Riemann manifold Langevin and Hamiltonian Monte Carlo methods

Clément Chadebec

ENS - MVA

January 9, 2020

Overview

Shortcomings of Monte Carlo algorithms

- Not scalable for target densities in high dimension
- Can demonstrate high correlations
- Can demonstrate low acceptation rates
  $\Longrightarrow$ Need for new methods

HMC

<u>Goal</u>: Simulate a random variable $\theta \in \mathbb{R}^D \sim \pi$ a target density.

- Introduction of an **independent** auxiliary variable $\mathbf{Y} \in \mathbb{R}^D \sim \nu = \mathcal{N}(0, M)$ where M is called the mass matrix
- The negative log-proba of the joint distribution follows:

$$H(\theta, \mathbf{Y}) = - \underbrace{\mathcal{L}(\theta)}_{\text{energy function}} + \frac{1}{2} \log[(2\pi)^D |M|] + \underbrace{\frac{1}{2} \mathbf{Y}^\top M^{-1} \mathbf{Y}}_{\text{kinetic energy}}$$

- The derivatives of H give

$$\frac{d\theta}{d\tau} = \frac{\partial H}{\partial \mathbf{Y}} = \mathbf{M}^{-\mathbf{1}} \mathbf{Y}$$
$$\frac{d\mathbf{Y}}{d\tau} = \frac{\partial H}{\partial \theta} = \nabla_\theta \mathcal{L}(\mathbf{X})$$

HMC

- Stormer - Verlet (leapfrog) integrator

$$\begin{aligned}\mathbf{Y}(\tau+\varepsilon/2) &= \mathbf{Y}(\tau)+\varepsilon\nabla_\theta\mathcal{L}(\theta)/2 \\ \theta(\tau+\varepsilon) &= \theta(\tau)+\varepsilon M^{-1}\mathbf{Y}(\tau+\varepsilon/2) \\ \mathbf{Y}(\tau+\varepsilon) &= \mathbf{Y}(\tau+\varepsilon/2)+\varepsilon\nabla_\theta\mathcal{L}(\theta(\tau+\varepsilon))/2\end{aligned} \qquad \text{(Stormer - Verlet)}$$

- HMC sampling of $\pi(\theta)$ as a Gibbs sampler:

$$\begin{aligned}\mathbf{Y}^{n+1}|\theta^n \sim \mathbf{Y}^{n+1} &\sim \mathcal{N}(0, M) \\ \theta^{n+1}|\mathbf{Y}^{n+1} &\sim \mu(\theta^{n+1}|\mathbf{Y}^{n+1})\end{aligned}$$

- $\mu(\theta^{n+1}|\mathbf{Y}^{n+1})$ simulated using Stormer - Verlet scheme and $(\tilde{\theta}, \tilde{\mathbf{Y}})$ is accepted with with probability $\min\{1, \exp(-H(\tilde{\theta}, \tilde{\mathbf{Y}}) + H(\theta^n, \mathbf{Y}^{n+1}))\}$
- $\Longrightarrow$ produces an ergodic, time reversible Markov Chain with stationary density π
- $\Longrightarrow$ Difficulty to select M regardless of the target density

Leapfrog Impact

Sampling of $\mathcal{N}(\mathbf{5}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.8 \end{pmatrix}$

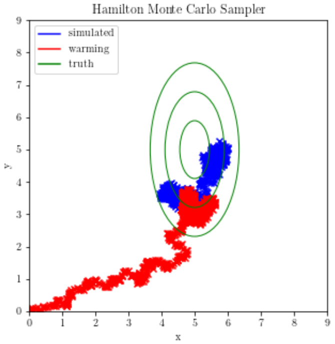

Figure: Leapfrog steps $= 2$, $\varepsilon = 0.01$

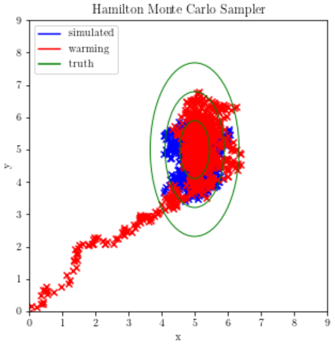

Figure: Leapfrog steps $= 5$, $\varepsilon = 0.01$

Leapfrog Impact

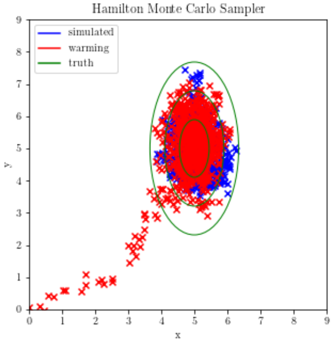

Figure: Leapfrog steps $= 10$, $\varepsilon = 0.01$

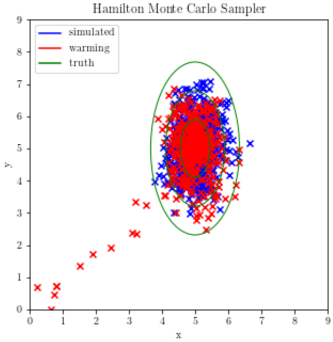

Figure: Leapfrog steps $= 20$, $\varepsilon = 0.01$

Acceptance Ratio

- Dimensions ranging from $D = 1$ to 50
- Sampling of $\mathcal{N}(\mathbf{5}, \mathbf{I})$

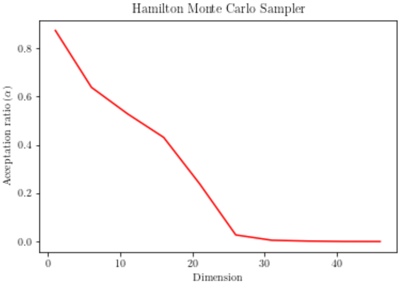

Figure: Leapfrog steps = 20, $\varepsilon = 0.01$

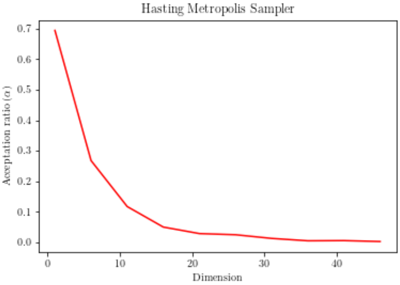

Figure: Sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$

Auto-correlation

- Sampling of $\mathcal{N}(\mathbf{5}, \mathbf{I})$
- Leapfrog steps $= 20$, $\varepsilon = 0.01$ (HMC)
- Sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (HM)

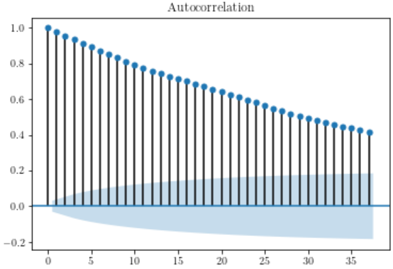

Figure: HMC ($D = 2$)

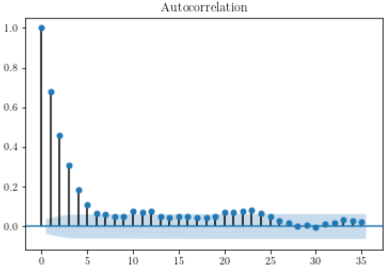

Figure: HM ($D = 2$)

Auto-correlation

- Sampling of $\mathcal{N}(\mathbf{5}, \mathbf{I})$
- Leapfrog steps $= 20$, $\varepsilon = 0.01$ (HMC)
- Sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (HM)

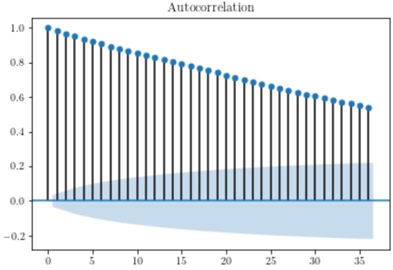

Figure: HMC ($D = 5$)

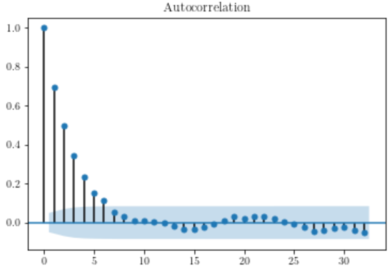

Figure: HM ($D = 5$)

Auto-correlation

- Sampling of $\mathcal{N}(\mathbf{5}, \mathbf{I})$
- Leapfrog steps $= 20$, $\varepsilon = 0.01$ (HMC)
- Sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (HM)

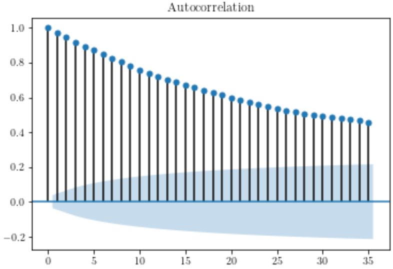

Figure: HMC ($D = 10$)

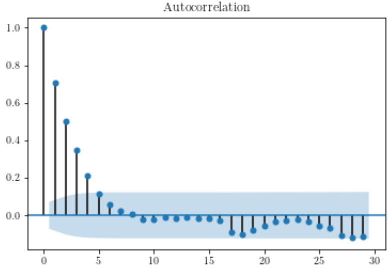

Figure: HM ($D = 10$)

Auto-correlation

- Sampling of $\mathcal{N}(\mathbf{5}, \mathbf{I})$
- Leapfrog steps $= 20$, $\varepsilon = 0.01$ (HMC)
- Sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (HM)

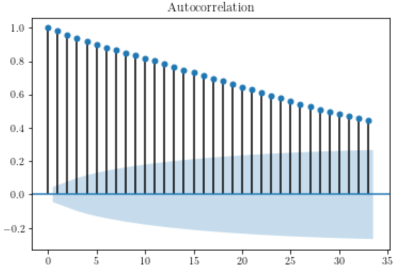

Figure: HMC ($D = 20$)

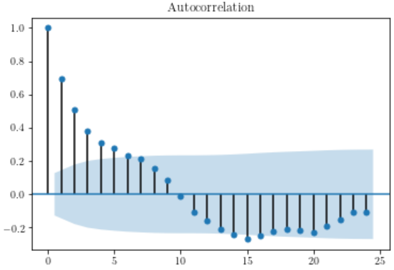

Figure: HM ($D = 20$)

Influence of M

- Leapfrog steps $= 20$, $\varepsilon = 0.01$
- warm start $= 5000$

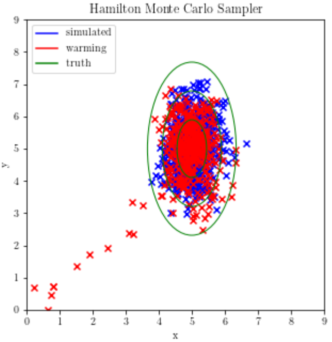

Figure: $M = \mathbf{I}$

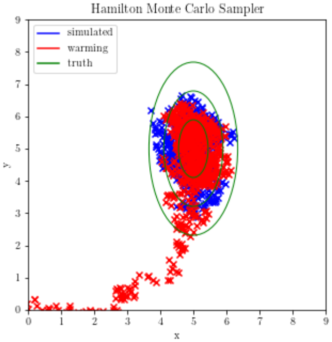

Figure: $M = 10\mathbf{I}$

Influence of M

- Leapfrog steps $= 20$, $\varepsilon = 0.01$
- warm start $= 5000$

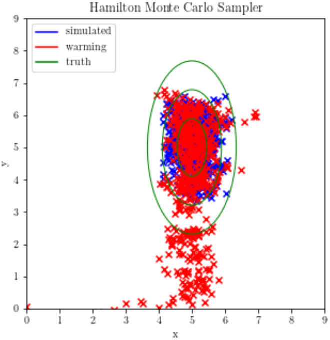

Figure: $M = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix}$.

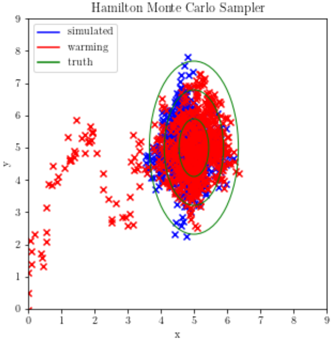

Figure: $M = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}$.

Riemann Manifold HMC

- Distance between two density functions $\Longrightarrow p(\mathbf{X};\theta)$ and $p(\mathbf{X};\theta+\delta\theta) = \delta\theta^\top \mathbf{G}(\theta)\delta\theta$ where $G(\theta)$ is the expected Fisher information matrix

$$\mathbf{G}(\theta) = -\mathbb{E}_{\mathbf{X}|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log(p(\mathbf{X}|\theta))\right] = \mathrm{cov}\left[\frac{\partial}{\partial\theta}\log(p(\mathbf{X}|\theta))\right]$$

$\Longrightarrow$ position specific metric on a Riemann manifold

- The Hamiltonian follows

$$H(\theta,\mathbf{Y}) = -\underbrace{\mathcal{L}(\theta)}_{\text{energy function}} + \frac{1}{2}\log[(2\pi)^D|G(\theta)|] + \underbrace{\frac{1}{2}\mathbf{Y}^\top\mathbf{G}(\theta)^{-\mathbf{1}}\mathbf{Y}}_{\text{kinetic energy}}$$

- Baysian approach $\Longrightarrow \mathbf{G}(\theta) = -\mathbb{E}_{\mathbf{X}|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log(p(\mathbf{X},\theta))\right]$

Riemann Manifold HMC

- Again the Riemann HMC sampling of $\pi(\theta)$ can be seen as a Gibbs sampler:
$$\mathbf{Y}^{n+1}|\theta^n \sim \mathcal{N}(0, \mathbf{G}(\theta^n))$$
$$\theta^{n+1}|\mathbf{Y}^{n+1} \sim \mu(\theta^{n+1}|\mathbf{Y}^{n+1})$$
- $\mu(\theta^{n+1}|\mathbf{Y}^{n+1})$ simulated using generalized Stormer - Verlet scheme and $(\tilde{\theta}, \tilde{\mathbf{Y}})$ is accepted with with probability $\min\{1, \exp(-H(\tilde{\theta}, \tilde{\mathbf{Y}}) + H(\theta^n, \mathbf{Y}^{n+1}))\}$
- $\Longrightarrow$ produces an ergodic, time reversible Markov Chain with stationary density π
- M mass matrix replaced by position specific metric $\mathbf{G}(\theta) \longrightarrow$ no need to tune the M coefficients
$\Longrightarrow$ Need for a new time-reversible numerical integrator for solving the non-separable Hamiltonian
- How to choose the metric $\mathbf{G}$?

Example of Bayesian Logistic Regression

- The model:
  - Let $\mathbf{X} \in \mathbb{R}^{N \times D}$ be the design matrix
  - $\beta \in \mathbb{R}^D$ regression parameter with $\beta \sim \pi = \mathcal{N}(0, \alpha \mathbf{I})$ with α given
  - We look for β such that $\mathbf{Y}_n = s(\mathbf{X}_n^\top \beta^\top)$ where s is the sigmoid function
- The metric tensor follows:

$$
\begin{aligned}
\mathbf{G}(\beta) &= -\mathbb{E}_{\mathbf{Y}|\beta}\Big[\frac{\partial^2}{\partial \beta^2} \log(p(\mathbf{Y}, \beta))\Big] \\
&= \underbrace{\mathbb{E}_{\mathbf{Y}|\beta}\Big[\frac{\partial^2}{\partial \beta^2} \log(p(\mathbf{Y}|\beta))\Big]}_{\text{Fisher-Rao}} - \underbrace{\frac{\partial^2}{\partial \beta^2} \log(\pi(\beta))}_{NegativeHessian} \\
&= \mathbf{X}_n^\top \Lambda \mathbf{X} + \alpha^{-1} \mathbf{I}
\end{aligned}
$$

where Λ is diagonal and $\Lambda_{n,n} = s(\beta^\top \mathbf{X}_n^\top)(1 - s(\beta^\top \mathbf{X}_n^\top))$

Comparison

- Models considered
  - Component-wise adaptive MH
  - Joint updating Gibbs
  - MALA
  - HMC
  - RHMC - Student
  - Iterated weighted least squares

Name	*Covariates* (D)	*Data points* (N)	*Dimension of* $\boldsymbol{\beta}$ (b)
Pima Indian	7	532	8
Australian credit	14	690	15
German credit	24	1000	25
Heart	13	270	14
Ripley	2	250	7

Figure: Dataset

Results

- Criteria: $ESS = N(1 + 2\sum_k \gamma(k))$ on each covariate.
  N: the number of posterior samples
  $\sum_k \gamma(k)$: sum of the K monotone sample auto-correlations.

Method	*Time*	*ESS (min, avg, max)*	*s/min ESS*	*Relative speed*
Metropolis	23.4	(167, 613, 1015)	0.140	13.3
Mala	3.5	(95.5, 316, 667)	0.037	50.3
HMC	117.9	(3182, 3632, 3986)	0.037	50.3
IWLS	7.8	(4.2, 9.9, 69)	1.1862	1
RHMC - S	257.4	(3981, 4934, 5000)	0.065	28.6
RMHMC	**246.6**	**(4757, 5000, 5000)**	**0.052**	**35.8**

Table: Results ($D = 24$, $N = 1000$)

Conclusion

- Strong demonstrated results
- Choice of the metric to be further investigated (Student, ...)
- Choice of the kinetic energy to be further investigated
- What about even bigger dimensions $(100, 1000, ...)$?
  $\Longrightarrow$ Computation cost scaling

Proposition

Proposition: The transition kernel:

$$P(\theta, A) = \int_{\mathcal{Y}} \mathbf{1}(\tilde{\Phi}^N(\theta, y))\alpha((\theta, y, \Phi^N(\theta, y))\nu(y)dy$$
$$+ \mathbf{I}_\theta(A) \int_{\mathcal{Y}} (1 - \alpha((\theta, y, \Phi^N(\theta, y))\nu(y)dy$$

where Φ^N is the outcome of N leapfrog step and $\tilde{\Phi}(\theta, y) = \theta$

Proposition: π is stationary for P

Proof

Proof let f be a Borel function

$$
\begin{aligned}
\mathbb{E}\Big[f(\theta^{n+1})|\mathcal{F}_n\Big] &= \mathbb{E}_{\mathbf{Y}}\Big[\mathbb{E}\Big[f(\theta^{n+1}|\mathcal{F}_n, \mathbf{Y}_{n+1}\Big]\Big] \\
&= \int_{\mathcal{Y}} \mathbb{E}_{(U,\theta)}\Big[f(\tilde{\Phi}^N(\theta, y)\mathbf{I}_{\{U\leq\alpha((\theta,y,\Phi^N(\theta,y))\}} \\
&+ f(\theta^n)\mathbf{I}_{\{U>\alpha((\theta,y,\Phi^N(\theta,y))\}}\Big]\nu(y)dy \\
&= \int_\theta \int_{\mathcal{Y}} f(\tilde{\Phi}^N(\theta, y))\alpha((\theta, y, \Phi^N(\theta, y))\nu(y)dyd\theta \\
&+ \int_\theta \delta_{\theta^n}(f) \int_{\mathcal{Y}} (1 - \alpha((\theta, y, \Phi^N(\theta, y)))\nu(y)dyd\theta
\end{aligned}
$$

Proof

Sketch of proof: We use the balanced equation
"$\pi(\theta)P(\theta', \theta) = \pi(\theta')P(\theta, \theta')$"

- $$\pi(\theta) * \alpha((\theta, y, \theta', y') * \nu(y) = \pi(\theta) * \min\left(1, \frac{\pi(\theta')\nu(y')}{\pi(\theta)\nu(y)}\right) * \nu(y)$$
  $$= \pi(\theta') * \alpha(\theta', y', \theta, y) * \nu(y')$$

- $$\int_{A\times B} \pi(d\theta)P_2(\theta, d\theta') = \int_{A\cap B} \pi(d\theta)h(\theta, \cdot)$$
  $$= \int_{A\cap B} \pi(d\theta')h(\theta, \cdot)$$
  $$= \int_{A\times B} \pi(d\theta')P_2(\theta', d\theta)$$