MCMC Methods for Multiscale Measures

Andrew Stuart[1]

[1]Mathematics Institute and
Centre for Scientific Computing
University of Warwick

INI-HOP
Cambridge, July 2^{th}–6^{th}

Collaboration with
A. Beskos, M. Hairer, G. Roberts and J. Voss (Warwick),
D. White (Warwick) and F. Pinski (Cincinnati)

Funded by EPSRC, ONR

WARWICK

Outline

Outline

Introductory Example

- Sample the pdf

$$\pi(x) \propto \exp(-x^2/2)\mathbb{I}_{(3,4)}(x).$$

- Use MCMC method.
- True density in green, empirical histogram in red.

Metropolis-Hastings Algorithm

- Metropolis (1953), Hastings (1970).
- Propose move $x^{(j)} \longrightarrow y$ (using Markov kernel).
- Accept move ($x^{(j+1)} = y$) with probability $\alpha(x^{(j)}, y)$.
- Reject move ($x^{(j+1)} = x^{(j)}$) with probability $1 - \alpha(x^{(j)}, y)$.
- Generates a new Markov chain $\{x^{(j)}\}_{j=0}^{J-1}$.
- If Markov chain is ergodic then its histogram approaches π.
- Optimize proposal (within a class) to minimize cost.
- IDEAS FROM NUMERICAL ANALYSIS ARE KEY

Outline

1. Bridge Path Sampling

- **Find** $x(u) \in \mathcal{H} = L^2([0,1], \mathbb{R}^d)$:

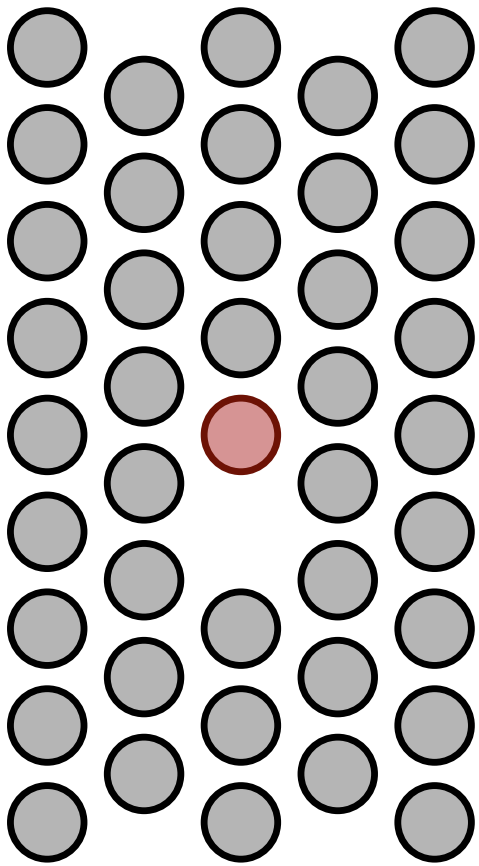
$$\frac{dx}{du} = f(x) + \gamma \frac{dB}{du}.$$

- **Given**

$$x(0) = X^- \quad \& \quad x(1) = X^+.$$

  - $\frac{dB}{du}$ is Gaussian white noise.
  - Randomness of B induces randomness in x.
  - Need to sample a measure on H.
  - Applications: **econometrics, rare event MD, Parisi-Wu**.

2. Vacancy Diffusion



Crystal lattice with vacancy.

3. Signal Processing

- **Find** $x(u) \in \mathcal{H} = L^2([0,1], \mathbb{R}^d)$:

$$\frac{dx}{du} = f(x) + \gamma \frac{dB_1}{du}, \quad X(0) \sim \zeta.$$

- **Given** $y(u)$:

$$\frac{dy}{du} = g(x,y) + \sigma \frac{dB_2}{du}, \quad Y(0) = 0.$$

  - $\frac{dB_i}{du}$ denote Gaussian white noises.
  - Randomness of B_1, B_2 induces randomness in $x|y$.
  - Need to sample a measure on H.
  - Applications: **signal processing, data assimilation.**

4. Lagrangian Data Assimilation

- **Find** velocity field $x(u) \in \mathcal{H} = L^2(\mathbb{T}^d, \mathbb{R}^d)$, initial condition for $X(u, s)$:

$$\begin{aligned} \frac{\partial X}{\partial s} &= F(X), \\ X(u, 0) &= x(u) \sim \mathcal{N}(\bar{v}, \mathcal{C}). \end{aligned}$$

- **Given** noisy observations $\{z_{j,k}\}$ of Lagrangian tracers

$$\begin{aligned} \frac{dy_j}{ds} &= X(y_j, s), \quad j = 1, \ldots, J, \\ z_{j,k} &= y_j(k\delta) + \mathcal{N}(0, \Sigma_{j,k}) \\ z &= H(y) + \mathcal{N}(0, \Sigma). \end{aligned}$$

CK

Outline

Structure of the Target

- IID Product in $\mathbb{R}^n$

$$\pi(x) = \Pi_{i=1}^n f(x_i).$$

- Scaled Product in $\mathbb{R}^n$

$$\pi(x) = \Pi_{i=1}^n \frac{1}{\lambda_i} f\left(\frac{x_i}{\lambda_i}\right).$$

- Change of Measure in $\mathbb{R}^n$

$$\pi(x) = \exp\Big(-\Phi_n(x)\Big)\Pi_{i=1}^n \frac{1}{\lambda_i} f\left(\frac{x_i}{\lambda_i}\right).$$

- Change of Measure from Gaussian in $\mathcal{H}$

$$\pi(x) = \exp\left(-\Phi(x) + \frac{1}{2}\langle x, \mathcal{L}x\rangle\right).$$

Langevin Proposals

- The Langevin S(P)DE

$$\frac{dx}{dt} = \mathcal{A}\nabla \log \pi(x) + \sqrt{2\mathcal{A}}\frac{dW}{dt}$$

  has π as invariant for $\mathcal{A} = \mathcal{A}^*, \mathcal{A} > 0$.
- Exact solution $x \longrightarrow y$ with $x = x(0)$ and $y = x(\Delta t)$ would be a perfect proposal.
- Discretization gives a good family of candidate proposals.
- Optimize over form of discretization, form of $\mathcal{A}$ and choice of Δt.

Cost

- Choose largest scaling of $\Delta t(n)$ (size of move):

$$\liminf_{n\to\infty} \mathbb{E}\alpha > 0.$$

- Large Δt improves decorrelation in samples.
- Denote number of iterations for mixing by $M(n)$.
- Then $M(n) = \mathcal{O}(1/\Delta t(n))$.
- Denote by $I(n)$ cost of each step.
- Complexity is $C(n) = I(n) \times M(n)$.

IID Products

$$\pi(x) = \Pi_{i=1}^{n} f(x_i).$$

Proposal $$\frac{y-x}{\Delta t} = \beta \nabla \log \pi(x) + \sqrt{\frac{2}{\Delta t}}\xi, \quad \xi \sim \mathcal{N}(0, I).$$

Theorem 1. (Roberts et al 97, 98)

- $\beta = 0 : M(n) = \mathcal{O}(n)$.
- $\beta = 1 : M(n) = \mathcal{O}(n^{1/3})$.

Steepest Descents Impacts Complexity

Scaled Products

$$\pi(x) = \Pi_{i=1}^{n} \frac{1}{\lambda_i} f\left(\frac{x_i}{\lambda_i}\right).$$

Proposal $$\frac{y - x}{\Delta t} = \mathcal{A}\nabla \log \pi(x) + \sqrt{\frac{2\mathcal{A}}{\Delta t}}\xi, \quad \xi \sim \mathcal{N}(0, I).$$

Theorem 2. $\lambda_i = i^{-k}, k > 0$. Set $C = \text{diag}\{\lambda_i^2\}$.

- $\mathcal{A} = I : M(n) = \mathcal{O}(n^{2k+1/3})$.
- $\mathcal{A} = C : M(n) = \mathcal{O}(n^{1/3})$.

Preconditioning Impacts Complexity

Change of Measure

$$\pi(x) = \exp\Big(-\Phi_n(x)\Big)\Pi_{i=1}^n \frac{1}{\lambda_i} f\Big(\frac{x_i}{\lambda_i}\Big).$$

Proposal $$\frac{y-x}{\Delta t} = \mathcal{A}\nabla \log \pi(x) + \sqrt{\frac{2\mathcal{A}}{\Delta t}}\xi, \quad \xi \sim \mathcal{N}(0, I).$$

Theorem 3. $\lambda_i = i^{-k}, k > 0$. $C = \text{diag}\{\lambda_i^2\}$.

$$\sup_{n,x}\Big(|\Phi_n(x)| + |\nabla\Phi_n(x))|\Big) < \infty.$$

- $\mathcal{A} = I : M(n) = \mathcal{O}(n^{2k+1/3})$.
- $\mathcal{A} = C : M(n) = \mathcal{O}(n^{1/3})$.

Change of Meausure Does Not Affect Scalings

Change of Measure from Gaussian

$$\pi(x) = \exp\Big(-\Phi(x) + \frac{1}{2}\langle x, \mathcal{L}x\rangle\Big).$$

$$\frac{y-x}{\Delta t} = \mathcal{A}\Big(\theta\mathcal{L}y + (1-\theta)\mathcal{L}x\Big) - \mathcal{A}\nabla\Phi(x) + \sqrt{\frac{2\mathcal{A}}{\Delta t}}\xi, \quad \xi \sim \mathcal{N}(0, I).$$

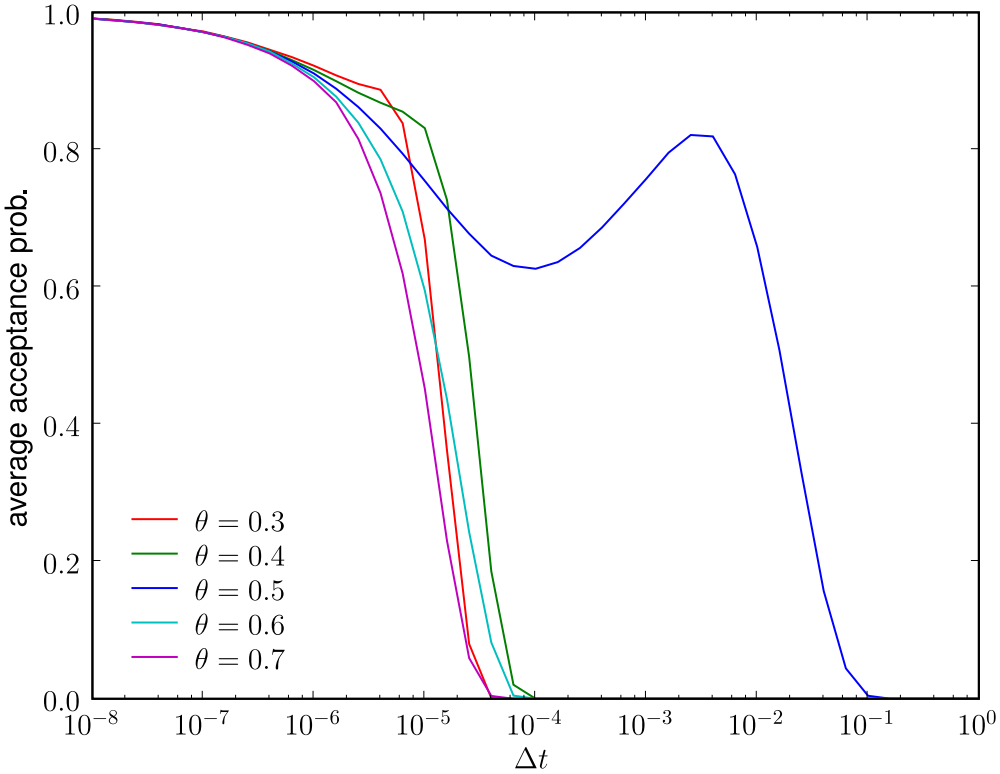
Theorem 4. $C = -\mathcal{L}^{-1}, \quad \sup_x\Big(|\Phi(x)| + |\nabla\Phi(x)|\Big) < \infty.$

- $\theta = \frac{1}{2}$ $\mathcal{A} = I, C : M(n) = \mathcal{O}(1).$
- $\theta \neq \frac{1}{2}$ $\mathcal{A} = C : M(n) = \mathcal{O}(n^{1/3}).$
- $\theta \neq \frac{1}{2}$ $\mathcal{A} = I : M(n) = \mathcal{O}(n^{2k+1/3}).$

Implicitness Impacts Complexity

WARWICK

Example. Bridge Sampling; d=1



Average acceptance probability in equilibrium for the double-well process ($n = 1000$).

WARWICK

Outline

What We Have Shown

We have shown that:

- **Applications:** Measures which have density with respect to a Gaussian arise naturally in many applications where the solution is a measure on functions.
- **Algorithms:** Using these SPDEs, MCMC methods can be constructed in function space.
- **SPDEs:** Langevin SPDEs form natural basis for MCMC propoals.
- **Numerical Analysis:** Ideas such as steepest descents, peconditioning and implicitness have crucial impact on complexity of MCMC algorithms.

What Remains Open

- **Algorithms:** optimize by methods such as blocking, low-rank approximation.
- **Applications:** are numerous in physics, data assimilation, signal processing and econometrics.
- **Evaluation:** of methods introduced here in comparison with other methods.
- **Analysis**: extend our theory to non-gradient SDEs, state-dependent noise, degenerate noise, non-Gaussian noise.
- **Numerical Analysis**: develop an approximation theory (boundary conditions, nonlinear Dirac sources, preserving symmetry of inverse covariance).
- **Statistics:** incoporate into (Gibbs) sampler to estimate parameters as well as functions.

References

- A.M.Stuart, P. Wiberg and J. Voss. ”Conditional Path Sampling of SDEs and the Langevin MCMC Method.” Comm. Math. Sci. 2(2004), 685–697.
- M. Hairer, A.M.Stuart, P. Wiberg and J. Voss. ”Analysis of SPDEs Arising in Path Sampling. Part 1: The Gaussian Case.” Comm. Math. Sci. 3(2005), 587–603
- M. Hairer, A.M.Stuart and J. Voss. ”Sampling the posterior: an approach to non-Gaussian data assimilation.” PhysicaD, **230**(2007), 50–64.

References (Continued)

- M. Hairer, A.M.Stuart and J. Voss. "Analysis of SPDEs Arising in Path Sampling. Part 2: The Nonlinear Case." Ann. Appl. Prob., to appear.
- A. Beskos, G.O. Roberts, A.M. Stuart and J. Voss. "An MCMC Method for diffusion bridges." Submitted.
- A. Beskos and A.M. Stuart. "Scalings for local Metropolis-Hastings chains on non-product targets." In preparation.
- For all papers see:

  http : *//www.maths.warwick.ac.uk/* ∼ *stuart/sample.html*

References

- A. Gelman, W.R. Gilks and G.O. Roberts, *Weak convergence and optimal scaling of random walk Metropolis algorithms*. Ann. Appl. Prob. **7**(1997), 110–120.
- G.O. Roberts and J. Rosenthal, *Optimal scaling of discrete approximations to Langevin diffusions*. JRSSB **60**, 255–268.
- G.O. Roberts and J. Rosenthal, *Optimal scaling for various Metropolis-Hastings algorithms*. Statistical Science **16**, 351–367.

Langevin SPDE – Example

- Recall the Signal Processing case:

$$\frac{dx}{du} = f(x) + \frac{dB_1}{du}, \quad X(0) \sim \zeta$$
$$\frac{dy}{du} = g(x, y) + \sigma \frac{dB_2}{du}, \quad Y(0) = 0.$$

- We want to sample the distribution on paths x, given a single path y.
- x is is the signal. y is the observation.
- u parameterizes the path.

Langevin SPDE – Example

- The SPDE for $x(u,t)$ is:

$$\begin{aligned}\frac{\partial x}{\partial t} =& \frac{\partial^2 x}{\partial u^2} - (\nabla f(x) - \nabla f(x)^*)\frac{\partial x}{\partial u} - \nabla_x \Phi(x) + \sqrt{2}\frac{\partial W}{\partial t} \\ &+ dg(x,y)^T(\sigma\sigma^T)^{-1}\left(\frac{dy}{du} - g(x,y)\right) - \frac{1}{2}\nabla_x\left(\nabla_y \cdot g(x,y)\right) \\ \frac{\partial x}{\partial u} =& \left(f(x) - \nabla_x \ln \zeta(x)\right), \quad u = 0, \qquad \frac{\partial x}{\partial u} = f(x), \quad u = 1.\end{aligned}$$

- t is algorithm time.
- In statistical equilibrium ($t \gg 1$) $x(\cdot, t)$ samples signal x, given the observation $y(u)$.