Dynamics-based MCMC Methods

Chang Liu

Tsinghua University

June 1, 2016

Outline

1. Introduction
2. D-MCMCs with Deterministic Dynamics
   - Deterministic dynamics and its density evolution
   - Stationary distribution
   - Some discussions on HMC
3. SG-MCMCs with Stochastic Dynamics
   - Stochastic dynamics and its density evolution
   - Framework for the dynamics with desired stationary distribution
   - Stochastic dynamics and SG-MCMCs
4. Convergence Properties for Integrators of SG-MCMCs
   - Numerical integrators for a stochastic dynamics
   - Setup for describing an integrator
   - Convergence Analysis
   - Practical Numerical Integrators

Introduction

- What is dynamics-based MCMC methods (D-MCMCs)?
  A kind of MCMC methods that generate samples by simulating a dynamics.
- What is a dynamics?
  A rule for a state to evolve over time. We only consider dynamics under which a state evolves is continuous with respect to time.
  - deterministic: can be described by differential equations. E.g. Hamilton dynamics.
  - stochastic: can be described by stochastic differential equations. E.g. Brownian motion, Langevin dynamics.
- Why use dynamics?
  - Arbitrarily sophisticated differentiable target distributions.
  - Samples with high mixing rate (low correlation).

- Why dynamics can generate desired samples?
  - Only samples from the stationary distribution (may not be unique) of the dynamics can be drawn.
    Stationary: suppose at some time there are a lot of states satisfying a certain distribution. Each of the states evolves under the dynamics for a same time interval and after that all the states behaves another distribution. If the latter distribution is identical to the initial one for any time interval, then the initial distribution is stationary.
  - If the current sample is from the stationary distribution, then its evolved state is also from the stationary distribution. So it can be the next sample.

D-MCMCs with Deterministic Dynamics

Based on (Neal, 2011)

Deterministic dynamics

- A deterministic dynamics described by a differential equation: $\dot{z} = f(z)$. $z = z(t)$: "the position of a particle", a function of t.
  E.g. Hamilton dynamics: $z = (q, p) \in \mathbb{R}^{2n}$, for a Hamiltonian $H(z) \in \mathbb{R}$, $(\dot{q}, \dot{p}) = (\nabla_p H(z), -\nabla_q H(z))$.
- A distribution on states: $\pi(z, t)$. z: "a position in space", independent on t.
  A distribution is stationary under the dynamics: $\frac{\partial}{\partial t}\pi(z, t) = 0$. Thus it can be denoted as $\pi^s(z)$.
  Interpretation: for a large number of independent particles $\{1, \ldots, N\}$ with their position evolution $\{z_i(t)\}_{i=1}^N$, at some time point t_1, their positions $\{z_i(t_1)\}_{i=1}^N$ obeys the distribution $\pi(z, t_1)$. At another time point t_2, their positions $\{z_i(t_2)\}_{i=1}^N$ obeys the distribution $\pi(z, t_2)$.
- How it evolves over time when each $z_i(t)$ evolves under a deterministic dynamics?

Evolution of $\pi(z,t)$

We choose some bounded region in the state space and bind the region with the particles it contains and let the region evolve as all the contained particles evolve. Denote the evolving region as $\Omega(t)$. Then

$$0 \xlongequal[\text{conservation}]{\text{particle number}} \frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega(t)} \pi(z,t) \mathrm{d}z$$

$$\xlongequal[\dot{z}=\dot{z}(z,t)\text{: a vector field; } \pi(z,t)\text{: a scalar field}]{\text{Reynolds transport theorem \& Gauss theorem}} \int_{\Omega(t)} \left(\frac{\partial \pi}{\partial t} + \nabla \cdot (\pi \dot{z}) \right) \mathrm{d}z$$

holds for any $\Omega(t)$, so

$$\frac{\partial \pi}{\partial t} = -\nabla \cdot (\pi \dot{z}) \xlongequal{\text{calculus}} -\nabla \pi \cdot \dot{z} - \pi \nabla \cdot \dot{z}.$$

Refer to the dynamics description $\dot{z} = f(z)$,

$$\frac{\partial}{\partial t} \pi(z,t) = -\nabla \cdot \big(\pi(z,t) f(z) \big),$$

which is the Fokker-Planck equation (FPE) with no diffusion.

Condition for a stationary distribution

$$\frac{\partial \pi}{\partial t} = -\nabla \pi \cdot \dot{z} - \pi \nabla \cdot \dot{z}.$$

- $\nabla \cdot \dot{z} = 0 \Leftrightarrow$
  Liouville's theorem holds $\Leftrightarrow$ "Volume preservation" in (Neal, 2011) $\Leftrightarrow$
  The Jacobian determinant of infinitesimal time evolution is 1.
  - Liouville's theorem: $\frac{\mathrm{d}\pi(z(t),t)}{\mathrm{d}t} \overset{\text{calculus}}{=\!=} \frac{\partial \pi}{\partial t} + \dot{z} \cdot \nabla \pi \overset{\text{iff } \nabla \cdot \dot{z} = 0}{=\!=} 0.$
  - Volume preservation: $\frac{\mathrm{d}V}{\mathrm{d}t} \triangleq \frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega(t)} \mathrm{d}z \overset{\substack{\text{Reynolds transport theorem} \\ \text{Gauss theorem}}}{=\!=\!=} \int_{\Omega(t)} \nabla \cdot \dot{z} \mathrm{d}z \overset{\text{iff } \nabla \cdot \dot{z} = 0}{=\!=} 0.$
- Hamilton dynamics satisfies Liouville's theorem (thus $\nabla \cdot \dot{z} = 0$, or volume preservation):

$$\nabla \cdot \dot{z} = \nabla_q \cdot \dot{q} + \nabla_p \cdot \dot{p} = \nabla_q \cdot \nabla_p H - \nabla_p \cdot \nabla_q H = 0$$

More generally, a symplectic dynamics preserves volume. Hamilton dynamics is symplectic.

Condition for a stationary distribution

$$\frac{\partial \pi}{\partial t} = -\nabla \pi \cdot \dot{z} - \pi \nabla \cdot \dot{z}.$$

- If the dynamics conserves some scalar $H(z)$ (not explicitly involving t, thus $\partial H/\partial t = 0$), i.e. $\mathrm{d}H/\mathrm{d}t = 0$, and if at some instant $t = t_0$, $\pi(z, t_0)$ takes the form $g\big(H(z)\big)$ for some differentiable $g : \mathbb{R} \to \mathbb{R}$, then $\pi(z, t)$ satisfies $\nabla \pi \cdot \dot{z} = 0$ at the instant:

$$\nabla \pi \cdot \dot{z} = g'(H(z)) \nabla H \cdot \dot{z} \overset{\text{calculus}}{=\!=\!=} g'(H(z)) \left(\frac{\mathrm{d}H}{\mathrm{d}t} - \frac{\partial H}{\partial t} \right) = 0.$$

- Hamilton dynamics conserves Hamiltonian $H(z)$:

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \dot{z} \cdot \nabla H = \dot{q} \cdot \nabla_q H + \dot{p} \cdot \nabla_p H = \nabla_p H \cdot \nabla_q H - \nabla_q H \cdot \nabla_p H = 0.$$

Condition for a stationary distribution

In all,

- For a deterministic dynamics $\dot{z} = f(z)$ with $\nabla \cdot f(z) = 0$ and some scalar $H(z)$ conserved, distributions in the form $\pi(z,t) = g\big(H(z(t))\big)$ with g differentiable are its stationary distributions, thus can be denoted as $\pi^s(z)$.
- Hamilton dynamics has stationary distributions in the form $\pi^s(z) = g\big(H(z)\big)$.

Discussions on the stationary distribution

Some issues on "Hamilton dynamics has stationary distributions in the form $\pi^s(z) = g\big(H(z)\big)$".

- From the form $\pi = g\big(H(z)\big)$, we already have $\partial\pi/\partial t = 0$ i.e. stationary, so why bother?
  $\partial\pi/\partial t = 0$ iff stationary, but the form $g\big(H(z)\big)$ should be interpreted as an initial distribution $\pi(z, 0)$. Before we conclude that beginning with this distribution, the distribution afterwards remains the same (i.e. the distribution is stationary), we cannot say that $\pi(z, t) = g\big(H(z)\big)$ holds for all $t > 0$.
- Now that the stationary distribution is not unique for Hamilton dynamics, which stationary distribution are we sampling from when using HMC?
  It depends on the initial state, which determines the proportions over different energies. (States with the same energy are uniformly preferred.) In HMC, switching between different energies according to our desired proportion is done by resampling the momentum p from $\mathcal{N}(0, M)$. This determines the stationary distribution of HMC operations is $\exp\{-H(z)\}$, as desired.

Discussions on the stationary distribution

Some issues on "Hamilton dynamics has stationary distributions in the form $\pi^s(z) = g\big(H(z)\big)$".

- Intuition: we begin with $g\big(H(z)\big)$, or many sets of particles that particles in the same set have the same Hamiltonian (energy) and uniformly distributed on states with the Hamiltonian. The proportion of particles in the set with Hamiltonian E is $g(E)$. The particles evolve under Hamilton dynamics. Since the dynamics conserves Hamiltonian, particles in different sets do not mix while evolving, so the set with Hamiltonian E shares weight $g(E)$ constantly. Since Liouville's theorem $\mathrm{d}\pi/\mathrm{d}t = 0$ holds for Hamilton dynamics, i.e. for each particle at any time the density around its evolved state is the same as the density around its initial state, so at any time every state of the same Hamiltonian is occupied by some particle (otherwise the particle number is changed) thus at any time the density over states of the same Hamiltonian keeps uniform. So the distribution is stationary.

Some Ideas

- Other forms of the Hamiltonian? Especially other forms of the kinetic energy?
- Other forms of $g(H)$ apart from $g(H) = \exp\{-H\}$?
- Switch different energies directly by resampling energy from its proper distribution $g(E)\Omega(E)$ ($\Omega(E) = \frac{\mathrm{d}}{\mathrm{d}E}\int_{H(z)\leq E}\pi(z)\mathrm{d}z$) instead of resampling the momentum, or other more efficient ways to randomly switch energy?

Simulation

- From the conclusions before, the key properties for a numerical integrator to simulate the dynamics are volume preservation and conservation of Hamiltonian.
- The change in Hamiltonian can be corrected by an Metropolis-Hastings test.
- The Euler integrator is not volume-preserving, so is not a proper one. The leap-frog integrator is volume-preserving.

SG-MCMCs with Stochastic Dynamics

Following (Ma et al., 2015)

Stochastic dynamics and its density evolution

- Stochastic gradient MCMC methods (SG-MCMCs)
- Commonly, a stochastic dynamics is a continuous Markov process, which can generally be described by the stochastic differential equation:

$$\mathrm{d}z = f(z)\mathrm{d}t + \sqrt{2D(z)}\mathrm{d}W(t), \tag{1}$$

where $f(z)$ denotes the deterministic drift, $W(t)$ is the Wiener process, and $D(z)$ is a positive semidefinite diffusion matrix. The wiener process $W(t)$ is the standard Brownian motion: a stochastic process satisfying 1) $W(0) = 0$ with probability 1; 2) $W(t+h) - W(t) \sim \mathcal{N}(0, h)$ and is independent of $W(\tau)$ for $\tau \leq t$. Thus $\mathrm{d}W(t)$ is usually written as $\mathcal{N}(0, \mathrm{d}t)$ informally and $\sqrt{2D(z)}\mathrm{d}W(t)$ as $\mathcal{N}(0, 2D(z)\mathrm{d}t)$.

- The evolution of $\pi(z,t)$ under the dynamics is given by the Fokker-Planck Equation (FPE)

$$\frac{\partial}{\partial t}\pi(z,t) = -\nabla \cdot \big(\pi(z,t)f(z)\big) + \nabla\nabla : \big(\pi(z,t)D(z)\big),$$

where $\nabla\nabla : \big(\pi(z,t)D(z)\big) = \sum_{i,j} \frac{\partial^2}{\partial z_i \partial z_j}\big(\pi(z,t)D_{ij}(z)\big)$. For intuitive derivations, see e.g. `http://aforrester.bol.ucla.edu/educate/Research/Derive_FokkerPlanck.pdf` or `https://me.ucsb.edu/~moehlis/moehlis_papers/appendix.pdf`.

Framework for the dynamics with desired stationary distribution

- If our target distribution is in the form $\pi(z) \propto \big(-H(z)\big)$, which should be the stationary distribution of the dynamics, then what the dynamics looks like?
  The answer is, for the stochastic dynamics of Eqn. (1),
  $\pi(z) \propto \exp\big(-H(z)\big)$ is its stationary distribution if it satisfies

$$f(z) = -\big(D(z) + Q(z)\big)\nabla H(z) + \Gamma(z), \Gamma_i(z) = \sum_j \frac{\partial}{\partial z_j}\big(D_{ij}(z) + Q_{ij}(z)\big), \tag{2}$$

  where Q is some skew-symmetric curl matrix. Additionally, if $D(z)$ is positive definite, or if ergodicity can be shown, then this stationary distribution is unique.

Proof: We only need to show that for $\pi(z, t=0) \propto \exp\big(-H(z)\big)$, $\frac{\partial}{\partial t}\pi(z,t) = 0$ for $t > 0$. This can be done by FPE. Denote for simplicity $\frac{\partial}{\partial z_i}$ as ∂_i, then at $t = 0$,

$$
\begin{aligned}
\frac{\partial \pi}{\partial t} &\overset{\text{FPE}}{=\!=} -\sum_i \partial_i \big(\pi(z,0) f_i(z)\big) + \sum_{i,j} \partial_i \partial_j \big(\pi(z,0) D_{ij}(z)\big) \\
&\overset{\text{Eqn. (2)}}{=\!=} \sum_i \partial_i \Big[\pi \sum_j (D+Q)_{ij} \partial_j H\Big] - \sum_i \partial_i \Big[\pi \sum_j \partial_j (D+Q)_{ij}\Big] + \sum_{i,j} \partial_i \partial_j (\pi D_{ij}) \\
&\overset{\partial_i \pi = -\pi \partial_i H}{=\!=} -\sum_{i,j} \partial_i \big[(D+Q)_{ij} \partial_j \pi\big] - \sum_{i,j} \partial_i \big[\pi \partial_j (D+Q)_{ij}\big] + \sum_{i,j} \partial_i \partial_j (\pi D_{ij}) \\
&\overset{\text{calculus}}{=\!=} -\sum_{i,j} \partial_i \partial_j \big[\pi (D+Q)_{ij}\big] + \sum_{i,j} \partial_i \partial_j (\pi D_{ij}) \\
&\overset{Q^\top = -Q}{=\!=} -\sum_{i,j} \partial_i \partial_j (\pi D_{ij}) + \sum_{i,j} \partial_i \partial_j (\pi D_{ij}) \\
&= 0.
\end{aligned}
$$

Framework for the dynamics with desired stationary distribution

- Completeness of the framework: For the SDE of Eqn. (1), suppose its stationary distribution $\pi(z)$ uniquely exists, and that $\left[\pi(z)f_i(z) - \sum_j \partial_j \left(\pi(z)D_{ij}(z)\right)\right]$ is integrable with respect to the Lebesgue measure, then there exists a skew-symmetric $Q(z)$ such that Eqn. (2) holds.
- Any continuous Markov process with stationary distribution $\pi(z)$ can be written as in Eqn. (1), which gives us $D(z)$. If the continuous Markov process further meets the conditions above, then the curl matrix $Q(z)$ can be constructed.

How to use stochastic gradient to simulate?

- A Bayesian model with random variable q: prior $\pi(q)$, likelihood $\pi(x_d|q)$, i.i.d. data $\mathcal{D} = \{x_d\}_{d=1}^D$. Posterior $\pi(q|\mathcal{D}) \propto \pi(q)\prod_{d=1}^D \pi(x_d|q)$. For SG-MCMCs, $z = (q, r)$ with some augmenting variable r and the Hamiltonian is usually in the form $H(z) = T(z) + U(q)$. To make the marginalized stationary distribution $\int \exp\{-H(z)\}\mathrm{d}r$ as our target $\pi(q|\mathcal{D})$, define $U(q) \triangleq -\log \pi(q|\mathcal{D}) = -\log \pi(q) - \sum_{d=1}^D \log \pi(x_d|q)$ and require $\int \exp\{-T(z)\}\mathrm{d}r = \text{const}$. Samples from $\exp\{-H(z)\}$ can be drawn by simulating the dynamics.
- Simulate the dynamics by $\mathrm{d}z = f(z)\mathrm{d}t + \mathcal{N}(0, 2D(z)\mathrm{d}t)$: $f(z)$ uses exact gradient.
- Simulate the dynamics by $\mathrm{d}z = \tilde{f}(z)\mathrm{d}t + \mathcal{N}(0, 2D(z)\mathrm{d}t - B(z)\mathrm{d}t^2)$: $\tilde{f}(z)$ uses stochastic gradient: with a randomly selected subset $\mathcal{S}$, $\nabla_q \tilde{U}(q) = \nabla_q \log \pi(q) - \frac{D}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} \nabla_q \log \pi(x|q) \overset{\text{CLT}}{=\!=} \nabla_q \tilde{U}(q) + \mathcal{N}(0, V(q))$, and

$$\tilde{f}(z) = f(z) + \mathcal{N}(0, B(z)). \tag{3}$$

D-MCMCs viewed in the framework

HMC : $z = (q, p)$, $H(z) = \frac{1}{2} p^\top M^{-1} p + U(q)$, $D(z) = 0$,
$Q(z) = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$. Dynamics: $\begin{cases} \mathrm{d}q &= M^{-1} p \mathrm{d}t \\ \mathrm{d}p &= -\nabla U(q) \mathrm{d}t \end{cases}$

SGHMC : $z = (q, p)$, $H(z) = \frac{1}{2} p^\top M^{-1} p + U(q)$, $D(z) = \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$,
$Q(z) = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$. Dynamics:
$\begin{cases} \mathrm{d}q &= M^{-1} p \mathrm{d}t \\ \mathrm{d}p &= -\nabla \tilde{U}(q) \mathrm{d}t - C M^{-1} p \mathrm{d}t + \mathcal{N}(0, 2C\mathrm{d}t - B\mathrm{d}t^2) \end{cases}$

SGLD : $z = q$, $H(z) = U(q)$, $D(z) = D$, $Q(z) = 0$. Dynamics:
$\mathrm{d}q = -D\nabla \tilde{U}(q)\mathrm{d}t + \mathcal{N}(0, 2D\mathrm{d}t - B\mathrm{d}t^2)$.

SGRLD : $z = q$, $H(z) = U(q)$, $D(z) = G(q)^{-1}$, $Q(z) = 0$. Dynamics:
$\mathrm{d}q = -G(q)^{-1}\nabla \tilde{U}(q)\mathrm{d}t - \Gamma(q)\mathrm{d}t + \mathcal{N}(0, 2G(q)^{-1}\mathrm{d}t - B\mathrm{d}t^2)$,
where $\Gamma_i(q) = \sum_j \partial_j D_{ij}(q)$.

D-MCMCs viewed in the framework

SGNHT : $z = (q, p, \xi)$, $H(z) = \frac{1}{2}p^\top p + U(q) + \frac{1}{2d}(\xi - A)^2$,

$$D(z) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & AI & 0 \\ 0 & 0 & 0 \end{pmatrix}, Q(z) = \begin{pmatrix} 0 & -I & 0 \\ I & 0 & p/d \\ 0 & -p^\top/d & 0 \end{pmatrix}, \text{ where}$$

$q, p \in \mathbb{R}^d, \xi, A \in \mathbb{R}$. Dynamics:

$$\begin{cases} \mathrm{d}q &= p\mathrm{d}t \\ \mathrm{d}p &= -\nabla\tilde{U}(q)\mathrm{d}t - \xi p\mathrm{d}t + \mathcal{N}(0, 2A\mathrm{d}t - B\mathrm{d}t^2) \\ \mathrm{d}\xi &= \left(p^\top p/d - 1\right)\mathrm{d}t \end{cases}$$

SGRHMC : $z = (q, p)$, $H(z) = \frac{1}{2}p^\top p + U(q)$, $D(z) = \begin{pmatrix} 0 & 0 \\ 0 & G(q)^{-1} \end{pmatrix}$,

$Q(z) = \begin{pmatrix} 0 & -G(q)^{-1/2} \\ G(q)^{-1/2} & 0 \end{pmatrix}$. Dynamics:

$$\begin{cases} \mathrm{d}q = & G(q)^{-1/2}p\mathrm{d}t \\ \mathrm{d}p = & -G(q)^{-1/2}\nabla\tilde{U}(q)\mathrm{d}t - \Gamma(q)\mathrm{d}t \\ & + G(q)^{-1}p\mathrm{d}t + \mathcal{N}(0, 2G(q)^{-1} - B\mathrm{d}t^2) \end{cases},$$

where $\Gamma_i(q) = \sum_j \frac{\partial}{\partial q_j}(G(q)^{-1/2})_{ij}$.

Convergence Properties for Integrators of SG-MCMCs

Following (Chen et al., 2015)

Numerical integrators (simulation methods) for a stochastic dynamics

- Closed form solution $\pi\big(z(t)\big), z(t) \in \mathbb{R}^n$ for a stochastic dynamics of Eqn. (1) $\mathrm{d}z = f(z)\mathrm{d}t + \sqrt{2D(z)}\mathrm{d}W(t)$ is almost always intractable. To simulate, a numerical integrator has to be designed.
- Two approximations that a numerical integrator has to make for tractability and scalability where simulation error may come from: (Appr1) local generator approximation, and (Appr2) stochastic gradient approximation.

Setup for describing an integrator

- Idea of weak convergence analysis: focusing on the effect of the sample path, instead of on the sample path itself.
- An (infinitesimal) *generator* $\mathcal{L}$ is used for analysis. It is defined for any compactly supported twice differentiable function $\phi : \mathbb{R}^n \to \mathbb{R}$ as

$$\mathcal{L}\phi(z(t)) \triangleq \lim_{h \to 0^+} \frac{\mathbb{E}[\phi(z(t+h))] - \phi(z(t))}{h}$$
$$\underline{\underline{\text{Eqn. (1)}}} \Big(f(z(t)) \cdot \nabla + 2D(z) : \nabla\nabla \Big) \phi(z(t)), \tag{4}$$

or in matrix representation, $f^\top \nabla\phi + \mathrm{tr}\big(D(z)(\nabla\nabla^\top \phi)\big)$. One key property is, for an **e**xact sample $z^e(T)$ at a fixed time T,

$$\mathbb{E}[\phi(z^e(T))] = e^{T\mathcal{L}}\phi(z(0)).$$

(Informal intuition: $\mathcal{L}\phi(z) = \mathrm{d}\mathbb{E}[\phi(z)]/\mathrm{d}t, \mathcal{L}\mathrm{d}t = \mathrm{d}\mathbb{E}[\phi(z)]/\phi(z)$, $\mathcal{L}\mathrm{d}t = \mathrm{d}\log\mathbb{E}[\phi(z)]$.) The Kolmogorov operator $e^{t\mathcal{L}}$ is an exact operator for the evolution of $\mathbb{E}[\phi(z(t))]$ for all t (global).

Setup for describing an integrator

To practically simulate the dynamics: the analytical form of $e^{t\mathcal{L}}$ is almost always unknown, so we need Appr1; for scalability, we need Appr2. So

$$\begin{aligned}\mathbb{E}[\phi(z^e(T))] &= e^{T\mathcal{L}}\phi(z(0)) = e^{h\mathcal{L}} \circ \cdots \circ e^{h\mathcal{L}}\phi(z(0)) \\ &\overset{\text{Appr2}}{\approx} e^{h\tilde{\mathcal{L}}_L} \circ \cdots \circ e^{h\tilde{\mathcal{L}}_1}\phi(z(0)) \\ &\overset{\text{Appr1}}{\approx} \tilde{P}_h^L \circ \cdots \circ \tilde{P}_h^1\phi(z(0)) = \mathbb{E}[\phi(\tilde{z}^n(T))],\end{aligned}$$

where time interval T is split into L pieces with each of size h, and:

- a set of local generators $\{\tilde{\mathcal{L}}_l\}_{l=1}^L$ are introduced that each uses the stochastic gradient based on the minibatch selected for step l; Denote $\tilde{\mathcal{L}}_l = \mathcal{L} + \Delta V_l$, where ΔV_l is their difference and is $(\nabla_q \tilde{U}_l - \nabla_q U) \cdot \nabla_p$ for SGHMC.
- a set of tractable approximate local operators $\{\tilde{P}_h^l\}_{l=1}^L$ are introduced that each uses stochastic gradient and $\tilde{P}_h^l \approx e^{h\tilde{\mathcal{L}}_l}$; $\tilde{P}_h^l$ corresponds to a **Kth-order** local integrator of $\tilde{\mathcal{L}}_l$ if $\tilde{P}_h^l\phi(z) = e^{h\tilde{\mathcal{L}}_l}\phi(z) + O(h^{K+1})$.
- a **n**umerical sample path $\{\tilde{z}^n(lh)\}_{l=1}^L$ can be drawn by successively applying $\{\tilde{P}_h^l\}_{l=1}^L$ on current state: $\mathbb{E}[\phi(\tilde{z}^n(lh))] = \tilde{P}_h^l\phi(\tilde{z}^n((l-1)h))$.

Convergence Analysis

- Finite-time error analysis
- Stationary invariant measures (asymptotic sample distribution)
- Convergence analysis with decreasing step sizes

The gradient noise (embodied as the matrix $B(z)$ in Eqn. (3)) is ignored (set $B(z) = 0$) in the simulating dynamics, since in practice $B(z)$ is almost always unknown and the noise is embodied in ΔV_l in our analysis here.

Finite-time error analysis

- Finite-time error analysis
  In practice we are only interested in the effect of samples: how the sample average $\hat{\phi} \triangleq \frac{1}{L}\sum_{l=1}^{L}\phi(\tilde{z}^n(lh))$ is close to the true expectation $\bar{\phi} \triangleq \int \phi(z)\pi(z)\mathrm{d}z$. For this, we have the following theorems:

Theorem (Bias bound)

Under Assumption 1 in (Chen et al., 2015), the bias of an SG-MCMC with a Kth-order integrator at time $T = hL$ can be bounded as:

$$\left|\mathbb{E}\hat{\phi} - \bar{\phi}\right| = O\left(\frac{1}{Lh} + \frac{\sum_l \|\mathbb{E}\Delta V_l\|}{L} + h^K\right).$$

Theorem (MSE bound)

Under Assumption 1 in (Chen et al., 2015) and the requirement of smoothness of ϕ, the MSE of an SG-MCMC with a Kth-order integrator at time $T = hL$ is bounded, for some $C > 0$ independent of (L, h), as

$$\mathbb{E}\left(\hat{\phi} - \bar{\phi}\right)^2 \leq C\left(\frac{\frac{1}{L}\sum_l \mathbb{E}\|\Delta V_l\|^2}{L} + \frac{1}{Lh} + h^{2K}\right).$$

Finite-time error analysis

- Finite-time error analysis
  From the theorems, we have:
  - $|\mathbb{E}\hat{\phi} - \bar{\phi}|$ might diverge when $\sum_l \|\mathbb{E}\Delta V_l\|$ grows faster than $O(L)$. But for most SG-MCMCs $\sum_l \|\mathbb{E}\Delta V_l\|$ is assumed to vanish since the stochastic gradient is regarded as unbiased.
  - In case where $\sum_l \|\mathbb{E}\Delta V_l\|$ vanishes, an optimal bound for the bias $O(L^{-K/(K+1)})$ is achieved when $h \propto L^{-1/(K+1)}$, and for the MSE $O(L^{-2K/(2K+1)})$ when $h \propto L^{-1/(2K+1)}$. So a higher order integrater basically performs better.

Stationary invariant measures

- Asymptotic properties: for the two bounding theorems above, fix h and let $L \to \infty$:
  Empirical expectation: $\mathbb{E}\hat{\phi} = \bar{\phi} + O(h^K)$, empirical variance:
  $\mathbb{E}(\hat{\phi} - \mathbb{E}\hat{\phi})^2 = O(h^{2K})$
- How the stationary distribution of the numeric simulation dynamics is close to the stationary distribution of the exact dynamics? Define the distance between two measures π_1 and π_2 as
  $d(\pi_1, \pi_2) \triangleq \sup_\phi \left| \mathbb{E}_{\pi_1}[\phi(z)] - \mathbb{E}_{\pi_2}[\phi(z)] \right|$.

Theorem (Error of asymptotic sample distribution)

Assume that a Kth-order integrator is geometric ergodic and its invariant measures exist, then any invariant measure $\tilde{\pi}_h$ is close to the invariant measure π of the exact dynamics with an error up to $O(h^K)$, i.e. $d(\tilde{\pi}, \pi) \leq Ch^K$ for some $C \geq 0$.

Only orders of numerical approximations (Appr1) but not the stochastic gradient approximation (Appr2) affect the asymptotic invariant measure of an SG-MCMC algorithm.

Convergence analysis with decreasing step sizes

Assumption 2

The step sizes $\{h_l\}$ are decreasing, and 1) $\sum_{l=1}^{\infty} h_l = \infty$; and 2) $\lim_{L\to\infty} \frac{\sum_{l=1}^{L} h_l^{K+1}}{\sum_{l=1}^{L} h_l} = 0$.

Denote the finite sum of step sizes as $S_L \triangleq \sum_{l=1}^{L} h_l$ and modify the sample average as $\tilde{\phi} \triangleq \frac{1}{S_L} \sum_{l=1}^{L} h_l \phi\big(\tilde{z}^n(S_l)\big)$. Assume $\mathbb{E}\Delta V_l = 0$. Then we have the following theorem:

Convergence analysis with decreasing step sizes

Theorem (Bias bound and MSE bound for decreasing step sizes)

Under Assumption 1 in (Chen et al., 2015) and Assumption 2, for a smooth test function ϕ, the bias and MSE of a decreasing-step-size SG-MCMC with a Kth-order integrator at time S_L are bounded as:

$$\textit{BIAS:} \left|\mathbb{E}\tilde{\phi} - \bar{\phi}\right| = O\left(\frac{1}{S_L} + \frac{\sum_{l=1}^{L} h_l^{K+1}}{S_L}\right)$$

$$\textit{MSE:}\ \mathbb{E}\left(\tilde{\phi} - \bar{\phi}\right)^2 \leq C\left(\frac{1}{S_L^2}\sum_{l=1}^{L} h_l^2 \mathbb{E}\left[\|\Delta V_l\|^2\right] + \frac{1}{S_L} + \frac{(\sum_{l=1}^{L} h_l^{K+1})^2}{S_L^2}\right)$$

Thus $\lim_{L\to\infty}\left|\mathbb{E}\tilde{\phi} - \bar{\phi}\right| = 0$, *and if* $\lim_{L\to\infty}\frac{1}{S_L^2}\sum_{l=1}^{L} h_l^2 = 0$ *(which the assumption* $\sum_{l=1}^{\infty} h_l^2 < \infty$ *satisfies),* $\lim_{L\to\infty}\mathbb{E}\left(\tilde{\phi} - \bar{\phi}\right)^2 = 0.$

Convergence analysis with decreasing step sizes

Corollary

Using the step size sequences $h_l \propto l^{-\alpha}$ for $0 < \alpha < 1$, all the assumptions in the above theorem are satisfied thus $\tilde{\phi}$ is asymptotically consistent with $\bar{\phi}$.

Remark

For the step size scheme $h_l \propto l^{-\alpha}$, an optimal bound for the bias is achieved when $\alpha = 1/(K+1)$ and for MSE when $\alpha = 1/(2K+1)$, both agree with the optimal step size scheme for fixed-step-size case.

The decreasing step size scheme enjoys the theoretical advantage that the empirical estimation is asymptotically unbiased, but in practice this benefit might not be significant since the exploration over the whole sample space may be inefficient. (Thus SGHMC proposes fixed step size.)

Practical Numerical Integrators

- Euler integrator: $\tilde{z}^n(S_l) = f\big(\tilde{z}^n(S_{l-1})\big)h_l + \mathcal{N}\Big(0, 2D\big(\tilde{z}^n(S_{l-1})\big)h_l\Big)$.
  It is a first-order integrator.
- Symmetric Splitting Integrator (SSI): split the local generator $\tilde{\mathcal{L}}_l$ into several sub-generators that can be solved analytically.
  Example of SSI for SGHMC:

$$\begin{pmatrix} \mathrm{d}q \\ \mathrm{d}p \end{pmatrix} = \begin{pmatrix} M^{-1}p \\ -\nabla_q \tilde{U}_l(q) - CM^{-1}p \end{pmatrix} \mathrm{d}t + \mathcal{N}\left(0, \begin{pmatrix} 0 & 0 \\ 0 & 2C\mathrm{d}t \end{pmatrix}\right),$$

So according to Eqn. (4),

$$\tilde{\mathcal{L}}_l = \begin{pmatrix} (M^{-1}p)\cdot\nabla_q \\ -\big(\nabla_q \tilde{U}_l(q)\big)\cdot\nabla_p - \big(CM^{-1}p\big)\cdot\nabla_p + 2C : \nabla_p\nabla_p \end{pmatrix}.$$

Split $\tilde{\mathcal{L}}_l$ into: $\tilde{\mathcal{L}}_l = \mathcal{L}_A + \mathcal{L}_B + \mathcal{L}_{O_l}$, where

$$\mathcal{L}_A = \begin{pmatrix} (M^{-1}p)\cdot\nabla_q \\ 0 \end{pmatrix}, \ \mathcal{L}_B = \begin{pmatrix} 0 \\ -(CM^{-1}p)\cdot\nabla_p \end{pmatrix},$$
$$\mathcal{L}_{O_l} = \begin{pmatrix} 0 \\ -(\nabla_q \tilde{U}_l(q))\cdot\nabla_p + 2C:\nabla_p\nabla_p \end{pmatrix}.$$

According to Eqn. (4), the corresponding sub-SDEs are

$$A: \begin{cases} \mathrm{d}q = M^{-1}p\mathrm{d}t \\ \mathrm{d}p = 0 \end{cases}, \ B: \begin{cases} \mathrm{d}q = 0 \\ \mathrm{d}p = -CM^{-1}p\mathrm{d}t \end{cases},$$
$$O_l: \begin{cases} \mathrm{d}q = 0 \\ \mathrm{d}p = -\nabla_q \tilde{U}_l(q)\mathrm{d}t + \mathcal{N}(0, 2C\mathrm{d}t) \end{cases}.$$

which can be easily solved in closed form, yielding $e^{t\mathcal{L}_A}$, $e^{t\mathcal{L}_B}$ and $e^{t\mathcal{L}_{O_l}}$. Then construct the approximate local operator as
$\tilde{P}_h^l \triangleq e^{\frac{h}{2}\mathcal{L}_A} \circ e^{\frac{h}{2}\mathcal{L}_B} \circ e^{h\mathcal{L}_{O_l}} \circ e^{\frac{h}{2}\mathcal{L}_B} \circ e^{\frac{h}{2}\mathcal{L}_A}$.
The SSI scheme is a second-order local integrator, i.e. $\tilde{P}_h^l = e^{h\tilde{\mathcal{L}}_l} + O(h^3)$.

References

Neal, Radford M. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2, 2011.

Ma, Yi-An and Chen, Tianqi and Fox, Emily. A Complete Recipe for Stochastic Gradient MCMC. In *Advances in Neural Information Processing Systems*, pp. 2899–2907, 2015.

Chen, Changyou and Ding, Nan and Carin, Lawrence. On the Convergence of Stochastic Gradient MCMC Algorithms with High-Order Integrators. In *Advances in Neural Information Processing Systems*, pp. 2269–2277, 2015.

Thanks!