Introduction to Markov Chains

From sampling methods to Markov chain Monte Carlo

Computational Statistics

Recap: generating samples from a given distribution

Last week, we discussed methods for generating random samples.

- **Pseudorandom number generators:** generate pseudorandom numbers from a uniform distribution.
- **Box–Muller:** use uniform samples to generate normal samples.
- **Inverse transform sampling:** use an invertible cumulative distribution function.
- **Rejection sampling:** use a suitable proposal distribution and a constant C.

Rejection sampling: the proposal and the constant

Rejection sampling is flexible in principle, provided we have:

- A suitable proposal distribution from which we can sample.
- A finite constant C giving the required bound.

Required bound

For target density p and proposal density q,

$$p(x) \leq Cq(x)$$

throughout the target's support.

From univariate to multivariate normal samples

The Box–Muller method generates normal samples.

- Start with independent samples from a univariate standard normal distribution.
- Use a transformation to obtain multidimensional, correlated normal samples.
- This allows sampling from normal distributions in one or more dimensions.

The point of the recap

For normal distributions, sampling is relatively straightforward.

Inverse transform sampling: scope and limitations

We considered inverse transform sampling when the cumulative distribution function is strictly increasing and invertible.

- The basic version considered works directly for a **univariate distribution**.
- This is a one-dimensional sampling problem.
- Computing the inverse can also involve computational cost.

Main limitation discussed

The direct version introduced in the lecture applies to one dimension.

Rejection sampling can become inefficient

Rejection sampling can work in multiple dimensions, but its performance can deteriorate as dimensionality increases.

Example

Suppose we propose **100 samples**, but only **three are accepted**.

- Most proposals are rejected, giving a low acceptance rate.
- In some problems, three, four or five dimensions may be practical, while ten may already be difficult.

These numbers are illustrative, not a universal dimensional limit.

Practical problems can be very high-dimensional

In applications, we may need samples from distributions with:

- A hundred or several hundred dimensions.
- Thousands of dimensions.

Large and complicated data sets can lead to high-dimensional problems.

Why the earlier methods may be insufficient

They may not be sufficiently efficient or effective in this setting, especially rejection sampling.

We need a more flexible sampling approach

We want methods that can:

- Deal with high-dimensional problems.
- Generate samples from a broad range of target distributions.

Box–Muller, together with a suitable transformation, can handle high-dimensional normal distributions.

We also need to sample from more general distributions.

Markov chain Monte Carlo

The approach we will study

Markov chain Monte Carlo, usually abbreviated to **MCMC**.

- A widely used approach to sampling from different distributions and in different dimensions.
- A **family of algorithms**, with many variants sharing a similar underlying idea.
- We first need to understand what a **Markov chain** is, both mathematically and intuitively.

MCMC is not automatically efficient for every target.

A Markov chain is a stochastic process

A Markov chain is a particular kind of **stochastic process**, also called a **random process**.

- We will discuss **discrete-time** stochastic processes.
- Continuous-time stochastic processes also exist, but are outside this discussion.

Definition used in the lecture

A stochastic process is a collection of random variables indexed by time.

A finite or an infinite sequence

A process can consist of a finite sequence, for example

$$X_1,\ X_2,\ X_3.$$

It can also consist of an infinite sequence:

$$X_1,\ X_2,\ \ldots,\ X_n,\ \ldots.$$

The important feature

The random variables are indexed by time. The time index set may be finite or infinite.

The order of the random variables matters

In the process

$$X_1,\ X_2,\ X_3,\ \ldots,$$

X_1 comes first, followed by X_2, followed by X_3.

- The indices usually represent successive observation times.
- Reordering the observations changes their time order.
- More general index sets can describe, for example, locations in space.

In this module, we interpret the index as time.

Examples: stock prices and temperatures

Stochastic processes can model physical, chemical and biological phenomena.

Financial example

The price of a particular stock over successive days.

Weather examples

Temperature, or average temperature, on day one, day two, and so on.

Alternatively, temperature this hour, next hour, and the hour after that.

Why these are stochastic-process models

1. **The values are uncertain.**

   We cannot determine future stock prices or temperatures completely before they occur.

2. **The observations are indexed by time.**

   There is a clear order: one observation follows another.

These two features allow us to model the sequences as stochastic processes.

A special kind of stochastic process

In this discrete-time discussion, we refer to a Markov process as a **Markov chain**.

Change of notation

From now on, use a superscript for the time index:

$$X^1,\ X^2,\ X^3,\ \ldots.$$

A Markov chain is a stochastic process with a particular property of its conditional distributions.

Looking at the process up to time t

Consider the random variables

$$X^1,\ X^2,\ \ldots,\ X^t.$$

These are t random variables, so we can discuss their joint distribution, marginal distributions and conditional distributions.

The conditional distribution of interest

The distribution of X^t given all the previous states:

$$p(x^t \mid x^{t-1}, x^{t-2}, \ldots, x^1).$$

The Markov property

Definition

The process is a Markov chain if

$$\boxed{p(x^t \mid x^{t-1}, x^{t-2}, \ldots, x^1) = p(x^t \mid x^{t-1}).}$$

The distribution given the entire previous history equals the distribution given only the immediately preceding state.

This must hold at each relevant time, for histories where the conditional distributions are defined.

Here p denotes the appropriate conditional probability mass function or density.

Stock-price example: predicting day 10

Suppose we know a particular stock's prices on:

$$\text{day } 1, \text{day } 2, \ldots, \text{day } 9.$$

We want to use this information to predict the price on **day 10**.

Using the conditional model

Knowing the observed history and the conditional distribution allows us to describe the distribution of the price on day 10.

Prediction describes uncertainty

Knowing the conditional distribution does not generally tell us the exact next price with certainty.

- We cannot generally assert: "The price will be 100."
- We can describe which prices, or ranges of prices, are more or less likely.

What we predict

The conditional probability distribution of the price on day 10, given the prices on days 1 through 9.

What changes under the Markov assumption?

To predict the price distribution on day 10, all we need from the observed history is the price on day 9.

- Day 9 is the present, and day 10 is the future.
- Given today's price and the conditional model, earlier prices add no predictive information.
- The predictive distribution is the same as when we use the full observed history.

The example illustrates the Markov assumption. It does not establish that actual stock prices satisfy it.

The future depends on the present state

Interpretation of the Markov property

Once the present state is given, the earlier history adds no information for predicting the next state.

Knowing what is happening today is enough to predict tomorrow's distribution under the model.

We do not also need to know what happened yesterday or the day before yesterday.

To predict at time t, all we need from the past is the state at time $t - 1$.

What values can the process take?

At a fixed time, X^t is a random variable.

- It may be **continuous**.
- It may be **discrete**, with a finite or countably infinite set of possible values.

Our restriction from now on

Each X^t will be discrete and take values in a **finite set**.

A finite state space keeps the discussion simple

For example, each X^t might take only the values

$$1,\ 2,\ 3,$$

or only the values

$$1,\ 2,\ 3,\ 4,\ 5.$$

- We impose this restriction to simplify the discussion and proofs.
- A Markov chain does not have to have a finite state space.

Time indices and state labels

Write the finite state space as

$$\mathcal{X} = \{x_1, x_2, \ldots, x_S\}.$$

Superscript: time

X^t is the random variable at time t.

Subscript: state label

x_i is the ith possible value in the state space.

X^t is a random variable, while x_i is a possible value of that variable.

How do we describe a Markov chain?

A finite state space makes the description simpler. We need two ingredients:

1. A distribution at the starting time.
2. The conditional distributions describing how the chain moves from one time to the next.

First: the marginal distribution

We write the distribution at a given time as a **row vector**.

The marginal distribution at time t

Write the marginal distribution as an S-dimensional row vector:

$$\boxed{\pi^t = (\pi_1^t, \pi_2^t, \ldots, \pi_S^t),}$$

where

$$\pi_i^t = \mathbb{P}(X^t = x_i), \qquad i = 1, \ldots, S.$$

- The first entry is the probability that $X^t = x_1$.
- The second entry is the probability that $X^t = x_2$, and so on.

One probability for each possible state

The vector describes the whole distribution at time t, because the state space is finite.

- If $S = 3$, the state space has three elements, so the probability vector has three entries.
- If $S = 4$, the state space has four elements, so the probability vector has four entries.

The entries correspond to all the possible states.

The probabilities must sum to one

Because π^t is a probability distribution,

$$\pi_i^t \geq 0 \quad \text{for every } i,$$

and

$$\boxed{\sum_{i=1}^{S} \pi_i^t = 1.}$$

Adding the probabilities of all possible states gives one.

The first ingredient

Specifying the initial distribution tells us how the chain starts.

The second ingredient: the conditional distribution

The conditional distribution tells us how the chain behaves over time.

- Knowing today's state, we can predict the next day's distribution.
- Using the next day's state, we can then predict the day after that.

Evolving the Markov chain

The conditional transition probabilities allow us to make predictions forward in time.

The next state conditional on the present

By the Markov property, we only need

$$\boxed{p(x^t \mid x^{t-1}).}$$

We do not need a separate next-state distribution conditional on the whole history.

Finite state space

These conditional probabilities have a simple representation as a matrix, called a **transition matrix**.

The transition matrix

The transition matrix for the step from time $t-1$ to time t is an $S \times S$ matrix:

$$\boxed{P^{t-1} = \left(P_{ij}^{t-1}\right)_{i,j=1}^{S}.}$$

Its entries are

$$\boxed{P_{ij}^{t-1} = \mathbb{P}(X^t = x_j \mid X^{t-1} = x_i).}$$

Each entry is the probability of the next state being x_j, given that the current state is x_i.