Diffusions with position-dependent volatility and the Metropolis-adjusted Langevin algorithm

Sam Livingstone
With T. Xifara & C. Sherlock (Lancaster), S. Byrne & M. Girolami (UCL)

November 19, 2013
Research Interests

Practice
Develop general purpose methods to fit complex statistical models

\[ L(\alpha, \beta|y) = \int f(y|x, \beta)p(x|\alpha)dx \]

Theory
Understand what methods work well where and why

\[ \|P^n(x_0, \cdot) - \pi(\cdot)\|_{TV} \leq M(x_0)r^n \]
Goal of talk

- Diffusions as a basis for MCMC proposals
Goal of talk

- Diffusions as a basis for MCMC proposals
- Generalise a diffusion-based MCMC method
Goal of talk

- Diffusions as a basis for MCMC proposals
- Generalise a diffusion-based MCMC method
- Explain difference with literature
Goal of talk

- Diffusions as a basis for MCMC proposals
- Generalise a diffusion-based MCMC method
- Explain difference with literature
- Geometry & Stochastic processes
Monte Carlo

\[ \mathbb{E}_\pi[f(\theta)] = \int_{\Omega} f(\theta) \pi(\theta) d\theta \approx \frac{1}{n} \sum_i f(\theta_i), \quad \theta_i \sim \pi. \]
Monte Carlo

\[ \mathbb{E}_\pi[f(\theta)] = \int_\Omega f(\theta)\pi(\theta)d\theta \approx \frac{1}{n} \sum_i f(\theta_i), \quad \theta_i \sim \pi. \]

Only $c\pi$ known?

**Re-sampling**

- Draw from some distribution $q$
- Adjust samples to learn about $\pi$
Monte Carlo

\[ E_\pi[f(\theta)] = \int_\Omega f(\theta)\pi(\theta)d\theta \approx \frac{1}{n}\sum_i f(\theta_i), \ \theta_i \sim \pi. \]

Only \( c\pi \) known?

**Re-sampling**

- Draw from some distribution \( q \)
- Adjust samples to learn about \( \pi \)
- Key issue \( q \approx \pi \) (difficult in high dimensions)
Markov Chain Monte Carlo

- Draw dependent samples from $q$ (Markov chain)
Markov Chain Monte Carlo

- Draw dependent samples from $q$ (Markov chain)
- Easy to construct a chain which *converges* to $\pi$
Markov Chain Monte Carlo

- Draw dependent samples from $q$ (Markov chain)
- Easy to construct a chain which converges to $\pi$
- ‘local’ proposals remove difficulty in high dimensions
Markov Chain Monte Carlo

- Draw dependent samples from $q$ (Markov chain)
- Easy to construct a chain which converges to $\pi$
- ‘local’ proposals remove difficulty in high dimensions

At iteration $i$:

1. Draw $\theta' \sim q(\cdot|\theta^{(i-1)})$
2. Set

$$
\theta^{(i)} = \begin{cases} 
\theta' & \text{w.p. } \alpha = \min \left(1, \frac{\pi(\theta')q(\theta'|\theta')}{\pi(\theta)q(\theta'|\theta)} \right) \\
\theta^{(i-1)} & \text{w.p. } 1 - \alpha 
\end{cases}
$$
Markov Chain Monte Carlo

- Draw dependent samples from $q$ (Markov chain)
- Easy to construct a chain which converges to $\pi$
- ‘local’ proposals remove difficulty in high dimensions

At iteration $i$:

1. Draw $\theta' \sim q(.|\theta^{(i-1)})$
2. Set

$$
\theta^{(i)} = \begin{cases} 
\theta' & \text{w.p. } \alpha = \min \left(1, \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)} \right) \\
\theta^{(i-1)} & \text{w.p. } 1 - \alpha
\end{cases}
$$

Good choice for $q(.|\theta^{(i-1)})$?
A continuous time Markov process \((X_t)_{t \geq 0}\) with (almost surely) continuous sample paths.
Diffusions

- A continuous time Markov process $(X_t)_{t \geq 0}$ with (almost surely) continuous sample paths
- For any fixed $t$, $X_t$ is a random variable
Diffusions

- A *continuous time Markov process* $(X_t)_{t \geq 0}$ *with (almost surely) continuous sample paths*
- For any fixed $t$, $X_t$ is a random variable
- Realisation: a draw from a distribution over paths, where $\text{Prob}(\text{discontinuous path}) = 0$
Diffusions

- A continuous time Markov process \((X_t)_{t \geq 0}\) with (almost surely) continuous sample paths
- For any fixed \(t\), \(X_t\) is a random variable
- Realisation: a draw from a distribution over paths, where \(\text{Prob}(\text{discontinuous path}) = 0\)

\[
dX_t = b(X_t) \, dt + \sigma(X_t) \, dB_t, \quad X_0 = x_0,
\]

BM: Gaussian process with independent increments.

\[
B_t + h - B_t \sim N(0, hI)
\]
Diffusions

- A continuous time Markov process \((X_t)_{t \geq 0}\) with (almost surely) continuous sample paths
- For any fixed \(t\), \(X_t\) is a random variable
- Realisation: a draw from a distribution over paths, where \(\text{Prob}(\text{discontinuous path}) = 0\)

\[
dX_t = b(X_t) \, dt + \sigma(X_t) \, dB_t, \quad X_0 = x_0,
\]

\[
X_{t+h} \approx X_t + b(X_t)h + \sigma(X_t)\mathcal{N}(0, hI)
\]
Diffusions

- A continuous time Markov process $(X_t)_{t \geq 0}$ with (almost surely) continuous sample paths.
- For any fixed $t$, $X_t$ is a random variable.
- Realisation: a draw from a distribution over paths, where $\text{Prob}(\text{discontinuous path}) = 0$

\[
dX_t = b(X_t) \, dt + \sigma(X_t) \, dB_t, \quad X_0 = x_0,
\]

\[
X_{t+h} \approx X_t + b(X_t)h + \sigma(X_t)N(0, hI)
\]

- BM: Gaussian process with independent increments.
\[
B_{t+h} - B_t \sim \mathcal{N}(0, hI)
\]
Densities

- Fokker-Planck equation:

\[
\frac{\partial}{\partial t} u(x, t) = - \sum_i \frac{\partial}{\partial x_i} [b_i(x) u(x, t)] + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} [V_{ij}(x) u(x, t)],
\]

\[V = \sigma \sigma^T, \quad u = \text{density of } X_t.\]
Densities

- Fokker-Planck equation:

\[
\frac{\partial}{\partial t} u(x, t) = - \sum_i \frac{\partial}{\partial x_i} [b_i(x) u(x, t)] + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} [V_{ij}(x) u(x, t)],
\]

\[V = \sigma \sigma^T, \; u = \text{density of } X_t.\]

- \(\partial u / \partial t = 0 \implies \text{Invariant distribution!}\)
Densities

- **Fokker-Planck equation:**

\[
\frac{\partial}{\partial t} u(x, t) = - \sum_i \frac{\partial}{\partial x_i} [b_i(x) u(x, t)] + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} [V_{ij}(x) u(x, t)],
\]

\(V = \sigma \sigma^T, \ u = \) density of \(X_t\).

- \(\partial u/\partial t = 0 \implies \) Invariant distribution!

\[
\implies X_t \sim \pi \implies X_{t+\tau} \sim \pi \ \forall \tau \geq 0
\]
Densities

- Fokker-Planck equation:

\[
\frac{\partial}{\partial t} u(x, t) = - \sum_i \frac{\partial}{\partial x_i} [b_i(x)u(x, t)] + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} [V_{ij}(x)u(x, t)],
\]

\(V = \sigma \sigma^T, \ u = \text{density of } X_t.\)

- \(\frac{\partial u}{\partial t} = 0 \Rightarrow \text{Invariant distribution!}\)
  - \(X_t \sim \pi \Rightarrow X_{t+\tau} \sim \pi \ \forall \tau \geq 0\)
  - Convergence to \(\pi\) under mild conditions
Densities

Fokker-Planck equation:

\[
\frac{\partial}{\partial t} u(x, t) = -\sum_i \frac{\partial}{\partial x_i} [b_i(x)u(x, t)] + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} [V_{ij}(x)u(x, t)],
\]

\(V = \sigma \sigma^T, \ u = \text{density of } X_t.\)

\(\frac{\partial u}{\partial t} = 0 \implies \text{Invariant distribution!}\)

\(X_t \sim \pi \implies X_{t+\tau} \sim \pi \ \forall \tau \geq 0\)

Convergence to \(\pi\) under mild conditions

\[
\sum_i \frac{\partial}{\partial x_i} [b_i(x)\pi(x)] = \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} [V_{ij}(x)\pi(x)]. \quad (1)
\]
Langevin Diffusion

Stochastic process that converges to $\pi$ by design.

$$dX_t = \frac{1}{2} \nabla \log \pi(X_t) dt + dB_t, \quad X_0 = x_0.$$
Langevin Diffusion

Stochastic process that converges to $\pi$ by design.

$$dX_t = \frac{1}{2} \nabla \log \pi(X_t) dt + dB_t, \quad X_0 = x_0.$$  

Substituting into (1):

$$b_i(x)\pi(x) = \frac{1}{2} \frac{\partial}{\partial x_i} [\log \pi(x)] \pi(x),$$

$$= \frac{1}{2} \frac{\partial}{\partial x_i} \pi(x),$$

$$= \frac{1}{2} \sum_j \frac{\partial}{\partial x_j} [V_{ij}(x)\pi(x)].$$
Metropolis-adjusted Langevin Algorithm

$$\theta' | \theta \sim \mathcal{N} \left( \theta + \frac{\lambda^2}{2} \nabla \log \pi(\theta), \lambda^2 I \right)$$

- MH accept rate just controls for discretisation error
- Proposals directed towards posterior mode
Can’t deal with correlations
Langevin diffusion with volatility

\[ dX_t = \frac{1}{2} A \nabla \log \pi(X_t) dt + \sqrt{A} dB_t \]

Positive-definite \( A \) now ‘warps’ proposals.
Langevin diffusion with volatility

\[ d\mathbf{X}_t = \frac{1}{2} A \nabla \log \pi(\mathbf{X}_t) dt + \sqrt{\lambda} dB_t \]

Positive-definite \( A \) now ‘warps’ proposals.

Substituting into (1):

\[ b_i(\mathbf{x})\pi(\mathbf{x}) = \frac{1}{2} \sum_j A_{ij} \frac{\partial}{\partial x_j} [\log \pi(\mathbf{x})] \pi(\mathbf{x}), \]

\[ = \frac{1}{2} \sum_j \frac{\partial}{\partial x_j} [A_{ij} \pi(\mathbf{x})], \]
Pre-conditioned MALA

\[ \theta' | \theta \sim \mathcal{N} \left( \theta + \frac{\lambda^2}{2} A \nabla \log \pi(\theta), \lambda^2 A \right) \]

- Acceptance rate just controls for discretisation error
- Proposals are now ellipses
Pre-conditioned MALA: elliptical proposals
Local Correlations
Position-dependent volatility?

\[ dX_t = \frac{1}{2} A(X_t) \nabla \log \pi(X_t) dt + \sqrt{A(X_t)} dB_t \]
Position-dependent volatility?

\[ d\mathbf{X}_t = \frac{1}{2} A(\mathbf{X}_t) \nabla \log \pi(\mathbf{X}_t) dt + \sqrt{A(\mathbf{X}_t)} dB_t \]

RHS of (1) becomes:

\[ \frac{1}{2} \left( \sum_j A_{ij}(\mathbf{x}) \frac{\partial}{\partial x_j} [\pi(\mathbf{x})] + \pi(\mathbf{x}) \sum_j \frac{\partial}{\partial x_j} [A_{ij}(\mathbf{x})] \right) \]

This term breaks Fokker-Planck!
Position-dependent MALA

\[ d\mathbf{X}_t = \frac{1}{2} A(\mathbf{X}_t) \nabla \log \pi(\mathbf{X}_t) dt + \Lambda(\mathbf{X}_t) dt + \sqrt{A(\mathbf{X}_t)} dB_t \]

Extra drift term:

\[ \Lambda_i(\mathbf{X}_t) = \frac{1}{2} \sum_j \frac{\partial}{\partial x_j} A_{ij}(\mathbf{X}_t) \]
Position-dependent MALA

\[ dX_t = \frac{1}{2} A(X_t) \nabla \log \pi(X_t) dt + \Lambda(X_t) dt + \sqrt{A(X_t)} dB_t \]

Extra drift term:

\[ \Lambda_i(X_t) = \frac{1}{2} \sum_j \frac{\partial}{\partial x_j} A_{ij}(X_t) \]

Metropolis-Hastings proposals become:

\[ \theta' | \theta \sim \mathcal{N} \left( \theta + \frac{\lambda^2}{2} A(\theta) \nabla \log \pi(\theta) + \lambda^2 \Lambda(\theta), \lambda^2 A(\theta) \right) \]
Previous results are different to this!


\[ d\mathbf{X}_t = \frac{1}{2} G^{-1}(\mathbf{X}_t) \nabla \log \pi(\mathbf{X}_t) dt + \Omega(\mathbf{X}_t) dt + \sqrt{G^{-1}(\mathbf{X}_t)} d\mathbf{B}_t \]
Previous results are different to this!


\[ d\mathbf{X}_t = \frac{1}{2} G^{-1}(\mathbf{X}_t) \nabla \log \pi(\mathbf{X}_t) dt + \Omega(\mathbf{X}_t) dt + \sqrt{G^{-1}(\mathbf{X}_t)} dB_t \]

Extra drift term:

\[ \Omega_i(\mathbf{X}_t) = |G(\mathbf{X}_t)|^{-\frac{1}{2}} \sum_j \frac{\partial}{\partial X_j} [||G(\mathbf{X}_t)||^{\frac{1}{2}} G^{-1}_{ij}(\mathbf{X}_t)] \]

\[ dX_t = \frac{1}{2} G^{-1}(X_t) \nabla \log \pi(X_t) dt + \Omega(X_t) dt + \sqrt{G^{-1}(X_t)} dB_t \]

Extra drift term:

\[ \Omega_i(X_t) = |G(X_t)|^{-\frac{1}{2}} \sum_j \frac{\partial}{\partial X_j} [|G(X_t)|^{\frac{1}{2}} G^{-1}_{ij}(X_t)] \]

1. Where does this diffusion come from?
Previous results are different to this!


\[ dX_t = \frac{1}{2} G^{-1}(X_t) \nabla \log \pi(X_t) dt + \Omega(X_t) dt + \sqrt{G^{-1}(X_t)} dB_t \]

Extra drift term:

\[ \Omega_i(X_t) = |G(X_t)|^{-\frac{1}{2}} \sum_j \frac{\partial}{\partial X_j} [||G(X_t)|^{\frac{1}{2}} G_{ij}^{-1}(X_t)] \]

1. Where does this diffusion come from?
2. Is \( \pi \) the invariant density?
Differential Geometry

- Smooth manifold $M$: most general space on which we can do Calculus
Differential Geometry

- Smooth manifold $M$: most general space on which we can do Calculus
- Distances in a parameter space can be re-defined
  - MCMC: faster convergence, better mixing
Differential Geometry

- Smooth manifold $M$: most general space on which we can do Calculus
- Distances in a parameter space can be re-defined
  - MCMC: faster convergence, better mixing
- Gradients can still be exploited
Differential Geometry

- Smooth manifold $M$: most general space on which we can do Calculus
- Distances in a parameter space can be re-defined
  - MCMC: faster convergence, better mixing
- Gradients can still be exploited
Riemannian metrics

Manifold can be defined by mapping:

\[ r(\theta) : \mathbb{R}^n \rightarrow M \subset \mathbb{R}^{n+m} \]
Riemannian metrics

Manifold can be defined by mapping:

\[ r(\theta) : \mathbb{R}^n \rightarrow M \subset \mathbb{R}^{n+m} \]

Distance between points \( r(\theta) \) and \( r(\theta + \triangle \theta) \) given by:

\[ d(r(\theta), r(\theta + \triangle \theta)) \approx \sqrt{\triangle \theta^T G(\theta) \triangle \theta} \]
Riemannian metrics

Manifold can be defined by mapping:

\[ r(\theta) : \mathbb{R}^n \rightarrow M \subset \mathbb{R}^{n+m} \]

Distance between points \( r(\theta) \) and \( r(\theta + \triangle \theta) \) given by:

\[ d(r(\theta), r(\theta + \triangle \theta)) \approx \sqrt{\triangle \theta^T G(\theta) \triangle \theta} \]

Matrix \( G(\theta) \) given by:

\[ G_{ij}(\theta) = \left\langle \frac{\partial r}{\partial \theta_i}, \frac{\partial r}{\partial \theta_j} \right\rangle_{\mathbb{R}^{n+m}} \]
Riemannian metrics

Manifold can be defined by mapping:

\[ r(\theta) : \mathbb{R}^n \rightarrow M \subset \mathbb{R}^{n+m} \]

Distance between points \( r(\theta) \) and \( r(\theta + \triangle \theta) \) given by:

\[ d(r(\theta), r(\theta + \triangle \theta)) \approx \sqrt{\triangle \theta^T G(\theta) \triangle \theta} \]

Matrix \( G(\theta) \) given by:

\[ G_{ij}(\theta) = \left\langle \frac{\partial r}{\partial \theta_i}, \frac{\partial r}{\partial \theta_j} \right\rangle_{\mathbb{R}^{n+m}} \]

Any positive definite \( G(\theta) \) defines a manifold!
Riemannian metrics

Manifold can be defined by mapping:

\[ r(\theta) : \mathbb{R}^n \rightarrow M \subset \mathbb{R}^{n+m} \]

Distance between points \( r(\theta) \) and \( r(\theta + \triangle \theta) \) given by:

\[ d(r(\theta), r(\theta + \triangle \theta)) \approx \sqrt{\triangle \theta^T G(\theta) \triangle \theta} \]

Matrix \( G(\theta) \) given by:

\[ G_{ij}(\theta) = \left\langle \frac{\partial r}{\partial \theta_i}, \frac{\partial r}{\partial \theta_j} \right\rangle_{\mathbb{R}^{n+m}} \]

Any positive definite \( G(\theta) \) defines a manifold!
We choose the distances we want, and manifold is implicitly defined...
Diffusions on Manifolds


\[ dX_t = \frac{1}{2} \nabla \log \pi (X_t) \, dt + dB_t \]

Becomes this?

\[ dX_t = \frac{1}{2} G^{-1}(X_t) \nabla \log \pi (X_t) \, dt + \Omega(X_t) \, dt + G^{-1}(X_t) \, dB_t \]
Diffusions on Manifolds


- Objective: write a diffusion on $\mathbb{R}^n$ s.t. image under a smooth mapping $r : \mathbb{R}^n \to M$ is a Langevin diffusion on $M$. 

\[
\frac{d}{dt} X_t = \frac{1}{2} \nabla \log \pi(X_t) \ dt + dB_t
\]

Becomes this?

\[
\frac{d}{dt} X_t = \frac{1}{2} G^{-1}(X_t) \nabla \log \pi(X_t) \ dt + \Omega(X_t) \ dt + G^{-1}(X_t) dB_t
\]
Diffusions on Manifolds

- Objective: write a diffusion on $\mathbb{R}^n$ s.t. image under a smooth mapping $r : \mathbb{R}^n \rightarrow M$ is a Langevin diffusion on $M$.

$$dX_t = \frac{1}{2} \nabla \log \pi(X_t) dt + dB_t$$
Diffusions on Manifolds

- R & S (2002), G & C (2011) approach: generalise Langevin diffusion to \( M \)

- Objective: write a diffusion on \( \mathbb{R}^n \) s.t. image under a smooth mapping \( r : \mathbb{R}^n \rightarrow M \) is a Langevin diffusion on \( M \).

\[
dX_t = \frac{1}{2} \nabla \log \pi(X_t) dt + dB_t
\]

Becomes this?

\[
dX_t = \frac{1}{2} G^{-1}(X_t) \nabla \log \pi(X_t) dt + \Omega(X_t) dt + G^{-1}(X_t) dB_t
\]
Two subtle points

1. Mapping $x \mapsto r(x)$ alters invariant density - change of variables ($J = |G(x)|^{1/2}$)

2. Transcription error (Kent, 1978): $\tilde{\Omega}_i(X_t) = \frac{1}{2}|G(X_t)|^{-\frac{1}{2}} \sum_j \frac{\partial}{\partial X_j} \left[ \frac{1}{2} G^{-1}(X_t)_{ij}(X_t) \right]$
Two subtle points

1. Mapping \( x \rightarrow r(x) \) alters invariant density - change of variables (\( J = |G(x)|^{1/2} \))
   - If we aim to preserve \( \pi(x) \) we will actually preserve \( \pi(x)|G(x)|^{1/2} \)
Two subtle points

1. Mapping $\mathbf{x} \rightarrow r(\mathbf{x})$ alters invariant density - change of variables ($J = |G(\mathbf{x})|^{1/2}$)
   - If we aim to preserve $\pi(\mathbf{x})$ we will actually preserve $\pi(\mathbf{x}) |G(\mathbf{x})|^{1/2}$

2. Transcription error (Kent, 1978):

   $$\tilde{\Omega}_i(\mathbf{X}_t) = \frac{1}{2} |G(\mathbf{X}_t)|^{-\frac{1}{2}} \sum_j \frac{\partial}{\partial X_j} [\|G(\mathbf{X}_t)\|^{\frac{1}{2}} G_{ij}^{-1}(\mathbf{X}_t)]$$
Incorporating these makes diffusions equivalent!

\[
\frac{1}{2} G^{-1}(X_t) \nabla \log \left( \pi(X_t) | G(X_t)^{-\frac{1}{2}} \right) dt + \tilde{\Omega}(X_t) dt
\]

reduces to:

\[
\frac{1}{2} G^{-1}(X_t) \nabla \log \pi(X_t) dt + \Lambda(X_t) dt
\]
Experimental results

Bayesian Logistic Regression: PMALA = 143.7, MMALA = 116.4
Stochastic Volatility Model: PMALA = 0.063, MMALA = 0.05
Nonlinear Differential Equation: PMALA = 0.74, MMALA = 0.68
Conclusion

- Simpler, more accurate version of position dependent MALA
Conclusion

- Simpler, more accurate version of position dependent MALA
- Corrected error which has propagated through literature
Conclusion

- Simpler, more accurate version of position dependent MALA
- Corrected error which has propagated through literature
- Understanding of stochastic processes on manifolds
Thanks for listening

Special thanks:
Co-authors, Alex Beskos, Mike Betancourt, Michael Epstein.