Markov chain Monte Carlo (MCMC) inference

Jiali Lin

Virginia Tech

December 4, 2016

Outline

Introduction

Gibbs sampling

Metropolis Hastings algorithm

Speed and accuracy of MCMC

Auxiliary variable MCMC

Introduction

- **Markov chain Monte Carlo (MCMC)**: iterative sampling algorithm walks in high-demnsinal distributions.
- **Idea:** construct a Markov chain on the state space $\mathbb{X}$ whose stationary distribution is the target density $p(\boldsymbol{x})$ of interest.
- How? Perform a random walk on the state space, in such a way that the fraction of time we spend in each state $\boldsymbol{x}$ is proportional to $p(\boldsymbol{x})$.
- The **advantages** of sampling are:
  1. Easier to implement.
  2. Applicable to a broader range of models, such as models without nice conjugate priors.
  3. Can be faster than variational methods in large datasets.
- The **disadvantages**:
  1. Computationally demanding, often limiting their use to small-scale problems.
  2. Hard to know whether a sampling scheme is generating independent samples.

Outline

Gibbs sampling

Gibbs sampling is easy to sample $\boldsymbol{x}^s$. However, we need to know $p(x_i|\boldsymbol{x}_{-i})$.

Initialize x_0.

for $i = 1 : S$ **do**

1. $x_1^{s+1} \sim p(x_1|x_2^s, \ldots, x_p^s)$.
2. $x_2^{s+1} \sim p(x_2|x_1^{s+1}, \ldots, x_p^s)$.
3. $\ldots$.
4. $x_p^{s+1} \sim p(x_p|x_1^{s+1}, \ldots, x_{p-1}^{s+1})$.

return $x_1^s, \ldots, x_p^s$.

- Gibbs sampling could be very slow sometimes.
- **Collapsed Gibbs sampling:** we can analytically integrate out some of the unknown quantities, and just sample the rest.
- **Blocking Gibbs sampling:** we can efficiently sample groups of variables at a time.

Outline

Introduction

Gibbs sampling

Metropolis Hastings algorithm

Speed and accuracy of MCMC

Auxiliary variable MCMC

Metropolis Hastings algorithm

- **Idea:** at each step, we propose to move from the current state $\boldsymbol{x}$ to a new state $\boldsymbol{x}^*$ with probability $q(\boldsymbol{x}^*|\boldsymbol{x})$ (**proposal distribution**).
- Having proposed a move to $\boldsymbol{x}^*$, we then decide whether to accept this proposal or not according to some formula.
- It ensures that the fraction of time spent in each state is proportional to $p(\boldsymbol{x})$.
- If the proposal is accepted, the new state is $\boldsymbol{x}^*$, otherwise the new state is the same as the current state, $\boldsymbol{x}$.
- MH does not "discard" samples but "repeats" sample.

Initialize x_0.

for $i = 1 : S$ **do**

1. Sample $x^* \sim q(x^*|x)$.
2. Compute $\alpha = \frac{p(x^*)q(x|x^*)}{p(x)q(x^*|x)} = \frac{\tilde{p}(x^*)q(x|x^*)}{\tilde{p}(x)q(x^*|x)}$ where $p(x) = \frac{1}{z}\tilde{p}(x)$.
3. $r = \min(1, \alpha)$.
4. Sample $u \sim U(0, 1)$.
5. $x^{s+1} = x^*$ if $u < r$. Otherwise, $x^{s+1} = x^s$.

return $x_1^s, \ldots, x_p^s$.

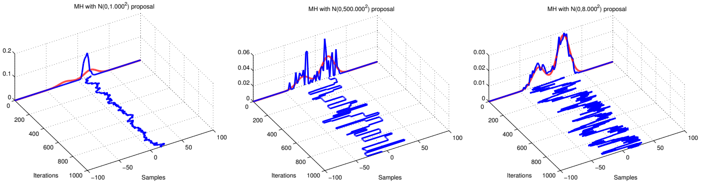
How MH works?

- We want: required distribution $p(\boldsymbol{x})$ is invariant is to choose the transition probabilities.
- A sufficient (but not necessary) condition: **detailed balance**, defined by
$$p(\boldsymbol{x})p(\boldsymbol{x}^*|\boldsymbol{x}) = p(\boldsymbol{x}^*)p(\boldsymbol{x}|\boldsymbol{x}^*)$$
- A Markov chain that respects detailed balance is **reversible**.
- If a chain satisfies detailed balance, then p is its **stationary**.
- **Goal:** show MH algorithm defines a transition function that satisfies detailed balance and hence that p is its stationary distribution (It is not true the otherway around).

$$\begin{aligned}
p(\boldsymbol{x})q(\boldsymbol{x}^*|\boldsymbol{x})\alpha(\boldsymbol{x}^*) &= p(\boldsymbol{x})q(\boldsymbol{x}^*|\boldsymbol{x})\min(1, \frac{p(\boldsymbol{x}^*)}{q(\boldsymbol{x}|\boldsymbol{x}^*)}) \\
&= \min(p(\boldsymbol{x})q(\boldsymbol{x}^*|\boldsymbol{x}), p(\boldsymbol{x}^*)q(\boldsymbol{x}|\boldsymbol{x}^*)) \\
&= p(\boldsymbol{x}^*)q(\boldsymbol{x}|\boldsymbol{x}^*)\min(1, \frac{p(\boldsymbol{x})q(\boldsymbol{x}^*|\boldsymbol{x})}{p(\boldsymbol{x}^*)q(\boldsymbol{x}|\boldsymbol{x}^*)}) \\
&= p(\boldsymbol{x}^*)q(\boldsymbol{x}|\boldsymbol{x}^*)\alpha(\boldsymbol{x})
\end{aligned}$$

Illustration

Figure: An example of the MH for sampling from a mixture of two 1D Gaussians ($\mu = (-20, 20), \pi = (0.3, 0.7), \sigma = (100, 100)$), using a Gaussian proposal with variances of $v \in \{1, 500, 8\}$. Figure generated by `McmcGmmDemo`.



(a) (b) (c)

- When $v = 1$, the chain gets trapped near the starting state and fails to sample from the mode at $\mu = -20$.
- When $v = 500$, the chain is very "sticky", so its effective sample size is low.
- Using a variance of $v = 8$ is just right and leads to a good approximation of the true distribution (shown in red).

Gibbs sampling is a special case of MH

- Gibbs sampling is a special case of MH.
- We move to a new state where x_i is sampled from its full conditional.
- But $\boldsymbol{x}_{-i}$ is left unchanged.
- The acceptance rate of each such proposal

$$\alpha = \frac{p(\boldsymbol{x}')q(\boldsymbol{x}|\boldsymbol{x}')}{p(\boldsymbol{x})q(\boldsymbol{x}'|\boldsymbol{x})} = \frac{p(x_i'|\boldsymbol{x}'_{-i})p(\boldsymbol{x}'_{-i})p(x_i|\boldsymbol{x}'_{-i})}{p(x_i|\boldsymbol{x}_{-i})p(\boldsymbol{x}_{-i})p(x_i'|\boldsymbol{x}_{-i})} = 1$$

Proposal distributions

- A **valid** proposal q gives a non-zero probability of moving to the states that have non-zero probability in the target.
- Example: Gaussian random walk proposal.
- For a Gaussian random walk proposal, it is very important to set the variance of the proposal v correctly.
  - If the v is too low, the chain will only explore one of the modes.
  - If the v is too large, most of the moves will be rejected, and the chain will stay in the same state for a long time.
  - If we set the proposal's variance just right, the samples clearly explore the support of the target distribution.
- **Optimal acceptance rate:** between 25% and 40%.

Outline

Speed and accuracy of MCMC

- **Burn-in phase:** Samples collected before the chain has reached its stationary distribution do not come from p^*, and are thrown away.
- **Mixing time:** the amount of time a Markov chain takes to converge to the stationary distribution, and forget its initial state.
- **Trace plot:** shows the values the parameter took during the runtime of the chain.
- **Accuracy of MCMC**: samples produced by MCMC are auto-correlated, thus can not be used for estimation.

Outline

Slice sampling

- Sometimes we can sample by introducing dummy auxiliary variables.
- Require require that $\sum_z p(x,z) = p(x)$ and $p(x,z)$ is easier to sample from than just $p(x)$.
- Consider sampling from a univariate, but multimodal, distribution $\tilde{p}(x)$.
- Add an auxiliary variable u. We define the joint distribution

$$\hat{p}(x,u) = \begin{cases} 1/Z_p, & \text{if } 0 \le u \le \tilde{p}(x) \\ 0, & \text{otherwise} \end{cases}$$

  where $Z_p = \int \tilde{p}(x)dx$.
- The marginal distribution over x is given by

$$\int \hat{p}(x,u)du = \int_0^{\tilde{p}(x)} \frac{1}{Z_p} du = \frac{\tilde{p}(x)}{Z_p} = p(x)$$

Slice sampling (cont'd)

We can sample from $p(x)$ by sampling from $\hat{p}(x, u)$ and then ignoring u. The full conditionals have the form

$$p(u|x) = U_{[0,\tilde{p}(x)]}(u)$$
$$p(x|u) = U_A(x)$$

where $A = \{x : \tilde{p}(x) \geq u\}$ it the set of points on or above u.

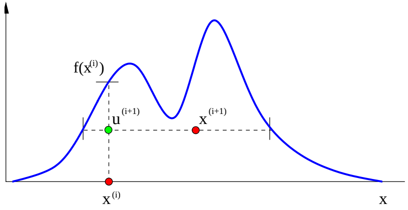


Figure: Illustration of the principle behind slice sampling. Given a previous sample x^i , we sample u^{i+1} uniformly on $[0, f(x^i)]$, which then defines a 'slice' through the distribution. We then sample x^{i+1} along the slice where $f(x) \geq u^{i+1}$. Figure generated by `SliceSamplingDemo1d`.