we get an approximation of C.

Step 1:

- We have the unnormalized function
  $\hat{f}(x) = x^{0.2}(1-x)^{1.3}$, $0 \le x \le 1$
- and the density is
  $f(x) = C \cdot x^{0.2}(1-x)^{1.3}$ on $[0,1]$

Step 2: Use the fact that f is a probability density:

Since f is a density on $[0,1]$, it must integrate to 1.

$$\int_0^1 f(x)\,dx = 1$$

$$\Rightarrow \int_0^1 C \cdot x^{0.2}(1-x)^{1.3}\,dx = 1$$

$$\Rightarrow C \cdot \int_0^1 x^{0.2}(1-x)^{1.3}\,dx = 1$$

Define $I := \int_0^1 x^{0.2}(1-x)^{1.3}\,dx$

Then the normalization constant is:

$$C = \frac{1}{I}$$

Step 3: Rewrite integral as expectation:

Let U be a random variable with distribution Uniform(0,1).

Its density is $g(x) = 1$ on $[0,1]$.

For any function $h(x)$,

$$\boxed{h(U) = \tilde{f}(x)}$$

$$E[h(U)] = \int_0^1 h(x) g(x) dx = \int_0^1 h(x) \cdot 1 \, dx$$

Here, choose $h(x) = x^{0.2}(1-x)^{1.3}$, then

$$I = \int_0^1 x^{0.2}(1-x)^{1.3} dx = E[h(U)]$$

Step 4: Monte Carlo estimate of the integral:

To approximate $E[h(U)]$, we:

- Draw i.i.d samples $U_1, \dots, U_m \sim$ Uniform(0,1)
- Compute $h(U_i) = U_i^{0.2}(1-U_i)^{1.3}$ for each sample.
- Use the sample mean as an estimate of the expectation: $\hat{I} = \frac{1}{m} \cdot \sum_{i=1}^{m} h(U_i)$ $\boxed{h(U) = \tilde{f}(x)}$

For large m (ex $m = 100000$), $\hat{I}$ will be close to I.

Step 5: Estimate C from the Monte Carlo integral:

Since $C = \frac{1}{I}$, we approximate with:

$$\hat{C} = \frac{1}{\hat{I}} = \frac{1}{\frac{1}{n} \cdot \sum_{i=1}^{n} x^{0.2}(1-x)^{1.3}}$$

Problem 2

2 2. [4p]

Follow the calculation from the lecture notes where we derive the logistic regression and implement the final loss function inside the class `ProportionalSpam`. You can use the Test cell to check that it gives the correct value for a test-point.

Made this derivation in the code in a markdown text cell.

Problem 2

4 4. [3p]

Use the trained model `problem2_ps` and the calibrator `problem2_calibrator` to make final predictions on the testing data, store the prediction in `problem2_final_predictions`.

Compute the $0 - 1$ test-loss and store it in `problem2_01_loss` and provide a 99% confidence interval of it.
Store this interval in the variable `problem2_interval` (this should again be a tuple as in Problem 1).

Conceptually:

- problem2_ps.predict(problem2_X_test) gives us the model's probabilities.
- Then we reshape to (n_test, 1) because the

calibrator expects a 2D input.

- problem2_calibrator.predict(...) outputs calibrated probabilities:

$$\hat{p}_i \approx \mathbb{P}(Y_i = 1 \mid X_i)$$

for each test email i.

So: After each line, problem2_final_predictions is a 1D-array of probabilities in $[0,1]$, one per test sample.

Step 2: Turn probabilities into hard 0/1 predictions:

We now need decisions (spam/not spam) to compute the 0-1 loss.

For 0-1 loss with equal cost for FP and FN, the Bayes classifier is:

- Predict $\hat{y}_i = 1$ if $\hat{p}_i \geq 0.5$
- Predict $\hat{y}_i = 0$ if $\hat{p}_i < 0.5$

Step 3:

The 0-1 loss for a single observation is:

$$l_i = \begin{cases} 1, & \text{if } \hat{y}_i \neq y_i \\ 0, & \text{if } \hat{y}_i = y_i \end{cases} = \mathbb{1}\{\hat{y}_i \neq y_i\}.$$

The test 0-1 loss is the average error:

$$\hat{L}_{0/1} = \frac{1}{n_{test}} \cdot \sum_{i=1}^{n_{test}} l_i$$

Step 4: 99% confidence interval Hoeffding inequality

Define the random variables

$$L_i = \mathbb{1}\{\hat{Y}_i \neq Y_i\}, \quad i = 1, \dots, n_{test}.$$

Each L_i takes values in $[0,1]$, either 0 or 1, and we assume i.i.d:

- True (unknown test error):

$$L = E[L_i]$$

- Empirical test error (what we computed):

$$\hat{L} = \frac{1}{n} \cdot \sum_{i=1}^{n} L_i = \text{problem2\_01\_loss}(\,)$$

Hoeffding's inequality for bounded i.i.d variables $L_i \in [a,b]$ says:

$$\mathbb{P}(|\hat{L} - L| \geq \varepsilon) \leq 2e^{\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)}$$

In our case:

- $a = 0$, $b = 1$, (0-1 errors)
- So $b - a = 1$
- Let $n = n_{test} = $ len(problem2_Y_test())

We want a 99% confidence interval, i.e., probability that $|\hat{L} - L|$ is less than ε is at least 99%. That means:

$$\mathbb{P}(|\hat{L} - L| \geq \varepsilon) \leq \alpha = 0.01$$

Set Hoeffding's upper bound equal to α:

$$2e^{(-2n\varepsilon^2)} = 0.01$$

$$\Rightarrow e^{(-2n\varepsilon^2)} = \frac{0.01}{2}$$

$$\Rightarrow -2n\varepsilon^2 = \ln\left(\frac{0.01}{2}\right)$$

$$\Rightarrow \varepsilon^2 = \frac{\ln\left(\frac{0.01}{2}\right)}{-2n}$$

$$\Rightarrow \varepsilon = \sqrt{\frac{\ln\left(\frac{0.01}{2}\right)}{-2n}}$$

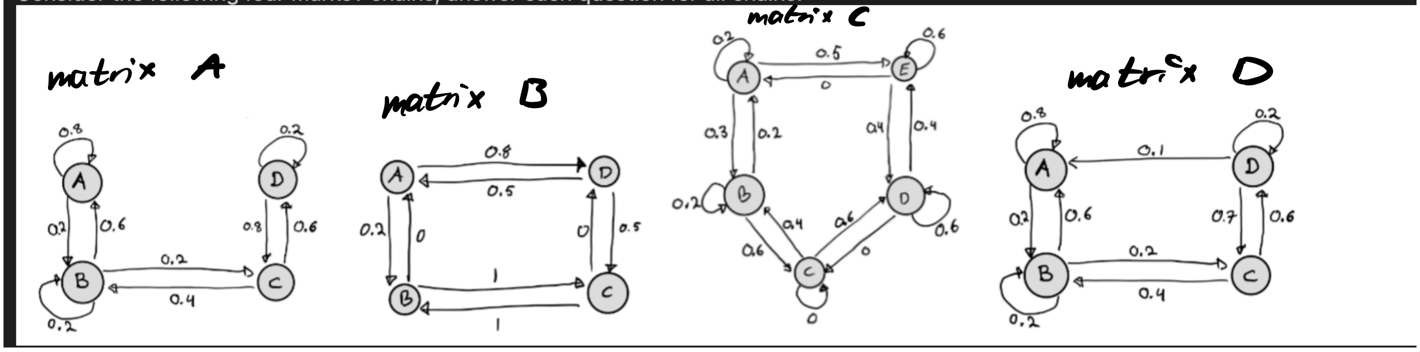
So: Then the 99% confidence interval for the true test 0-1 loss is:

$$(\hat{L} - \varepsilon, L + \varepsilon)$$

Problem 3

1

Consider the following four Markov chains, answer each question for all chains:



1. [2p]

What is the transition matrix?

Matrix A:

	A	B	C	D
A	0.8	0.2	0	0
B	0.6	0.2	0.2	0
C	0	0.4	0	0.6
D	0	0	0.8	0.2

Matrix B:

	A	B	C	D
A	0	0.2	0	0.8
B	0	0	1	0
C	0	1	0	0
D	0.5	0	0.5	0

Matrix C:

	A	B	C	D	E
A	0.2	0.3	0	0	0.5
B	0.2	0.2	0.6	0	0
C	0	0.4	0	0.6	0
D	0	0	0	0.6	0.4
E	0	0	0	0.4	0.6

Matrix D:

	A	B	C	D
A	0.8	0.2	0	0
B	0.6	0.2	0.2	0
C	0	0.4	0	0.6
D	0.1	0	0.7	0.2

Problem 3

2

2. [2p]

Is the Markov chain irreducible?

- A Markov chain is irreducible if:
  - You can get from every state to every other state
  - As long as there is some path with positive probability.

* If matrix has no zeros at all, it is trivially irreducible, because all transitions can occur in 1 step.

* This does not need to necessarily happen in one step

So: No zeros in matrix -> irreducible!
And we need to be able to come back to the starting state, it have to be a complete link, for every state in the Markov chain.

So:

Matrix A: irreducible

Matrix B: No!

Matrix C: No!

Matrix D: irreducible

Problem 3

3. 3. [3p]

Is the Markov chain aperiodic?
What is the period for each state?

Hint: Recall our definition of period:
Let

$$T := \{t \in \mathbb{N} : P^t(x,x) > 0\}$$

and the greatest common divisor of T is the period.

(i) Is the Markov chains aperiodic?

(ii) What is the period of each state?

- (ii): The period tells us:
  "What is the greatest common divisor (GCD) of all times t where you can return to that same state exactly at time t.
- (i): Matrix is aperiodic if every state have period 1.

So: We must also include all possible paths to calculate period, and not only the smallest path.

Matrix A:

- **Period**: From: $A \to A$ => Period = 1
  From: $B \to B$ => Period = 1
  From: $C \to C$ => Period = 2, 3 => GCD(2,3) = 1
  From: $D \to D$ => Period = 1

=> GCD(1, 1, 1, 1) = 1

=> Period = 1 => aperiodic!

Matrix B:

- **Period**: From: $A \to A$ => Period = 2, 4 => GCD(2,4) = 2

From : B -> B => Period = 2, 4 => GCD(2,2) = 2
From : C -> C => Period = 2, 4 => -||- 2
From : D -> D => Period = 2, 4 => -||- 2
=> GCD(2,2,2,2) = 2
=> Period 2 => NOT aperiodic

Matrix C:

- Period : From : A -> A => Period = 1
  From : B -> B => Period = 1
  From : C -> C => Period = 1
  From : D -> D => Period = 1
  From : E -> E => Period = 1
  => GCD(1,1,1,1,1) = 1
  => Period = 1 => aperiodic

Matrix D:

- Period : From : A -> A => Period = 1
  From : B -> B => Period = 1
  From : C -> C => Period = 2, 3 => GCD(2,3) = 1
  From : D -> D => Period = 1
  => GCD(1,1,1,1) = 1
  => Period = 1 => aperiodic

Problem ③

④ 4. [3p]

Does the Markov chain have a stationary distribution, and if so, what is it?

We label the stationary distribution as

$\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ for Matrix A:

Transition matrix is still the same:

	A	B	C	D
A	0.8	0.2	0	0
B	0.6	0.2	0.2	0
C	0	0.4	0	0.6
D	0	0	0.8	0.2

Stationary distribution satisfies: $\pi \cdot P = \pi$

$$\Rightarrow \begin{cases} \pi_1 = 0.8\pi_1 + 0.6\pi_2 \\ \pi_2 = 0.2\pi_1 + 0.2\pi_2 + 0.4\pi_3 \\ \pi_3 = 0.2\pi_2 + 0.8\pi_4 \\ \pi_4 = 0.6\pi_3 + 0.2\pi_4 \end{cases}$$

and we also know that: $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$

Put these 5 equations into wolframalpha, then we get: Wolfram not working, use matrixes:

Matrix B:

We label the stationary distribution as

$\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ for **Matrix B**:

Transition matrix is still the same:

	A	B	C	D
A	0	0.2	0	0.8
B	0	0	1	0
C	0	1	0	0
D	0.5	0	0.5	0

Stationary distribution satisfies: $\pi \cdot P = \pi$

$$\Rightarrow \begin{cases} \pi_1 = 0.5\pi_4 \\ \pi_2 = 0.2\pi_1 + 1 \cdot \pi_3 \\ \pi_3 = 1 \cdot \pi_2 + 0.5\pi_4 \\ \pi_4 = 0.8\pi_1 \end{cases}$$

and we also know that: $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$

Put into matrixes and solve:

Same for matrix C and D:

Problem ③

⑤

5. [3p]

Is the Markov chain reversible?

General procedure

For **each** Markov chain:

1. **Take the stationary distribution** π (you already computed it).
2. For every **pair of states** (i, j), compute the **flow**:
   - Forward flow: $\text{flow}_{i \to j} = \pi_i P_{ij}$
   - Backward flow: $\text{flow}_{j \to i} = \pi_j P_{ji}$
3. If for **all** pairs (i, j) you get
   $\pi_i P_{ij} = \pi_j P_{ji}$,
   then the chain **is reversible**.
4. If you find **one** pair where the equality fails, the chain is **not reversible**.

Note: if both $P_{ij} = 0$ and $P_{ji} = 0$, then both sides are 0, so that pair is automatically fine.

Matrix A:

	A	B	C	D
A	0.8	0.2	0	0
B	0.6	0.2	0.2	0
C	0	0.4	0	0.6
D	0	0	0.8	0.2

Stationary distribution:

$$\pi^{(A)} = (\underbrace{0.6154}_{A}, \underbrace{0.2051}_{B}, \underbrace{0.1026}_{C}, \underbrace{0.0769}_{D})$$

So:

1) Check detailed balance on edges where transitions are non-zero.

- Pair A<->B:

- **Forward**: $\pi_A \cdot P_{AB} \approx 0.6154 \cdot 0.2 = 0.1231$

From A to B probability

- **Backward**: $\pi_B \cdot P_{BA} \approx 0.2051 \cdot 0.6 = 0.1231$

- **Equal**: ✓

- Pair $B \leftrightarrow C$:
  - Forward: $\pi_B P_{BC} \approx 0.2051 \cdot 0.2 = 0.0410$
  - Backward: $\pi_C P_{CB} \approx 0.1026 \cdot 0.4 = 0.0410$
  - Equal ✅
- Pair $C \leftrightarrow D$:
  - Forward: $\pi_C P_{CD} \approx 0.1026 \cdot 0.6 = 0.0615$
  - Backward: $\pi_D P_{DC} \approx 0.0769 \cdot 0.8 = 0.0615$
  - Equal ✅

All other pairs either have both transitions 0 or are already implied.

So **A is reversible**.

We do this for <u>all</u> possible paths.

So this you should also check, not only check in order:

Pair: $A \Leftrightarrow C$

$A \Leftrightarrow D$

If they were not 0 (zero).