Building MCMC From Hastings-Peskun to meta-algorithms

Omiros Papaspiliopoulos
ICREA, UPF & BGSE
http://www.econ.upf.edu/~omiros
Material based on part of a chapter of a book with Gareth O. Roberts

Key Ideas

The same algorithmic framework on different targets:

- Hastings-Peskun framework
  - Sub-optimality vs computational efficiency
- Think big:
  - Hypo-dimensional MCMC and moves on manifolds
  - Groups and generalised Gibbs sampling
- Algorithm augmentation: meta-algorithms
  - Intractable targets and proposals
  - Augmenting algorithms

Outline

Invariant measures

A σ-finite measure π on $\mathfrak{B}(\mathfrak{X})$ with the property

$$\pi(A) = \int_{\mathfrak{X}} \pi(dx) P(x, A), \qquad A \in \mathfrak{B}(\mathfrak{X})$$

will be called *invariant*.

Reversibility

We say that a Markov chain P is *reversible* with respect to a probability measure π if

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx)$$

the equality being understood as an equality of the two measures as defined on $\mathfrak{B}(\mathfrak{X}) \otimes \mathfrak{B}(\mathfrak{X})$.

Markov transition kernels based on proposals and rejections

$$P(x, dy) = Q(x, dy)\alpha(x, y) + r(x)\delta_x(dy)$$

where

$$r(x) = \int (1 - \alpha(x, u))Q(x, du)\,.$$

hence, reversibility becomes

$$\pi(dx)Q(x, dy)\alpha(x, y) = \pi(dy)Q(y, dx)\alpha(y, x)\,,$$

Now define the Radon-Nikodym derivative,

$$t(x, y) = \frac{\pi(dx)Q(x, dy)}{\pi(dy)Q(y, dx)}$$

Hastings[1]-Peskun[2] framework

Re-arranging the reversibility equation we get

$$\alpha(x, y)t(x, y) = \alpha(y, x)$$

and since

$$s(x, y) = \alpha(x, y) + \alpha(y, x)$$

is symmetric by construction, we obtain that *any* reversible-inducing acceptance probability should be

$$\alpha(x, y) = \frac{s(x, y)}{1 + t(x, y)}$$

where $s(x, y)$ is symmetric such that $0 \leq \alpha(x, y) \leq 1$

---

[1] Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109

[2] Peskun, P. H. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60:607–612

From above we get

$$\alpha(x, y) \leq \min\{1, t(y, x)\}$$

Proposition

Any valid acceptance probability can be expressed as

$$\alpha(x, y) = \min\{1, t(y, x)\}\tilde{s}(x, y)\,,$$

where $\tilde{s}$ is symmetric, and $0 \leq \tilde{s}(x, y) \leq 1$.

Metropolis-Hastings rule:

$$\alpha(x, y) = \min\{1, t(y, x)\} = \min\left\{1, \frac{\pi(dy)Q(y, dx)}{\pi(dx)Q(x, dy)}\right\}$$

Barker's algorithm:

$$s(x, y) = 1$$

MCMC for computationally expensive and measures on Hilbert spaces

$$\pi(x) = \pi_2(x)\pi_1(x)$$

where one is expensive and other cheap to compute:

- likelihood/prior
- $\pi_1 = \tilde{\pi}$ and $\pi_2 = \pi/\tilde{\pi}$

$$t_1(x,y) = \frac{\pi_1(dx)Q(x,dy)}{\pi_1(dy)Q(y,dx)}\,, \quad t_2(x,y) = \frac{\pi_2(x)}{\pi_2(y)}\,, \quad t = t_1 \times t_2$$

Practically cheaper to decide according to

$$\min\{1, t_1(y,x)\} \times \min\{1, t_2(y,x)\},$$

as opposed to

$$\min\{1, t_1(y,x)t_2(y,x)\}$$

Easy to check that this is a special case of generic with

$$s(x,y) = (1 + t(x,y)) \min\{1, t_1(y,x)\} \min\{1, t_2(y,x)\},$$

Essence behind delayed acceptance[3] and certain methods in graphical models [4]

---

[3] Christen, J. A. and Fox, C. (2005). Markov chain Monte Carlo using an approximation. *J. Comput. Graph. Statist.*, 14(4):795–810

[4] Green, P. J. and Thomas, A. (2013). Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, 100(1):91–110

Rejection probability at the first step is zero when $Q(x, dy)$ is *reversible* wrt π_1. Then overall:

$$\min\{1, \pi_2(y)/\pi_2(x)\}$$

Attractive, e.g. when $\pi_1(dx) \equiv N(0, C)$, see latent Gaussian models of Neal[5] and distributions on Hilbert spaces[6]

E.g.

$$y = \sqrt{1-\rho^2}x + \rho L\xi\,, \quad \xi \sim N(0, I)\,, \quad \rho \in [-1, 1]\,, \quad LL^* = C\,,$$

---

[5] Neal, R. M. (1999). Regression and classification using Gaussian process priors. In *Bayesian statistics, 6 (Alcoceber, 1998)*, pages 475–501. Oxford Univ. Press, New York

[6] Beskos, A., Roberts, G. O., Stuart, A. M., and Voss, J. (2008b). MCMC methods for diffusion bridges. *Stochastics and Dynamics*, 8(3):319–350

Outline

1. Hastings-Peskun framework
2. Hypo-dimensional MCMC
3. Auxiliary expansions
4. Exact population MCMC

Hypo-dimensional MCMC

Perspective: view proposal as deterministic transform of current and random seeds and expand the state-space of the Markov chain. Design moves on manifolds.

- (original) state-space $\mathfrak{X} \subseteq \mathbb{R}^d$, noise space $\mathfrak{U} \subseteq \mathbb{R}^m$
- $\pi(\mathbf{x}, \mathbf{u}) = \pi(\mathbf{x})\pi_{\mathbf{x}}(\mathbf{u})$
- Involution $T : \mathfrak{X} \times \mathfrak{U} \to \mathfrak{X} \times \mathfrak{U}$
- $(\mathbf{Y}, \mathbf{V}) = T(\mathbf{X}, \mathbf{U})$

(most perturbations you can think of are special case of this)

For example in random-walk Metropolis, $(\mathbf{Y}, \mathbf{V}) = (\mathbf{X} + \mathbf{U}, -\mathbf{U})$

Appealing to the generic Hastings-Peskun framework we get

$$t(\mathbf{x},\mathbf{u},\mathbf{y},\mathbf{v}) = \frac{\pi(\mathbf{x})\pi_{\mathbf{x}}(\mathbf{u})}{\pi(\mathbf{y})\pi_{\mathbf{y}}(\mathbf{v})|\det J_T(\mathbf{x},\mathbf{u})|}\,, \quad \text{where} \quad (\mathbf{y},\mathbf{v}) = T(\mathbf{x},\mathbf{u})\,.$$

Common implementations of Metropolis-Hastings would discard $\mathbf{V}$ before the next step of the algorithm. However, it can be beneficial not to do so, e.g. within Hamiltonian MCMC or Reversible Jump MCMC [7]

A strict subset of this framework is Hastings-within-Gibbs

[7] Brooks, S. P., Giudici, P., and Roberts, G. (2001). Efficient rjmcmc proposals. *submitted for publication*

Example: random walk on a hypersurface

Aim: perturb locally $\mathbf{x}$ while keeping $h(\mathbf{x})$ constant. Then:

$$T(\mathbf{x}^{(-d)}, x^{(d)}, \mathbf{u}) = (\mathbf{y}^{(-d)} = \mathbf{x}^{(-d)} + \mathbf{u}, y^{(d)} = f(\mathbf{x}^{(-d)} + \mathbf{u}, h(\mathbf{x})), \mathbf{v} = -\mathbf{u})$$

with Hastings-Peskun ratio

$$t(\mathbf{x}, \mathbf{u}, \mathbf{y}, \mathbf{v}) = \frac{\pi(\mathbf{x}) |f_d(\mathbf{x}^{(-d)}, h(\mathbf{x}))|}{\pi(\mathbf{y}) |f_d(\mathbf{y}^{(-d)}, h(\mathbf{x}))|}$$

Generalised Gibbs sampler

Q: is there a choice of $\pi_x(\mathbf{u})$ s.t. $t(\mathbf{x}, \mathbf{u}, \mathbf{y}, \mathbf{v}) = 1$?
A: Yes!

(a) $\mathfrak{U}$ equipped with a multiplication operator, $\cdot$, is a locally compact topological group and the left and right Haar measures associated to it have Lebesgue densities m_L and m_R respectively.

(b) The transformation T takes the following generic form:

$$T(\mathbf{x}, \mathbf{u}) = (S(\mathbf{x}, \mathbf{u}), \mathbf{u}^{-1})$$

where S is continuously differentiable function, and $\mathbf{u}^{-1}$ is the inverse of $\mathbf{u}$ according to the group.

(c) For any $\mathbf{u}, \mathbf{v}$

$$S(S(\mathbf{x}, \mathbf{u}), \mathbf{v}) = S(\mathbf{x}, \mathbf{v} \cdot \mathbf{u}) \, .$$

Note that this assumption, together with (b)-(c) above make T an involution and also imply that $T(\mathbf{x}, \mathbf{e}) = (\mathbf{x}, \mathbf{e})$ for all $\mathbf{x}$.

Haar densities

The topological group structure implies the existence of densities

$$m_L(\mathbf{u}\cdot\mathbf{w})\left|\det\frac{\partial(\mathbf{u}\cdot\mathbf{w})}{\partial\mathbf{w}}\right| = m_L(\mathbf{w}) \quad \forall\ \mathbf{w},\mathbf{u}\in\mathfrak{U}$$

$$m_R(\mathbf{w}\cdot\mathbf{u})\left|\det\frac{\partial(\mathbf{w}\cdot\mathbf{u})}{\partial\mathbf{w}}\right| = m_R(\mathbf{w}) \quad \forall\ \mathbf{w},\mathbf{u}\in\mathfrak{U}$$

with

$$m_L(\mathbf{u}^{-1})\left|\det\frac{d\mathbf{u}^{-1}}{d\mathbf{u}}\right| = m_R(\mathbf{u})$$

Effectively, in the spaces weighted by these densities the transformations $\mathbf{w}\to\mathbf{u}\cdot\mathbf{w}$ and $\mathbf{w}\to\mathbf{w}\cdot\mathbf{u}$ are volume preserving

Example: scale transformations

$\mathfrak{X} = \mathbb{R}^d$, $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})$, $A \subseteq \{1, 2, \ldots, d\}$; $\mathfrak{U} \in \mathbb{R}_+^m$, for $m = |A|$, and for convenience take the elements of $\mathbf{u}$ to be indexed by the indices in A

Element-wise multiplication and $\mathbf{u} \cdot \mathbf{v}$ is an m-dimensional vector with elements $u^{(j)} \times v^{(j)}$ for $j \in A$.

In this group, $\mathbf{e} = \mathbf{1}$ and $\mathbf{u}^{-1}$ has elements $1/u^{(j)}$ for $j \in A$.

$T(\mathbf{x}, \mathbf{u}) = (\mathbf{y}, \mathbf{u}^{-1})$, with $y^{(j)} = x^{(j)} \times u^{(j)}$ if $j \in A$ and $y^{(j)} = x^{(j)}$ otherwise. Left and right Haar densities can be taken to be the same, $m_L(\mathbf{u}) = m_R(\mathbf{u}) = \prod_{j \in A}(1/u^{(j)})$

Example: scale-affine transformations

Let $\mathfrak{X} \in \mathbb{R}_+ \times \mathbb{R}$, $\mathbf{x} = (x^{(1)}, x^{(2)})$, $\mathfrak{U} = \mathbb{R}_+ \times \mathbb{R}$ equipped with the assortative multiplication:

$$\mathbf{u} \cdot \mathbf{v} = \left(u^{(1)} \times v^{(1)}, u^{(1)} \times u^{(2)} + v^{(2)}\right),$$

where $\mathbf{e} = (1, 0)$, and $\mathbf{u}^{-1} = (1/u^{(1)}, -u^{(2)}/u^{(1)})$.

Then, the transformation is

$$T(\mathbf{x}, \mathbf{u}) = \left(u^{(1)} \times x^{(1)}, u^{(1)} \times x^{(2)} + u^{(2)}, \mathbf{u}^{-1}\right).$$

In this example left and Haar densities differ; we can take $m_L(\mathbf{u}) = \left(1/u^{(1)}\right)^2$ and $m_R(\mathbf{u}) = 1/u^{(1)}$.

Theorem

Suppose that Assumptions (a)-(c) hold, and that

$$c(\mathbf{x}) := \int \pi(S(\mathbf{x},\mathbf{u}))\left|\det S_1(\mathbf{x},\mathbf{u})\right| m_L(\mathbf{u})d\mathbf{u}$$

is such that $0 < c(\mathbf{x}) < \infty$ *for all* $\mathbf{x}$. *Then, by choosing*

$$\pi_{\mathbf{x}}(\mathbf{u}) = c(\mathbf{x})^{-1}\pi(S(\mathbf{x},\mathbf{u}))\left|\det S_1(\mathbf{x},\mathbf{u})\right| m_L(\mathbf{u})$$

the acceptance probability of the proposed move $(\mathbf{x},\mathbf{u}) \to T(\mathbf{x},\mathbf{u})$ *is 1.*

Outline

1. Hastings-Peskun framework
2. Hypo-dimensional MCMC
3. Auxiliary expansions
4. Exact population MCMC

Auxiliary expansions

"Purposely constructing unobserved/unobservable variables offers an extraordinarily flexible and powerful framework for both scientific modeling and computation and is one of the central statistical contributions to natural, engineering, and social sciences."[8].

[8] Meng, X.-L. (2000). Missing data: dial M for ???
J. Amer. Statist. Assoc., 95(452):1325–1330

A taxonomy

1. Missing data/data augmentation
2. Vertical expansion: if

$$(X, Z) \sim U\left(\{(x, z) : x \in \mathfrak{X}, z \leq \pi(x)\}\right)$$

   then marginally $X \sim \pi(dx)$ (related to slice sampling)
3. Expansion using simulation variables: expand state-space to include random variables used in simulation algorithms that target $\pi(dx)$ or its conditionals; & then use the Hastings-Peskun framework
   - Boost efficiency (particles)
   - Widen applicability (intractable)

   Demonstrate the idea using two poplar algorithms: Multiple Try Metropolis, and Pseudo-marginal

Algorithm 1 Multiple-try Metropolis-Hastings algorithm

Initialisation: Choose X_0; Choose M; Choose N; Set $n = 0$

while $n < N + 1$ **do**

Sample $Y_{n+1}^{(m)} \sim Q(X_n, \cdot),\;\; m = 1, \dots, M$

Set $L_{n+1} = m$ with probability proportional to $w(X_n, Y_{n+1}^{(m)})$

Sample K uniformly in the set $\{1, \dots, M\}$

Set $X_{n+1}^{(K)} = X_n$

Sample $X_{n+1}^{(m)} \sim Q(Y_{n+1}^{(L_{n+1})}, \cdot),\;\; m \in \{1, 2, \dots, N\} - \{K\}$

Draw $U_{n+1} \sim U(0, 1)$

if $U_{n+1} < \alpha$ **then**

$X_{n+1} \leftarrow Y_{n+1}^{(L_{n+1})}$

else

$X_{n+1} \leftarrow X_n$

end if

$n \leftarrow n + 1$

end while

Define auxiliary expansion

$$\pi(dx, dy_*, \ell) = \pi(dx) \prod_{m=1}^{M} Q(x, dy^{(m)}) \frac{w(x, y^{(\ell)})}{\sum_m w(x, y^{(m)})}.$$

Consider now (among various alternatives) the proposed move:

$$(x, y_*, \ell) \to (y^{(\ell)}, X_*, K)$$

with $X_*^{(K)} = x$, $X_*^{(m)} \sim Q(y^{(\ell)}, \cdot)$ for $m \neq K$, and K drawn uniformly between 1 and M.

joint measure of current, say (x, y_*, ℓ), and proposed, say (y, x_*, k),

$$\pi(dx) \prod_k Q(x, dy^{(k)}) \frac{w(x, y^{(\ell)})}{\sum_n w(x, y^{(k)})} \times \delta_{y^{(\ell)}}(dy) \prod_{m \neq k} Q(y, dx^{(m)}) \delta_x(dx^{(k)}) \frac{1}{M}$$

Hence:

$$t((x, y_*, \ell), (y, x_*, k)) = \frac{\pi(dx) Q(x, dy)}{\pi(dy) Q(y, dx)} \frac{w(x, y)}{w(y, x)} \frac{\sum_m w(y, x^{(m)})}{\sum_m w(x^{(m)}, y)}$$

Pseudo-marginal algorithm

$$\pi(dx) = \kappa\pi_u(dx) = \kappa\pi_u(x)\nu(dx)$$

- κ is a normalising constant
- $\nu(dx)$ a dominating measure

Assume

$$\exists\ h(x,z) \geq 0\,\forall x, z, \quad \int h(x,z)q_x(dz) = \pi_u(x)$$

Then, auxiliary expansion

$$\pi(dx, dz) = h(x,z)q_x(dz)\nu(dx)\,,$$

Apply the Hastings-Peskun machinery; e.g. (among others)

$$(x, z) \to (Y, W)\,, \quad Y \sim Q(x, dy)\,, W|Y = y \sim q_y(dw)$$

Thus, joint measure of current and proposed state is:

$$h(x, z)q_x(dz)\nu(dx) \times Q(x, dy)q_y(dw)\,,$$

hence:

$$t(x, y) = \frac{h(x, z)\nu(dx)Q(x, dy)}{h(y, w)\nu(dy)Q(y, dx)} = \frac{\hat{\pi}(dx)Q(x, dy)}{\hat{\pi}(dy)Q(y, dx)}\,.$$

The name originates from a particular instance of this framework, with target $\overline{\pi}(d\theta)$ as a marginal to $\overline{\pi}(d\theta, dx) = \overline{\pi}(\theta)\nu(d\theta)\overline{\pi}_\theta(dx)$

If

$$\frac{\overline{\pi}_\theta(dx)}{q_\theta(dx)} = h_0(\theta, x)$$

$$h(\theta, Z) := \frac{1}{M}\sum_{m=1}^{M} h_0(\theta, Z^{(m)}) =: \widehat{\overline{\pi}}(\theta)\,, \quad Z = (Z^{(1)}, \ldots, Z^{(M)})$$

is a positive unbiased estimator of $\overline{\pi}(\theta)$, provided the $Z^{(m)}$ are *marginally* drawn from $q_\theta(\cdot)$, and resultant ratio is

$$t((\theta, z), (\phi, w)) = \frac{\widehat{\overline{\pi}}(d\theta)Q(\theta, d\phi)}{\widehat{\overline{\pi}}(d\phi)Q(\phi, d\theta)}$$

Meta-algorithms

Algorithms built on top of simpler, potentially not too efficient in isolation algorithms for sampling π. (weak learners)

New generation of the data augmentation paradigm within MCMC and allows the MCMC toolbox, while using the same classical Hastings-Peskun framework, to incorporate important developments in other areas of Monte Carlo, such as particle filters[9], exact simulation of stochastic processes [10] or Bernoulli factories [11].

---

[9] Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(3):269–342

[10] Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(3):333–382. With discussions and a reply by the authors

[11] Łatuszyński, K., Kosmidis, I., Papaspiliopoulos, O., and Roberts, G. O. (2011). Simulating events of unknown probabilities via reverse time martingales. *Random Structures Algorithms*, 38(4):441–452

Outline

1. Hastings-Peskun framework
2. Hypo-dimensional MCMC
3. Auxiliary expansions
4. Exact population MCMC

Adaptive direction sampling

This is another instance of hypo-dimensional MCMC and of Generalised Gibbs Sampling (although, again, not being conceived like this in the past)

Possibilities for building adaptation in MCMC while perserving Markovianity

Aim: sample from π on $\mathbb{R}^d$

Ingredients: set of active particles $X_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. Target instead

$$\pi(\mathbf{x}_1)\pi(\mathbf{x}_2)\ldots\pi(\mathbf{x}_k)\,.$$

p: dimensionality of proposed move

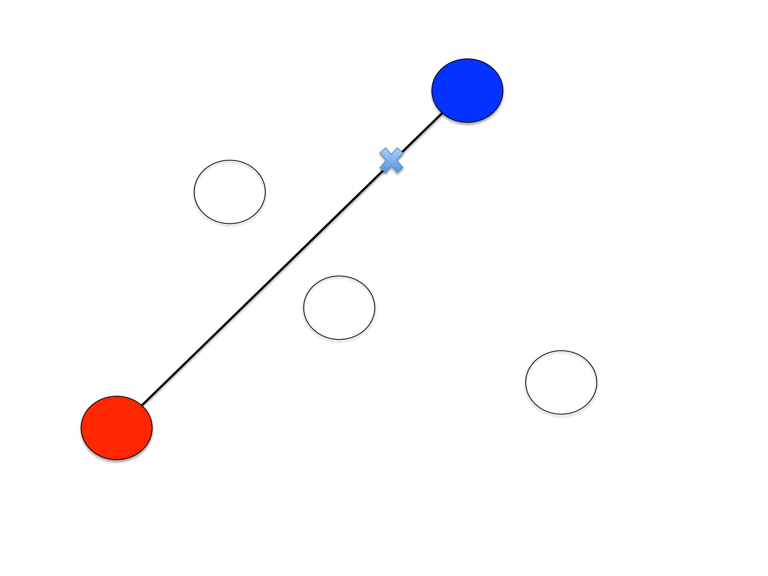
Algorithm 2 Adaptive Direction Sampling (ADS)

Initialisation: Choose X_0; Choose N; Set $n = 0$
while $n < N + 1$ **do**
Choose $\mathbf{x}_n^{(c)}$ uniformly at random from X_n. Let $C_n = X_n - \{\mathbf{x}_n^{(c)}\}$
Generate $\mathbf{a}$ from $D_v(C_n)$, B from $D_M(C_n)$
Sample $\mathbf{u} \in \mathbb{R}^p$ according to the density

$$\pi_{\mathbf{x}_n^{(c)},\mathbf{b},A}(\mathbf{u}) \propto \pi(\mathbf{x}_n^{(c)}(1 + \mathbf{a}^T\mathbf{u}) + B\mathbf{u})|1 + \mathbf{a}^T\mathbf{u}|^{d-p}$$

Let $\mathbf{y}_n = \mathbf{x}_n^{(c)}(1 + \mathbf{a}^T\mathbf{u}) + B\mathbf{u}$
$X_{n+1} = X_n - \{\mathbf{x}_n^{(c)}\} \cup \{\mathbf{y}_n\}$
$n \leftarrow n + 1$
end while

An example: the snooker algorithm (p=1)



Justification: (population) Generalised Gibbs sampling

The algorithm is based on a *non-commutative* group structure on $\mathbb{R}^p$ with identity element $\mathbf{0}$ and group operation

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u} + \mathbf{v} + \mathbf{a}^T \mathbf{u}\, \mathbf{v}$$

with

$$\mathbf{u}^{-1} = -\frac{1}{1 + \mathbf{a}^T \mathbf{u}} \mathbf{u}\,.$$

Notice that the group structure does not depend on B chosen as part of the algorithm. Finally, it is straightforward to check that

$$m_L(\mathbf{u}) = |1 + \mathbf{a}^T \mathbf{u}|^{-p}$$

Outline

1. Hastings-Peskun framework
2. Hypo-dimensional MCMC
3. Auxiliary expansions
4. Exact population MCMC

Summary

Highlighted 3 basic principles that underly the vast majority of developments in MCMC

- Hastings-Peskun framework: classic but increasingly relevant in exchanging statistical for computational efficiency
- Hypo-dimensional MCMC: deterministic moves in higher-dimensional spaces
- Common framework for developing and justifying algorithms: auxiliary expansions