Markov chain Monte-Carlo

Sampling Complex distributions

Babak Maboudi - day 2 - Jyväskylä summer school 2025

Monte Carlo Integration

- Let X be an $\mathbb{R}^n$-valued random variable, i.e., a random variable which takes values in $\mathbb{R}^n$, and f be an integrable function. Then

$$\mathbb{E}(f(X)) = \int_{\mathbb{R}^n} f(\mathbf{x})\pi_X(\mathbf{x})\ d\mathbf{x} \approx \sum_{j=1}^{N} w_j f(\mathbf{x}_j),$$

- In Monte Carlo Integration, $\mathbf{x}_j$ are i.i.d. realization of π_X, then the approximator becomes the ergodic average:
  - Mean approximation

$$\mathbf{m} = \mathbb{E}(X) \approx \sum_{j=1}^{N} \frac{1}{N}\mathbf{x}_j$$

  - Variance approximation

$$v = \text{Var}(X) = \mathbb{E}(\|X - m\|_2^2) \approx \sum_{j=1}^{N-1} \frac{1}{N-1}\|\mathbf{x}_j - \mathbf{m}\|_2^2$$

Monte Carlo Integration

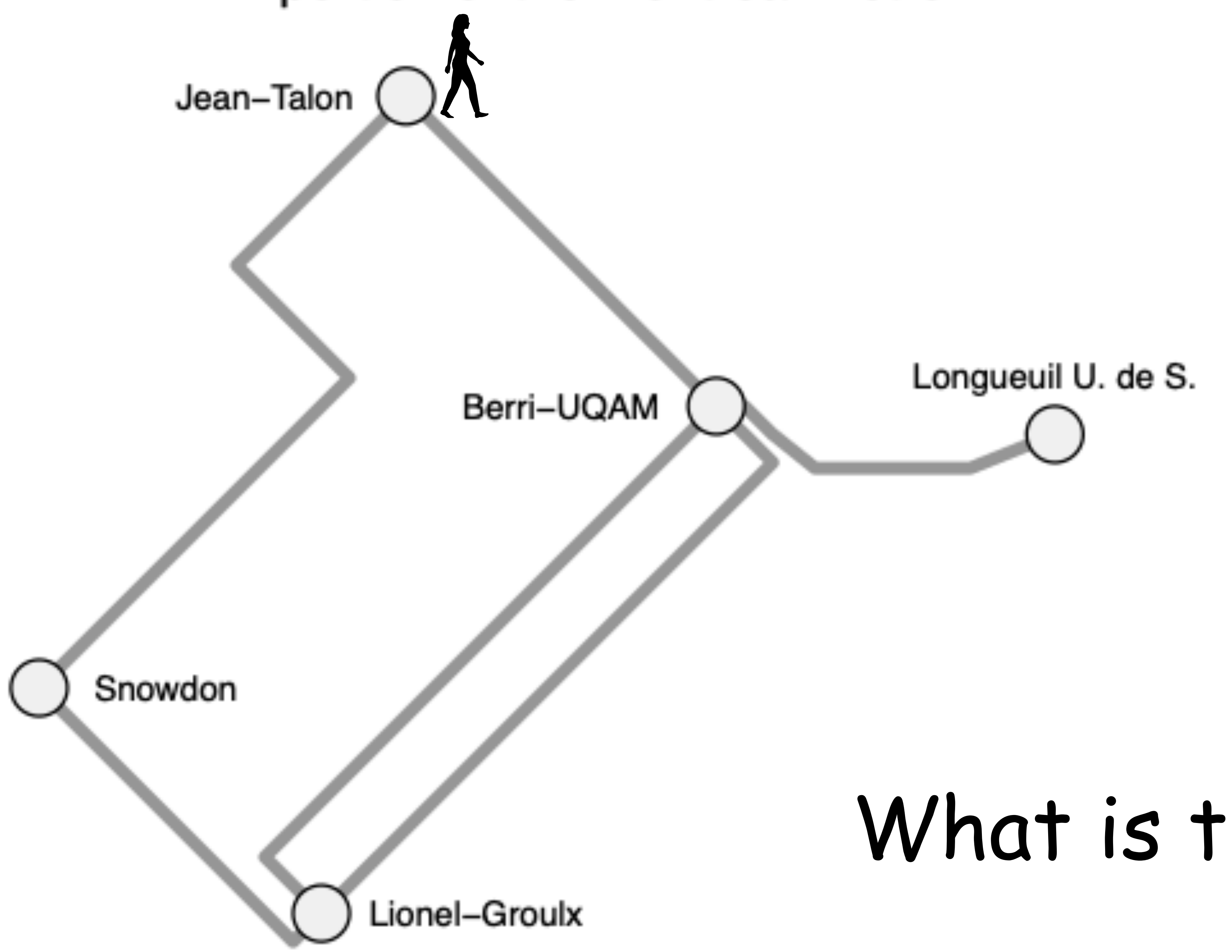
- However, the foundation for Monte Carlo estimation is independent realizations of the distribution of X.
- In Inverse Problems, we rarely have access to the distribution of X, this requires a complete knowledge of the density function.
- However, dependent sampling is possible! This is the principle idea of Markov-chain Monte Carlo methods.

Markov chains

Montréal Metro Map

Exercise from Art B. Owen (2013)

A portion of the Montréal métro



What is the distribution of Alice's location?

Markov chains

Introducing notations

- Let $\Omega = \{\omega_1, \ldots, \omega_M\}$ be the *State Space*.
- Let X be a Ω-*valued random variable*.

Markov chains

Introducing notations

- Definition: A *Markov chain* is a sequence $X_0, X_1, X_2, \ldots, X_N$ (or $\{X_i\}_{i \leq N}$) of random variables with Markov property:
  - $\mathbb{P}(X_{i+1} \in A \,|\, X_j = x_j,\ 0 \leq j \leq i) = \mathbb{P}(X_{i+1} \in A \,|\, X_i = x_i)$
  - Here A is a set of states.
  - This is referred to being memoryless.

Markov chains

Further conditions

- A Markov chain is *(time-)homogeneous* if

$$\mathbb{P}(X_{i+1} = y \,|\, X_i = x) = \mathbb{P}(X_1 = y \in A \,|\, X_0 = x)$$

- A *transition probability* is the probability of going from state ω_i to ω_j:

$$p_{i \leftarrow j} = p_{ij} = \mathbb{P}(X_1 = \omega_i \in A \,|\, X_0 = \omega_j)$$

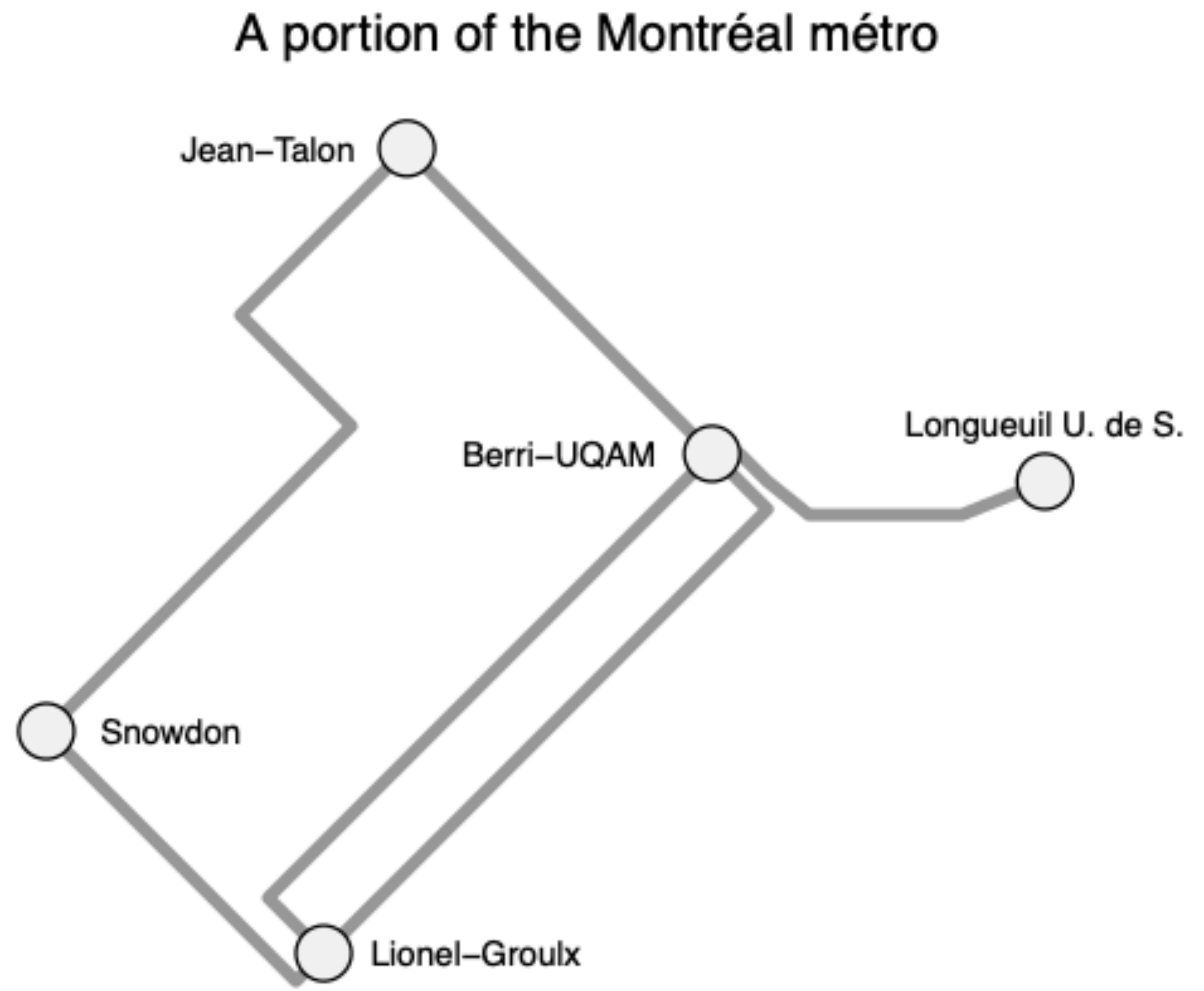
- A *transition matrix* is when you collect all transition probabilities in a matrix.

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1M} \\ p_{21} & p_{22} & \cdots & p_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1M} & p_{M2} & \cdots & p_{MM} \end{pmatrix}$$

Markov chains

Exercise 1 - `exercise_1.py`

- Choose a starting station
- Apply the function `roam` to move to the next station
- Draw 10000 samples from Montréal métro problem.
- Plot the histogram of the samples.
- Which station is the most likely destination?



Markov chains

Exercise 2 - `exercise_2.py`

- Now suppose that the initial state is not deterministic:
  $p_0(\omega_j)$: the probability of being at the jth station on the first step
- Similarly we define:
  $p_n(\omega_j)$: the probability of being at the jth station on the nth step
- What is the probability of being in the second station after 1 step?
  $$p_1(\omega_2) = p_0(\omega_1)p_{21} + p_0(\omega_2)p_{22} + \ldots + p_0(\omega_M)p_{2M} = \sum_j p_{2j}p_0(\omega_j)$$
- Complete the python code `exercise_2.py` to compute $p_1(\omega_2)$ when you are initially at any given station with equal probability, i.e.,
  $$p_0(\omega_j) = 1/5, \quad j = 1,\ldots,5,$$
- Can you write an operation between P and p_0 that gives you all the probabilities $p_1(\omega_j)$, for $j = 1,\ldots,5$?
- What is the sum of the elements in p_1? Why?

Markov chains

- We have
$$p_1 = Pp_0,$$
then,
$$p_2 = Pp_1,$$
and
$$p_n = Pp_{n-1}.$$

- What is the transition matrix Q for doing 2 steps? i.e., what is the matrix Q that gives you:
$$p_2 = Qp_0$$
Hint: look at the Markovian principle (the recursive definitions above).

Markov chains

Exercise 3

- Compute the transition matrix, P^2, for 2 steps?
- Compute the transition matrix, P^{200}, for 200 steps?
- What does the pattern in P^{200} mean?
  Hint: the component $[P^{200}]_{ij}$, i.e., the element on the ith row and the jth column of P^{200}, means the probability of starting at the station j and after 200 steps of the Markov chain arriving at station i.

Sampling using a Markov chain

Explanation,

- What ever value X_0 has it will be almost forgotten (independent) in X_{100}.
- What ever value X_{100} has, it will be forgotten (independent) in X_{200}.
- If we take a widely separated sequence of equi-spaced samples we should get a nearly i.i.d. samples.
- Repeat the Markov chain sampling in `exercise_1.py`, but this time select only every 100 samples. Create a histogram and normalize it.
- Find the distribution p_{1000}, i.e., $P^{1000}p_0$. Compare the histogram with p_{1000}.

Markov Chain

Stationary distribution

- We say π is a stationary distribution when:
$$P\pi = \pi$$
In other words, the transition matrix doesn't change the distribution.

Irreducible and periodic transition kernels

- What can you say about these transition matrices? Do they have a unique stationary distribution?

$$P_1 = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}, \qquad P_1 = \begin{pmatrix} 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{pmatrix}.$$

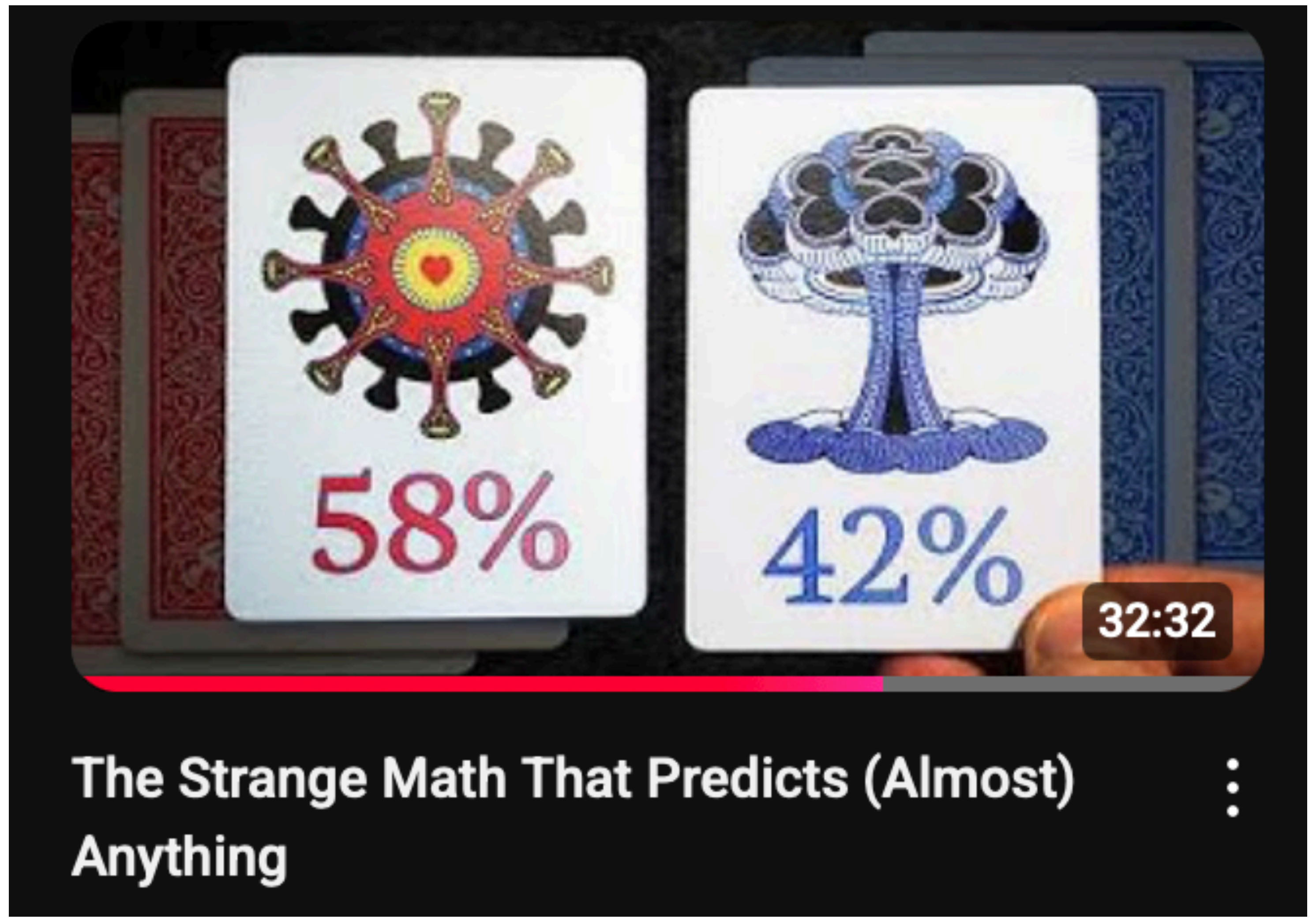
Irreducible and aperiodic transition kernels

- Theorem: If a transition matrix P is irreducible and aperiodic, and has a stationary distribution π then:

$$\lim_{n \to \infty} \mathbb{P}_{\omega_0}(X_n = \omega) = \pi(\omega)$$

- This "means" that the Markov chain method can arrive at the stationary distribution.

Veritasium video on Markov and Markov chains

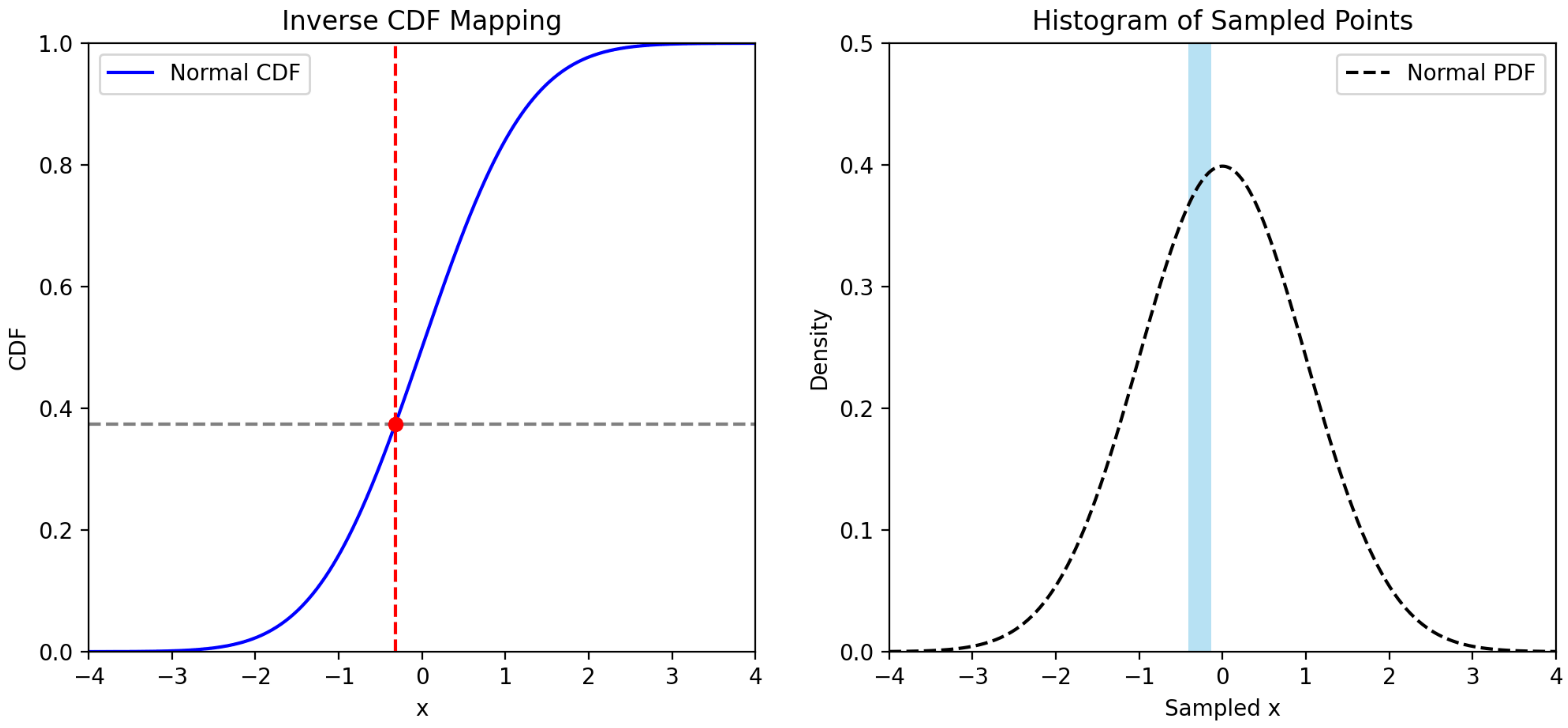


Acceptance/Rejection Sampling

Sampling from a Distribution

The inverse CDF method

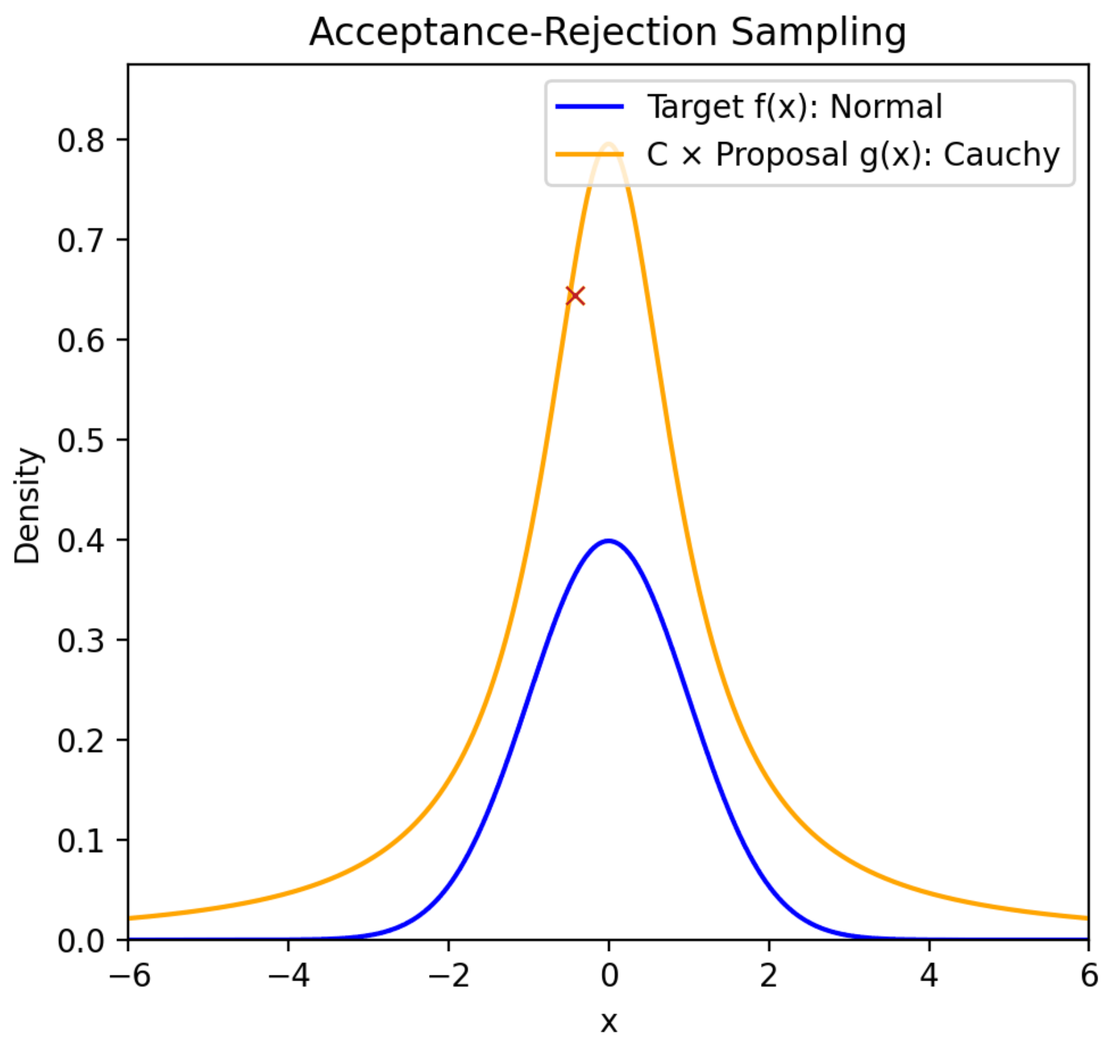
$$F_X(x) = \mathbb{P}(X < x) = \int_{-\infty}^{x} \pi_X(x)\ dx$$

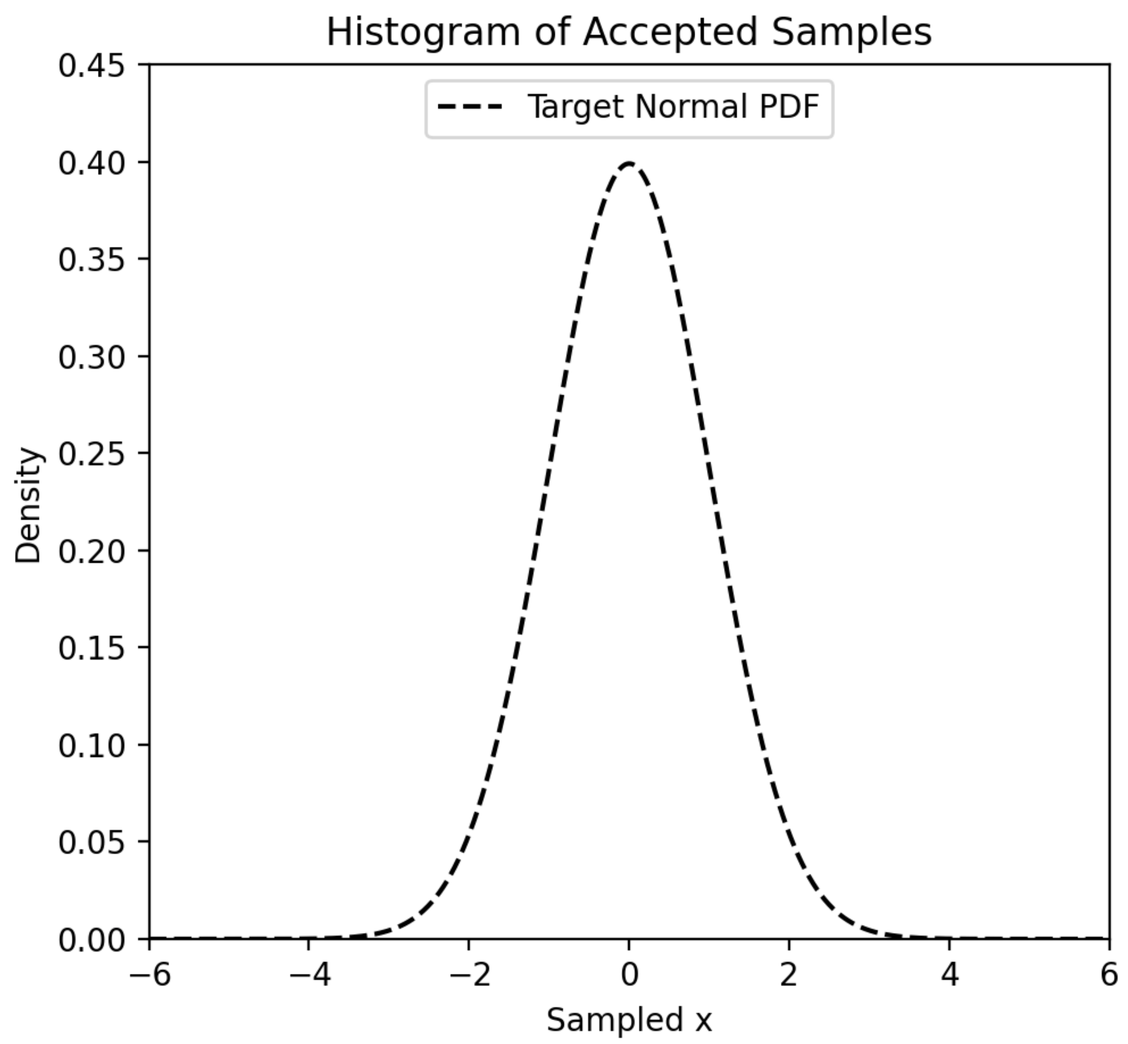


Sampling from a Distribution

Acceptance/Rejection method

- f is the density of the target distribution
- g is a density of a distribution that is easy t sample from (e.g. using the inverse CDF method).
- Algorithm:
  1. Sample y according to g
  2. Sample u according to $U(0,1)$
  3. If $u \leq f(y)/(Cg(y))$ accept, otherwise reject
  4. Repeat until desired samples achieved.





Acceptance/Rejection Sampling

Exercise (HW2)

- Get the Python script `exercise_4.py` from day 2 folder.
- Write Python functions `f(x)` and `g(x)` that computes the density functions of target Gaussian distribution and proposal Cauchy distribution:

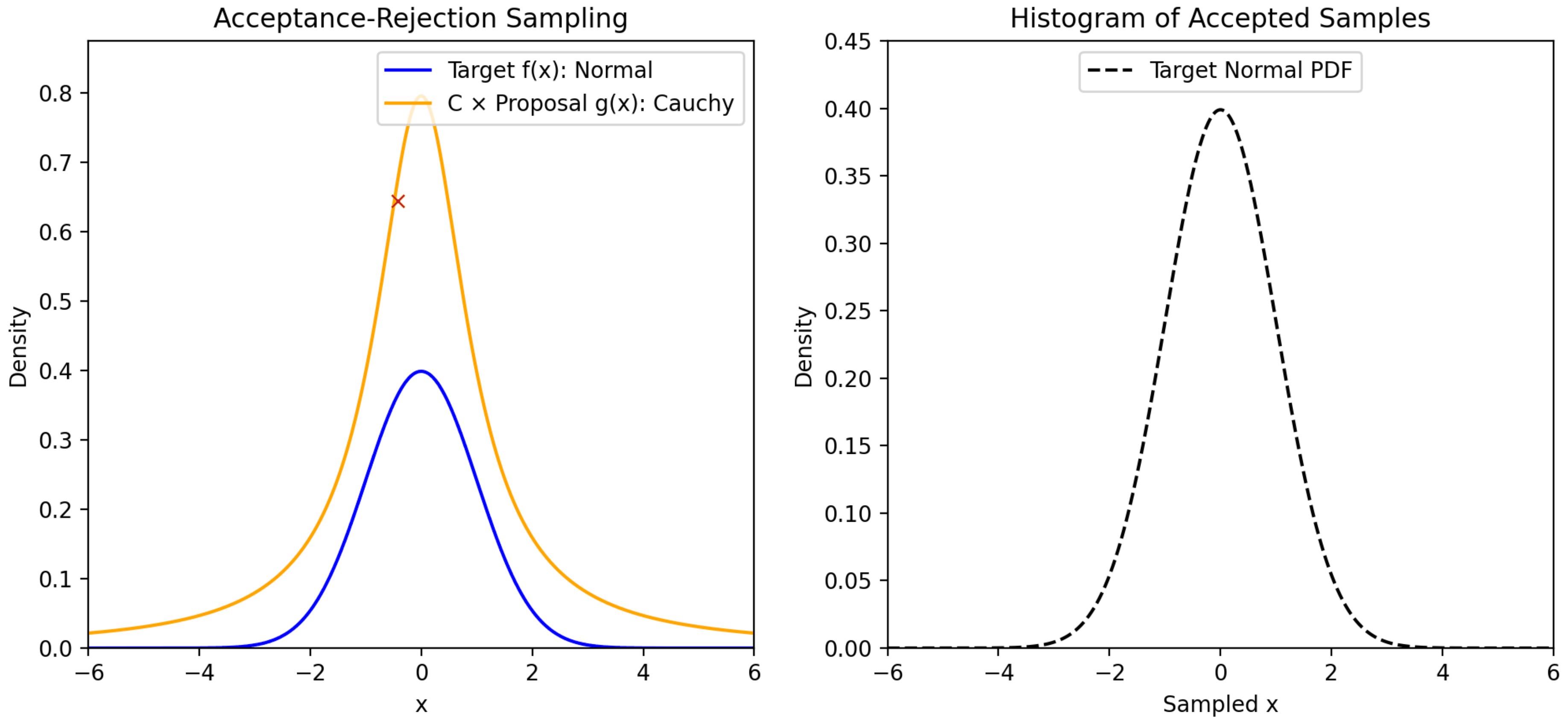
$$f(x) = \frac{1}{\sqrt{(2\pi)}} \exp(-\frac{x^2}{2}), \qquad g(x) = \frac{1}{\pi(1+x^2)}$$

- Set $c = 2$ and perform acceptance/rejection to draw 2000 samples from the distribution of f, i.e.,
  - draw a sample $x^\star$ from the proposal distribution g.
  - Draw a number from the uniform distribution $u \sim U(0,1)$
  - If $cg(x)u \leq f(x)$ accept $x^\star$ as a sample from f, otherwise reject.
- Plot the histogram of the samples and show that they approximate a standard-normal distribution.
- Choose the "step-size" $c = 1{,}1.52$, and 2 and repeat the sampling. Compute the number of accepted samples. Which value is the best and why?

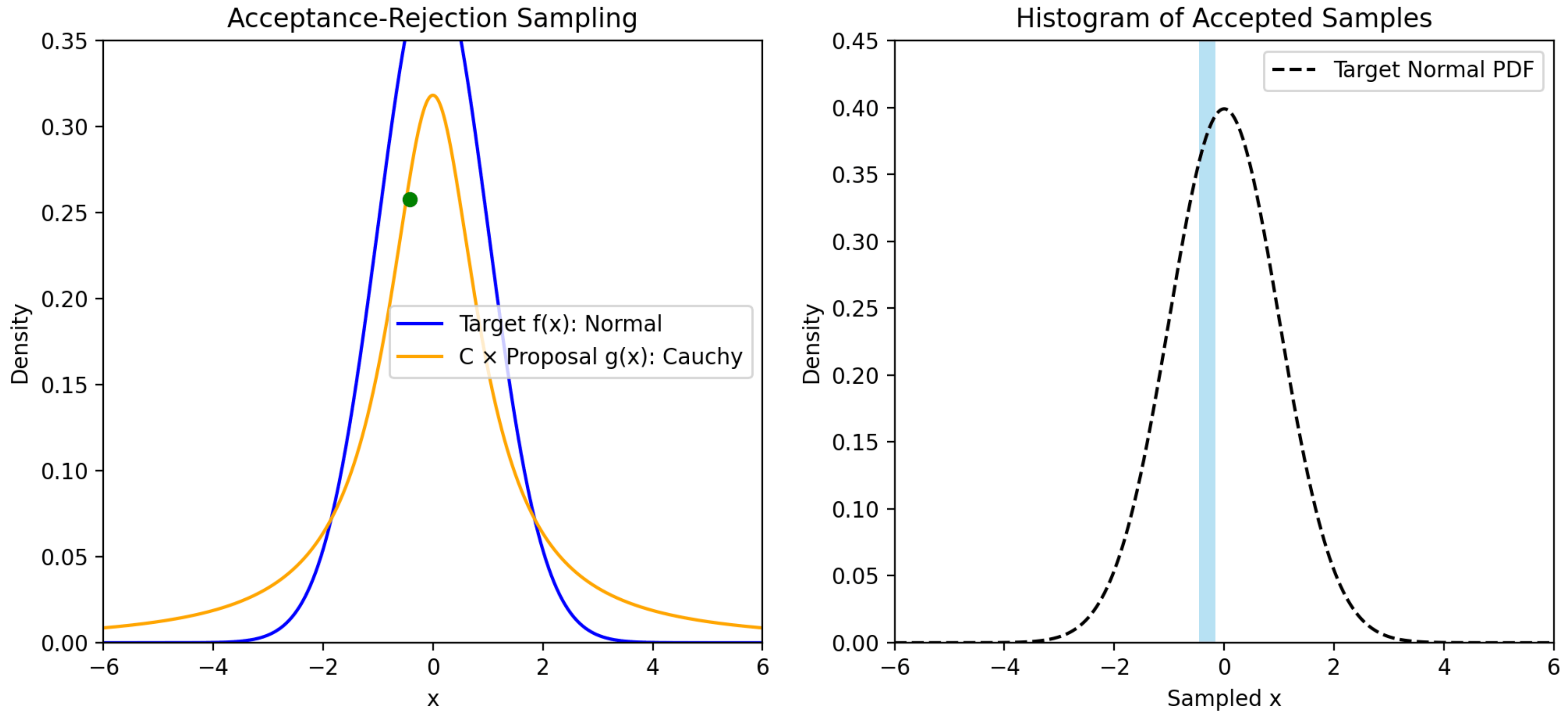
Sampling from a Distribution

Acceptance/Rejection method

$c = 2.5$ acc. rate $= 0.39$



$c = 1$ acc. rate $= 0.71$



$c = 1.52$ acc. rate $= 0.63$