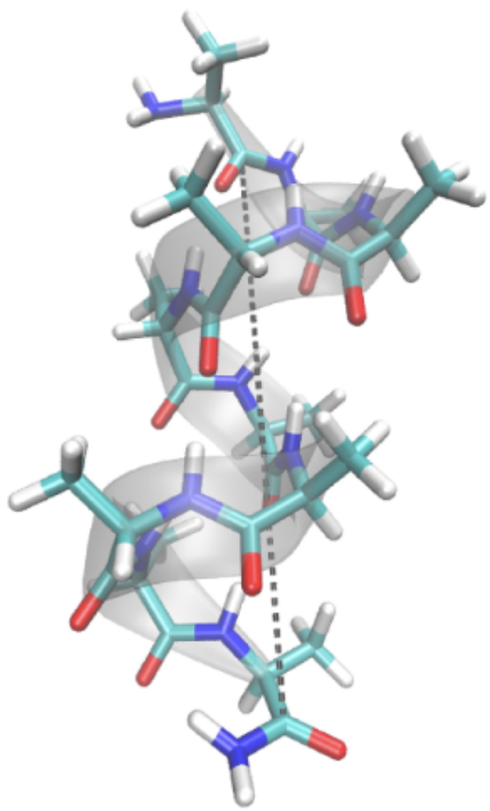
Thermostatic Controls for Noisy Gradient Systems and Applications to Machine Learning

Ben Leimkuhler
University of Edinburgh

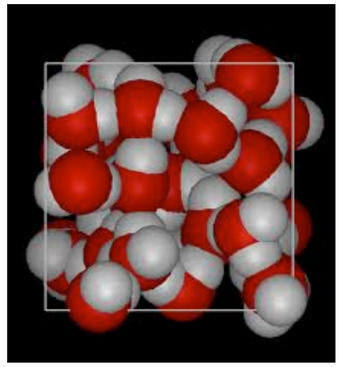
Joint work with **C. Matthews** (Chicago),
G. Stoltz (ENPC-Paris), **M. Tretyakov** (Nottingham)
X. Shang (PhD student, Edinburgh)

Our Group

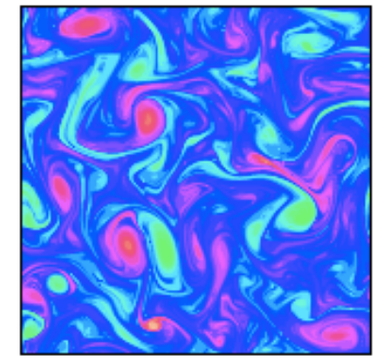


Molecular Dynamics Algorithms:

Gibbs sampling, numerical methods
coarse graining/mesocale modelling,
stochastic differential equations,
multiscale modelling, nonequilibrium



Water!



Software and Implementation in Consortium Code

ExTASY
Extensible Tools for Advanced Sampling and Analysis

NAMD
Scalable Molecular Dynamics

OpenMM & Amber

And don't forget: MD•M

The Father of Data Science

advising the president on
how to plan for a nuclear catastrophe

Bayesian Learning Application

Find best choice of parameters q given observations X

$$X = \{x_1, x_2, \ldots x_N\}$$

Challenges: data set very large
Ex: Netflix: 480000 users, 17000 ratings ⇒ 100M ratings!

Posterior probability density (from Bayes' Theorem):

$$p(q|X) \propto \exp(-U(q)), \qquad U(q) = -\log p(X|q) - \log p(q)$$

Data Scientist Thomas Bayes, U of Edinburgh, Class of 1721

Use Maximum Likelihood Estimate/"Subsampling":

$$\log p(X|q) \approx \frac{N}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \log p(x_i|q) \qquad \tilde{N} << N$$

The Sampling Problem

In high dimensions, the sampling problem cannot be solved using a direct integration method.

$\therefore$ Most sampling procedures are one of two types

***Monte-Carlo**:* Draw samples from a "prior" distribution accept or reject according to a *Metropolis test.*

***Discrete Dynamics**:* First define a Stochastic Differential Equation whose invariant distribution is the desired target; discretize the SDE to produce a Markov chain that approximates the desired distribution.

Problem: use stochastic dynamics to accurately sample a distribution with given positive smooth density

$$\rho \propto \exp(-U)$$

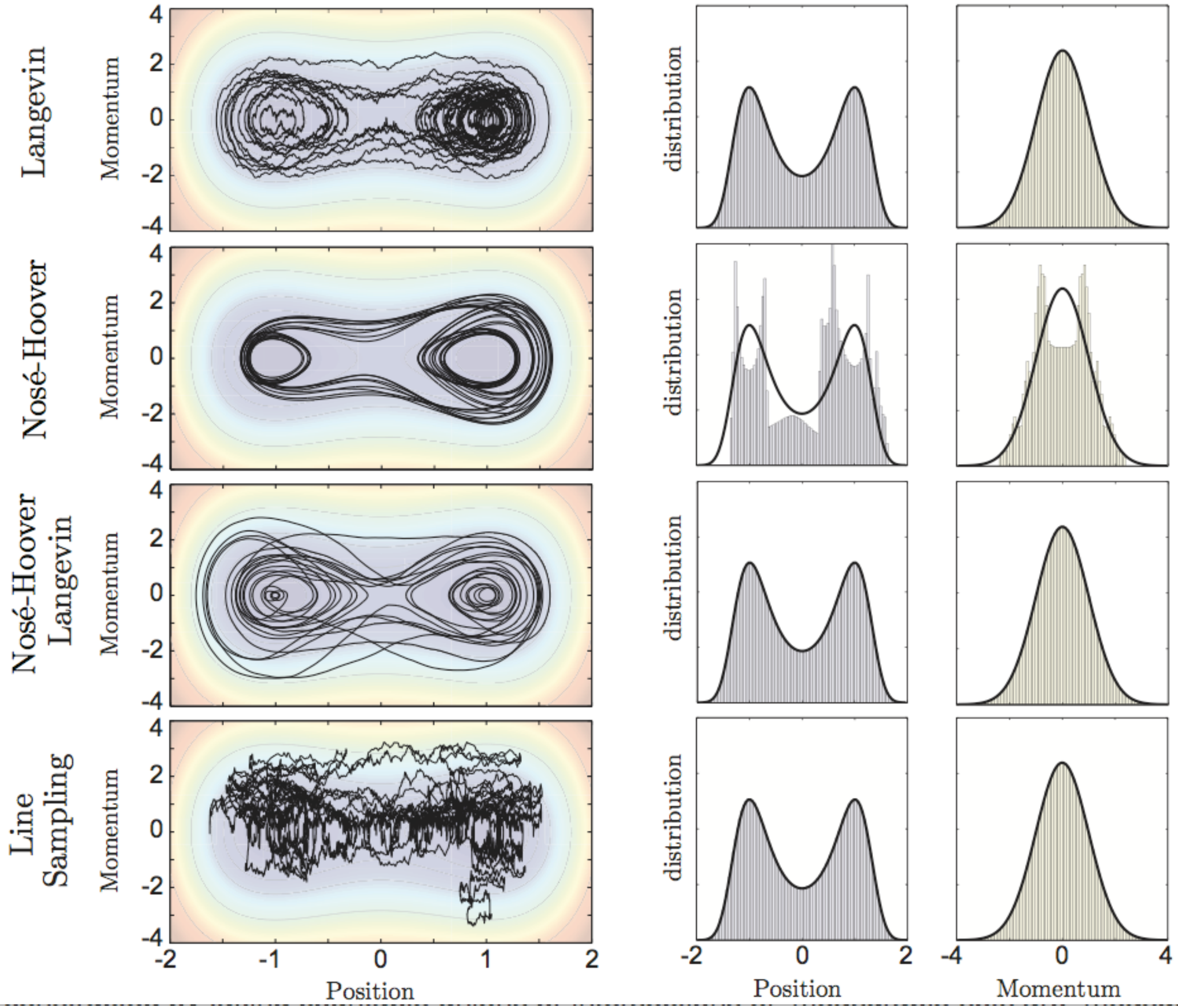
in case the force $-\nabla U$ can only be computed approximately

Examples:

Multiscale models
several flavors of hybrid **ab initio MD Methods**
QM/MM methods

…Many applications in **Bayesian Inference & Big Data Analytics**

From *L., Physical Review E, 2010*



With a clean gradient: $F(x) = -\nabla U(x)$

<u>Brownian Dynamics</u>

$$\mathrm{d}x = F(x)\mathrm{d}t + \sqrt{2}\mathrm{dW}$$

- SDEs which can be solved to generate a path $x(t)$
- Under typical conditions, for almost all paths,

$$\lim_{\tau\to\infty} \tau^{-1} \int_0^\tau \varphi(x(t))\mathrm{d}t = \int_\Omega \varphi(x)\rho(x)\mathrm{d}x$$

How to discretize?

Euler-Maruyama? Stochastic Heun?

Euler-Maruyama Method

$$x_0 \to x_1 \to \ldots \to x_N, \qquad Nh = \tau$$

discrete Brownian path

$$x_{n+1} = x_n + \cancel{hF(x_n)} + \sqrt{2h}R_n$$

$$R_n \sim \mathcal{N}(0,1)$$

Leimkuhler-Matthews Method [L. & Matthews, AMRX, 2013]

$$x_{n+1} = x_n + hF(x_n) + \sqrt{h/2}(R_n + R_{n-1})$$

[L., Matthews & Stoltz, IMA J. Num. Anal., 2015]

[L., Matthews & Tretyakov, Proc Roy Soc A, 2014]

Theorem ***[BL-CM-MT Proc Roy Soc A 2014]***
For the **L-M method**, under suitable conditions,

$$\mathbf{E}\varphi(\mathrm{X}_x(\tau)) - \mathbf{E}\varphi(\mathrm{X}_N) = C_0(\tau, x)h + C(\tau, x)h^2$$

$$|C_0(\tau, x)| \leq K_0(1 + |x|^\eta)e^{-\lambda_0 \tau}$$

$$\lambda_0, \lambda > 0$$

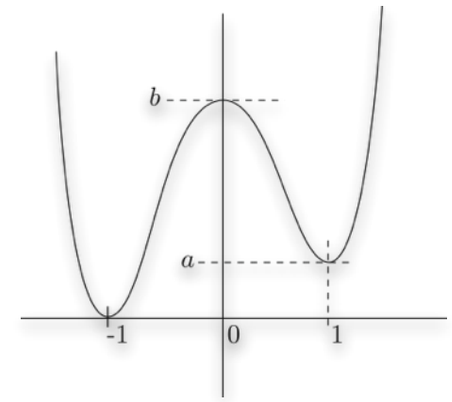
$$|C(\tau, x)| \leq K(1 + |x|^\eta e^{-\lambda \tau})$$

Weak first order -> **weak asymptotic second order**

exponentially fast in time

with constants that can be estimated using
Kolmogorov equations

Uneven Double Well

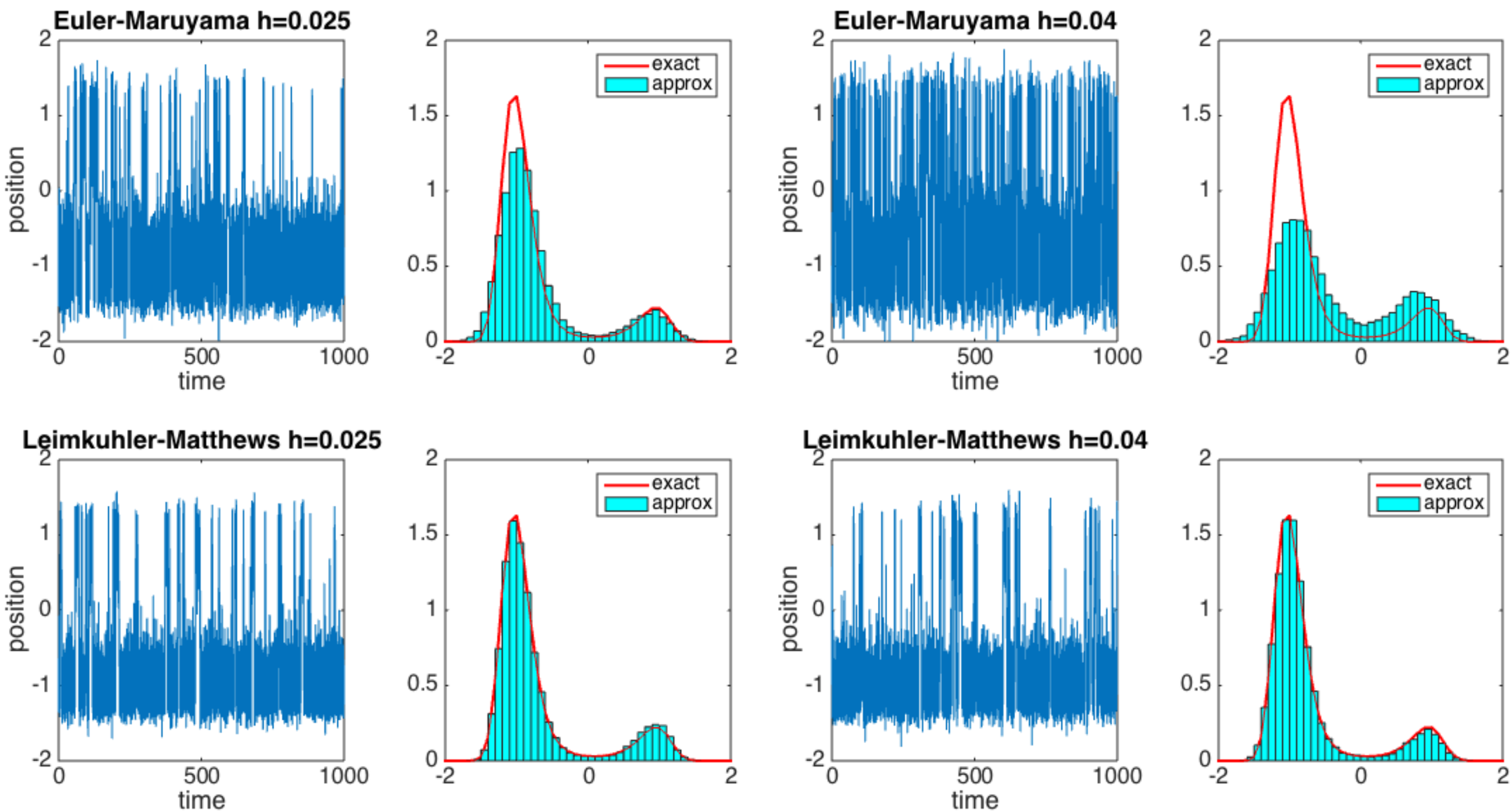


small stepsize

large stepsize

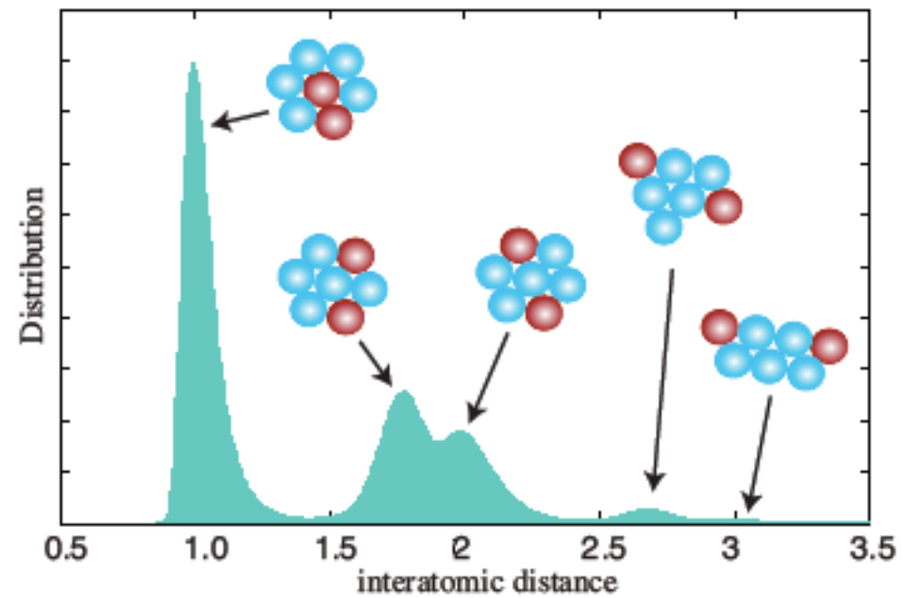
E-M

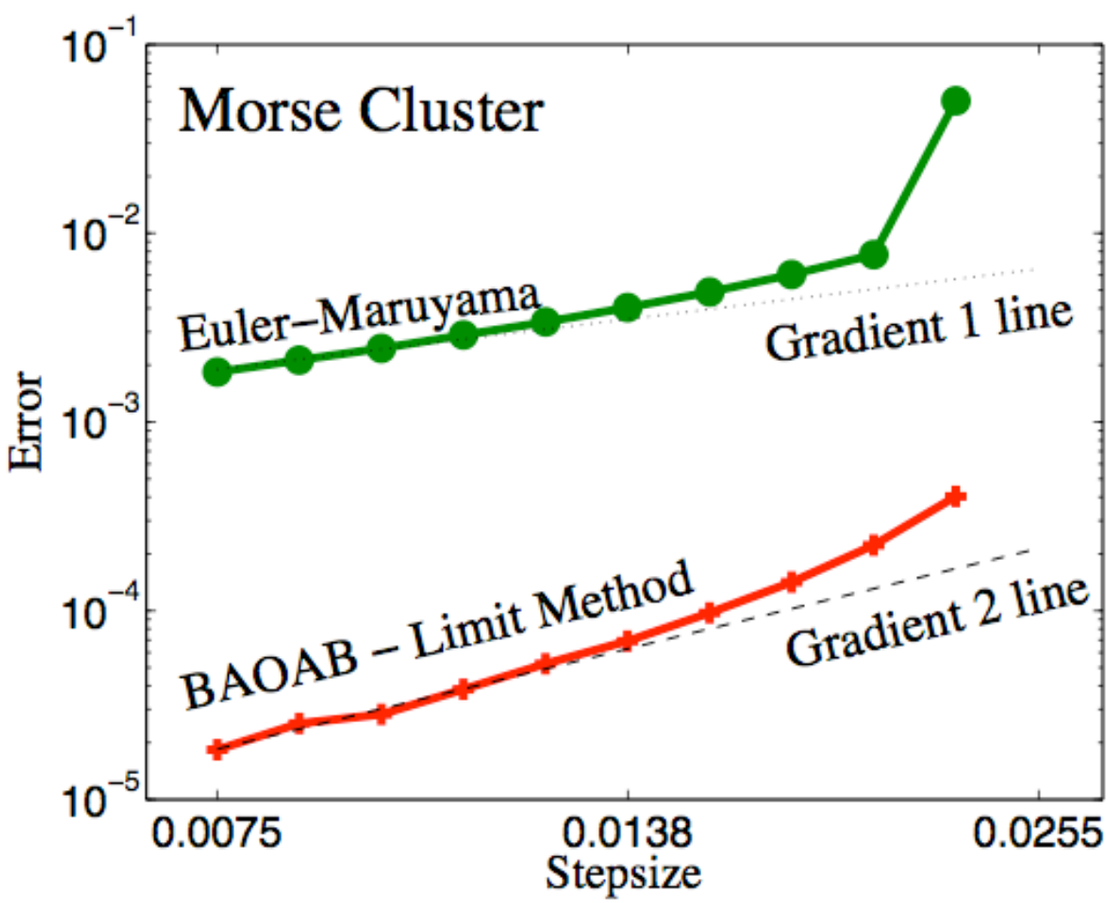
L-M

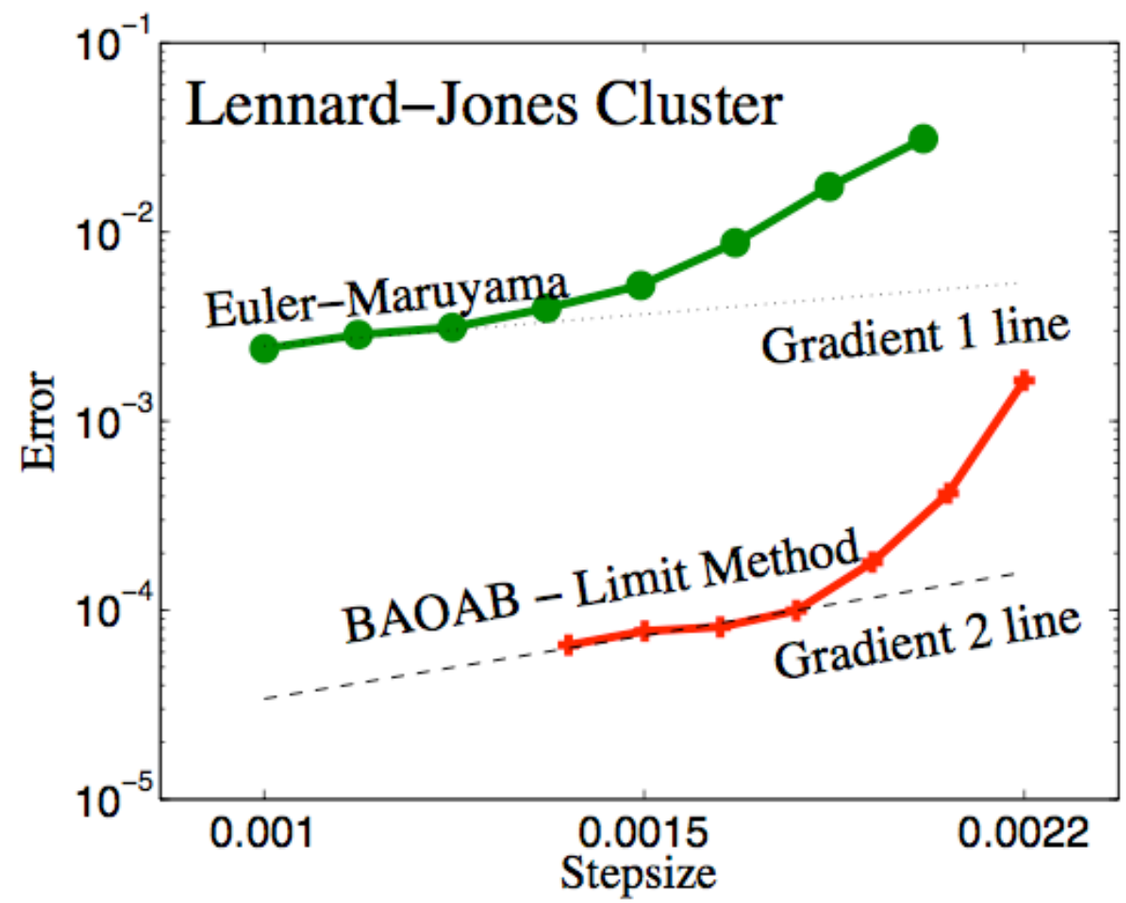


Morse and Lennard Jones Clusters

binned radial density for comparison







Accuracy ≠ Sampling Efficiency

Most sampling calculations are performed in the pre-converged regime (not at infinite time).

The challenge is often effective search in a high dimensional space riddled with entropic and energetic barriers

Brownian (first order) dynamics is ***"non-inertial"***

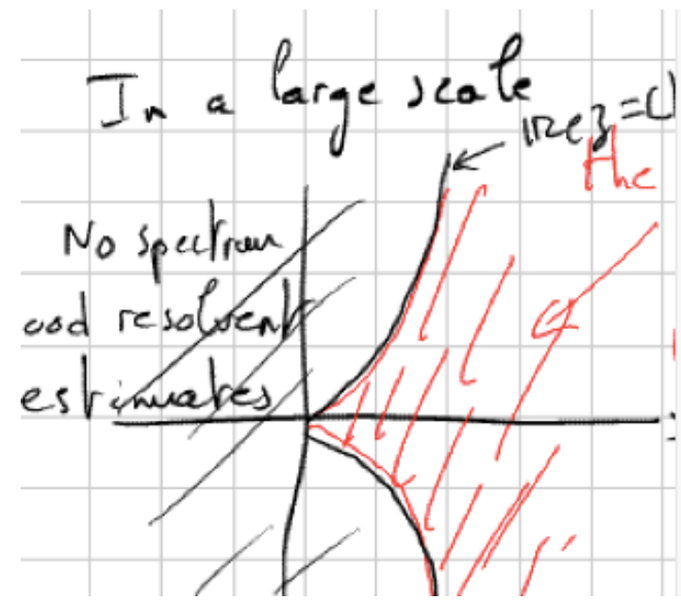
Langevin (inertial) stochastic dynamics, at low or modest friction, can enhance diffusion in systems with rough landscapes.

Langevin Dynamics

$$\mathrm{d}x = M^{-1} p \mathrm{d}t$$

$$\mathrm{d}p = -\nabla U \mathrm{d}t - \gamma M^{-1} p \mathrm{d}t + \sqrt{2\beta^{-1}\gamma}\mathrm{dW}$$

With Periodic Boundary Conditions and smooth potential, ergodic sampling of the canonical distribution with density



courtesy F.Nier

$$\rho_{\mathrm{can}} \propto e^{-\beta(p^T M^{-1} p/2 + U(q))}$$

Hamiltonian $H(x,p)$

Splitting Methods for Langevin Dynamics

$$\mathscr{L} = \mathscr{L}_{\mathrm{A}} + \mathscr{L}_{\mathrm{B}} + \mathscr{L}_{\mathrm{O}}$$

$$\mathscr{L}_{\mathrm{A}} = (M^{-1}p) \cdot \nabla_x$$

$$\mathscr{L}_{\mathrm{B}} = -\nabla U(x) \cdot \nabla_p$$

$$\mathscr{L}_{\mathrm{O}} = -\gamma (M^{-1}p) \cdot \nabla_p + \gamma \beta^{-1} \Delta_p$$

$$e^{h\hat{\mathscr{L}}_{\mathrm{BAOAB}}} = e^{\frac{h}{2}\mathscr{L}_{\mathrm{B}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{A}}} e^{h\mathscr{L}_{\mathrm{O}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{A}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{B}}}$$

Expansion of the invariant distribution

$$[\mathscr{L}^\dagger + h^2 \mathscr{L}_2^\dagger + \ldots] e^{-\beta(H + h^2 f_2 + \ldots)} = 0$$

Leading order:

$$\mathscr{L}^\dagger(\rho_{\mathrm{can}} f_2) = \beta^{-1} \mathscr{L}_2^\dagger \rho_{\mathrm{can}}$$

L. & Matthews, AMRX, 2013
L., Matthews, & Stoltz, IMA J. Num. Anal. 2015

- detailed treatment of all 1st and 2nd order splittings
- estimates for the operator inverse and justification of the expansion
- treatment of nonequilibrium (e.g. transport coefficients)

Configurational Sampling

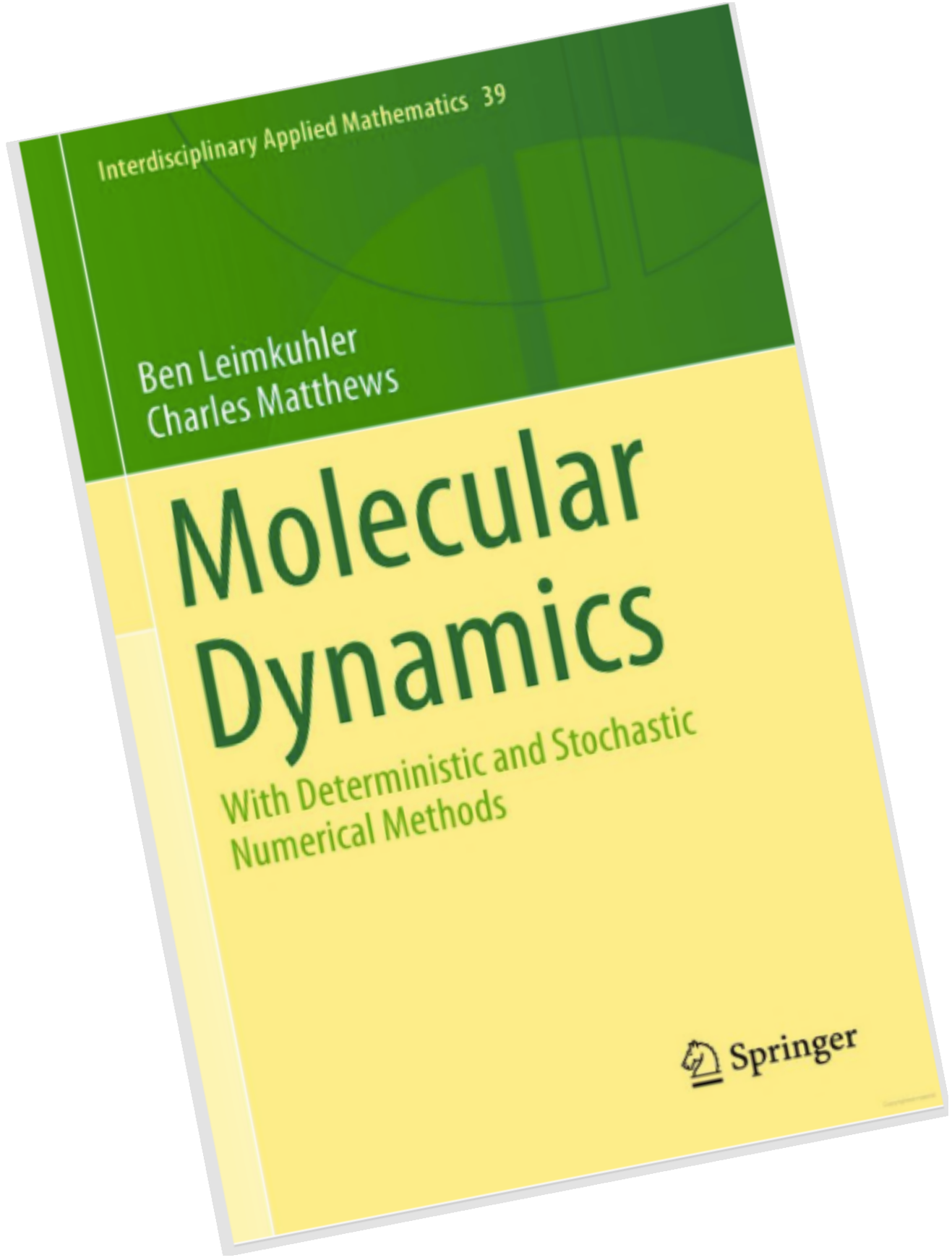
The Magic Cancellation: *[L. & Matthews 2013]*
The marginal (configurational) distribution of the **BAOAB** method has an expansion of the form

$$\bar{\rho}_h = e^{-\beta U}[1 + O(h^2/\gamma^2) + O(h^4)]$$

In the high friction limit: 4th order, and with just one force evaluation per timestep.

Weak accuracy order = 2 but for high friction,
4th order in the invariant measure.

Hardbound or
via SpringerLink



but….

What to do about the force error?

$$\tilde{F}(x) = -\nabla U(x) + \eta(x)$$

a sampling error… it seems natural to take

$$\eta(x) \sim \mathcal{N}(0, \sigma(x))$$

and also, at least in the first stage, to assume $\sigma(x) \approx \sigma$

$$\begin{aligned} x_{n+1} &= x_n + hF(x_n) + h\sigma\tilde{R}_n + \sqrt{2h}R_n \\ &= x_n + hF(x_n) + \sqrt{h}\sqrt{h\sigma^2 + 2}\hat{R}_n \end{aligned}$$

Like Euler-Maruyama discretization of

$$\mathrm{d}x = F(x)\mathrm{d}t + \sqrt{2 + \sigma h}\mathrm{dW}$$

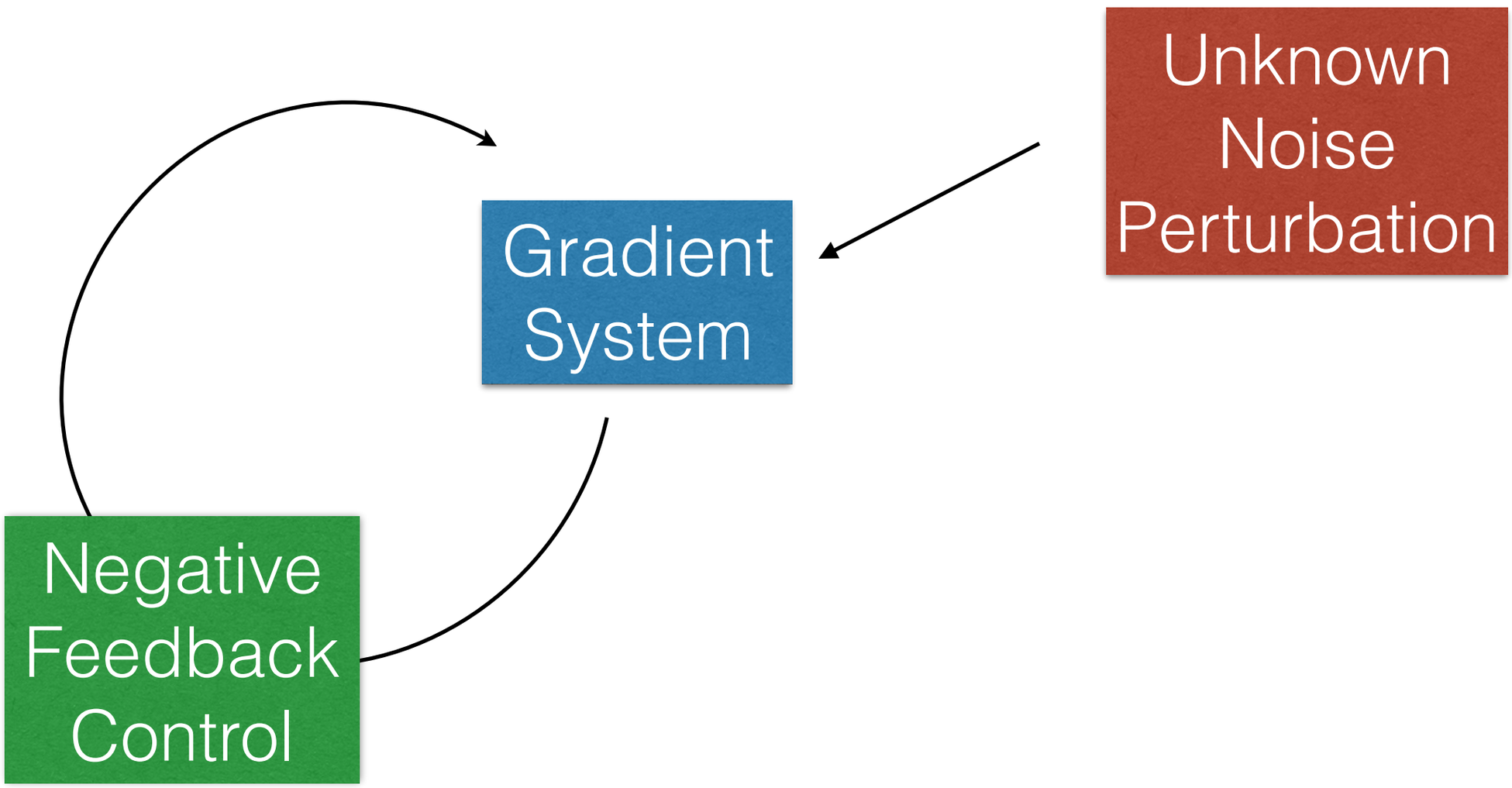
$$\mathrm{d}x = F(x)\mathrm{d}t + \sqrt{2 + \sigma h}\mathrm{dW}$$

1. Stepsize-dependent dynamics (like in B.E.A.)
2. Distorts temperature
3. Possible to correct - if we know σ
4. Computing/estimating σ can be difficult in practice

Options:

Monte-Carlo based approach [Ceperley et al, 'Quantum Monte Carlo' 1999]
Stochastic Gradient Langevin Dynamics [Welling, Teh, 2011]
Adaptive Thermostats [Jones and L., 2011]

control of thermodynamic observables



Adaptive Thermostats

Jones & L., J. Chem. Phys. 2011

Applying Nosé-Hoover Dynamics to a system which is driven by white noise restores the canonical distribution.

Adaptive (Automatic) Langevin

$$\mathrm{d}x = M^{-1}p\mathrm{d}t$$

$$\mathrm{d}p = -\nabla U\mathrm{d}t - \sqrt{h}\sigma\mathrm{d}W - \xi p\mathrm{d}t + \sigma_A\mathrm{d}W_A$$

$$\mathrm{d}\xi = \mu^{-1}[p^T M^{-1}p - n\beta^{-1}]\mathrm{d}t$$

$$\tilde{\rho} = e^{-\beta[p^T M^{-1}p/2+U(x)]} \times e^{-\beta\mu(\xi-\gamma)^2/2}$$ ergodic!

Shift in auxiliary variable by $\gamma = \dfrac{\beta(h\sigma^2 + \sigma_A^2)}{2\mathrm{Tr}(M)}$

<u>**Discretization**</u> [With ***X. Shang, 2015***]

generator: $\mathscr{L} = \mathscr{L}_{\mathrm{A}} + \mathscr{L}_{\mathrm{B}} + \mathscr{L}_{\mathrm{O}} + \mathscr{L}_{\mathrm{D}}$

$$\mathscr{L}_{\mathrm{A}} = (M^{-1}p) \cdot \nabla_x$$

$$\mathscr{L}_{\mathrm{B}} = -\nabla U(x) \cdot \nabla_p + \frac{h\sigma^2}{2}\Delta_p$$

$$\mathscr{L}_{\mathrm{O}} = -\xi p \cdot \nabla_p + \frac{\sigma_{\mathrm{A}}^2}{2}\Delta_p$$

$$\mathscr{L}_{\mathrm{D}} = G(p)\frac{\partial}{\partial \xi}$$

define related operator by composition, e.g. **BADODAB**

$$e^{h\hat{\mathscr{L}}} = e^{\frac{h}{2}\mathscr{L}_{\mathrm{B}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{A}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{D}}} e^{h\mathscr{L}_{\mathrm{O}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{D}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{A}}} e^{\frac{h}{2}\mathscr{L}_{\mathrm{B}}}$$

BADODAB ≈ BAOAB

BAOAB has remarkable sampling properties:

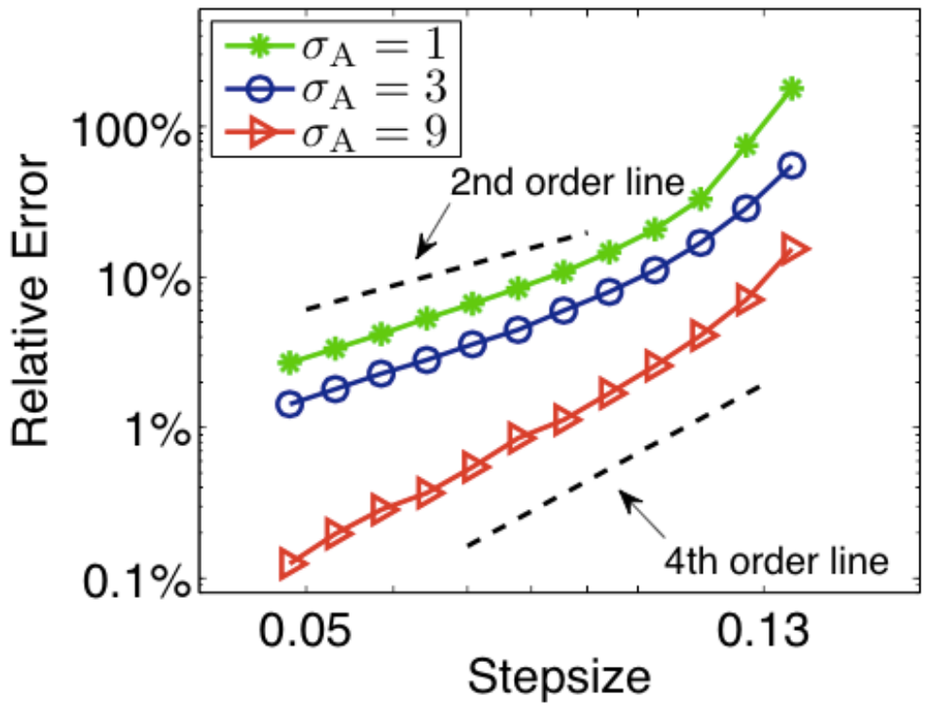
- superconvergence in the high friction limit
- exact sampling (in x) for harmonic systems

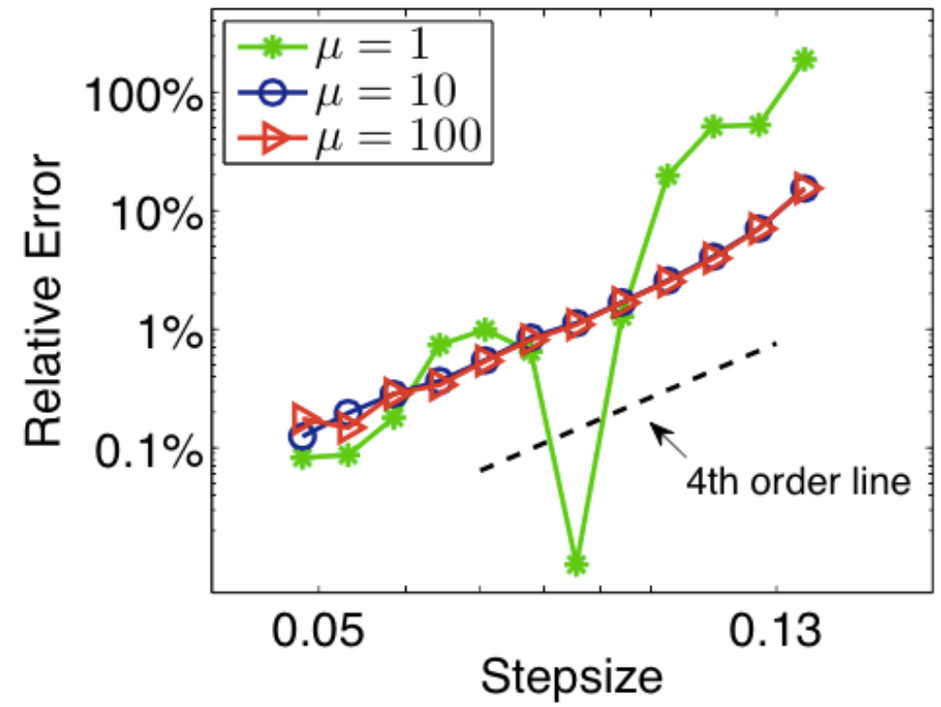
By taking large $\mu \propto \sigma_A^2$ we can make BADODAB behave like BAOAB after averaging over the auxiliary variable.

This can be viewed as a **projection method** for the Fokker-Planck stationary problem.

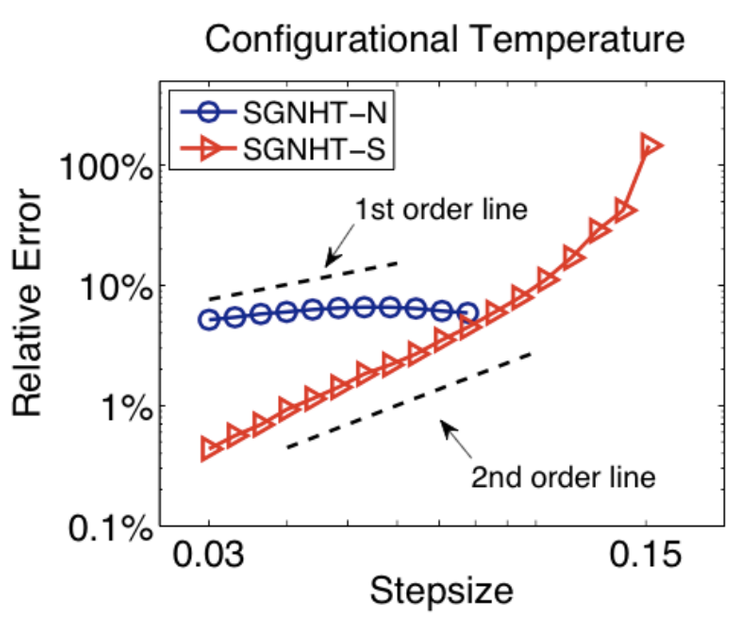
500 Lennard-Jones particles, clean gradient

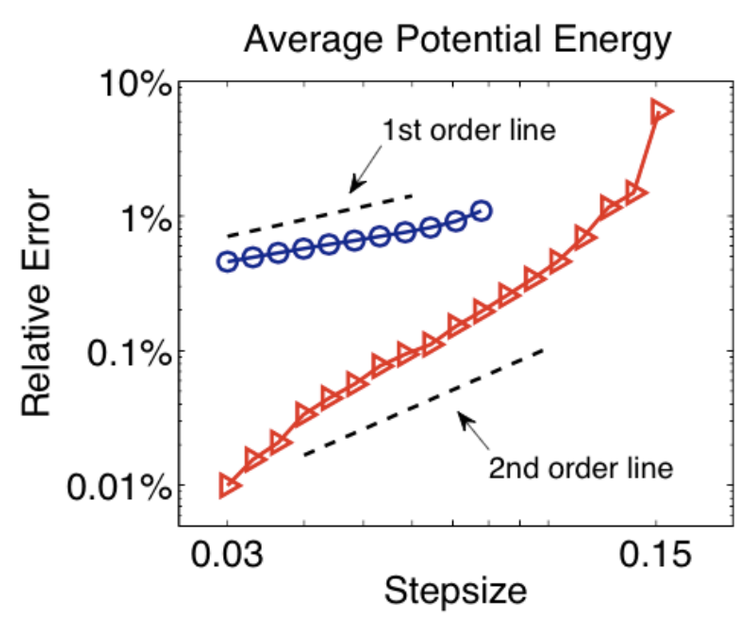
configurational temperature





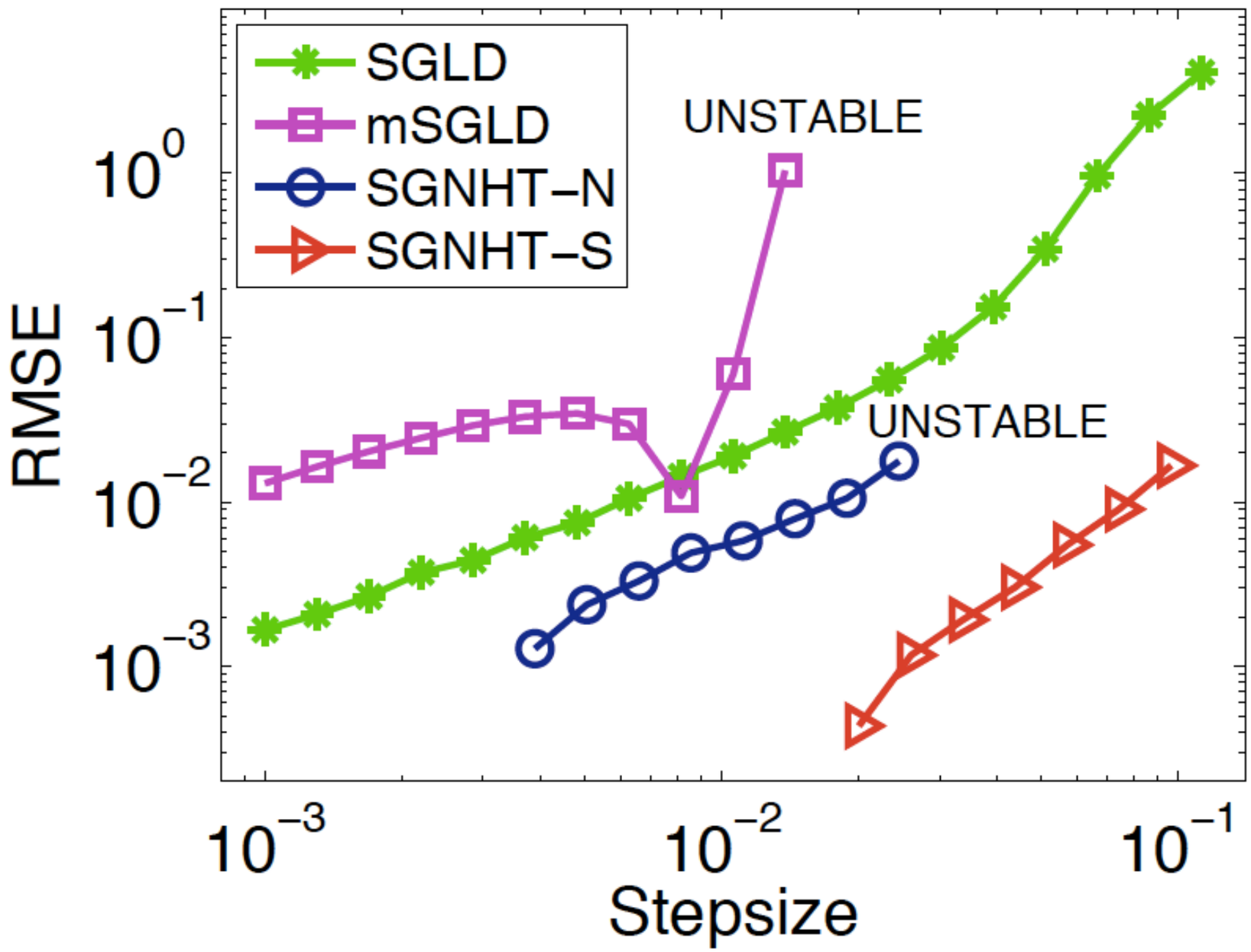
Comparison with Chen et al. (Google)





Bayesian Logistic Regression

(small model)

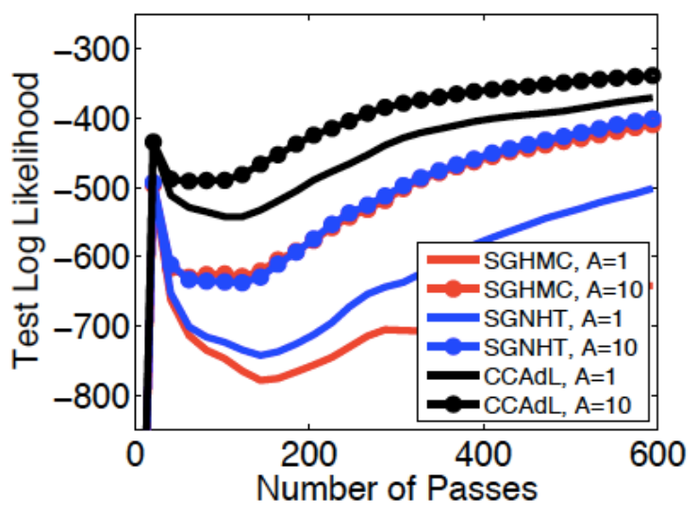


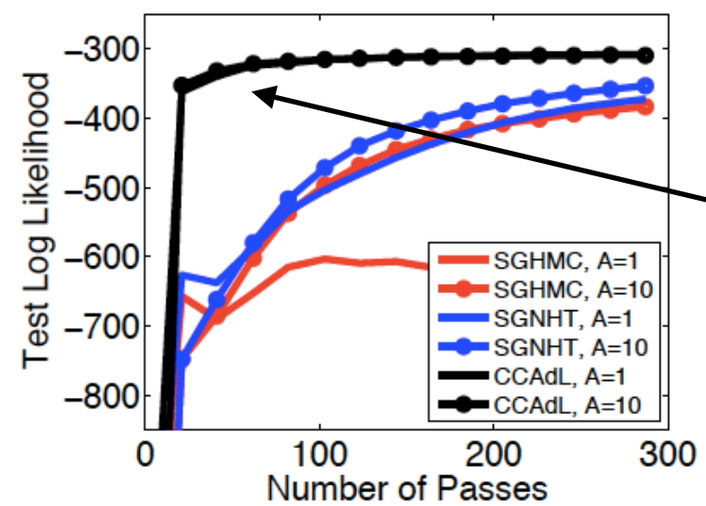
Teaser!

New variant of the SGNHT scheme

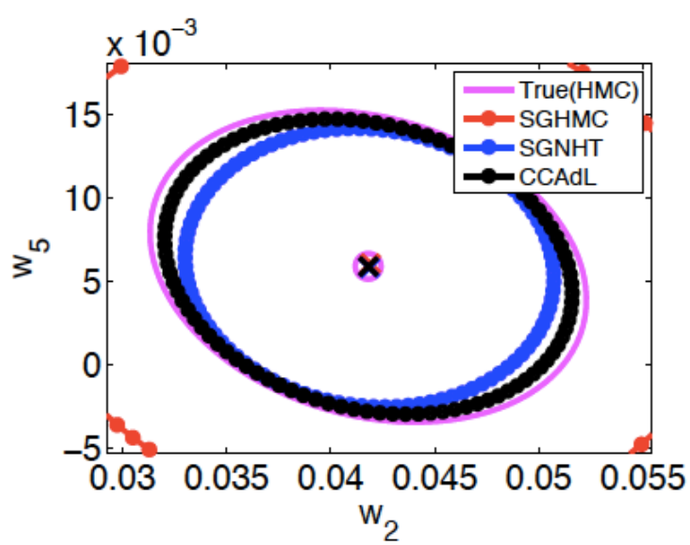
w. X. Shang, A. Storkey & Z. Zhu

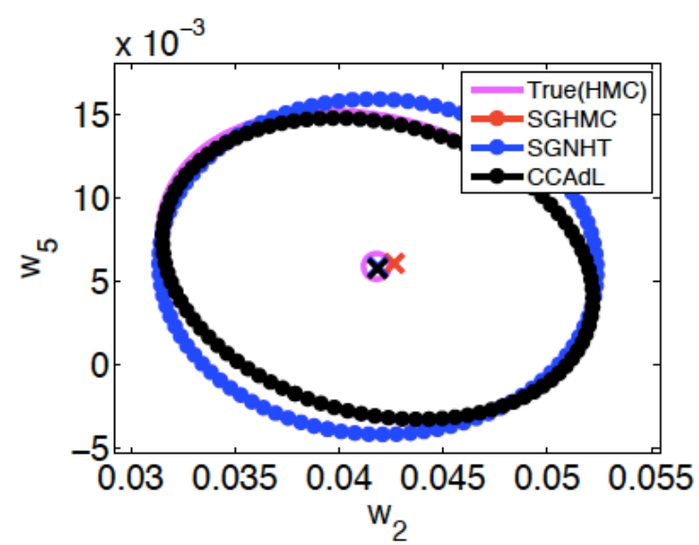
MNIST 7 or 9?





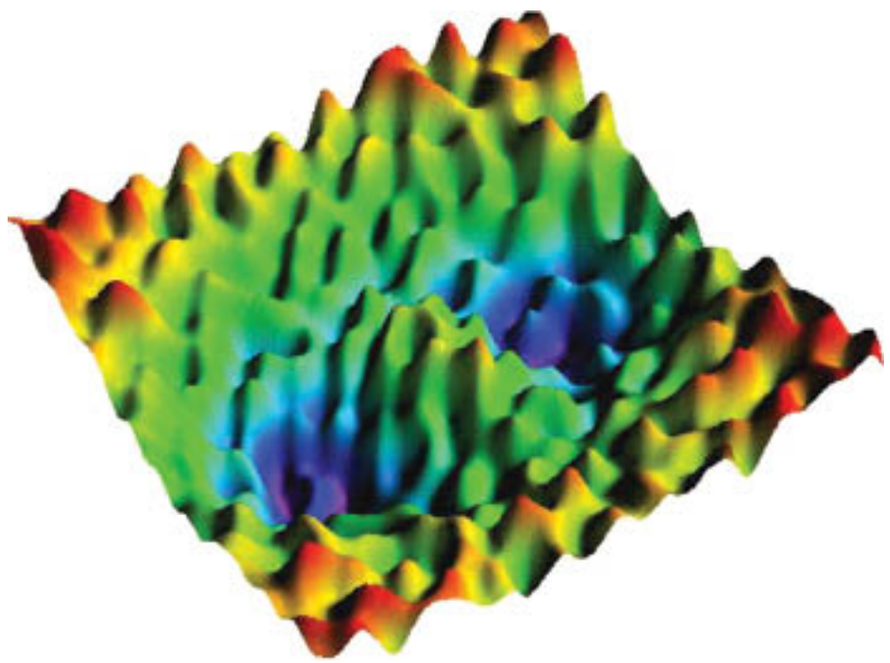
(!)

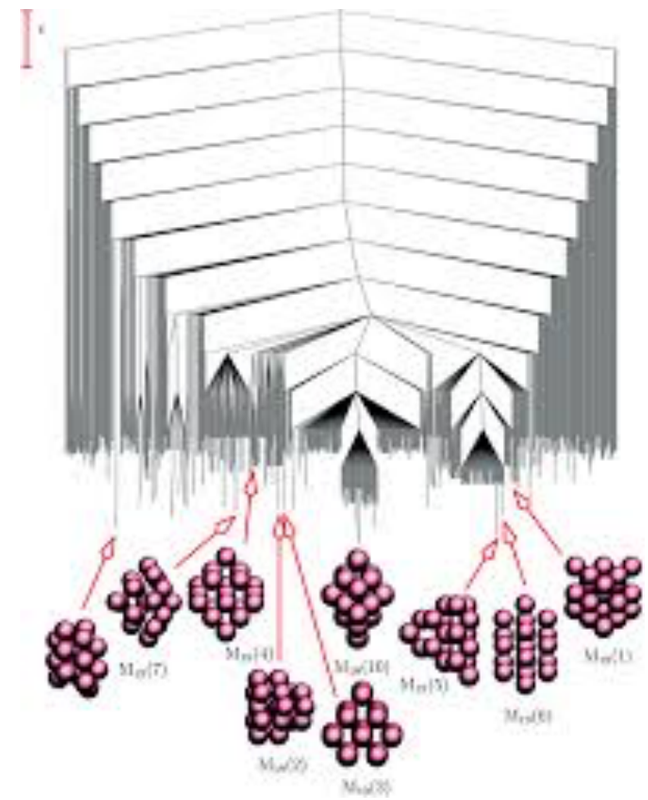




(left) $h = 0.2 \times 10^{-4}$ (right) $h = 1 \times 10^{-4}$

Multimodal Landscapes





Problem: sample all the basins accessible
at a given temperature in a realistic simulation time.

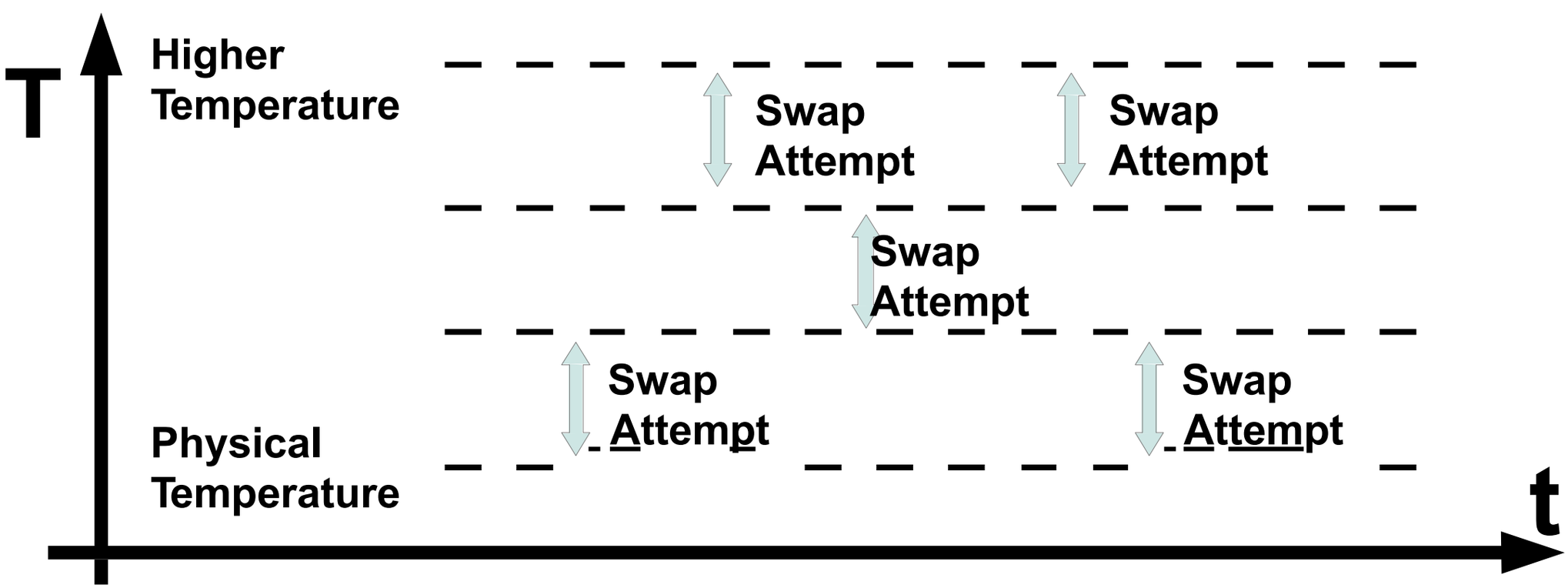
Continuous Tempering

G. Gobbo & L., Phys Rev E 2015

- **Tempering Approaches:**

  At higher temperature transitions are more likely to happen (Simulated Tempering, Replica Exchange, etc.)

Replica Exchange



Continuous Tempering

1. Add a degree of freedom that directly controls temperature

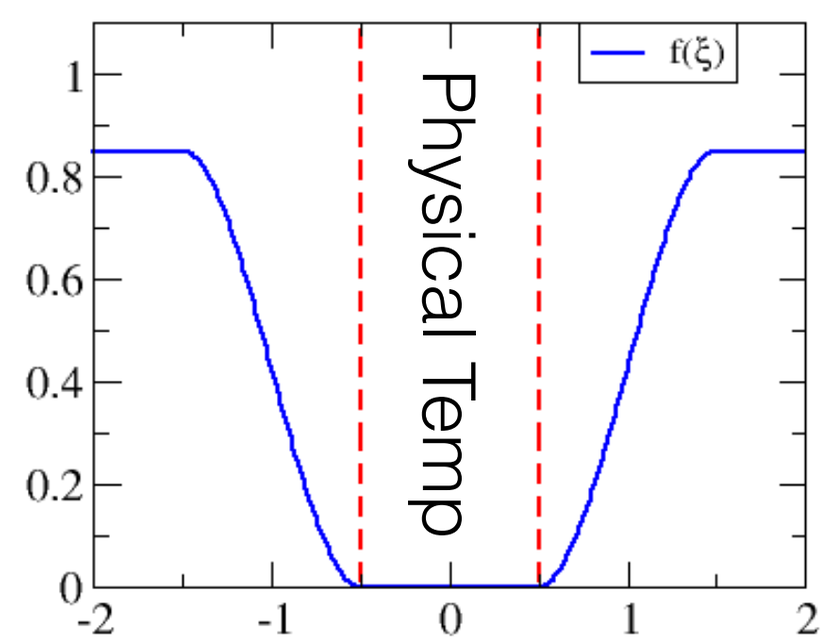
$$\hat{H}(q,p,\xi,p_\xi) = (1-f(\xi))H(q,p) + \frac{p_\xi^2}{2m_\xi} + \varphi(\xi)$$

2. The stationary distribution for the extended system is

$$\hat{\rho}(q,p,\xi,p_\xi) \propto e^{-\beta\hat{H}} = e^{-\beta_{\text{eff}}H} \times \hat{\rho}(\xi,p_\xi)$$

$$\beta_{\text{eff}} = \beta(1-f(\xi))$$

3. Draw samples only for physical values of the temperature



Application: MIST Implementation

We have implemented our method using **MIST**

http://www.extasy-project.org/mist

ExTASY

Extensible Tools for Advanced Sampling and Analysis

NSF-EPSRC Project (~$4M)

Edinburgh EPCC Mathematics

Duke Mathematics

Rice Chemistry

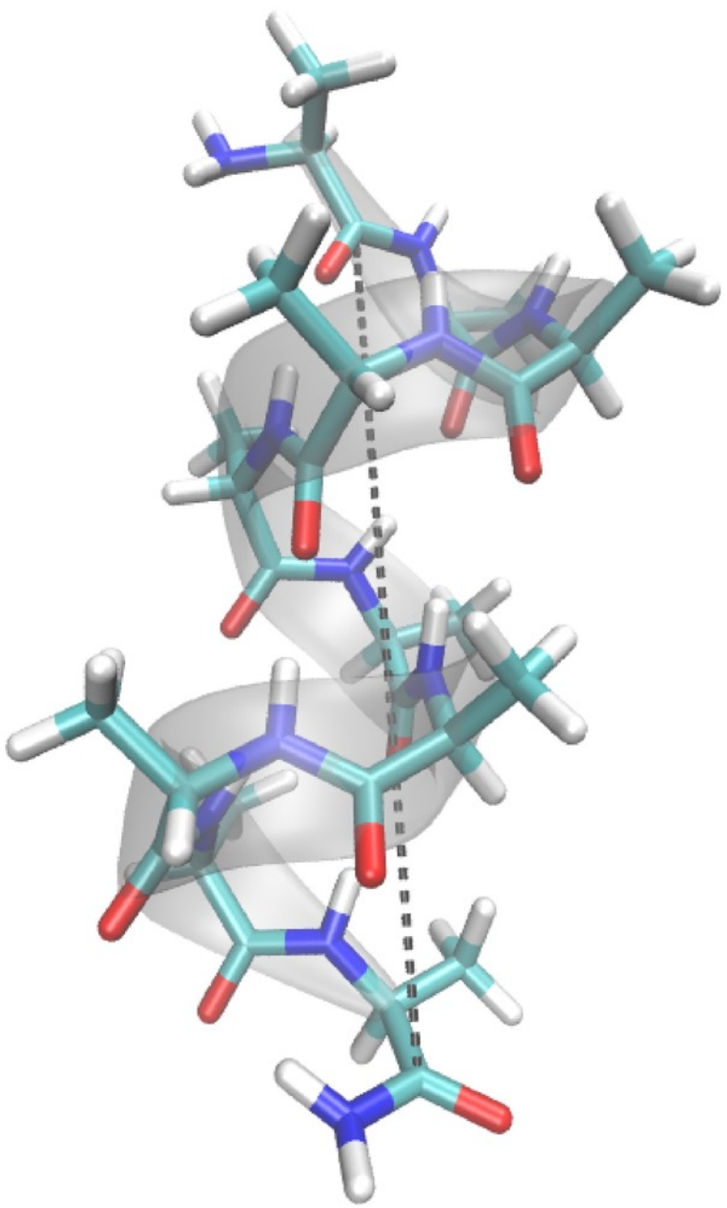
Nottingham Pharma-Chem

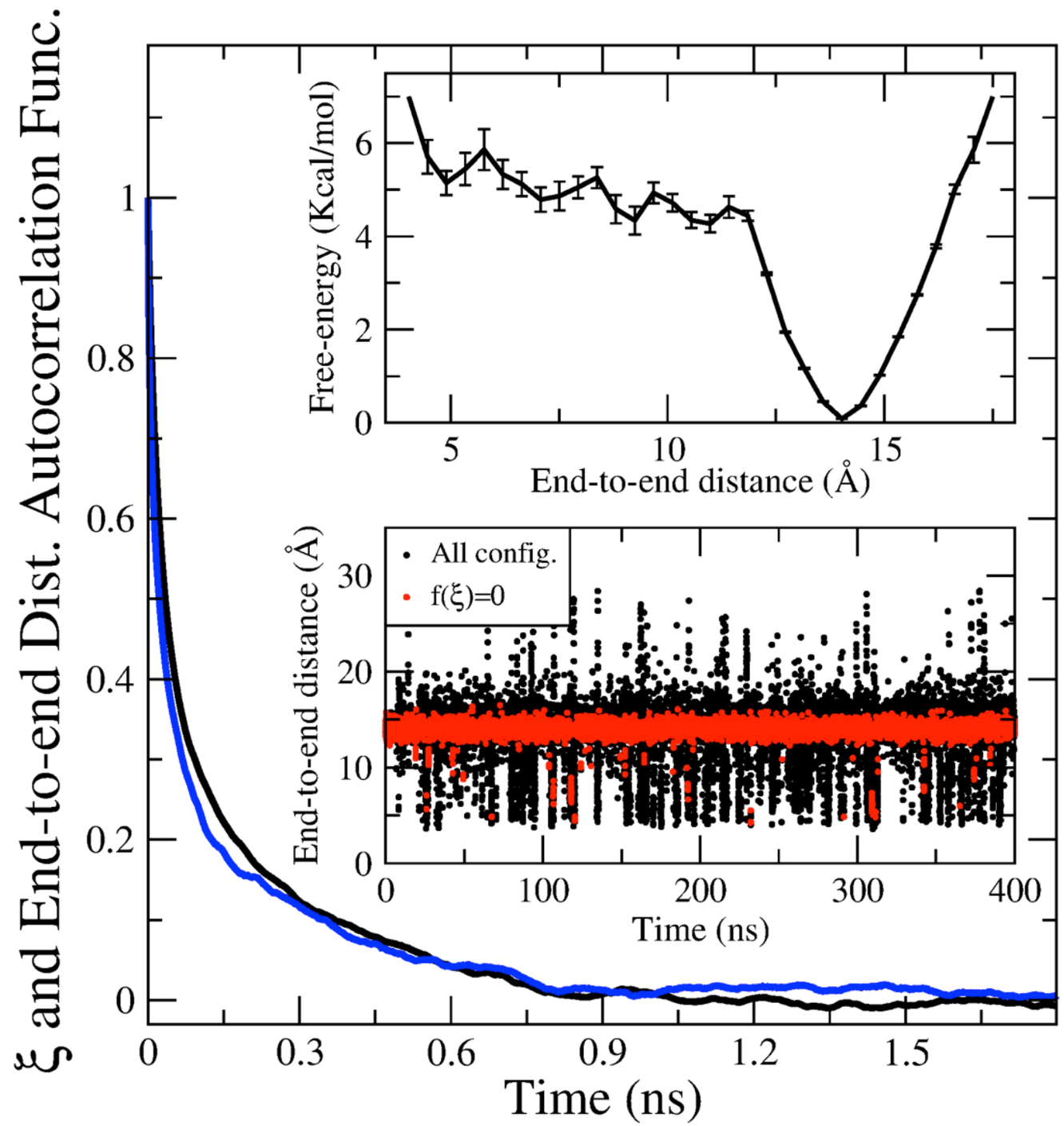
Imperial College Computer Sci

Rutgers Computer Sci

Gromacs Version Now Available

Application: Ala_{10}





Free-energy profile compatible with Comer et. al, J. Chem.Theory Comp. (2014)

Summary

High Accuracy Discrete Dynamics: *the perfect sampling bias in discretized SDEs can be reduced dramatically using the right choice of numerical method.*

Noisy Gradients: *Carefully designed feedback controls allow correct sampling despite error in gradients*

Continuous Tempering: *A simple and thermodynamically consistent approach to global sampling of corrugated landscapes.*

Questions: Structure of Bayesian Landscapes?
Analogues of multiscale models/free energies?
Role of implicit methods? Variable stepsizes?
Use of geometric information? …