Bayesian Sampling using Stochastic Gradient Thermostats
–history, development and extension

Changyou Chen

Joint work with Nan Ding, Youhan Fang, Ryan Babbush, Robert Skeel, Hartmut Neven
October 3, 2014
Outline

1. History and Preliminary

2. Discretization simulation of SDEs

3. Stochastic Gradient Thermostats

4. Information-Geometric MCMC using Diffusions
Four developments of classical mechanics:

- Classical mechanics:
  - Newton's second law of motion

- Lagrangian mechanics:
  - Generalized coordinate systems

- Hamiltonian mechanics:
  - With auxiliary conjugate coordinate

- Statistical mechanics:
  - Introduces probability into systems
Newton’s second law of motion – a second order ordinary differential equation (ODE):

\[ m \frac{d^2 \theta(t)}{dt^2} = f(\theta(t)), \]

where \( m \) denotes the mass, \( \theta \) the location, \( f \) the force and \( t \) the time.
Lagrangian mechanics

- Transformation of the classical mechanics using generalized coordinates:

\[ \frac{d}{dt} \left( \frac{\partial L}{\partial \dot{\theta}} \right) = \frac{\partial L}{\partial \theta}, \]

\( L = \) kinetic energy – potential energy, \( \{\theta_j\}'s \) are the generalized coordinates, \( \{\dot{\theta}_j\} \) the generalized velocities.

- The use of generalized coordinates may usually simplify a system’s analysis.
Hamiltonian mechanics

- A Legendre transformation of the Lagrangian mechanics with a set of canonical coordinates \( r = (\theta, p) \).
- The time evolution of the system:
  \[
  \frac{d\theta}{dt} = \frac{\partial H}{\partial p} \tag{1}
  \]
  \[
  \frac{dp}{dt} = - \frac{\partial H}{\partial \theta} \tag{2}
  \]

where \( H = H(\theta, p, t) \) is the Hamiltonian, which often corresponds to the total energy of the system: \( H = \frac{1}{2}p^T M^{-1} p + V(\theta) \).
- It is a first order differential equation.
- \( V(\theta) \) would be the log-likelihood of the model in machine learning.
Introduces probability theory into classical mechanics:
- define probability over all possible states of the system

Bridges the gap between mechanical laws and everyday life experiences:
- e.g., not necessary to know at a microscopic level of each molecule while performing a chemical reaction

Explains the thermodynamic behavior of large systems:
- e.g., system of a jug of gas with large amount of molecules
Statistical mechanics: Liouville equation

- Statistical mechanics allows us to describe the time evolution of the state probabilities of a mechanical system.
- Consider the differential equation of the form:
  \[
  \frac{dx}{dt} = v(x,t), \tag{3}
  \]
  where \( x \) is variables of the system:
  - in Hamiltonian mechanics, \( x = \{\theta, p\} \)
- Let \( \rho(x,t) \) be the probability distribution of the states at time \( t \). Liouville equation:
  \[
  \frac{\partial \rho}{\partial t} + \nabla_x \cdot (\rho v) = 0 \quad \tag{4}
  \]
  \[
  \iff \frac{\partial \rho}{\partial t} + \sum_i \frac{\partial (\rho v_i)}{\partial x_i} = 0. \quad \tag{5}
  \]
In Hamiltonian mechanics, \( H = \frac{1}{2} p^T M^{-1} p + V(\theta) \).

Liouville equation becomes:

\[
\frac{\partial}{\partial t} \rho + \sum_i \left( \dot{p}_j \frac{\partial}{\partial p_j} \rho + \dot{\theta}_i \frac{\partial}{\partial \theta_i} \rho \right) = 0.
\]

It is easy to check that \( \tilde{\rho} \propto e^{-H} \) is the stationary distribution of the system, i.e., with \( t \rightarrow \infty \).
Consider a statistical thermodynamics (also called statistical mechanics) surrounded by a large heat bath, the motion of particles will usually affected by exchanging energy with the heat bath.

One can introduce stochastic processes into the system to simulate this phenomena.

A typical solution is to add random Gaussian noise into the ODE’s, resulting in stochastic differential equations (SDE) of the form:

$$dx_t = v(x_t, t)dt + \sqrt{2D(x_t, t)}dW_t,$$

where \(v\) is called the drift coefficient, \(\sqrt{2D}\) the diffusion coefficient. \(W_t\) is the Wiener process/Brownian motion.
How does the random noise affect the equilibrium distribution of the system?

\[ \text{d}x_t = v(x_t, t) \text{d}t + \sqrt{2D(x_t, t)} \text{d}W_t \]

- The time evolution of state distribution \( \rho(x_t, t) \) can be described by a generalization of the Liouville equation – Fokker-Planck equation (FP):

\[
\frac{\partial}{\partial t} \rho + \nabla_x \cdot (\rho v) - \nabla_x \nabla_x : (\rho D) = 0, \tag{7}
\]

where \( X : Y = \text{trace}(X^T Y) \).

- The FP equation is particularly useful in identifying the correctness of the equilibrium distribution of the system:
  - in machine learning, it corresponds to model posterior
Outline

1. History and Preliminary
2. Discretization simulation of SDEs
3. Stochastic Gradient Thermostats
4. Information-Geometric MCMC using Diffusions
Most SDEs (as well as ODEs) can not be solved exactly. Discretized simulation is required:

- for some particular ODEs, e.g., Hamiltonian mechanics, a second order of discretization error method called leapfrog method is preferable
- for others, the most popular one is Euler’s method, but with a first order of discretization error.
Discretization simulation of SDEs

ODEs

\[ \frac{dx}{dt} = v(x, t) \]

- **Euler’s method** (error = \(O(h)\)):
  \[ x^t = x^{t-1} + hv(x^{t-1}) , \tag{8} \]
  where \(h\) is the step size.

- **(Generalized) leapfrog method** (error = \(O(h^2)\)):
  - divide \(x\) into two parts \(x = (x_1, x_2)\)
  \[ x^{t+\frac{1}{2}}_1 = x^t_1 + \frac{h}{2} v(x^{t+\frac{1}{2}}_1, x^t_2) , \tag{9} \]
  \[ x^{t+1}_2 = x^t_2 + \frac{h}{2} \left[ v(x^{t+\frac{1}{2}}_1, x^t_2) + v(x^{t+\frac{1}{2}}_1, x^{t+1}_2) \right] , \tag{10} \]
  \[ x^{t+1}_1 = x^{t+\frac{1}{2}}_1 + \frac{h}{2} v(x^{t+\frac{1}{2}}_1, x^{t+1}_2) . \tag{11} \]
\[ \mathbf{d}x_t = \mathbf{v}(x_t, t)dt + \sqrt{2D(x_t, t)}dW_t \]

- **Euler’s method:**
  \[ x^t = x^{t-1} + hv^{t-1} + \sqrt{2D(x^{t-1})} \cdot \xi, \quad (12) \]
  where \( h \) is the step size and \( \xi \) are a set of \( i.i.d. \) normal random variables.

- **(Generalized) leapfrog method:**
  - some stochastic versions of the leapfrog method exist, \( e.g. \), [Qiang & Habib, 2000]
  - seems not popularly used, not sure if is useful
Outline

1. History and Preliminary
2. Discretization simulation of SDEs
3. Stochastic Gradient Thermostats
4. Information-Geometric MCMC using Diffusions
Bayesian methods in machine learning is popular.

Difficult when dealing with large scale data.

A Bayesian version of the popular stochastic gradient descend algorithm is needed.
Existing works using subsets of data

- Use a subset of data to evaluate the acceptance probabilities in the Metropolis-Hastings algorithm.
  - *e.g.*, [Balan *et al.*, 2014, Bardenet *et al.*, 2014]
  - not cover here

- Use stochastic gradient in Bayesian models.
  - *e.g.*, [Welling & Teh, 2011, Patterson & Teh, 2013, Chen *et al.*, 2014]
  - ideas not taken from statistical thermodynamics/statistical mechanics, empirically find lack of stabilization

- We propose a novel stochastic gradient thermostats algorithm – **stochastic Nosé-Hoover thermostat**:
  - deal with big data in Bayesian learning
  - generalizes the ideas from Hamiltonian mechanics to statistical mechanics
  - get better performances than previous works
Stochastic gradient Hamiltonian Monte Carlo (SGHMC) [Chen et al., 2014]

Hamiltonian: \[
\frac{d\theta}{dt} = \frac{\partial H}{\partial p}, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial \theta}, \quad H = V(\theta) + \frac{1}{2} p^T M^{-1} p
\]

- In SGHMC, first, we calculate the stochastic gradient with a subset of data $S \subset \{x_1, \cdots, x_N\}$:
  \[
  \tilde{V}(\theta) = -N \frac{1}{|S|} \sum_{x_i \in S} \log p(x_i | \theta) - \log p(\theta)
  \] (13)

- Let $\tilde{v}(\theta) = \nabla_{\theta} \tilde{V}(\theta)$, with step size $h$, assume the following assumption:
  \[
  hv(\theta) \approx h\tilde{v}(\theta) + N(0, 2hB(\theta)) .
  \] (14)

- A naive construction of SGHMC would be (assume $M = I$):
  \[
  d\theta = pdt, \quad dp = \tilde{v}(\theta)dt + N(0, 2B(\theta)dt)
  \] (15)
Stochastic gradient Hamiltonian Monte Carlo (SGHMC) [Chen et al., 2014]

\[
\begin{align*}
\mathrm{d}\theta &= p\,\mathrm{d}t, \\
\mathrm{d}p &= \tilde{v}(\theta)\,\mathrm{d}t + N(0, 2B(\theta)\,\mathrm{d}t)
\end{align*}
\]  \hspace{1cm} (16)

However, using the Fokker-Planck equation, we can check that the equilibrium distribution of system (16) is not the model posterior.

SGHMC overcome this by adding an additional friction term into the system:

\[
\begin{align*}
\mathrm{d}\theta &= p\,\mathrm{d}t, \\
\mathrm{d}p &= \tilde{v}(\theta)\,\mathrm{d}t - B(\theta)p\,\mathrm{d}t + N(0, 2B(\theta)\,\mathrm{d}t)
\end{align*}
\]  \hspace{1cm} (17)

(17) is an SDE with the model posterior \(V(\theta)\) as its equilibrium distribution.
Some potential disadvantages of SGHMC:
- the friction term $B(\theta)$ is hard to estimate
- it might encounter ergodicity issue

In thermodynamics, the system interacts with its surroundings:
- a typical example is a heat bath, which is held at a prescribed temperature, regardless of how much heat might be drawn from it

Intuitively, a heat bath provides energy to make the atom travel around the whole state space, making it less prone to the ergodicity issue, though no theoretical guarantee.

In addition, our propose stochastic gradient thermostats adaptively updates the friction term.
In thermodynamics, a system in thermal equilibrium with a heat bath at fixed temperature $T$ has the following thermal equilibrium condition [Tuckerman, 2010]:

$$\frac{k_B T}{2} \approx \frac{p^T p}{n},$$

where $k_B$ is known as the Boltzmann constant and $n$ the degree of freedom of the system:
- in Bayesian statistics, $k_B T = 1$, $n$ is the dimension of $\theta$

We want to incorporate this condition into our algorithm.
As a result, we generalize the SGHMC to SGNHT:

\[
\begin{align*}
    d\theta &= pdt, \\
    dp &= \tilde{v}(\theta)dt - \xi pdt + N(0, 2Cdtdt), \\
    d\xi &= \left(\frac{1}{n}p^T p - 1\right)dt
\end{align*}
\]  

The formulas without the stochastic term $N(0, 2Cdtdt)$ is known in thermodynamics as Nosé-Hoover thermostat.

With the Fokker-Planck equation we can show that

**Theorem**

The system (18)-(20) has exactly the model posterior as its thermostatic equilibrium distribution.
**The algorithm**

**Input:** Parameters $h$, $A$.
Initialize $\theta(0) \in \mathbb{R}^n$, $p(0) \sim N(0, I)$, and $\xi(0) = C$;

**for** $t = 1, 2, \ldots$ **do**

- Evaluate $\tilde{v}(\theta_{(t-1)})$ with a random subset of data;
- $p(t) = p(t-1) - \xi(t-1)p(t-1)h - \tilde{v}(\theta_{(t-1)})h + \sqrt{2}CN(0, h)$;
- $\theta(t) = \theta(t-1) + p(t)h$;
- $\xi(t) = \xi(t-1) + \left(\frac{1}{n}p(t)^{\top}p(t) - 1\right)h$;

**end**

**Algorithm 1:** Stochastic Gradient Nosé-Hoover Thermostat
Illustration on Double-Well potential

Figure: Plots of density $p(\theta)$ and kinetic energy $\frac{p^T p}{2}$ with $\xi = 1$ (1st col), 10 (2nd col), 0.1 (3rd col), SGNHT (4th col).
Bayesian Neural Network [Chen et al., 2014]:
- MNIST dataset, 50,000 for training, 10,000 for validation, 10,000 for testing
- 100 hidden layer

Bayesian Matrix Factorization [Salakhutdinov & Mnih, 2008]:
- Movielens dataset, 6,050 users, 3,883 movies, about 1M ratings
- Netflix dataset, 480,046 users, 17,000 movies, about 100M ratings
- 80% for training, 20% for testing for each dataset

Latent Dirichlet Allocation [Blei et al., 2003]:
- 5 year ICML abstracts, 765 documents (small)
- 80% for training, 20% for testing
Results

MNIST ($\eta = 2 \times 10^{-7}$)

Test Error

iterations

$\cdot 10^4$

Netflix ($\eta = 6 \times 10^{-7}$)

Test RMSE

iterations

$\cdot 10^5$

MovieLens1M ($\eta = 6 \times 10^{-7}$)

Test RMSE

iterations

$\cdot 10^5$

ICML ($\eta = 2 \times 10^{-5}$)

Test Perplexity

iterations

$\cdot 10^4$
Some open theoretical issues

- The discretization of continuous SDEs introduces bias into the solution.

- Some work such as [Sato & Nakagawa, 2014] has proved a finite time weak convergence for the stochastic gradient Langevin dynamics (SGLD):
  - still not satisfactory, we need long time convergence, i.e., an invariant measure with $t \to \infty$
  - potential solution might borrow ideas from [Leimkuhler et al., 2014], which shows an invariant measure for SGLD

- The discretization error can also be corrected by using MCMC based method:
  - use the discretized solution as a proposal in the Metropolis-Hastings algorithm
  - disadvantage: prohibit from large scale learning
Outline

1. History and Preliminary
2. Discretization simulation of SDEs
3. Stochastic Gradient Thermostats
4. Information-Geometric MCMC using Diffusions
Metropolis-Hastings algorithm

**Input:** Candidate transition kernel \( Q(x, \cdot) \) with density \( q(\cdot | x) \)

**Input:** A target distribution \( \pi(x) \)

**Output:** Generate samples from \( \pi(x) \)

Random initialize \( x_0 \);

for \( i = 1, 2, \cdots \) do

- Draw \( x' \sim Q(x_{i-1}, \cdot) \);
- Draw \( r \sim \text{Uniform}[0, 1] \);
- Set \( \alpha(x_{i-1}, x') = 1 \wedge \frac{\pi(x')q(x_{i-1}|x')}{\pi(x_{i-1})q(x'|x_{i-1})} \);
- if \( r < \alpha(x_{i-1}, x') \) then
  - Set \( x_i = x' \);
- else
  - Set \( x_i = x_{i-1} \);
end

Return \( \{x_1, x_2, \cdots\} \) as samples from \( \pi(x) \);
Key step: proposals

- Usually a random walk proposal is adopted:
  - \( i.e., Q(x, \cdot) = \mathcal{N}(x, \lambda^2 \Sigma) \)
  - does not use the posterior/gradient information, leads to slow mixing

- Recent advances in MCMC use gradient based proposals to improve the mixing of MH:
  - Hamiltonian MCMC uses a proposal generated from ODEs
  - Stochastic gradient thermostats based MCMC uses a proposal generated from SDEs\(^1\):
    - Langevin dynamics based MH, Hamiltonian dynamics based MH, Thermodynamics based MH, etc.
    - would also benefit from considering the information geometric of the posteriors [Girolami & Calderhead, 2011]

\(^1\)Haven’t been done in general.
We will be using proposals from the solutions of diffusions. We focus on time-homogeneous Itô diffusions government by the following SDE:

\[
d\theta_t = v(\theta_t)dt + \sigma(\theta_t)dW_t, \quad \theta_0 = x_0
\]  

- \(W_t\) is the Wiener process, drift vector \(v\) and volatility matrix \(\sigma\) are Lipschitz continuous
- same form as our stochastic gradient thermostats

Using the discretized solutions of SDEs as proposals, we get the class of Metropolis-adjusted thermodynamics algorithm, e.g., the Metropolis-adjusted Langevin algorithm [Girolami & Calderhead, 2011].

However, we want more:
- using the Manifold information of the target posterior to get more effective proposals
Manifolds

- A topological space $M$ that is locally Euclidean:
  - uniquely determined by its tangent bundle
  $$TM = \bigcup_{x \in M} \{(x, y) | y \in \text{tangent space of } x\}$$

- Let $x := \{x_1, x_2\} = r^{-1} \circ \gamma : [0, 1] \to \mathbb{R}^2$, then the distance between $A$ and $B$ is:
  $$L(A, B) = \int_0^1 \sqrt{\dot{x}(t)^T G(x) \dot{x}(t)} \, dt, \quad (22)$$

  where $G_{ij}(x) = \langle \frac{\partial r}{\partial x_i}, \frac{\partial r}{\partial x_j} \rangle$ being a local metric matrix.
Diffusions on Manifolds

- Assume the following diffusion (in machine learning):

\[
d\theta_t = \nabla \log \pi(\theta_t) dt + dW_t \tag{23}
\]

- It can be shown that in a manifold \( M \) with local metric matrix \( G(\theta) \):

\[
\nabla \log \pi(\theta_t) \mapsto \nabla \log \pi(\theta)|G(\theta)|^{-\frac{1}{2}} \tag{24}
\]

\[
dW_t \mapsto \Omega(\theta_t)dt + \sqrt{G^{-1}(\theta_t)}dW_t \tag{25}
\]

with \( \Omega_i(\theta) = \frac{1}{2} |G(\theta)|^{-\frac{1}{2}} \sum_{j=1}^{n} \frac{\partial}{\partial \theta_j} \left(|G(\theta)|^{\frac{1}{2}} \{G^{-1}(\theta)\}_{ij}\right) \tag{26} \)

- Diffusions on manifolds are obtained by substituting (24) and (25) into (23).
We see the key step in using manifold based MCMC is to define the local metric matrix $G(\theta)$.

No need to know the transformation $r(\cdot)$ and $\gamma(\cdot)$ explicitly.

Viewing a class of distributions $p(y|\theta)$ (the likelihood in machine learning) as elements of a Riemannian manifold, a suitable choice of $G(\theta)$ is the Fisher metric [Rao, 1945]:

$$G(\theta) = \mathbb{E}_{y|\theta} \left[ -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(y|\theta) \right]$$

Other metrics are also applicable, an open problem.
An open question for future work

Would it be possible to apply the information geometric methods in the stochastic thermostats framework (without doing a Metropolis-Hastings step)?

- need to guarantee the thermal equilibrium distribution is the model posterior\(^2\)
- need to prove the existence of an invariant measure after applying the discretization simulation for SDEs

\(^2\)A potential solution has appeared [Xifara et al., 2014].
Thanks for your attention!!!
References I

Qiang, J., Habib, S.:
A second-order stochastic leap-frog algorithm for Langevin simulation.
XX International Linac Conference (2000)

Balan, K., Chen, Y., Welling, M.:
Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget.
International Conference on Machine Learning (2014)

Bardenet, R., Doucet, A., Holmes, C.:
International Conference on Machine Learning (2014)

Sato, I., Nakagawa, H.:
Approximation Analysis of Stochastic Gradient Langevin Dynamics by using Fokker-Planck Equation and Ito Process.
International Conference on Machine Learning (2014)

Patterson, S., Teh, Y. W.:
Stochastic gradient Riemannian Langevin dynamics on the probability simplex.
Leimkuhler, B., Matthews, C., Stoltz, G.:
The Computation of Averages from Equilibrium and Nonequilibrium Langevin Molecular Dynamics.
IMA Journal of Numerical Analysis (2014)

Rao, R.:
Information and accuracy attainable in the estimation of statistical parameters.
Bulletin of the Calcutta Mathematical Society (1945)

Welling, M., Teh, Y. W.:
Bayesian learning via stochastic gradient Langevin dynamics.
International conference on machine learning (2011)

Girolami, M., Calderhead, B.:
Riemann manifold Langevin and Hamiltonian Monte Carlo methods.

Chen, T., Fox, E., Guestrin, C.:
Stochastic gradient Hamiltonian Monte Carlo.
International Conference on Machine Learning (2014)

Tuckerman, M., E.:
Statistical Mechanics: Theory and Molecular Simulation.
Oxford University Press (2010)
