Sampling from Rough Energy Landscapes

Gideon Simpson

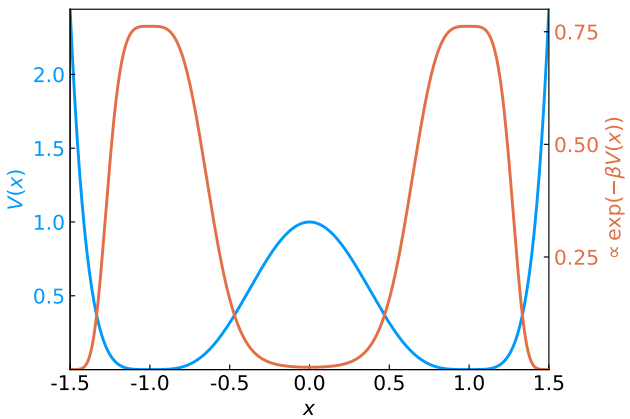
Department of Mathematics
Drexel University

September 18, 2018

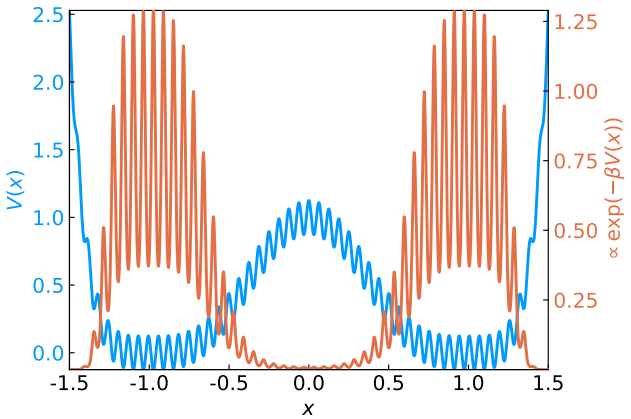
Joint with P. Plechac (U. Delaware)

Rough Energy Landscapes

and their Distributions



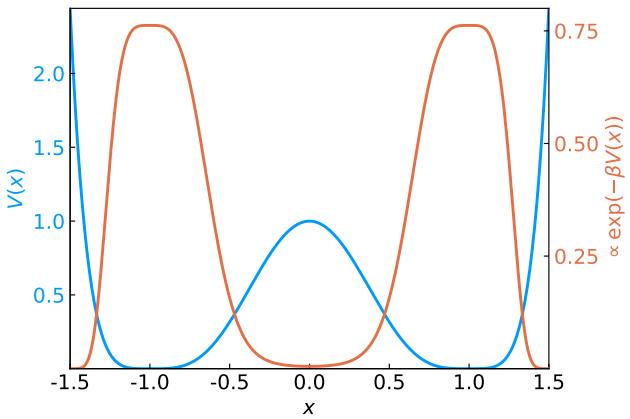
(a)



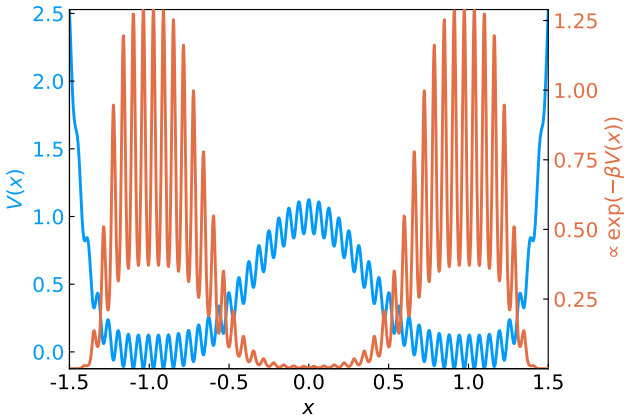
(b)

- (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



(c)



(d)

- **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

$$V(x) = V_0(x) + V_1(x, x/\epsilon) \tag{1}$$

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,

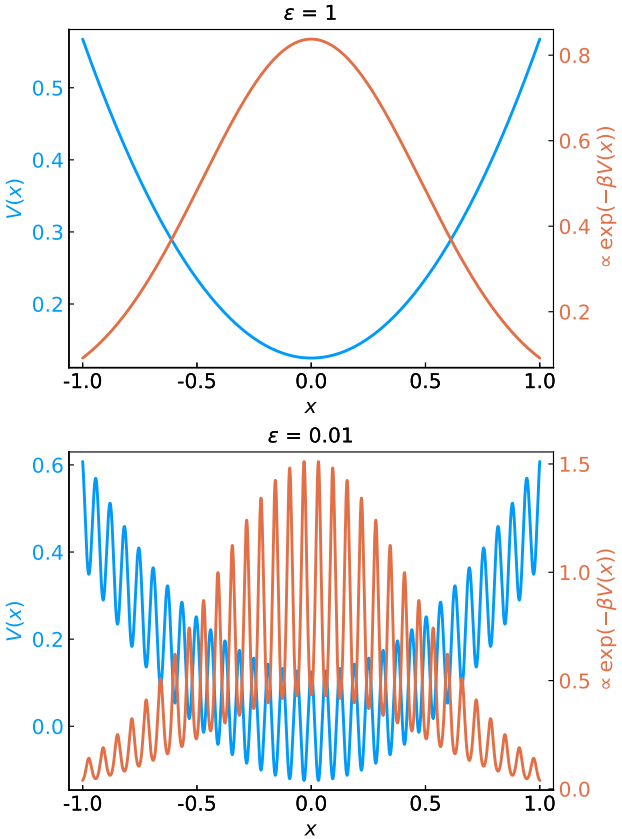
$$X_{n+1}^{\mathrm{p}} = x_n - \nabla V(X_n)\Delta t + \sqrt{2\beta^{-1}\Delta t}\xi_{n+1} \tag{2}$$

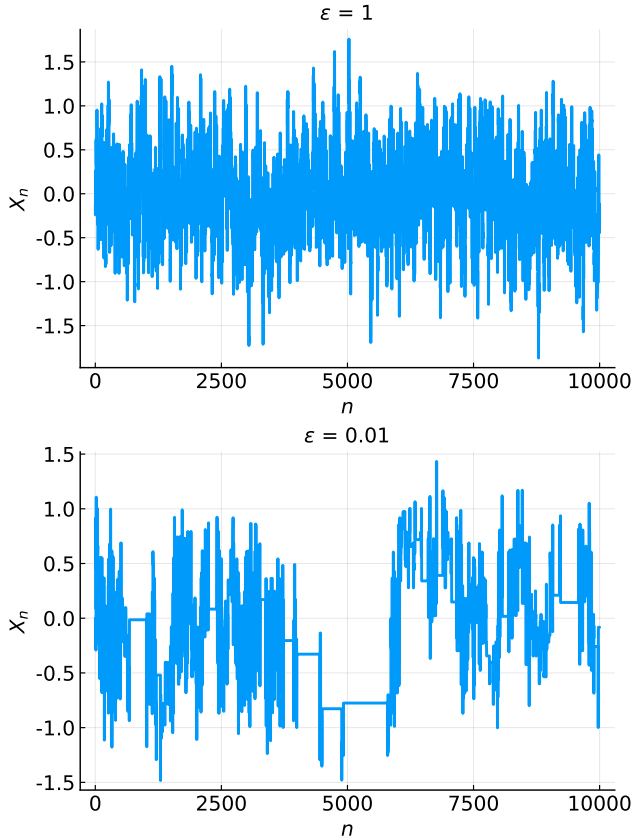
$$X_{n+1} = \begin{cases} X_{n+1}^{\mathrm{p}} & \text{with probability } 1 \wedge e^{R(X_n, X_{n+1}^{\mathrm{p}})} \\ X_n & \text{with probability } 1 - 1 \wedge e^{R(X_n, X_{n+1}^{\mathrm{p}})} \end{cases} \tag{3}$$

$$R(x,y) = \log \frac{e^{-\beta V(y)} q(y \to x)}{e^{-\beta V(x)} q(x \to y)} \tag{4}$$

- Use $\Delta t = 0.1$

MALA Example, Continued





- Increasing stagnation (poorer sampling) as $\epsilon \to 0$

Sample of Sampling Methods

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

MALA $X^{\mathrm{p}}_{n+1} = X_n - \nabla V(X_n) + \sqrt{2\Delta t}\xi_{n+1}$

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

Metropolized Langevin $(X^{\mathrm{p}}_{n+1}, P^{\mathrm{p}}_{n+1})$ from second order Langevin – marginalize out momentum

HMC $(X^{\mathrm{p}}_{n+1}, P^{\mathrm{p}}_{n+1})$ 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, $1 \wedge e^{R(x,x^{\mathrm{p}})}$,

$$R(x,y) = V(x) - V(y) + \log \frac{q(y \to x)}{q(x \to y)} \tag{5}$$

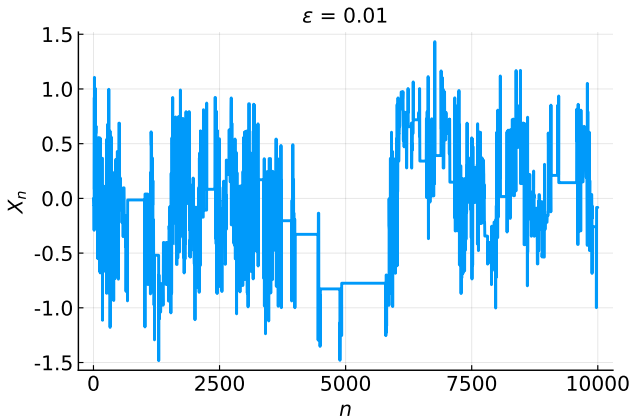
Choices of Parameters

RWM/MALA Need to choose step size Δt

Precond. MALA Need to choose Δt and preconditioning matrix

Langevin For a given splitting (there are many) need to choose Δt, damping, mass

HMC Δt, time of Hamiltonian trajectory, mass



Poor choice of parameter (Δt)?

Optimal Tuning

- Maximizing acceptance rate is the wrong objective
- Sending $\Delta t \to 0$ always sends the mean acceptance rate to 1
  For RWM

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

$$\text{Mean Rejection Rate} = \mathbb{E}[1 - a(x, y)] \leq \sqrt{\mathbb{E}[|R(x, y)|^2]} \lesssim \sqrt{\Delta t}$$

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

$$\text{MSD} = \mathbb{E}[|X_1^{(1)} - X_0^{(1)}|^2] \tag{6}$$

Results on Tuning in Equilibirum

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:

$$V_d(x) = \sum_{i=1}^{d} v(x_i), \quad v : \mathbb{R} \to \mathbb{R}, \quad e^{-V_d(x)} = \prod_{i=1}^{d} e^{-v(x_i)} \tag{7}$$

- d.o.f.'s only interact during accept/reject step
- As $d \to \infty$, ensemble average acceptance and MSD can be predicted:

$$A(\ell) = 2\Phi\left(-\frac{\ell^p}{2}\sqrt{K}\right), \quad \text{MSD} = \underbrace{\ell^2 d^{-q}}_{\Delta t} A(\ell) = h(\ell) d^{-q} \tag{8}$$

  with K is a functional of v (independent of d)
- Maximize h over ℓ (independent of d)
- In some cases $h(\ell)$ appears in a $d \to \infty$ limiting diffusion in **one** d.o.f.

$$dy_t = -\frac{1}{2} h(\ell) v'(y_t) + \sqrt{h(\ell)} dw_t \tag{9}$$

Results on Tuning, Continued

- Optimal ℓ for RWM corresponds to $A = .234$ and $\Delta t \propto d^{-1}$
- Optimal ℓ for MALA corresponds to $A = .574$ and $\Delta t \propto d^{-1/3}$
- Optimal ℓ 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

$$\Delta t_\star = \frac{\ell_\star^2}{d}, \quad \ell_\star \sim \frac{1}{\sqrt{K}} \tag{10}$$

- Tends to zero as $d \to \infty$ or v becomes rough
- For $v = v(x, x/\epsilon)$, $K \sim \epsilon^{-2}$ so $\Delta t_\star \sim \epsilon^2$:

$$\text{MSD} = \ell_\star^2 d^{-1} A(\ell_\star) = \mathrm{O}(\epsilon^2)$$

- 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

$$d\mu \propto e^{-\Phi(x)} d\mu_0, \quad \mu_0 \text{ a product measure} \tag{11}$$

- For multiscale $V = V_0(x) + V_1(x, x/\epsilon)$ in finite d (ridged densities) limiting process is

$$dX_t = \mathcal{D}_{V,\sigma^2}(X_t)dt + \sigma(X_t)dW_t \tag{12}$$

$$\sigma^2(x) = \ell^2 a_0(x, \ell), \quad \text{Conditional Acceptance Rate} \tag{13}$$

Harmonic Potential in 1D

Mathematica

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

- For RWM

$$A(\delta) = \frac{2}{\pi}\arctan\sqrt{\frac{2}{\delta}} \tag{15}$$

$$F(\delta) = 2\delta A(\delta) - \frac{4\sqrt{2}\delta^{3/2}}{\pi(2+\delta)} \tag{16}$$

- for MALA

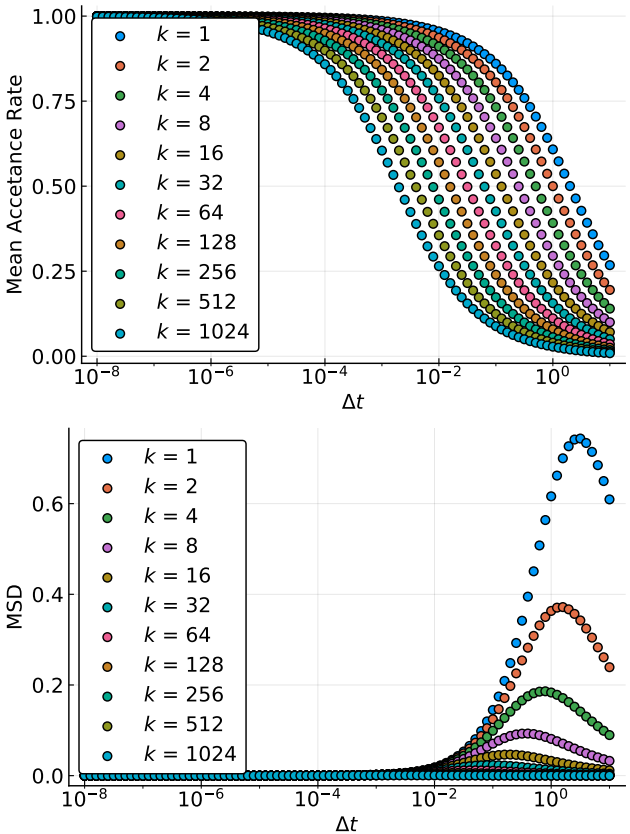
$$A(\delta) = \frac{2}{\pi}\arctan\sqrt{\frac{8}{\delta^3}} \tag{17}$$

$$F(\delta) = \delta(2+\delta)A(\delta) - \frac{4\sqrt{2}\delta^{5/2}}{\pi(4+\delta(-2+\delta)} \tag{18}$$

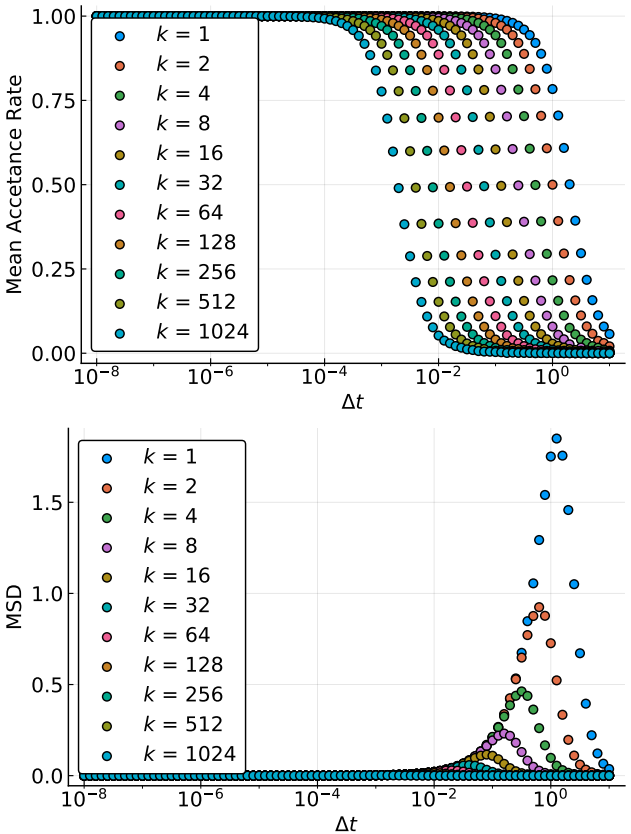
- MSD $= k^{-1}F$

Harmonic Potential in 1D, Continued

RWM

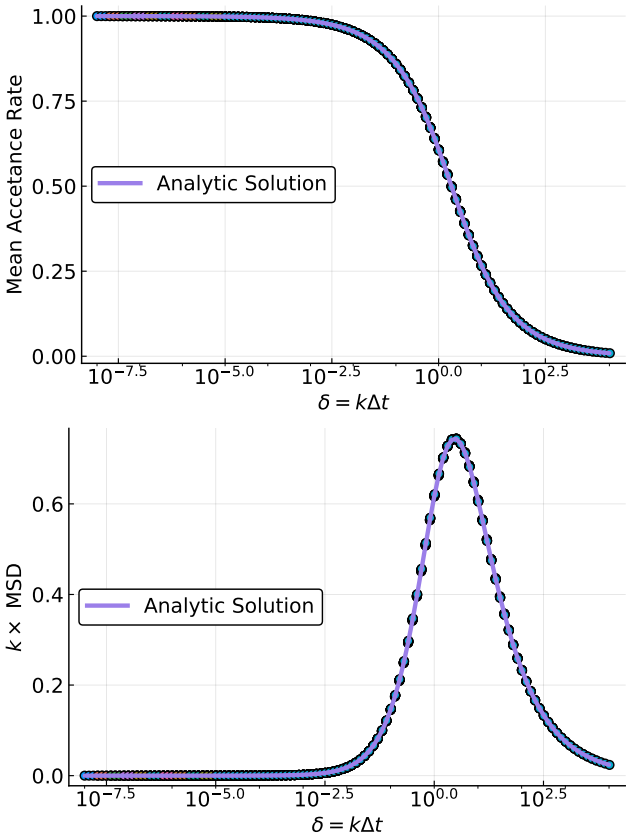


MALA

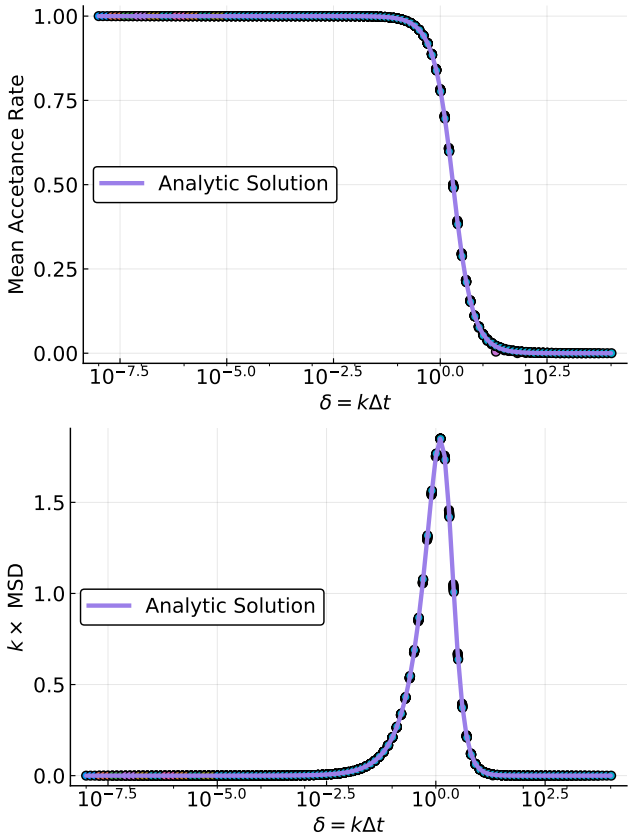


Harmonic Potential in 1D, Continued

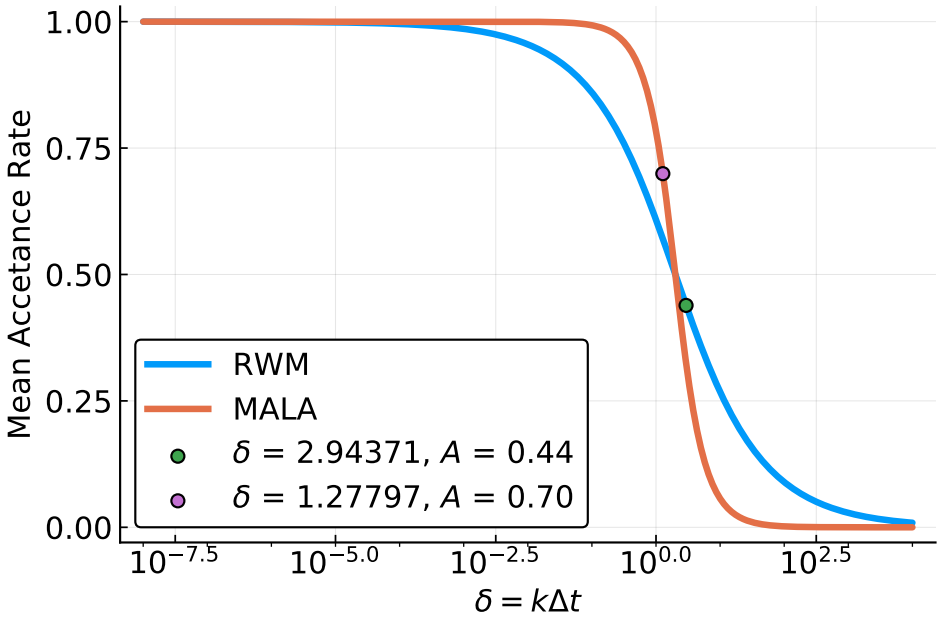
RWM

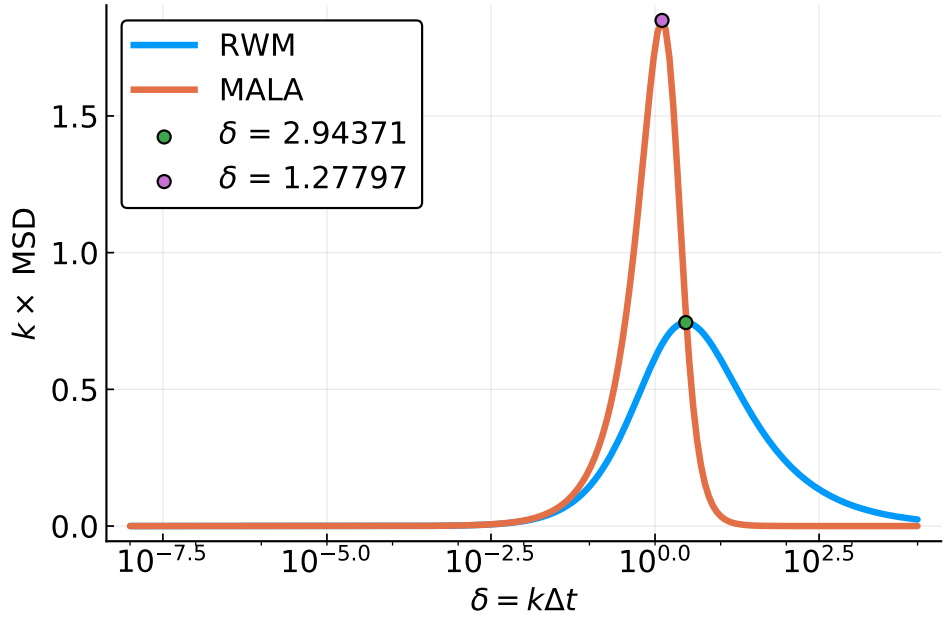


MALA



Harmonic Potential in 1D, Continued





- Optimal RWM $\Delta t > 2\times$ Optimal MALA Δt (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 $k = \epsilon^{-1} \to \infty$

Homogenization with a Two Scale Potential

Duncan, Kalliadasis, Pavliotis, Pradas '16, Ben Arous, Owhadi '03, Owhadi '03

- Assume

$$V(x) = V(x, x, \epsilon) = V_0(x) + V_1(x, x/\epsilon) \tag{19}$$

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

$$dX_t = -\mathcal{M}(X_t)\nabla \log Z(X_t)dt + \nabla \cdot \mathcal{M}(X_t)dt + \sqrt{2\mathcal{M}(X_t)}dW_t \tag{20}$$

  - Does not address sampling
  - Suggests effective dynamics – position dependent proposals on a smoothed landscape

Naive Dissection in MC Methods

$$\begin{aligned} R(x,y) &= V(x) - V(y) + \log \frac{q(y \to x)}{q(x \to y)} \\ &= \underbrace{(V(x) - U(x))}_{\Delta(x)} - (V(y) - U(y)) \\ &\quad + \underbrace{\log \left(\frac{e^{-U(y)} q(y \to x)}{e^{-U(x)} q(x \to y)} \right)}_{\tilde{R}(x,y)} \end{aligned} \tag{21}$$

- 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

$$\begin{aligned} R(x,y) &= \Delta(x) - \Delta(y) + \tilde{R}(x,y) \\ &\geq -\sup_{x'} \Delta(x') + \inf_{y'} \Delta(y') + \tilde{R}(x,y) = -\operatorname{osc} \Delta + \tilde{R}(x,y) \end{aligned} \tag{22}$$

- Lower bounds on acceptance and MSD

$$1 \wedge e^{R(x,y)} \geq e^{-\operatorname{osc} Delta} 1 \wedge e^{\tilde{R}(x,y)} \tag{23}$$

$$\begin{aligned} \mathrm{MSD} = \mathbb{E}[(X_1 - X_0)^2] &= \mathbb{E}[(y-x)^2 1 \wedge e^{R(x,y)}] \\ &\geq e^{-\operatorname{osc} \Delta} \mathbb{E}[(y-x)^2 1 \wedge e^{\tilde{R}(x,y)}] \end{aligned} \tag{24}$$

- In high d product case, $\Delta = d\delta$ – ineffective lower bound

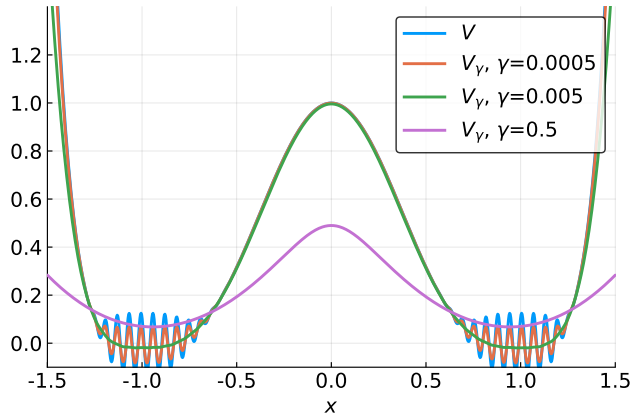
Local Entropy Smoothing

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

$$V_\gamma(x) = -\beta^{-1} \log N(0, \gamma) * e^{-\beta V(x)} \tag{25}$$

- $V = V_\gamma + (V - V_\gamma)$



- Need to estimate a fast scale $\sqrt{\gamma}$
- Need an efficient method for estimating V_γ

Proposed Sampling Strategy

Thermostatted version of Chaudhari et al., '16, Chaudhari et al. '17

- Run short minibatch of

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

and use these to estimate

$$\begin{aligned}\nabla V_\gamma(x) &= \gamma^{-1}\int (x-y)\rho(y;x)dy \\ &= \gamma^{-1}\int (x-y)Z(x,\gamma)^{-1}e^{-V(y)-\frac{1}{2\gamma}|y-x|^2}dy \\ &\approx \frac{1}{M}\sum_k \gamma^{-1}(x - Y_\tau^{(k)})\end{aligned} \tag{27}$$

Then Metropolize against V

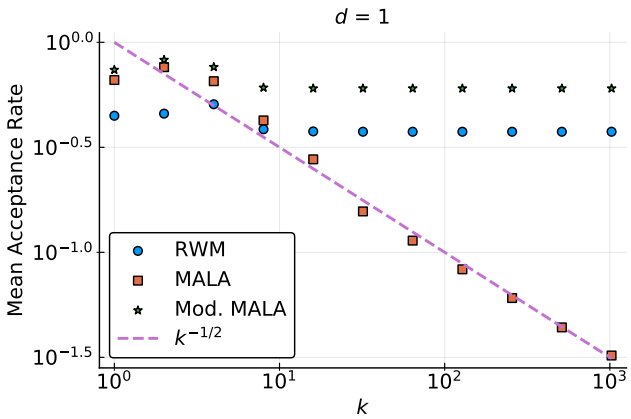
Problem Setup

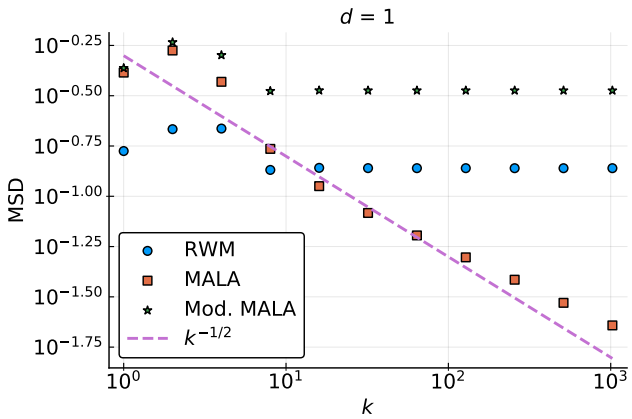
- Additive oscillatory term:

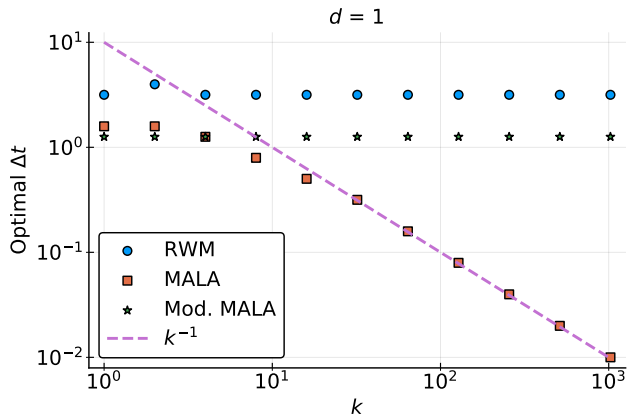
$$V(x) = \tfrac{1}{2}x^2 + \tfrac{1}{8}\cos(kx) \tag{28}$$

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

Results

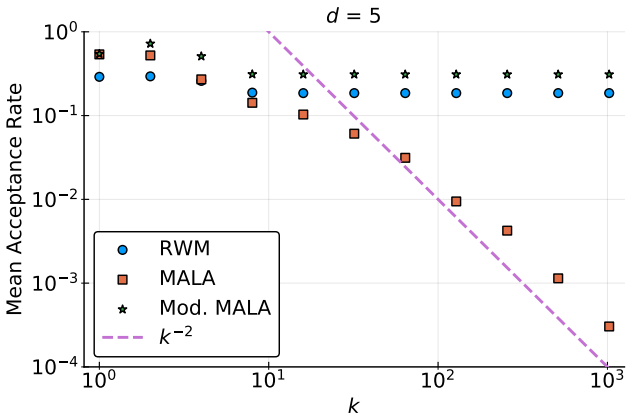


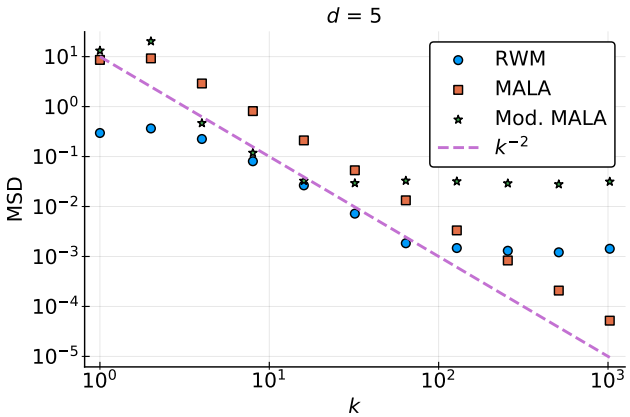


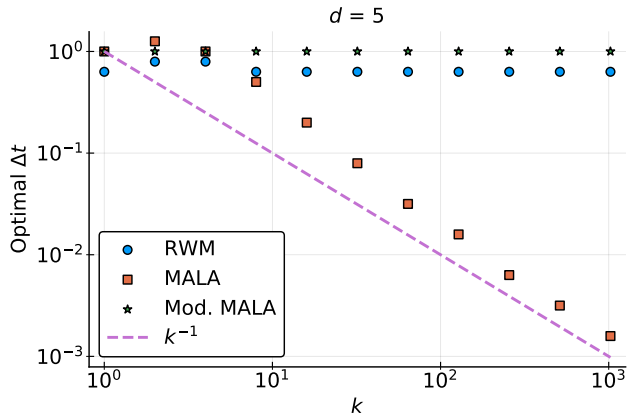


- As $k \to \infty$, Mod. MALA $>$ RWM $>$ MALA
- Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance

Results

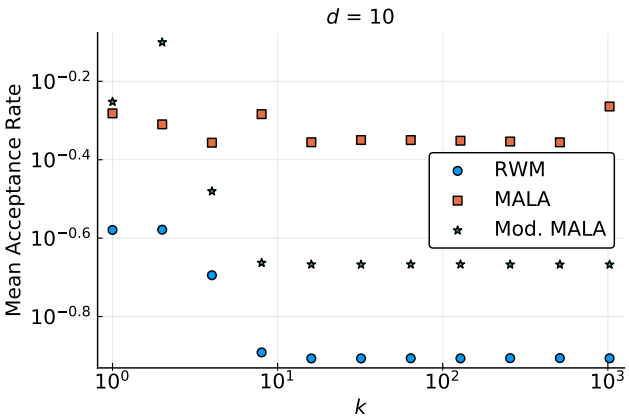


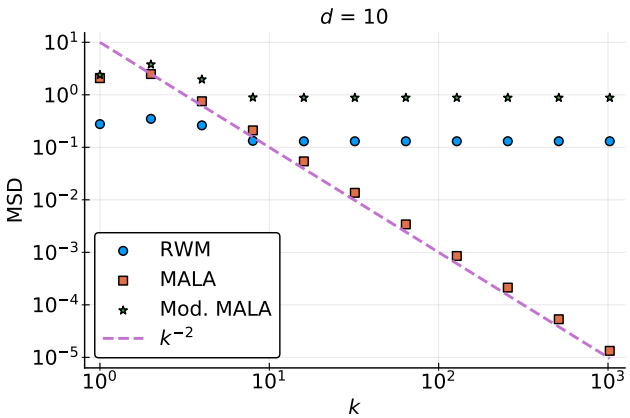


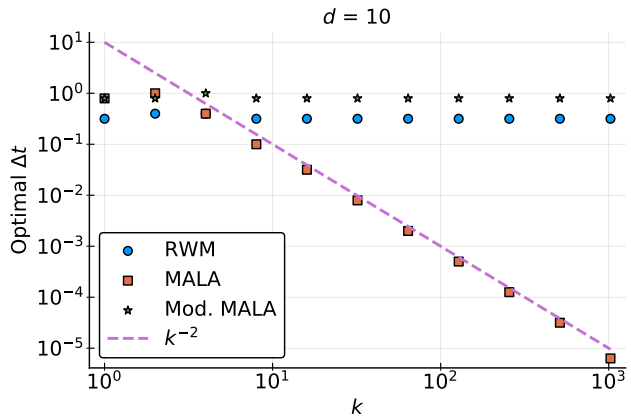


- As $k \to \infty$, Mod. MALA $>$ RWM $>$ MALA
- Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance

Results

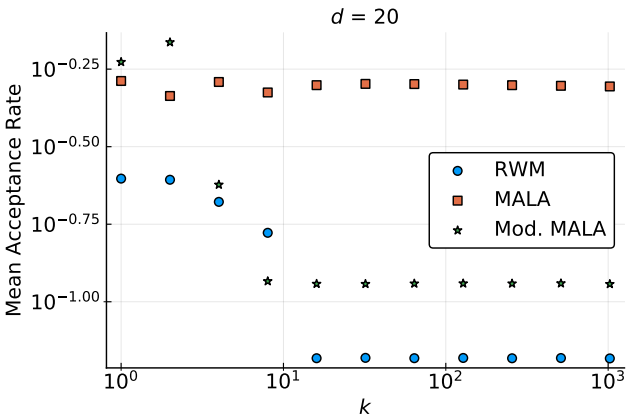


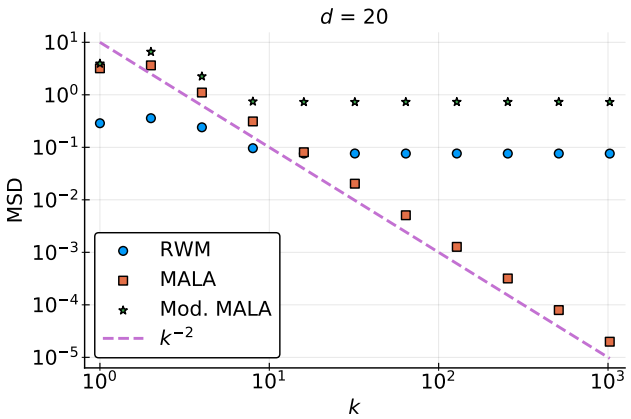


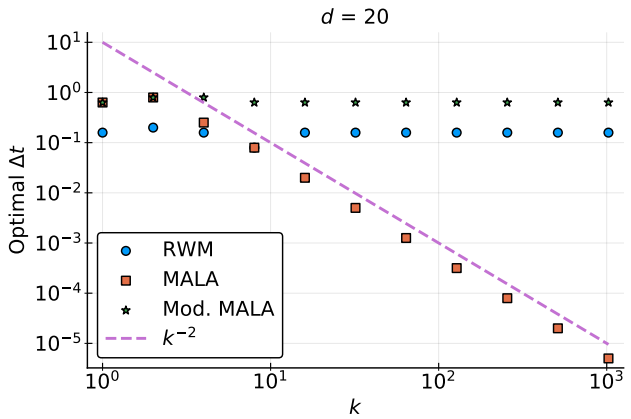


- As $k \to \infty$, Mod. MALA $>$ RWM $>$ MALA
- Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance

Results

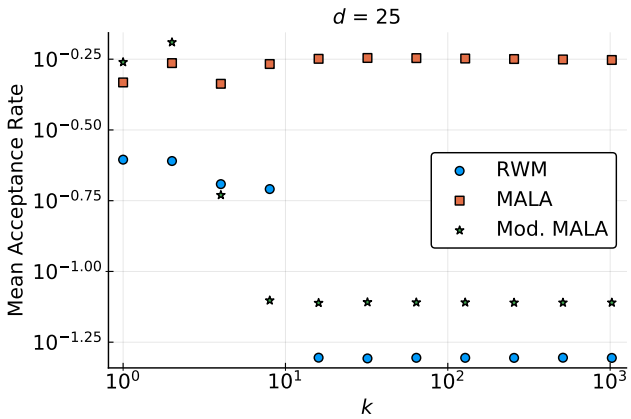


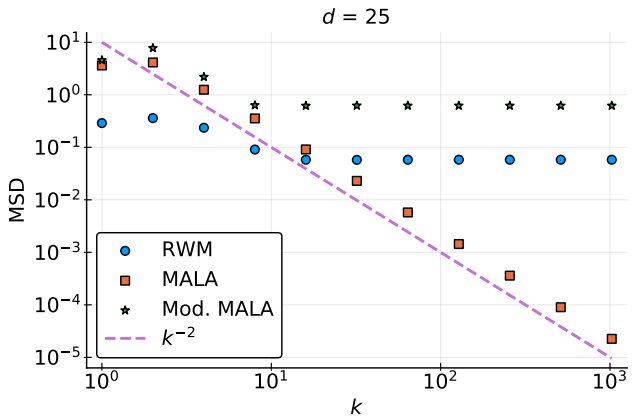


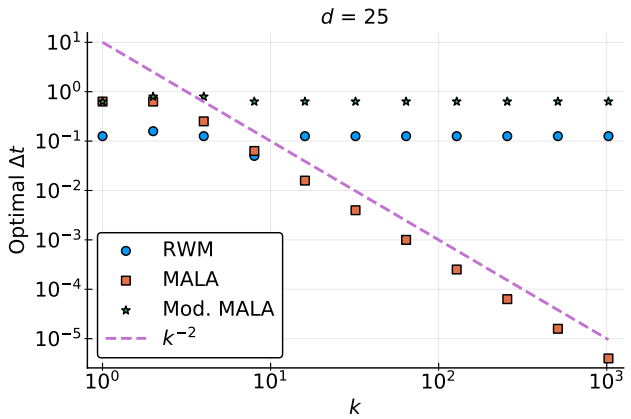


- As $k \to \infty$, Mod. MALA $>$ RWM $>$ MALA
- Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance

Results

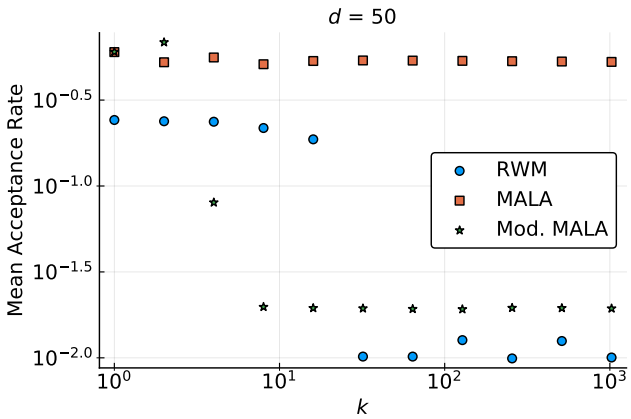


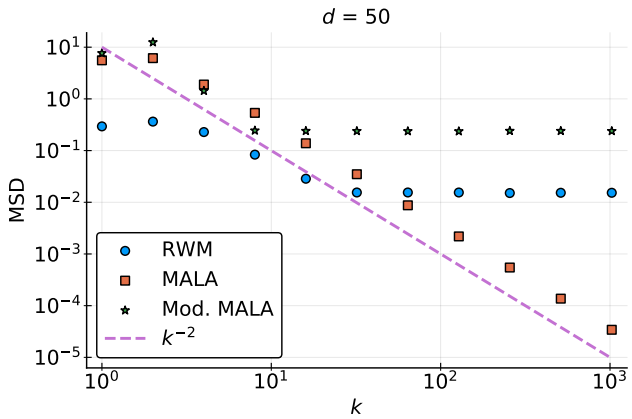


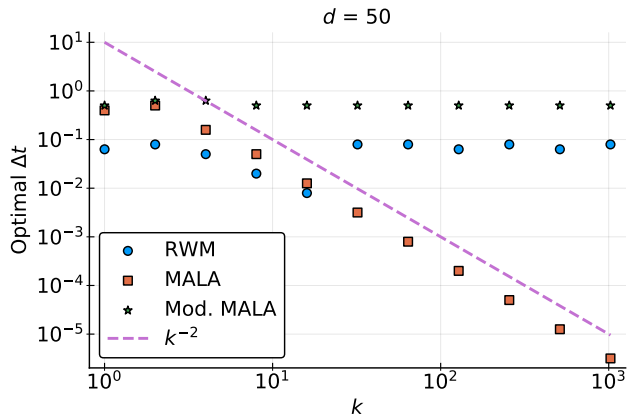


- As $k \to \infty$, Mod. MALA $>$ RWM $>$ MALA
- Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance

Results

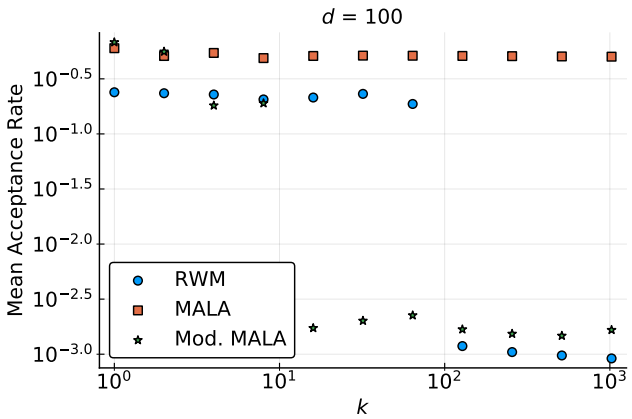


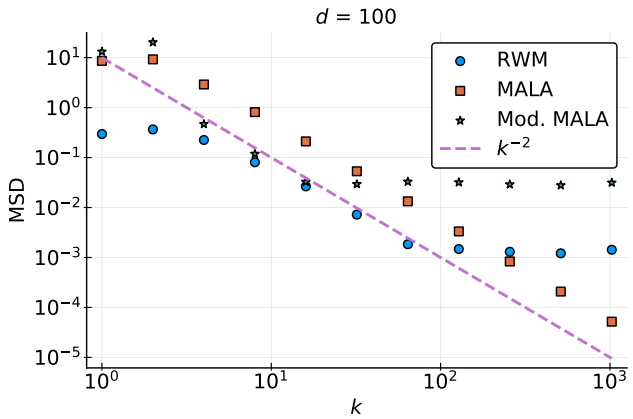


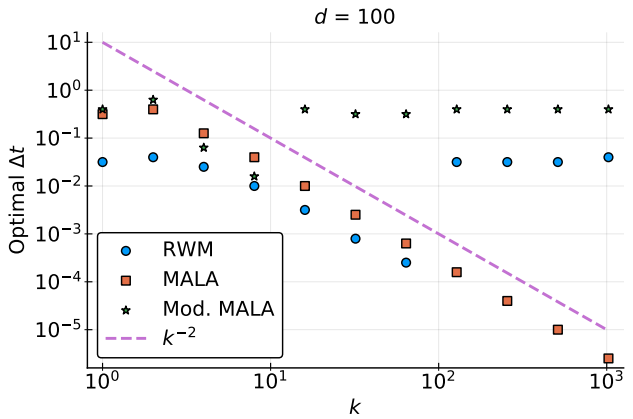


- As $k \to \infty$, Mod. MALA $>$ RWM $>$ MALA
- Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance

Results

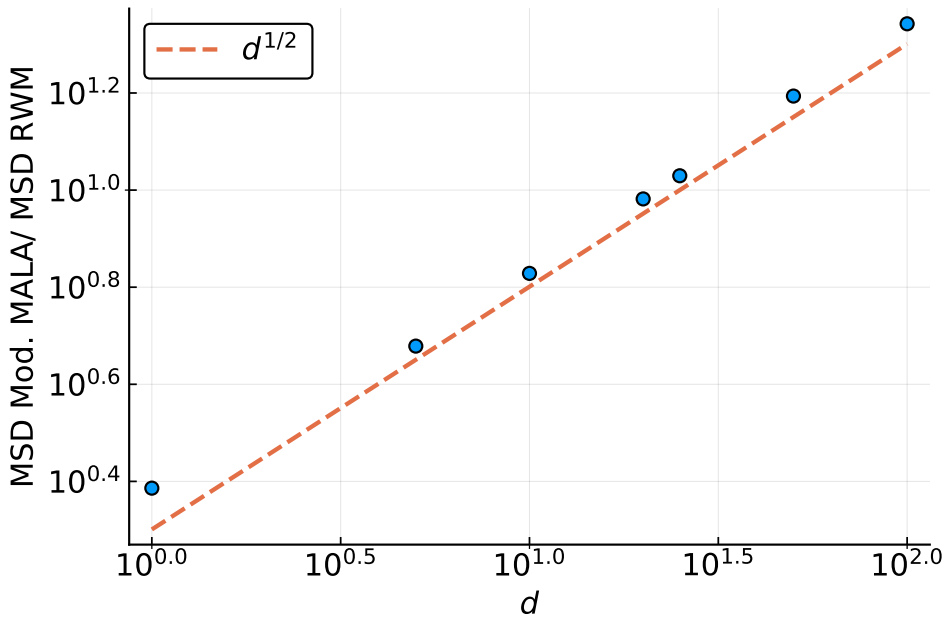






- As $k \to \infty$, Mod. MALA $>$ RWM $>$ MALA
- Even with low acceptance rates, larger time steps of Mod. MALA and RWM yield better performance

Results, Continued



- As d increases, the Mod. MALA scheme continues to outperform RWM

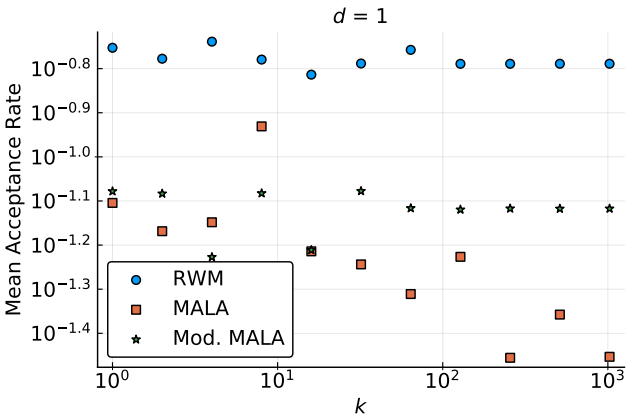
Problem Setup

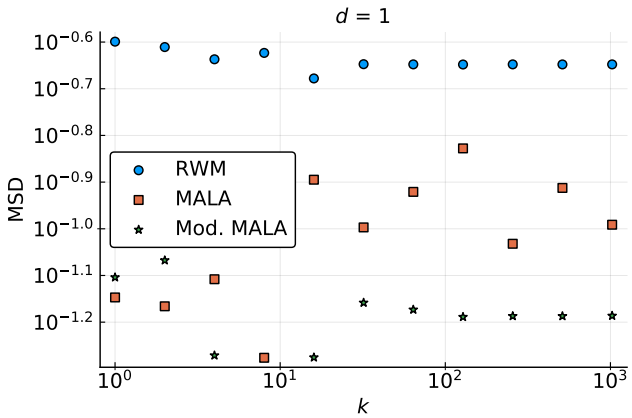
- Additive oscillatory term:

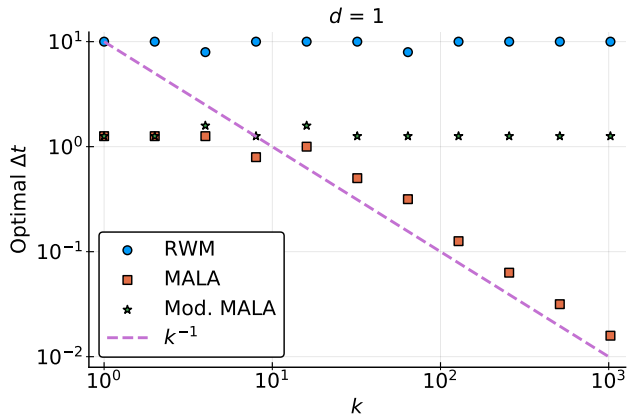
$$V(x) = (x^2 - 1)^2 + \tfrac{1}{8}\cos(kx) \tag{29}$$

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

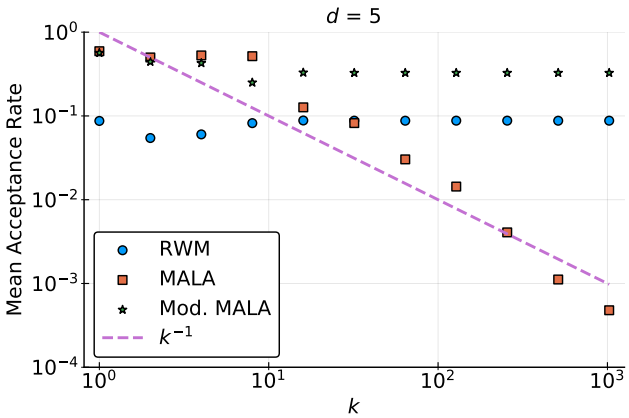
Results

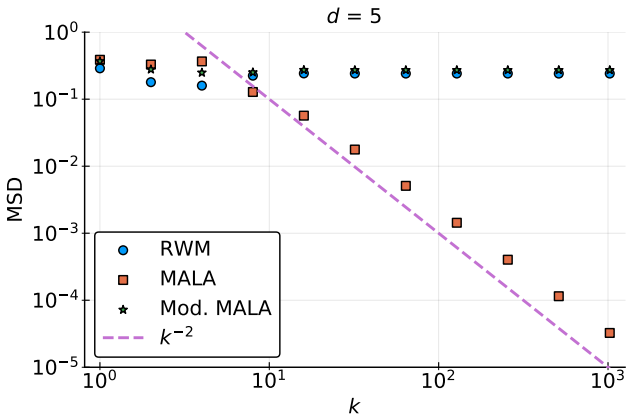

$d = 1$
Mean Acceptance Rate
$10^{-0.8}$
$10^{-0.9}$
$10^{-1.0}$
$10^{-1.1}$
$10^{-1.2}$
$10^{-1.3}$
$10^{-1.4}$
RWM
MALA
Mod. MALA
10^0
10^1
10^2
10^3
k

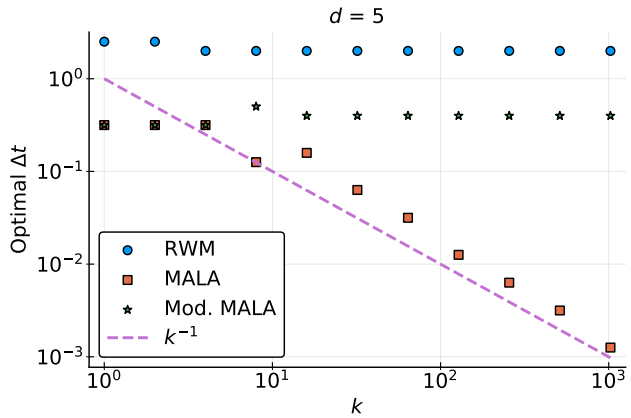

$d = 1$
MSD
$10^{-0.6}$
$10^{-0.7}$
$10^{-0.8}$
$10^{-0.9}$
$10^{-1.0}$
$10^{-1.1}$
$10^{-1.2}$
RWM
MALA
Mod. MALA
10^0
10^1
10^2
10^3
k


$d = 1$
Optimal Δt
10^1
10^0
10^{-1}
10^{-2}
RWM
MALA
Mod. MALA
k^{-1}
10^0
10^1
10^2
10^3
k

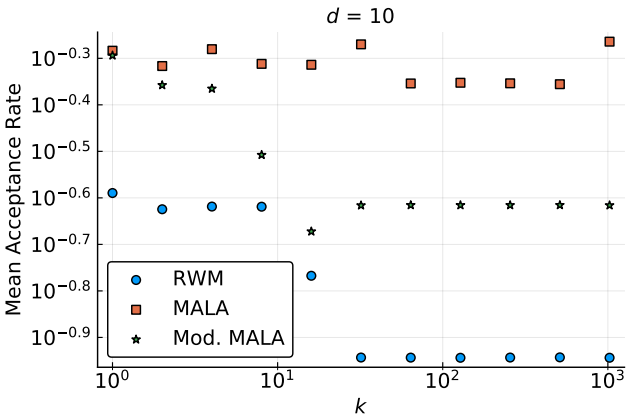
Results

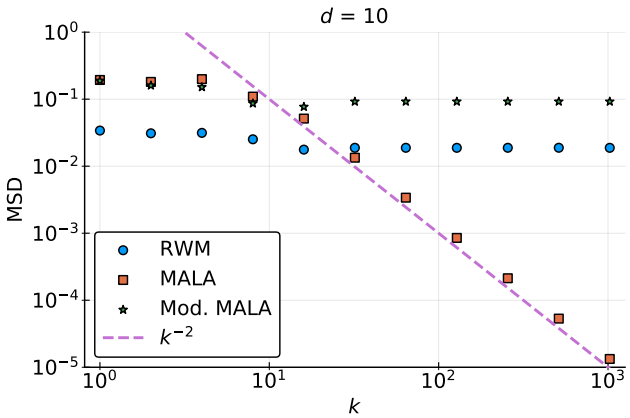

$d = 5$
Mean Acceptance Rate
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
RWM
MALA
Mod. MALA
k^{-1}
10^0
10^1
10^2
10^3
k

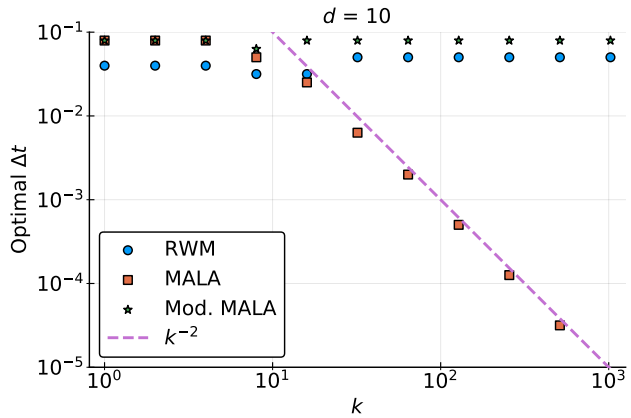

$d = 5$
MSD
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
RWM
MALA
Mod. MALA
k^{-2}
10^0
10^1
10^2
10^3
k


$d = 5$
Optimal Δt
10^0
10^{-1}
10^{-2}
10^{-3}
RWM
MALA
Mod. MALA
k^{-1}
10^0
10^1
10^2
10^3
k

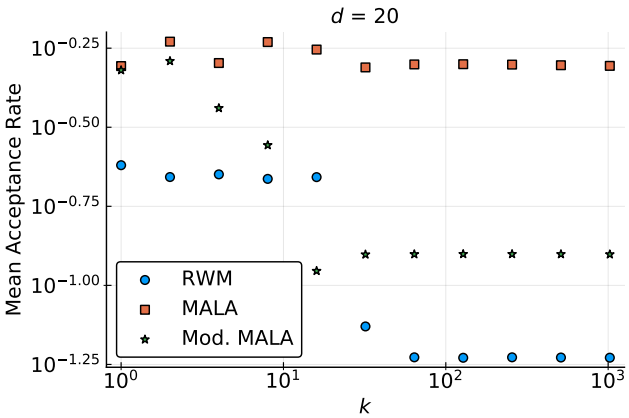
Results

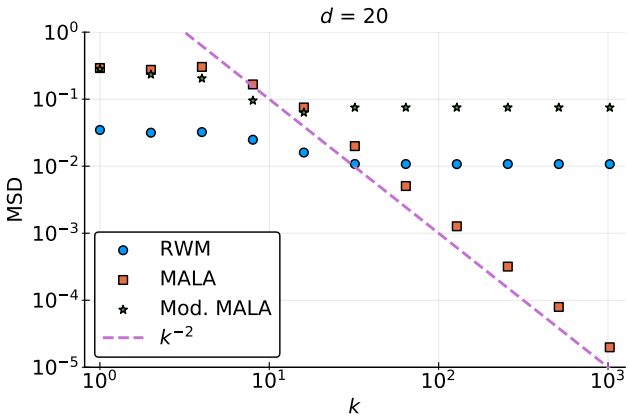

$d = 10$
Mean Acceptance Rate
$10^{-0.3}$
$10^{-0.4}$
$10^{-0.5}$
$10^{-0.6}$
$10^{-0.7}$
$10^{-0.8}$
$10^{-0.9}$
RWM
MALA
Mod. MALA
10^0
10^1
10^2
10^3
k

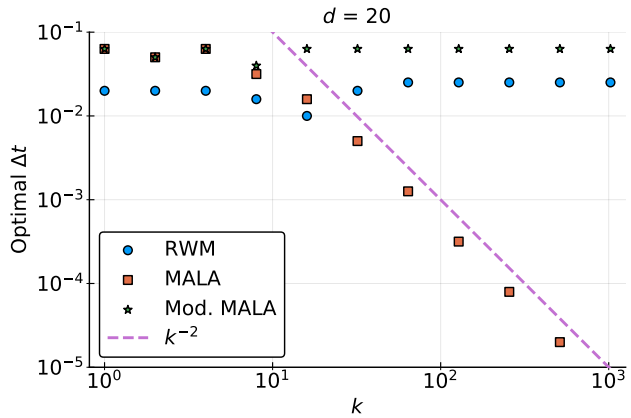

$d = 10$
MSD
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
RWM
MALA
Mod. MALA
k^{-2}
10^0
10^1
10^2
10^3
k


$d = 10$
Optimal Δt
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
RWM
MALA
Mod. MALA
k^{-2}
10^0
10^1
10^2
10^3
k

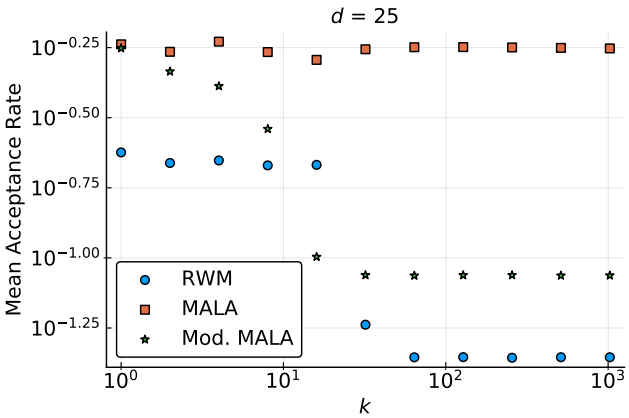
Results

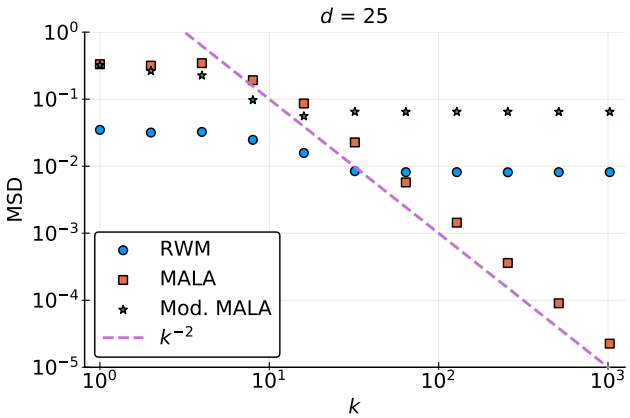

d = 20
Mean Acceptance Rate
$10^{-0.25}$
$10^{-0.50}$
$10^{-0.75}$
$10^{-1.00}$
$10^{-1.25}$
10^0
10^1
10^2
10^3
k
RWM
MALA
Mod. MALA

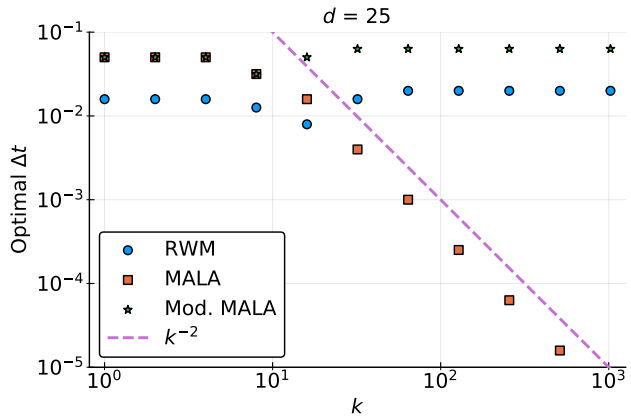

d = 20
MSD
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^0
10^1
10^2
10^3
k
RWM
MALA
Mod. MALA
k^{-2}


d = 20
Optimal Δt
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^0
10^1
10^2
10^3
k
RWM
MALA
Mod. MALA
k^{-2}

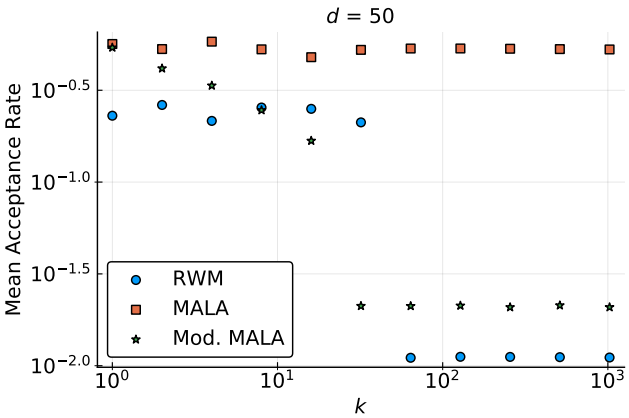
Results

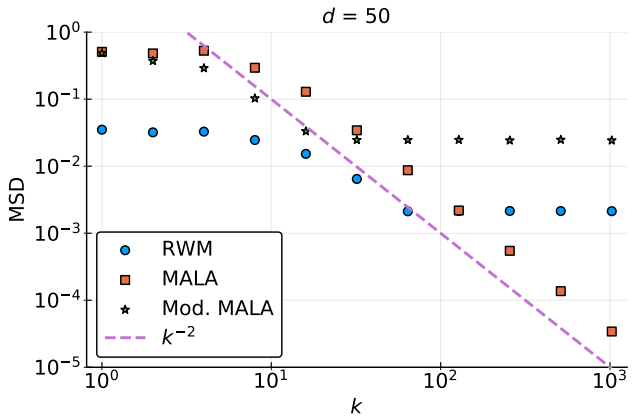

d = 25
Mean Acceptance Rate
$10^{-0.25}$
$10^{-0.50}$
$10^{-0.75}$
$10^{-1.00}$
$10^{-1.25}$
RWM
MALA
Mod. MALA
10^0
10^1
10^2
10^3
k

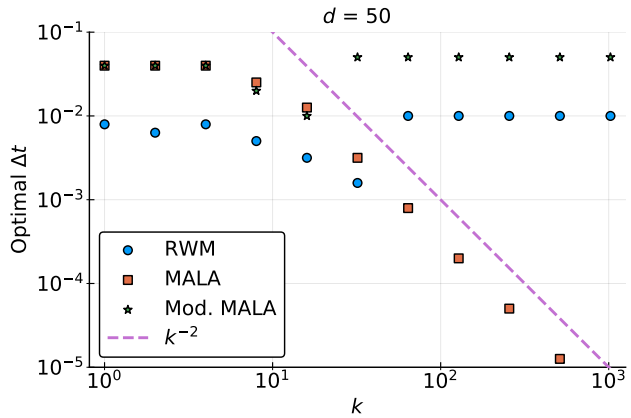

d = 25
MSD
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
RWM
MALA
Mod. MALA
k^{-2}
10^0
10^1
10^2
10^3
k


d = 25
Optimal Δt
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
RWM
MALA
Mod. MALA
k^{-2}
10^0
10^1
10^2
10^3
k

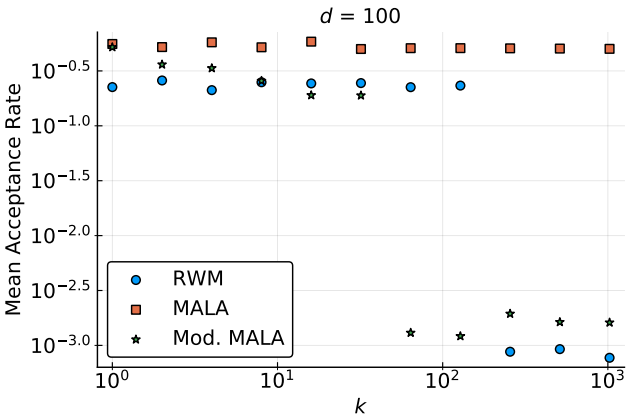
Results

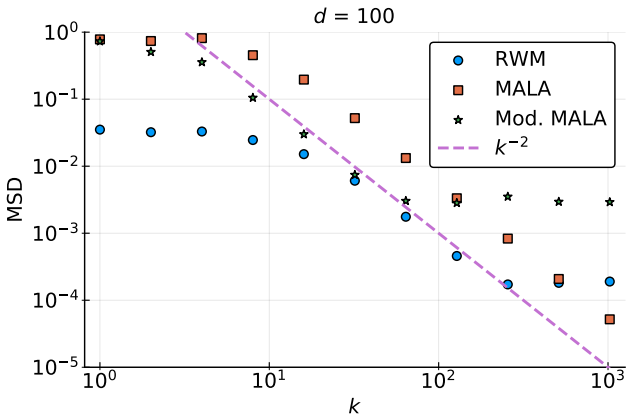

d = 50
Mean Acceptance Rate
$10^{-0.5}$
$10^{-1.0}$
$10^{-1.5}$
$10^{-2.0}$
RWM
MALA
Mod. MALA
10^0
10^1
10^2
10^3
k

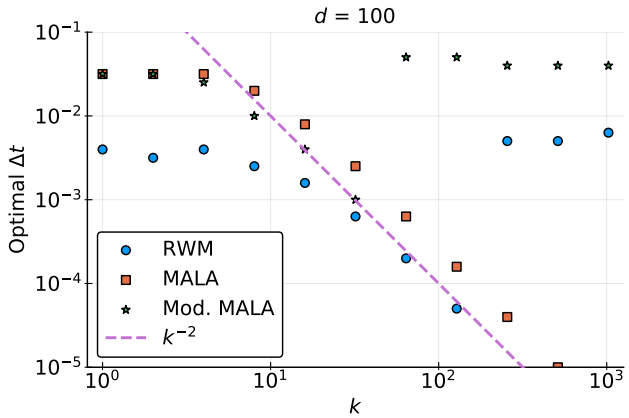

d = 50
MSD
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
RWM
MALA
Mod. MALA
k^{-2}
10^0
10^1
10^2
10^3
k


d = 50
Optimal Δt
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
RWM
MALA
Mod. MALA
k^{-2}
10^0
10^1
10^2
10^3
k

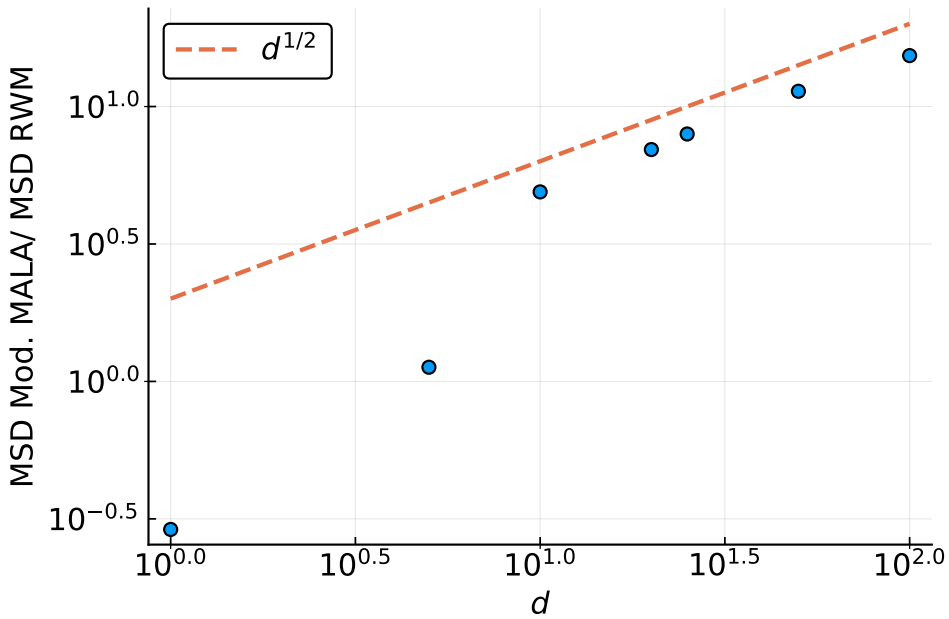
Results


d = 100
Mean Acceptance Rate
k
RWM
MALA
Mod. MALA


d = 100
MSD
k
RWM
MALA
Mod. MALA
k^{-2}


d = 100
Optimal Δt
k
RWM
MALA
Mod. MALA
k^{-2}

Results, Continued



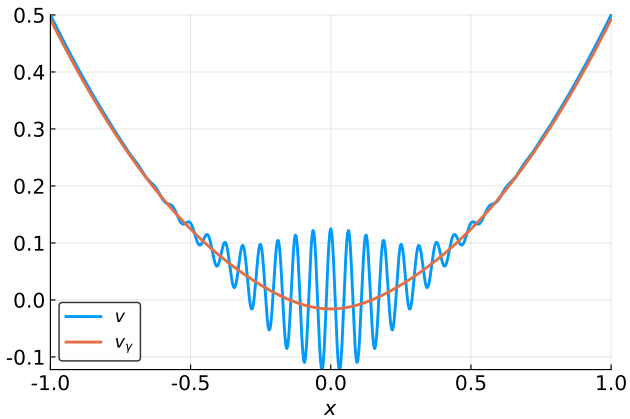
- As d increases, the Mod. MALA scheme continues to outperform RWM

Problem Setup

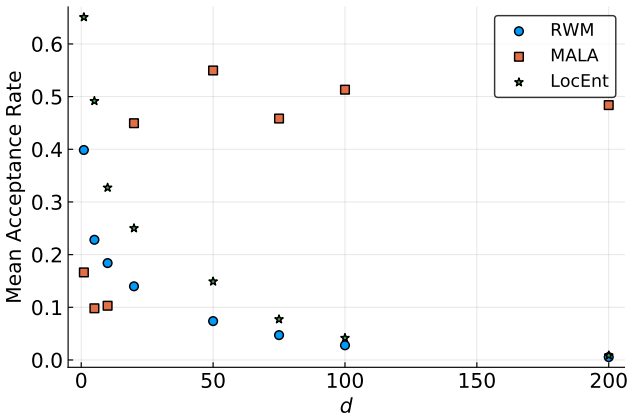
- Multiplicative oscillatory term:

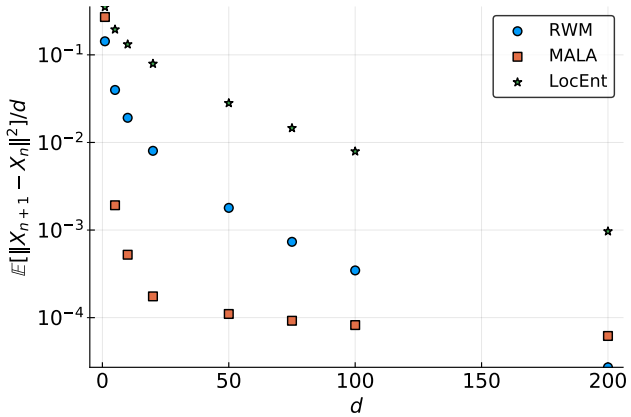
$$V(x) = \frac{1}{2}x^2 + \frac{1}{8}e^{-10x^2}\cos(100x) \tag{30}$$

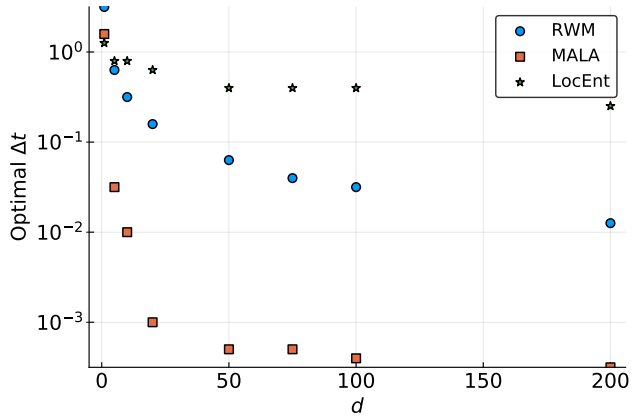
- Product measures
- Precompute V_γ by quadrature with $\gamma = 0.02$



Results

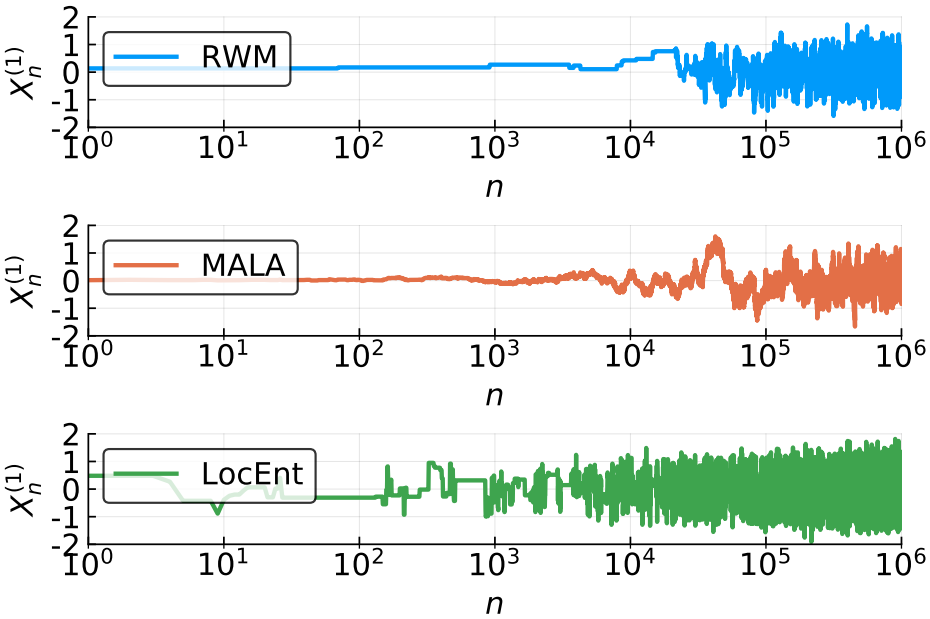


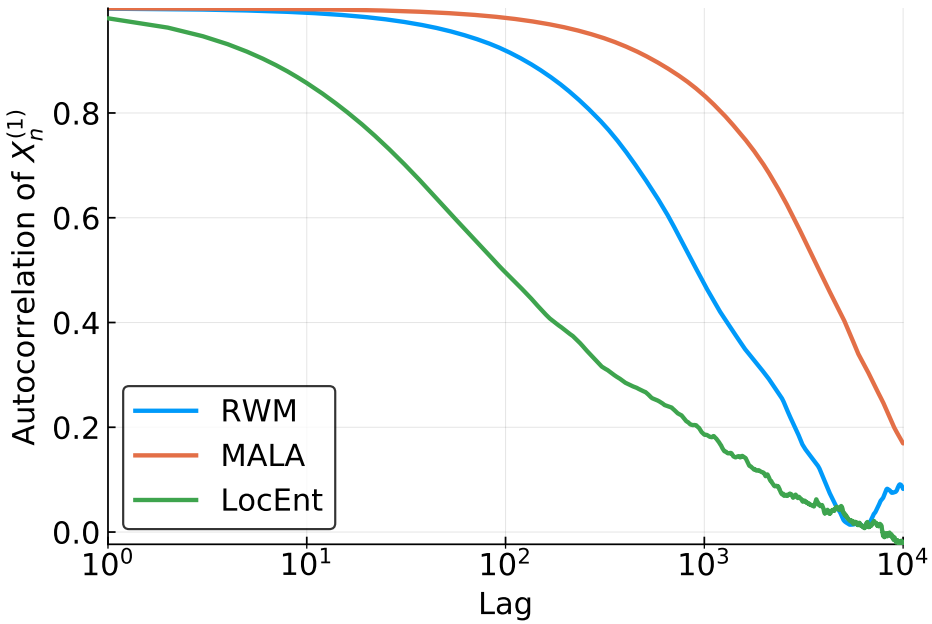




- Performance gain improves wiht d

Results, Continued





- $d = 100$ case

Remarks & Open Problems

- Performance of MALA suffers on multiscale energy landcapes
- Conjecture: Similar challenges with other methods involving ∇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/