COMPSCI 688: Probabilistic Graphical Models

Lecture 16: MCMC Practical Aspects and Bayesian Inference Intro

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)

MCMC Practical Aspects

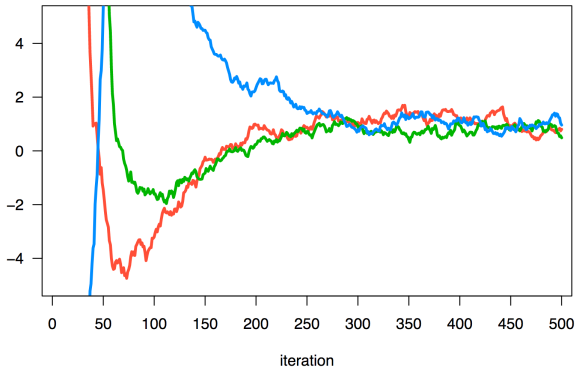
Issues with MCMC

- **Burn-in:** The underlying Markov chains take time to converge to the distribution of interest. The time needed to reach the stationary distribution of the chain is called the *burn-in time*.
- **Autocorrelation:** Consecutive samples drawn from the chain at equilibrium may be highly correlated with each other. The time lag between samples that are approximately independent of each other is called the *autocorrelation time* of the chain.

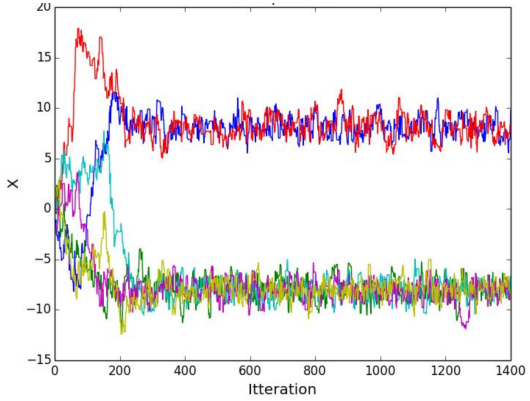
Burn-in Time

- The most fundamental issue with burn-in is that, in the absence of a theoretical lower bound, you can never be exactly sure that the chain has converged to the equilibrium distribution.
- MCMC practitioners usually rely on heuristic convergence diagnostics to assess burn-in time.
- One of the most useful heuristics is to run multiple chains from different starting points and track one or more scalar functions of the state of the chain (the log probability of the data is often a good choice).
- The distribution of values of these functions will all converge to the same mean and variance at equilibrium.

Example: Burn-in Time



Example: Burn-in Time



Autocorrelation Time

- At or near equilibrium, different samplers can traverse the state space at different rates.
- The autocorrelation time of a sampler is the number of sampling iterations we must apply at equilibrium to obtain two samples that are approximately independent.
- Practitioners sometimes use estimates of autocorrelation at different lags to estimate an "effective sample size" of S MCMC samples

Practical Aspects Summary

- There are many different diagnostics. Useful to learn some of them
- It's often easy to diagnose clear failures
- It's basically impossible to diagnose success
- Some practitioners advocate just running one chain for a very long time

Bayesian Inference

Bayesian Inference Example

Suppose we observe data $x^{(1)}, \ldots, x^{(n)}$ which we assume to come from a Bernoulli model

$$p(x^{(n)}|\theta) = \begin{cases} \theta & x^{(n)} = 1 \\ 1-\theta & x^{(n)} = 0 \end{cases}$$

- Maximum-likelihood says to find θ by solving $\max_\theta \frac{1}{n} \sum_{n=1}^{N} \log p(x^{(n)}|\theta)$

When might we want something different?

Example: you go on a three-day trip to Australia and want to learn about the weather

$$X = \begin{cases} 1 & \text{rain} \\ 0 & \text{no rain} \end{cases}$$

Observe $x^{(1)} = 1, x^{(2)} = 1, x^{(3)} = 1$

MLE learning $\rightarrow \hat{\theta} = 1$

It rains every day in Australia. What went wrong?

Being Bayesian

A Bayesian says: give me the **probability** of θ given the data. What does this mean?

$$p(\theta|\text{Data}) = \frac{p(\theta)p(\text{Data}|\theta)}{p(\text{Data})}$$

- $p(\theta)$ is the **prior**. It encodes beliefs (either subjective or objective) about θ **prior** to seeing any evidence. We need one!
- $p(\text{Data}|\theta) = \prod_{n=1}^{N} p(x^{(n)}|\theta)$ is the **likelihood**. It incorporates evidence.
- $p(\text{Data}) = \int p(\theta)p(\text{Data}|\theta)d\theta$ is the **marginal likelihood** or **evidence**. We usually don't need to compute it.
- $p(\theta|\text{Data})$ is the **posterior**. What we believe about θ after observing data.

Why Be Bayesian?

- Philosophy: Update subjective prior beliefs based on evidence.
- Practical: deal with small samples
- Practical: excellent tools exist (MCMC, stan)

Making our Model Bayesian

$$\theta \sim \text{Uniform}([0,1])$$
$$x^{(n)} \sim \text{Bernoulli}(\theta)$$

Bayesian Modeling: Implications

- We now have a **joint probability model** $p(\theta, x)$

$$p(\theta, x) = p(\theta)p(x|\theta)$$

- θ is now a **random variable** instead of a fixed but unknown parameter
- Learning is replaced by **posterior inference**
  - Learning: $\max_\theta \mathcal{L}(\theta|x^{(1)}, \ldots, x^{(N)})$
  - Posterior inference: compute $p(\theta|x^{(1)}, \ldots, x^{(N)})$

Posterior Inference

$$\begin{aligned} p(\theta|x^{(1:N)}) &= \frac{p(\theta)p(x^{(1:N)}|\theta)}{p(x^{(1:N)})} \\ &\propto p(\theta)p(x^{(1:N)}|\theta) \\ &= \prod_{n=1}^{N} \theta^{\mathbb{I}[x^{(n)}=1]}(1-\theta)^{\mathbb{I}[x^{(n)}=0]} \\ &= \theta^{\#(X=1)}(1-\theta)^{\#(X=0)} \end{aligned}$$

E.g., use MCMC to sample from density on $[0, 1]$ proportional to this

General inference strategy: use MCMC to sample from density proportional to $p(\theta)p(\text{Data}\,|\,\theta)$

But in some *special* cases the problem is easy to solve without MCMC... (next time)

A Little History...

- Bayes is the OG statistics! (Bayes, Laplace ~late 1700s)
  - But suspicious about priors ("objective Bayes"), often used flat ones
- Early 1900s: frequentist stats, e.g., MLE (Fisher, Neyman)
  - But inference requires imagining repeated data from same distribution
- Subjective Bayes (Savage, de Finetti)
  - Prior encodes subjective personal beliefs
- Mid 1900s: objective Bayes (Jeffreys)
  - Pragmatic view, combines frequentist ideas
  - E.g., does posterior converge to truth as we get more data?
- 2000s: Computational Bayes?

More Motivation: de Finetti

Exchangeability: A sequence of random variables $x_1, x_2, x_3, \ldots$ is exchangeable if for any n and any permutation π

$$p(x_1, x_2, \ldots, x_n) = p(x_{\pi_1}, x_{\pi_2}, \ldots, x_{\pi_n})$$

(the joint probability is invariant to any permutation of the indices)

Examples:

- $x_1, x_2, x_3, \ldots$ where x_i are iid (sequence is iid)
- $x_0 + x_1, x_0 + x_2, x_0 + x_3, \ldots$ (sequence is exchangeable but not iid)

de Finetti Theorem (1930s): A sequence of random variables $x_1, x_2, \ldots$ is exchangeable if and only if for all n

$$p(x_1, x_2, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) p(\theta) d\theta$$

for some $p(\theta)$. (Actually, θ doesn't need to have a density.)

Implications: if exchangeable,

- There must exist a parameter θ
- There must exist a likelihood $p(x|\theta)$
- There must exist a distribution $p(\theta)$
- The data is conditionally independent given θ