HW3: due next ~~Wed~~ → Fri

Quiz 6: Friday

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

A Quiz Question

$p_\theta(x) = \exp(\theta\, \mathbb{I}[x_1=1, x_2=1] - A(\theta))$

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	$\mathbb{I}[x_1=1, x_2=1]$
1	1	1
1	0	0
1	1	1
0	1	0

avg: $\frac{1}{2}$

"data exp" = "model exp"

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

$$\underbrace{\hat{\mathbb{E}}[\mathbb{I}[x_1=1, x_2=1]]}_{= 1/2} = \mathbb{E}_{p_{\theta^*}}[\mathbb{I}[x_1=1, x_2=1]]$$

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

X r.v.

$$\mathbb{E}[\mathbb{I}[X \in A]] = \int p(x) \cdot \mathbb{I}[x \in A]\, dx$$
$$= \int_A p(x)\, dx$$
$$= \Pr(X \in A)$$

Application Example

Covid Model

Showed Covid modeling example w/ NumPyro. See Jupyter notebook

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 $X_1, \dots, X_N \sim p(x)$, then

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

$f(X) = X$
$f(X) = \mathbb{I}[X \in [a,b]]$

Nice properties:

- Unbiased
- Variance decreases like $\frac{1}{N}$. ⇐ "error" $\sim \frac{1}{\sqrt{N}}$
- Measure arbitrary properties by choosing f.

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

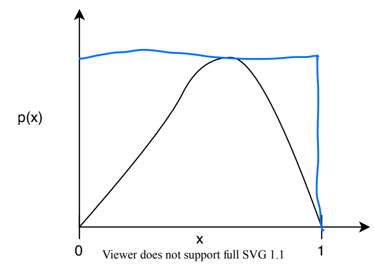
Given $p(x)$, may be hard to sample $x \sim p(x)$

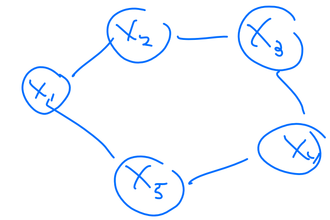
Examples

- inverse transform sampling
- rejection sampling

} low dimensional

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: an algorithm

Gibbs Sampling

Markov Chain Monte Carlo Overview

Input: $p(x)$ – density – pmf (could be unnormalised)

- 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

Input: $p(x)$ $\in \mathbb{R}^d$

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

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

Gibbs sampler

14.7 . . . . . 4.7

-13 3.5

0 0

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

$p(x_1 | x_2=0, x_3=0, \dots, x_D=0)$

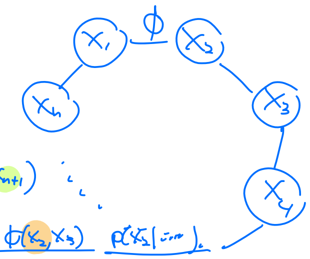
$p(x_2 | x_1=-13, x_3=0, \dots, x_D=0)$

Subproblem usually "easy" because 1-D dist

One approximate 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)



Update x_2

free fixed

$p(x_2 | x_1, x_3, x_4, \dots, x_n)$

$\propto \frac{1}{Z} \phi(x_1, x_2)\phi(x_2, x_3)\phi(x_3, x_4) \cdots \phi(x_n, x_{n+1})$

$\propto \phi(x_1, x_2)\phi(x_2, x_3)$

x_2	$\phi(x_1,x_2)\phi(x_2,x_3)$	$p(x_2 \mid \dots)$
0	3	.3
1	7	.7

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

$\frac{1}{N} \sum_{n=1}^{N} f(x^{(n)})$

- 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.