Sampling from Rough Energy Landscapes

Gideon Simpson

Department of Mathematics
Drexel University

September 18, 2018

Joint with P. Plechac (U. Delaware)
1 Motivation & Background
   - Motivating Example
   - Sampling Strategies
   - Tuning Algorithms
   - An Explicit Computation

2 Managing Roughness
   - Two Scale Potentials
   - Dissection of the Potential
   - Dissection on the Fly

3 Numerical Experiments
   - Rough Harmonic Potential
   - Rough Doublewell Potential
   - Local Entropy Smoothing

4 Summary & Acknowledgements
Motivation & Background

Motivating Example

Rough Energy Landscapes
and their Distributions

- (b) is potential from (a) + $A \cos(x/\epsilon)$
- Expect (b) to be more difficult to sample than (a) – Quantification?
- Can we modify algorithms to improve sampling on (b)?
- Inspired by superbasin/low energy barrier problems
Considerations

- **CAVEAT:** This is work in progress
- Focus on unbiased samplers (i.e., include MH step)
- Finite $d$ vs. $d \to \infty$
- Stationary vs. nonstationary data
- May or may not have explicit scale separation

\[
V(x) = V_0(x) + V_1(x, x/\epsilon) \tag{1}
\]
Motivation & Background

Motivating Example

Metropolis Adjusted Langevin (MALA) Example

- $V(x) = \frac{1}{2}x^2 + \frac{1}{8}\cos(x/\epsilon)$
- Sample $e^{-\beta V}$ at $\beta = 5$ by MALA,

$$X_{n+1}^p = x_n - \nabla V(X_n)\Delta t + \sqrt{2\beta^{-1}\Delta t}\xi_{n+1}$$

$$X_{n+1} = \begin{cases} X_{n+1}^p & \text{with probability } 1 \wedge e^{R(X_n,X_{n+1}^p)} \\ X_n & \text{with probability } 1 - 1 \wedge e^{R(X_n,X_{n+1}^p)} \end{cases}$$

$$R(x,y) = \log \frac{e^{-\beta V(y)}q(y \rightarrow x)}{e^{-\beta V(x)}q(x \rightarrow y)}$$

- Use $\Delta t = 0.1$
Increasing stagnation (poorer sampling) as $\epsilon \to 0$
Sample of Sampling Methods

**RWM** \( X_{n+1}^p = X_n + \sqrt{2\Delta t} \xi_{n+1} \) – no information about \( V \) in proposal (cheap)

**MALA** \( X_{n+1}^p = X_n - \nabla V(X_n) + \sqrt{2\Delta t} \xi_{n+1} \)

**Precond. MALA** \( X_{n+1}^p = X_n - P \nabla V(X_n) + \sqrt{2\Delta t} P \xi_{n+1} \) – need a preconditioning matrix \( P \)

**Metropolized Langevin** \( (X_{n+1}^p, P_{n+1}^p) \) from second order Langevin – marginalize out momentum

**HMC** \( (X_{n+1}^p, P_{n+1}^p) \) from Hamiltonian flow – velocities are Gaussian, marginalize out momentum

**Others** Riemannian Manifold methods (MALA, Langevin), Irreversible & biased methods, ...

Focus on methods with accept/reject step, \( 1 \wedge e^{R(x,x^p)} \),

\[
R(x, y) = V(x) - V(y) + \log \frac{q(y \rightarrow x)}{q(x \rightarrow y)} 
\]

(5)
Choices of Parameters

RWM/MALA  Need to choose step size $\Delta t$

Precond. MALA  Need to choose $\Delta t$ and preconditioning matrix

Langevin  For a given splitting (there are many) need to choose $\Delta t$, damping, mass

HMC  $\Delta t$, time of Hamiltonian trajectory, mass

Poor choice of parameter ($\Delta t$)?
Maximizing acceptance rate is the wrong objective
Sending $\Delta t \to 0$ always sends the mean acceptance rate to 1
For RWM

$$1 - a(x, y) = 1 - 1 \wedge e^R \leq R(x, y)^-$$

Mean Rejection Rate $\leq \sqrt{\mathbb{E}[|R(x, y)|^2]} \lesssim \sqrt{\Delta t}$

Try to maximize “mixing”
Proxy for mixing: One Step Mean Square Displacement (per d.o.f.):

$$MSD = \mathbb{E}[|X_1^{(1)} - X_0^{(1)}|^2]$$
Results on Tuning in Equilibrium

High Dimensional Product Measures – Roberts, Rosenthal ’97, Roberts, Gelman, Gilks ’98, Beskos, Roberts, Stuart ’09, Beskos et al. ’13, Bou-Rabee, Sanz-Serna,’18, ...

- Product measure ansatz:
  \[ V_d(x) = \sum_{i=1}^{d} v(x_i), \quad v: \mathbb{R} \rightarrow \mathbb{R}, \quad e^{-V_d(x)} = \prod_{i=1}^{d} e^{-v(x_i)} \quad (7) \]

- d.o.f.’s only interact during accept/reject step
- As \( d \rightarrow \infty \), ensemble average acceptance and MSD can be predicted:
  \[ A(\ell) = 2\Phi \left( -\frac{\ell^p}{2} \sqrt{K} \right), \quad \text{MSD} = \ell^2 d^{-q} \frac{A(\ell)}{\Delta t} = h(\ell) d^{-q} \quad (8) \]

  with \( K \) is a functional of \( v \) (independent of \( d \))
  - Maximize \( h \) over \( \ell \) (independent of \( d \))
  - In some cases \( h(\ell) \) appears in a \( d \rightarrow \infty \) limiting diffusion in one d.o.f.
    \[ dy_t = -\frac{1}{2} h(\ell) v'(y_t) + \sqrt{h(\ell)} dw_t \quad (9) \]
Results on Tuning, Continued

- Optimal $\ell$ for RWM corresponds to $A = .234$ and $\Delta t \propto d^{-1}$
- Optimal $\ell$ for MALA corresponds to $A = .574$ and $\Delta t \propto d^{-1/3}$
- Optimal $\ell$ for HMC (with Verlet) corresponds to $A = .651$ and $\Delta t \propto d^{-1/4}$
Known Results on Tuning, Continued

- For RWM, $p = 1$, $q = -1$ and $K = \mathbb{E}[|v'|^2]$ – Optimal choice
  
  \[
  \Delta t^*_k = \frac{\ell^*_k}{d}, \quad \ell^*_k \sim \frac{1}{\sqrt{K}}
  \]

- Tends to zero as $d \to \infty$ or $v$ becomes rough
- For $v = v(x, x/\epsilon)$, $K \sim \epsilon^{-2}$ so $\Delta t^*_k \sim \epsilon^2$:
  
  \[
  \text{MSD} = \ell^*_k d^{-1} A(\ell^*_k) = O(\epsilon^2)
  \]

- An optimal choice exists, but performance degrades with roughness
Out of Equilibrium and Non-Product Results
Jourdain, Lelièvre, Miasojedow ’14,’15, Beskos, Roberts, Stuart, ’09, Beskos, Roberts, Thiery, Pillai, ’15

- RWM with nonstationary data similar to stationary limit. MALA more complicated, with no single optimal choice.
- Perturbations of the product measure case
  \[ d\mu \propto e^{-\Phi(x)}d\mu_0, \quad \mu_0 \text{ a product measure} \] (11)
- For multiscale \( V = V_0(x) + V_1(x, x/\epsilon) \) in finite \( d \) (ridged densities) limiting process is
  \[ dX_t = D_{V,\sigma^2}(X_t)dt + \sigma(X_t)dW_t \] (12)
  \[ \sigma^2(x) = \ell^2 a_0(x, \ell), \quad \text{Conditional Acceptance Rate} \] (13)
Harmonic Potential in 1D

Mathematica

\[ V(x) = \frac{k}{2} x^2, \quad \delta = k \Delta t \] (14)

- For RWM

\[ A(\delta) = \frac{2}{\pi} \arctan \sqrt{\frac{2}{\delta}} \] (15)
\[ F(\delta) = 2\delta A(\delta) - \frac{4\sqrt{2}\delta^{3/2}}{\pi(2+\delta)} \] (16)

- For MALA

\[ A(\delta) = \frac{2}{\pi} \arctan \sqrt{\frac{8}{\delta^3}} \] (17)
\[ F(\delta) = \delta(2 + \delta)A(\delta) - \frac{4\sqrt{2}\delta^{5/2}}{\pi(4+\delta(-2+\delta))} \] (18)

- MSD = \( k^{-1}F \)
Harmonic Potential in 1D, Continued

**RWM**

**MALA**

- **Mean Acceptance Rate**
  - For different values of $k$: 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024

- **Mean Squared Displacement (MSD)**
  - For different values of $k$: 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024

- **Time ($\Delta t$)**: 0.0, 0.25, 0.50, 0.75, 1.00
Harmonic Potential in 1D, Continued

RWM

MALA

Mean Acceptance Rate

$\delta = k \Delta t$

$k \times MSD$

$10^{-7.5} \ 10^{-5.0} \ 10^{-2.5} \ 10^{0.0} \ 10^{2.5}$

$10^{-7.5} \ 10^{-5.0} \ 10^{-2.5} \ 10^{0.0} \ 10^{2.5}$
Optimal RWM $\Delta t > 2 \times$ Optimal MALA $\Delta t$ (inside EM stability region)

Optimal MALA MSD $> 2 \times$ Optimal RWM MSD

Optimal acceptance rates deviate from $d \to \infty$ limit

RWM and MALA both have MSD $\to 0$ as $k = \epsilon^{-1} \to \infty$
1 Motivation & Background

2 Managing Roughness
   - Two Scale Potentials
   - Dissection of the Potential
   - Dissection on the Fly

3 Numerical Experiments

4 Summary & Acknowledgements
Assume

\[ V(x) = V(x, x, \epsilon) = V_0(x) + V_1(x, x/\epsilon) \]  \hspace{1cm} (19)

where

- \( V_0 \) is large scale, trapping contribution
- \( V_1 \) is bounded, rough contribution

In the case that \( V_1(x, y) \) is periodic in \( y \) in 1D, homogenization of overdamped Langevin leads to

\[ dX_t = -\mathcal{M}(X_t)\nabla \log Z(X_t)dt + \nabla \cdot \mathcal{M}(X_t)dt + \sqrt{2\mathcal{M}(X_t)}dW_t \]  \hspace{1cm} (20)

- Does not address sampling
- Suggests effective dynamics – position dependent proposals on a smoothed landscape
Naive Dissection in MC Methods

\[ R(x, y) = V(x) - V(y) + \log \frac{q(y \rightarrow x)}{q(x \rightarrow y)} \]

\[ = \left( V(x) - U(x) \right) - \left( V(y) - U(y) \right) \]

\[ \Delta(x) \]

\[ + \log \left( \frac{e^{-U(y)} q(y \rightarrow x)}{e^{-U(x)} q(x \rightarrow y)} \right) \]

\[ \tilde{R}(x, y) \]

Pick \( U \) and \( q \) such that:

1. \( U \) is captures the smooth, large scale features, and \( V - U \) is the bounded, rough contribution
2. \( q \) is a “good” proposal for \( e^{-U} \)

Smooth proposals on \( U \) and corrected by Metropolis for \( V \)
Naive Dissection in MC Methods

Lower Bound on Performance

- Lower bound on $R$

\[
R(x, y) = \Delta(x) - \Delta(y) + \tilde{R}(x, y) \\
\geq -\sup_{x'} \Delta(x') + \inf_{y'} \Delta(y') + \tilde{R}(x, y) = -\text{osc } \Delta + \tilde{R}(x, y)
\]

(22)

- Lower bounds on acceptance and MSD

\[
1 \land e^{R(x, y)} \geq e^{-\text{osc } \Delta} 1 \land e^{\tilde{R}(x, y)}
\]

(23)

\[
\text{MSD} = \mathbb{E}[(X_1 - X_0)^2] = \mathbb{E}[(y - x)^2 1 \land e^{R(x, y)}] \\
\geq e^{-\text{osc } \Delta} \mathbb{E}[(y - x)^2 1 \land e^{\tilde{R}(x, y)}]
\]

(24)

- In high $d$ product case, $\Delta = d\delta$ – ineffective lower bound
Local Entropy Smoothing
Chaudhari et al., '16, Chaudhari et al. '17

- Unlikely to have $V(x) = V_0(x) + V_1(x, x/\epsilon)$
- Inspired by works in nonconvex, nonlinear optimization (machine learning)
- Use Local Entropy approximation of $V$
  \[ V_\gamma(x) = -\beta^{-1} \log N(0, \gamma) * e^{-\beta V(x)} \]  
  (25)
- $V = V_\gamma + (V - V_\gamma)$
- Need to estimate a fast scale $\sqrt{\gamma}$
- Need an efficient method for estimating $V_\gamma$
Proposed Sampling Strategy
Thermostatted version of Chaudhari et al., '16, Chaudhari et al. '17

- Run short minibatch of

\[
dY_t^{(k)} = -\nabla V(Y_t^{(k)})dt - \gamma^{-1}(Y_t^{(k)} - x)dt + \sqrt{2}dW_t^{(k)}, \quad (26)
\]

and use these to estimate

\[
\nabla V_\gamma(x) = \gamma^{-1} \int (x - y)\rho(y; x)dy
\]

\[
= \gamma^{-1} \int (x - y)Z(x, \gamma)^{-1}e^{-V(y)-\frac{1}{2\gamma}\|y-x\|^2}dy \quad (27)
\]

\[
\approx \frac{1}{M} \sum_k \gamma^{-1}(x - Y_t^{(k)})
\]

Then Metropolize against \( V \).
Motivation & Background

Managing Roughness

Numerical Experiments
- Rough Harmonic Potential
- Rough Doublewell Potential
- Local Entropy Smoothing

Summary & Acknowledgements
Problem Setup

- Additive oscillatory term:
  \[ V(x) = \frac{1}{2} x^2 + \frac{1}{8} \cos(kx) \]  

- Product measures
- Use \( V_0 = \frac{1}{2} x^2 \) for modified MALA proposals
- Compute over a range of \( \Delta t \) to empirically identify the optimal value for different \( d \) and \( k \)
- \( 10^8 \) iterations per run.
Results

- As $k \to \infty$, Mod. MALA > RWM > MALA
- Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance
Results

- As $k \to \infty$, Mod. MALA $> \text{RWM} > \text{MALA}$
- Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance
As \( k \to \infty \), Mod. MALA > RWM > MALA

Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance
As $k \to \infty$, Mod. MALA $> \text{RWM} > \text{MALA}$

Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance
As $k \to \infty$, Mod. MALA $>\ RWM >\ MALA$

- Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance
Results

- As \( k \to \infty \), Mod. MALA \( \geq \) RWM \( \geq \) MALA
- Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance
As $k \to \infty$, Mod. MALA > RWM > MALA

Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance
As $d$ increases, the Mod. MALA scheme continues to outperform RWM.
Problem Setup

- Additive oscillatory term:
  
  \[ V(x) = (x^2 - 1)^2 + \frac{1}{8} \cos(kx) \]  

- Product measures
- Use \( V_0 = (x^2 - 1)^2 \) for modified MALA proposals
- Compute over a range of \( \Delta t \) to empirically identify the optimal value for different \( d \) and \( k \)
- \( 10^8 \) iterations per run.
Numerical Experiments

Rough Doublewell Potential

Results

- Mean Acceptance Rate
- MSD
- Optimal $t$

$\text{d} = 1$

$k$ values: $10^0, 10^1, 10^2, 10^3$

Graphs show data for RWM, MALA, and Mod. MALA.
Results

- Mean Acceptance Rate
  - $d = 5$
  - $RWM$, $MALA$, $Mod. MALA$
  - $k = 1, 2, 3, 4, 5$

- MSD
  - $d = 5$
  - $RWM$, $MALA$, $Mod. MALA$
  - $k = 1, 2, 3, 4, 5$

- Optimal $t$
  - $d = 5$
  - $RWM$, $MALA$, $Mod. MALA$
  - $k = 1, 2, 3, 4, 5$
Results

Numerical Experiments

Rough Doublewell Potential

Mean Acceptance Rate

\( d = 10 \)

RWM
MALA
Mod. MALA

Optimal \( \Delta t \)

\( d = 10 \)

RWM
MALA
Mod. MALA

MSD

\( d = 10 \)

RWM
MALA
Mod. MALA

k

100 101 102 103

k

10 5
10 4
10 3
10 2
10 1

Simpson (Drexel)

Rough Landscapes–CIRM 2018

September 18, 2018 28 / 34
Numerical Experiments

Rough Doublewell Potential

Results

![Graphs showing numerical experiments results for Mean Acceptance Rate, MSD, and Optimal t](image-url)
Results
Results
Results

Numerical Experiments

Rough Doublewell Potential

Mean Acceptance Rate

\[ d = 100 \]

\[ k \]

RWM
MALA
Mod. MALA

Mean Absolute Deviation (MSD)

\[ d = 100 \]

\[ k \]

Optimal \( t \)

\[ d = 100 \]

\[ k \]
As $d$ increases, the Mod. MALA scheme continues to outperform RWM
Problem Setup

- Multiplicative oscillatory term:

\[ V(x) = \frac{1}{2}x^2 + \frac{1}{8}e^{-10x^2}\cos(100x) \]  

(30)

- Product measures
- Precompute \( V_\gamma \) by quadrature with \( \gamma = 0.02 \)
Results

- Performance gain improves with $d$
Numerical Experiments

Local Entropy Smoothing

Results, Continued

- $d = 100$ case
Remarks & Open Problems

- Performance of MALA suffers on multiscale energy landscapes
- Conjecture: Similar challenges with other methods involving $\nabla V$, with $V$ a multiscale potential
- Certain limiting cases of MALA (1D Harmonic, and $d \to \infty$ product measure) show that roughness sends performance to zero – Is there a general result in finite $d$/non-product case?
- Can local entropy smoothing be made practical and exploited?
- Joint $\epsilon \to 0$ and $d \to \infty$ limit
- Restricted Observables
Acknowledgements

Collaborators  P. Plechac (U. Delaware)

Funding  NSF 1818716, US DOE DE-SC0012733

http://www.math.drexel.edu/~simpson/