COMPSCI 688: Probabilistic Graphical Models

Lecture 13: Introduction to 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)

A Quiz Question Monte Carlo Methods Gibbs Sampling Markov Chain Theory

A Quiz Question

A Quiz Question Monte Carlo Methods Gibbs Sampling Markov Chain Theory

A Quiz Question

Consider an exponential family on $x_1, x_2 \in \{0, 1\}$ with $T(x_1, x_2) = \mathbb{I}[x_1 = 1, x_2 = 1]$.
Suppose you use the data below to estimate maximum likelihood parameters:

x_1	x_2
1	1
1	0
1	1
0	1

data exp. = model exp.

$\hat{\mathbb{E}}[T(x)] = \mathbb{E}_{p_{\theta^*}}[T(x)]$

$\hat{\mathbb{E}}[\mathbb{I}[x_1=1, x_2=1]] = \mathbb{E}_{p_{\theta^*}}[\mathbb{I}[X_1=1, X_2=1]]$

$= 1/2$

At the maximum likelihood estimate θ^*, what will be $P_{\theta^*}(X_1 = 1, X_2 = 1)$? $1/2$

A Quiz Question Monte Carlo Methods Gibbs Sampling Markov Chain Theory

Monte Carlo Methods

Motivation

Computing expectations is important!

$$\mathbb{E}_{p(x)}[f(X)] = \int p(x)f(x)dx$$

Example: suppose $p(\mathbf{x})$ is an MRF, then

$$P(X_u = a, X_v = b) = \mathbb{E}_{p(\mathbf{x})}\left[\mathbb{I}[X_u = a, X_v = b]\right]$$

In general, computing expectations is hard, so we need an approximation.

Monte Carlo methods

In a Monte Carlo method, we approximate an expected value by a sample average. Draw N samples $X_1, \ldots, X_N \sim p(x)$, then

$$\mathbb{E}_{p(x)}[f(X)] \approx \frac{1}{N}\sum_{n=1}^{N} f(X_n).$$

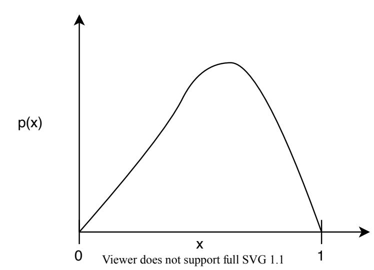
Nice properties:

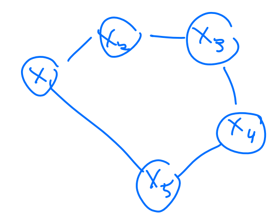
- Unbiased
- Variance decreases like $\frac{1}{N}$.
- Measure arbitrary properties by choosing f.

Not nice properties: **sampling is algorithmically/computationally hard** in general

Examples

Suppose we have $p(x) = 12(x^2 - x^3)$, where $x \in [0, 1]$. Or suppose we have an MRF with a cycle.





Question: How do we sample from these distributions? A: some algorithm

Gibbs Sampling

Markov Chain Monte Carlo Overview

- Markov chain Monte Carlo (MCMC) methods *iteratively* construct samples from a given "target distribution" $p(\mathbf{x})$
- They require only access to the *unnormalized* distribution, so can apply easily to models like MRFs.
- Formally, they work by constructing a *Markov chain* that has the target distribution $p(\mathbf{x})$ as its limiting distribution.
- We'll introduce one MCMC method today, and then start to develop some of the theory needed to understand the algorithm.
- Importance / applications: statistical physics, econometrics, ecology, epidemiology, weather modeling, ...

The Gibbs Sampler

A simple and powerful algorithm! Assume $\mathbf{X} = (X_1, \ldots, X_D)$.

Initialize all variables arbitrarily, then repeatedly update each variable by sampling from its conditional distribution given all other variables.

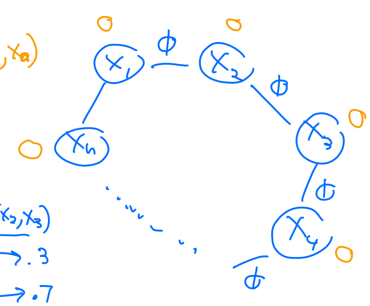
Gibbs sampler

- Initialize $x_1, \ldots, x_D$
- Repeat
  - For $i = 1$ to D, resample $x_i \sim p(X_i \,|\, \mathbf{X}_{-i} = \mathbf{x}_{-i})$
  - Record $\mathbf{x} = (x_1, \ldots, x_D)$ as one sample

One sample is generated after each loop through all of the variables.

Example: Cycle MRF

Suppose $p(\mathbf{x}) = \prod_{i=1}^{n} \phi(x_i, x_{i+1})$ (mod n)



Then $p(x_i|\mathbf{x}_{-i}) \propto \phi(x_{i-1}, x_i)\phi(x_i, x_{i+1})$ (factor reduction!)

For a general MRF: $p(x_i|\mathbf{x}_{-i}) \propto \prod_{c: i \in c} \phi_c(x_i, \mathbf{x}_{c \setminus i})$

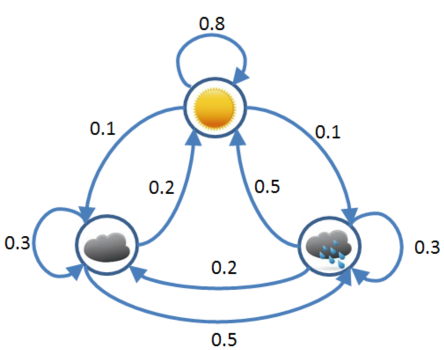
The Gibbs Sampler: Properties

- The Gibbs sampler eventually draws samples from the target distribution $p(\mathbf{x})$ regardless of how it is initialized.
- It can take time to converge to the target distribution $p(\mathbf{x})$. This phase of the algorithm is referred to as the "burn-in" phase of the algorithm.
- Convergence to the target distribution needs to be tested empirically in most cases using convergence diagnostics.
- Even after convergence, the samples **are not independent**, but can still be used in Monte Carlo averages. The degree of correlation of the samples affects the rate of convergence of Monte Carlo averages.

Markov Chain Theory

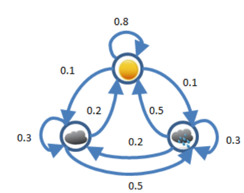
Markov Chains

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.



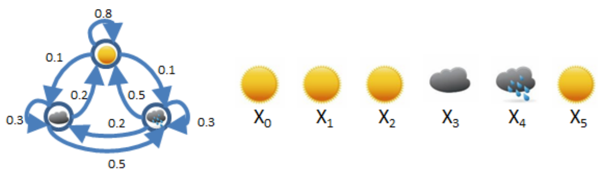
To

From

			
	0.8	0.1	0.1
	0.2	0.3	0.5
	0.5	0.2	0.3

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 $T_{ij} = P(X_{t+1} = j \,|\, X_t = i)$ for all $i, j \in \{1, 2, \dots, D\}$

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

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

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 $x_0, \dots, x_N$?

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:

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

The t-Step Distribution for Fixed x_0

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:

$$p(x_1|x_0) =$$

$$p(x_2|x_0) =$$

In general, we have the recursive expression:

$$p(x_t|x_0) =$$

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:

$$p(x_1) =$$

$$p(x_2) =$$

In general:

$$p(x_t) =$$

t-Step Recurrence as Matrix-Vector Multiplication

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

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

t-Step Distribution as Matrix Power

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

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

Thus

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

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

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