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

Software and Implementation in Consortium Code

And don’t forget:
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 $\Rightarrow$ 100M ratings!

Posterior probability density (from Bayes’ Theorem):

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

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

Use Maximum Likelihood Estimate/“Subsampling”:

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

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

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: \( F(x) = -\nabla U(x) \)

**Brownian Dynamics**

\[
dx = F(x)\,dt + \sqrt{2}\,dW
\]

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

\[
\lim_{\tau \to \infty} \tau^{-1} \int_0^\tau \varphi(x(t))\,dt = \int_\Omega \varphi(x)\rho(x)\,dx
\]

**How to discretize?**

Euler-Maruyama? Stochastic Heun?
Euler-Maruyama Method

\[ x_0 \rightarrow x_1 \rightarrow \ldots \rightarrow x_N, \quad N h = \tau \]

Discrete Brownian path

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

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

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

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

For the **L-M method**, under suitable conditions,

\[ E\varphi(X_x(\tau)) - E\varphi(X_N) = C_0(\tau, x)h + C(\tau, x)h^2 \]

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

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

**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

Euler-Maruyama $h=0.025$

Leimkühler-Matthews $h=0.025$

L-M

Large stepsize

Euler-Maruyama $h=0.04$

Leimkühler-Matthews $h=0.04$
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

\[dx = M^{-1}p \, dt\]

\[dp = -\nabla U \, dt - \gamma M^{-1}p \, dt + \sqrt{2\beta^{-1}} \gamma dW\]

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

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

\[\text{Hamiltonian } H(x, p)\]
Splitting Methods for Langevin Dynamics

\[ \mathcal{L} = \mathcal{L}_A + \mathcal{L}_B + \mathcal{L}_O \]

\[ \mathcal{L}_A = (M^{-1}p) \cdot \nabla_x \]

\[ \mathcal{L}_B = -\nabla U(x) \cdot \nabla p \]

\[ \mathcal{L}_O = -\gamma (M^{-1}p) \cdot \nabla p + \gamma \beta^{-1} \Delta p \]

\[ e^{h\hat{\mathcal{L}}_{BAOAB}} = e^{\frac{h}{2} \mathcal{L}_B} e^{\frac{h}{2} \mathcal{L}_A} e^{h\mathcal{L}_O} e^{\frac{h}{2} \mathcal{L}_A} e^{\frac{h}{2} \mathcal{L}_B} \]
Expansion of the invariant distribution

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

Leading order:

\[ \mathcal{L}^\dagger (\rho_{\text{can}} f_2) = \beta^{-1} \mathcal{L}_2^\dagger \rho_{\text{can}} \]

*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

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

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.
Molecular Dynamics
With Deterministic and Stochastic Numerical Methods

Ben Leimkuhler
Charles Matthews

Interdisciplinary Applied Mathematics 39

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 \( \sigma(x) \approx \sigma \)

\[ x_{n+1} = x_n + hF(x_n) + h\sigma \tilde{R}_n + \sqrt{2hR_n} \]

\[ = x_n + hF(x_n) + \sqrt{h} \sqrt{h\sigma^2 + 2\hat{R}_n} \]

Like Euler-Maruyama discretization of

\[ dx = F(x)dt + \sqrt{2 + \sigma} hdW \]
\[ dx = F(x)dt + \sqrt{2 + \sigma}hdW \]

1. Step-size-dependent dynamics (like in B.E.A.)
2. Distorts temperature
3. Possible to correct - if we know \( \sigma \)
4. Computing/estimating \( \sigma \) 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

- Gradient System
- Unknown Noise Perturbation
- Negative Feedback Control
Adaptive Thermostats

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

\[ \frac{dx}{dt} = M^{-1} p dt \]
\[ \frac{dp}{dt} = -\nabla U dt - \sqrt{\hbar} \sigma dW - \xi dt + \sigma_A dW_A \]
\[ d\xi = \mu^{-1} \left[ p^T M^{-1} p - n \beta^{-1} \right] dt \]

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

Shift in auxiliary variable by \( \gamma = \frac{\beta (\hbar \sigma^2 + \sigma_A^2)}{2 \text{Tr}(M)} \)
Discretization  [With X. Shang, 2015]

generator: \( \mathcal{L} = \mathcal{L}_A + \mathcal{L}_B + \mathcal{L}_O + \mathcal{L}_D \)

\( \mathcal{L}_A = (M^{-1}p) \cdot \nabla_x \)

\( \mathcal{L}_B = -\nabla U(x) \cdot \nabla p + \frac{h\sigma^2}{2} \Delta p \)

\( \mathcal{L}_O = -\xi p \cdot \nabla p + \frac{\sigma^2_A}{2} \Delta p \)

\( \mathcal{L}_D = G(p) \frac{\partial}{\partial \xi} \)

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

\( e^{h\mathcal{L}} = e^{\frac{h}{2} \mathcal{L}_B} e^{\frac{h}{2} \mathcal{L}_A} e^{\frac{h}{2} \mathcal{L}_D} e^{h\mathcal{L}_O} e^{\frac{h}{2} \mathcal{L}_D} e^{\frac{h}{2} \mathcal{L}_A} e^{\frac{h}{2} \mathcal{L}_B} \)
BADODAB \approx 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?

![Graphs showing test log likelihood over the number of passes for different algorithms and their variants, with annotations and axes labels.](image)
Problem: sample all the basins accessible at a given temperature in a realistic simulation time.
Continuous Tempering


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

Replica Exchange

Higher Temperature

Physical Temperature

Swap Attempt

Swap Attempt

Swap Attempt

Swap Attempt
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

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

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? …