Non-reversible Metropolis-Hastings

Joris Bierkens

DynStoch, Warwick University, 10 September 2014

`http://jbierkens.nl`

SNN Adaptive Intelligence

Radboud University

Outline

- Motivation
- Theory of non-reversible Metropolis-Hastings
- Example
- Application to spin systems (work in progress)
- General state spaces (work in progress)

Monte Carlo methods

Let $\mu = \frac{\pi}{Z}$ a probability distribution, with $\pi : S \to \mathbb{R}$ known and normalization constant Z possibly unknown.

Examples

- Gibbs density $\mu(x) \propto \exp(-\beta H(x))$ for a Hamiltonian H and inverse temperature β;
- Bayesian posterior $\mu(\theta) \propto \prod_{i=1}^N f(x_i|\theta)\pi_0(\theta)$ for observations $(x_i)_{i=1}^N$ and prior distribution π_0.

Goal

Compute $\mathbb{E}_\mu\left[\varphi(X)\right] = \int_S \varphi(x)\, d\mu(x)$

Monte Carlo method

- Obtain samples $(X_1, \ldots, X_K)$ from the distribution μ
- Estimate $\int \varphi(x)\, d\mu \approx \frac{1}{K}\sum_{k=1}^K \varphi(x_k)$

Markov Chain Monte Carlo

Markov Chain Monte Carlo method

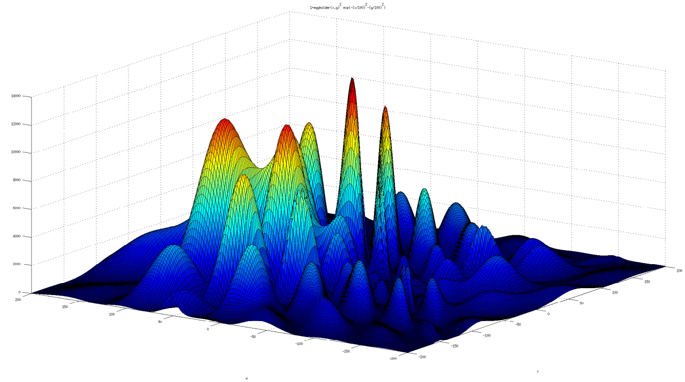
- Construct a Markov chain with transition matrix P that has μ as its invariant distribution.
- Obtain a sample path $(X_1, \ldots, X_K)$ of P
- Estimate

$$\mathbb{E}_\mu\left[\varphi(X)\right] \approx \frac{1}{K}\sum_{k=1}^{K}\varphi(X_k).$$

Examples

Metropolis-Hastings, Gibbs sampling, Glauber dynamics

The challenge



Reversibility

A Markov chain with transition density $p(x, y)$ is reversible with respect to $\pi(x)$ if

$$\pi(x)p(x, y) = p(y, x)\pi(y) \quad \forall x, y.$$

Other terminology: “satisfies detailed balance”, “symmetrizable”.

Symmetrizable

Let $Pf(x) = \int p(x, y)f(y)\ dy$ and $(f, g)_\pi := \int f(x)g(x)\pi(x)$. Then

$$\text{reversibility} \quad \Leftrightarrow \quad P = P^\star.$$

- Key in correctness proof of Metropolis-Hastings.

Reversibility

A Markov chain with transition density $p(x, y)$ is reversible with respect to $\pi(x)$ if

$$\pi(x)p(x, y) = p(y, x)\pi(y) \quad \forall x, y.$$

Other terminology: "satisfies detailed balance", "symmetrizable".

Symmetrizable

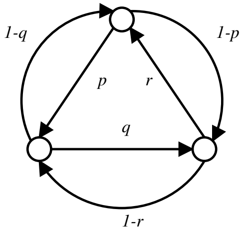
Let $Pf(x) = \int p(x, y)f(y)\, dy$ and $(f, g)_\pi := \int f(x)g(x)\pi(x)$. Then

$$\text{reversibility} \quad \Leftrightarrow \quad P = P^\star.$$

- Key in correctness proof of Metropolis-Hastings.

Non-reversible processes are better!

Example



- Transition matrix $P = \begin{pmatrix} 0 & p & 1-p \\ 1-q & 0 & q \\ r & 1-r & 0 \end{pmatrix}$.
- Choose p, q and r such that $(\frac{2}{5}, \frac{2}{5}, \frac{1}{5})$ is invariant distribution.
- The resulting transition matrix is

$$P = \begin{pmatrix} 0 & \frac{3}{4} + \frac{1}{2}\gamma & \frac{1}{4} - \frac{1}{2}\gamma \\ \frac{3}{4} - \frac{1}{2}\gamma & 0 & \frac{1}{4} + \frac{1}{2}\gamma \\ \frac{1}{2} + \gamma & \frac{1}{2} - \gamma & 0 \end{pmatrix}.$$

Example, continued

$$P = \begin{pmatrix} 0 & \frac{3}{4} + \frac{1}{2}\gamma & \frac{1}{4} - \frac{1}{2}\gamma \\ \frac{3}{4} - \frac{1}{2}\gamma & 0 & \frac{1}{4} + \frac{1}{2}\gamma \\ \frac{1}{2} + \gamma & \frac{1}{2} - \gamma & 0 \end{pmatrix}.$$

Spectral gap: $1 - \max(|\lambda_-|, |\lambda_+|)$

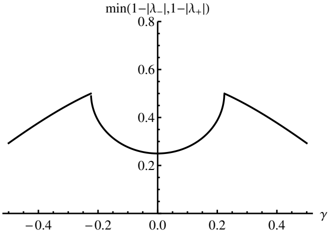


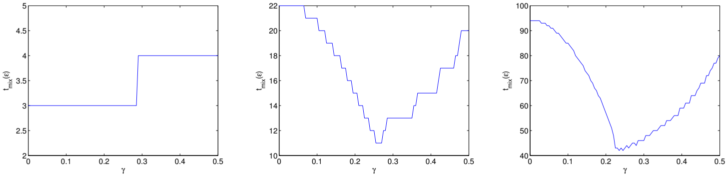
Figure : Spectral gap as a function of γ

Difficult to relate to notion of mixing time in non-reversible case

Example, continued

Mixing time

- Total variation distance: $||\mu - \nu||_{\text{TV}} := \max_{A \subset S} |\mu(A) - \nu(A)| = \frac{1}{2} \sum_{x \in S} |\mu(x) - \nu(x)|$.
- Define $d(t) := \max_x ||P^t(x, \cdot) - \mu||$, where μ is invariant for P
- Mixing time: $t_{\text{mix}}(\varepsilon) := \inf\{t \geq 0 : d(t) < \varepsilon\}$.



(a) $\varepsilon = 0.25$ (b) $\varepsilon = 10^{-3}$ (c) $\varepsilon = 10^{-12}$

Asymptotic variance

$$Y_T = \frac{1}{T}\sum_{t=1}^{T}\varphi(X_t) - \mathbb{E}_\pi\varphi,$$

Asymptotic variance

$$\sigma_\varphi^2 = \lim_{T\to\infty} T\mathbb{E}_x\left[Y_T^2\right]$$

Theorem

Let P be a Markov transition matrix.
Let K be its self-adjoint part with respect to $(\cdot,\cdot)_\pi$.
Then $\sigma^2_{\varphi,K} \geq \sigma^2_{\varphi,P}$ and there exists a φ for which strict inequality holds if $P \neq K$.

Non-reversible Metropolis-Hastings

[JB, Non-reversible Metropolis-Hastings, 2014]

Target distribution π.

Lemma

Let $P \in \mathbb{R}^{n \times n}$ Markov transition matrix. Define

$$\Gamma(x,y) = \pi(x)P(x,y) - \pi(y)P(y,x). \tag{1}$$

(i) Γ is skew-symmetric.

(ii) π is invariant for P iff $\sum_y \Gamma(x,y) = 0$ for all x

(iii) P is reversible w.r.t. π iff $\Gamma \equiv 0$.

Idea

- Let Γ be a matrix satisfying (i) and (ii)
- Construct a Markov chain P such that (1) holds.

Non-reversible Metropolis-Hastings

[JB, Non-reversible Metropolis-Hastings, 2014]

Ingredients

- Target distribution π.
- Γ satisfying
  - (i) Γ is skew-symmetric.
  - (ii) $\sum_y \Gamma(x, y) = 0$ for all x
- Proposal chain Q

Non-reversible Metropolis-Hastings

- Propose state y according to $Q(x, \cdot)$
- Accept with probability $A(x, y) = \min\left(1, \frac{\Gamma(x,y)+\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right)$

Resulting chain P satisfies $\Gamma(x, y) = \pi(x)P(x, y) - \pi(y)P(y, x)$.

Therefore π is invariant for P!

Non-reversible Metropolis-Hastings

Ingredients

π, Γ skew-symmetric with zero row sums, Q

Non-reversible Metropolis-Hastings

Propose y according to $Q(x,\cdot)$, accept with probability
$A(x,y) = \min\left(1, \frac{\Gamma(x,y)+\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right)$

Claim: $\Gamma(x,y) = \pi(x)P(x,y) - \pi(y)P(y,x)$

Proof: Suppose $\frac{\Gamma(x,y)+\pi(y)Q(y,x)}{\pi(x)Q(x,y)} > 1$. Rearranging gives

$$\begin{aligned}\Gamma(x,y)+\pi(y)Q(y,x) > \pi(x)Q(x,y) &\Leftrightarrow \pi(y)Q(y,x) > -\Gamma(x,y)+\pi(x)Q(x,y)\\ &\Leftrightarrow \pi(y)Q(y,x) > \Gamma(y,x)+\pi(x)Q(x,y)\end{aligned}$$

Remarks on NRMH

NRMH can construct 'all' Markov chains

Markov chain Q, with invariant distribution π and vorticity matrix

$$\Gamma(x,y) = \pi(x)Q(x,y) - \pi(y)Q(y,x).$$

With Q as proposal chain,

$$A(x,y) = \min\left(1, \frac{\Gamma(x,y) + \pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right) = 1.$$

Compatibility requirement

$$A(x,y) = \min\left(1, \frac{\Gamma(x,y) + \pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right)$$

Require $A \geq 0$. In particular

$$\Gamma(x,y) = 0 \quad \text{whenever} \quad Q(x,y) = 0.$$

Vorticity matrices

Essential in non-reversible Metropolis-Hastings: matrices $\Gamma \in \mathbb{R}^{n\times n}$ such that (i) $\Gamma = -\Gamma^T$, (ii) $\Gamma \mathbb{1} = 0$.

Lemma

(a) Let $u, v \in \mathbb{R}^n$ satisfy $u \perp v$ and $u, v \perp \mathbb{1}$. Then $\Gamma_{u,v} := uv^T - vu^T$ satisfies (i), (ii).

(b) Let $u_1, u_2, \ldots, u_{n-1}$ be an orthonormal base of $\mathbb{1}^\perp$ in $\mathbb{R}^n$ and write $\Gamma_{i,j} := \Gamma_{u_i,u_j} = u_i u_j^T - u_j u_i^T$. Then $\Gamma_{i,j} \perp \Gamma_{k,l}$ whenever $\{i,j\} \neq \{k,l\}$.

Corollary

$\{\Gamma_{i,j} : i = 1, \ldots, n-1, j = 1, \ldots, i-1\}$ is an orthonormal base of $\mathcal{V}$, so $|\mathcal{V}| = \frac{1}{2}(n-1)(n-2)$.

Compatibility

Graph $G = (S, E)$; edges represent positive transition probabilities in Q.

(i) $\Gamma = -\Gamma^T$.

(ii) $\Gamma \mathbb{1} = 0$, i.e. $\sum_{j=1}^n \Gamma(i,j) = 0$ for all $i = 1, \ldots, n$.

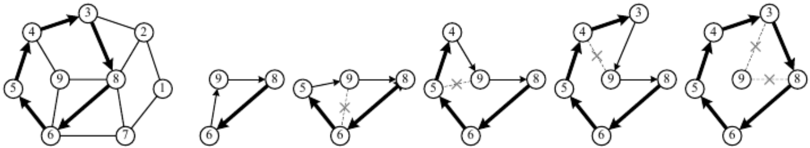
(iii) Compatibility: $\Gamma(i,j) = 0$ whenever (i,j) is not an edge.

Proposition Let $x, y \in S$

Γ satisfying (i) - (iii) exists and $\Gamma(x,y) > 0$
$\Leftrightarrow$ G contains a cycle with (x,y) as an edge.

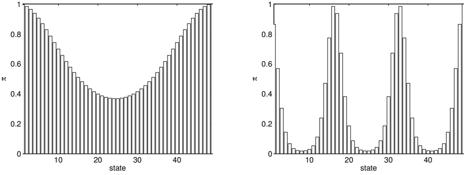
Cycle calculus

Image: [Sun, Gomez, Schmidhuber]



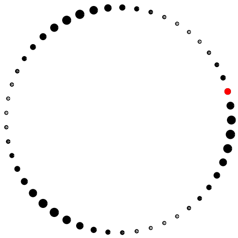
Example: n-cycle

$$Q = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \dots & \dots & 0 & \frac{1}{2} \\ \frac{1}{2} & \ddots & \ddots & \ddots & & & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & & & \ddots & \ddots & \ddots & \frac{1}{2} \\ \frac{1}{2} & 0 & \dots & \dots & 0 & \frac{1}{2} & 0 \end{pmatrix} \qquad \Gamma = \begin{pmatrix} 0 & 1 & 0 & \dots & \dots & 0 & -1 \\ -1 & \ddots & \ddots & \ddots & & & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & & & \ddots & \ddots & \ddots & 1 \\ 1 & 0 & \dots & \dots & 0 & -1 & 0 \end{pmatrix}$$

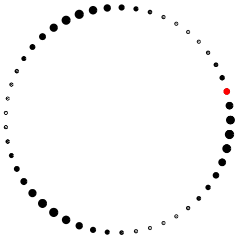


(a) $M = 1$, $\beta = 1$. (b) $M = 3$, $\beta = 4$.

$$M = 3, \quad \beta = 4.$$



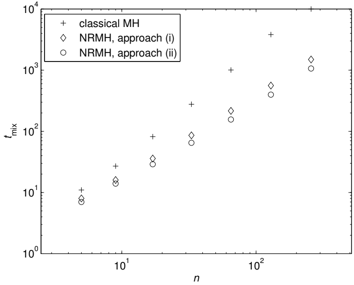
(a) Classical Metropolis-Hastings



(b) Non-reversible Metropolis-Hastings

Numerical results

M	β	spectral gap	NRMH	MH	mixing time	NRMH	MH
0	0		0.00814	0.00205		116	456
1	2		0.0132	0.00907		92	164
1	4		0.0205	0.0122		100	159
2	2		0.0141	0.00248		83	310
2	4		0.00703	0.000598		176	1189
3	2		0.0125	0.00375		91	275
3	4		0.00592	0.000943		188	1055



Example: Spin systems

Fundamental model in statistical physics, theoretical neuroscience and machine learning

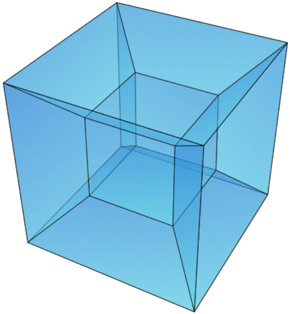
- $G = (V, E)$ a finite graph
- $w : E \to \mathbb{R}$ interaction between vertices
- $h : V \to \mathbb{R}$ external field
- $S = \{+, -\}^V$ set of possible spin configurations (state space)
- $H : S \to \mathbb{R}$ energy function

$$H(x) = -\sum_{v_1 v_2 \in E} w(v_1 v_2) x(v_1) x(v_2) - \sum_{v \in V} h(v) \sigma(v), \quad x \in S,$$

- β 'inverse temperature'
- $\mu_\beta(x) = \exp(-\beta H(x))/Z$ *Boltzmann distribution*

MCMC for spin systems

- State space $S = \{+, -\}^n$.
- Markov chain on S: flipping one bit at a time.
- Corresponds to Markov chain on the n-dim. hypercube



- Proposal chain Q: random walk on hypercube.

Compatible vorticity matrices for hypercube

Lemma

The dimension a_n of space of compatible vorticity matrices for n-dimensional hypercube satisfies

$$a_{n+1} = 2a_n + \left(2^n - 1\right), \quad a_1 = 0,$$

with solution $a_n = 1 + (\frac{1}{2}n - 1)2^n$.

Examples

- Every face of the hypercube
- Hamiltonian circuit (Gray code)
- For $A \in \mathbb{R}^{n \times n}$ skew-adjoint,

$$\Gamma_A(x,y) = \begin{cases} x_i \sum_{j=1}^n a_{ij} x_j & \text{if } y \text{ equals } x \text{ with bit } i \text{ flipped,} \\ 0 & \text{otherwise.} \end{cases}$$

A long story short

Recall

$$A(x,y) = \min\left(1, \frac{\Gamma(x,y) + \pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right)$$

For given proposal chain Q, target distribution π, and compatible vorticity matrix Γ_0, for what range of γ is $\Gamma = \gamma\Gamma_0$ suitable?

- Some (technical) results in estimating this range.
- Only modest improvements in mixing time so far.
- **what is the effect of ‘vorticity’ on mixing time?**

Estimating mixing time

- Very limited results on mixing time for (classical) Metropolis-Hastings [Diaconis, Saloff-Coste, 1998]
- Poincaré inequality: Does not capture improvement over reversible chain
- [James Fill (1991)]: Does not capture improvement over reversible chain

Estimating mixing time

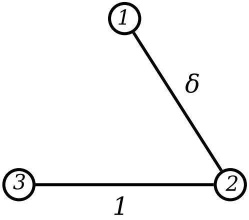
- Very limited results on mixing time for (classical) Metropolis-Hastings [Diaconis, Saloff-Coste, 1998]
- Poincaré inequality: Does not capture improvement over reversible chain
- [James Fill (1991)]: Does not capture improvement over reversible chain

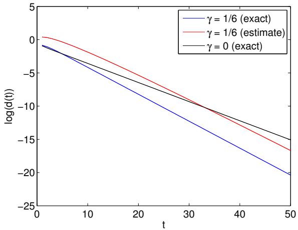
Estimating mixing time

- Very limited results on mixing time for (classical) Metropolis-Hastings [Diaconis, Saloff-Coste, 1998]
- Poincaré inequality: Does not capture improvement over reversible chain
- [James Fill (1991)]: Does not capture improvement over reversible chain

Estimating mixing time

- Very limited results on mixing time for (classical) Metropolis-Hastings [Diaconis, Saloff-Coste, 1998]
- Poincaré inequality: Does not capture improvement over reversible chain
- [James Fill (1991)]: Does not capture improvement over reversible chain
- Path coupling / optimal transport / discrete Ricci curvature:





Overview

- Non-reversible chains are better (in some sense)...
- ... but so far Metropolis-Hastings created reversible chains.
- **Non-reversible Metropolis-Hastings removes this limitation**

Overview

- Non-reversible chains are better (in some sense)...
- ... but so far Metropolis-Hastings created reversible chains.
- **Non-reversible Metropolis-Hastings removes this limitation**

Use:

- If you have a good (fast mixing) non-reversible chain, use it as proposal chain in NRMH

Overview

- Non-reversible chains are better (in some sense)...
- ... but so far Metropolis-Hastings created reversible chains.
- **Non-reversible Metropolis-Hastings removes this limitation**

Use:

- If you have a good (fast mixing) non-reversible chain, use it as proposal chain in NRMH

Main challenge:

- understanding mixing time for non-reversible chains

Overview

- Non-reversible chains are better (in some sense)...
- ... but so far Metropolis-Hastings created reversible chains.
- **Non-reversible Metropolis-Hastings removes this limitation**

Use:

- If you have a good (fast mixing) non-reversible chain, use it as proposal chain in NRMH

Main challenge:

- understanding mixing time for non-reversible chains

END

Vorticity measures on general state spaces

- $(S, \mathscr{S})$ measurable space.
- $P(x, dy)$ Markov transition kernel with invariant distribution π
- Forward $F_P(dx, dy) := \pi(dx)P(x, dy)$ and backward $B_P(dx, dy) = \pi(dy)P(y, dx)$ ergodic flow

Vorticity measure

$$\Gamma(dx, dy) = F_P(dx, dy) - B_P(dx, dy).$$

Then Γ is a signed measure on $S \times S$, satisfying

- $\Gamma(A \times B) = -\Gamma(B \times A)$ for all $A, B \in \mathscr{S}$,
- $\Gamma(A, S) = 0$ for all $A \in \mathscr{S}$.

Non-reversible Metropolis-Hastings in general spaces

Let Γ be a signed measure on $S \times S$, satisfying

- $\Gamma(A \times B) = -\Gamma(B \times A)$ for all $A, B \in \mathscr{S}$,
- $\Gamma(A, S) = 0$ for all $A \in \mathscr{S}$.

Let

- $Q(x, dy)$ be a proposal chain,
- $F_Q(dx, dy) = \pi(dx) Q(x, dy)$,
- $B_Q(dx, dy) = \pi(dy) Q(y, dx)$.
- Symmetric structure: F_Q and B_Q equivalent (i.e. mutually absolutely continuous)

Hastings Ratio

$$R(x, y) := \frac{d\Gamma}{dF_Q}(x, y) + \frac{dB_Q}{dF_Q}(x, y).$$

Acceptance probability

$$A(x, y) := \min(1, R(x, y)).$$

General state spaces; absolutely continuous case

- Proposal chain $Q(x, dy) = q(x, y)\lambda(dy)$, where λ is some reference measure.
- Target distribution $\pi(dx) = \rho(x)\ d\lambda(x)$
- Symmetric structure: $\rho(x)q(x, y) = 0 \Leftrightarrow \rho(y)q(y, x) = 0$
- $\gamma : S \times S \to \mathbb{R}$, satisfying
  - $\gamma(x, y) = -\gamma(y, x)$
  - $\int_{A\times S} \gamma(x, y)\ \lambda(dx)\ \lambda(dy) = 0$ for all $A \in \mathscr{S}$.
  - $\gamma(x, y) = 0$ whenever $\rho(x)q(x, y) = 0$.
- Hastings ratio:

$$R(x, y) = \begin{cases} \frac{\gamma(x,y)+\rho(y)q(y,x)}{\rho(x)q(x,y)}, & \rho(x)q(x, y) \neq 0, \\ 1, & \rho(x)q(x, y) = 0. \end{cases}$$

Example: Ornstein Uhlenbeck process

$$dX(t) = AX(t)\ dt + B\ dW(t).$$

- Reversible if and only if $BB^T A^T = ABB^T$
- Invariant distribution covariance satisfies $AQ_\infty + Q_\infty A^T = -BB^T$
- Wieldy expression available for vorticity density
- To do: Relate to Lelièvre, Nier, Pavliotis

Convergence to equilibrium

Different quantifications:

- Let
$$d(t) := \max_x ||P^t(x,\cdot) - \mu(\cdot)||_{\mathrm{TV}}.$$
The ε-mixing time is $\inf\{t \geq 0 : d(t) \leq \varepsilon\}$.
- spectral gap: $1 - \max\{|\lambda| : \lambda \in \sigma(P), \lambda \neq 1\}$
- asymptotic variance:
$$\sigma^2(\varphi) := \lim_{T\to\infty} T \operatorname{var}\left(\frac{1}{T}\sum_{t=1}^{T} \varphi(X_t)\right).$$