HMC and SGHMC

-Summary-

Preliminary – HMC

- 2D analogy of HMC (hockey puck without friction) :
  - Let θ = current puck position, r = momentum of the puck, M = mass of the puck
  - A scalar function governing dynamics of the puck :
    - Hamiltonian $H(\theta, r) = U(\theta) + \frac{1}{2} r^T M^{-1} r$

      where $U(\theta)$ is the potential energy of the puck

- Now, we can propose samples (θ, r) from Hamiltonian dynamics:

$$\begin{cases} d\theta = & M^{-1} r \, dt \\ dr = & -\nabla U(\theta) dt \end{cases}$$

Preliminary – HMC

- However, we want to simulate MCMC by Hamiltonian dynamics:
  - Note that $p(\theta|\mathcal{D}) \propto \exp\big(-U(\theta)\big)$, where $U(\theta) = -\sum_{x\in\mathcal{D}} \log p(x|\theta) - \log p(\theta)$
  - **[Fact]** Then, Hamiltonian dynamics $\begin{cases} d\theta = & M^{-1} r\, dt \\ dr = & -\nabla U(\theta) dt \end{cases}$ simulates samples from a joint distribution of (θ, r) defined by $\pi(\theta, r) \propto \exp\left(-U(\theta) - \frac{1}{2} r^T M^{-1} r\right)$,which is a stationary distribution.
  - Since $\pi(\theta, r) \propto \exp\big(-U(\theta)\big) \cdot \exp\left(-\frac{1}{2} r^T M^{-1} r\right)$, by independency, we can take samples $\theta|\mathcal{D}$ from HMC samples (θ, r) by discarding r.

Preliminary – HMC

- Problems & Solutions :

  1. With initial momentum r_0, the $H(\theta, r)$ remains constant ($\because$ potential E + kinetic E remains constant assuming no external force)

  $\Rightarrow$ Resamples the momentum r during HMC iterations

  2. Discretization of continuous dynamics $\begin{cases} d\theta = & M^{-1}r\,dt \\ dr = & -\nabla U(\theta)dt \end{cases}$ to realize HMC:

     $\Rightarrow$ Use 'leap-frog' discretization method with MH step (for stationary distribution guarantee of $\pi(\theta, r)$).

Preliminary – HMC

<MHC algorithm>

Algorithm 1: Hamiltonian Monte Carlo

Input: Starting position $\theta^{(1)}$ and step size ϵ
for $t = 1, 2 \cdots$ **do**
  Resample momentum r
  $r^{(t)} \sim \mathcal{N}(0, M)$
  $(\theta_0, r_0) = (\theta^{(t)}, r^{(t)})$
  Simulate discretization of Hamiltonian dynamics in Eq. (4):
  $r_0 \leftarrow r_0 - \frac{\epsilon}{2}\nabla U(\theta_0)$
  for $i = 1$ ***to*** m **do**
    $\theta_i \leftarrow \theta_{i-1} + \epsilon M^{-1} r_{i-1}$
    $r_i \leftarrow r_{i-1} - \epsilon \nabla U(\theta_i)$
  end
  $r_m \leftarrow r_m - \frac{\epsilon}{2}\nabla U(\theta_m)$
  $(\hat{\theta}, \hat{r}) = (\theta_m, r_m)$
  Metropolis-Hastings correction:
  $u \sim \mathrm{Uniform}[0, 1]$
  $\rho = e^{H(\hat{\theta}, \hat{r}) - H(\theta^{(t)}, r^{(t)})}$
  if $u < \min(1, \rho)$, **then** $\theta^{(t+1)} = \hat{\theta}$
end

Resampling the momentum r

'leap-frog' discretization

MH step for achieving stationary $\pi(\theta, r)$

Algorithm (SGHMC)

- As in SGLD, we want stochastic version of HMC to avoid intractable calculation of $U(\theta) = \sum_{x \in \mathcal{D}} \log p(x|\theta) - \log p(\theta)$, which requires whole iteration of dataset.

- Let $\widetilde{\mathcal{D}}$ be a batch sampled randomly from $\mathcal{D}$, and $\nabla \widetilde{U}(\theta) = -\frac{|\mathcal{D}|}{|\widetilde{\mathcal{D}}|} \sum_{x \in \widetilde{\mathcal{D}}} \nabla \log p(x|\theta) - \nabla \log p(\theta)$ be the unbiased estimate of $\nabla U(\theta)$.

- As the batch size $\left|\widetilde{\mathcal{D}}\right|$ become sufficiently large ($\sim 10^2$ is sufficient in practice), we can use central limit theorem to approximate the noise from approximation of $\nabla U(\theta)$ by $\nabla \widetilde{U}(\theta)$.

Algorithm (SGHMC)

- Thus, $-\epsilon\nabla\widetilde{U}(\theta) = -\epsilon\nabla U(\theta) + N\left(0, \epsilon^2 V(\theta)\right)$, where $V(\theta)$ is the variance from noisy estimate of $\nabla U(\theta)$, and it gives the following continuous SDE:

$$\begin{cases} d\theta = \;\; M^{-1} r \, dt \\ dr = \;\; -\nabla\widetilde{U}(\theta)dt + \boldsymbol{N(0, 2B(\theta)dt)} \end{cases}$$

where $B(\theta) \coloneqq \frac{1}{2}\epsilon V(\theta)$ is the diffusion matrix contributed by gradient noise.

- Analogy in 2D : hockey puck without friction, <u>but with random wind blowing</u>.

- **[Fact]** However, the distribution $\pi(\theta, r)$ is no longer invariant under the above dynamics!

  (It can be verified by showing that $\partial_t H\left(p_t(\theta, r)\right) \geq 0$ under some assumptions)

Algorithm (SGHMC)

- One strategy to add MH step for each iteration, which leads to long simulation runs with low acceptance probabilities.

- Instead, we can minimize the defect of the injected noise from $\nabla\widetilde{U}(\theta)$ by adjusting the dynamics itself : ⇒ Add 'friction' term to the momentum update:

$$\begin{cases} d\theta = & M^{-1}r\,dt \\ dr = & -\nabla\widetilde{U}(\theta)dt - \boldsymbol{BM^{-1}rdt} + N(0{,}2Bdt) \end{cases}$$

where $B = B(\theta)$ can be interpreted as a friction coefficient.

(This dynamical system is commonly referred to as <u>2nd order Langevin dynamics</u> in Physics)

Algorithm (SGHMC)

- **[Fact]** $\pi(\theta, r) \propto \exp\big(-H(\theta, r)\big)$ is the unique stationary (invariant) distributions of the given dynamics (with friction).

  (It can be verified by showing that the distribution evolution $\partial_t p_t(\theta, r) = 0$.)

- Problem and Solution:

$\boldsymbol{\hat{V}} :\cong \frac{|\mathcal{D}|^2}{|\widetilde{\mathcal{D}}|} \cdot \sum_{i=1}^{|\widetilde{\mathcal{D}}|} (\boldsymbol{s_i} - \boldsymbol{\bar{s}})(\boldsymbol{s_i} - \boldsymbol{\bar{s}})^T$***, where*** $\boldsymbol{s_i} = \boldsymbol{\nabla \log p(x_i|\theta)} + \frac{1}{|\mathcal{D}|}\boldsymbol{\nabla \log p(\theta)}$

1. We do not known the exact value of $B = B(\theta)$ (noise from $\nabla\widetilde{U}(\theta)$)

   $\Rightarrow$ Take an estimate $\hat{B}$ of B (ex : $\hat{B} = 0$ or $\frac{1}{2}\epsilon\hat{V}$) and set user-specified friction term $C \succcurlyeq \hat{B}$:

$$\begin{cases} d\theta = & M^{-1} r\, dt \\ dr = & -\nabla\widetilde{U}(\theta)dt - \boldsymbol{CM^{-1}rdt} + N\big(0, 2\big(C - \hat{B}\big)dt\big) + N(0, 2Bdt) \end{cases}$$

This dynamics gives stationary $\boldsymbol{\pi(\theta, r) \propto exp\big(-H(\theta, r)\big)}$ if $\boldsymbol{\hat{B} = B}$

Algorithm (SGHMC)

- Take an estimate $\hat{B}$ of B (ex : $\hat{B} = 0$ or $\frac{1}{2}\epsilon\hat{V}$) and set user-specified friction term $C \succcurlyeq \hat{B}$:

$$\begin{cases} d\theta = & M^{-1} r \, dt \\ dr = & -\nabla\tilde{U}(\theta) dt - \boldsymbol{CM^{-1} r dt} + N\big(0, 2\big(C - \hat{B}\big) dt\big) + N(0, 2B dt) \end{cases}$$

Algorithm 2: Stochastic Gradient HMC

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

optionally, resample momentum r *as*

$r^{(t)} \sim \mathcal{N}(0, M)$

$(\theta_0, r_0) = (\theta^{(t)}, r^{(t)})$

simulate dynamics in Eq.(13):

for $i = 1$ ***to*** m **do**

$\theta_i \leftarrow \theta_{i-1} + \epsilon_t M^{-1} r_{i-1}$

$r_i \leftarrow r_{i-1} - \epsilon_t \nabla\tilde{U}(\theta_i) - \epsilon_t C M^{-1} r_{i-1} + \mathcal{N}(0, 2(C - \hat{B})\epsilon_t)$

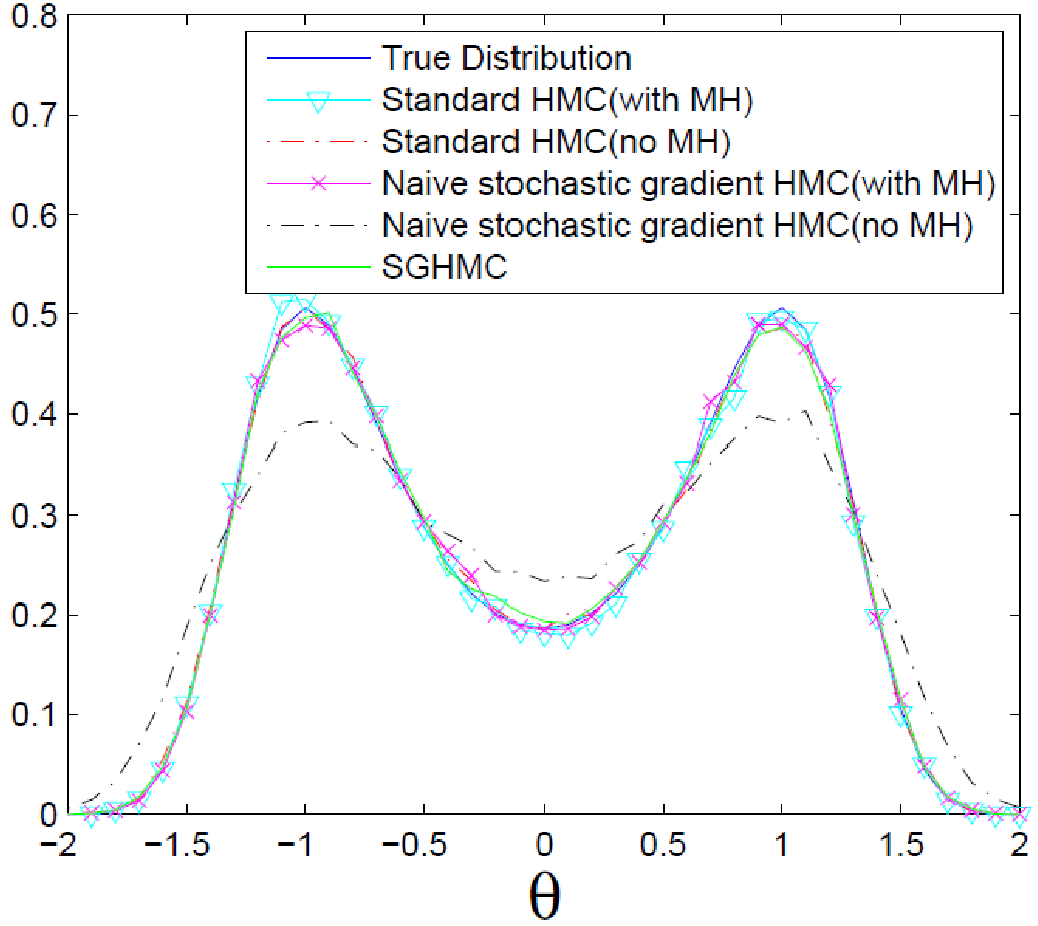
end

$(\theta^{(t+1)}, r^{(t+1)}) = (\theta_m, r_m)$, no M-H step

end

Experiments (SGHMC)

- Empirical distributions associated with various sampling algorithms
  - Target distribution with $U(\theta) = -2\theta^2 + \theta^4$ ($\Leftrightarrow p(\theta|\mathcal{D}) \propto \exp\left(2\theta^2 - \theta^4\right)$)



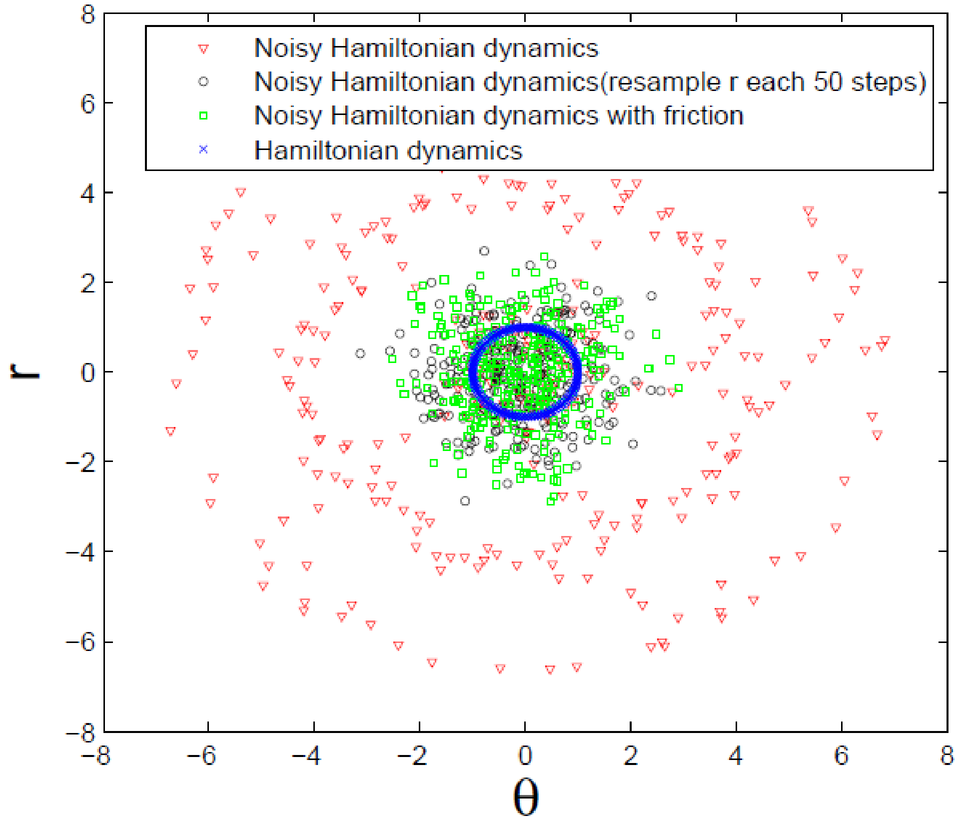
Note:

- **Naïve HMC (w/o MH) fails to achieve target distribution.**
- **Standard HMC achieves target distribution regardless of MH step (as the theory suggested).**

Experiments (SGHMC)

- Points (θ, r) simulated from discretizations of various Hamiltonian dynamics

  using $U(\theta) = \frac{1}{2}\theta^2$, and replace gradient by $\nabla\tilde{U}(\theta) = \theta + N(0,4)$



Note:

- **Target distribution $p(\theta|\mathcal{D}) \propto exp\left(-\frac{1}{2}\theta^2\right)$**
- **Noisy HMC w/o friction has divergent samples (red)**
- **Resampling r helps control divergence, but associated HMC stationary distribution is not correct (as before)**
- **Noisy HMC w/ friction achieves samples similar to those from HMC**