COMPSCI 688: Probabilistic Graphical Models

Lecture 14: Markov Chain Monte Carlo

Dan Sheldon

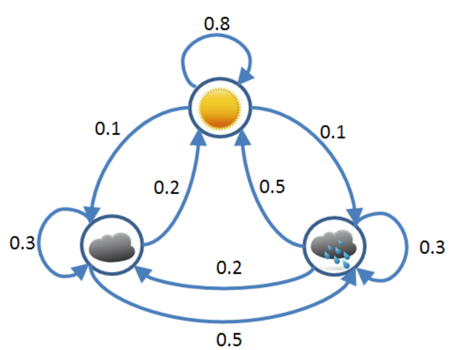
Manning College of Information and Computer Sciences
University of Massachusetts Amherst

Partially based on materials by Benjamin M. Marlin (marlin@cs.umass.edu) and Justin Domke (domke@cs.umass.edu)

Markov Chain Theory

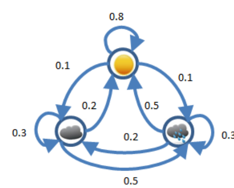
Markov Chains

A discrete Markov chain is a **set of states** with **transition probabilities** between each pair of states. **Example** (note: not a graphical model!)

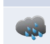


Transition Matrix

- The probabilistic transitions in the state diagram can also be represented by an equivalent matrix of transition probabilities.
- The "from" states are rows and the"to" states are columns.



To

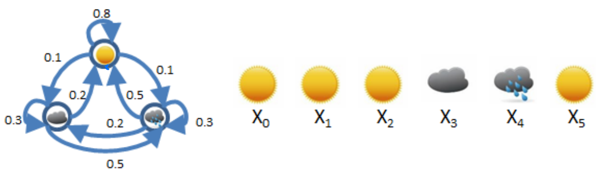
From			
	0.8	0.1	0.1
	0.2	0.3	0.5
	0.5	0.2	0.3

transition matrix $= T$

T_{ij} = prob transition from state i to state j

Markov Chains: Simulation and State Sequences

- To simulate a Markov chain, we draw $x_0 \sim p_0$, then repeatedly sample x_{t+1} given the current state x_t according to the transition probabilities T.



Markov Chain: Formal Definition

By repeatedly making random transitions from a starting state, we generate a *chain* of random variables $X_0, X_1, X_2, X_3, \dots$.

Formally, a Markov chain is specified by:

- A set of states $\{1, 2, \dots, D\}$
- A starting distribution p_0 with $p_0(i) = P(X_0 = i)$.
- Transition probabilities $T_{ij} = P(X_{t+1} = j \mid X_t = i)$ for all $i, j \in \{1, 2, \dots, D\}$

A Markov chain **assumes the Markov property**:

$$P(X_t = x_t \mid X_0 = x_0, X_1 = x_1, \dots, X_{t-1} = x_{t-1}) = P(X_t = x_t \mid X_{t-1} = x_{t-1})$$

Markov Chain Questions

Three important questions:

1. What is the joint probability of a sequence of states of length N?
2. What is the marginal probability distribution over states after a given number of steps t?
3. What happens to the probability distribution over states in the limit as t goes to infinity?

Markov Chain Factorization

Question: What is the joint probability over the state sequence $x_0, ..., x_N$?

Answer: by the Markov property:

$$P(X_1 = x_1, \dots, X_N = x_N | X_0 = x_0) = P(X_1 = x_1 | X_0 = x_0) \times P(X_2 = x_2 | X_1 = x_1) \times \cdots$$
$$\times P(X_N = x_N | X_{N-1} = x_{N-1})$$

Shorter version:

$= p(x_1|x_0)\,p(x_2|x_1,\cancel{x_0})\,p(x_3|x_2,\cancel{x_1},\cancel{x_0}) \cdots$

$$\begin{aligned} p(x_1, x_2, \dots, x_N | x_0) &= p(x_1|x_0)p(x_2|x_1)\dots p(x_N|x_{N-1}) \\ &= T_{x_0 x_1} \times T_{x_1 x_2} \times \cdots \times T_{x_{N-1} x_N} \end{aligned}$$

The t-Step Distribution for Fixed x_0 , $X_1, X_2, X_3, \ldots, X_t$

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:

$$p(x_1|x_0) = T_{x_0 x_1}$$

$$p(x_2|x_0) = \sum_{x_1} p(x_1, x_2|x_0) = \sum_{x_1} p(x_1|x_0)\, p(x_2|x_1)$$

In general, we have the recursive expression:

$$p(x_t|x_0) = \sum_{x_{t-1}} p(x_{t-1}, x_t|x_0) = \sum_{x_{t-1}} p(x_{t-1}|x_0)\, p(x_t|x_{t-1})$$

(assume know $p(x_{t-1}|x_0)$) $$= \sum_{x_{t-1}} p(x_{t-1}|x_0)\, T_{x_{t-1} x_t}$$

The t-Step Distribution for Random X_0

marginals $p(x_1), p(x_2), \ldots$

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:

$$p(x_1) = \sum_{x_0} p(x_0, x_1) = \sum_{x_0} p_0(x_0)\, T_{x_0 x_1}$$

$$p(x_2) = \sum_{x_1} p(x_1, x_2) = \sum_{x_1} p(x_1)\, T_{x_1 x_2}$$

In general:

$$p(x_t) = \sum_{x_{t-1}} p(x_{t-1})\, T_{x_{t-1} x_t}$$

recurrence for marginals $p(x_t)$

t-Step Recurrence as Matrix-Vector Multiplication

$p(x_t)$

The recurrences for the t-step distributions can be expressed using matrix-vector multiplciation. 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

$$p_t = p_{t-1} T.$$

$$\sum_{x_{t-1}} p(x_{t-1})\, T_{x_{t-1} x_t}$$

$$\underbrace{[P(X_t=1) \;\cdots\; P(X_t=x_t) \;\cdots\; P(X_t=D)]}_{p_t} \underset{=}{=} \underbrace{[P(X_{t-1}=1) \;\cdots\; P(X_{t-1}=D)]}_{p_{t-1}} \underbrace{\begin{array}{c} x_t \\ \begin{bmatrix} T_{1x_t} \\ T_{2x_t} \\ \vdots \\ T_{Dx_t} \end{bmatrix} \end{array}}_{T}$$

$$P(X_t = x_t) = \sum_{x_{t-1}} P(X_t = x_{t-1})\, T_{x_{t-1} x_t}$$

$$\boxed{p_t = p_{t-1} T}$$

t-Step Distribution as Matrix Power

$p_t = p_{t-1} T$

By unrolling the recurrence, the t-step distribution can be obtained as a matrix power

$$
\begin{aligned}
p_t &= p_{t-1} T \\
&= (p_{t-1}) T \\
&= (p_{t-2} T) T \\
&= (p_{t-2}) T T \\
&= (p_{t-3} T) T T \\
&\ \vdots \\
&= p_0 \underbrace{T T \ldots T}_{t \text{ times}}. \quad = p_0 T^t
\end{aligned}
$$

Markov Chain Theory | Understanding MCMC | Detailed Balance

Thus

$$\boxed{p_t = p_0 T^t.}$$

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

$$(T^t)_{ij} = P(X_t = j | X_0 = i) = P(X_{s+t} = j | X_s = i)$$

$\|$

prob transition from i to j in t steps

Markov Chain Theory | Understanding MCMC | Detailed Balance

One-Slide Summary So Far

- Markov chain: defined by initial distribution $p_0 \in \mathbb{R}^D$, transition matrix $T \in \mathbb{R}^{D \times D}$

$$p_0(i) = P(X_0 = i), \qquad T_{ij} = P(X_t = j \mid X_{t-1} = j)$$

- Defines distribution of chain $X_0, X_1, X_2, \ldots, X_t, \ldots$ (with Markov assumption)
- Joint probability

$T_{x_1 x_0}$ $T_{x_2 x_1}$ $\cdots$ $T_{x_{N-1} x_N}$

$$p(x_1, x_2, \ldots, x_N | x_0) = p(x_1|x_0) p(x_2|x_1) \cdots p(x_{N-1}|x_N)$$

- Recurrence for t-step distribution: $p(x_t) = \sum_{x_{t-1}} p(x_{t-1}) T_{x_{t-1} x_t}$
- Recurrence as matrix-vector multiplication. Let $p_t \in \mathbb{R}^D$ with $p_t(i) = P(X_t = i)$. Then

$$\boxed{p_t = p_{t-1} T}$$

- **Next**: what happens as $t \to \infty$?

Markov Chain Theory | Understanding MCMC | Detailed Balance

Limiting Distribution

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

$$\boxed{\lim_{t \to \infty} p_t = \pi} \qquad \text{(limiting distribution)}$$

The algorithmic idea of Markov chain Monte Carlo is:

- Suppose π 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 $\overset{\pi}{p_t} = \overset{\pi}{p_{t+1}} = \pi$. Since $p_{t+1} = p_t T$, this implies

$$\boxed{\pi = \pi T} \qquad \text{(stationary distribution)}$$

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

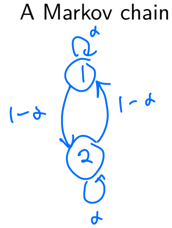
Stationary and Limiting Distributions

We reason about *limiting distributions* via *stationary distributions*:

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



Spse $X_0 = 1 \Rightarrow X_1 = 2,\ X_2 = 1,\ X_3 = 2, \ldots$

$p_0 = [1, 0],\ p_1 = [0, 1],\ p_2 = [1, 0],\ p_3 = [0, 1], \ldots$ (does not converge)

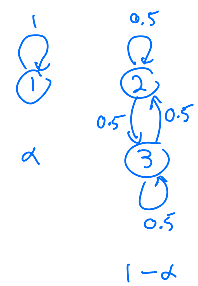
On the other hand, if $p_0 = [\frac{1}{2}, \frac{1}{2}]$

$p_0 = [\frac{1}{2}, \frac{1}{2}],\ p_1 = [\frac{1}{2}, \frac{1}{2}], \ldots$

$\mathrm{diag}(T^5) > 0 \Rightarrow$ not periodic

What Could Go Wrong: Reducibility

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



$\pi_\alpha = [\alpha, \frac{1-\alpha}{2}, \frac{1-\alpha}{2}]$ is stationary

for every $\alpha \in [0, 1]$

$t = 5,\quad T^5_{ij} > 0$

Regularity

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

$$Pr(X_{t+s}=j \mid X_s=i) = (T^t)_{ij} > 0,$$

- 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 π and $\lim_{t\to\infty} p_t = \pi$ for all starting distributions p_0.

(We can sample from the unique stationary distribution π 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 π such that $\pi T = \pi$. If you start from π 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.

  T^{10}
- **Convergence to stationary distribution**: if T is regular, the chain converges to a unique stationary distribution π for any starting distribution.

Understanding MCMC

High-Level Idea

Know: $p(x) \cdot C$

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

$$\mathbb{E}_{p(X)}[f(X)] \approx \frac{1}{S}\sum_{t=1}^{S} f(x_t).$$

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

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

could be bias $\approx .9999^S$

$p_1 \quad p_2 \quad p_3 \quad p_4 \qquad \longrightarrow \qquad p$

$X_1, X_2, X_3, X_4, \ldots\ldots$

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 distibution?
- How long to run the chain ?
- How to initialize the chain?
- What is the best Markov chain?

practical

MCMC for Multivariate Distributions

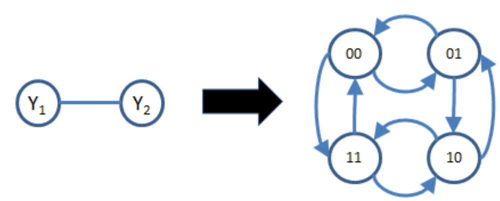
$x = (x_1, \ldots, x_D)$

- To sample from a multivariate distribution $p(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^D$, an MCMC algorithm generates a sequence of *states*

$$\overset{000}{\mathbf{x}_1}, \overset{101}{\mathbf{x}_2}, \overset{111}{\mathbf{x}_3}, \ldots, \mathbf{x}_S$$

- Each $\mathbf{x}_t = (x_{t1}, \ldots, x_{tD})$ is a full vector — with a setting for each variable
- The state space of the Markov chain is the full domain $\mathbf{x} \in \mathrm{Val}(\mathbf{X})$. E.g., with D binary variables, the Markov chain has 2^D states. $\Rightarrow T \in \mathbb{R}^{2^d \times 2^d}$
- 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 $\pi(\mathbf{x})$? given

We will first introduce **detailed balance**, a sufficient condition for $\pi(\mathbf{x})$ 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 $\pi(\mathbf{x})$

These together will imply that the chain converges to π, which is the unique stationary distribution

Detailed Balance

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

p_t

$$\pi(x)T(x'|x) = \pi(x')T(x|x').$$

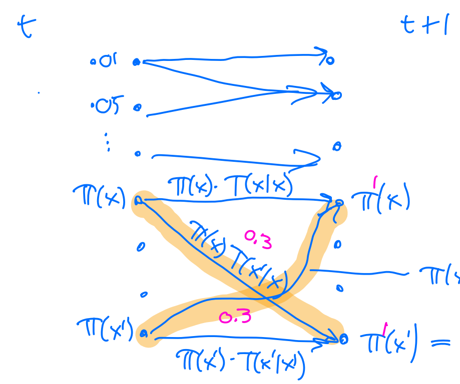
Suppose $X_t \sim \pi$ "flow $x \to x'$"

LHS $= Pr(X_t = x) \cdot Pr(X_{t+1} = x' \mid X_t = x) = Pr(X_t = x, X_{t+1} = x')$

RHS $= Pr(X_t = x', X_{t+1} = x)$

"flow $x' \to x$"

Detailed Balance Interpretation



$\pi'(x') = \sum_x \underbrace{\pi(x)T(x'|x)}_{\text{flow}}$

Assume $p_t = \pi$ Then $p_{t+1} = \pi'$

Detailed Balance $\Longrightarrow$ Stationary

Theorem: If T satisfies detailed balance with respect to π then π is a stationary distribution of T.

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

$$\pi'(x') = \sum_x \pi(x) \cdot T(x'|x)$$

detailed balance

$$= \sum_x \pi(x') \cdot T(x|x')$$

$$= \pi(x') \sum_x T(x|x')$$

$$= \pi(x') \qquad \square$$