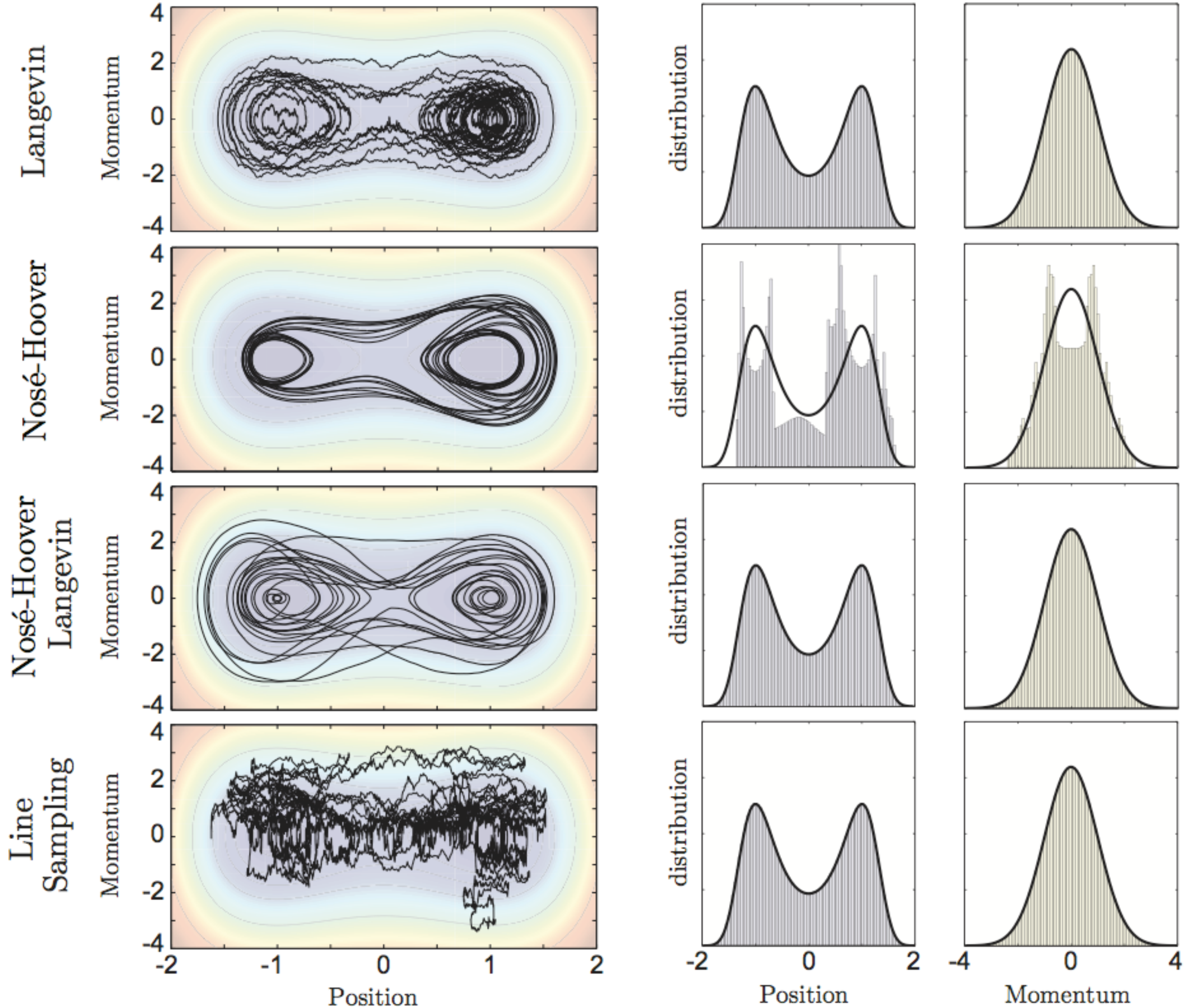
Molecular Dynamics
Day 5

Ben Leimkuhler

general thermostats and SDEs
Nosé-Hoover methods
adaptive thermostats
noisy gradients
ensemble preconditioning

Peking 2018

From *L., Generalized Bulgak-Kusnezov Thermostats, PRE, 2010*



Finding the "Right" Dynamics for the Job

There are many different stochastic models that can be used in MD, but they can have very different efficiencies for a particular task.

Overdamped Langevin Dynamics

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

great for sampling well scaled multivariate Gaussian distribution,
awful for a highly corrugated landscape

Nosé-Hoover

gentle - good for autocorrelation functions in systems with strong internal mixing properties…
not ergodic - lousy for nucleic acid simulations in implicit solvent

Thermostats

Gibbs distribution

Overdamped Langevin

Langevin Dynamics

Nosé-Hoover Langevin

Nosé-Hoover Dynamics

Generalized
Bulgac-Kusnezov

Preconditioned Methods
Ensemble Quasi-Newton

A generic sampling dynamics

$$\mathrm{d}x = [J(x) + S(x)]\nabla \log \pi(x) + \nabla \cdot [J(x) + S(x)] + \sqrt{2S(x)}\mathrm{dW}$$

***J(x)* antisymmetric**

***S(x)* symmetric**

Includes, e.g.,

SDEs like **Brownian** and **Langevin** dynamics

non-reversible perturbation methods

various **ensemble sampling** schemes

Questions:

Which approach **converges most rapidly**? (small IAT)

What is the **sampling bias under discretization**?

How to effectively **combine with extension**?

Generalized Sampling

Up to know we have assumed the situation of a known distribution with invariant density

$$\rho \propto e^{-\beta U(q)}$$

What if we don't know U or cannot exactly resolve the force?

Multiscale models, e.g. ab initio MD Methods and **QM/MM methods** (heating due to force mismatch)

Nonequilibrium MD (e.g **Shear Flows**)

Applications in **Bayesian Inference & Big Data Analytics**

Problems for today

1. How to **gently perturb Hamiltonian dynamics** in order to achieve thermal equilibriation.

2. How to handle **noisy gradient systems** and driven systems efficiently, in particular with momentum constraints for shear flow applications.

3. How to **accelerate convergence** to equilibrium by use of an **ensemble of "particles"** (walkers).

Additivity

The thermostats can be **combined** in most cases without altering their effectiveness (often improving it).

$$\dot{x} = f(x) + g(x)$$

$$\mathscr{L}_{f+g} = \mathscr{L}_f + \mathscr{L}_g$$

$$\left.\begin{array}{l} \mathscr{L}_f^\dagger \rho = 0 \\ \mathscr{L}_g^\dagger \rho = 0 \end{array}\right. \Rightarrow \mathscr{L}_{f+g}^\dagger \rho = 0$$

Works for **SDEs** too…

Extension

Many schemes make good use of the concept of ***extension***

$$\int \pi(x)\tilde{\pi}(y)\mathrm{d}y \propto \pi(x)$$

This looks banal but the key point is that although x and y decouple in the invariant distribution, they may be tightly coupled in the associated SDEs.

Example: **Langevin dynamics**

$$\int e^{-\beta p^2/2} e^{-\beta U(x)} \mathrm{d}p \propto e^{-\beta U(x)}$$

Remote control of thermal equilibration

Ex: Langevin dynamics

$$\text{Conservative dynamical system} \rightleftarrows \text{Ornstein-Uhlenbeck SDE}$$

Preserves

$$\rho_\beta = \boxed{e^{-\beta p^T M^{-1} p}} e^{-\beta U}$$

Ergodic for

$$\bar{\rho} = e^{-\beta p^T M^{-1} p}$$

- The two systems are both ***compatible*** with ρ_β
- Sufficient mixing

The ergodicity of the OU process implies ergodicity of the full system

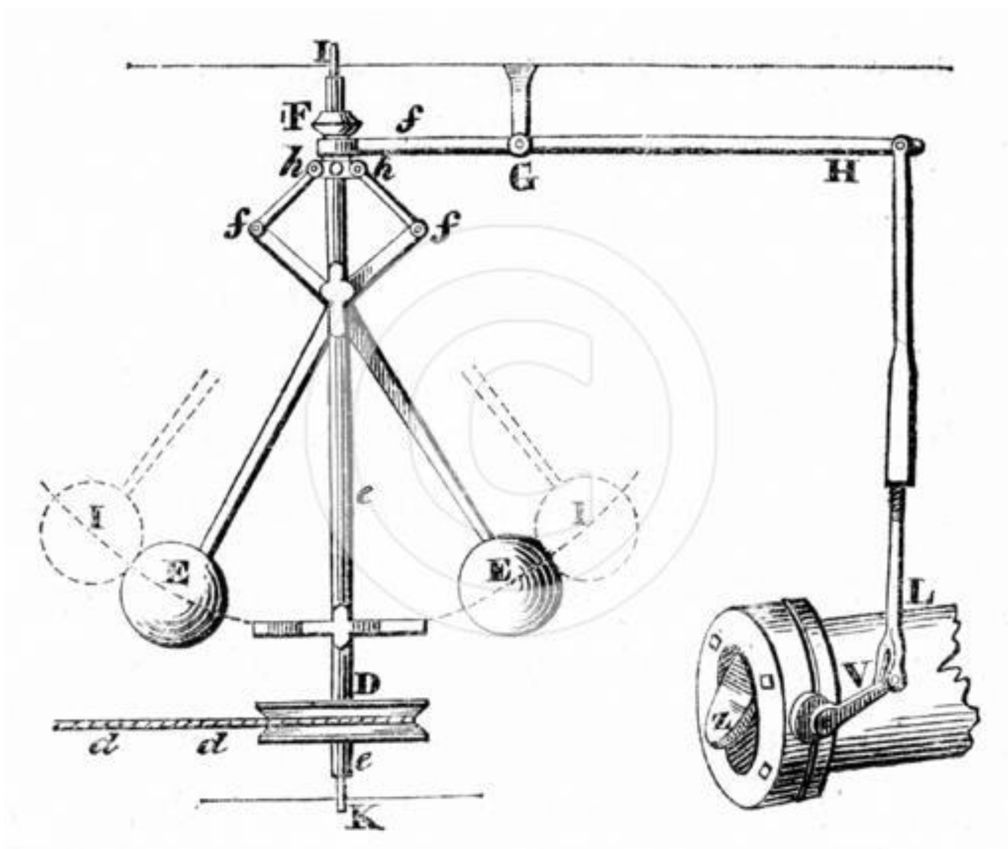
Nosé-Hoover
and gentle equilibration

Where I learned about Nosé Dynamics

Seminar, Cambridge University 1997:
Nosé Dynamics

Sir Henry Peter Francis Swinnerton-Dyer, 16th Baronet KBE FRS
Number Theorist, Student of Littlewood, Polya and Sylvester Prizeholder
Vice-Chancellor of Cambridge University 1973-83

James Watt’s Engine



Too fast: balls move to outside, opening valve, releasing steam, reducing pressure, reducing speed

Too slow: balls fall to inside, closing valve, leading to an increase in pressure, increasing speed

Nose-Hoover dynamics - a “Gibbs Governor”

$$
\begin{aligned}
\dot{q} &= p \\
\dot{p} &= -\nabla U(q) - \xi p \\
\dot{\xi} &= p^2 - kT
\end{aligned}
$$

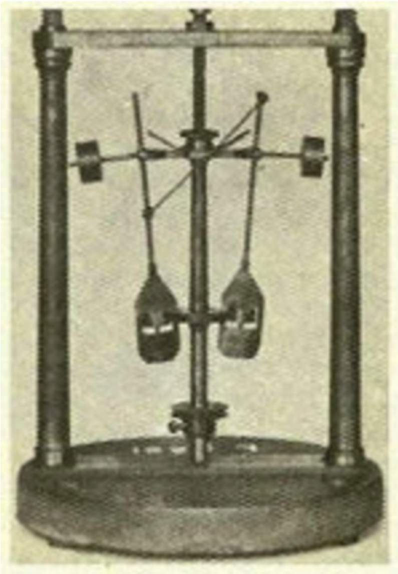
Preserves

$$e^{-\beta[p^2/2+U(q)]} \times e^{-\beta\xi^2/2}$$

Problems with the Gibbs Governor

It doesn't actually work.

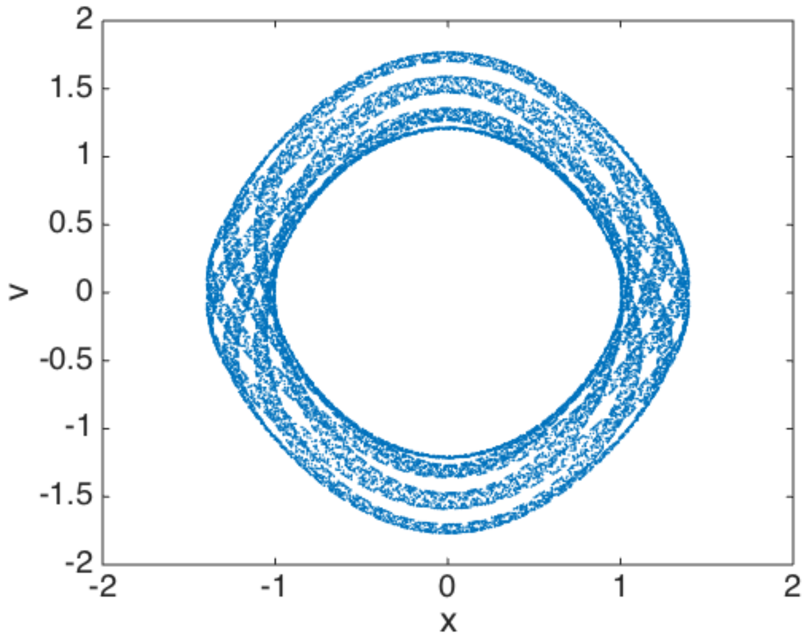
It's not *the Gibbs Governor.* This is:



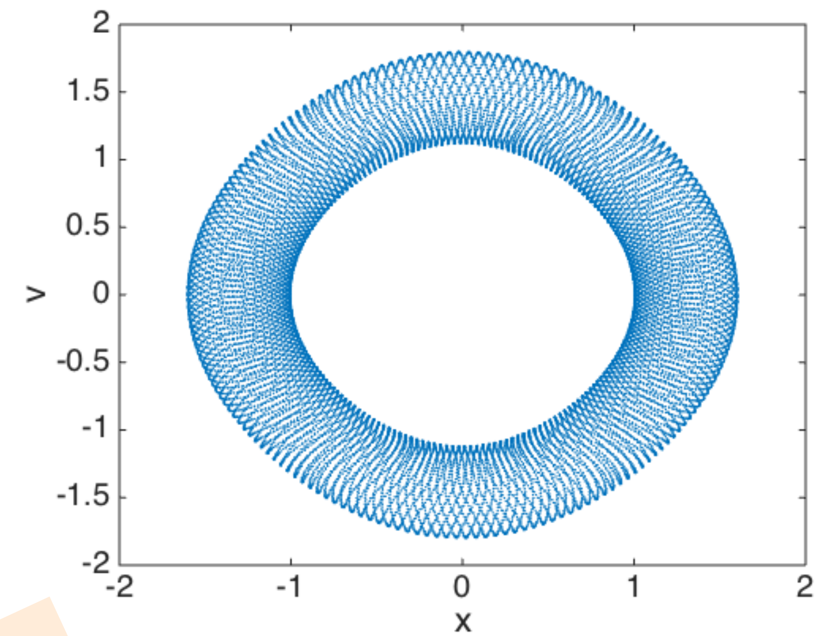
Undergraduate research project of Josiah Willard Gibbs

Nosé-Hoover Dynamics for Harmonic Oscillator

$\mu = 1$

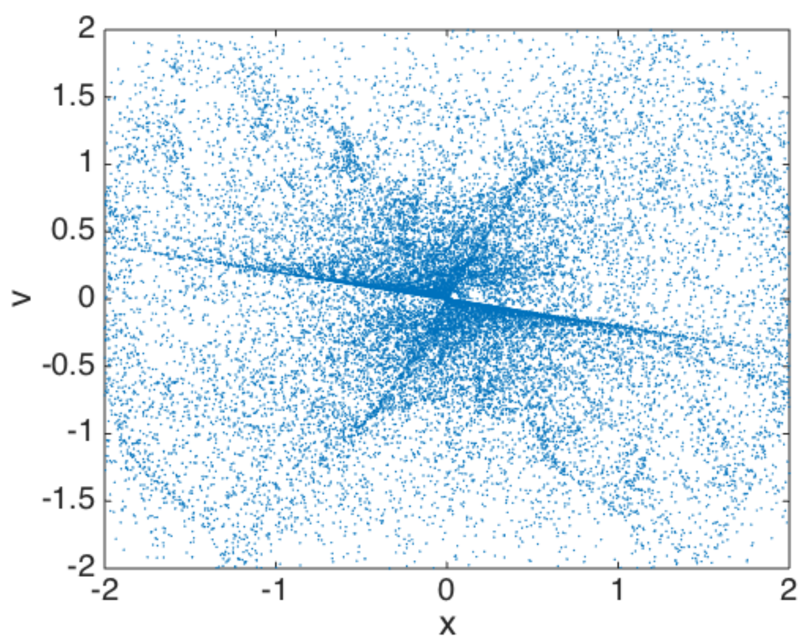


$\mu = 2$

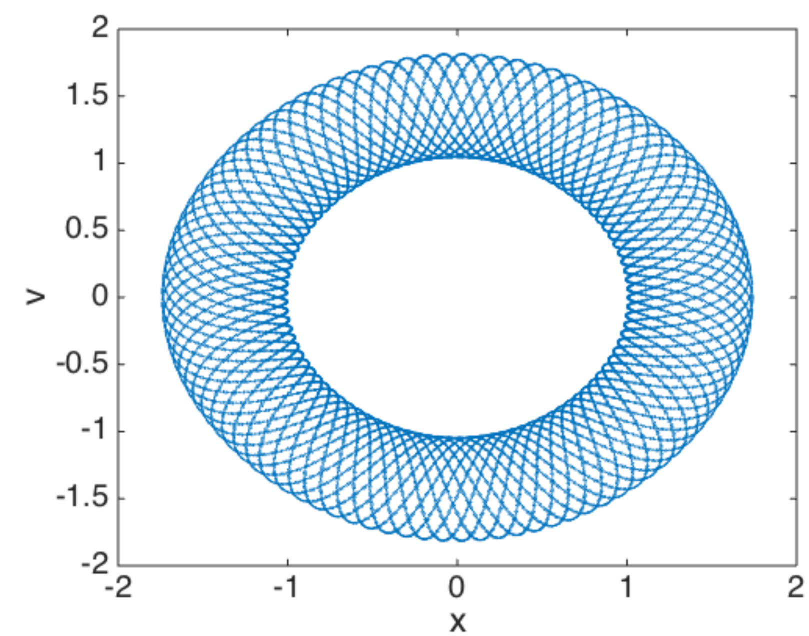


All Wrong!

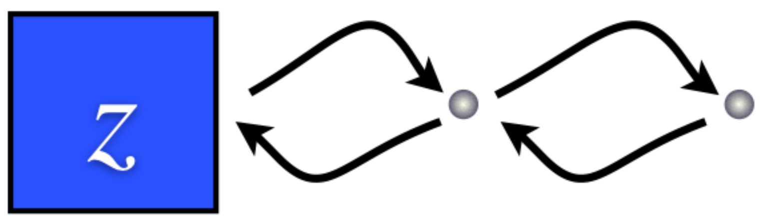
$\mu = 1/2$



$\mu = 4$

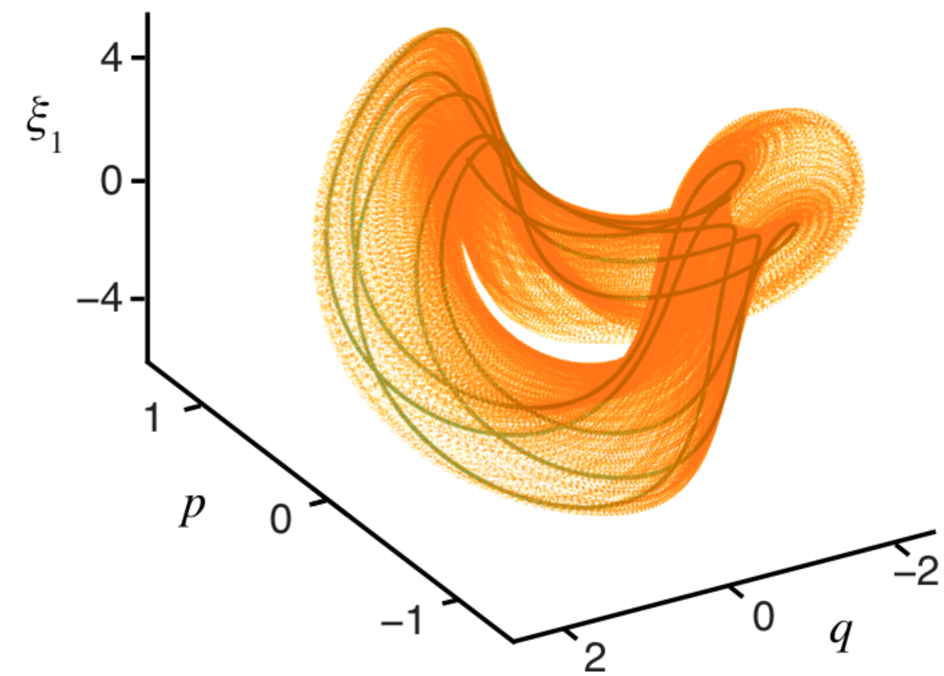


Nosé-Hoover Chains also are not ergodic..

Chains: 

$$\dot{q} = p$$
$$\dot{p} = -q - \xi p$$
$$\dot{\xi}_1 = \mu_1^{-1}(p^2 - kT) - \xi_1 \xi_2$$
$$\dot{\xi}_2 = \mu_2^{-1}(\mu_1 \xi_1^2 - kT)$$

$$\mu_1 = 0.2, \qquad \mu_2 = 1$$



Stochastic version: ***Nosé-Hoover-Langevin*** *dynamics*

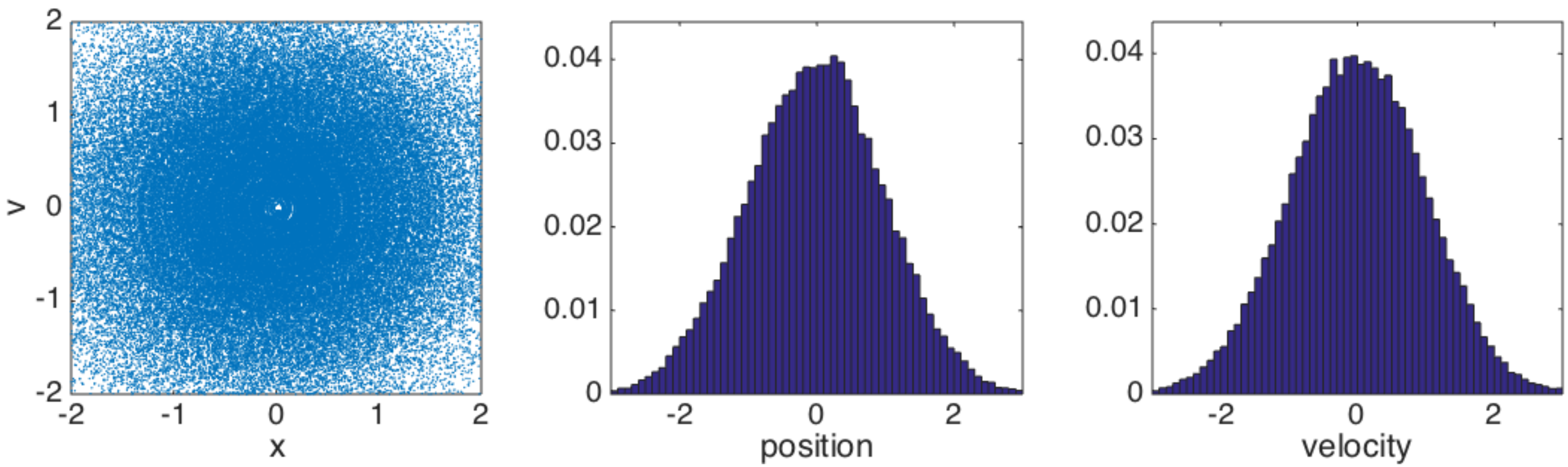
$$\dot{q} = p$$

$$\dot{p} = -\nabla U(q) - \xi p$$

$$\dot{\xi} = p^2 - kT \boxed{- \gamma\xi + \sigma\eta(t)}$$

scalar OU process

'Histograms'



matches theoretical behavior

Nosé-Hoover-Langevin

$$\mathrm{d}q = M^{-1}p\mathrm{d}t$$
$$\mathrm{d}p = -\nabla U(q)\mathrm{d}t - \xi p\mathrm{d}t$$
$$\mathrm{d}\xi = [p^T M^{-1}p - nk_\mathrm{B}T]\mathrm{d}t - \gamma\xi\mathrm{d}t + \sqrt{2\beta^{-1}\gamma}\mathrm{dW}_t$$

$\dot{q} = M^{-1}p$ $\dot{p} = -\nabla U$	**preserves**	$e^{-\beta p^T M^{-1}p/2}e^{-\beta U(q)}$
$\dot{p} = -\xi p$ $\dot{\xi} = p^T M^{-1}p - nk_\mathrm{B}T$	**preserves**	$e^{-\beta p^T M^{-1}p/2}e^{-\beta\xi^2/2}$
$\mathrm{d}\xi = -\gamma\xi\mathrm{d}t + \sqrt{2\beta^{-1}\gamma}\mathrm{dW}_t$	**preserves ergodically**	$e^{-\beta\xi^2/2}$

Ergodicity of NHL

NHL is clearly compatible with an extended Gibbs distribution meaning that

$$\mathcal{L}^{\dagger}_{\mathrm{NHL}}[\rho_\beta e^{-\beta\xi^2/2}] = 0$$

We can also prove it is ergodic by using the theory developed for Langevin dynamics and explained in the previous lectures.

[L., Noorizadeh, Theil 2009]

Harmonic system w/o resonance

$$H = \frac{p^T M^{-1} p}{2} + \frac{q^T B q}{2} \qquad q, p \in \mathbb{R}^d$$

$$A = M^{-1} B, \qquad A\varphi_k = \omega_k \varphi_k$$

$$\omega_k \neq \omega_l, \qquad k \neq l$$

Theorem: **NHL is ergodic** on a large part of $\mathbb{R}^{2d+1}$

$$f_0 = \begin{bmatrix} M^{-1}p \\ -Bq - \xi p \\ p^T M^{-1} p - d\beta^{-1} - \xi \end{bmatrix}, \qquad f_1 = \mathbf{e}_{2d+1}$$

Example: clamped harmonic spring chain.

Marginalization

If a system is ergodic for

$$\tilde{\rho}(q,p,\xi) = e^{-\beta(p^T M^{-1} p/2 + U(q))} e^{-\beta \xi^2/2}$$

then the paths of the system provide Gibbs-weighted averages

$$\lim_{\tau\to\infty} \int_0^\tau \varphi(q(t),p(t),\xi(t))\mathrm{d}t = \int_\Omega \varphi(q,p,\xi)\tilde{\rho}(q,p,\xi)\mathrm{d}q\mathrm{d}p\mathrm{d}\xi$$

and

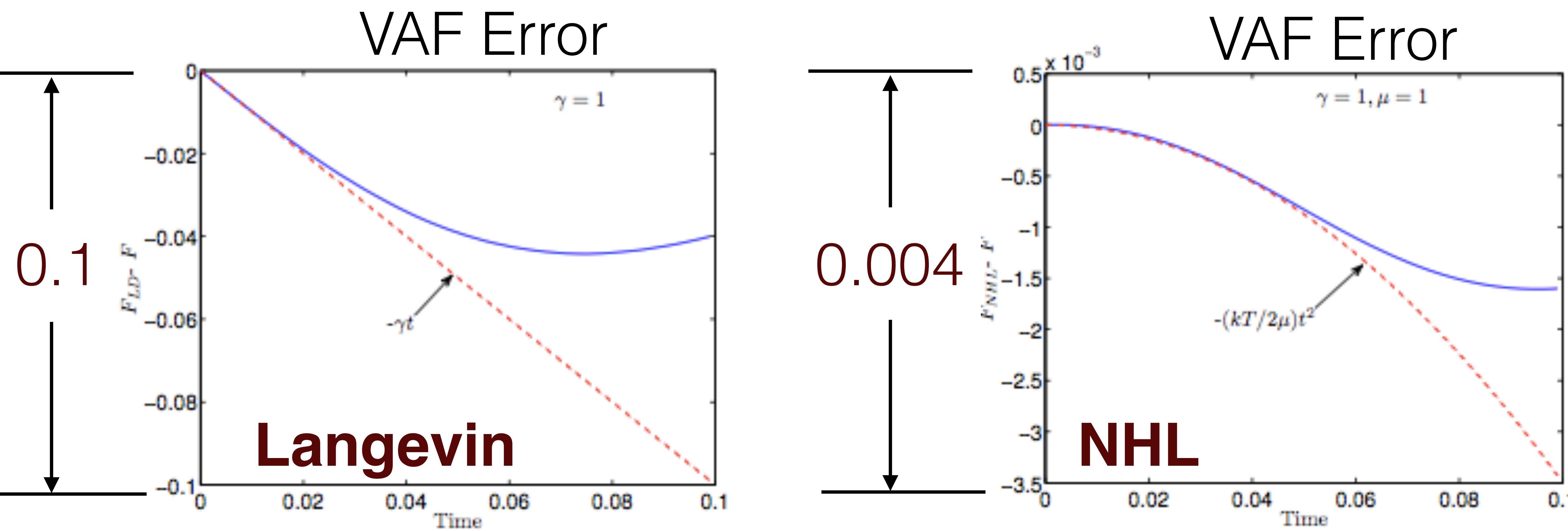
$$\lim_{\tau\to\infty} \int_0^\tau \psi(q(t),p(t))\mathrm{d}t = \int_\Omega \psi(q,p)\rho_\beta(q,p)\mathrm{d}q\mathrm{d}p$$

and

$$\lim_{\tau\to\infty} \int_0^\tau \eta(q(t))\mathrm{d}t = \int_\Omega \eta(q)e^{-\beta U(q)}\mathrm{d}q$$

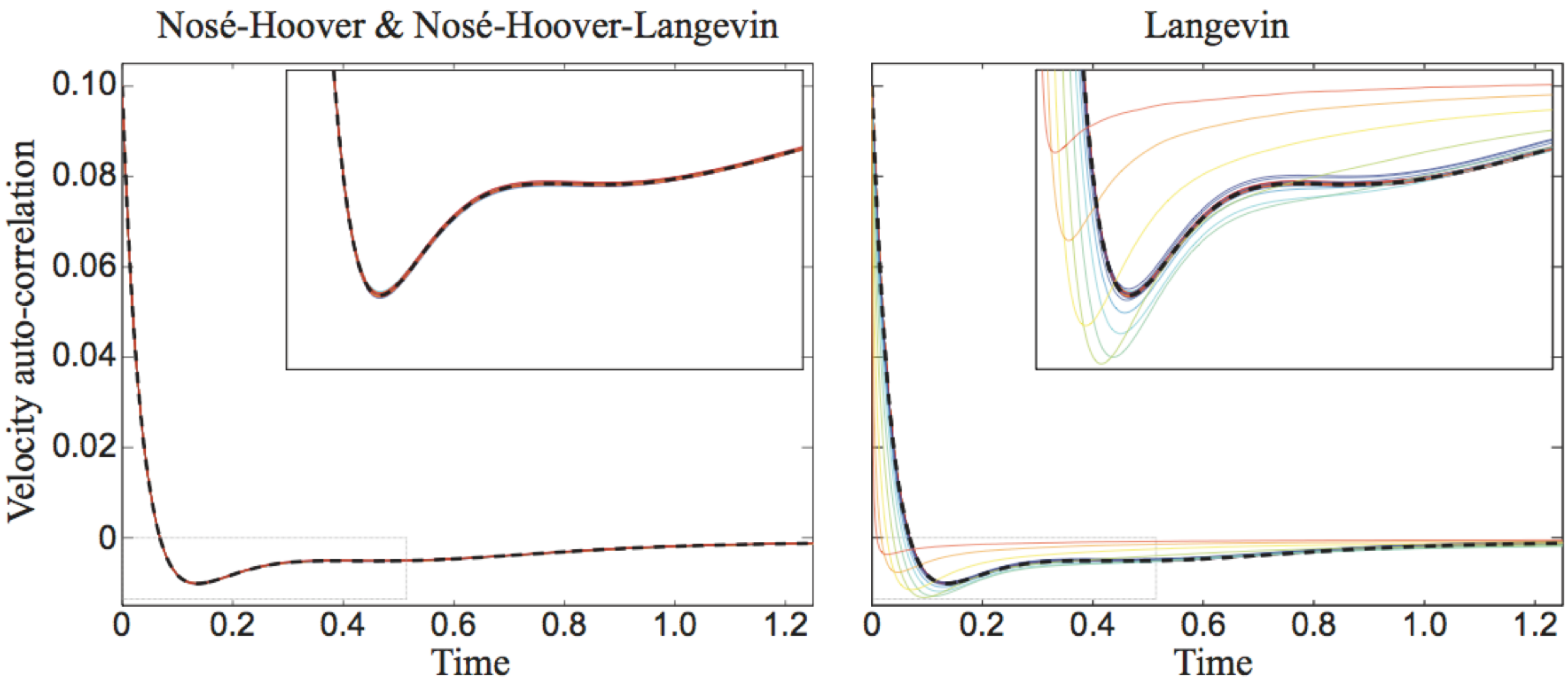
“Gentle” property of NHL

We can show that NHL is a **“gentle”** thermostat: dynamical properties are mildly perturbed for a given rate of convergence of kinetic energy.
[**L., Noorizadeh and Penrose,** J. Stat. Phys., 2011]



Similar (but less smooth) for “Stochastic Velocity Rescaling” of G. Bussi, D. Donadio and M. Parinello

Autocorrelation functions (LJ System)



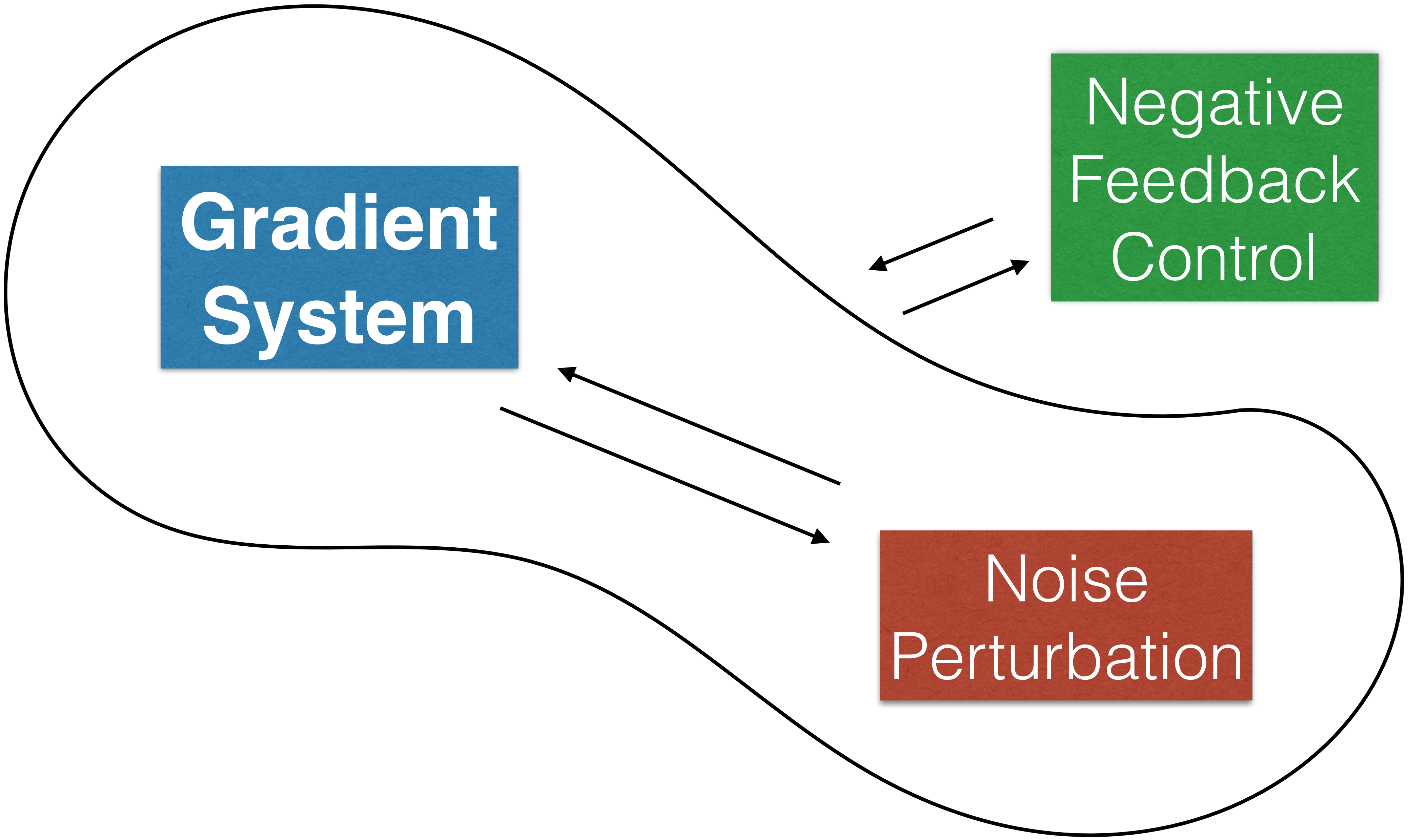
parameter dependence

Even the deterministic method seems to work for complicated systems.

Adaptive Thermostats
for Coarse Grained Models

Adaptive Thermostats

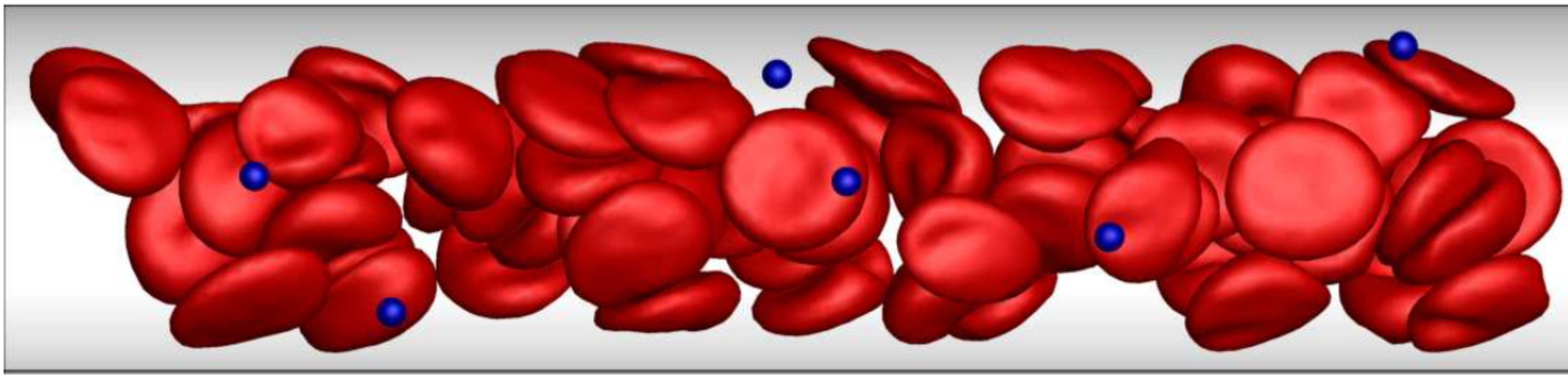
Jones & L., J. Chemical Physics, 2011



Use negative feedback loop control to stabilize the system against force perturbation (even unknown)

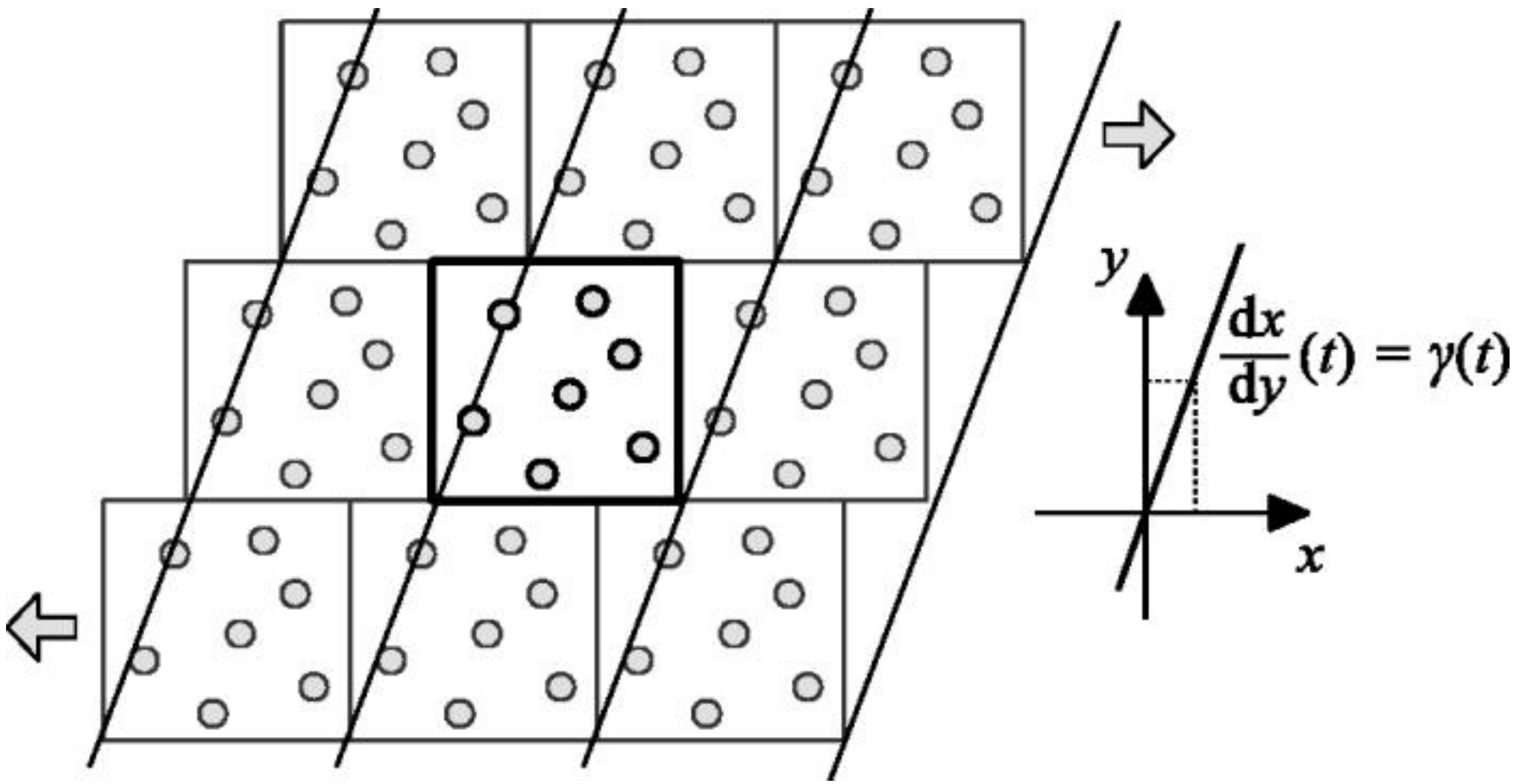
Blood Flow under Shear

Dissipative Particle Dynamics simulation under shear flow



Understanding particle margination in blood flow —- a step toward optimised drug delivery systems, K Mueller, D. Fedosov, and G. Gompper, 2015
Predicting human blood viscosity in silico, D. Fedosov et al, PNAS, 2011

Shear Flow



Implemented using Lees-Edwards Boundary Conditions

DPD Thermostat

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

$$\mathrm{d}\mathbf{p} = \mathbf{F}(\mathbf{q}, t)\mathrm{d}t - \gamma\mathbf{\Gamma}(\mathbf{q})\mathbf{p}\mathrm{d}t + \sqrt{2\gamma k_{\mathrm{B}}T}\mathbf{\Sigma}(\mathbf{q})\mathrm{d}\mathbf{W}$$

DPD friction and noise

1. **Preserves canonical distribution (equilibrium)**
2. **Preserves momentum ("hydrodynamics")**

[1] P. Hoogerbrugge and J. Koelman. Simulating microscopic hydrodynamic phenomena with dissipative particle dynamics. *Europhysics Letters*, 19(3):155, 1992.
[2] P. Espanol and P. Warren. Statistical mechanics of dissipative particle dynamics. *Europhysics Letters*, 30(4):191, 1995.

Integrators for DPD

Joint work with X. Shang

$$\mathrm{d}\begin{bmatrix}\mathbf{q}_i\\ \mathbf{p}_i\end{bmatrix} = \underbrace{\begin{bmatrix} m_i^{-1}\mathbf{p}_i \\ \mathbf{0}\end{bmatrix}\mathrm{d}t}_{\mathrm{A}} + \underbrace{\begin{bmatrix}\mathbf{0}\\ \mathbf{F}_i^C\end{bmatrix}\mathrm{d}t}_{\mathrm{B}} + \underbrace{\begin{bmatrix}\mathbf{0}\\ -\gamma \mathbf{V}_i \mathrm{d}t + \sigma \mathbf{R}_i\end{bmatrix}}_{\mathrm{O}}$$

OBAB (Shardlow)

$$\exp\left(h\hat{\mathcal{L}}_{\text{DPD-S1}}\right) = \exp\left(h\mathcal{L}_{\mathrm{O}}\right)\exp\left(\frac{h}{2}\mathcal{L}_{\mathrm{B}}\right)\exp\left(h\mathcal{L}_{\mathrm{A}}\right)\exp\left(\frac{h}{2}\mathcal{L}_{\mathrm{B}}\right)$$

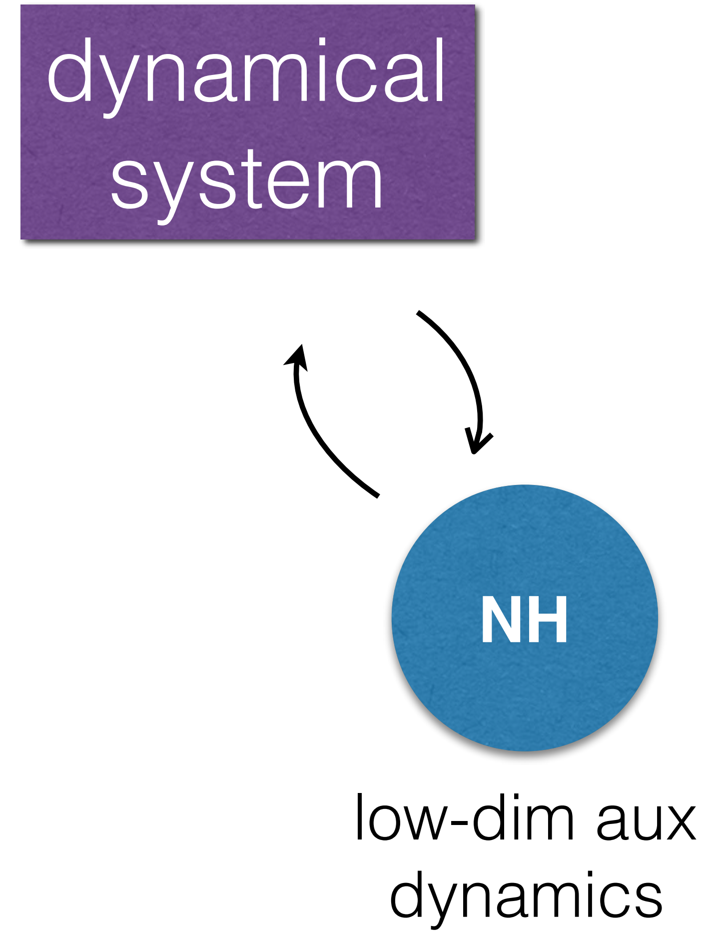
$$\exp\left(h\hat{\mathcal{L}}_{\mathrm{O}}\right) = \exp\left(h\mathcal{L}_{\mathrm{O}_{N-1,N}}\right)\cdots\exp\left(h\mathcal{L}_{\mathrm{O}_{1,3}}\right)\exp\left(h\mathcal{L}_{\mathrm{O}_{1,2}}\right)$$

Others:

DPD-Trotter = A(B+O)A (Coveney et al)

Also Lowe-Anderson, NHLA, …

Adaptive Alternatives to DPD



PNHL L. & Shang, *JCP*, 2015

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

$$\mathrm{d}\mathbf{p} = \mathbf{F}(\mathbf{q}, t)\mathrm{d}t - \xi\boldsymbol{\Gamma}(\mathbf{q})\mathbf{M}^{-1}\mathbf{p}\mathrm{d}t$$

$$\mathrm{d}\xi = G(\mathbf{q}, \mathbf{p})\mathrm{d}t - \tilde{\gamma}\xi\mathrm{d}t + \tilde{\sigma}\mathrm{d}W$$

Nose-Hoover-Like (“gentle noise”)

PAdL L. & Shang, JCP, 2016

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

$$\mathrm{d}\mathbf{p} = \mathbf{F}(\mathbf{q}, t)\mathrm{d}t - \xi\boldsymbol{\Gamma}(\mathbf{q})\mathbf{M}^{-1}\mathbf{p}\mathrm{d}t + \sigma\boldsymbol{\Sigma}(\mathbf{q})\mathrm{d}\mathbf{W}$$

$$\mathrm{d}\xi = G(\mathbf{q}, \mathbf{p})\mathrm{d}t$$

“Adaptive variant of DPD”

Splittings for PNHL

$$
\mathrm{d}\begin{bmatrix} \mathbf{q}_i \\ \mathbf{p}_i \\ \xi \end{bmatrix} = \underbrace{\begin{bmatrix} m_i^{-1}\mathbf{p}_i \\ \mathbf{0} \\ 0 \end{bmatrix}}_{\mathrm{A}} \mathrm{d}t + \underbrace{\begin{bmatrix} \mathbf{0} \\ \mathbf{F}_i^C \\ 0 \end{bmatrix}}_{\mathrm{B}} \mathrm{d}t + \underbrace{\begin{bmatrix} \mathbf{0} \\ -\xi\mathbf{V}_i \\ 0 \end{bmatrix}}_{\mathrm{C}} \mathrm{d}t + \underbrace{\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ G \end{bmatrix}}_{\mathrm{D}} \mathrm{d}t + \underbrace{\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ -\gamma^*\xi\mathrm{d}t + \sigma^*\mathrm{dW} \end{bmatrix}}_{\mathrm{O}}
$$

PNHL-S = ABCDODCBA

$$
e^{h\hat{\mathcal{L}}_{\text{PNHL-S}}} = e^{\frac{h}{2}\mathcal{L}_{\mathrm{A}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{B}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{C}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{D}}} e^{h\mathcal{L}_{\mathrm{O}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{D}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{C}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{B}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{A}}}
$$

PNHL-N = ABCDODCAB *not symmetric!*

$$
e^{h\hat{\mathcal{L}}_{\text{PNHL-N}}} = e^{\frac{h}{2}\mathcal{L}_{\mathrm{A}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{B}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{C}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{D}}} e^{h\mathcal{L}_{\mathrm{O}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{D}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{C}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{A}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{B}}}
$$

1st order but ***2nd order!*** for $\phi(\mathbf{q}, \mathbf{p}) = \mathbf{p}^{2k}\vartheta(\mathbf{q})$

Proof - uses a projection technique for the FP operator

Splittings for PAdL

$$\mathrm{d}\begin{bmatrix}\mathbf{q}_i\\ \mathbf{p}_i\\ \xi\end{bmatrix} = \underbrace{\begin{bmatrix}m_i^{-1}\mathbf{p}_i\\ \mathbf{0}\\ 0\end{bmatrix}\mathrm{d}t}_{\mathrm{A}} + \underbrace{\begin{bmatrix}\mathbf{0}\\ \mathbf{F}_i^{\mathrm{C}}\\ 0\end{bmatrix}\mathrm{d}t}_{\mathrm{B}} + \underbrace{\begin{bmatrix}\mathbf{0}\\ -\xi\mathbf{V}_i\mathrm{d}t + \sigma\mathbf{R}_i\\ 0\end{bmatrix}}_{\mathrm{O}} + \underbrace{\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ G\end{bmatrix}\mathrm{d}t}_{\mathrm{D}}$$

PAdL-S = ABODOBA

$$e^{h\hat{\mathcal{L}}_{\mathrm{PAdL-S}}} = e^{\frac{h}{2}\mathcal{L}_{\mathrm{A}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{B}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{O}}} e^{h\mathcal{L}_{\mathrm{D}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{O}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{B}}} e^{\frac{h}{2}\mathcal{L}_{\mathrm{A}}}$$

This splitting is not at all BAOAB!

What to measure?

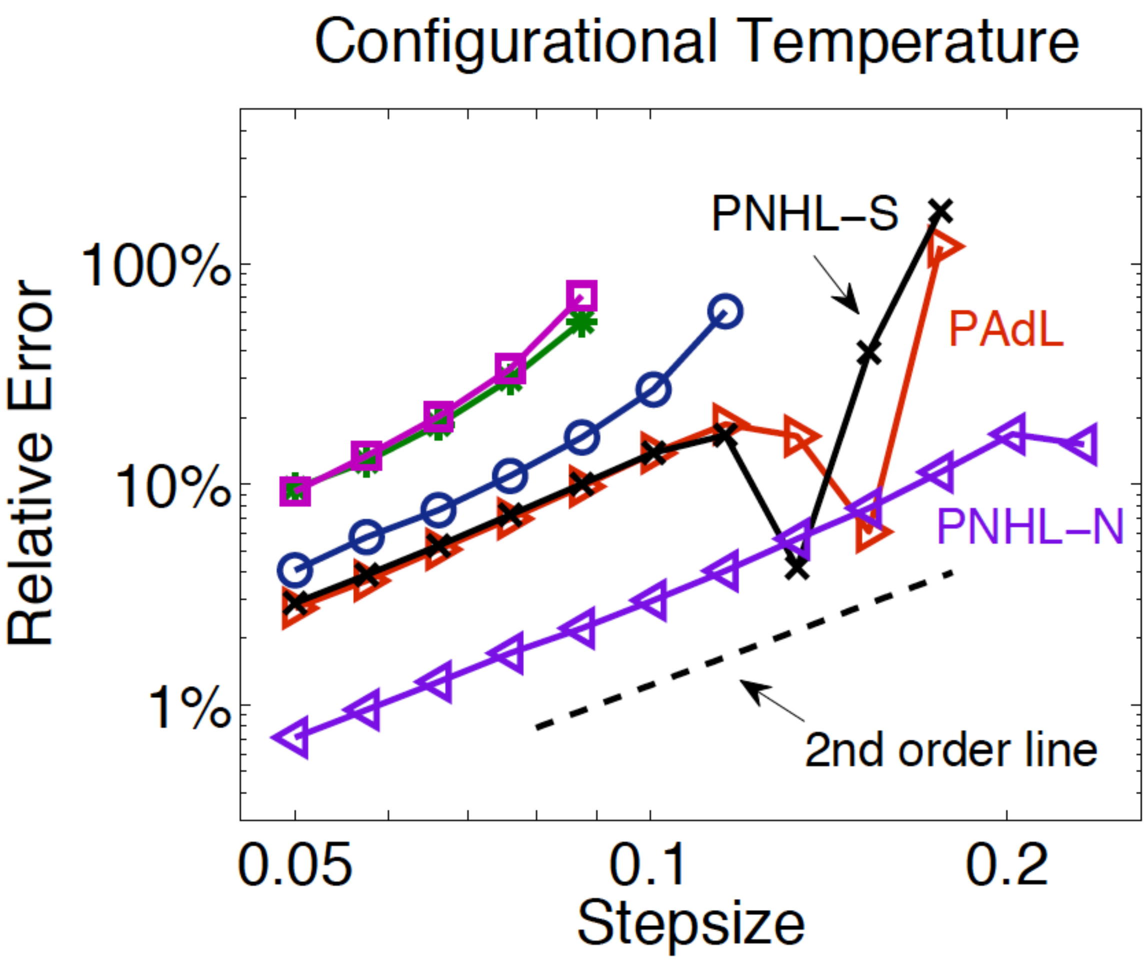
1. Configurational Temperature (after Rugh 1989)

$$k_B T_c = \frac{\sum_i \langle \|\nabla_i U\|^2 \rangle}{\sum_i \langle \nabla_i^2 U \rangle}$$

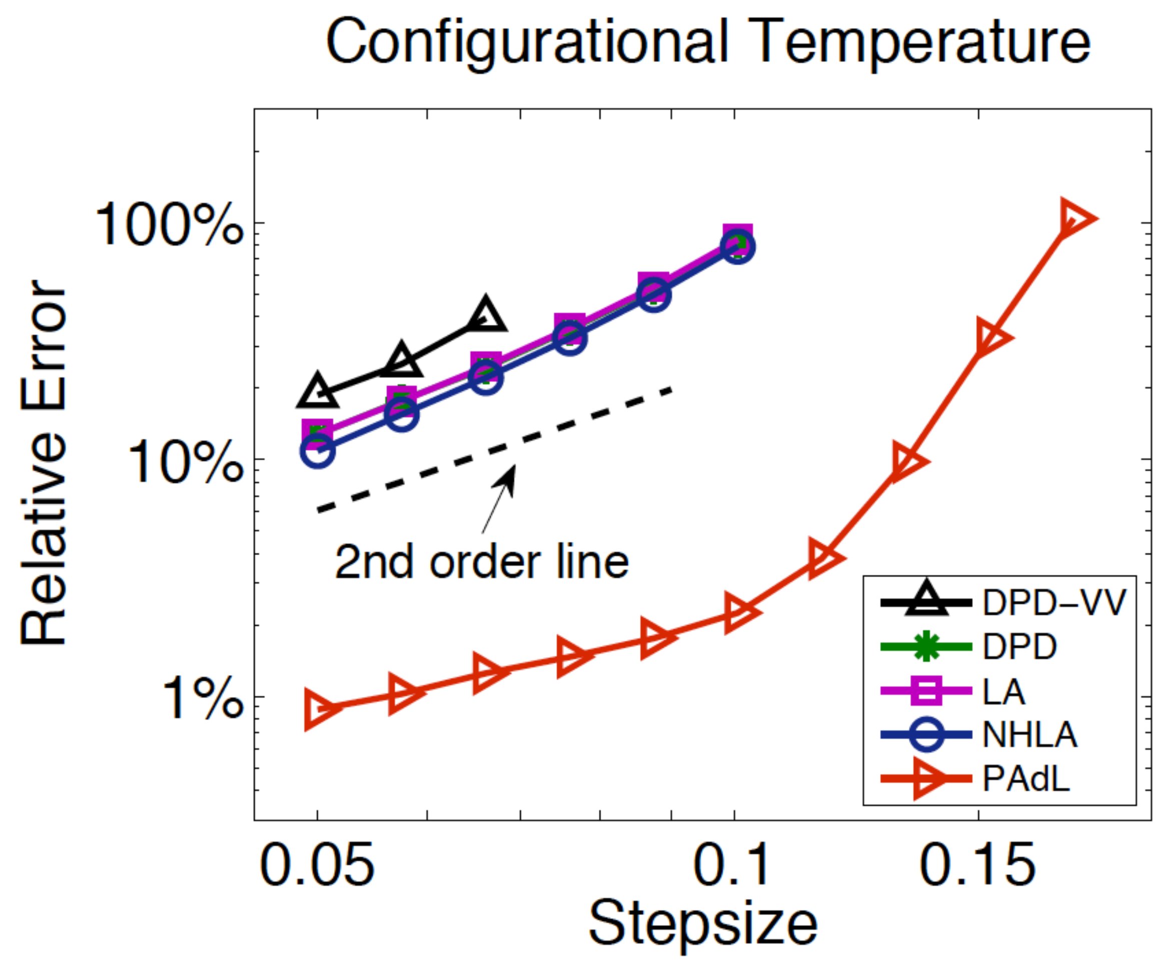
2. Velocity Autocorrelation Functions (averaged dynamics)

$$\psi(t) = \langle \mathbf{v}_i(t) \cdot \mathbf{v}_i(0) \rangle$$

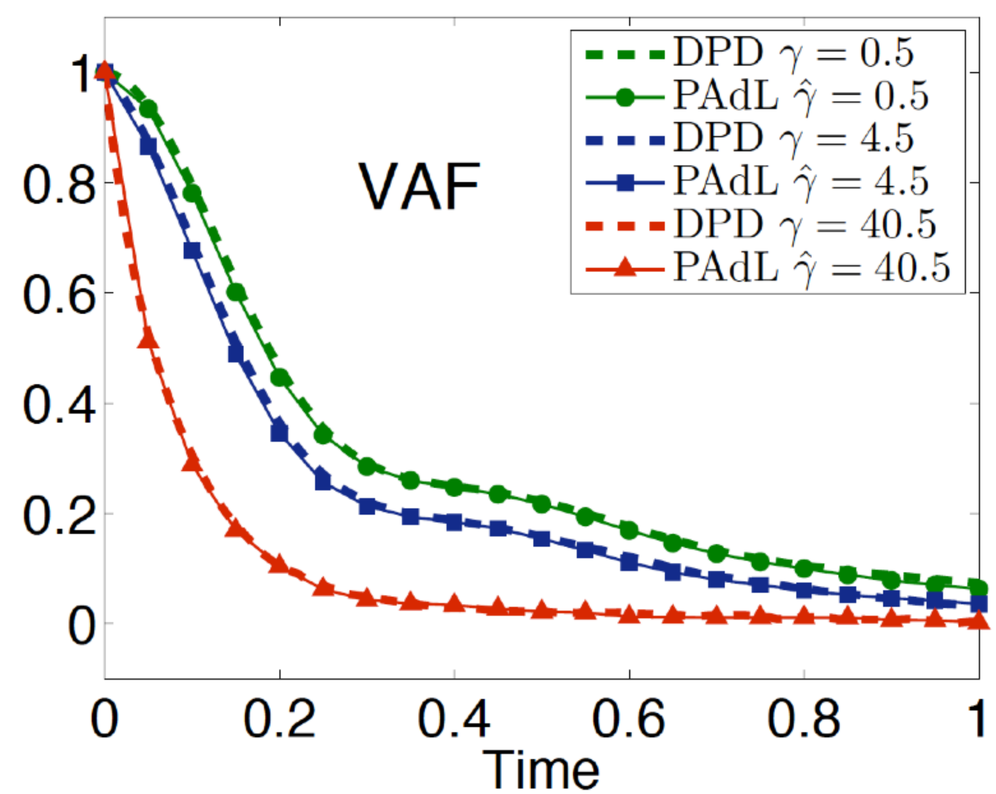
$\gamma = 0.5$



$\gamma = 40.5$

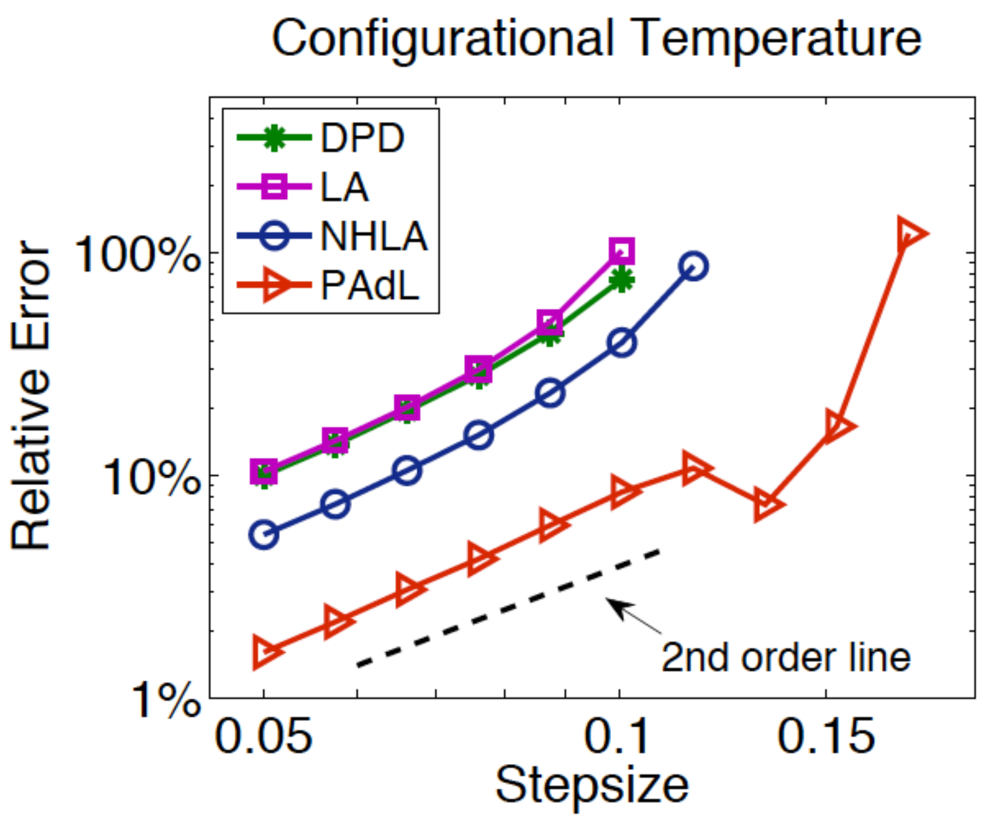


PAdL mimics DPD VAFs

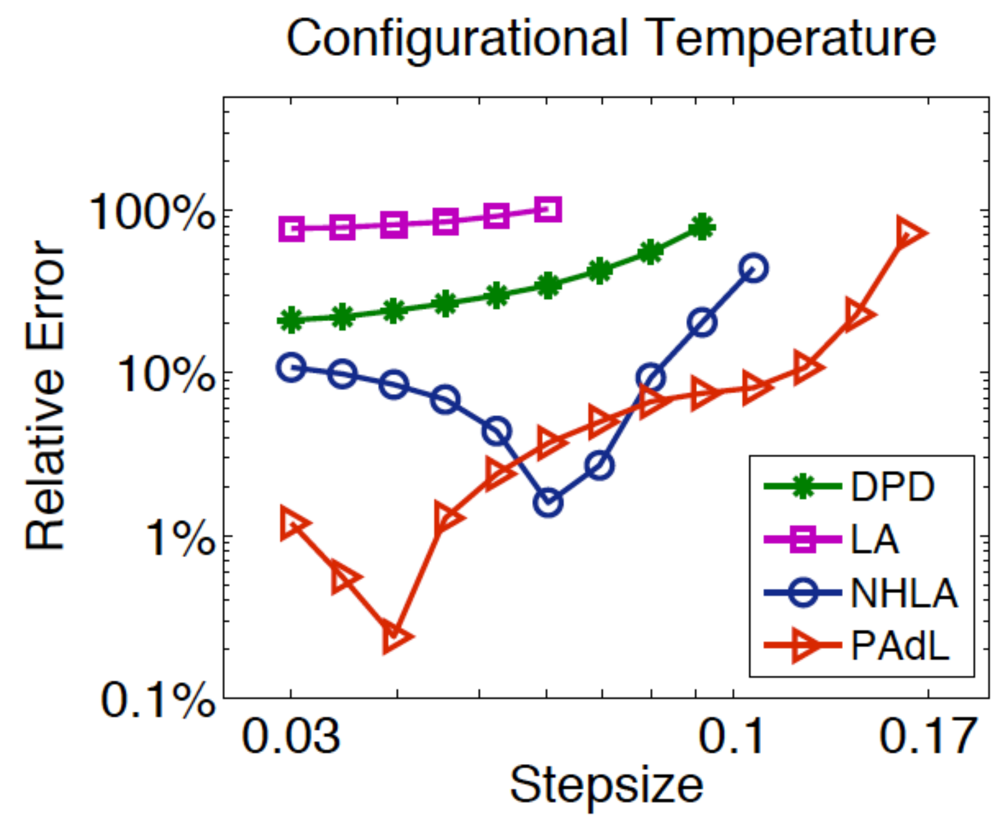


Sheared (Nonequilibrium MD) Simulations

Low Shear Rate



High Shear Rate



Method	Critical stepsize	Maximal stepsize	CPU time	Scaled efficiency
DPD-VV	0.05	0.10	20.212	100.0%
DPD-S1	0.05	0.11	20.618	98.0%
DPD-Trotter	0.05	0.11	21.451	94.2%
Peters	0.05	0.11	21.274	95.0%
LA	0.05	0.10	18.048	112.0%
NHLA	0.07	0.13	18.691	151.4%
PAdL	0.13	0.17	23.103	227.5%

DPD/NEMD

- Much **higher accuracy** for certain methods

- For driven systems (e.g. systems in shear flow) some **Adaptive DPD algorithms are much more stable** than alternatives.

  Recent work with M. Kroeger (ETH-Zurich), systematic study of DPD approaches with application to polymer models under various conditions (weak and strong shear rates).

Noisy gradients

$$\dot{q} = \frac{p}{m}\,, \qquad \dot{p} = F(q) - \xi p + \sigma_{\text{heat}}\,\dot{w}_{\text{heat}}\,, \qquad \dot{\xi} = \left[\frac{p^2}{m} - kT\right]/Q$$

$$\mathscr{L}^{\dagger}\left(\rho_\beta(q,p) \times \hat{\rho}(\xi)\right) = 0, \qquad \hat{\rho}(\xi) = \exp\left(-\,\beta Q(\xi - \xi_{\text{heat}})^2/2\right)$$

& ergodic for this equilibrium distribution

Bayesian Sampling

Understand choice of parameters q given observations X

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

Posterior probability density **(from Bayes' Theorem):**

$$p(X|q)p(q) = p(q|X)$$

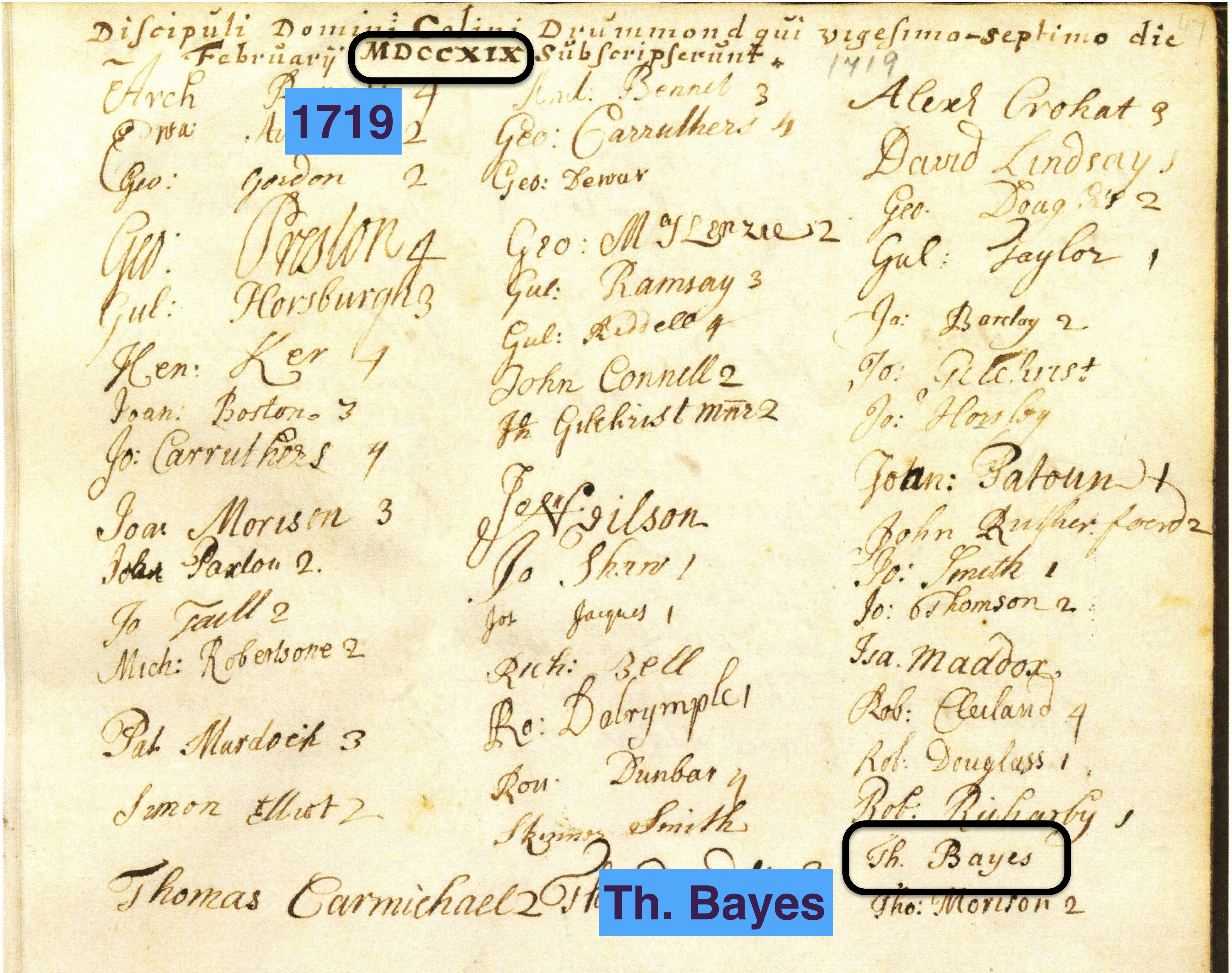
$$p(q|X) \propto \exp(-U(q)), \qquad U(q) = -\underbrace{\log p(X|q)}_{\textit{model}} - \underbrace{\log p(q)}_{\textit{prior}}$$

Use **Maximum Likelihood Estimate**/"Subsampling":

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

GRADIENT NOISE

Student Roll, University of Edinburgh



Discipuli Domini Colini Drummond qui vigesimo-septimo die Februarij MDCCXIX Subscripserunt.

1719

Arch [illegible] 4
[illegible] A[illegible] 2
Geo: Gordon 2
Gul: Preston 4
Gul: Horsburgh 3
Hen: Ker 4
Joan: Boston 3
Jo: Carruthers 4
Joan Morison 3
John Paxton 2
Jo Faull 2
Mich: Robertsone 2
Pat Murdoch 3
Simon Elliot 2
Thomas Carmichael 2

And: Bennet 3
Geo: Carruthers 4
Geo: Dewar
Geo: McKenzie 2
Gul: Ramsay 3
Gul: Riddell 4
John Connell 2
Jo Gilchrist mnr 2
Jo Wilson
Jo Shaw 1
Jos Jaques 1
Rich: Bell
Ro: Dalrymple 1
Rou: Dunbar 4
Skinner Smith
Th[illegible]

Alex Crokat 3
David Lindsay 1
Geo Douglas 2
Gul: Taylor 1
Ja: Barclay 2
Jo: Gilchrist
Jo: Horsley
Joan: Patoun 1
John Rutherford 2
Jo: Smith 1
Jo: Thomson 2
Isa. Maddox
Rob: Cleland 4
Rob: Douglass 1
Rob: Richardby 1
Th. Bayes
Tho: Morison 2

Th. Bayes

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**

What to do about the force error?

Methods for Gibbs sampling with a noisy gradient

- Ignore the perturbation
- Estimate the perturbation/correct for it
- SGLD (Langevin with a diminishing stepsize sequence)
- Adaptive thermostats Ad-L, Ad-NH,…

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

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. Easy 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 Thermostat [Jones and L., 2011]

The Adaptive Property

Jones & L. 2011

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

<u>Adaptive (Automatic) Langevin</u>

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

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

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

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

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

Discretization

[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**

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

typically anticipate 2nd order (IM)

Superconvergence

BAOAB, in the high friction limit, gives a superconvergence property for configurational quantities.

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

Effectively the extra driving noise implements a projection to the case of Langevin dynamics, **but large driving noise also implies large friction so restricted phase space exploration** (even if better accuracy). So caution is needed…

$$\mathscr{L}^\dagger(f_2\tilde{\rho}) = \mathscr{L}_2^\dagger\tilde{\rho}$$

$$\mathscr{L}^\dagger = -p\partial_q + U'(x)\partial_p + \xi\partial_p(p\cdot) + \frac{\gamma}{\beta}\partial_{pp} - \frac{1}{\mu}(p^2 - \beta^{-1})\partial_\xi$$

$$\varepsilon = 1/\hat{\gamma} = 1/\mu$$

$$\left(\bar{\mathcal{L}}_{\mathrm{O}}^\dagger + \varepsilon\mathcal{L}_{\mathrm{H}}^\dagger\right)\left(\hat{f}_{2,0} + \varepsilon\hat{f}_{2,1} + O(\varepsilon^2)\right)\rho_\beta = -\varepsilon\mathcal{P}\mathcal{L}_2^\dagger\tilde{\rho}_\beta$$

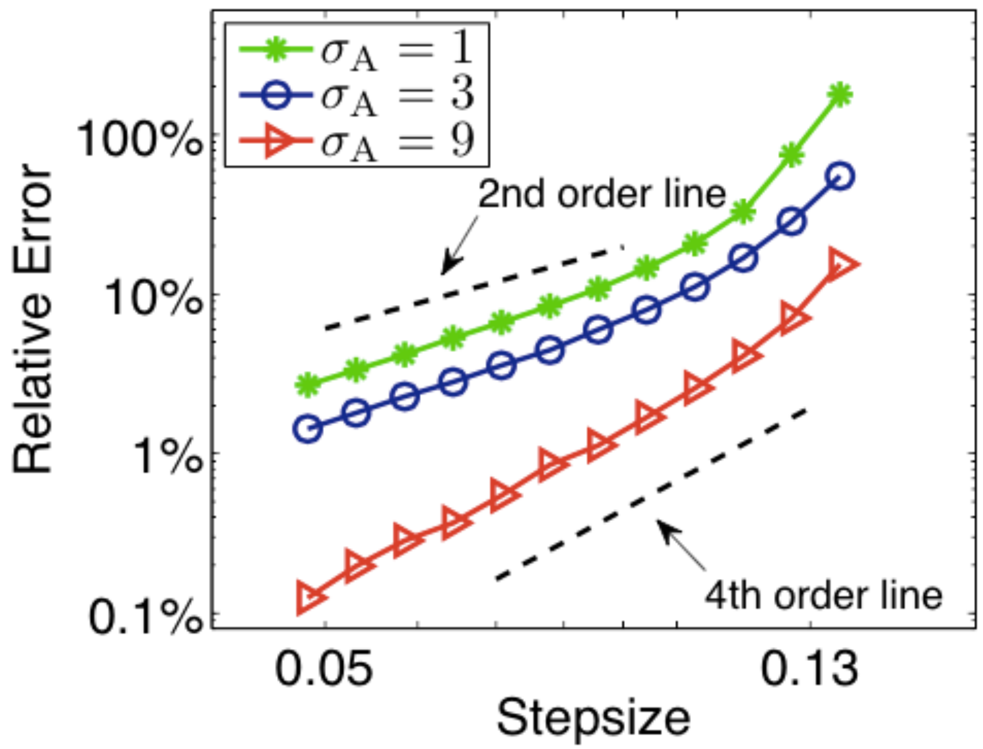
$$\mathcal{P}\mathcal{L}_2^\dagger\tilde{\rho}_\beta = \left(\frac{\beta}{12}\left[3pU'(x)U''(x) - p^3U'''(x)\right] + \frac{\hat{\gamma}}{12}\left[3U''(x) - 3\beta p^2U''(x) + \frac{1}{\mu}\left(6\beta p^4 - 28p^2 + 10\beta^{-1}\right)\right]\right)\rho_\beta$$

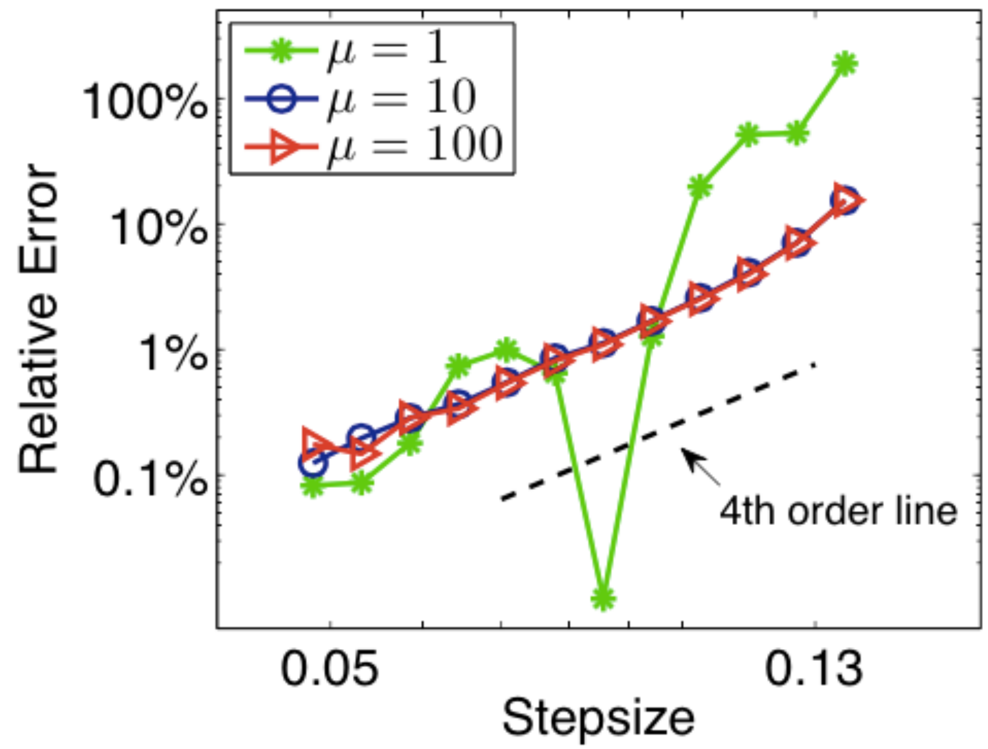
$$\hat{f}_{2,0} \equiv \hat{f}_{2,0}^{\mathrm{BADODAB}} = \frac{1}{8}\left(U''(x) - \beta p^2U''(x)\right)$$

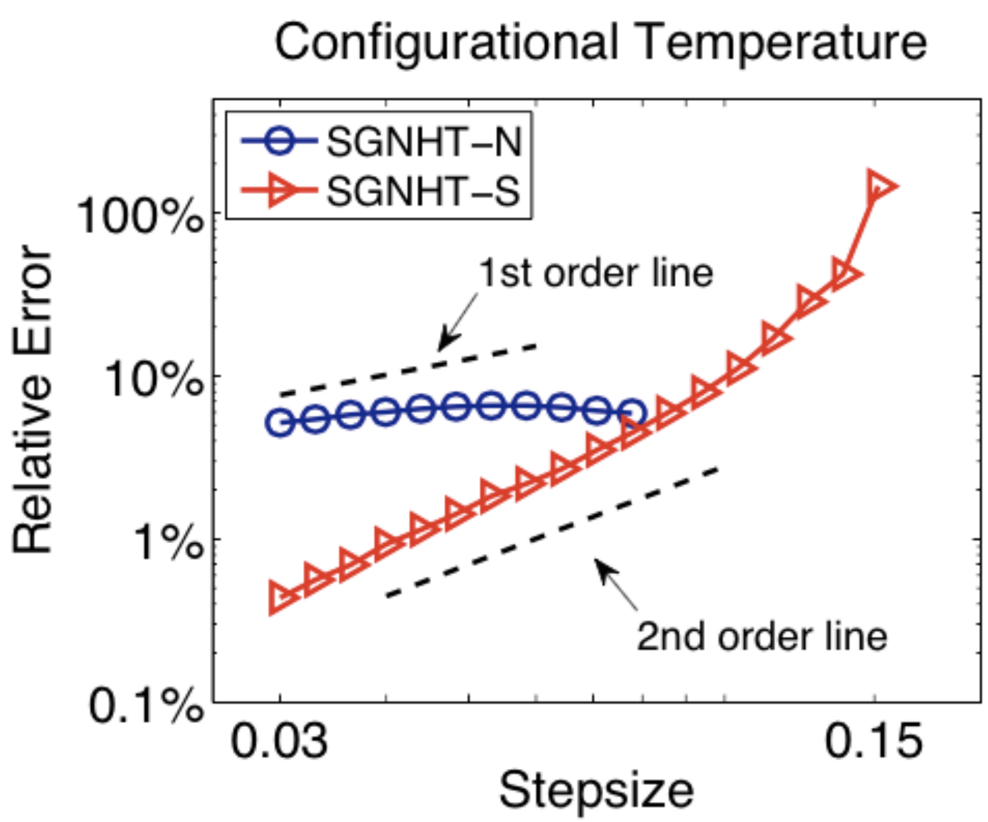
$$\langle\phi(x)\rangle_{\mathrm{BADODAB}} = \langle\phi(x)\rangle + h^2\langle\phi(x)\hat{f}_{2,0}^{\mathrm{BADODAB}}\rangle + O(\varepsilon h^2 + h^4)$$

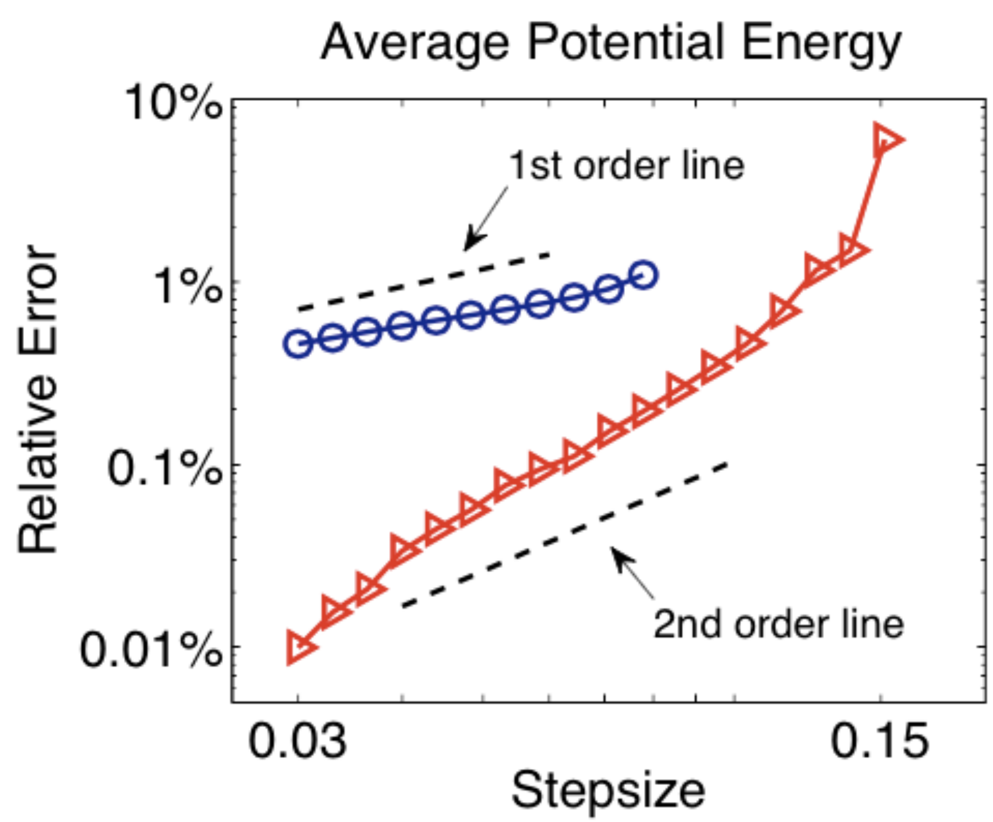
500 LJ particles, clean gradient

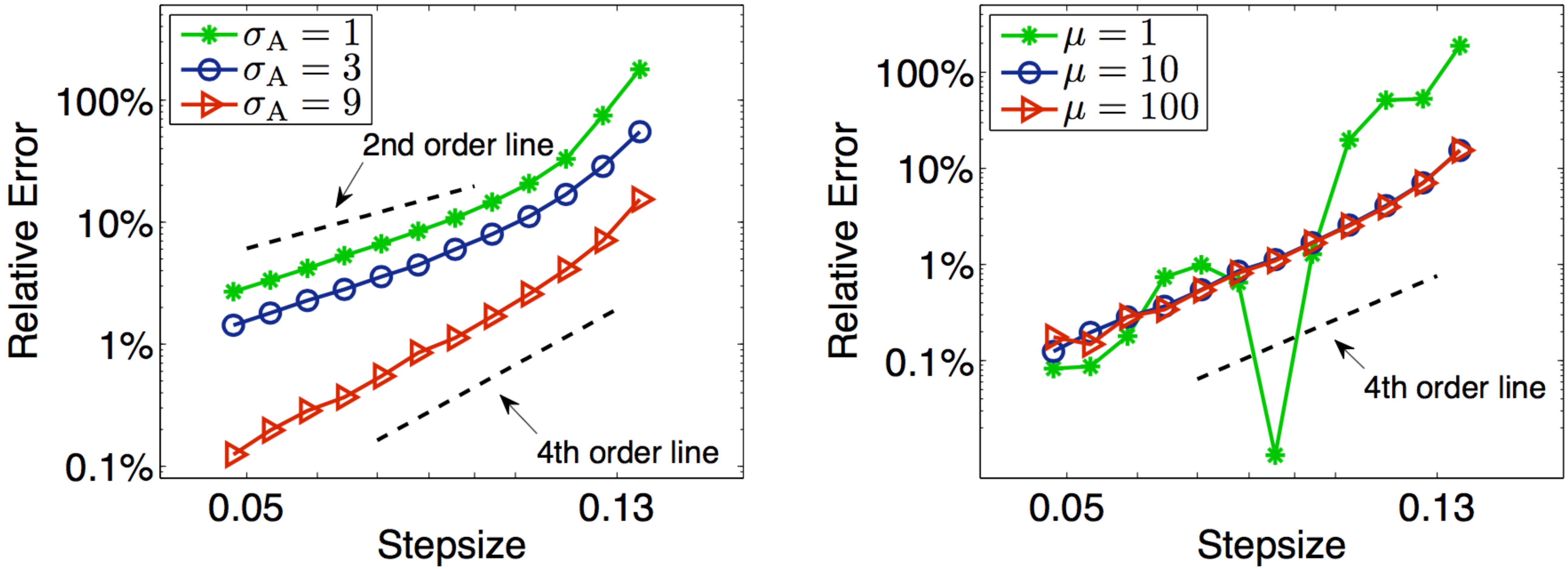
configurational temperature











- Fourth order convergence to the invariant measure
- Large friction ($\hat{\gamma} \propto \sigma_{\mathrm{A}}^2$) and thermal mass ($\mu$) limits
- Only one force calculation required at each step

Bayesian Logistic Regression

$$\pi(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = f(y_i \boldsymbol{\beta}^T \mathbf{x}_i)$$

f: logistic function

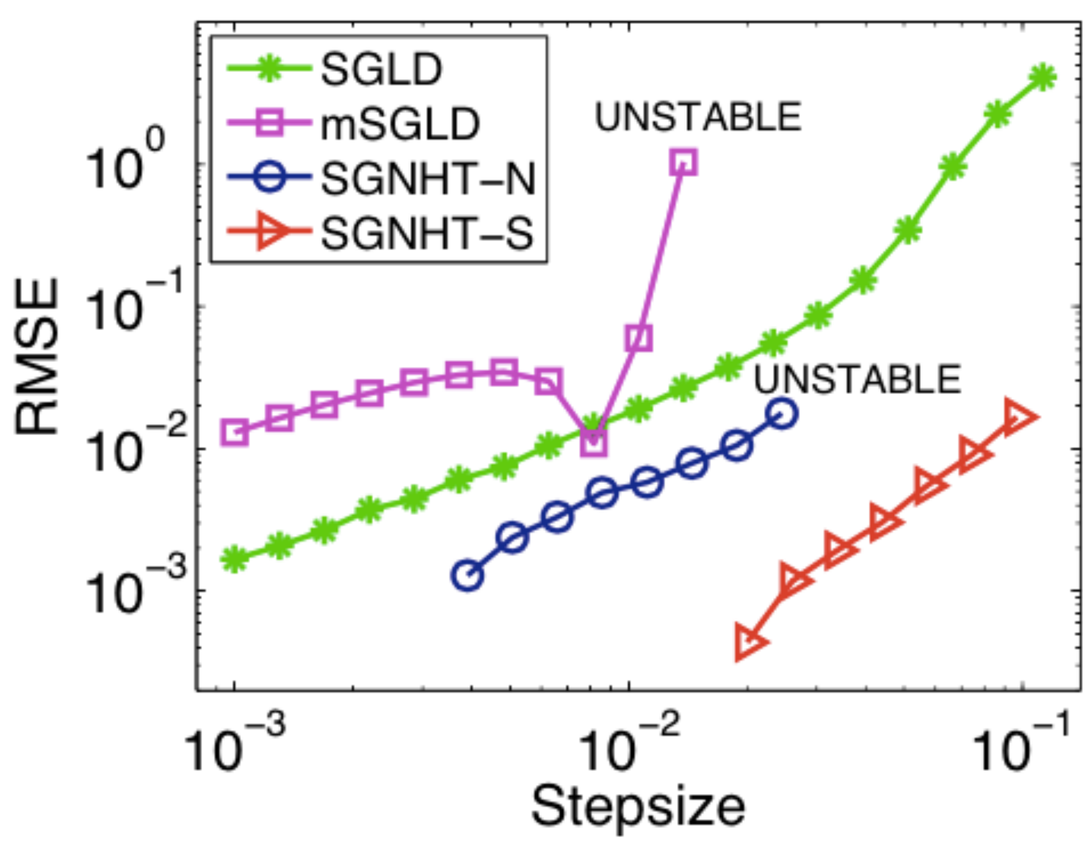
covariates e.g. age, income, …

data e.g. voting intention

posterior parameter distribution

$$\pi(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{\beta}\|^2\right) \prod_{i=1}^{N} f(y_i \boldsymbol{\beta}^T \mathbf{x}_i)$$

Gaussian prior



Covariance-Controlled Adaptive Langevin Dynamics

Shang, Zhu, Leimkuhler and Storkey, *NIPS* 2015

In the typical case, the noise may have a multivariate Gaussian distribution but with unknown (and evolving) covariance.

If we assume that we can obtain a covariance estimator then we can use this to enhance the accuracy of the SDEs.

CCAdL=
"Covariance Controlled Adaptive Langevin Dynamics" incorporates such a correction term together with an adaptive Langevin thermostat…

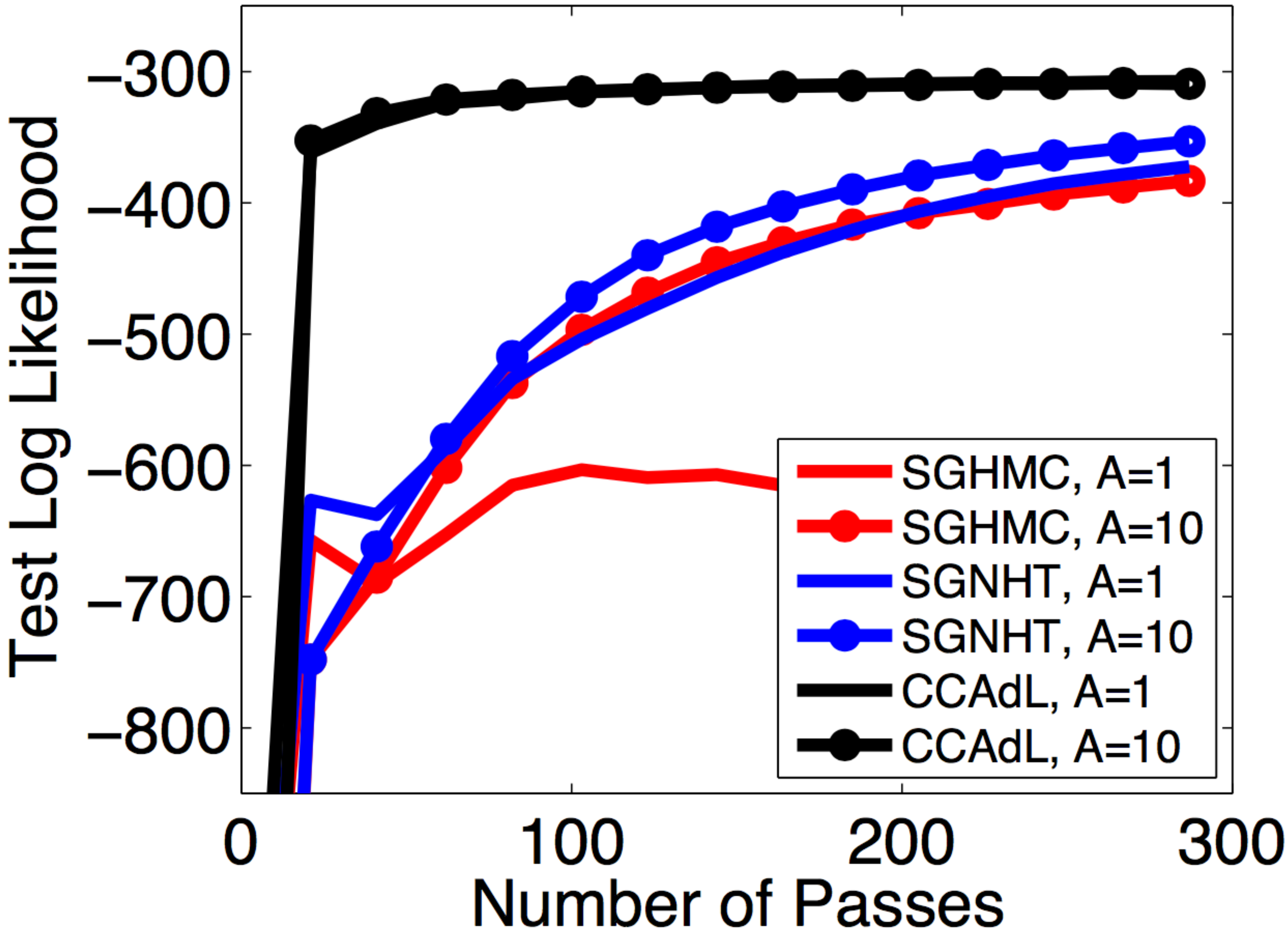
- Formulation

$$
\begin{aligned}
\mathrm{d}\mathbf{q} &= \mathbf{M}^{-1}\mathbf{p}\mathrm{d}t \\
\mathrm{d}\mathbf{p} &= -\nabla U(\mathbf{q})\mathrm{d}t + \sqrt{h\boldsymbol{\Sigma}(\mathbf{q})}\mathbf{M}^{1/2}\mathrm{d}\mathbf{W} - (h/2)\beta\boldsymbol{\Sigma}(\mathbf{q})\mathbf{p}\mathrm{d}t \\
&\quad - \xi\mathbf{p}\mathrm{d}t + \sqrt{2\hat{\gamma}\beta^{-1}}\mathbf{M}^{1/2}\mathrm{d}\mathbf{W}_{\mathrm{A}} \\
\mathrm{d}\xi &= \mu^{-1}\left[\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} - N_{\mathrm{d}}k_{\mathrm{B}}T\right]\mathrm{d}t
\end{aligned}
$$

with invariant distribution

$$
\tilde{\rho}_\beta(\mathbf{q},\mathbf{p},\xi) \propto \exp\left(-\beta H(\mathbf{q},\mathbf{p})\right)\exp\left(-\frac{\beta\mu}{2}\left(\xi-\hat{\gamma}\right)^2\right)
$$

- Parameter-dependent noise effectively dissipated by the additional covariance control term.



Binary classification of handwritten digits 7 and 9.

Very recent related work: NOGIN - C. Matthews and J. Weare

Ensemble preconditioning MCMC simulation

Goal: redesign the dynamics (and integrator) to enhance the rate of convergence for typical observables f

$$\langle f \rangle = \int f(x)\rho(x)\mathrm{d}x = \lim_{N\to\infty} N^{-1}\sum_{t=1}^{N} f(x_t)$$

figure of merit = **Integrated Autocorrelation Time (IAT)**

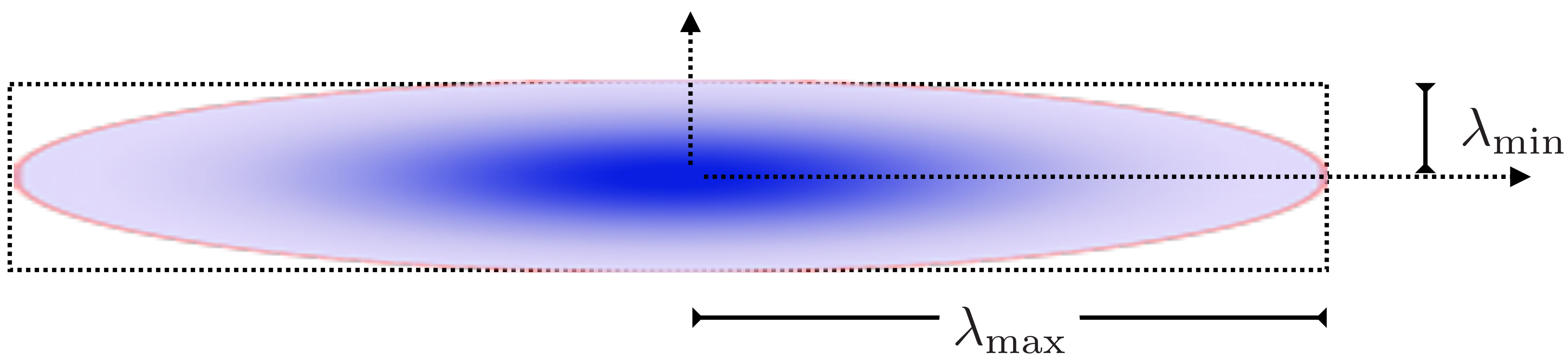
$$\tau_f = 1 + 2\sum_{t=1}^{\infty} \mathrm{cor}(f(x_t), f(x_0))$$

We would like to have τ_f as small as possible

Motivating Example:

$$\pi(x) = \exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)$$

Eigenvalues: $0 < \lambda_{\text{min}} =< \ldots < \lambda_{\text{max}} = \rho(\Sigma)$



For MCMC schemes like Euler-Maruyama or Leimkuhler-Matthews, **stability requires** $h = O(\lambda_{\text{min}})$

But for $f(x) = x \cdot \mathbf{e}_{\text{max}}$ $\tau_f = O(\lambda_{\text{max}}/\lambda_{\text{min}})$

Poor Scaling $\implies$ **Slow Convergence**

Ensemble Preconditioning

More generally, we wish to sample problems with complicated energy functions, where each basin or local approximation may be very poorly scaled.

Related Concepts

Stochastic Newton schemes

BFGS Method

MC Hammer (Goodman and Weare)

Compare to: Riemannian Manifold HMC (Girolami et al)

Ensemble Preconditioning

Use local information to estimate inverse Hessian matrix; precondition (rescale) dynamics to enhance convergence (reduce IAT)

Wishlist:

- Increase efficiency by reducing the IAT
- Compute the preconditioning based on a local ensemble approximation
- Allow for inertial effects (underdamped Langevin/HMC)

Idea: Use a collection of "walkers" to generate local covariance information and use this to estimate the inverse Hessian adaptively

Procedure

Use an ensemble of L walkers:

$$Q = (q_1, q_2, \ldots, q_L) \in \mathbb{R}^{dL}, \qquad P = (p_1, p_2, \ldots, p_L) \in \mathbb{R}^{dL},$$

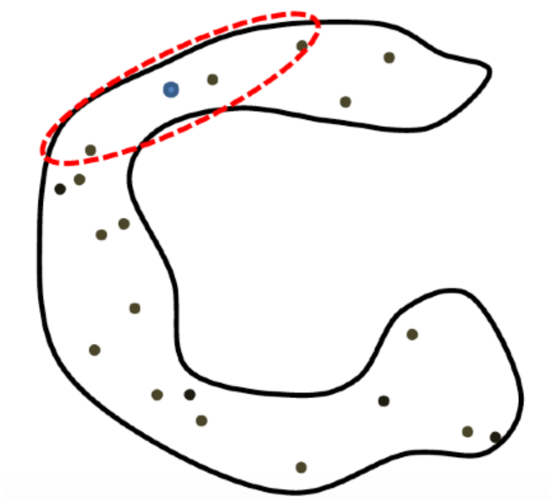
$$\bar{\pi}(Q,P) = \prod_{i=1}^{L} \hat{\pi}(q_i, p_i), \qquad \int \bar{\pi}(Q,P)\underset{\cdot}{\mathrm{P}} = \prod_{i=1}^{L} \pi(q_i).$$

each walker samples the same target distribution $\pi(q)$

We construct an dynamics in the extended space and compute ensemble averages by marginalisation over the individual walkers.

$$\dot{Q} = B(Q)P,$$
$$\dot{P} = B(Q)^{\mathrm{T}}\nabla\log(\pi(Q)) + \mathrm{div}(B(Q)^{\mathrm{T}}) - \gamma P + \sqrt{2\gamma}\eta(t).$$

$$B(Q) = \mathrm{diag}(B_1(Q), B_2(Q), \ldots, B_L(Q))$$



$$B_i(Q) = \sqrt{I_{\mathrm{d}} + \eta\,\mathrm{wcov}(Q_{[i]}, \omega_{\lambda(Q_{[i]}, q_i)})}$$

collects covariance info of nearby walkers

blend with identity (robustness)

$$Q_{[i]} = (q_1, q_2, \ldots, q_{i-1}, q_{i+1}, \ldots, q_L)$$

Basing B_i on other walkers only
eliminates the problems of multiplicative noise

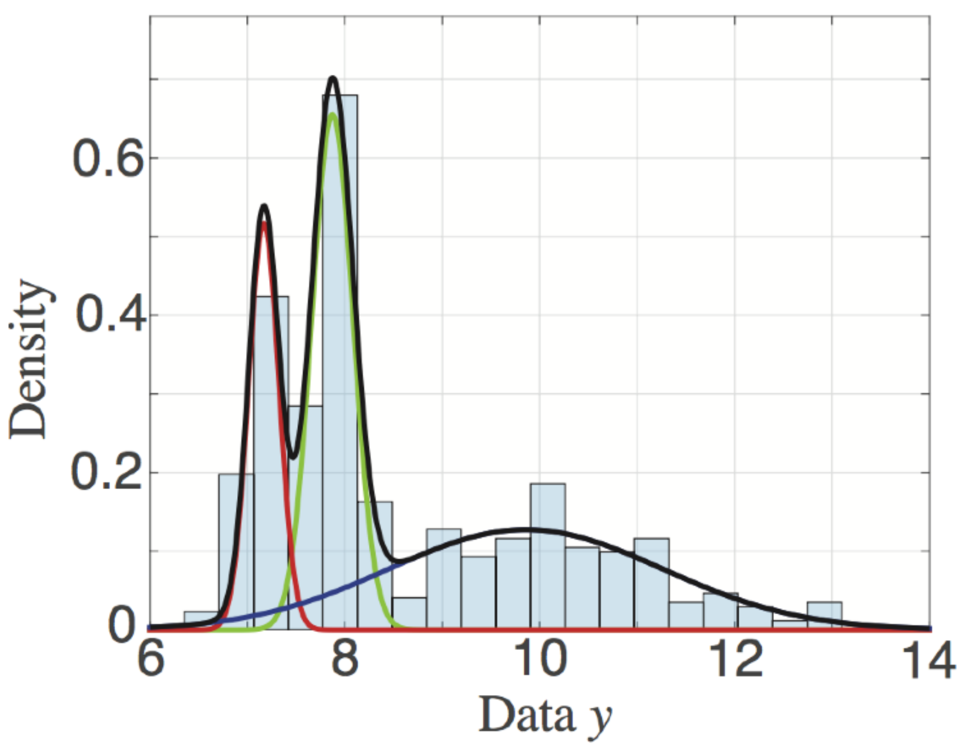
Discretization of SDEs: similar to **BAOAB**

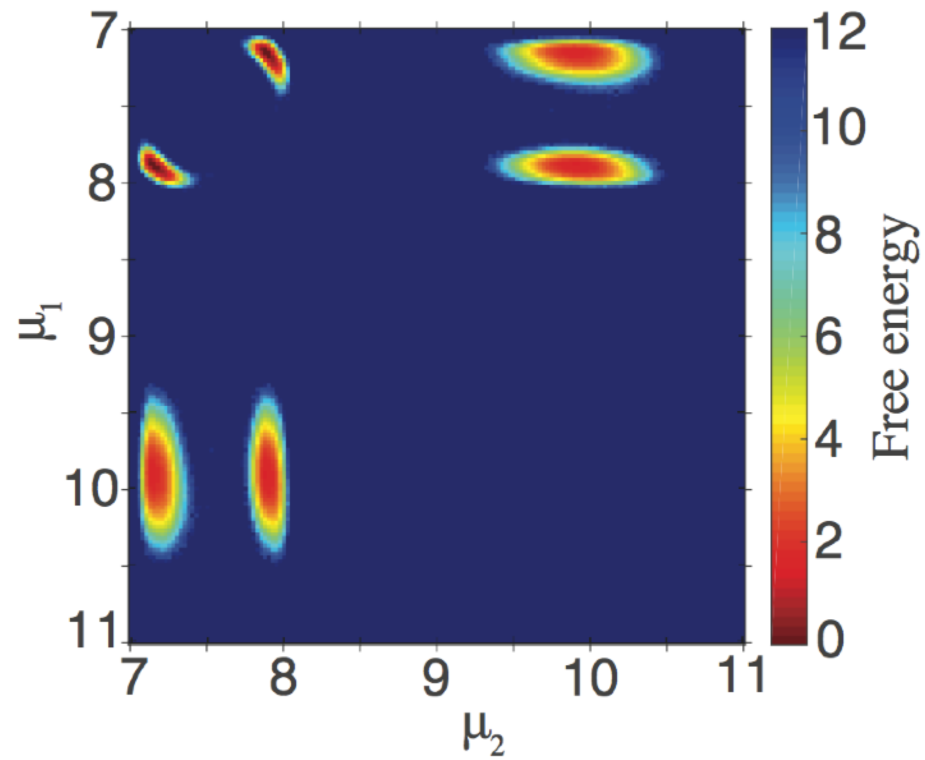
Gaussian Mixture Model: Hidalgo Stamps

"Adventures in Stamp Collecting"

$$\sum_{n=1}^{3} z_n \mathcal{N}(\mu_n, \lambda_n^{-2})$$

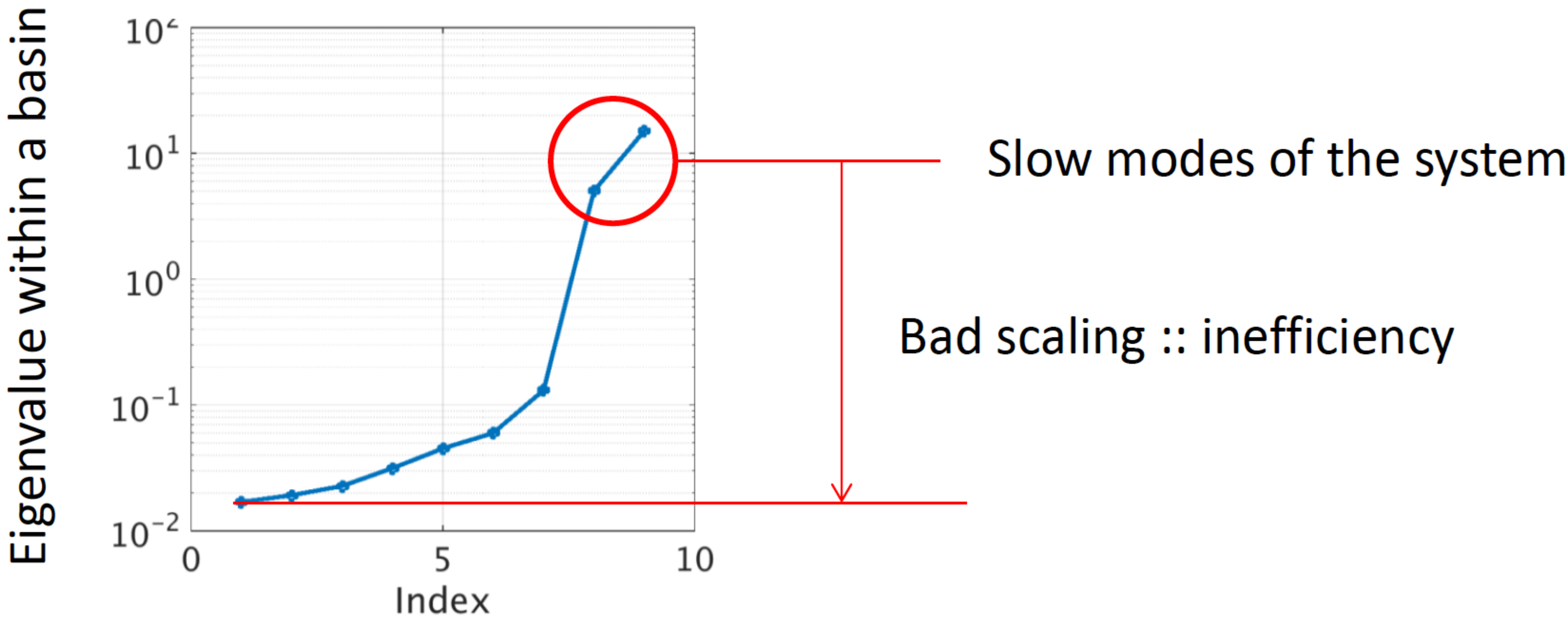
Dataset: Thickness of 485 stamps from Mexico in 1872.





- (moderately) poorly scaled basins
- multimodal due to "label switching" symmetry

Within one basin, we have bad scaling:



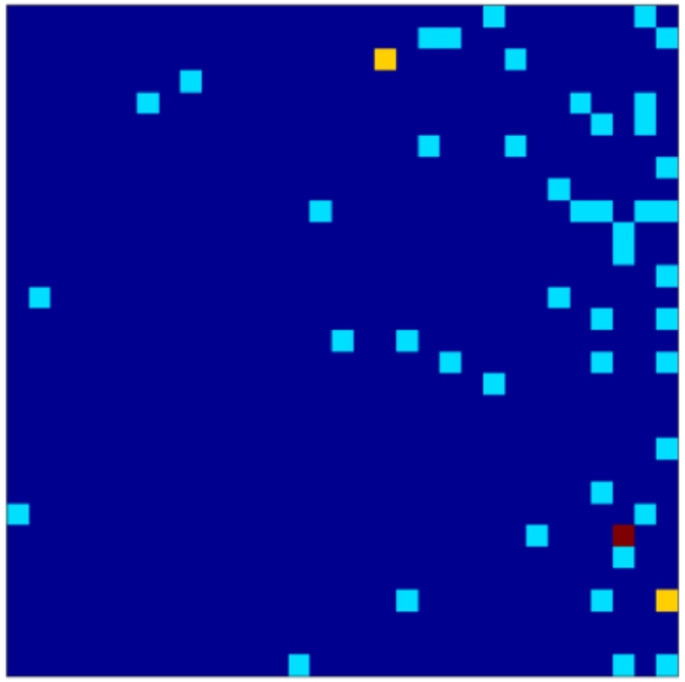
But the eigensystems are **different in different basins**, so the localized covariance is needed…

Gaussian Mixture Model: Hidalgo Stamps

Integrated Autocorrelation Times of Different Schemes

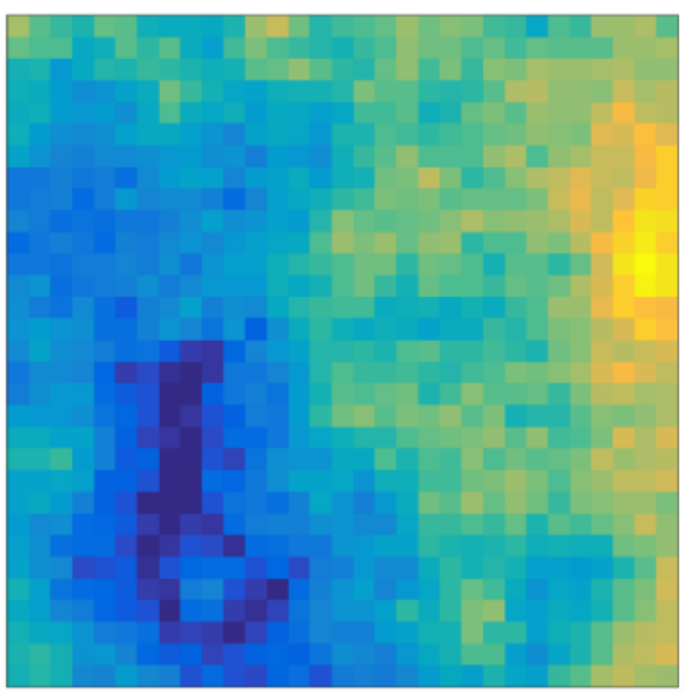
Scheme	$\min(z)$	$\max(\lambda)$	$\min(\mu)$	β
HMC	21495	**42935**	27452	7148
Langevin Dynamics	6825	**13279**	8384	4641
Ensemble Q-N	69	83	98	**115**

Numerical test: Log-Gaussian Cox model



Observations X

Means $\exp(Y_{i,j})$



Break $[0,1]^2$ into a 32x32 grid.

Observed intensity in box (i,j) is $X_{i,j}$, Poisson distributed with mean

$$\Lambda(i,j)/32^2, \qquad \Lambda(i,j) = \exp(Y_{i,j})$$

where $Y \sim N(\mu, \Sigma)$, with

$$\Sigma_{(i,j),(i',j')} = \sigma^2 \exp[-\sqrt{(i-i')^2 + (j-j')^2}/(32\beta)]$$

We generate synthetic data X using

$$\sigma^2 = 1.91, \quad \beta = 1/33, \quad \mu = \log(126) - \sigma^2/2$$

We fix μ and aim to infer likely Y, using hyperparameters σ^2, β, with prior $\mathrm{gamma}(2, \frac{1}{2})$

Log Gaussian Cox Model

Scheme	x	σ^2	β	Efficiency
HMC	800.7	1041.6	**1318.7**	1.0
RMHMC	2158.9	34.0	**1502.0**	0.15
LD	405.1	140.6	**435.3**	3.5
$\cdots$ (no Metropolis)	81.6	20.5	**136.5**	11.2
EQN	71.9	49.2	**239.5**	5.4
$\cdots$ (no Metropolis)	64.4	8.8	**47.8**	26.8

Ensemble Quasi-Newton python package

http://bitbucket.org/c_matthews/ensembleqn