Robust Monte Carlo Sampling using Riemannian Nosé-Poincaré Hamiltonian Dynamics

Anirban Roychowdhury ROYCHOWDHURY.7@OSU.EDU
Ohio State University, Columbus, OH 43210

Brian Kulis BKULIS@BU.EDU
Boston University, Boston, MA 02215

Srinivasan Parthasarathy SRINI@CSE.OHIO-STATE.EDU
Ohio State University, Columbus, OH 43210

Appendices

This document contains supplementary material for the submission "Robust Monte Carlo Sampling using Riemannian Nosé-Poincaré Hamiltonian Dynamics".

A. Proof of Theorem 1

The Riemann-augmented Nosé-Poincaré Hamiltonian can be written as

$$H(\boldsymbol{\theta},\mathbf{p},s,q) = s\Bigg(-\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2}\left(\frac{\mathbf{p}}{s}\right)^T \mathbf{G}(\boldsymbol{\theta})^{-1}\frac{\mathbf{p}}{s} + \frac{1}{2}|\mathbf{G}(\boldsymbol{\theta})|^{-1}q^2 + \frac{1+kT}{2}\log\left\{(2\pi)^D|\mathbf{G}(\boldsymbol{\theta})|\right\} + gkT\log s - H_0\Bigg). \tag{1}$$

We need to prove the following about this Hamiltonian:

Theorem 1. *The dynamical system derived from the Riemannian Nosé-Poincaré Hamiltonian (1) generates samples from the canonical ensemble.*

Proof. First we denote $H_{gc}(\boldsymbol{\theta},\mathbf{p}) = -\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^T\mathbf{G}(\boldsymbol{\theta})^{-1}\mathbf{p}$.

We will show that we can integrate out s, q from $p(\boldsymbol{\theta},\mathbf{p},s,q) \propto \exp\left(-H(\boldsymbol{\theta},\mathbf{p},s,q)\right)$ to get $p(\boldsymbol{\theta},\mathbf{p}) \propto \exp\left(-H_{gc}(\boldsymbol{\theta},\mathbf{p})/kT\right)$. The integration of s essentially follows (Bond et al., 1999).

The probability of any $(\boldsymbol{\theta},\mathbf{p},q)$ can be written as

$$\begin{aligned} p(\boldsymbol{\theta},\mathbf{p},q) &\propto \int_s \delta[H-H_0] \\ &\propto \int \delta\Bigg[s\Big(H_{gc}\left(\boldsymbol{\theta},\left(\frac{\mathbf{p}}{s}\right)\right) + \frac{1}{2}|\mathbf{G}(\boldsymbol{\theta})|^{-1}q^2 \\ &\quad + \frac{1+kT}{2}\log\left\{(2\pi)^D|\mathbf{G}(\boldsymbol{\theta})|\right\} + gkT\log s - H_0\Big)\Bigg] \\ &\propto \int s^{N_f}\delta\Bigg[s\Big(H_{gc} + \frac{1}{2}|\mathbf{G}(\boldsymbol{\theta})|^{-1}q^2 \\ &\quad + \frac{1+kT}{2}\log\left\{(2\pi)^D|\mathbf{G}(\boldsymbol{\theta})|\right\} + gkT\log s - H_0\Big)\Bigg]. \end{aligned}$$

See (Leimkuhler & Reich, 2004) for details of the last step. The argument of the δ-function has a root at $s_0 = \exp\Bigg[-\frac{1}{gkT}\Bigg(H_{gc} + \frac{1}{2}|\mathbf{G}(\boldsymbol{\theta})|^{-1}q^2 + \frac{kT}{2}\log\left\{(2\pi)^D|\mathbf{G}(\boldsymbol{\theta})|\right\} - H_0\Bigg)\Bigg]$.

We can therefore write $p(\boldsymbol{\theta},\mathbf{p},q)$ as

$$\begin{aligned} p(\boldsymbol{\theta},\mathbf{p},q) &\propto \int \frac{s^{N_f}}{gkT}\delta\Bigg[s - \exp\Bigg(-\frac{1}{gkT}\Bigg(H_{gc} + \frac{1}{2}|\mathbf{G}(\boldsymbol{\theta})|^{-1}q^2 \\ &\quad + \frac{kT}{2}\log\left\{(2\pi)^D|\mathbf{G}(\boldsymbol{\theta})|\right\} - H_0\Bigg)\Bigg)\Bigg] \\ &\propto \exp\left(\frac{N_f H_0}{gkT}\right)\exp\Bigg(-\frac{N_f}{gkT}\Bigg(H_{gc} \\ &\quad + \frac{1}{2}|\mathbf{G}(\boldsymbol{\theta})|^{-1}q^2 + \frac{kT}{2}\log\left\{(2\pi)^D|\mathbf{G}(\boldsymbol{\theta})|\right\}\Bigg)\Bigg) \\ &\propto \exp\Bigg(-\frac{1}{kT}\Bigg(H_{gc} + \frac{1}{2}|\mathbf{G}(\boldsymbol{\theta})|^{-1}q^2 \\ &\quad + \frac{kT}{2}\log\left\{(2\pi)^D|\mathbf{G}(\boldsymbol{\theta})|\right\}\Bigg)\Bigg), \text{ using } g = N_f. \end{aligned}$$

See (Leimkuhler & Reich, 2004) for the details of the first step.

With s integrated out, we are left with

$$p(\boldsymbol{\theta}, \mathbf{p}, q) \propto \exp\left(-H_{gc}/kT\right) \exp\Big[-\frac{1}{2kT}|\mathbf{G}(\boldsymbol{\theta})|^{-1}q^2 - \frac{1}{2}\log\left\{(2\pi)^D|\mathbf{G}(\boldsymbol{\theta})|\right\}\Big].$$

We can easily integrate out q from the second exponential term on the right to get the desired form for $p(\boldsymbol{\theta}, \mathbf{p})$. □

Note that the integration over s works along the energy slice $H = H_0$, where H_0 is the initial value of the potential / Hamiltonian. The desired marginal density over $(\boldsymbol{\theta}, \mathbf{p})$ however is invariant to H_0, since H_0 is delegated to the proportionality constant during the integration over s, and cancels out under normalization.

B. Discretized Dynamics

B.1. Generalized Leapfrog in the Stochastic case

We propose the following dynamics to incorporate stochastic noise correction terms into the deterministic updates:

$$\begin{aligned}
\dot{\boldsymbol{\theta}} &= \left(\frac{\mathbf{p}}{s}\right)^T \mathbf{G}(\boldsymbol{\theta})^{-1} \\
\dot{\mathbf{p}} &= s\frac{1}{2}\left(\frac{\mathbf{p}}{s}\right)^T \mathbf{G}(\boldsymbol{\theta})^{-1}\nabla\mathbf{G}(\boldsymbol{\theta})\mathbf{G}(\boldsymbol{\theta})^{-1}\left(\frac{\mathbf{p}}{s}\right) \\
&\quad + s\frac{1}{2}q^2\mathbf{G}(\boldsymbol{\theta})^{-1}\nabla\mathbf{G}(\boldsymbol{\theta})\mathbf{G}(\boldsymbol{\theta})^{-1} - \sqrt{s}B(\boldsymbol{\theta})\left(\frac{\mathbf{p}}{s}\right)^T \mathbf{G}(\boldsymbol{\theta})^{-1} \\
&\quad - \frac{s}{2}(1+kT)\text{tr}\left(\mathbf{G}(\boldsymbol{\theta})^{-1}\nabla\mathbf{G}(\boldsymbol{\theta})\right) + s\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}) \\
\dot{s} &= sq\mathbf{G}(\boldsymbol{\theta})^{-1} \\
\dot{q} &= -gkT + \left(\frac{\mathbf{p}}{s}\right)^T \mathbf{G}(\boldsymbol{\theta})^{-1}\frac{\mathbf{p}}{s} - \tilde{H}_{\text{inner}} - A(\boldsymbol{\theta})sq\mathbf{G}(\boldsymbol{\theta})^{-1}.
\end{aligned} \tag{2}$$

As mentioned in the main paper, we use the generalized leapfrog algorithm to discretize the continuous differential equations. The generalized leapfrog algorithm is a composition of a symplectic first-order Euler integrator with its adjoint. For the dynamics (2), the update equations can be written as

$$\begin{aligned}
p_i^{(t+\epsilon/2)} &= p_i^{(t)} - \frac{\epsilon}{2}\nabla_{\theta_i}H\left(\boldsymbol{\theta}^{(t)}, s^{(t)}, \mathbf{p}^{(t+\epsilon/2)}, q^{(t+\epsilon/2)}\right) \\
q^{(t+\epsilon/2)} &= q^{(t)} - \frac{\epsilon}{2}\Bigg[As^{(t)}q^{(t)}G(\boldsymbol{\theta}^{(t)})^{-1} \\
&\quad + \tilde{H}_{\text{inner}} + gkT - \left(\frac{\mathbf{p}^{(t+\epsilon/2)}}{s^{(t)}}\right)^T G(\boldsymbol{\theta}^{(t)})^{-1}\left(\frac{\mathbf{p}^{(t+\epsilon/2)}}{s^{(t)}}\right)\Bigg] \\
\theta_i^{(t+\epsilon} &= \theta_i^{(t)} \\
&\quad + \frac{\epsilon}{2}\Bigg[\left(\frac{\mathbf{p}^{(t+\epsilon/2)}}{s^{(t)}}\right)^T G(\boldsymbol{\theta}^{(t)})^{-1} \\
&\quad + \left(\frac{\mathbf{p}^{(t+\epsilon/2)}}{s^{(t+\epsilon)}}\right)^T G(\boldsymbol{\theta}^{(t+\epsilon)})^{-1}\Bigg]_i \\
s^{(t+\epsilon)} &= s^{(t)} \\
&\quad + \frac{\epsilon}{2}\left[s^{(t)}q^{(t+\epsilon/2)}|G(\boldsymbol{\theta}^{(t)})|^{-1} + s^{(t+\epsilon)}q^{(t+\epsilon/2)}|G(\boldsymbol{\theta}^{(t+\epsilon)})|^{-1}\right] \\
p_i^{(t+\epsilon)} &= p_i^{(t+\epsilon/2)} - \frac{\epsilon}{2}\nabla_{\theta}H\left(\boldsymbol{\theta}^{(t+\epsilon)}, s^{(t+\epsilon)}, \mathbf{p}^{(t+\epsilon/2)}, q^{(t+\epsilon/2)}\right) \\
q^{(t+\epsilon)} &= q^{(t+\epsilon/2)} - \frac{\epsilon}{2}\Bigg[As^{(t+\epsilon)}q^{(t+\epsilon/2)}G(\boldsymbol{\theta}^{(t+\epsilon)})^{-1} \\
&\quad + \tilde{H}_{\text{inner}} + gkT - \left(\frac{\mathbf{p}^{(t+\epsilon/2)}}{s^{(t+\epsilon)}}\right)^T G(\boldsymbol{\theta}^{(t+\epsilon)})^{-1}\left(\frac{\mathbf{p}^{(t+\epsilon/2)}}{s^{(t+\epsilon)}}\right)\Bigg]
\end{aligned} \tag{3}$$

where $\nabla_{\theta_i}H(\boldsymbol{\theta}, s, \mathbf{p}, q)$

$$\begin{aligned}
&= \frac{\epsilon}{2}\Bigg[-\frac{1}{2}s\left(\frac{\mathbf{p}}{s}\right)^T G(\boldsymbol{\theta})^{-1}\left(\frac{\partial}{\partial\theta_i}G(\boldsymbol{\theta})\right)G(\boldsymbol{\theta})^{-1}\left(\frac{\mathbf{p}}{s}\right) \\
&\quad - s\frac{q^2}{2}|G(\boldsymbol{\theta})|^{-1}\text{tr}\left\{G(\boldsymbol{\theta})^{-1}\frac{\partial}{\partial\theta_i}G(\boldsymbol{\theta})\right\} + \sqrt{s}B\left(\frac{\mathbf{p}}{s}\right)^T G(\boldsymbol{\theta})^{-1} \\
&\quad + \frac{s}{2}(1+kT)\text{tr}\left\{G(\boldsymbol{\theta})^{-1}\frac{\partial}{\partial\theta_i}G(\boldsymbol{\theta})\right\} - s\frac{\partial}{\partial\theta_i}\tilde{\mathcal{L}}(\boldsymbol{\theta})\Bigg].
\end{aligned}$$

Using the tensor $G(\boldsymbol{\theta}) = \text{diag}(\boldsymbol{\theta})^{-1}$, (assuming $G(\boldsymbol{\theta}) \succ 0$), the implicit system for $\mathbf{p}^{t+\epsilon/2}$ and $q^{t+\epsilon/2}$ reduces to the following quadratic system:

$$\begin{aligned}
&\frac{\epsilon}{4}\frac{p_i^2}{s^{(t)}} + p_i - p_i^{(t)} - \frac{\epsilon}{2}s^{(t)}\Bigg[\frac{1}{\theta_i^{(t)}}\left(1 + kT - |G(\boldsymbol{\theta}^{(t)})|^{-1}\frac{q^2}{2}\right) \\
&\qquad\qquad + \frac{\partial}{\partial\theta_i^{(t)}}\mathcal{L}(\boldsymbol{\theta^{(t)}})\Bigg] = 0
\end{aligned}$$

$$\begin{aligned}
&\frac{\epsilon}{4}|G(\boldsymbol{\theta}^{(t)})|^{-1}q^2 + q - q^{(t)} + \frac{\epsilon}{2}\Bigg[-\mathcal{L}(\boldsymbol{\theta^{(t)}}) \\
&+ gkT(1+\log s) + \frac{1+kT}{2}\log\left\{(2\pi)^D|\mathbf{G}(\boldsymbol{\theta})|\right\} \\
&- \frac{1}{2}\left[\frac{\mathbf{p}}{s^{(t)}}\right]^T G(\boldsymbol{\theta}^{(t)})^{-1}\left[\frac{\mathbf{p}}{s^{(t)}}\right]\Bigg] = 0
\end{aligned}$$

where we have omitted the $t+\epsilon/2$ superscripts for clarity (i.e. $p_i = p_i^{(t+\epsilon/2)}$). The Jacobian in this case is an arrowhead matrix. We should mention here that the Newton iterations can be performed for any choice of metric tensor; we choose the diagonal tensor in our real-data experiments for simplicity of implementation.

C. Proof of Theorem 2

The Fokker-Planck equation describes the evolution of the probability distribution of the parameters of a differential equation under stochastic forces. For a stochastic differential equation with diffusion coefficient $D(\boldsymbol{\theta})$, written as $\dot{\boldsymbol{\theta}} = f(\boldsymbol{\theta}) + N(0, 2D(\boldsymbol{\theta}))$, with the distribution of $\boldsymbol{\theta}$ being $p(\boldsymbol{\theta})$, the Fokker-Planck equation can be written as

$$\frac{\partial}{\partial t}\boldsymbol{p}(\boldsymbol{\theta}) = -\frac{\partial}{\partial \boldsymbol{\theta}}\left[f(\boldsymbol{\theta})p(\boldsymbol{\theta})\right] + \frac{\partial^2}{\partial \boldsymbol{\theta}^2}\left[D(\boldsymbol{\theta})p(\boldsymbol{\theta})\right], \quad (4)$$

where the notation $\frac{\partial^2}{\partial \boldsymbol{\theta}^2}$ denotes $\sum_{i,j}\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}$.

Theorem 2. *The dynamics defined in (2) leave the probability distribution defined by* $p(\boldsymbol{\theta},\boldsymbol{p},s,q) \propto \exp\left(-H(\boldsymbol{\theta},\boldsymbol{p},s,q)\right)$ *invariant.*

Proof. Let us start with the *deterministic* Hamiltonian dynamics, and replace the log-likelihood terms therein with their stochastic versions, without any corrections. Following the notation of (Yin & Ao, 2006), the dynamics can be represented in the following format:

$$\begin{bmatrix}\dot{\boldsymbol{\theta}}\\ \dot{\mathbf{p}}\\ \dot{s}\\ \dot{q}\end{bmatrix} = -\begin{bmatrix}0 & 0 & 0 & -I\\ 0 & 0 & I & 0\\ 0 & -I & 0 & 0\\ I & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\frac{\partial}{\partial s}H(\boldsymbol{\theta},\mathbf{p},s,q)\\ \frac{\partial}{\partial q}H(\boldsymbol{\theta},\mathbf{p},s,q)\\ \frac{\partial}{\partial \boldsymbol{\theta}}H(\boldsymbol{\theta},\mathbf{p},s,q)\\ \frac{\partial}{\partial \mathbf{p}}H(\boldsymbol{\theta},\mathbf{p},s,q)\end{bmatrix} + \boldsymbol{N} \quad (5)$$

where $\boldsymbol{N} = [0, N(0, 2\sqrt{s}B(\boldsymbol{\theta})), 0, N(0, 2A(\boldsymbol{\theta}))]^T$. Let us denote the anti-symmetric matrix above by X. Then, denoting $\nabla = [\partial/\partial\boldsymbol{\theta}; \partial/\partial\mathbf{p}; \partial/\partial s; \partial/\partial q]$, it is easy to see that $\text{tr}\left\{\nabla^T\nabla X y\right\} = 0$ for any $y(\boldsymbol{\theta},\mathbf{p},s,q)$.

Therefore the right hand side of the Fokker-Planck equation (4) can be written as

$$\begin{aligned}&-\text{tr}\nabla^T\left\{p(\boldsymbol{\theta},\mathbf{p},s,q)X\nabla H\right\} + \text{tr}\left\{\nabla^T D\nabla p(\boldsymbol{\theta},\mathbf{p},s,q)\right\}\\ &= -\text{tr}\nabla^T\left\{p(\boldsymbol{\theta},\mathbf{p},s,q)X\nabla H\right\}\\ &+\text{tr}\left\{(D+X)\nabla^T\nabla p(\boldsymbol{\theta},\mathbf{p},s,q)\right\}.\end{aligned}$$

Here we have used the shorthand ∇H to refer to the second matrix on the right hand side of equation (5), and

$$D = \begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 0 & 0 & \sqrt{s}B(\boldsymbol{\theta})\\ 0 & 0 & 0 & 0\\ 0 & A(\boldsymbol{\theta}) & 0 & 0\end{bmatrix}.$$

Note that $\nabla p = -p\nabla H$, since $p \propto \exp(-H)$. Therefore, if we simply replace X with $D+X$ in (5), the right hand side of the Fokker-Plank equation reduces to zero. Using $D+X$ in (5) is equivalent to the dynamics (2). □

As an aside, one can intuitively see that it is possible to incorporate any arbitrary dynamics into this formulation, by making appropriate assumptions about the diffusion terms in the noise and making corresponding corrections to the deterministic dynamics. Indeed, this fact has been investigated more thoroughly in (Ma et al., 2015).

D. Experimental Addenda

D.1. Parameter Estimation in Bayesian Logistic Regression

Tables 1 and 3 show the RMSE for the two parameters w_1 and w_2 for the synthetic Bayesian logistic regression experiment, for SGR-NPHMC and SG-NHT respectively.

Table 1. RMSE and auto-correlation times of the parameter samples from SGR-NPHMC runs on the synthetic Bayesian logistic regression dataset.

{A,B}	RMSE (w_1)	RMSE (w_2)
0.01	0.2684	0.1833
0.001	0.1927	0.1932
0.0001	0.1965	0.2125

Table 2. RMSE and auto-correlation times of the parameter samples from SG-NHT runs on the synthetic Bayesian logistic regression dataset.

{A}	RMSE (w_1)	RMSE (w_2)
0.1	0.2071	0.1884
1	0.2023	0.1850
10	0.2151	0.1907

D.2. Perplexity for Topic Modeling

We use the perplexity metric from (Zhou & Carin, 2015), where it is defined as

$$\text{Perplexity} = \exp\left(-\frac{1}{y_{..}}\sum_{n=1}^{N_{test}}\sum_{v=1}^{V} y_{nv}\log m_{nv}\right)$$

where y_{nv} =the count of vocabulary term v in held-out test document n, and $y_{..} = \sum_{n=1}^{N_{test}}\sum_{v=1}^{V} y_{nv}$.

m_{nv} is the mean of the collected samples of $\mathbf{\Theta} \times \mathbf{\Phi}$ normalized over the vocabulary, defined as

$$m_{nv} = \sum_{s=1}^{S}\sum_{k=1}^{K} \phi_{vk}^{(s)}\theta_{kn}^{(s)} \Big/ \sum_{v=1}^{V}\sum_{s=1}^{S}\sum_{k=1}^{K} \phi_{vk}^{(s)}\theta_{kn}^{(s)}.$$

Here s and k index the samples and latent topics respectively.

D.3. Runtime comparisons

We present the runtimes for SGR-NPHMC and SG-NHT for the synthetic and real-data experiments mentioned in the main paper [1]. We note that for toy datasets with low parameter dimensionality, our algorithm can be upto an order of magnitude slower than SG-NHT, albeit producing results with higher accuracy and lower variance. The runtime gap is closed when transitioning to real-life datasets with parameter dimensionality in the hundreds.

Table 3. Runtimes for SGR-NPHMC and SG-NHT for synthetic and real datasets (MS = milliseconds)

{DATASET}	SGR-NPHMC	SG-NHT
SYN-GAUSSIAN	4.4MS	0.36MS
SYN-BAYES LR	8.5MS	0.45MS
REAL-20 NEWS	0.95S	0.42S
REAL- REUTERS	3.8S	2.7S

References

Bond, S. D., Lemkuhler, B. J., and Laird, B. B. The Nosé-Poincaré Method for Constant Temperature Molecular Dynamics. *J. Comput. Phys*, 151:114–134, 1999.

Leimkuhler, B. and Reich, S. *Simulating Hamiltonian Dynamics*. Cambridge University Press, 2004.

Ma, Y., Chen, T., and Fox, E. B. A Complete Recipe for Stochastic Gradient MCMC. In *NIPS*, 2015.

Yin, L. and Ao, P. Existence and Construction of Dynamical Potential in Nonequilibrium Processes without Detailed Balance. *Journal of Physics A: Mathematical and General*, 39(27):8593, 2006.

Zhou, M. and Carin, L. Negative Binomial Process Count and Mixture Modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):307–320, 2015.

[1] All runtimes were gathered on a Windows 7 laptop with a 2.3Ghz Core i7 processor and 8GB RAM.