Markov chain Monte Carlo methods

Youssef Marzouk

Department of Aeronautics and Astronatics
Massachusetts Institute of Technology
ymarz@mit.edu

22 June 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

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, $\mathbb{E}_\pi f$.

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

$$\mathbb{E}_\pi f \approx \frac{1}{n} \sum_i^n f\left(x^{(i)}\right)$$

- $x^{(i)}$ will be asymptotically distributed according to π
- $x^{(i)}$ will **not** be i.i.d. In other words, we must pay a price!

Construction of an MCMC sampler

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 $K(x_n, x_{n+1})$.

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

(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* π independently of the initial distribution (starting point).

Construction of an MCMC sampler (cont.)

Goal: choose transition kernel K such that chain converges asymptotically to the *target distribution* π 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/`

Metropolis-Hastings algorithm

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 π, via the Metropolis acceptance step.
- q does not have to be symmetric. If the proposal is symmetric, the acceptance probability simplifies (a "Hastings" proposal).
- π need be evaluated only up to a multiplicative constant

Metropolis-Hastings algorithm

What is the **transition kernel** of the Markov chain we have just defined?

- *Hint:* it is not q!

Metropolis-Hastings algorithm

What is the **transition kernel** of the Markov chain we have just defined?

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

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

Metropolis-Hastings algorithm

What is the **transition kernel** of the Markov chain we have just defined?

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

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

- More precisely, we have:

$$\begin{aligned} K\left(x_n, x_{n+1}\right) &= p\left(x_{n+1}|x_n\right) \\ &= q(x_{n+1}|x_n)\alpha(x_n, x_{n+1}) + \delta_{x_n}\left(x_{n+1}\right) r(x_n), \\ \text{where } r(x_n) &\equiv \int q(y|x_n)\left(1-\alpha(x_n, y)\right) dy \end{aligned}$$

Metropolis-Hastings algorithm

Now, some theory. What are the key questions?

1. Is π a stationary distribution of the chain? (Is the chain π-invariant?)
   - Stationarity: π 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 π?
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 π a stationary distribution of the chain? (Is the chain π-invariant?)
   - Stationarity: π 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 π?
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 π a stationary distribution of the chain? (Is the chain π-invariant?)
   - Stationarity: π 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 π?
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.

Metropolis-Hastings algorithm

Beyond π-invariance, we also need to establish (2) and (3) from the previous slide. This leads to additional technical requirements:

- π-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"

Metropolis-Hastings algorithm

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

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

   for every initial distribution μ.

   - 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 π) we have

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

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

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).

See [Roberts & Rosenthal 2004] for an excellent survey of MCMC convergence results.

Metropolis-Hastings algorithm

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

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

MCMC

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

where

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

is the **integrated autocorrelation**.

Metropolis-Hastings algorithm

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 π as limiting distribution
- Use a convex combination: $K^* = \sum_i \nu_i K_i$
- ν_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 $q_i(y|\mathbf{x})$ updates only component i
- Walk through components of the state sequentially, $i = 1 \ldots d$:
  - Propose a new value for component i using

$$q_i\left(y^i | x_{n+1}^1, \ldots, x_{n+1}^{i-1}, x_n^i, x_n^{i+1}, \ldots, x_n^d\right)$$

  - Accept $(x_{n+1}^i = y^i)$ or reject $(x_{n+1}^i = x_n^i)$ this component update with acceptance probability

$$\alpha_i(\mathbf{x}_i, \mathbf{y}_i) = \min\left\{1, \frac{\pi(\mathbf{y}_i) q_i(x_n^i | \mathbf{y}_i)}{\pi(\mathbf{x}_i) q_i(y^i | \mathbf{x}_i)}\right\}$$

    where $\mathbf{x}_i$ and $\mathbf{y}_i$ differ only in component i

$$\begin{aligned} \mathbf{y}_i &\equiv \left(x_{n+1}^1, \ldots, x_{n+1}^{i-1}, y, x_n^{i+1}, \ldots, x_n^d\right) \text{ and} \\ \mathbf{x}_i &\equiv \left(x_{n+1}^1, \ldots, x_{n+1}^{i-1}, x_n^i, x_n^{i+1}, \ldots, x_n^d\right) \end{aligned}$$

Gibbs sampling

- One very useful *cycle* is the Gibbs sampler.
- Requires the ability to sample directly from the *full conditional distribution* $\pi(x_i|\mathbf{x}_{\sim i})$.
  - $\mathbf{x}_{\sim i}$ denotes all components of $\mathbf{x}$ other than x_i
  - In problems with appropriate *structure*, generating independent samples from the full conditional may be feasible while sampling from π 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**!

$$\begin{aligned}\alpha_i(\mathbf{x}_i, \mathbf{y}_i) &= \min\left\{1, \frac{\pi(\mathbf{y}_i)\, q_i(x_n^i|\mathbf{y}_i)}{\pi(\mathbf{x}_i)\, q_i(y^i|\mathbf{x}_i)}\right\} \\ &= \min\left\{1, \frac{\pi(y_i|\mathbf{x}_{\sim i})\pi(\mathbf{x}_{\sim i})\, \pi(x_n^i|\mathbf{x}_{\sim i})}{\pi(x_n^i|\mathbf{x}_{\sim i})\pi(\mathbf{x}_{\sim i})\, \pi(y^i|\mathbf{x}_{\sim i})}\right\} = 1.\end{aligned}$$

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:

$$\begin{aligned} x_1|x_2 &\sim N\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2-\mu_2), (1-\rho^2)\sigma_1^2\right)\\ x_2|x_1 &\sim \ldots \end{aligned}$$

See computational demo

Gibbs sampling example

Bayesian linear regression with a variance hyperparameter

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

- 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:

$$\begin{aligned} \mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2 &\sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right) \\ \boldsymbol{\beta} \,|\, \sigma^2 &\sim N\left(0, \tau^2 \sigma^2 \mathbf{I}_d\right) \\ 1/\sigma^2 &\sim \Gamma\left(\alpha, \gamma\right) \end{aligned}$$

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:

$$\begin{aligned}
\pi\left(\boldsymbol{\beta}, \sigma^2\right) &\equiv p\left(\boldsymbol{\beta}, \sigma^2 | \mathcal{D}_n\right) \\
&\propto \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^T\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right) \\
&\quad \frac{1}{(\tau\sigma)^d} \exp\left(-\frac{1}{2\tau^2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \\
&\quad \left(\frac{1}{\sigma^2}\right)^{\alpha-1} \exp\left(-\gamma/\sigma^2\right)
\end{aligned}$$

- Full conditionals $\boldsymbol{\beta}|\sigma^2, \mathcal{D}_n$ and $\sigma^2|\boldsymbol{\beta}, \mathcal{D}_n$ 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 $\boldsymbol{\beta}$ is Gaussian:

$$\boldsymbol{\beta} \,|\, \sigma^2, \mathcal{D}_n \sim N\left(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma}\right)$$

where

$$\boldsymbol{\Sigma}^{-1} = \left(\frac{1}{\tau^2}\mathbf{I}_d + \mathbf{X}^T\mathbf{X}\right) \text{ and } \boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{X}^T\mathbf{y}.$$

- Full conditional for $1/\sigma^2$ is Gamma:

$$1/\sigma^2 \,|\boldsymbol{\beta}, \mathcal{D}_n \sim \Gamma\left(\hat{\alpha}, \hat{\gamma}\right)$$

where

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

- Alternately sample from these FCs in order to simulate the joint posterior.
- Also, this is an example of the use of **conjugate priors**.

Metropolis-within-Gibbs

What if we cannot sample from the full conditionals?

Metropolis-within-Gibbs

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 π as its stationary distribution
- Discretize the SDE (e.g., Euler-Maruyama)

$$X^{t+1} = X^t + \frac{\sigma^2}{2}\nabla \log \pi(X^t) + \sigma\epsilon^t,\ \epsilon^t \sim N(0, I)$$

  - Discretized process X^t no longer has π 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.
  - σ^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 $\mathbf{A}$ to the Langevin SDE:

$$dX_t = \frac{1}{2}\mathbf{A}\nabla \log \pi(X_t)dt + \mathbf{A}^{1/2}dW_t$$

- Let $\mathbf{A}$ reflect covariance structure of target
- For example: let $\mathbf{A}$ be the local inverse Hessian of the log-posterior, or the inverse Hessian at the posterior mode

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 $p(x,w) = \frac{1}{Z}\exp\left(-H(x,w)/T\right)$:
  - 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 π
- Adaptive Metropolis scheme of [Haario *et al.* 2001]:
  - Covariance matrix at step n

$$C_n^* = s_d \, \mathrm{Cov}\,(x_0, \ldots, x_n) + s_d \epsilon I_d$$

    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 π. 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 ψ.
- Key idea: minimize Kullback-Leibler divergence between this proposal and the target distribution:

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

- 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

Samplers in the IPython demo

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)

MCMC practicalities

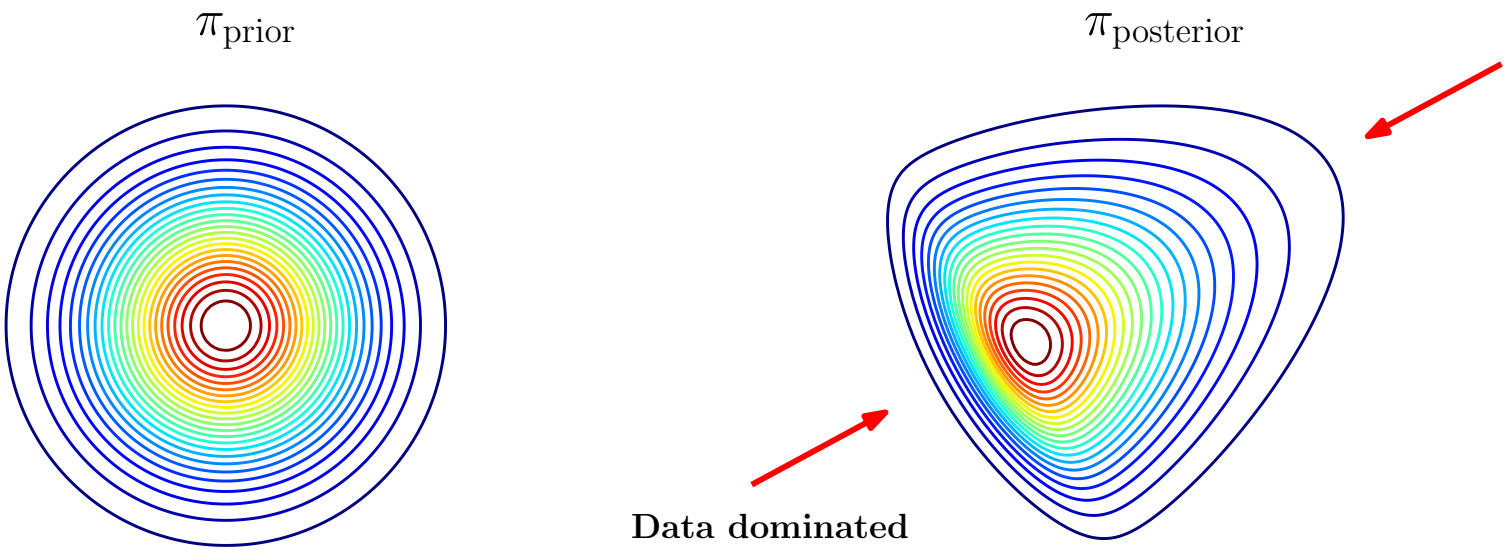
Effective use of MCMC still requires some (problem-specific) experience. Some useful rules of thumb:

- 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.
  - Example: Cui, Martin, Marzouk, Solonen, Spantini, "Likelihood-informed dimension reduction for nonlinear inverse problems," *Inverse Problems* 30: 114015 (2014).



Approximations in MCMC

Efficient sampling is great, but what if each posterior evaluation is very expensive?

Approximations in MCMC

Efficient sampling is great, but what if each posterior evaluation is very expensive?

Obvious answer: *approximate* the expensive part, e.g., the forward model.

Approximations in MCMC

Efficient sampling is great, but what if each posterior evaluation is very expensive?

Obvious answer: *approximate* the expensive part, e.g., the forward model.

This raises many interesting issues:

- What kind of approximation scheme to use? What properties of the forward model/likelihood are being exploited?
- When to construct the approximation (offline versus online) and what kind of accuracy to demand from it?
- What is the accuracy of the resulting posterior? Bias in posterior estimates? Can/should we correct for these?

Approximations in MCMC

Much work has been done on this topic.

Approximations in MCMC

Much work has been done on this topic.

Approximation schemes: coarse-grid PDE models, polynomial expansions, Gaussian process emulators, reduced-basis methods and reduced-order models, simplified physics, etc.

Approximations in MCMC

Much work has been done on this topic.

Approximation schemes: coarse-grid PDE models, polynomial expansions, Gaussian process emulators, reduced-basis methods and reduced-order models, simplified physics, etc.

Construction schemes:

- Surrogates accurate over the prior (e.g., convergent in $L^2_{\pi_{\text{prior}}}$ sense) versus *posterior-focused* (and hence data-driven) surrogates
- Constructed offline or *online* during posterior sampling

Approximations in MCMC

Much work has been done on this topic.

Approximation schemes: coarse-grid PDE models, polynomial expansions, Gaussian process emulators, reduced-basis methods and reduced-order models, simplified physics, etc.

Construction schemes:

- Surrogates accurate over the prior (e.g., convergent in $L^2_{\pi_{\text{prior}}}$ sense) versus *posterior-focused* (and hence data-driven) surrogates
- Constructed offline or *online* during posterior sampling

Errors and correction:

- Convergence rate of the forward model approximation transfers to the posterior it induces (Marzouk & Xiu 2009; Cotter, Dashti, Stuart 2010)
- Can always correct using a *delayed-acceptance* scheme (Christen & Fox 2005), but at a price
- Recent work in *asymptotically exact*, online, and posterior-focused approximations (Conrad, Marzouk, Pillai, Smith 2015)

MCMC references

A small selection of useful "general" MCMC references.

- C. Andrieu, N. de Freitas, A. Doucet, M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning* 50 (2003) 5–43.
- S. Brooks, A. Gelman, G. Jones and X. Meng, editors. *Handbook of MCMC*. Chapman & Hall/CRC, 2011.
- A. Gelman, J. B. Carlin, H. S. Stern, D. Dunson, A. Vehtari, D. B. Rubin. *Bayesian Data Analysis.* Chapman & Hall CRC, 3rd edition, 2013.
- P. J. Green. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82: 711–732, 1995.
- H. Haario, M. Laine, A. Mira, and E. Saksman. "DRAM: Efficient adaptive MCMC." *Statistics and Computing*, 16(4): 339–354, 2006.
- C. P. Robert, G. Casella, *Monte Carlo Statistical Methods*, 2nd Edition, Springer, 2004.
- G. Roberts, J. Rosenthal. "General state space Markov chains and MCMC algorithms." *Probability Surveys*, 1: 20–71, 2004.

MCMC with surrogate modeling (1)

Disclaimer: this is a hopelessly incomplete list!

- J. A. Christen and C. Fox, "Markov chain Monte Carlo using an approximation." *J. Comp. Graph. Stat.* 14: 795–810, 2005.
- P. Conrad, Y. Marzouk, N. Pillai, and A. Smith, "Accelerating asymptotically exact MCMC for computationally intensive models via local approximations." Submitted, arXiv:1402.1694, 2015.
- T. Cui, C. Fox, and M. J. O'Sullivan, "Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm." *Water Resources Research* 47: W10521, 2011.
- T. Cui, Y. Marzouk, and K. E. Willcox, "Data-driven model reduction for the Bayesian solution of inverse problems." *Int. J. Num. Meth. Eng.*, 102: 966–990, 2015.
- V. Hoang, Ch. Schwab, A. Stuart, "Complexity analysis of accelerated MCMC methods for Bayesian inversion." *Inverse Problems* 29: 085010, 2013.
- J. Li and Y. Marzouk, "Adaptive construction of surrogates for the Bayesian solution of inverse problems." *SIAM J. Sci. Comp.*, 36: A1163–A1186, 2014.
- Y. Marzouk, H. Najm, L. Rahn, "Stochastic spectral methods for efficient Bayesian solution of inverse problems." *J. Comp. Phys.* 224: 560–586, 2007.

MCMC with surrogate modeling (2)

- Y. Marzouk, H. Najm. "Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems." *J. Comp. Phys.*, 228: 1862–1902, 2009.
- Y. Marzouk, D. Xiu. "A stochastic collocation approach to Bayesian inference in inverse problems." *Comm. Comp. Phys.*, 6(4): 826–847, 2009.

Advanced posterior sampling for inverse problems (1)

Disclaimer: this is a hopelessly incomplete list!

- S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, "MCMC methods for functions: modifying old algorithms to make them faster." *Statistical Science*, 28: 424–446, 2013.
- T. Cui, K. Law, and Y. Marzouk, "Dimension-independent likelihood-informed MCMC." Submitted, arXiv:1411.3688, 2014.
- T. Cui, J. Martin, Y. Marzouk, A. Solonen, and A. Spantini, "Likelihood informed dimension reduction for nonlinear inverse problems." *Inverse Problems*, 30 (2014), 114015.
- M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods." *J. Roy. Stat. Soc. B*, 73: 123–214, 2011.
- C. Ketelsen, R. Scheichl, A. Teckentrup, "A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow." arXiv:1303.7343, 2013.
- J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas, "A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion." *SIAM J. Sci. Comp.* 34: A1460–A1487, 2012.

Advanced posterior sampling for inverse problems (2)

- T. A. Moselhy and Y. Marzouk, "Bayesian inference with optimal maps." *J. Comp. Phys.*, 231: 7815–7850, 2012.
- M. Parno, Y. Marzouk, "Transport map accelerated Markov chain Monte Carlo." Submitted, arXiv:1412.5492, 2015.
- N. Petra, J. Martin, G. Stadler, and O. Ghattas, "A computational framework for infinite-dimensional Bayesian inverse problems: Part II. Stochastic Newton MCMC with application to ice sheet inverse problems." *SIAM J. Sci. Comp.*, 36: A1525–A1555, 2014.
- C. Schillings, Ch. Schwab, "Sparse adaptive Smolyak quadratures for Bayesian inverse problems." *Inverse Problems* 29: 065011, 2013.