Optimal Scaling of MCMC Algorithms

Natesh S. Pillai
Department of Statistics
Harvard University

INFORMS 2016

Outline

- What is Optimal Scaling?
- Gareth Roberts' work: Random walk Metropolis and Langevin Algorithm
- Hybrid Monte Carlo Algorithm
- Optimal Scaling in Infinite Dimensions
- Conclusion

Outline

- What is Optimal Scaling?
- Gareth Roberts' work: Random walk Metropolis and Langevin Algorithm
- Hybrid Monte Carlo Algorithm
- Optimal Scaling in Infinite Dimensions
- Conclusion

Outline

- What is Optimal Scaling?
- Gareth Roberts' work: Random walk Metropolis and Langevin Algorithm
- Hybrid Monte Carlo Algorithm
- Optimal Scaling in Infinite Dimensions
- Conclusion

Outline

- What is Optimal Scaling?
- Gareth Roberts' work: Random walk Metropolis and Langevin Algorithm
- Hybrid Monte Carlo Algorithm
- Optimal Scaling in Infinite Dimensions
- Conclusion

Outline

- What is Optimal Scaling?
- Gareth Roberts' work: Random walk Metropolis and Langevin Algorithm
- Hybrid Monte Carlo Algorithm
- Optimal Scaling in Infinite Dimensions
- Conclusion

Outline

- What is Optimal Scaling?
- Gareth Roberts' work: Random walk Metropolis and Langevin Algorithm
- Hybrid Monte Carlo Algorithm
- Optimal Scaling in Infinite Dimensions
- Conclusion

Metropolis Algorithm: RGG97

- Seminal paper: Roberts, Gelman and Gilks, 1997. (1000 citations)
- Target density: i.i.d components

$$\pi^N(x) = \prod_{i=1}^{N} f_i(x_i) \propto \prod_{i=1}^{N} e^{-g_i(x_i)}$$

- Simple Random walk Proposal:

$$y = x + \sqrt{\ell\,\delta} Z_N$$

- $Z_N \sim \mathrm{No}(0, I_N)$.
- ℓ - "optimisation" parameter.
- $\delta = \delta(N)$ is the SCALE.

Metropolis Algorithm: RGG97

- Seminal paper: Roberts, Gelman and Gilks, 1997. (1000 citations)
- Target density: i.i.d components

$$\pi^N(x) = \prod_{i=1}^{N} f_i(x_i) \propto \prod_{i=1}^{N} e^{-g_i(x_i)}$$

- Simple Random walk Proposal:

$$y = x + \sqrt{\ell\,\delta} Z_N$$

- $Z_N \sim \mathrm{No}(0, I_N)$.
- ℓ - "optimisation" parameter.
- $\delta = \delta(N)$ is the SCALE.

Metropolis Algorithm: RGG97

- Seminal paper: Roberts, Gelman and Gilks, 1997. (1000 citations)
- Target density: i.i.d components

$$\pi^N(x) = \prod_{i=1}^{N} f_i(x_i) \propto \prod_{i=1}^{N} e^{-g_i(x_i)}$$

- Simple Random walk Proposal:

$$y = x + \sqrt{\ell\,\delta} Z_N$$

- $Z_N \sim \mathrm{No}(0, I_N)$.
- ℓ - "optimisation" parameter.
- $\delta = \delta(N)$ is the SCALE.

Metropolis Algorithm: RGG97

- Seminal paper: Roberts, Gelman and Gilks, 1997. (1000 citations)
- Target density: i.i.d components

$$\pi^N(x) = \prod_{i=1}^{N} f_i(x_i) \propto \prod_{i=1}^{N} e^{-g_i(x_i)}$$

- Simple Random walk Proposal:

$$y = x + \sqrt{\ell\,\delta} Z_N$$

- $Z_N \sim \mathrm{No}(0, I_N)$.
- ℓ - “optimisation” parameter.
- $\delta = \delta(N)$ is the SCALE.

Metropolis Algorithm: RGG97

- Seminal paper: Roberts, Gelman and Gilks, 1997. (1000 citations)
- Target density: i.i.d components

$$\pi^N(x) = \prod_{i=1}^{N} f_i(x_i) \propto \prod_{i=1}^{N} e^{-g_i(x_i)}$$

- Simple Random walk Proposal:

$$y = x + \sqrt{\ell\,\delta} Z_N$$

- $Z_N \sim \mathrm{No}(0, I_N)$.
- ℓ - "optimisation" parameter.
- $\delta = \delta(N)$ is the SCALE.

Traditionally, ...

- Study of Mixing times
- Time to attain Stationarity
- 'Burn in time'
- Spectral gap

Hard problems ...

For practical MCMC arguably optimisation questions (find the best algorithm from a class) are more important

The new perspective in Roberts, Gelman and Gilks, 1997

- Study the Markov chain AFTER Stationarity
  - thus complementing work on convergence, robustness to starting values etc..
- Scale the proposal as a function of the dimension.
- Goldilocks Principle (attributed to Rosenthal, J.)

The new perspective in Roberts, Gelman and Gilks, 1997

- Study the Markov chain AFTER Stationarity
  - thus complementing work on convergence, robustness to starting values etc..
- Scale the proposal as a function of the dimension.
- Goldilocks Principle (attributed to Rosenthal, J.)

The new perspective in Roberts, Gelman and Gilks, 1997

- Study the Markov chain AFTER Stationarity
  - thus complementing work on convergence, robustness to starting values etc..
- Scale the proposal as a function of the dimension.
- Goldilocks Principle (attributed to Rosenthal, J.)

Acceptance Probability

- Acceptance Probability = $\min(1, \frac{\pi(y)}{\pi(x)})$.
- If $y \approx x$, $\pi(y) \approx \pi(x)$, and thus acceptance probability is equal is very high.
- If y is far away from x, then $\frac{\pi(y)}{\pi(x)} \ll 1$!

Goldilock's Principle; Figure courtesy: Roberts and Rosenthal, 2001.

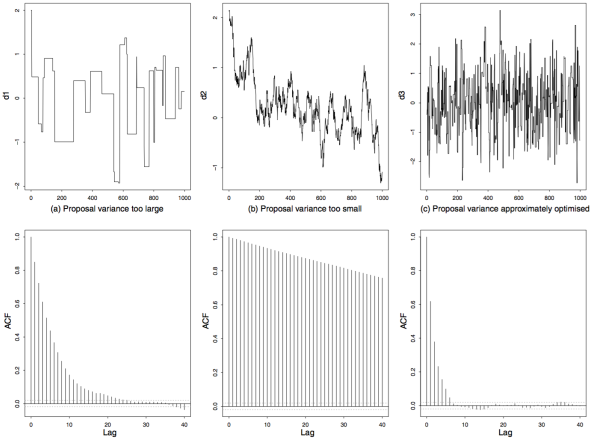


FIG. 2. *Simple Metropolis algorithm with* (a) *too-large variance (left plots),* (b) *too-small variance (middle) and* (c) *appropriate variance (right). Trace plots (top) and autocorrelation plots (below) are shown for each case.*

Key Technical Idea.

- Choose the scale such that

$$\mathbb{E}(\text{acc prob}) = \mathcal{O}(1)$$

- For large N,

$$\mathbb{E}(\text{acc prob}) = a(N) \approx a$$

- Optimise a, to obtain “best" acceptance probability.

Roberts, Gilks and Gelman, 1997:

Theorem: (for distributions with exponential moments + mild conditions)

- $\delta = \delta(N) = \frac{1}{N}$.
- A SINGLE component (rescaled) : $x^k \Rightarrow X_t$

$$dX_t = -h(\ell)\,\nabla g(X_t)dt + \sqrt{2h(\ell)}\,dW_t$$

- $\mathbb{E}(\text{acc prob}) \to 2\Phi(-\frac{\ell}{\sqrt{2}})$
- Expected Squared Jumping Distance:

$$h(\ell) = \mathbb{E}(x^{k+1} - x^k)^2 \to 2\ell^2\Phi(-\frac{\ell}{\sqrt{2}})$$

- Optimal acceptance probability: Maximizes the expected squared jumping distance:

$$\widehat{a} = 0.234$$

Roberts, Gilks and Gelman, 1997:

Theorem: (for distributions with exponential moments + mild conditions)

- $\delta = \delta(N) = \frac{1}{N}$.
- A SINGLE component (rescaled) : $x^k \Rightarrow X_t$

$$dX_t = -h(\ell)\, \nabla g(X_t) dt + \sqrt{2h(\ell)}\, dW_t$$

- $\mathbb{E}(\text{acc prob}) \to 2\Phi(-\frac{\ell}{\sqrt{2}})$
- Expected Squared Jumping Distance:

$$h(\ell) = \mathbb{E}(x^{k+1} - x^k)^2 \to 2\ell^2 \Phi(-\frac{\ell}{\sqrt{2}})$$

- Optimal acceptance probability: Maximizes the expected squared jumping distance:

$$\widehat{a} = 0.234$$

Roberts, Gilks and Gelman, 1997:

Theorem: (for distributions with exponential moments + mild conditions)

- $\delta = \delta(N) = \frac{1}{N}$.
- A SINGLE component (rescaled) : $x^k \Rightarrow X_t$

$$dX_t = -h(\ell)\,\nabla g(X_t)dt + \sqrt{2h(\ell)}\,dW_t$$

- $\mathbb{E}(\text{acc prob}) \to 2\Phi(-\frac{\ell}{\sqrt{2}})$
- Expected Squared Jumping Distance:

$$h(\ell) = \mathbb{E}(x^{k+1} - x^k)^2 \to 2\ell^2\Phi(-\frac{\ell}{\sqrt{2}})$$

- Optimal acceptance probability: Maximizes the expected squared jumping distance:

$$\widehat{a} = 0.234$$

Roberts, Gilks and Gelman, 1997:

Theorem: (for distributions with exponential moments + mild conditions)

- $\delta = \delta(N) = \frac{1}{N}$.
- A SINGLE component (rescaled) : $x^k \Rightarrow X_t$

$$dX_t = -h(\ell)\,\nabla g(X_t)dt + \sqrt{2h(\ell)}\,dW_t$$

- $\mathbb{E}(\text{acc prob}) \to 2\Phi(-\frac{\ell}{\sqrt{2}})$
- Expected Squared Jumping Distance:

$$h(\ell) = \mathbb{E}(x^{k+1} - x^k)^2 \to 2\ell^2\Phi(-\frac{\ell}{\sqrt{2}})$$

- Optimal acceptance probability: Maximizes the expected squared jumping distance:

$$\widehat{a} = 0.234$$

Roberts, Gilks and Gelman, 1997:

Theorem: (for distributions with exponential moments + mild conditions)

- $\delta = \delta(N) = \frac{1}{N}$.
- A SINGLE component (rescaled) : $x^k \Rightarrow X_t$

$$dX_t = -h(\ell)\,\nabla g(X_t)dt + \sqrt{2h(\ell)}\,dW_t$$

- $\mathbb{E}(\text{acc prob}) \to 2\Phi(-\frac{\ell}{\sqrt{2}})$
- Expected Squared Jumping Distance:

$$h(\ell) = \mathbb{E}(x^{k+1} - x^k)^2 \to 2\ell^2\Phi(-\frac{\ell}{\sqrt{2}})$$

- Optimal acceptance probability: Maximizes the expected squared jumping distance:

$$\widehat{a} = 0.234$$

Optimal Acceptance Probability; Figure courtesy: Roberts and Rosenthal, 2001

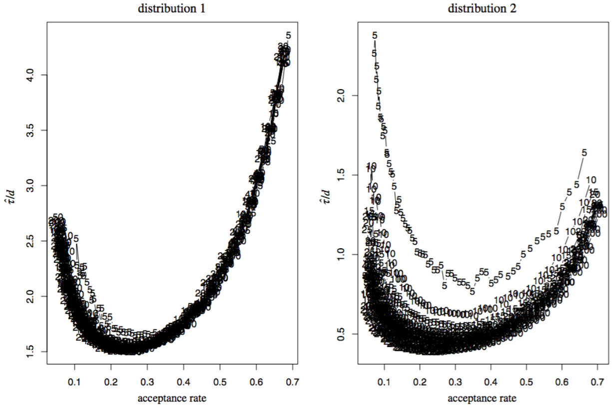


FIG. 5. *Convergence times for Metropolis algorithms as a function of their acceptance rates. The plotting symbol indicates the dimension of the simulation.*

Diffusion limit: Insights

- Why maximize $\mathbb{E}(x^{k+1} - x^k)^2$?
- At stationarity

$$\begin{aligned}\mathbb{E}(x^{k+1} - x^k)^2 &= \mathbb{E}(x^{k+1})^2 + \mathbb{E}(x^k)^2 - 2\mathrm{Cov}(x^{k+1}, x^k) \\ &= 2M - 2\mathrm{Cov}(x^{k+1}, x^k)\end{aligned}$$

- Why only lag-1 correlation? Higher orders?
- The quantity $h(\ell)$ is the Speed of the diffusion.
- Thus, because of the diffusion limit, maximizing $h(\ell)$ leads to minimizing the asymptotic variance.

Diffusion limit: Insights

- Why maximize $\mathbb{E}(x^{k+1} - x^k)^2$?
- At stationarity

$$\begin{aligned}\mathbb{E}(x^{k+1} - x^k)^2 &= \mathbb{E}(x^{k+1})^2 + \mathbb{E}(x^k)^2 - 2\mathrm{Cov}(x^{k+1}, x^k) \\ &= 2M - 2\mathrm{Cov}(x^{k+1}, x^k)\end{aligned}$$

- Why only lag-1 correlation? Higher orders?
- The quantity $h(\ell)$ is the Speed of the diffusion.
- Thus, because of the diffusion limit, maximizing $h(\ell)$ leads to minimizing the asymptotic variance.

Diffusion limit: Insights

- Why maximize $\mathbb{E}(x^{k+1} - x^k)^2$?
- At stationarity

$$\begin{aligned}\mathbb{E}(x^{k+1} - x^k)^2 &= \mathbb{E}(x^{k+1})^2 + \mathbb{E}(x^k)^2 - 2\mathrm{Cov}(x^{k+1}, x^k) \\ &= 2M - 2\mathrm{Cov}(x^{k+1}, x^k)\end{aligned}$$

- Why only lag-1 correlation? Higher orders?
- The quantity $h(\ell)$ is the Speed of the diffusion.
- Thus, because of the diffusion limit, maximizing $h(\ell)$ leads to minimizing the asymptotic variance.

Diffusion limit: Insights

- Why maximize $\mathbb{E}(x^{k+1} - x^k)^2$?
- At stationarity

$$\begin{aligned}\mathbb{E}(x^{k+1} - x^k)^2 &= \mathbb{E}(x^{k+1})^2 + \mathbb{E}(x^k)^2 - 2\mathrm{Cov}(x^{k+1}, x^k) \\ &= 2M - 2\mathrm{Cov}(x^{k+1}, x^k)\end{aligned}$$

- Why only lag-1 correlation? Higher orders?
- The quantity $h(\ell)$ is the Speed of the diffusion.
- Thus, because of the diffusion limit, maximizing $h(\ell)$ leads to minimizing the asymptotic variance.

Diffusion limit: Insights

- Why maximize $\mathbb{E}(x^{k+1} - x^k)^2$?
- At stationarity

$$\begin{aligned}\mathbb{E}(x^{k+1} - x^k)^2 &= \mathbb{E}(x^{k+1})^2 + \mathbb{E}(x^k)^2 - 2\mathrm{Cov}(x^{k+1}, x^k)\\ &= 2M - 2\mathrm{Cov}(x^{k+1}, x^k)\end{aligned}$$

- Why only lag-1 correlation? Higher orders?
- The quantity $h(\ell)$ is the Speed of the diffusion.
- Thus, because of the diffusion limit, maximizing $h(\ell)$ leads to minimizing the asymptotic variance.

Practical Conclusion of Diffusion Limit

Suppose we want to estimate $\int f(u)\pi(du)$.

- Given the precision ϵ, find T and compute

$$\widehat{f} = \frac{1}{T}\int f(X_t)dt$$

- Diffusion Limit + Optimal Scaling implies that

$$\widehat{f}_N = \frac{1}{T}\sum_{k=1}^{\lfloor T/\delta \rfloor} f(X_k) \qquad \delta = O(N^{-1})$$

  has the same precision as $\widehat{f}$.
- The mixing time of the RWM is $O(N)$.

Langevin Algorithm

- Recall Langevin diffusion: $dx_t = -\nabla g(x_t)dt + \sqrt{2}dW_t$.
- Langevin Proposal:

$$y = x - \nabla g(x)\ell\,\delta + \sqrt{2\ell\,\delta}Z_N$$

- Need a Metropolis Accept/Reject mechanism.
- x^k is the Langevin Markov chain on $\mathbb{R}^N$ for iid target.
- Theorem (Roberts + Rosenthal 1998): The scale is $\delta(N) = N^{-1/3}$ and after rescaling the first component of the Markov chain $\{x^k\}$ converges in distribution to X_t:

$$dX_t = -h_1(\ell)\,\nabla g(X_t)dt + \sqrt{2h_1(\ell)}\,dW_t\ .$$

- Optimal Acceptance Probability $= 0.574$.

Langevin Algorithm

- Recall Langevin diffusion: $dx_t = -\nabla g(x_t)dt + \sqrt{2}dW_t$.
- Langevin Proposal:

$$y = x - \nabla g(x)\ell\,\delta + \sqrt{2\ell\,\delta}Z_N$$

- Need a Metropolis Accept/Reject mechanism.
- x^k is the Langevin Markov chain on $\mathbb{R}^N$ for iid target.
- Theorem (Roberts + Rosenthal 1998): The scale is $\delta(N) = N^{-1/3}$ and after rescaling the first component of the Markov chain $\{x^k\}$ converges in distribution to X_t:

$$dX_t = -h_1(\ell)\,\nabla g(X_t)dt + \sqrt{2h_1(\ell)}\,dW_t\ .$$

- Optimal Acceptance Probability $= 0.574$.

Langevin Algorithm

- Recall Langevin diffusion: $dx_t = -\nabla g(x_t)dt + \sqrt{2}dW_t$.
- Langevin Proposal:

$$y = x - \nabla g(x)\ell\,\delta + \sqrt{2\ell\,\delta}Z_N$$

- Need a Metropolis Accept/Reject mechanism.
- x^k is the Langevin Markov chain on $\mathbb{R}^N$ for iid target.
- Theorem (Roberts + Rosenthal 1998): The scale is $\delta(N) = N^{-1/3}$ and after rescaling the first component of the Markov chain $\{x^k\}$ converges in distribution to X_t:

$$dX_t = -h_1(\ell)\,\nabla g(X_t)dt + \sqrt{2h_1(\ell)}\,dW_t\ .$$

- Optimal Acceptance Probability $= 0.574$.

Langevin Algorithm

- Recall Langevin diffusion: $dx_t = -\nabla g(x_t)dt + \sqrt{2}dW_t$.
- Langevin Proposal:

$$y = x - \nabla g(x)\ell\,\delta + \sqrt{2\ell\,\delta}Z_N$$

- Need a Metropolis Accept/Reject mechanism.
- x^k is the Langevin Markov chain on $\mathbb{R}^N$ for iid target.
- Theorem (Roberts + Rosenthal 1998): The scale is $\delta(N) = N^{-1/3}$ and after rescaling the first component of the Markov chain $\{x^k\}$ converges in distribution to X_t:

$$dX_t = -h_1(\ell)\,\nabla g(X_t)dt + \sqrt{2h_1(\ell)}\,dW_t\ .$$

- Optimal Acceptance Probability $= 0.574$.

Comparing RWM vs. Langevin

- Recall RWM had complexity of $O(N)$
- Langevin has complexity of $O(N^{1/3})$.
- Thus optimal scaling gives a nice way to compare algorithms.

So far:

- Summary:
  - optimal scaling: tuning proposals.
  - Diffusion limits for RWM and Langevin
- Hybrid Monte Carlo Algorithm
- Infinite Dimensional Result

So far:

- Summary:
  - optimal scaling: tuning proposals.
  - Diffusion limits for RWM and Langevin
- Hybrid Monte Carlo Algorithm
- Infinite Dimensional Result

So far:

- Summary:
  - optimal scaling: tuning proposals.
  - Diffusion limits for RWM and Langevin
- Hybrid Monte Carlo Algorithm
- Infinite Dimensional Result

Hybrid Monte Carlo

- Algorithm from Physics, (Duane et. al. (1987))
- Based on Hamiltonian Dynamics, conservation of energy.

Hamiltonian Dynamics

- Location x, velocity v; total energy,

$$H(x, v) = g(x) + \frac{1}{2} v^2$$

- Hamiltonian equations

$$\frac{dx}{dt} = v; \quad \frac{dv}{dt} = -\nabla g(x)$$

- They give rise to solution operator

$$\phi^T : (x_0, v_0) \mapsto (x_T, v_T)$$

  that preserves total energy.
- Equivalently the joint density

$$\exp\{-H(x, v)\} = \exp\{-g(x) - \frac{1}{2} v^2\}$$

  is preserved.

Hamiltonian Dynamics

- Location x, velocity v; total energy,

$$H(x, v) = g(x) + \frac{1}{2} v^2$$

- Hamiltonian equations

$$\frac{dx}{dt} = v; \quad \frac{dv}{dt} = -\nabla g(x)$$

- They give rise to solution operator

$$\phi^T : (x_0, v_0) \mapsto (x_T, v_T)$$

  that preserves total energy.

- Equivalently the joint density

$$\exp\{-H(x, v)\} = \exp\{-g(x) - \frac{1}{2} v^2\}$$

  is preserved.

Hamiltonian Dynamics

- Location x, velocity v; total energy,

$$H(x, v) = g(x) + \frac{1}{2} v^2$$

- Hamiltonian equations

$$\frac{dx}{dt} = v; \ \frac{dv}{dt} = -\nabla g(x)$$

- They give rise to solution operator

$$\phi^T : (x_0, v_0) \mapsto (x_T, v_T)$$

  that preserves total energy.

- Equivalently the joint density

$$\exp\{-H(x, v)\} = \exp\{-g(x) - \frac{1}{2} v^2\}$$

  is preserved.

Hamiltonian Dynamics

- Location x, velocity v; total energy,

$$H(x, v) = g(x) + \frac{1}{2}v^2$$

- Hamiltonian equations

$$\frac{dx}{dt} = v; \quad \frac{dv}{dt} = -\nabla g(x)$$

- They give rise to solution operator

$$\phi^T : (x_0, v_0) \mapsto (x_T, v_T)$$

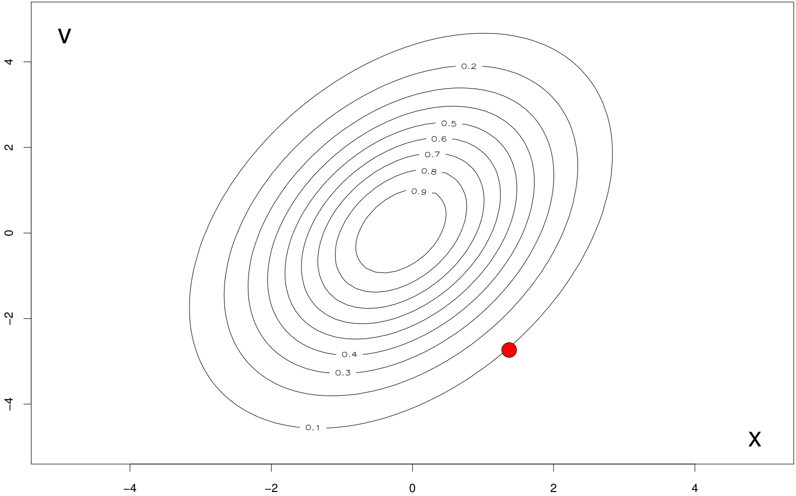
  that preserves total energy.

- Equivalently the joint density

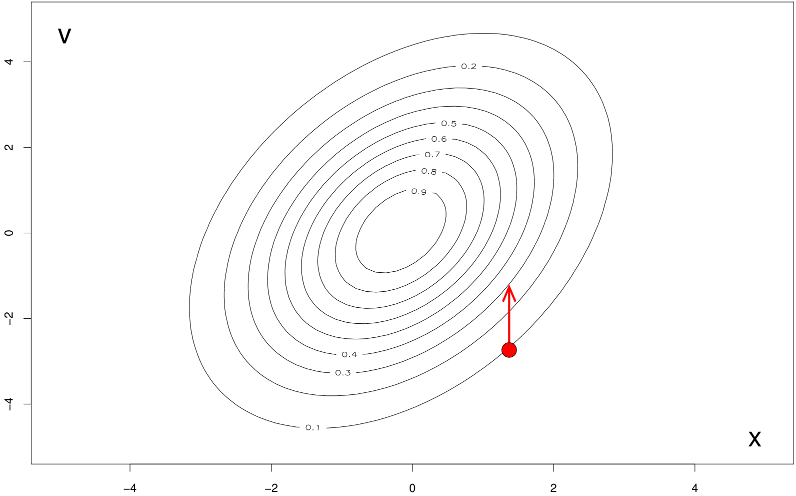
$$\exp\{-H(x, v)\} = \exp\{-g(x) - \frac{1}{2}v^2\}$$

  is preserved.

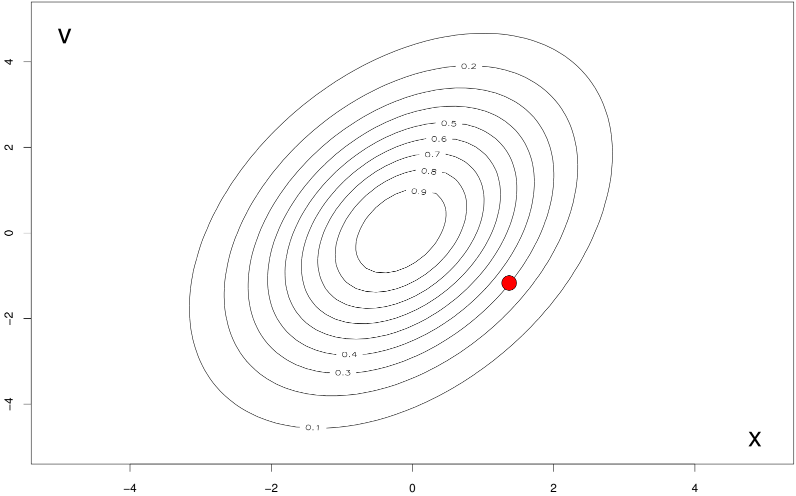
‘Exact’ Hamiltonian Dynamics



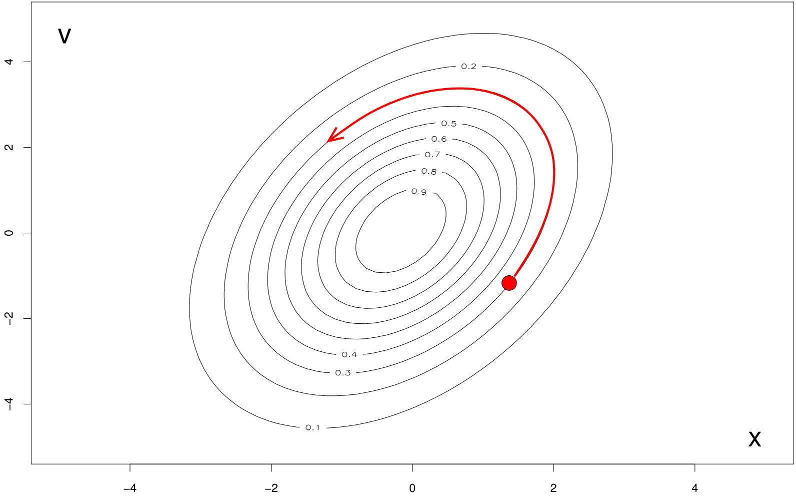
‘Exact’ Hamiltonian Dynamics



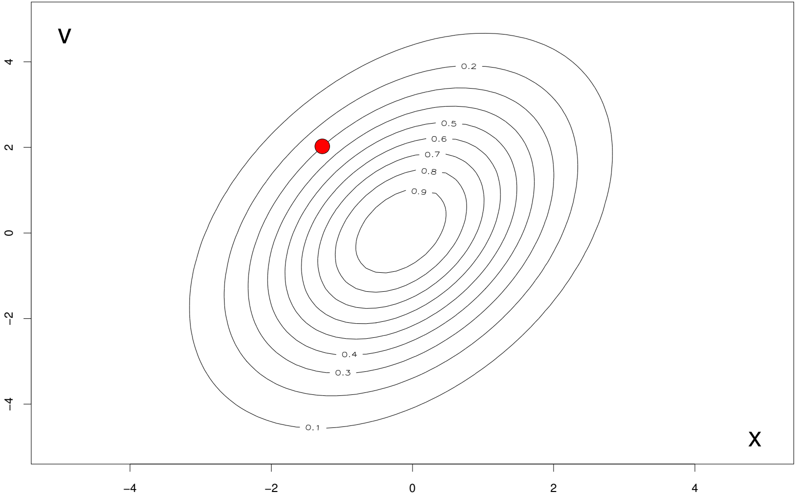
‘Exact’ Hamiltonian Dynamics



‘Exact’ Hamiltonian Dynamics



'Exact' Hamiltonian Dynamics



Leapfrog Discretisation

- In practice, dynamics are approximated:

$$\phi^T \approx \phi^{T,h} .$$

- For initial state (x_0, v_0): Leapfrog Discretisation,

$$\begin{aligned} v_{h/2} &= v_0 - \tfrac{h}{2} \nabla g(x_0) \\ x_h &= x_0 + h\, v_{h/2} \\ v_h &= v_0 - \tfrac{h}{2} \nabla g(x_h) \end{aligned}$$

- $\phi^{T,h}$ is obtained by composing $\frac{T}{h}$ leapfrog steps.
- The crucial properties of this approach are that it is volume preserving and reversible.

Acceptance probability

- (x_0, v_0) : Initial position
- (x_T, v_T): Final position
- Accept with probability

$$1 \wedge \exp\{H(x_0, v_0) - H(x_T, v_T)\}$$

- Acceptance probability $= 1$, if Hamilton's Differential Equations can be solved explicitly.

Acceptance probability

- (x_0, v_0) : Initial position
- (x_T, v_T): Final position
- Accept with probability

$$1 \wedge \exp\{H(x_0, v_0) - H(x_T, v_T)\}$$

- Acceptance probability $= 1$, if Hamilton's Differential Equations can be solved explicitly.

Lingyu Chen, Zhaohui Qin, Jun S. Liu, 2000 (2001)

"Although HMC has been found useful for Bayesian computations, many important issues remain open. For example, how to choose tuning parameters in HMC, e.g., the step-size and the number of the leapfrog iterations, is still a difficult problem. A rule of thumb is to maintain an acceptance rate of nearly 70%. But there seems to be no clear theoretical basis for this rule. "

Main result

- P., Beskos, Roberts, Sanz-Serna, Stuart, 2013.
- Target density
$$\pi^N(x) = \prod_{i=1}^{N} e^{-g_i(x_i)}$$
- Theorem : For any fixed integration length T, the step size which maximizes the expected squared distance : $h = h(N) = \frac{1}{N^{1/4}}$.
- For any Fixed integration length T, optimal scaling leads to a complexity of $O(N^{1/4})$.
- For any Volume preserving, time reversible second order numerical integrator, the optimal acceptance probability is 0.651.

Main result

- P., Beskos, Roberts, Sanz-Serna, Stuart, 2013.
- Target density

$$\pi^N(x) = \prod_{i=1}^{N} e^{-g_i(x_i)}$$

- Theorem : For any fixed integration length T, the step size which maximizes the expected squared distance : $h = h(N) = \frac{1}{N^{1/4}}$.
- For any Fixed integration length T, optimal scaling leads to a complexity of $O(N^{1/4})$.
- For any Volume preserving, time reversible second order numerical integrator, the optimal acceptance probability is 0.651.

Main result

- P., Beskos, Roberts, Sanz-Serna, Stuart, 2013.
- Target density

$$\pi^N(x) = \prod_{i=1}^{N} e^{-g_i(x_i)}$$

- Theorem : For any fixed integration length T, the step size which maximizes the expected squared distance : $h = h(N) = \frac{1}{N^{1/4}}$.
- For any Fixed integration length T, optimal scaling leads to a complexity of $O(N^{1/4})$.
- For any Volume preserving, time reversible second order numerical integrator, the optimal acceptance probability is 0.651.

Complexity of HMC

- HMC, the optimal behavior is not Diffusive!
- The complexity of HMC is $O(N^{\frac{1}{4}})$.

Complexity of HMC

- HMC, the optimal behavior is not Diffusive!
- The complexity of HMC is $O(N^{\frac{1}{4}})$.

Universality: How general?

- This first result was proved for IID targets only.
- Empirically seen to be robust well beyond IID case!
- What are the challenges for the Non-Product case?
  - progress made by Roberts, Rosenthal, Sherlock, Neal, Bedard and others ...

Universality: How general?

- This first result was proved for IID targets only.
- Empirically seen to be robust well beyond IID case!
- What are the challenges for the Non-Product case?
  - progress made by Roberts, Rosenthal, Sherlock, Neal, Bedard and others ...

Universality: How general?

- This first result was proved for IID targets only.
- Empirically seen to be robust well beyond IID case!
- What are the challenges for the Non-Product case?
  - progress made by Roberts, Rosenthal, Sherlock, Neal, Bedard and others ...

Infinite Dimensional Distribution

- Let $\mathcal{H}$ be an infinite dimensional Hilbert space, $\pi_0 \sim \mathrm{N}(0, C)$.
- Our target measure:

$$\pi(f) \propto \exp\{-\Psi(f)\}\pi_0(f)$$

- For N large, we take N dimensional projection:

$$\pi^N \approx \pi$$

- π is NOT a product measure.

Target Measure: Radon Nikodym Derivative w.r.t to Gaussian

- $\mathcal{H}$ Hilbert space, $\pi_0 \sim \mathrm{N}(0, C)$.
- Target:

$$\pi(f) \propto \exp\{-\Psi(f)\}\pi_0(f)$$

- Diffusion bridges. (Girsanov)
- Constructive Quantum Field Theory, $P(\phi_2^4)$ model.

Target Measure: Radon Nikodym Derivative w.r.t to Gaussian

- $\mathcal{H}$ Hilbert space, $\pi_0 \sim \mathrm{N}(0, C)$.
- Target:

$$\pi(f) \propto \exp\{-\Psi(f)\}\pi_0(f)$$

- Diffusion bridges. (Girsanov)
- Constructive Quantum Field Theory, $P(\phi_2^4)$ model.

Target Measure: Radon Nikodym Derivative w.r.t to Gaussian

- $\mathcal{H}$ Hilbert space, $\pi_0 \sim \mathrm{N}(0, C)$.
- Target:

$$\pi(f) \propto \exp\{-\Psi(f)\}\pi_0(f)$$

- Diffusion bridges. (Girsanov)
- Constructive Quantum Field Theory, $P(\phi_2^4)$ model.

Target Measure: Radon Nikodym Derivative w.r.t to Gaussian

- $\mathcal{H}$ Hilbert space, $\pi_0 \sim \mathrm{N}(0, C)$.
- Target:

$$\pi(f) \propto \exp\{-\Psi(f)\}\pi_0(f)$$

- Diffusion bridges. (Girsanov)
- Constructive Quantum Field Theory, $P(\phi_2^4)$ model.

Infinite Dimensional Result

- Recall f is a function, target π measure on $\mathcal{H}$.

$$\pi(f) \propto \exp\{-\Psi(f)\}\,\pi_0(f), \qquad \pi_0(f) \sim \text{No}(0, C)$$

- Proposal $y = x + Z_N, \quad Z_N \sim \text{No}(0, C^N)$.
  Mattingly, P., Stuart (Annals of App. Prob., 2012)
- $\{x^k\}$ is the Random Walk Metropolis Markov chain on $\mathbb{R}^N$.
- Theorem: For the scaling $\delta(N) = \frac{1}{N}$ the (rescaled) Markov chain $\{x^k\}$ converges in distribution to an infinite dimensional diffusion (SPDE) X_t

$$dX_t = (-X_t - C\nabla\Psi(X_t))dt + \sqrt{2C}\,dW_t.$$

- Weak Convergence in $C([0, T], \mathcal{H})$.
- Optimal Acceptance Probability $= 0.234$.

Infinite Dimensional Result

- Recall f is a function, target π measure on $\mathcal{H}$.

$$\pi(f) \propto \exp\{-\Psi(f)\}\,\pi_0(f), \qquad \pi_0(f) \sim \text{No}(0, C)$$

- Proposal $y = x + Z_N, \quad Z_N \sim \text{No}(0, C^N)$.
  Mattingly, P., Stuart (Annals of App. Prob., 2012)
- $\{x^k\}$ is the Random Walk Metropolis Markov chain on $\mathbb{R}^N$.
- Theorem: For the scaling $\delta(N) = \frac{1}{N}$ the (rescaled) Markov chain $\{x^k\}$ converges in distribution to an infinite dimensional diffusion (SPDE) X_t

$$dX_t = (-X_t - C\nabla\Psi(X_t))dt + \sqrt{2C}\,dW_t.$$

- Weak Convergence in $C([0, T], \mathcal{H})$.
- Optimal Acceptance Probability $= 0.234$.

Infinite Dimensional Result

- Recall f is a function, target π measure on $\mathcal{H}$.

$$\pi(f) \propto \exp\{-\Psi(f)\}\,\pi_0(f), \qquad \pi_0(f) \sim \mathrm{No}(0, C)$$

- Proposal $y = x + Z_N, \quad Z_N \sim \mathrm{No}(0, C^N)$.
  Mattingly, P., Stuart (Annals of App. Prob., 2012)
- $\{x^k\}$ is the Random Walk Metropolis Markov chain on $\mathbb{R}^N$.
- Theorem: For the scaling $\delta(N) = \frac{1}{N}$ the (rescaled) Markov chain $\{x^k\}$ converges in distribution to an infinite dimensional diffusion (SPDE) X_t

$$dX_t = (-X_t - C\nabla\Psi(X_t))dt + \sqrt{2C}\,dW_t.$$

- Weak Convergence in $C([0, T], \mathcal{H})$.
- Optimal Acceptance Probability $= 0.234$.

Infinite Dimensional Result

- Recall f is a function, target π measure on $\mathcal{H}$.

$$\pi(f) \propto \exp\{-\Psi(f)\}\,\pi_0(f), \qquad \pi_0(f) \sim \mathrm{No}(0, C)$$

- Proposal $y = x + Z_N, \quad Z_N \sim \mathrm{No}(0, C^N)$.
  Mattingly, P., Stuart (Annals of App. Prob., 2012)
- $\{x^k\}$ is the Random Walk Metropolis Markov chain on $\mathbb{R}^N$.
- Theorem: For the scaling $\delta(N) = \frac{1}{N}$ the (rescaled) Markov chain $\{x^k\}$ converges in distribution to an infinite dimensional diffusion (SPDE) X_t

$$dX_t = (-X_t - C\nabla\Psi(X_t))dt + \sqrt{2C}\,dW_t.$$

- Weak Convergence in $C([0, T], \mathcal{H})$.
- Optimal Acceptance Probability $= 0.234$.

Weak Convergence to SPDE: Proof Sketch

- Decompose the Markov Chain into Drift + Noise

$$x^{k+1} = x^k + \mathbb{E}(x^{k+1} - x^k | x^k) + \sqrt{2\ell\delta}\,\Gamma^k \ .$$

- Obtain Drift and Diffusion Estimates

$$\mathbb{E}(x^{k+1} - x^k | x^k) \approx -\nabla\Psi(x^k)\,\delta$$

- Martingale Central Limit Theorem, noise satisfies an invariance principle.
- Continuity of the Ito map : $\Theta : C([0,T],\mathcal{H}) \mapsto C([0,T],\mathcal{H})$, $\Theta(W) = X$:

$$dX_t = (-X_t - C\nabla\Psi(X_t))dt + \sqrt{2C}\,dW_t.$$

- Continuous mapping theorem, concludes the proof.
- Connection to the Euler-Maruyama Scheme!

Weak Convergence to SPDE: Proof Sketch

- Decompose the Markov Chain into Drift + Noise

$$x^{k+1} = x^k + \mathbb{E}(x^{k+1} - x^k | x^k) + \sqrt{2\ell\delta}\,\Gamma^k \ .$$

- Obtain Drift and Diffusion Estimates

$$\mathbb{E}(x^{k+1} - x^k | x^k) \approx -\nabla\Psi(x^k)\,\delta$$

- Martingale Central Limit Theorem, noise satisfies an invariance principle.
- Continuity of the Ito map : $\Theta : C([0,T],\mathcal{H}) \mapsto C([0,T],\mathcal{H})$, $\Theta(W) = X$:

$$dX_t = (-X_t - C\nabla\Psi(X_t))dt + \sqrt{2C}\,dW_t.$$

- Continuous mapping theorem, concludes the proof.
- Connection to the Euler-Maruyama Scheme!

Weak Convergence to SPDE: Proof Sketch

- Decompose the Markov Chain into Drift + Noise

$$x^{k+1} = x^k + \mathbb{E}(x^{k+1} - x^k | x^k) + \sqrt{2\ell\delta}\,\Gamma^k\ .$$

- Obtain Drift and Diffusion Estimates

$$\mathbb{E}(x^{k+1} - x^k | x^k) \approx -\nabla\Psi(x^k)\,\delta$$

- Martingale Central Limit Theorem, noise satisfies an invariance principle.
- Continuity of the Ito map : $\Theta : C([0,T],\mathcal{H}) \mapsto C([0,T],\mathcal{H})$, $\Theta(W) = X$:

$$dX_t = (-X_t - C\nabla\Psi(X_t))dt + \sqrt{2C}\,dW_t.$$

- Continuous mapping theorem, concludes the proof.
- Connection to the Euler-Maruyama Scheme!

Weak Convergence to SPDE: Proof Sketch

- Decompose the Markov Chain into Drift + Noise

$$x^{k+1} = x^k + \mathbb{E}(x^{k+1} - x^k | x^k) + \sqrt{2\ell\delta}\,\Gamma^k\ .$$

- Obtain Drift and Diffusion Estimates

$$\mathbb{E}(x^{k+1} - x^k | x^k) \approx -\nabla\Psi(x^k)\,\delta$$

- Martingale Central Limit Theorem, noise satisfies an invariance principle.
- Continuity of the Ito map : $\Theta : C([0,T],\mathcal{H}) \mapsto C([0,T],\mathcal{H})$, $\Theta(W) = X$:

$$dX_t = (-X_t - C\nabla\Psi(X_t))dt + \sqrt{2C}\,dW_t.$$

- Continuous mapping theorem, concludes the proof.
- Connection to the Euler-Maruyama Scheme!

Weak Convergence to SPDE: Proof Sketch

- Decompose the Markov Chain into Drift + Noise

$$x^{k+1} = x^k + \mathbb{E}(x^{k+1} - x^k | x^k) + \sqrt{2\ell\delta}\,\Gamma^k \ .$$

- Obtain Drift and Diffusion Estimates

$$\mathbb{E}(x^{k+1} - x^k | x^k) \approx -\nabla\Psi(x^k)\,\delta$$

- Martingale Central Limit Theorem, noise satisfies an invariance principle.
- Continuity of the Ito map : $\Theta : C([0,T],\mathcal{H}) \mapsto C([0,T],\mathcal{H})$, $\Theta(W) = X$:

$$dX_t = (-X_t - C\nabla\Psi(X_t))dt + \sqrt{2C}\,dW_t.$$

- Continuous mapping theorem, concludes the proof.
- Connection to the Euler-Maruyama Scheme!

Weak Convergence to SPDE: Proof Sketch

- Decompose the Markov Chain into Drift + Noise

$$x^{k+1} = x^k + \mathbb{E}(x^{k+1} - x^k | x^k) + \sqrt{2\ell\delta}\,\Gamma^k \ .$$

- Obtain Drift and Diffusion Estimates

$$\mathbb{E}(x^{k+1} - x^k | x^k) \approx -\nabla\Psi(x^k)\,\delta$$

- Martingale Central Limit Theorem, noise satisfies an invariance principle.
- Continuity of the Ito map : $\Theta : C([0,T],\mathcal{H}) \mapsto C([0,T],\mathcal{H})$, $\Theta(W) = X$:

$$dX_t = (-X_t - C\nabla\Psi(X_t))dt + \sqrt{2C}\,dW_t.$$

- Continuous mapping theorem, concludes the proof.
- Connection to the Euler-Maruyama Scheme!

Open problems

- Combining behavior at transience + behavior at stationarity.
- Recall that, at stationarity, the scaling for Langevin is $N^{-1/3}$.
- For RWM, Langevin, before reaching stationarity, the scaling is N^{-1} (O.F. Christensen, G.O. Roberts, and J.S. Rosenthal, 2003.)
- Combine spectral gap analysis + optimal scaling.

Open problems

- Combining behavior at transience + behavior at stationarity.
- Recall that, at stationarity, the scaling for Langevin is $N^{-1/3}$.
- For RWM, Langevin, before reaching stationarity, the scaling is N^{-1} (O.F. Christensen, G.O. Roberts, and J.S. Rosenthal, 2003.)
- Combine spectral gap analysis + optimal scaling.

Open problems

- Combining behavior at transience + behavior at stationarity.
- Recall that, at stationarity, the scaling for Langevin is $N^{-1/3}$.
- For RWM, Langevin, before reaching stationarity, the scaling is N^{-1} (O.F. Christensen, G.O. Roberts, and J.S. Rosenthal, 2003.)
- Combine spectral gap analysis + optimal scaling.

Open problems

- Combining behavior at transience + behavior at stationarity.
- Recall that, at stationarity, the scaling for Langevin is $N^{-1/3}$.
- For RWM, Langevin, before reaching stationarity, the scaling is N^{-1} (O.F. Christensen, G.O. Roberts, and J.S. Rosenthal, 2003.)
- Combine spectral gap analysis + optimal scaling.

Conclusion

- Optimal scaling is an important idea, with deep practical implications.
- Lots more to do!
- “Dimension" can be different things.

Thank you!

Thanks to:

- Gareth O. Roberts
- Jeffrey S. Rosenthal
- Organizers + Applied Probability Society

References

- Weak convergence and optimal scaling of random walk Metropolis algorithms, Roberts, G.O., Gelman A., and Gilks, W.R., 1997, Annals of Applied Probability.
- Optimal scaling of discrete approximations to Langevin diffusions, Roberts, G.O. and Rosenthal, J.S., 1998, JRSS, Series B.
- Optimal Scaling of various Metropolis-Hastings algorithms, Roberts, G.O. and Rosenthal, J.S., 2001, Statistical Science.

References

- Diffusion Limits of the Random Walk Metropolis Algorithm in High Dimensions, Mattingly, J.C., Pillai, N.S., Stuart, A.M., 2012, Annals of Applied Probability.
- Optimal tuning of the Hybrid Monte-Carlo Algorithm, Beskos, A., Pillai, N.S., Roberts, G.O., Sanz-Serna, J.M., Stuart, A.M. 2013, Bernoulli.