COMP 330: Optimization–Markov Chain Monte Carlo

Chris Jermaine and Kia Teymourian

Rice University

Bayesian: Review

MCMC closely linked with Bayesian ML...

A Bayesian uses data X to update the prior on the parameter set Θ

- Resulting distribution—$P(\Theta|X)$ is called the "posterior"

Update is accomplished via "Bayes' Rule"

$$P(\Theta|X) = \frac{P(\Theta)P(X|\Theta)}{P(X)}$$

Can usually drop $P(X)$ as a constant, so we have

$$P(\Theta|X) \propto P(\Theta)P(X|\Theta)$$

Bayesian Can Be Easy

Recall that sometimes, things are easy!

- Conjugate prior
- Scores so far: $\{99, 92, 94, 94, 88\}$
- Mean score $\mu \sim \textbf{Normal}(50, 5)$
- Each score $x_i \sim \textbf{Normal}(\mu, 4)$
- Look up in table, that $P(\mu|\textbf{scores})$ is Normal, mean $\left(\frac{50}{25} + \frac{467}{16}\right) / \left(\frac{1}{25} + \frac{5}{16}\right)$, var $\left(\frac{1}{25} + \frac{5}{16}\right)^{-1}$

Sometimes, Hard

Back to our coin toss example

- Have a bag with two coins
- First has prob $p_1 \sim \textbf{Beta}(1, 1)$ of heads
- Second has prob $p_2 \sim \textbf{Beta}(1, 1)$ of heads
- I repeatedly reach in, pull out a coin
- Identity is $z_i \in \{1, 2\}$
- Flip it 10 times and observe x_i heads
- How to compute posterior $P(\{p_1, p_2\}|X)$

Conjugate prior not enough here!

MCMC

Markov Chain Monte Carlo

- Family of alogirithms
- Idea: define a Markov Chain
- Random walk on graph of states
- Each state has a label
- So that "stationary distribution" of chain is target $P(.)$
- Simulate the chain, use labels generated as samples from $P(.)$

Amazing, like magic!

Detailed Balance

Most MCMC algorithms based on this idea

Theorem: Let $P(\theta \leftarrow \theta')$ be the prob of transitioning from state θ to state θ'

Then $P(\theta)$ is the stationary distribution of the chain if:

$$P(\theta)P(\theta \leftarrow \theta') = P(\theta')P(\theta' \leftarrow \theta)$$

Leads Directly to Metropolis Hastings Algorithm

Say we want to sample $\theta \sim P(\theta)$

That's hard, but we have a distribution $Q(\theta_{\text{next}}|\theta)$ that's easy

```
initialize θ;
for i = 1 to big do:
  generate θ_next ~ Q(θ_next|θ);
  compute a = P(θ_next)Q(θ|θ_next) / P(θ)Q(θ_next|θ);
  if (a ≥ 1)
    θ ← θ_next;
  else
    flip coin where prob of heads is a;
    if heads
      θ ← θ_next;
```

Example Metropolis Hastings

We want to generate samples from high dimensional Normal(μ, Σ)

- μ is a vector
- Σ a matrix
- Not trivial

Then define random transition Q as:

- pick a random dimension d
- add a Normal$(0, 1)$ to the dth dimension of the current answer
- This gives you θ_{next}
- Then $Q(\theta_{\text{next}}|\theta) = Q(\theta_{\text{next}}|\theta) = \textbf{Normal}(\textbf{proposed change in dim } d|0, 1)$

That's it!!

Why Does This Work?

Start with detailed balance:

$$P(\theta)P(\theta \leftarrow \theta') = P(\theta')P(\theta' \leftarrow \theta)$$

Why Does This Work?

Start with detailed balance:

$$P(\theta)P(\theta \leftarrow \theta') = P(\theta')P(\theta' \leftarrow \theta)$$

- So for detailed balance to hold:

$$\frac{P(\theta \leftarrow \theta')}{P(\theta' \leftarrow \theta)} = \frac{P(\theta')}{P(\theta)}$$

- Use $Q(\theta'|\theta)P(\texttt{heads}|\theta \leftarrow \theta')$ for $P(\theta \leftarrow \theta')$. Then:

$$\frac{Q(\theta'|\theta)P(\texttt{heads}|\theta \leftarrow \theta')}{Q(\theta|\theta')P(\texttt{heads}|\theta' \leftarrow \theta)} = \frac{P(\theta')}{P(\theta)}$$

- So:

$$\frac{P(\texttt{heads}|\theta \leftarrow \theta')}{P(\texttt{heads}|\theta' \leftarrow \theta)} = \frac{P(\theta')Q(\theta|\theta')}{P(\theta)Q(\theta'|\theta)}$$

Gibbs Sampling

Sometimes, designing proposal not so easy

- Getting number of rejections right is important
- Too many: what happens?
- Too few: what happens?

Gibbs sampling is a flavor of MH with no rejections

Gibbs Sampling

Basic idea

- Say I want to sample from $P(\theta)$ where $\theta = \langle \theta_1, \theta_2, ... \rangle$
- High dimensionality is the problem
- Coin flip example: $\theta = \langle p_1, p_2, z_1, z_2, \ldots, z_n \rangle$

Gibbs Sampling

Basic idea

- Say I want to sample from $P(\theta)$ where $\theta = \langle \theta_1, \theta_2, ... \rangle$
- High dimensionality is the problem
- Coin flip example: $\theta = \langle p_1, p_2, z_1, z_2, \ldots, z_n \rangle$

initialize each of the m elements of θ
for i = 1 to big **do:**
  for j = 1 to m **do:**
    $\theta_m \sim P(\theta_m | \theta_1, \theta_2, ..., \theta_{m-1}, \theta_{m+1}, \theta_{m+2}, ...)$

That's it!

- Requires only that we can sample from conditional distribution for θ_m

Gibbs Sampling Example

We need:

- $P(p_1|x_1, x_2, \ldots, x_n, p_2, z_1, z_2, \ldots, z_n)$ (p_2 is the same)
- $P(z_1|x_1, x_2, \ldots, x_n, p_1, p_2, z_1, z_2, \ldots, z_n)$ (all other z_i's are the same)

Gibbs Sampling Example

$$P(p_1|x_1, x_2, ..., x_n, p_2, z_1, z_2, ..., z_n)$$

- p_2 is not relevant. Why?
- Nor is any x_i where z_i is not 1. Why?
- So, using Bayes' rule, we have:

$$P(p_1|x_1, x_2, ..., x_n, p_2, z_1, z_2, ..., z_n) = \frac{\text{Beta}(p_1|1, 1) \prod_{i \text{ s.t. } z_i=1} \text{Binomial}(x_i|p_1, 10)}{P(x_1, x_2, ..., x_n, p_2, z_1, z_2, ..., z_n)}$$

- Dropping the denom, we have simply:

$$P(p_1|x_1, x_2, ..., x_n, p_2, z_1, z_2, ..., z_n) \propto \text{Beta}(p_1|1, 1) \prod_{i \text{ s.t. } z_i=1} \text{Binomial}(x_i|p_1, 10)$$

Gibbs Sampling Example

$$P(p_1|x_1, x_2, ..., x_n, p_2, z_1, z_2, ..., z_n)$$

- p_2 is not relevant. Why?
- Nor is any x_i where z_i is not 1. Why?
- So, using Bayes' rule, we have:

$$P(p_1|x_1, x_2, ..., x_n, p_2, z_1, z_2, ..., z_n) = \frac{\text{Beta}(p_1|1, 1) \prod_{i \text{ s.t. } z_i=1} \text{Binomial}(x_i|p_1, 10)}{P(x_1, x_2, ..., x_n, p_2, z_1, z_2, ..., z_n)}$$

- Dropping the denom, we have simply:

$$P(p_1|x_1, x_2, ..., x_n, p_2, z_1, z_2, ..., z_n) \propto \text{Beta}(p_1|1, 1) \prod_{i \text{ s.t. } z_i=1} \text{Binomial}(x_i|p_1, 10)$$

Turns out Beta is conjugate for Binomial!

- From table, we have

$$P(p_1|x_1, x_2, ..., x_n, p_2, z_1, z_2, ..., z_n) = \text{Beta}(1 + \sum_{i \text{ s.t. } z_i=1} x_i, 1 + \sum_{i \text{ s.t. } z_i=1} 10 - x_i)$$

Gibbs Sampling Example

$$P(z_1|x_1, x_2, ..., x_n, p_1, p_2, z_1, z_2, ..., z_n)$$

- This one is quite easy!
- Only care about x_1, p_1, p_2
- So, using Bayes' rule, we have:

$$P(z_1|x_1, x_2, ..., x_n, p_1, p_2, z_1, z_2, ..., z_n) = \frac{\text{Binomial}(x_1|p_1, 10)}{P(x_1, x_2, ..., x_n, p_1, p_2, z_2, ..., z_n)}$$

- Dropping everything but x_1 from the denom, we have:

$$P(z_1|x_1, x_2, ..., x_n, p_1, p_2, z_2, ..., z_n) = \frac{\text{Binomial}(x_1|p_1, 10)}{P(x_1)}$$

- This is:

$$\frac{\text{Binomial}(x_1|p_1, 10)}{0.5 \times \text{Binomial}(x_1|p_1, 10) + 0.5 \times \text{Binomial}(x_1|p_2, 10)}$$

That's it! We have our Gibbs sampler!

Questions?