Markov chain Monte Carlo methods

Youssef Marzouk

Department of Aeronautics and Astronautics
Center for Computational Engineering
Massachusetts Institute of Technology
http://uqgroup.mit.edu, ymarz@mit.edu

7 July 2015
Overview of the lecture

Markov chain Monte Carlo (MCMC)

- Metropolis-Hastings algorithm, transition kernels, ergodicity
- Mixture and cycles of kernels
- Gibbs sampling
- Gradient-exploiting MCMC, adaptive MCMC, other practicalities
- Using *approximations* (e.g., approximate likelihoods) within MCMC
In general, MCMC provides a means of sampling (“simulating”) from an arbitrary distribution.

- The density $\pi(x)$ need be known only up to a normalizing constant.
- Utility in inference and prediction: write both as posterior expectations, $\mathbb{E}_\pi f$. 
Why Markov chain Monte Carlo (MCMC)?

In general, MCMC provides a means of sampling ("simulating") from an arbitrary distribution.

- The density $\pi(x)$ need be known only up to a normalizing constant.
- Utility in inference and prediction: write both as posterior expectations, $\mathbb{E}_\pi f$.

Then

$$\mathbb{E}_\pi f \approx \frac{1}{n} \sum_{i}^{n} f(\mathbf{x}^{(i)})$$

- $\mathbf{x}^{(i)}$ will be asymptotically distributed according to $\pi$.
- $\mathbf{x}^{(i)}$ will not be i.i.d. In other words, we must pay a price!
Define a **Markov chain** (i.e., discrete time). For real-valued random variables, the chain has a continuous-valued state space (e.g., \( \mathbb{R}^d \)).

**Ingredients of the definition:**

- **Initial distribution,** \( x_0 \sim \pi_0 \)
- **Transition kernel** \( K (x_n, x_{n+1}) \).

\[
P (X_{n+1} \in A | X_n = x) = \int_A K (x, x') \, dx'
\]

(Analogy: consider matrix of transition probabilities for a finite state space.)

**Markov property:** \( X_{n+1} \) depends only on \( X_n \).

**Goal:** design transition kernel \( K \) such that chain converges asymptotically to the **target distribution** \( \pi \) independently of the initial distribution (starting point).
Goal: choose transition kernel $K$ such that chain converges asymptotically to the target distribution $\pi$ independently of the starting point.

- Use realizations of $X_n, X_{n-1}, \ldots$ in a Monte Carlo estimator of posterior expectations (an ergodic average)
- Would like to converge to the target distribution quickly and to have samples as close to independent as possible
- Price for non-i.i.d. samples: greater variance in MC estimates of posterior expectations
Metropolis-Hastings algorithm

A simple recipe!

1. Draw a proposal $y$ from $q(y|x_n)$
2. Calculate acceptance ratio
   \[
   \alpha(x_n, y) = \min\left\{ 1, \frac{\pi(y)q(x_n|y)}{\pi(x_n)q(y|x_n)} \right\}
   \]
3. Put
   \[
   x_{n+1} = \begin{cases} 
   y, & \text{with probability } \alpha(x_n, y) \\
   x_n, & \text{with probability } 1 - \alpha(x_n, y)
   \end{cases}
   \]
Metropolis-Hastings algorithm

A simple recipe!

1. Draw a proposal $y$ from $q(y|x_n)$
2. Calculate acceptance ratio
   \[
   \alpha(x_n, y) = \min\left\{ 1, \frac{\pi(y)q(x_n|y)}{\pi(x_n)q(y|x_n)} \right\}
   \]
3. Put
   \[
   x_{n+1} = \begin{cases} 
   y, & \text{with probability } \alpha(x_n, y) \\
   x_n, & \text{with probability } 1 - \alpha(x_n, y)
   \end{cases}
   \]

Very cool demo, thanks to Chi Feng (MIT):
http://chifeng.scripts.mit.edu/stuff/mcmc-demo/
Notes on the algorithm:

- If $q(y|x_n) \propto \pi(y)$ then $\alpha = 1$. Thus we “correct” for sampling from $q$, rather than from $\pi$, via the Metropolis acceptance step.
- $q$ does not have to be symmetric. If the proposal is symmetric, the acceptance probability simplifies (a “Hastings” proposal).
- $\pi$ need be evaluated only up to a multiplicative constant.
What is the transition kernel of the Markov chain we have just defined?

- *Hint*: it is not $q$!

Informally, it is

$$K(x_n, x_{n+1}) = p(x_{n+1} | \text{accept}) \cdot P[\text{accept}] + p(x_{n+1} | \text{reject}) \cdot P[\text{reject}]$$

More precisely, we have:

$$K(x_n, x_{n+1}) = p(x_{n+1} | x_n) = q(x_{n+1} | x_n) \cdot \alpha(x_n, x_{n+1}) + \delta x_n \cdot r(x_n)$$

where $r(x_n) \equiv \int q(y | x_n) \left(1 - \alpha(x_n, y)\right) dy$.
What is the **transition kernel** of the Markov chain we have just defined?

- **Hint**: it is not $q$!
- Informally, it is

\[
K(x_n, x_{n+1}) = p(x_{n+1}|\text{accept}) \mathbb{P}[\text{accept}] + p(x_{n+1}|\text{reject}) \mathbb{P}[\text{reject}]
\]
What is the **transition kernel** of the Markov chain we have just defined?

- *Hint*: it is not $q$!
- Informally, it is

$$K(x_n, x_{n+1}) = p(x_{n+1}|\text{accept}) \mathbb{P}[\text{accept}] + p(x_{n+1}|\text{reject}) \mathbb{P}[\text{reject}]$$

- More precisely, we have:

$$K(x_n, x_{n+1}) = p(x_{n+1}|x_n)$$

$$= q(x_{n+1}|x_n) \alpha(x_n, x_{n+1}) + \delta_{x_n}(x_{n+1}) r(x_n),$$

where $r(x_n) \equiv \int q(y|x_n) (1 - \alpha(x_n, y)) \, dy$
Metropolis-Hastings algorithm

Now, some theory. What are the key questions?

1. Is $\pi$ a stationary distribution of the chain? (Is the chain $\pi$-invariant?)
   - Stationarity: $\pi$ is such that $X_n \sim \pi \Rightarrow X_{n+1} \sim \pi$

2. Does the chain converge to stationarity? In other words, as $n \to \infty$, does $\mathcal{L}(X_n)$ converge to $\pi$?

3. Can we use paths of the chain in Monte Carlo estimates?
Metropolis-Hastings algorithm

Now, some theory. What are the key questions?

1. Is $\pi$ a stationary distribution of the chain? (Is the chain $\pi$-invariant?)
   - Stationarity: $\pi$ is such that $X_n \sim \pi \Rightarrow X_{n+1} \sim \pi$
2. Does the chain converge to stationarity? In other words, as $n \to \infty$, does $L(X_n)$ converge to $\pi$?
3. Can we use paths of the chain in Monte Carlo estimates?

A sufficient (but not necessary) condition for (1) is **detailed balance** (also called ‘reversibility’):

$$\pi(x_n)K(x_n, x_{n+1}) = \pi(x_{n+1})K(x_{n+1}, x_n)$$
Metropolis-Hastings algorithm

Now, some theory. What are the key questions?

1. Is $\pi$ a stationary distribution of the chain? (Is the chain $\pi$-invariant?)
   - Stationarity: $\pi$ is such that $X_n \sim \pi \Rightarrow X_{n+1} \sim \pi$

2. Does the chain converge to stationarity? In other words, as $n \to \infty$, does $\mathcal{L}(X_n)$ converge to $\pi$?

3. Can we use paths of the chain in Monte Carlo estimates?

A sufficient (but not necessary) condition for (1) is detailed balance (also called ‘reversibility’):

$$\pi(x_n)K(x_n,x_{n+1}) = \pi(x_{n+1})K(x_{n+1},x_n)$$

- This expresses an equilibrium in the flow of the chain
- Hence $\int \pi(x_n)K(x_n,x_{n+1}) \, dx_n = \int \pi(x_{n+1})K(x_{n+1},x_n) \, dx_n = \pi(x_{n+1}) \int K(x_{n+1},x_n) \, dx_n = \pi(x_{n+1})$.
- As an exercise, verify detailed balance for the M-H kernel defined on the previous slide.
Beyond $\pi$-invariance, we also need to establish (2) and (3) from the previous slide. This leads to additional technical requirements:

- **$\pi$-irreducibility**: for every set $A$ with $\pi(A) > 0$, there exists $n$ such that $K^n(x, A) > 0 \ \forall x$.
  
  *Intuition*: chain visits any measurable subset with nonzero probability in a finite number of steps. Helps you “forget” the initial condition. Sufficient to have $q(y|x) > 0$ for every $(x, y) \in \chi \times \chi$.

- **Aperiodicity**: “don’t get trapped in cycles”
When these requirements are satisfied (i.e., chain is *irreducible* and *aperiodic*, with *stationary* distribution $\pi$) we have

$$\lim_{n \to \infty} \left\| \int K^n(x, \cdot) \mu(dx) - \pi(\cdot) \right\|_{TV} = 0$$

for every initial distribution $\mu$.

- $K^n$ is the kernel for $n$ transitions
- This yields the law of $X_n$: $\int K^n(x, \cdot) \mu(dx) = \mathcal{L}(X_n)$
- The total variation distance $\|\mu_1 - \mu_2\|_{TV} = \sup_A |\mu_1(A) - \mu_2(A)|$ is the largest possible difference between the probabilities that the two measures can assign to the same event.
Metropolis-Hastings algorithm

When these requirements are satisfied (i.e., chain is *irreducible* and *aperiodic*, with *stationary* distribution $\pi$) we have

2. For $h \in L^1_\pi$,

$$
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} h(x^{(i)}) = \mathbb{E}_{\pi}[h] \text{ w.p. 1}
$$

This is a *strong law of large numbers* that allows computation of posterior expectations.

Obtaining a central limit theorem, or more generally saying anything about the *rate* of convergence to stationarity, requires additional conditions (e.g., geometric ergodicity).

What about the **quality** of MCMC estimates?
Metropolis-Hastings algorithm

What about the **quality** of MCMC estimates? What is the price one pays for correlated samples?

Compare Monte Carlo (iid) and MCMC estimates of $\mathbb{E}_\pi h$ (and for the latter, assume we have a CLT):

**Monte Carlo**

$$\text{Var} [\bar{h}_n] = \frac{\text{Var}_\pi [h(X)]}{n}$$

**MCMC**

$$\text{Var} [\bar{h}_n] = \frac{\text{Var}_\pi [h(X)]}{n} \theta$$

where

$$\theta = 1 + 2 \sum_{s>0} \infty \text{corr} (h(X_i), h(X_{i+s}))$$

is the **integrated autocorrelation**.
Now try a very simple computational demonstration: MCMC sampling from a univariate distribution.
More sophisticated Metropolis-Hastings

- M-H construction was extremely general.
- Achieving efficient sampling (good “mixing”) requires more exploitation of problem structure.
  1. Mixtures of kernels
  2. Cycles of kernels; Gibbs sampling
  3. Gradient-exploiting MCMC
  4. Adaptive MCMC
Mixtures and cycles

Mixtures of kernels

- Let $K_i$ all have $\pi$ as limiting distribution
- Use a convex combination: $K^* = \sum_i \nu_i K_i$
- $\nu_i$ is the probability of picking transition kernel $K_i$ at a given step of the chain
- Kernels can correspond to transitions that each have desirable properties, e.g., local versus global proposals

Cycles of kernels

- Split multivariate state vector into blocks that are updated separately; each update is accomplished by transition kernel $K_j$
- Need to combine kernels. **Cycle** = a systematic scan, $K^* = \prod_j K_j$
Componentwise Metropolis-Hastings

This is an example of using a cycle of kernels

- Let \( \mathbf{x} = (x^1, \ldots, x^d) \in \mathbb{R}^d \)
- Proposal \( q_i(y|x) \) updates only component \( i \)
- Walk through components of the state sequentially, \( i = 1 \ldots d \):
  - Propose a new value for component \( i \) using
    \[ q_i \left( y^i | x^1_{n+1}, \ldots, x^i_{n+1}, x_i^n, x^i_{n+1}, \ldots, x^d_n \right) \]
  - Accept \( (x^i_{n+1} = y^i) \) or reject \( (x^i_{n+1} = x^i_i) \) this component update with acceptance probability
    \[ \alpha_i(x_i, y_i) = \min \left\{ 1, \frac{\pi(y_i)q_i(x^i_i|y_i)}{\pi(x_i)q_i(y^i|x_i)} \right\} \]
    where \( \mathbf{x}_i \) and \( \mathbf{y}_i \) differ only in component \( i \)

\[ \mathbf{y}_i \equiv (x^1_{n+1}, \ldots, x^i_{n+1}, y^i, x^i_{n+1}, \ldots, x^d_n) \text{ and } \]
\[ \mathbf{x}_i \equiv (x^1_{n+1}, \ldots, x^i_{n+1}, x^i, x^i_{n+1}, \ldots, x^d_n) \]
Gibbs sampling

- One very useful cycle is the Gibbs sampler.
- Requires the ability to sample directly from the full conditional distribution \( \pi(x_i|x_{\sim i}) \).
  - \( x_{\sim i} \) denotes all components of \( x \) other than \( x_i \)
  - In problems with appropriate structure, generating independent samples from the full conditional may be feasible while sampling from \( \pi \) is not.
  - \( x_i \) can represent a block of the state vector, rather than just an individual component
- A Gibbs update is a proposal from the full conditional; the acceptance probability is identically one!

\[
\alpha_i(x_i, y_i) = \min \left\{ 1, \frac{\pi(y_i) q_i(x^n_i|y_i)}{\pi(x_i) q_i(y^i|x_i)} \right\} = \min \left\{ 1, \frac{\pi(y_i|x_{\sim i}) \pi(x_{\sim i}) \pi(x^n_i|x_{\sim i})}{\pi(x^n_i|x_{\sim i}) \pi(x_{\sim i}) \pi(y^i|x_{\sim i})} \right\} = 1.
\]
Gibbs sampling example

Correlated bivariate normal

\[ x \sim N \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix} \right) \]

Full conditionals are:

\[ x_1 | x_2 \sim N \left( \mu_1 + \frac{\sigma_1}{\sigma_2} \rho (x_2 - \mu_2), (1 - \rho^2) \sigma_1^2 \right) \]

\[ x_2 | x_1 \sim ... \]

See computational demo
Bayesian linear regression with a variance hyperparameter

\[ y_i = \mathbf{\beta}^T \mathbf{x}_i + \sigma z_i, \quad y_i \in \mathbb{R}; \quad \mathbf{\beta}, \mathbf{x}_i \in \mathbb{R}^d; \quad z_i \sim N(0, 1) \]

- This problem has a non-Gaussian posterior but is amenable to block Gibbs sampling
- Let the data consist of \( n \) observations \( \mathcal{D}_n \equiv \{(y_i, \mathbf{x}_i)\}_{i=1}^n \)
- Bayesian hierarchical model, likelihood and priors:

\[
\begin{align*}
\mathbf{y} \mid \mathbf{\beta}, \sigma^2 & \sim N(\mathbf{X}\mathbf{\beta}, \sigma^2 \mathbf{I}_n) \\
\mathbf{\beta} \mid \sigma^2 & \sim N(0, \tau^2 \sigma^2 \mathbf{I}_d) \\
1/\sigma^2 & \sim \Gamma(\alpha, \gamma)
\end{align*}
\]

where \( \mathbf{X} \in \mathbb{R}^{n \times d} \) has rows \( \mathbf{x}_i \) and \( \mathbf{y} \in \mathbb{R}^n \) is a vector of \( y_1 \ldots y_n \).
Gibbs sampling example (cont.)

Posterior density:
\[
\pi(\beta, \sigma^2) \equiv p(\beta, \sigma^2 | D_n)
\]
\[
\propto \frac{1}{\sigma^n} \exp \left( -\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) \right) \cdot \frac{1}{(\tau \sigma)^d} \exp \left( -\frac{1}{2\tau^2\sigma^2} \beta^T \beta \right) \\
\left( \frac{1}{\sigma^2} \right)^{\alpha-1} \exp \left( -\frac{\gamma}{\sigma^2} \right)
\]

Full conditionals \( \beta | \sigma^2, D_n \) and \( \sigma^2 | \beta, D_n \) have a closed form! Try to obtain by inspecting the joint density above. (See next page for answer.)
Gibbs sampling example (cont.)

- Full conditional for $\beta$ is Gaussian:
  
  $$\beta | \sigma^2, D_n \sim N(\mu, \sigma^2 \Sigma)$$

  where
  
  $$\Sigma^{-1} = \left( \frac{1}{\tau^2} I_d + X^T X \right)$$
  
  and
  
  $$\mu = \Sigma X^T y.$$ 

- Full conditional for $1/\sigma^2$ is Gamma:
  
  $$1/\sigma^2 | \beta, D_n \sim \Gamma(\hat{a}, \hat{\gamma})$$

  where
  
  $$\hat{a} = a + n/2 + d/2$$
  
  and
  
  $$\hat{\gamma} = \gamma + \frac{1}{2\tau^2} \beta^T \beta + \frac{1}{2} (y - X\beta)^T (y - X\beta).$$

- Alternately sample from these FCs in order to simulate the joint posterior.

- Also, this is an example of the use of conjugate priors.
What if we cannot sample from the full conditionals?
What if we cannot sample from the full conditionals?

- **Solution:** “Metropolis-within-Gibbs”
- This is just componentwise Metropolis-Hastings (which is where we started)
Langevin MCMC

- Intuitive idea: use gradient of the posterior to steer samples towards higher density regions
- Consider the SDE
  \[ dX_t = \frac{1}{2} \nabla \log \pi(X_t) dt + dW_t \]
  This SDE has \( \pi \) as its stationary distribution
- Discretize the SDE (e.g., Euler-Maruyama)
  \[ X^{t+1} = X^t + \frac{\sigma^2}{2} \nabla \log \pi(X^t) + \sigma \epsilon^t, \quad \epsilon^t \sim \mathcal{N}(0, I) \]
  - Discretized process \( X^t \) no longer has \( \pi \) as its stationary distribution! But we can use \( X^{t+1} \) as a proposal in the regular Metropolis-Hastings framework, and accept or reject it accordingly.
  - \( \sigma^2 \) (discretization time step) is an adjustable free parameter.
- Langevin schemes require access to the gradient of the posterior.
Preconditioned Langevin

- Introduce a positive definite matrix $A$ to the Langevin SDE:
  
  $$dX_t = \frac{1}{2} A \nabla \log \pi(X_t) dt + A^{1/2} dW_t$$

- Let $A$ reflect covariance structure of target

- For example: let $A$ be the local inverse Hessian of the log-posterior, or the inverse Hessian at the posterior mode, or posterior-averaged Hessian information, or some other estimate of the posterior covariance
Hamiltonian MCMC

- Let $x$ be “position” variables; introduce auxiliary “momentum” variables $w$
- Consider a separable Hamiltonian, $H(x, w) = U(x) + w^T M^{-1} w / 2$
- Hamiltonian dynamics are reversible and conserve $H$. Use them to propose new states $x$!
- In particular, sample from $p(x, w) = \frac{1}{Z} \exp \left( -\frac{H(x, w)}{T} \right)$:
  - First, sample the momentum variables $w$ from their Gaussian distribution
  - Second, integrate Hamilton’s equations to propose a new state $(x, w)$; then apply Metropolis accept/reject

**Features:**
- Enables faraway moves in $x$-space while leaving the value of the density (essentially) unchanged. Good mixing!
- Requires good symplectic integration methods, and access to derivatives
- Recent extension: Riemannian manifold HMC [Girolami & Calderhead JRSSB 2011]
Adaptive Metropolis

- Intuitive idea: learn a better proposal $q(y|x)$ from past samples.
  - Learn an appropriate proposal **scale**.
  - Learn an appropriate proposal **orientation** and anisotropy; this is *essential* in problems with strong correlation in $\pi$.
- Adaptive Metropolis scheme of [Haario et al. 2001]:
  - Covariance matrix at step $n$
    \[ C_n^* = s_d \text{Cov}(x_0, \ldots, x_n) + s_d \epsilon I_d \]
    where $\epsilon > 0$, $d$ is the dimension of the state, and $s_d = 2.4^2 / d$ (scaling rule-of-thumb).
  - Proposals are Gaussians centered at $x_n$. Use a fixed covariance $C_0$ for the first $n_0$ steps, then use $C_n^*$.
  - Chain is not Markov, and previous convergence proofs do not apply. Nonetheless, one can prove that the chain converges to $\pi$. See paper in references.
- Many other adaptive MCMC ideas have been developed in recent years.
Adaptive Metropolized independence samplers

- Independence proposal: does not depend on current state
- Consider a proposal $q(x; \psi)$ with parameter $\psi$.
- Key idea: minimize Kullback-Leibler divergence between this proposal and the target distribution:

$$\min_{\psi} D_{KL} (\pi(x) \| q(x; \psi))$$

- Equivalently, maximize $\int \pi(x) \log q(x; \psi) dx$

- Solve this optimization problem with successive steps of stochastic approximation (e.g., Robbins-Monro), while approximating the integral via MCMC samples
- Common choice: let $q$ be a mixture of Gaussians or other exponential-family distributions
MCMC in infinite dimensions

- Would like to construct a well-defined MCMC sampler for functions $u \in \mathcal{H}$.
- First, the posterior measure $\mu_y$ should be a well-defined probability measure on $\mathcal{H}$ (see Stuart *Acta Numerica* 2010). For simplicity, let the prior $\mu_0$ be $\mathcal{N}(0, C)$. 

Would like to construct a well-defined MCMC sampler for functions $u \in \mathcal{H}$.

First, the posterior measure $\mu_y$ should be a well-defined probability measure on $\mathcal{H}$ (see Stuart Acta Numerica 2010). For simplicity, let the prior $\mu_0$ be $\mathcal{N}(0, C)$.

Now let $q$ be the proposal distribution, and consider pair of measures

$$\nu(du, du') = q(u, du')\mu(du), \quad \nu^\perp(du, du') = q(u', du)\mu(du');$$

Then the MCMC acceptance probability is

$$\alpha(u_k, u') = \min\left\{1, \frac{d\nu^\perp}{d\nu}(u_k, u')\right\}$$

To define a valid transition kernel, we need absolute continuity $\nu^\perp \ll \nu$; in turn, this places requirements on the proposal $q$. 

One way to produce a valid transition kernel is the preconditioned Crank-Nicolson (pCN) proposal (Cotter et al. 2013):

\[ u' = (1 - \beta^2)^{1/2} u_k + \beta \xi_k, \quad \xi_k \sim \mathcal{N}(0, C), \quad \beta \in (0, 1). \]

Practical impact: sampling efficiency does not degenerate as discretization of \( u \) is refined

More sophisticated versions: combine pCN with Hessian/geometry information, e.g., DILI (dimension-independent likelihood-informed) proposals (Cui, Law, M 2015)
Recall Matt’s maple syrup example: http://nusselt.mit.edu/imauq

- **DR** = delayed rejection (Mira 2001)
- **AM** = adaptive Metropolis, **DR + AM** = **DRAM** (Haario et al. 2006)
- **NUTS** = no U-turn sampler (Hoffman & Gelman 2014), a variation of Hamiltonian/Hybrid MCMC (Neal 1996)
- **AMALA** = adaptive Metropolis-adjusted Langevin (Atchadé 2006)
- **MMALA** = simplified manifold MALA (Girolami & Calderhead 2011)

All are implemented in the MUQ (MIT Uncertainty Quantification) library, released at http://bitbucket.org/mituq/muq.
Effective use of MCMC still requires some (problem-specific) experience. Some useful rules of thumb:

- Adaptive schemes are not a panacea.
- Whenever possible, parameterize the problem in order to minimize posterior correlations.
- What to do, if anything, about “burn-in?”
- Visual inspection of chain components is often the first and best convergence diagnostic.
- Also look at autocorrelation plots. Run multiple chains from different starting points. Evaluate multivariate potential scale reduction factor (MPSRF, Gelman & Brooks), and other diagnostics.
**MCMC practicalities**

Additional advice:

- “The best Monte Carlo is a dead Monte Carlo”: If you can tackle any part of the problem analytically, do it.
A small selection of useful “general” MCMC references.


Disclaimer: this is a hopelessly incomplete list!


T. A. Moselhy and Y. Marzouk, “Bayesian inference with optimal maps.”

