MCMC

Rahul G. Krishnan

New York University

Lab 10, Nov 9, 2016

Outline

Overview

You have seen atleast two few different ways to estimate expectations:

- Importance Sampling
- Markov Chain Monte-Carlo
  - Metropolis Hastings
  - Gibbs Sampling
  - Hamiltonian Monte Carlo
- Code examples

Importance Sampling

- Approximate $\mathbb{E}_{p(x)}[f(x)]$ using another distribution $q(x)$
- Write as $\mathbb{E}_{q(x)}[f(x)w(x)]$ where Supp(p) $\subset$ Supp(q)
- Evaluate via Monte-Carlo
  $\mathbb{E}[f(x)] = \frac{1}{N}\sum_i f(x^{(i)})w(x^{(i)}) \qquad x^{(i)} \sim q, w(x^{(i)}) = \frac{p(x)}{q(x)}$
- Unbiased estimator
- Can work with unnormalized distributions
- Efficiency depends on how well one can approximate the typical set of the distribution

MCMC

Three conditions you need to construct a valid MCMC kernel

- Aperiodicity : Markov chain is not cyclic (non-zero probability of staying in the same state)
- Irreducibility : positive probability of visiting every state
- Aperiodicity and irreducibility imply that there exists a unique limiting or stationary distribution $\pi(x)$
- Detailed Balance:
  - If the kernel $\mathcal{T}$ satisfies $\pi(x)\mathcal{T}(x'|x) = \pi(x')\mathcal{T}(x|x')$
  - $\pi(x)$ is the stationary distribution
- All three together denote an *ergodic* Markov chain
- Ergodic Theorem: Samples from Markov chain will correspond to samples from the stationary distribution

Metropolis Hastings

- Given state x_a, sample from the proposal distribution $x_b \sim Q(x_b|x_a)$
- Compute $R = \min\left(1, \frac{p(x_b)Q(x_a|x_b)}{p(x_a)Q(x_b|x_a)}\right)$
- Accept with new state x_b with probability R otherwise x_a

Stationary Distribution

- The proposal distribution $Q(x_b|x_a)$ (or $K(x_a, x_b)$ in the lecture notes) and the acceptance rejection step form the transition kernel ($\mathcal{T}(x_b|x_a)$)
- Aperiodicity and irreducibility typically satisfied by our choice of the proposal distribution and rejection step
- Check detailed balance with respect to $\pi(x)$

- For the MH kernel we construct, $\mathcal{T}(x_a|x_b) = Q(x_a|x_b)p_{accept}(x_a|x_b)$
- Assume $\pi(x_b)Q(x_a|x_b) \geq \pi(x_a)Q(x_b|x_a)$
- Recall detailed balance: $\pi(x_b)\mathcal{T}(x_a|x_b) = \pi(x_a)\mathcal{T}(x_b|x_a)$.

$$RHS : \pi(x_a)Q(x_b|x_a)\min(1, \underbrace{\frac{\pi(x_b)Q(x_a|x_b)}{\pi(x_a)Q(x_b|x_a)}}_{\geq 1}) = \pi(x_a)Q(x_b|x_a)$$

$$LHS : \pi(x_b)Q(x_a|x_b)\min(1, \frac{\pi(x_a)Q(x_b|x_a)}{\pi(x_b)Q(x_a|x_b)})$$

$$LHS := \pi(x_b)Q(x_a|x_b)\frac{\pi(x_a)Q(x_b|x_a)}{\pi(x_b)Q(x_a|x_b)} = \pi(x_a)Q(x_b|x_a) = RHS$$

The MH kernel satisfies detailed balance

Gibbs Sampling

- Special case of MH
- The conditional distribution is used as the proposal distribution
- The proposal distribution is accepted with probability one

Proof

- The Gibbs sampling proposal distribution is $Q(x_i^b, x_{\neg i} | x_i^a, x_{\neg i}) = p(x_i^b | x_{\neg i}) = p(x_i^b, x_{\neg i}) p(x_{\neg i})$
- $R = \min\left(1, \frac{p(x_i^b, x_{\neg i}) Q(x_i^a, x_{\neg i} | x_i^b, x_{\neg i})}{p(x_i^a, x_{\neg i}) Q(x_i^b, x_{\neg i} | x_i^a, x_{\neg i})}\right)$
- Substitute proposal distribution
- $R = \min\left(1, \frac{p(x_i^b, x_{\neg i}) p(x_i^a, x_{\neg i}) p(x_{\neg i})}{p(x_i^a, x_{\neg i}) p(x_i^b, x_{\neg i}) p(x_{\neg i})}\right) = 1$
- Special case of MH

Hamiltonian

- Recall, high school/undergraduate physics.
- Imagine an object with position x and velocity p
- The energy of a system $H(x,p) = \underbrace{U(x)}_{\text{potential}} + \underbrace{K(p)}_{\text{kinetic}}$
- Hamiltonian Equations:
  - $\frac{\partial x_i}{\partial t} = \frac{\partial H}{\partial p_i}$
  - $\frac{\partial p_i}{\partial t} = -\frac{\partial H}{\partial x_i}$
  - i indexes a dimension of position and velocity
- If we have expressions for H and initial values of x, p, we can simulate the system
- We're going to use this to create an MCMC approach for inference with continuous random variables

Leapfrog Integrator

- At time t, pick some small δ
- Discretize time and approximate $p(t) \approx p(t-1) + \delta \frac{\partial p(t-1)}{\partial t}$
- $p_i(t + \frac{\delta}{2}) = p_i(t) - \frac{\delta}{2} \frac{\partial U}{\partial x_i(t)}$
- $x_i(t + \delta) = x_i(t) + \delta \frac{\partial K}{\partial p_i(t+\frac{\delta}{2})}$
- $p_i(t + \delta) = p_i(t + \frac{\delta}{2}) - \frac{\delta}{2} \frac{\partial U}{\partial x_i(t+\frac{\delta}{2})}$
- Allows us to simulate how the object moves through space

Inference

- We'll be doing inference to sample from a continuous distribution
- Potential energy: $U(x) \propto -\log p(x)$
- To use HMC, we'll define a set of auxillary random variables p to represent our velocities in each dimension
- We'll assume that $p(p) \sim \mathcal{N}(0, \mathbb{I})$
- Kinetic energy: $K(p) \propto -\log p(p) = p^T p$
- Canonical Distribution:

$$p(x, p) \propto \exp(-H(x, p)) = \exp(-U(x)) \exp(-K(p)) \propto p(x)p(p)$$

Algorithm

We'll construct a MCMC Kernel using the Hamiltonian

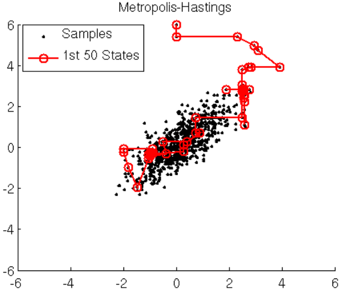
- Start from a random state x^0 for $t = 0$
- For $t = 1, \ldots, M$,
- $x_0 = x^{t-1}, p_0 \sim p(p)$ and simulate Leapfrog integrator for L steps. Yields x_*, p_*
- $R = \min(1, \frac{\exp(-H(x_*,p_*))}{\exp(-H(x_0,p_0))})$
- Accept new state $x^t = x_*$ with probability R, otherwise set $x^t = x^{t-1}$

Analysis

- Limitation: The leapfrog integrator requires you to be able to estimate gradients of $\log p(x)$ (which could be expensive)
- Advantage: The use of gradient information may lead to faster convergence and better scalability of the algorithm with high dimensional problems
- Stan uses a variant of HMC called NUTS to automate inference in any Bayesian network

Lets see this running on some code.

$$p(x) \sim \mathcal{N}(\mu; \Sigma)$$
$$\mu = [0; 0]$$
$$\Sigma = [10.8; 0.81]$$



References

- Blog Post https://theclevermachine.wordpress.com/2012/11/18/mcmc-hamiltonian-monte-carlo-a-k-a-hybrid-monte-
- The HMC Resource http://www.mcmchandbook.net/HandbookChapter5.pdf