Slice Sampling on Hamiltonian Trajectories

Benjamin Bloem-Reddy REDDY@STAT.COLUMBIA.EDU
John P. Cunningham JPC2181@COLUMBIA.EDU
Department of Statistics, Columbia University

Abstract

Hamiltonian Monte Carlo and slice sampling are amongst the most widely used and studied classes of Markov Chain Monte Carlo samplers. We connect these two methods and present Hamiltonian slice sampling, which allows slice sampling to be carried out along Hamiltonian trajectories, or transformations thereof. Hamiltonian slice sampling clarifies a class of model priors that induce closed-form slice samplers. More pragmatically, inheriting properties of slice samplers, it offers advantages over Hamiltonian Monte Carlo, in that it has fewer tunable hyperparameters and does not require gradient information. We demonstrate the utility of Hamiltonian slice sampling out of the box on problems ranging from Gaussian process regression to Pitman-Yor based mixture models.

1. Introduction

After decades of work in approximate inference and numerical integration, Markov Chain Monte Carlo (MCMC) techniques remain the gold standard for working with intractable probabilistic models, throughout statistics and machine learning. Of course, this gold standard also comes with sometimes severe computational requirements, which has spurred many developments for increasing the efficacy of MCMC. Accordingly, numerous MCMC algorithms have been proposed in different fields and with different motivations, and perhaps as a result the similarities between some popular methods have not been highlighted or exploited. Here we consider two important classes of MCMC methods: Hamiltonian Monte Carlo (HMC) (Neal, 2011) and slice sampling (Neal, 2003).

HMC considers the (negative log) probability of the intractable distribution as the potential energy of a Hamiltonian system, and samples a new point from the distribution by simulating a dynamical trajectory from the current sample point. With careful tuning, HMC exhibits favorable mixing properties in many situations, particularly in high dimensions, due to its ability to take large steps in sample space if the dynamics are simulated for long enough. Proper tuning of HMC parameters can be difficult, and there has been much interest in automating it; examples include Wang et al. (2013); Hoffman & Gelman (2014). Furthermore, better mixing rates associated with longer trajectories can be computationally expensive because each simulation step requires an evaluation or numerical approximation of the gradient of the distribution.

Similar in objective but different in approach is slice sampling, which attempts to sample uniformly from the volume under the target density. Slice sampling has been employed successfully in a wide range of inference problems, in large part because of its flexibility and relative ease of tuning (very few if any tunable parameters). Although efficiently slice sampling from univariate distributions is straightforward, in higher dimensions it is more difficult. Previous approaches include slice sampling each dimension individually, amongst others. One particularly successful approach is to generate a curve, parameterized by a single scalar value, through the high-dimensional sample space. Elliptical slice sampling (ESS) (Murray et al., 2010) is one such approach, generating ellipses parameterized by $\theta \in [0, 2\pi]$. As we show in section 2.2, ESS is a special case of our proposed sampling algorithm.

In this paper, we explore the connections between HMC and slice sampling by observing that the elliptical trajectory used in ESS is the Hamiltonian flow of a Gaussian potential. This observation is perhaps not surprising – indeed it appears at least implicitly in Neal (2011); Pakman & Paninski (2013); Strathmann et al. (2015) – but here we leverage that observation in a way not previously considered. To wit, this connection between HMC and ESS suggests that we might perform slice sampling along more general Hamiltonian trajectories, a method we introduce under the name Hamiltonian slice sampling (HSS). HSS is of theoretical interest, allowing us to consider model priors that will induce

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

simple closed form slice samplers. Perhaps more importantly, HSS is of practical utility. As the conceptual offspring of HMC and slice sampling, it inherits the relatively small amount of required tuning from ESS and the ability to take large steps in sample space from HMC techniques.

In particular, we offer the following contributions:

- We clarify the link between two popular classes of MCMC techniques.
- We introduce Hamiltonian slice sampling, a general sampler for target distributions that factor into two components (e.g., a prior and likelihood), where the prior factorizes or can be transformed so as to factorize, and all dependence structure in the target distribution is induced by the likelihood or the intermediate transformation.
- We show that the prior can be either of a form which induces analytical Hamiltonian trajectories, or more generally, of a form such that we can derive such a trajectory via a measure preserving transformation. Notable members of this class include Gaussian process and stick-breaking priors.
- We demonstrate the usefulness of HSS on a range of models, both parametric and nonparametric.

We first review HMC and ESS to establish notation and a conceptual framework. We then introduce HSS generally, followed by a specific version based on a transformation to the unit interval. Finally, we demonstrate the effectiveness and flexibility of HSS on two different popular probabilistic models.

2. Sampling via Hamiltonian dynamics

We are interested in the problem of generating samples of a random variable $\mathbf{f}$ from an intractable distribution $\pi^*(\mathbf{f})$, either directly or via a measure preserving transformation $r^{-1}(\mathbf{q}) = \mathbf{f}$. For clarity, we use $\mathbf{f}$ to denote the quantity of interest in its natural space as defined by the distribution π^*, and $\mathbf{q}$ denotes the transformed quantity of interest in Hamiltonian phase space, as described in detail below. We defer further discussion of the map $r(\cdot)$ until section 2.4, but note that in typical implementations of HMC and ESS, $r(\cdot)$ is the indentity map, i.e. $\mathbf{f} = \mathbf{q}$.

2.1. Hamiltonian Monte Carlo

HMC generates MCMC samples from a target distribution, often an intractable posterior distribution, $\pi^*(\mathbf{q}) := \frac{1}{Z}\tilde{\pi}(\mathbf{q})$, with normalizing constant Z, by simulating the Hamiltonian dynamics of a particle in the potential $U(\mathbf{q}) = -\log\tilde{\pi}(\mathbf{q})$. We are free to specify a distribution for the particle's starting momentum $\mathbf{p}$, given which the system evolves deterministically in an augmented state space $(\mathbf{q}, \mathbf{p})$ according to Hamilton's equations. In particular, at the i-th sampling iteration of HMC, the initial conditions of the system are $\mathbf{q}_0 = \mathbf{q}^{(i-1)}$, the previous sample, and $\mathbf{p}_0$, which in most implementations is sampled from $\mathcal{N}(0, M)$. M is a "mass" matrix that may be used to express a priori beliefs about the scaling and dependence of the different dimensions, and for models with high dependence between sampling dimensions, M can greatly affect sampling efficiency. An active area of research is investigating how to adapt M for increased sampling efficiency; for example Girolami & Calderhead (2011); Betancourt et al. (2016). For simplicity, in this work we assume throughout that M is set ahead of time and remains fixed. The resulting Hamiltonian is

$$H(\mathbf{q}, \mathbf{p}) = -\log\tilde{\pi}(\mathbf{q}) + \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p}. \tag{1}$$

When solved exactly, Hamilton's equations generate Metropolis-Hastings (MH) proposals that are accepted with probability 1. In most situations of interest, Hamilton's equations do not have an analytic solution, so HMC is often performed using a numerical integrator, e.g. Störmer-Verlet, which is sensitive to tuning and can be computationally expensive due to evaluation or numerical approximation of the gradient at each step. See Neal (2011) for more details.

2.2. Univariate and Elliptical Slice Sampling

Univariate slice sampling (Neal, 2003) generates samples uniformly from the area beneath a density $p(x)$, such that the resulting samples are distributed according to $p(x)$. It proceeds as follows: given a sample x_0, a threshold $h \sim \mathrm{U}(0, p(x_0))$ is sampled; the next sample x_1 is then sampled uniformly from the slice $S := \{x : p(x) > h\}$. When S is not known in closed form, a proxy slice S' can be randomly constructed from operations that leave the uniform distribution on S invariant. In this paper, we use the stepping out and shrinkage procedure, with parameters w, the width, and m, the step out limit, from Neal (2003).

ESS (Murray et al., 2010) is a popular sampling algorithm for latent Gaussian models, e.g., Markov random field or Gaussian process ($\mathcal{GP}$) models. For a latent multivariate Gaussian $\mathbf{f} \in \mathbb{R}^d$ with prior $\pi(\mathbf{f}) = \mathcal{N}(\mathbf{f}|0, \Sigma)$, ESS generates MCMC transitions $\mathbf{f} \to \mathbf{f}'$ by sampling an auxiliary variable $\boldsymbol{\nu} \sim \mathcal{N}(0, \Sigma)$ and slice sampling the likelihood supported on the ellipse defined by

$$\mathbf{f}' = \mathbf{f}\cos\theta + \boldsymbol{\nu}\sin\theta, \quad \theta \in [0, 2\pi]. \tag{2}$$

Noting the fact that the solutions to Hamilton's equations in an elliptical potential are ellipses, we may reinterpret ESS. The Hamiltonian induced by $\tilde{\pi}(\mathbf{q}) \propto \mathcal{N}(\mathbf{q}|0, \Sigma)$ and mo-

mentum distribution $\mathbf{p} \sim \mathcal{N}(0, M)$ is

$$H(\mathbf{q}, \mathbf{p}) = \frac{1}{2}\mathbf{q}^T\Sigma^{-1}\mathbf{q} + \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p}. \quad (3)$$

When $M = \Sigma^{-1}$, Hamilton's equations yield the trajectory

$$\mathbf{q}(t) = \mathbf{q}(0)\cos t + \Sigma\mathbf{p}(0)\sin t, \quad t \in (-\infty, \infty), \quad (4)$$

which is an ellipse. Letting $r(\cdot)$ be the identity map, and $\mathbf{q}(0)$ be the value of $\mathbf{f}$ from the previous iteration of an MCMC sampler, the trajectories of possible sample values given by (2) and (4) are distributionally equivalent with $\theta = t \mod 2\pi$ because $\boldsymbol{\nu} \overset{d}{=} \Sigma\mathbf{p}(0) = M^{-1}\mathbf{p}(0)$. The ESS variable $\boldsymbol{\nu}$ thus takes on the interpretation of the initial velocity of the particle. Accordingly, we have clarified the link between HMC and ESS. The natural next question is what this observation offers in terms of a more general sampling strategy.

2.3. Hamiltonian Slice Sampling

Using the connection between HMC and ESS, we propose the sampling algorithm given in Algorithm 1, called **Hamiltonian slice sampling** (HSS). The idea of HSS is the same as ESS: use the *prior* distribution to generate an analytic curve through sample space, then slice sample on the curve according to the *likelihood* in order to generate a sample from the *posterior* distribution. In the special distributions for which Hamilton's equations have analytic solutions and the resulting trajectories can be computed exactly, e.g. the multivariate Gaussian distribution, the trajectories have a single parameter, $t \in (-\infty, \infty)$. This simple parameterization is crucial: it enables us to use univariate slice sampling methods (Neal, 2003) in higher dimensions. The advantage of univariate slice sampling techniques is that they require little tuning; the slice width adapts to the distribution and sampler performance does not depend greatly on the sampler parameters.

We include in the Supplementary Materials some examples of distributions for which Hamilton's equations have analytical solutions. However, these special cases are in the minority, and most distributions do not admit analytic solutions. In such a case, a measure preserving transformation to a distribution that does have analytic solutions is necessary. Let $r(\cdot)$ denote such a transformation. We defer elaboration on specific transformations until the following section, but it is sufficient for $r(\cdot)$ to be differentiable and one-to-one. In particular, we will make extensive use of the inverse function $r^{-1}(\cdot)$.

We next show that Algorithm 1 is a valid MCMC sampler for a target distribution $\pi^*(\mathbf{f}|\mathcal{D}) = \frac{1}{Z}L(\mathcal{D}|\mathbf{f})\ \pi(\mathbf{f})$. The proof of validity is conceptually so similar to ESS that the proof in Murray et al. (2010) nearly suffices. For completeness, we here show that the dynamics used to generate new

Algorithm 1 Hamiltonian slice sampling

Input: Current state $\mathbf{f}$; differentiable, one-to-one transformation $\mathbf{q} := r(\mathbf{f})$ with Jacobian $J(r^{-1}(\mathbf{q}))$
Output: A new state $\mathbf{f}'$. If $\mathbf{f}$ is drawn from π^*, the marginal distribution of $\mathbf{f}'$ is also π^*.

1: Sample momentum for each component of $\mathbf{q}$:

$$\mathbf{p} \sim \mathcal{N}(0, M)$$

2: Obtain analytic solution $\mathbf{q}'(t)$ to Hamilton's equations, where $\mathbf{p}_0 = \mathbf{p}$ and $\mathbf{q}'_0 = \mathbf{q}$
3: Set slice sampling threshold:

$$u \sim \mathrm{U}[0, 1]$$
$$\log h \leftarrow \log L(\mathcal{D}|\mathbf{f}) - \log|J(r^{-1}(\mathbf{q}))| + \log u$$

4: Slice sample along $r^{-1}(\mathbf{q}'(t))$ for $t^* \in (-\infty, \infty)$, using the methods of (Neal, 2003) and threshold h on:

$$\log \pi^*(\mathbf{f}'|\mathbf{f}, \mathbf{p}, \mathcal{D}, u, t) \propto \log L(\mathcal{D}|\mathbf{f}'(t)) - \log|J(r^{-1}(\mathbf{q}(t)))| \quad (5)$$

5: **return** $\mathbf{f}'(t^*) = r^{-1}(\mathbf{q}'(t^*))$

states are reversible, and as such, π^* is a stationary distribution of the Markov chain defined by Algorithm 1. Furthermore, the sampler has non-zero probability of transitioning to any region that has non-zero probability under π^*. Taken together, these facts clarify that the sampler will converge to a unique stationary distribution that yields samples of $\mathbf{f}$ that are marginally distributed according to the target π^*.

Consider the joint distribution of the random variables in Algorithm 1 (suppressing hyperparameters for simplicity and denoting by $\{t_k\}$ the sequence of variables produced by the univariate slice sampling algorithm):

$$p(\mathbf{f}, \mathbf{p}, h, \{t_k\}) = \pi^*(\mathbf{f}|\mathcal{D})\ p(\mathbf{p})\ p(h|\mathbf{f}, \mathcal{D})\ p(\{t_k\}|\mathbf{f}, \mathbf{p}, h) \quad (6)$$

$$= \frac{1}{Z}L(\mathcal{D}|\mathbf{f})\ \pi(\mathbf{f})\ p(\mathbf{p})\ p(h|\mathbf{f}, \mathcal{D})\ p(\{t_k\}|\mathbf{f}, \mathbf{p}, h) \quad (7)$$

Now, remembering that for a differentiable, one-to-one transformation $r(\cdot)$, such that $\mathbf{f} = r^{-1}(\mathbf{q})$, with Jacobian $J(\cdot)$, the density of $\mathbf{q}$ is given by $\tilde{\pi}(\mathbf{q}) = \pi(r^{-1}(\mathbf{q}))|J(r^{-1}(\mathbf{q}))|$. Combining this with the density of the slice sampling threshold variable implied

by step 3 of Algorithm 1, $p(h|\mathbf{f},\mathcal{D}) = \frac{|J(r^{-1}(\mathbf{q}))|}{L(\mathcal{D}|\mathbf{f})}$, we have

$$
\begin{aligned}
&p(\mathbf{f},\mathbf{p},h,\{t_k\}) \\
&\quad = \frac{1}{Z} L(\mathcal{D}|\mathbf{f})\, \pi(r^{-1}(\mathbf{q})) \cdot \frac{|J(r^{-1}(\mathbf{q}))|}{|J(r^{-1}(\mathbf{q}))|} \cdots \\
&\qquad\quad \times p(\mathbf{p})\, p(h|\mathbf{f},\mathcal{D})\, p(\{t_k\}|\mathbf{f},\mathbf{p},h) \qquad (8) \\
&\quad \propto \tilde{\pi}(\mathbf{q})\, p(\mathbf{p})\, p(\{t_k\}|\mathbf{q},\mathbf{p},h) \qquad (9)
\end{aligned}
$$

Exact Hamiltonian dynamics for $(\mathbf{q},\mathbf{p})$ are reversible, keeping constant the factor $\tilde{\pi}(\mathbf{q})\ p(\mathbf{p})$, and they yield a trajectory in $\mathbf{q}$-space parameterized by t that can be transformed into $\mathbf{f}$-space as $\mathbf{f}(t) = r^{-1}(\mathbf{q}(t))$. Given the Hamiltonian trajectory, the univariate slice sampling methods, e.g. the step-out and shrinkage procedures from Neal (2003) can be applied to $\mathbf{f}(t)$; these, too, are reversible. We therefore have $p(\mathbf{f},\mathbf{p},h,\{t_k\}) = p(\mathbf{f}',\mathbf{p}',h,\{t'_k\})$ for any $\mathbf{f}'$ generated by starting at $\mathbf{f}$, which concludes the proof of validity for HSS.

2.4. HSS via the Probability Integral Transformation

The multivariate Gaussian distribution is a special case for which Hamilton's equations can be solved analytically, as demonstrated by ESS. This fact forms the basis for Pakman & Paninski (2013; 2014). The univariate uniform and exponential distributions also have analytic solutions. In the remainder of the paper, we focus on transformations of uniform random variables; see the Supplementary Materials for details on the exponential distribution and a few related transformations.

Hamiltonian dynamics under the uniform distribution $q \sim \mathrm{U}[0,1]$ are the extremely simple billiard ball dynamics: the particle moves back and forth across the unit interval at constant speed $\dot{q}_0 = p_0/m$, reflecting off the boundaries of the interval. The relevant quantities are the initial reflection time,

$$
R_0 = \frac{|\mathbb{1}\{\dot{q}_0 > 0\} - q_0|}{\dot{q}_0}\,, \qquad (10)
$$

where $\dot{q}_0$ is the initial velocity, and the period of a full traversal of the unit interval,

$$
T = \frac{1}{\dot{q}_0} = \frac{m}{p_0}\,. \qquad (11)
$$

These simple dynamics can be used to construct analytic curves on which to slice sample for any continuous prior with computable inverse cumulative distribution function (CDF). In particular, for a random variable f with CDF $G(f)$ and density $g(f)$, the distribution of $q := G(f)$ is $\mathrm{U}[0,1]$. The dynamics on the unit interval yield the trajectory $q(t)$; the resulting curve is $f(t) = G^{-1}(q(t))$. In this case the Jacobian is $1/g(f)$, so the slice sampling target

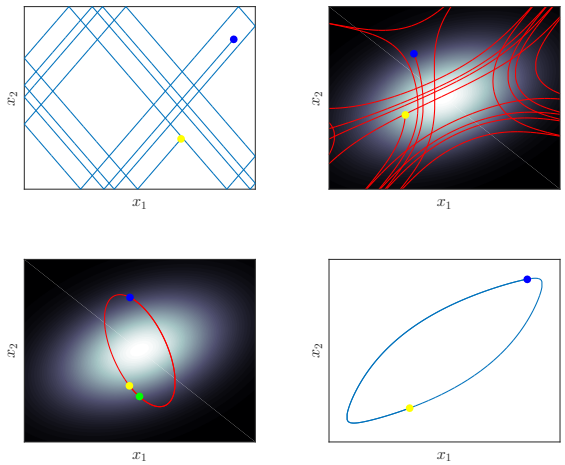

Figure 1. Example trajectories from HSS (top) and ESS (bottom). Yellow dots are starting points, blue dots are end points of Hamiltonian trajectories. *Top left:* Uniform trajectory in two independent dimensions. *Top right:* The resulting bivariate normal HSS trajectory via the probability integral transformation. *Bottom left:* Elliptical trajectory in two dimensions. The green dot in is $\boldsymbol{\nu}$ from ESS. *Bottom right:* Transformation of the elliptical trajectory to the unit square.

in Equation 5 of Algorithm 1 is exactly proportional to the posterior.

Although $f(t)$ is not a Hamiltonian flow in sample space, it is a measure preserving transformation of a Hamiltonian flow in the unit hypercube, and so preserves sampling validity. For the remainder of the paper, we use transformations to uniform random variables, and we refer to the resulting sampler as HSS. We note, however that it is only one instantiation of the general class of Hamiltonian slice sampling methods: any suitable transformation would produce different resulting dynamics in sample space.

Using this or any other transformation in Algorithm 1 generates curves through sample space only using information from the prior and the transformation. A useful line of inquiry is whether we can adapt the transformation or the prior to sample more efficiently. The problem of choosing or adapting to a prior that better captures the posterior geometry of the sample space is shared by all samplers based on Hamiltonian dynamics, including ESS and HMC. The issue is explored in depth in Girolami & Calderhead (2011); Betancourt et al. (2016). A possible approach to making the prior more flexible is via a pseudo-prior, as in Nishihara et al. (2014); Fagan et al. (2016). We do not explore such an approach here, but note that a pseudo-prior could be incorporated easily in Algorithm 1.

The basic requirement that must be satisfied in order to make our method applicable is that some component of the

target distribution can be expressed as a collection of conditionally independent variables. This is often the case in hierarchical Baysian models, as in section 3.2. As a simple example, whitening a collection of multivariate Gaussian random variables induces independence, enabling the simulation of independent uniform dynamics. An example of such dynamics on the unit square, along with the resulting trajectory in two-dimensional Gaussian space, is shown in Figure 1. For comparison, Figure 1 also shows an example ESS trajectory from the same distribution, along with its transformation to the unit square.

In principle, the HSS trajectory has access to different, possibly larger, regions of the sample space, as it is not constrained to a closed elliptical curve. It also spends relatively more time in regions of higher prior probability, as the particle's velocity in the original space is inversely proportional to the the prior density, i.e. $\dot{f}(t) = \dot{q}_0/g(f)$, as will be illustrated in section 3.2. This behavior also suggests that if the prior is sharply concentrated on some region, the sampler may get stuck. In order to avoid pathologies, relatively flat priors should be used. If the prior hyperparameters are also to be sampled, it should be done such that they make small moves relative to the parameters sampled by HSS. See Neal (2011, section 5.4.5) for discussion on this point in the context of HMC.

2.5. Related Work

HSS slice samples on a trajectory calculated using Hamiltonian dynamics. A similar idea underlies the No U-Turn Sampler (NUTS) (Hoffman & Gelman, 2014), which slice samples from the discrete set of states generated by the numerical simulation of Hamiltonian dynamics induced by the posterior distribution of interest. Other methods of multivariate slice sampling include those given in Neal (2003). Most often, each variable is slice sampled separately, which can lead to the same convergence problems as Gibbs sampling when the variables are highly dependent. Murray et al. (2010) review other sampling methods for latent $\mathcal{GP}$ models, of which HSS is another example. In so much as we derived HSS by noticing connections between existing methods, other commonalities may indeed exist.

ESS also fits into the framework of the preconditioned Crank-Nicholson (pCN) proposals studied in, e.g., Cotter et al. (2013). As those authors point out, the pCN proposal is a generalization of a random walk in the sample space, where "the target measure is defined with respect to a Gaussian." HSS also fits in this framework, with the target measure π^* defined with respect to non-Gaussian densities. ESS and HSS, like pCN, are dimension-free methods; their statistical efficiency is independent of the dimension of the sample space (Hairer et al., 2014). However, as more data is observed, and the posterior moves further from the prior, sampler performance may degrade – hence the motivation for adaptive techniques such as Nishihara et al. (2014); Fagan et al. (2016). Extending HSS to include pseudo-priors or Metropolis-adjusted Langevin-type variations, as in Cotter et al. (2013), is an interesting direction for future work.

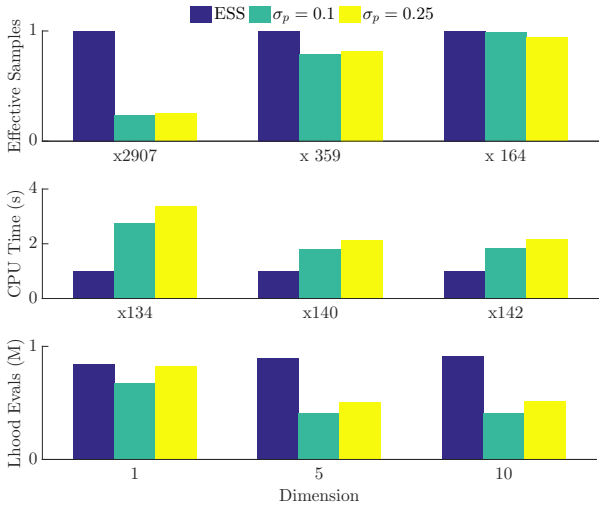

Figure 2. Effective number of samples for ESS and HSS from 10^5 iterations after 10^4 burn in on $\mathcal{GP}$ regression, averaged across 50 experiments. σ_p denotes the standard deviation of the normal distribution from which the Hamiltonian momentum variable $\mathbf{p}$ is sampled. *Top:* Effective samples. *Middle:* Evaluation time. *Bottom:* Number of likelihood evaluations.

3. Experiments

In order to test the effectiveness and flexibility of HSS, we performed experiments on two very different models. The first is latent $\mathcal{GP}$ regression on synthetic data, which allows us to make a direct comparison between HSS and ESS. The second is a non-parametric mixture of Gaussians with a stick-breaking prior, which demonstrates the flexibility of HSS. In both, we use the step-out and shrinkage slice sampling methods. See Neal (2003) for details.

3.1. $\mathcal{GP}$ Regression

We consider the standard $\mathcal{GP}$ regression model. We assume a latent $\mathcal{GP}$ f over an input space $\mathcal{X}$. Noisy Gaussian observations $\mathbf{y} = \{y_1, ..., y_n\}$ are observed at locations $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, and observations are conditionally independent with likelihood $L(\mathcal{D}|\mathbf{f}) := \prod_{i=1}^{n} \mathcal{N}(y_i|f(\mathbf{x}_i), \sigma_y^2)$, where σ_y^2 is the variance of the observation noise. Although the posterior is conjugate and can be sampled in closed form, the simplicity of the model allows us to test the performance of HSS against that of ESS on a model for which ESS is ideally suited, forming a highly conservative

baseline. Furthermore, we perform experiments in an increasing number of input dimensions, which lends insight to the scaling behavior of HSS.

Following Murray et al. (2010), we generated synthetic datasets with input dimension D from one to ten, each with $n = 200$ observations and inputs $\mathbf{x}_i$ drawn uniformly from the D-dimensional unit hypercube. Using the covariance function

$$\Sigma_{ij} = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}||\mathbf{x}_i - \mathbf{x}_j||_2^2\right),$$

the latent $\mathcal{GP}$ was drawn $\mathbf{f} \sim \mathcal{N}(0, \Sigma)$, with $\ell = 1$ and $\sigma_f^2 = 1$. Finally, observations y_i were drawn with noise variance $\sigma_y^2 = 0.3^2$.

Slice sampling parameters for HSS were set at $w = 0.5$ and $m = 8$. Momentum variables were sampled i.i.d. $p_i \sim \mathcal{N}(0, \sigma_p^2)$, with $\sigma_p \in \{0.1, 0.25\}$. We note that experiments were also performed with larger values of σ_p, but results are qualitatively similar. The HSS generated good posterior samples without any tuning of the parameters and it required very little tuning to achieve better efficiency.

Results for $D \in \{1, 5, 10\}$ are shown in Figure 2 (other dimensions are not shown for compactness). A commonly used measure of sample quality is the effective number of samples estimated from the log-likelihood, shown in the top panel using the R-CODA package (Plummer et al., 2006). In low dimensions, ESS is clearly more efficient than HSS. As the dimension increases, however, the performance gap closes; for $D = 10$, HSS samples nearly as efficiently as ESS. Again, this model is such that ESS should work well, and the fact that HSS quickly approaches its performance is encouraging. The computational time, shown in the second panel, is longer for HSS primarily due to the evaluation of the normal inverse CDF required to evaluate the likelihood. The number of likelihood evaluations is shown in the third panel. As it demonstrates, HSS requires relatively fewer slice sampling steps in higher dimensions, leading to fewer likelihood evaluations.

HSS is also easily applied to other latent $\mathcal{GP}$ models such as the log-Gaussian Cox process model or $\mathcal{GP}$ classification. In simple experiments (not shown), we have found implementation and tuning to be straightforward. Accordingly, even in situations where we anticipate good performance from ESS, HSS is a competitive alternative.

3.2. Mixture Models with Stick-breaking Priors

To demonstrate the flexibility of HSS, we investigate its performance on a markedly different type of model from $\mathcal{GP}$ regression. In particular, we test HSS on a widely used class of Bayesian nonparametric mixture models based on the Pitman-Yor ($\mathcal{PY}$) process, which may be formulated in terms of a stick-breaking prior (Ishwaran & James, 2001). The collection of independent stick breaks is amenable to sampling with HSS, which we embed in a Gibbs sampler for the overall model.

For observations Y_i and parameters $0 \leq \alpha < 1$ and $\theta > -\alpha$, the model can be written as follows:

$$\begin{aligned}
V_k|\alpha, \theta &\overset{\text{ind}}{\sim} \text{Beta}(1-\alpha, \theta + k\alpha), k = 1, 2, \ldots \\
\tilde{P}_k &= V_k \prod_{j=1}^{k-1}(1 - V_j), \\
\phi_k &\overset{\text{iid}}{\sim} \nu(\cdot), \\
Z_i|\{V_k\} &\overset{\text{iid}}{\sim} \sum_{k=1}^{\infty} \tilde{P}_k \delta_k(\cdot) \\
Y_i|Z_i, \{\phi_k\} &\overset{\text{ind}}{\sim} F(\cdot|\phi_{Z_i}).
\end{aligned} \tag{12}$$

When $\alpha = 0$, this is the ubiquitous Dirichlet Process (DP) mixture model. $\mathcal{PY}$ models have been used in a variety of applications, but in many cases sampling remains difficult, particularly when the posterior distribution of the random measure $\mu^*(\cdot) := \sum_{k=1}^{\infty} \tilde{P}_k \delta_k(\cdot)$ is of interest.

Two classes of samplers currently exist for sampling from $\mathcal{PY}$ mixture models. The so-called marginal class for DPs and certain other random measures (Favaro & Teh, 2013), marginalizes out the random measure μ^*. Marginal samplers are attractively simple and are widely used, but they do not allow direct inference of the random measure. A second class of samplers, the so-called conditional samplers, relies on a deterministic or probabilistic truncation of the random measure. Examples are Ishwaran & James (2001) and Kalli et al. (2011). Conditional samplers are often computationally inefficient, especially when the distribution of random weights $\{\tilde{P}_k)\}$ has heavy tails, as is the case when $\alpha > 0$. Their inefficiency stems from having to represent an infinite object in a finite amount of time and memory; either the truncation is a poor approximation, or the truncation has to occur at a very large k, which is computationally expensive.

In order to sample from the posterior efficiently, we embed HSS in a hybrid conditional-marginal blocked Gibbs sampler that alternates between sampling the individual cluster assignments $Z_i|Y_i, Z_{-i}, \{V_k\}, \{\phi_k\}$ and sampling all of the random measure parameters $\{V_k\}, \{\phi_k\}|\{Z_i\}$ simultaneously with HSS. In order to overcome the issues of the conditional sampler while still representing the entire random measure, we use a hybrid approach similar to that of Lomeli et al. (2015); only the weights of the "occupied" clusters are stored in memory; the weight of the unoccupied clusters is accumulated and used in the updates $Z_i|Y_i, Z_{-i}, \{V_k\}, \{\phi_k\}$, via a modification of the ReUse algorithm of Favaro & Teh (2013).

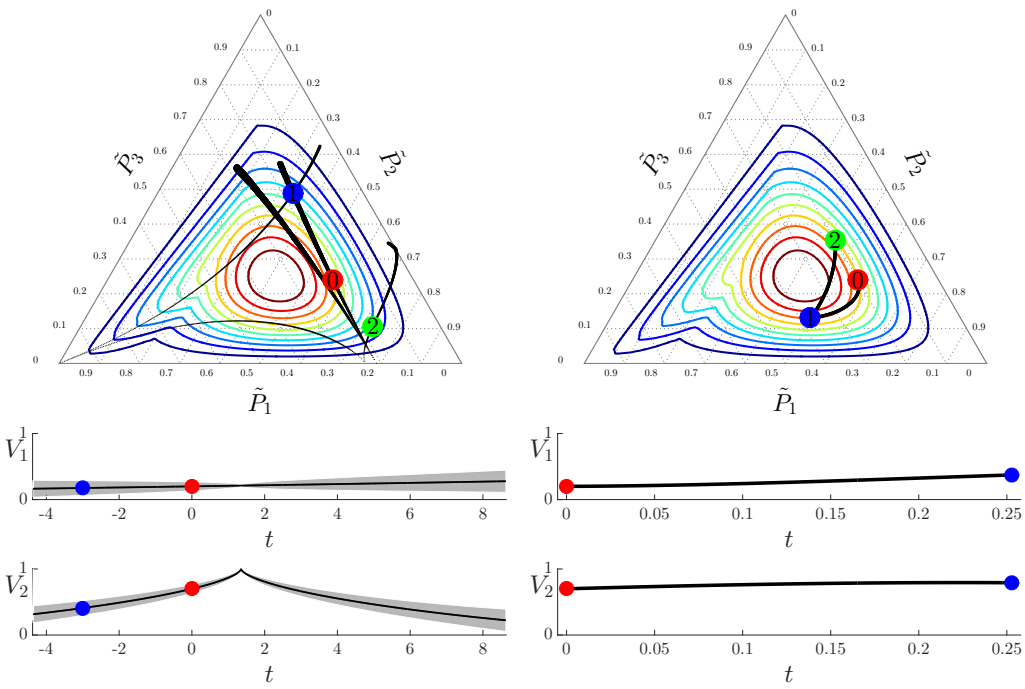

Figure 3. Top: Trajectories on the simplex used by HSS (left) and HMC (right) to generate two samples of stick weights V_k from a toy version of the model described in section 3.2, with $K = 3$. The contours of the posterior are shown, and order of the samples is indicated by the number on the sample point. *Bottom:* The trajectories for V_1 and V_2 corresponding to the first sample. (V_3 is fixed at 1.) The width of the trajectory (*top*) and shading (*bottom*) in the HSS plots is proportional to the inverse of the particle's velocity, and represents the probability mass placed by the slice sampler, which samples uniformly on the slice in t, on the corresponding part of sample space.

For comparison, we also used HMC to update the cluster parameters. HMC uses the gradient of the posterior and so should be expected to generate good posterior samples, if properly tuned. Figure 3 compares the behavior of HSS and HMC trajectories on the simplex in a toy example. To quatitatively compare HSS with HMC in a real data setting, we fit a simple one-dimensional Gaussian mixture model with a Gaussian prior on the cluster means, and a Gamma prior on the cluster precisions using the galaxy dataset, with $n = 82$ observations, that was also used in Favaro & Teh (2013); Lomeli et al. (2015).

In order to sample from the posterior distribution, we ran experiments with HSS or HMC embedded in the hybrid Gibbs sampler, fixing the hyperparameters at $\alpha = 0.3$ and $\theta = 1$. HSS was relatively easy to tune to achieve good convergence and mixing, with slice sampling parameters $w = 1$, and $m = \infty$. The different groups of latent variables in the model had very different scaling, and we found that $\sigma_p = 0.1$ and setting the mass equal to $m_v = 1$ for the $\{V_k\}$, and $m_n = m_\gamma = 10$ for the cluster means and precisions worked well. So as to compare the computation time between similar effective sample sizes (see section 3.3 below), we set the HMC parameters to achieve an acceptance ratio near 1, which required simulation steps of size $\epsilon \in [5 \times 10^{-5}, 15 \times 10^{-5}]$, sampled uniformly from that range for each sampling iteration. We ran HMC with $L \in \{60, 150\}$ steps for comparison. The results, displayed in Figure 4, show the relative efficiency of HSS; it achieves more effective samples per unit of computation time.

We note that HMC's performance improves, as measured by effective sample size, with larger simulation step sizes. However, larger simulation steps result in a non-trivial proportion of rejected HMC proposals, making impossible direct comparison with HSS due to the issue discussed in the following section. We also observed that the performance of HSS degraded with increasing sample size, because the posterior looks less like the prior, and the discussion in section 2.5 suggests. This behavior indicates that the use of a pseudo-prior would be beneficial in many situations.

3.3. Effective Sample Size in BNP Mixture Models

It is worth noting that the effective sample size (in log-joint probability) achieved by HSS in our experiments for the $\mathcal{PY}$ model are an order of magnitude lower than those reported by Favaro & Teh (2013); Lomeli et al. (2015), who calculated effective samples of K^*, the number of occupied clusters. While conducting preliminary tuning runs of the

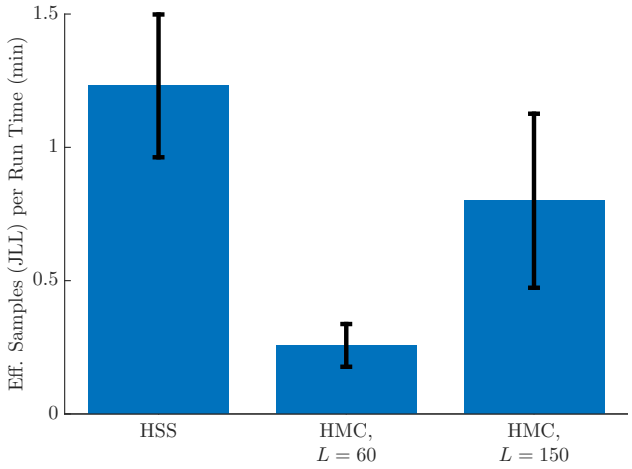

Figure 4. Effective number of samples per minute of run time for HSS and HMC from 10^5 samples taken at intervals of 20 interations, after 10^4 burn in iterations on the $\mathcal{PY}$ model, averaged across 10 experiments. L is the number of simulation steps used by HMC at each sampling iteration.

HMC sampler, we observed something odd: if the HMC simulations were not properly tuned, and thus almost every proposed HMC move was rejected, the resulting effective sample size in K^* was of the order of those previously reported. Further experiments in which we artificially limited HMC and HSS to take small steps produced the same effect: samplng efficiency, as measured either by K^* or log-joint probability, benefits from small (or no) changes in the cluster parameters. It seems that when the parameter clusters do a suboptimal job of explaining the data, the cluster assignment step destroys many of the clusters and creates new ones, sampling new parameters for each. This often happens when the parameter update step produces small or no changes to the cluster parameters. The result is steps in sample space that appear nearly independent, and correspondingly a large effective sample size, despite the undesireably small moves of the parameter updates. This underscores the importance of better measures of sample quality, especially for complicated latent variable models.

4. Discussion

Recognizing a link between two popular sampling methods, HMC and slice sampling, we have proposed Hamiltonian slice sampling, a general slice sampling algorithm based on analytic Hamiltonian trajectories. We described conditions under which analytic (possibly transformed) Hamiltonian trajectories can be guaranteed, and we demonstrated the simplicity and usefulness of HSS on two very different models: Gaussian process regression and mixture modeling with a stick-breaking prior. The former $\mathcal{GP}$ case is where ESS is particularly expected to perform, and in reasonable dimensionality we showed HSS performed competitively. The latter $\mathcal{PY}$ case is where we expect HMC to be more competitive (and where ESS does not apply). Here we found that HSS had a similar effective sample size but outperformed even carefully tuned HMC in terms of computational burden. As speed, scaling, and generality are always critical with MCMC methods, these results suggest HSS is a viable method for future study and application.

Acknowledgements

We thank Francois Fagan and Jalaj Bhandari for useful discussions, and for pointing out an error in an earlier draft; and anonymous referees for helpful suggestions. JPC is supported by funding from the Sloan Foundation and the McKnight Foundation.

References

Betancourt, M. J., Byrne, Simon, Livingstone, Samuel, and Girolami, Mark. The Geometric Foundations of Hamiltonian Monte Carlo. *Bernoulli*, (to appear), 2016.

Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D. MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster. *Statist. Sci.*, 28(3):424–446, 2013.

Fagan, Francois, Bhandari, Jalaj, and Cunningham, John P. Elliptical Slice Sampling with Expectation Propagation. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, volume (To appear), 2016.

Favaro, Stefano and Teh, Yee Whye. MCMC for Normalized Random Measure Mixture Models. *Statist. Sci.*, 28 (3):335–359, 2013.

Girolami, Mark and Calderhead, Ben. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011. ISSN 1467-9868.

Hairer, Martin, Stuart, Andrew M., and Vollmer, Sebastian J. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.*, 24(6): 2455–2490, 2014.

Hoffman, Matthew D. and Gelman, Andrew. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

Ishwaran, Hemant and James, Lancelot F. Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

Kalli, Maria, Griffin, Jim E., and Walker, Stephen G. Slice sampling mixture models. *Statistics and Computing*, 21 (1):93–105, 2011.

Lomeli, Maria, Favaro, Stefano, and Teh, Yee Whye. A hybrid sampler for Poisson-Kingman mixture models. In Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2152–2160. Curran Associates, Inc., 2015.

Murray, Iain, Adams, Ryan P., and MacKay, David J.C. Elliptical slice sampling. *JMLR: W&CP*, 9:541–548, 2010.

Neal, Radford M. Slice sampling. *Ann. Statist.*, 31(3):705–767, 2003.

Neal, Radford M. *Handbook of Markov Chain Monte Carlo*, chapter 5: MCMC using Hamiltonian dynamics, pp. 113–162. Chapman and Hall/CRC, 2011.

Nishihara, Robert, Murray, Iain, and Adams, Ryan P. Parallel MCMC with Generalized Elliptical Slice Sampling. *Journal of Machine Learning Research*, 15:2087–2112, 2014.

Pakman, Ari and Paninski, Liam. Auxiliary-variable Exact Hamiltonian Monte Carlo Samplers for Binary Distributions. In Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K.Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 2490–2498. Curran Associates, Inc., 2013.

Pakman, Ari and Paninski, Liam. Exact Hamiltonian Monte Carlo for Truncated Multivariate Gaussians. *Journal of Computational and Graphical Statistics*, 23 (2):518–542, 2014.

Plummer, Martyn, Best, Nicky, Cowles, Kate, and Vines, Karen. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6(1):7–11, 2006.

Strathmann, Heiko, Sejdinovic, Dino, Livingstone, Samuel, Szabo, Zoltan, and Gretton, Arthur. Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families. In Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 955–963. Curran Associates, Inc., 2015.

Wang, Ziyu, Mohamed, Shakir, and de Freitas, Nando. Adaptive Hamiltonian and Riemann Manifold Monte Carlo. In Dasgupta, Sanjoy and Mcallester, David (eds.), *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pp. 1462–1470. JMLR Workshop and Conference Proceedings, 2013.