Markov Chain Monte Carlo

Donald E. Brown

School of Data Science
University of Virginia
Charlottesville, VA 22904

Motivation for Markov Chain Monte Carlo

- Independent sampling methods, i.e., rejection sampling and importance sampling, don't scale to multidimensional problems and require finding proposal distributions that are close to the target
- MCMC is a framework for sampling from a large class of distributions
- MCMC scales well with the dimensionality of the sample space

MCMC Overview

- Use a proposal distribution, but now this distribution depends on the current sample or state, $q(\mathbf{x}|\mathbf{x}^{(\tau)})$ where $\mathbf{x}^{(\tau)}$ is the current state
- Note: This is a Markov chain and the sampling is no longer independent
- Again, the proposal distribution should be easy to sample from
- The target distribution should be easy to compute, at least up to a normalizing constant, so $p(\mathbf{x}) = \tilde{p}(\mathbf{x})/Z_p$ where Z_p may be unknown
- At each step of the sampling process, generate a candidate, $x^{(c)}$ from $q(\mathbf{x}|\mathbf{x}^{(\tau)})$ and accept, according to the criterion of the MCMC approach used

Basic Metropolis Algorithm

- Assume a symmetric proposal distribution, so $q(\mathbf{x}_A|\mathbf{x}_B) = q(\mathbf{x}_B|\mathbf{x}_A)$
- Compute

$$P(\mathbf{x}^{(c)}|\mathbf{x}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\mathbf{x}^{(c)})}{\tilde{p}(\mathbf{x}^{(\tau)})}\right)$$

- Choose $u \sim U(0,1)$ and accept the candidate, $\mathbf{x}^{(c)}$, if $P(\mathbf{x}^{(c)}|\mathbf{x}^{(\tau)}) > u$
- Notice that if $\tilde{p}(\mathbf{x}^{(c)}) \geq \tilde{p}(\mathbf{x}^{(\tau)})$, then the candidate will be accepted with certainty
- If the candidate is rejected, then draw a new sample from $q(\mathbf{x}|\mathbf{x}^{(\tau)})$

Metropolis-Hastings Algorithm

- No longer requires a symmetric proposal distribution, so
- we do not require $q(\mathbf{x}_A|\mathbf{x}_B) = q(\mathbf{x}_B|\mathbf{x}_A)$
  Suppose the current state is again, $\mathbf{x}^{(\tau)}$, then draw a candidate sample, $\mathbf{x}^{(c)}$ from $q(\mathbf{x}|\mathbf{x}^{(\tau)})$ and accept with probability

$$P(\mathbf{x}^{(c)}|\mathbf{x}^{(\tau)}) = \text{min}\left(1, \frac{\tilde{p}(\mathbf{x}^{(c)})q(\mathbf{x}^{(\tau)}|\mathbf{x}^{(c)})}{\tilde{p}(\mathbf{x}^{(\tau)})q(\mathbf{x}^{(c)}|\mathbf{x}^{(\tau)})}\right)$$

- Again, don't need to know Z_p
- If the proposal distribution is symmetric, then we have the basic Metropolis of slide 5

Resulting Markov Chain

- Result: The target distribution, $p(\mathbf{x})$ is the invariant distribution of the Markov Chain produced by the Metropolis-Hastings algorithm

Comments on the Metropolis-Hastings Algorithm

- Choices for proposal distribution
  - Continuous target: Gaussian or Cauchy centered on the current state
  - Discrete target: Uniform centered on the current state
- Since the samples are not independent, the algorithm does not immediately produce samples from $p(\mathbf{x})$, but requires start-up time, called *burn-in*
- Estimating burn-in time has been shown to be computationally intractable
- To obtain independent samples for $p(\mathbf{x})$, subsample every M^{th} point or generate multiple runs from different starting points and sample from them (after burn-in), called *thinning.*

Convergence

- The parameters of the proposal distribution have a strong influence on convergence
- For a Gaussian proposal distribution
  - If the variance is small then acceptance is high, but this will be a slow random walk through the state space
  - If the variance is large then the rejection rate will increase because we will now consider many states in which $p(\mathbf{x})$ is small

Example

- Suppose our target is a bivariate Gaussian with strong correlation between variables (shown on the next slide)
- Let $\sigma_{\max}$ be the scale in the longest axis and $\sigma_{\min}$ be the scale on the shortest
- Let ρ be the scale of the proposal distribution
- Suggests that we want $\rho \cong \sigma_{\min}$ to allow slow search in the longest dimension
- The number of steps to get past burn-in is of order $(\sigma_{\max}/\sigma_{\min})^2$

Example