Monte Carlo Markov Chain 2

MCMC

Problems with Gibbs Sampling:

- What if $P(X_i|\boldsymbol{x}_{-i})$ is not easy to sample from eg. in some continuous models?
- Gibbs chain involves changing one variable at a time.
- What if you need larger steps in the state space?

MCMC

- A finite-state Markov chain $\mathcal{T}$ is reversible if there exists a unique distribution π such that, for all $\boldsymbol{x}$, $\boldsymbol{x}' \in Val(\boldsymbol{X})$:

$$\pi(\boldsymbol{x})\mathcal{T}(\boldsymbol{x} \rightarrow \boldsymbol{x}') = \pi(\boldsymbol{x}')\mathcal{T}(\boldsymbol{x}' \rightarrow \boldsymbol{x})$$

- This equation is called the detailed balance
- Proposition 12.3: If $\mathcal{T}$ is regular and it satisfies the detailed balance equation relative to π, then π is the unique stationary distribution of $\mathcal{T}$

MCMC

Metropolis-Hastings Algorithm

- General construction that lets us build a reversible Markov chain with a particular stationary distribution
- Can't sample directly from target distribution for next state
- Uses a proposal distribution to generate next-state sample
- Corrects for proposal distribution by choosing to accept the proposed transition with some probability

MCMC

- Proposal distribution $\mathcal{T}^Q$:
  - transition model from state ***x*** to ***x'***
  - accept and transition to ***x'*** or stay at ***x***
- Acceptance probability $\mathcal{A}(\boldsymbol{x} \rightarrow \boldsymbol{x'})$
- The actual transition model is:

$$\mathcal{T}(\boldsymbol{x} \rightarrow \boldsymbol{x}') = \mathcal{T}^Q(\boldsymbol{x} \rightarrow \boldsymbol{x}')\mathcal{A}(\boldsymbol{x} \rightarrow \boldsymbol{x}') \quad \text{when } \boldsymbol{x} \neq \boldsymbol{x}'$$
$$\mathcal{T}(\boldsymbol{x} \rightarrow \boldsymbol{x}) = \mathcal{T}^Q(\boldsymbol{x} \rightarrow \boldsymbol{x}) + \sum_{\boldsymbol{x}' \neq \boldsymbol{x}} \mathcal{T}^Q(\boldsymbol{x} \rightarrow \boldsymbol{x}')(1 - \mathcal{A}(\boldsymbol{x} \rightarrow \boldsymbol{x}'))$$

5

5

MCMC

- Choice of proposal distribution can be arbitrary as long as it induces a regular chain
- A simple choice in discrete factored state spaces is to use a transition model $\mathcal{T}_i^Q$ which is uniform distribution over the values of X_i

6

6

MCMC

Given a transition model:

$$\mathcal{T}(\boldsymbol{x} \rightarrow \boldsymbol{x}') = \mathcal{T}^Q(\boldsymbol{x} \rightarrow \boldsymbol{x}')\mathcal{A}(\boldsymbol{x} \rightarrow \boldsymbol{x}') \quad \text{when } \boldsymbol{x} \neq \boldsymbol{x}'$$
$$\mathcal{T}(\boldsymbol{x} \rightarrow \boldsymbol{x}) = \mathcal{T}^Q(\boldsymbol{x} \rightarrow \boldsymbol{x}) + \sum_{\boldsymbol{x}' \neq \boldsymbol{x}} \mathcal{T}^Q(\boldsymbol{x} \rightarrow \boldsymbol{x}')(1 - \mathcal{A}(\boldsymbol{x} \rightarrow \boldsymbol{x}'))$$

The detailed balance equations assert that for all x ≠ x'

$$\pi(\boldsymbol{x})\mathcal{T}^Q(\boldsymbol{x} \rightarrow \boldsymbol{x}')\mathcal{A}(\boldsymbol{x} \rightarrow \boldsymbol{x}') = \pi(\boldsymbol{x}')\mathcal{T}^Q(\boldsymbol{x}' \rightarrow \boldsymbol{x})\mathcal{A}(\boldsymbol{x}' \rightarrow \boldsymbol{x})$$

And the acceptance probabilities satisfy:

$$\mathcal{A}(\boldsymbol{x} \rightarrow \boldsymbol{x}') = \min\left[1, \frac{\pi(\boldsymbol{x}')\mathcal{T}^Q(\boldsymbol{x}' \rightarrow \boldsymbol{x})}{\pi(\boldsymbol{x})\mathcal{T}^Q(\boldsymbol{x} \rightarrow \boldsymbol{x}')}\right]$$

7

7

MCMC

Let $\mathcal{T}^Q$ be any proposal distribution, and consider the Markov chain defined by the transition model (on previous slide) and acceptance probability (on previous slide).

If this Markov chain is regular, then it has the stationary distribution π

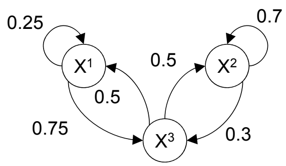
8

8

MCMC

Example of Metropolis-Hastings

Proposal distribution



Use the stationary distribution:

$\pi'(x^1) = 0.2$

$\pi'(x^2) = 0.5$

$\pi'(x^3) = 0.3$

Example of acceptance probabilities

$$A(x^1 \rightarrow x^3) = \min\left[1, \frac{\pi'(x^3)\mathrm{T}^Q(x^3 \rightarrow x^1)}{\pi'(x^1)\mathrm{T}^Q(x^1 \rightarrow x^3)}\right] = \min\left[1, \frac{(0.3)(0.5)}{(0.2)(0.75)}\right] = 1$$

$$A(x^3 \rightarrow x^1) = \min\left[1, \frac{\pi'(x^1)\mathrm{T}^Q(x^1 \rightarrow x^3)}{\pi'(x^3)\mathrm{T}^Q(x^3 \rightarrow x^1)}\right] = \min\left[1, \frac{(0.2)(0.75)}{(0.3)(0.5)}\right] = 1$$

MCMC

MCMC for graphical models

- Each local transition model $\mathcal{T}_i$ is defined via an associated proposal distribution $\mathcal{T}_i^{Q_i}$
- The acceptance probability for this chain is:

$$A(\boldsymbol{x}_{-i}, x_i \rightarrow \boldsymbol{x}_{-i}, x_i')$$
$$= min\left[1, \frac{\pi(\boldsymbol{x}_{-i}, x_i')T_i^{Q_i}(\boldsymbol{x}_{-i}, x_i' \rightarrow \boldsymbol{x}_{-i}, x_i)}{\pi(\boldsymbol{x}_{-i}, x_i)T_i^{Q_i}(\boldsymbol{x}_{-i}, x_i \rightarrow \boldsymbol{x}_{-i}, x_i')}\right]$$
$$= min\left[1, \frac{P_\Phi(\boldsymbol{x}_{-i}, x_i')T_i^{Q_i}(\boldsymbol{x}_{-i}, x_i' \rightarrow \boldsymbol{x}_{-i}, x_i)}{P_\Phi(\boldsymbol{x}_{-i}, x_i)T_i^{Q_i}(\boldsymbol{x}_{-i}, x_i \rightarrow \boldsymbol{x}_{-i}, x_i')}\right]$$

MCMC

Note that for graphical models:

$$\frac{P_\Phi(x_i', \boldsymbol{x}_{-i})}{P_\Phi(x_i, \boldsymbol{x}_{-i})} = \frac{P_\Phi(x_i' \mid \boldsymbol{x}_{-i})P_\Phi(\boldsymbol{x}_{-i})}{P_\Phi(x_i \mid \boldsymbol{x}_{-i})P_\Phi(\boldsymbol{x}_{-i})} = \frac{P_\Phi(x_i' \mid \boldsymbol{x}_{-i})}{P_\Phi(x_i \mid \boldsymbol{x}_{-i})}$$

In the case of Gibbs sampling (which is a special case of Metropolis-Hastings):

Define $\boldsymbol{U}_i$ = MarkovBlanket(X_i) and $\boldsymbol{u}_i = (\boldsymbol{x}_{-i})\langle \boldsymbol{Y}_i \rangle$

$$\frac{P_\Phi(x_i' \mid \boldsymbol{x}_{-i})}{P_\Phi(x_i \mid \boldsymbol{x}_{-i})} = \frac{P_\Phi(x_i' \mid \boldsymbol{u}_i)}{P_\Phi(x_i \mid \boldsymbol{u}_i)}$$

Assign the values of the evidence variables in $\boldsymbol{Y}_i$ to the nodes $\boldsymbol{x}_{-i}$

Using a Markov Chain

Using a Markov Chain

How do you use a Markov chain?

- Run chain till it converges to stationary distribution π
- Repeatedly sample from π to produce dataset ***D***
- Estimate probability from ***D***

But how do you know you are at the stationary distribution?

13

13

Using a Markov Chain

- Burn-in time T: the number of steps we take until we collect a sample from the chain
- Want T such that the Markov chain is close to the stationary distribution

14

14

Using a Markov Chain

Let $\mathcal{T}$ be a Markov chain. Let T_ε be the minimal T such that, for any starting distribution $P^{(0)}$, we have that:

$$\boldsymbol{D}_{\text{var}}(P^{(T)};\pi) \leq \varepsilon$$

Then T_ε is called the ε-mixing time of $\mathcal{T}$.

The variational distance $\boldsymbol{D}_{var}$ is defined as follows. Let P and Q be probability distributions defined over an event space S. Then

$$\boldsymbol{D}_{\text{var}}(P;Q) = \max_{\alpha \in S} | P(\alpha) - Q(\alpha) |$$

$$= \frac{1}{2}\|P - Q\|_1 = \sum_{x_1,\ldots,x_n} | P(x_1,\ldots,x_n) - Q(x_1,\ldots,x_n) |$$

15

15

Using a Markov Chain

- The mixing time can be very long!
- This happens when the state space looks like islands that are:
  - well-connected <u>within</u> the islands
  - but have low probability transitions <u>between</u> islands

16

16

Using a Markov Chain

- Let T be a Markov chain transition model and π its stationary distribution.
- The conductance of T is defined as follows:

$$\min_{S \subset Val(X)} \frac{P(\boldsymbol{S} \to \boldsymbol{S}^C)}{\pi(\boldsymbol{S})}, \quad \text{where } 0 < \pi(\boldsymbol{S}) \leq 1/2$$

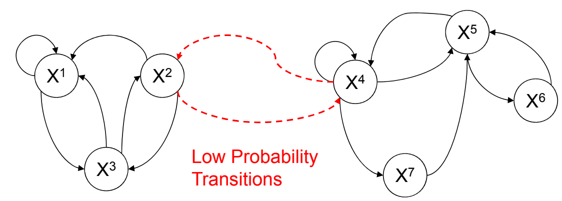
- Where
  - π(***S***) = probability assigned by the stationary distribution to the set of states ***S***
  - **S**^{C} = *Val(**X**)* – ***S***
  - $P(\boldsymbol{S} \to \boldsymbol{S}^C) = \sum_{x \in \boldsymbol{S}, x' \in \boldsymbol{S}^C} T(\boldsymbol{x} \to \boldsymbol{x}')$

17

17

Using a Markov Chain

- Intuitively, $P(\boldsymbol{S} \to \boldsymbol{S}^C)$ is the total “bandwidth” for transitioning from ***S*** to ***S***^{C}
- If conductance is low, if you are in some states ***S***, it is very hard to transition out of ***S***



18

18

Using a Markov Chain

- In graphical models, chains with low conductance most common in networks with deterministic or highly skewed parameterization
- Deterministic CPDs might lead to disconnected state spaces (a reducible chain)
- With positive distributions, might still have regions connected only by very low-probability transitions

19

19

Using Markov Chains

How do we obtain the ε-mixing time of a Markov chain?

- In general, it’s hard! Need to use heuristics
- Burn-in time is usually quite long

20

20

Collecting Samples

21

21

Collecting Samples

- Let $t = 0,\ldots,T$ be the burn-in phase
- Let $\boldsymbol{D} = \{\boldsymbol{x}^{(T+1)},\ldots,\boldsymbol{x}^{(T+M)}\}$ be M samples collected from stationary distribution π
  - Note that if $\boldsymbol{x}^{(T+1)}$ is from π then so are all M samples above
- If the chain has mixed, then for any function f, the following is an unbiased estimator for $\boldsymbol{E}_{\pi(\boldsymbol{X})}[f(\boldsymbol{X},\boldsymbol{e})]$:

$$\hat{\boldsymbol{E}}_D(f) = \frac{1}{M}\sum_{m=1}^{M} f(\boldsymbol{x}[m],\boldsymbol{e})$$

22

22

Collecting Samples

Theorem 12.6: Let $\mathcal{T}$ be a Markov chain and $\boldsymbol{X}[1]$, ..., $\boldsymbol{X}[M]$ a set of samples collected from $\mathcal{T}$ at its stationary distribution P. Then, since $M \to \infty$

$$\left(\hat{\boldsymbol{E}}_D(f) - \boldsymbol{E}_{X\sim P}[f(\boldsymbol{X})]\right) \to N(0;\sigma_f^2)$$

where

$$\sigma_f^2 = Var_{X\sim\mathcal{T}}[f(\boldsymbol{X})] +$$

$$2\sum_{l=1}^{\infty} \underbrace{Cov_{\mathcal{T}}[f(\boldsymbol{X}[m]); f(\boldsymbol{X}[m+l])]}_{\text{Autocovariance terms (due to correlated samples)}} < \infty$$

23

23

Collecting Samples

How do we use Theorem 12.6?

- Need to estimate variance from samples:

$$\sigma_f^2 = Var_{X\sim\mathcal{T}}[f(\boldsymbol{X})] \approx \frac{1}{M-1}\left[\sum_{m=1}^{M}\left(f(\boldsymbol{X}) - \hat{\boldsymbol{E}}_D(f)\right)^2\right]$$

- Need to estimate autocovariance terms:

$$Cov_{\mathcal{T}}[f(\boldsymbol{X}[m]); f(\boldsymbol{X}[m+l])] \approx$$

$$\frac{1}{M-l}\sum_{m=1}^{M-l}(f(\boldsymbol{X}[m] - \hat{\boldsymbol{E}}_D(f))(f(\boldsymbol{X}[m+l] - \hat{\boldsymbol{E}}_D(f))$$

24

24

Collecting Samples

How can we tell if the chain has mixed?

- Method 1: compute autocorrelation of lag l

$$p_l = \frac{Cov_{\mathcal{T}}[f(\boldsymbol{X}[m]); f(\boldsymbol{X}[m+l])]}{Var_{X\sim\mathcal{T}}[f(X)]}$$

- Autocorrelation should drop off exponentially with the length of the lag
- If you see high autocorrelation at distant lags, you have a poorly mixing chain
- Note: with large lags, you need more samples to estimate autocorrelation (otherwise you have large variance)

25

25

Collecting Samples

How can we tell if the chain has mixed?

- Method 2: Use multiple chains sampling the same distribution
- Suppose you have *K* chains run for *T*+*M* steps with different starting states
- Throw away the first *T* samples

26

26

Collecting Samples

- Let $\boldsymbol{X}_k[m]$ denote a sample from chain *k* after iteration *T*+*m*
- Compute the following:

$$\bar{f}_k = \frac{1}{T}\sum_{m=1}^{M} f(\boldsymbol{X}_k[m])$$

$$\bar{f} = \frac{1}{K}\sum_{m=1}^{M} \bar{f}_k$$

$$B = \frac{M}{K-1}\sum_{k=1}^{K}(\bar{f}_k - \bar{f})^2$$ Between-chain variance

$$W = \frac{1}{K}\frac{1}{T-1}\sum_{k=1}^{K}\sum_{m=1}^{M}(f(\boldsymbol{X}_k[m]) - \bar{f}_k)^2$$ Within-chain variance

27

27

Collecting Samples

- The following value *V* overestimates the variance of our estimate *f* based on the samples

$$V = \frac{M-1}{M}W + \frac{1}{M}B$$

- In the limit of $M \to \infty$, *W* and *V* converge to the true variance of the estimate
- Can use the following as a measure of disagreement between chains:

$$\hat{R} = \sqrt{\frac{V}{W}}$$

If equal to 1, all the chains have converged to either the true distribution or the same mode

28

28

Collecting Samples

Hybrid approach:

- Run small number of chains in parallel for a long time, diagnosing their behavior for mixing
- After burn-in phase, use multiple chains to estimate convergence and to generate multiple particles

29

29

Collecting Samples

Problems with MCMC methods

- Lots of hand-tuning:
  - Choosing proposal distribution
  - # of chains to run
  - Metrics for evaluating mixing
  - Lag between samples
  - Ways of making long-range moves in state space (eg. simulated annealing, block Gibbs sampling)
  - etc.
- This is more art than science!

30

30