Review

- Markov chain: defined by initial distribution $p_0 \in \mathbb{R}^D$, transition matrix $T \in \mathbb{R}^{D \times D}$
  \[ p_0(i) = P(X_0 = i), \quad T_{ij} = P(X_t = j | X_{t-1} = j) \]
- Defines distribution of chain $X_0, X_1, X_2, \ldots, X_t, \ldots$ (with Markov assumption)
- Joint probability
  \[ p(x_1, x_2, \ldots, x_N | x_0) = p(x_1 | x_0)p(x_2 | x_1) \cdot \ldots \cdot p(x_N | x_{N-1}) \]
- Next: $t$-step distributions $p(x_t | x_0)$ and $p(x_t)$

The $t$-Step Distribution for Fixed $x_0$

**Question:** What is the marginal probability distribution after $t$ steps given that the chain starts at $x_0$? I.e., what is $p(x_t | x_0)$?

**Examples:**
\[
\begin{align*}
p(x_1 | x_0) &= T_{x_0 x_1}, \\
p(x_2 | x_0) &= \sum_{x_1} p(x_1, x_2 | x_0) = \sum_{x_1} p(x_1 | x_0) T_{x_1 x_2}.
\end{align*}
\]

In general, we have the recursive expression:
\[
p(x_t | x_0) = \sum_{x_{t-1}} p(x_{t-1}, x_t | x_0) = \sum_{x_{t-1}} p(x_{t-1} | x_0) T_{x_{t-1} x_t}.
\]
The \( t \)-Step Distribution for Random \( X_0 \)

**Question:** What is the marginal probability distribution after \( t \) steps given that \( X_0 \sim p_0 \)? I.e., what is \( p(x_t) \)?

By similar logic:
\[
p(x_1) = \sum_{x_0} p(x_0, x_1) = \sum_{x_0} p(x_0)T_{x_0x_1},
\]
\[
p(x_1) = \sum_{x_0} p(x_1, x_2) = \sum_{x_1} p(x_1)T_{x_1x_2}.
\]

In general:
\[
p(x_t) = \sum_{x_{t-1}} p(x_{t-1}, x_t) = \sum_{x_{t-1}} p(x_{t-1})T_{x_{t-1}x_t}.
\]

---

\-Step Recurrence as Matrix-Vector Multiplication

The recurrences for the \( t \)-step distributions can be expressed using matrix-vector multiplication. Let \( p_t \) be the row-vector
\[
p_t = [P(X_t = 1), P(X_t = 2), \ldots, P(X_t = D)].
\]

Then, since \( T_{ij} = P(X_t = j|X_{t-1} = i) \), we can write the above recursive relationship as
\[
p_t = p_{t-1}T.
\]

---

\-Step Distribution as Matrix Power

By unrolling the recurrence, the \( t \)-step distribution can be obtained as a matrix power
\[
p_t = p_{t-1}T = (p_{t-1})T = (p_{t-2})T^2 = (p_{t-3})T^3 = \cdots = p_0 T^t.
\]
Thus

\[ p_t = p_0 T^t. \]

This also implies that \( T^t \) is the \( t \)-step transition matrix

\[ (T^t)_{ij} = P(X_t = j \mid X_0 = i) = P(X_{s+t} = j \mid X_s = i) \]

Limiting Distribution

What happens as \( t \) becomes large? Does \( p_t \) converge to a some limiting distribution \( \pi \)? That is, is there some \( \pi \) such that the following is true?

\[ \lim_{t \to \infty} p_t = \pi \] (limiting distribution)

The algorithmic idea of Markov chain Monte Carlo is:

- Suppose \( \pi \) is hard to sample from directly
- If we can design a Markov chain such that \( \lim_{t \to \infty} p_t = \pi \), then we can draw samples by simulating the Markov chain for many time steps
- It’s remarkable that this could be possible, but it can be done for very general target distributions!
- We need to reason about limiting distributions their properties

Stationary Distribution

Suppose a chain converges exactly, so that \( p_t = p_{t+1} = \pi \). Since \( p_{t+1} = p_t T \), this implies

\[ \pi = \pi T \] (stationary distribution)

- we call any such \( \pi \) a stationary distribution of the Markov chain
- If you start from \( \pi \) and run the chain for any number of steps, the distribution is unchanged.
- If \( \pi \) is a limiting distribution, it is a stationary distribution
- (Linear algebra connection: \( \pi \) is an eigenvector of \( T \) with eigenvalue 1. Useful for computing stationary distributions.)
Stationary and Limiting Distributions

We reason about limiting distributions via stationary distributions:

- If a Markov chain: (1) converges, and (2) has a unique stationary distribution \( \pi \), then it converges to \( \pi \).

- When can we guarantee (1) and (2)? What could go wrong?

What Could Go Wrong: Periodicity

A Markov chain can fail to converge by being periodic:

What Could Go Wrong: Reducibility

A Markov chain can fail to have a unique stationary distribution by being reducible:

Regularity

A Markov chain is regular if there exists a \( t \) such that, for all \( i,j \) pairs,

\[
(T^t)_{ij} > 0,
\]

- Recall that \( T^t \) is the \( t \)-step transition probability matrix. This means it is possible to get from any state \( i \) to any state \( j \) in exactly \( t \) steps.

- A regular Markov chain cannot be periodic or reducible (why?), and guarantees the desired computational property

**Theorem:** A regular Markov chain has a unique stationary distribution \( \pi \) and 
\[
\lim_{t \to \infty} p_t = \pi
\]

for all starting distributions \( p_0 \).

(We can sample from the unique stationary distribution by simulating the chain.)
Summary: Markov Chain Theory

- **t-step distribution**: Distribution of $X_t$, obtained by repeated multiplication with transition matrix: $p_t = p_0 T^t$
- **Limiting distribution**: the distribution of $\lim_{t \to \infty} p_t$, if it exists
- **Stationary distribution**: a distribution $\pi$ such that $\pi T = \pi$. If you start from $\pi$ and run the chain for any number of steps, the distribution is unchanged. Every limiting distribution is a stationary distribution.
- **Regularity**: if there is a $t$ such that $(T^t)_{ij} > 0$ for all $i, j$, a Markov chain is regular. It is possible to get from any state $i$ to any state $j$ in exactly $t$ steps.
- **Convergence to stationary distribution**: if $T$ is regular, the chain converges to a unique stationary distribution $\pi$ for any starting distribution.

High-Level Idea

Suppose we want to sample from $p$, but can’t do so directly. Instead, we can

- **Design a Markov chain** that has $p$ as a stationary distribution
- **Run it for a long time** to get a sequence of states $x_1, x_2, \ldots, x_S$
- **Approximate an expectation** as
  \[ E_{p(X)}[f(X)] \approx \frac{1}{S} \sum_{t=1}^{S} f(x_t). \]

If we run the chain long enough, the approximation will be good! We can often make the following guarantees:

- **Asymptotically correct**: $\lim_{S \to \infty} \frac{1}{S} \sum_{t=1}^{S} f(x_t) = E_{p(X)}[f(X)]$
- **Variance decreases like $1/S$**
- **The chain converges exponentially quickly** to the stationary distribution, so bias decreases quickly. (But in practice, we almost never know the rate!)
Some concerns:

- $X_1, X_2, \ldots$ are not true samples from $p$, especially early in the chain
- $X_1, X_2, \ldots, X_S$ are not independent
- How to create a Markov chain with $p$ as a stationary distribution?
- How to make sure that $p$ is the only stationary distribution?
- How long to run the chain?
- How to initialize the chain?
- What is the best Markov chain?

MCMC for Multivariate Distributions

- To sample from a multivariate distribution $p(x)$ for $x \in \mathbb{R}^D$, an MCMC algorithm generates a sequence of states $x_1, x_2, x_3, \ldots, x_S$
- Each $x_t = (x_{t1}, \ldots, x_{tD})$ is a full vector — with a setting for each variable
- The state space of the Markov chain is the full domain $x \in \text{Val}(X)$. E.g., with $D$ binary variables, the Markov chain has $2^D$ states.
- Because state spaces are huge, MCMC algorithms specify rules for random transitions between states without materializing the full transition matrix.

Example: Binary MRF

MRF: Two Binary-Valued
Random Variables

Markov Chain: One Random
Variable with Four States

Detailed Balance
The Burning Question

How to design a Markov chain with a stationary distribution $\pi(x)$?

We will first introduce detailed balance, a sufficient condition for $\pi(x)$ to be a stationary distribution of a Markov chain $T$.

Then we will design sampling algorithms (i.e., Markov chains) that, by construction

1. Are regular
2. Satisfy detailed balance with respect to $\pi(x)$

These together will imply that the chain converges to $\pi$, which is the unique stationary distribution.

Detailed Balance Interpretation

Detailed Balance

A Markov chain $T$ satisfies **detailed balance** with respect to a distribution $\pi$ if $\forall x, x'$,

$$\pi(x)T(x'|x) = \pi(x')T(x|x').$$
Detailed Balance $\implies$ Stationary

**Theorem:** If $T$ satisfies detailed balance with respect to $\pi$ then $\pi$ is a stationary distribution of $T$.

**Proof:** Let $\pi' = \pi T$ be the result of running the Markov chain for 1 iteration. Then

\[
\pi'(x') = \sum_x \pi(x)T(x'|x) \quad \text{(definition of $\pi' = \pi T$)}
\]
\[
= \sum_x \pi(x')T(x'|x') \quad \text{(detailed balance)}
\]
\[
= \pi(x') \sum_x T(x'|x') \quad \text{($\sum_x T(x'|x') = 1$)}
\]
\[
= \pi(x').
\]