Sequential Monte Carlo for Approximate Manifold Sampling

Mauro Camara Escudero, Christophe Andrieu, Mark Beaumont

University of Bristol

Abstract

Sampling from a probability density constrained to a manifold is of importance in numerous applications arising in statistical physics, statistics or machine learning. Sampling from such constrained densities, in particular using an MCMC approach, poses significant challenges and it is only recently that correct solutions have been proposed. The resulting algorithm can however be computationally expensive. We propose a relaxation of the problem and construct a bespoke and efficient parametrized family of MCMC kernels to sample from a small neighbourhood around the manifold, which we use as the transition kernel of an adaptive SMC sampler.

Distributions constrained on a Manifold

The co-area formula implies that the density of a distribution constrained on the level-set of a smooth function is the product of the unconstrained density and a correction term.

Theorem 0.1 (Constrained Density) *Let X be a $\mathbb{R}^n$-valued random variable with with distribution P_X and density p_X with respect to the Lebesgue measure. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth non-injective function with gradient $\nabla_x f(x)$. The density of P_X restricted to $\mathcal{M} = f^{-1}(0)$ is given by*

$$\overline{p}_X(x) = p_X(x) \left| \det \left(\nabla_x f(x) \nabla_x f(x)^\top \right) \right|. \tag{1}$$

Constrained Hamiltonian Monte Carlo (C-HMC) [1] introduces a Lagrange multiplier $\lambda \in \mathbb{R}$

$$H(x, v) = -\log \overline{p}_X(x) + \frac{1}{2} v^\top v + \lambda f(x) \tag{2}$$

and integrates the resulting separable Hamiltonian ODE system

$$\begin{aligned} \ddot{x} &= \nabla_x \log \overline{\pi}(x) - \lambda \nabla_x f(x) \\ f(x) &= 0 \end{aligned} \tag{3}$$

using RATTLE, a constrained version of the Leapfrog integrator, which is symplectic, reversible, second-order and explicit, but is costly as it requires optimization routines at each step to find λ and to enforce reversibility.

Distributions concentrated around a Manifold

A filamentary distribution is one that has a much larger scaling in directions tangential to the manifold, compared to directions orthogonal to it.

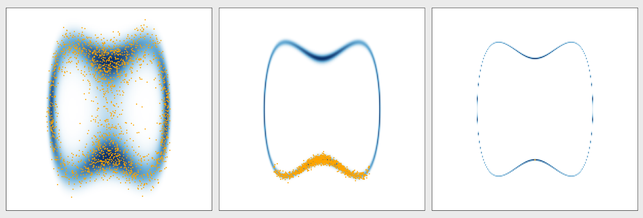
Definition 0.1.1 (Filamentary Distribution) *Let X be a $\mathbb{R}^n$-valued random variable with well-defined covariance matrix $\mathbb{V}[X]$, $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth function and $\mathcal{M} = f^{-1}(0)$. For any $x \in \mathcal{M}$ let $\widehat{\nabla} f(x)$ and $\mathsf{T}(x) = \{\widehat{t}_i(x)\}_{i=1}^n$ denote the normalized gradient and a tangent basis at x. The distribution of X is filamentary if*

$$\widehat{\nabla} f(x)^\top \mathbb{V}[X] \widehat{\nabla} f(x) \ll \widehat{t}_i(x)^\top \mathbb{V}[X] \widehat{t}_i(x) \qquad \forall\, x \in \mathcal{M} \;\; \text{and} \; \forall\, \widehat{t}_i(x) \in \mathsf{T}(x) \tag{4}$$

Usually, the filamentary density $\widehat{p}_X$ is the product of an unconstrained density and a smoothing kernel $k_\epsilon^{\mathcal{M}}(\cdot)$ with bandwidth ϵ

$$\widehat{p}_X(x) = p_X(x) k_\epsilon^{\mathcal{M}}(x),$$

and $\widehat{p}_X \to \overline{p}_X$ as the distribution concentrates more and more around the manifold. Sampling from these densities is challenging with standard MCMC kernels as shown below. As the filamentary density (blue) concentrates more tightly around $\mathcal{M}$, HMC (orange) fails to explore it efficiently.



Hug kernel for Filamentary Distributions

Theorem 0.2 (Metropolis-Hastings for Volume-Preserving Involutions) *If there exists a distribution with density $\rho(x, v) = \widehat{p}(x) p(v \mid x)$ and a volume-preserving involution $\phi(x, v)$ then this is a valid MH algorithm targeting $\widehat{p}_X(x)$:*

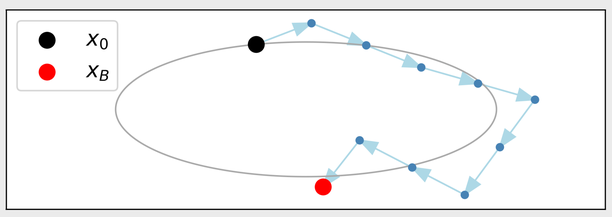
1. *Given x, sample auxiliary variable $v \sim p(v \mid x)$*
2. *Accept $(x', v') = \phi(x, v)$ with probability $1 \wedge \rho(x', v')/\rho(x, v)$ otherwise stay at (x, v).*

The involution of the Hug kernel for filamentary distributions is an approximate discretization of (3) that is volume-preserving (but not symplectic), explicit, and approximately second-order. It samples an auxiliary velocity variable $v \sim \mathcal{N}(0, \mathrm{I})$ and then performs B integrator steps with stepsize $\delta > 0$.

Algorithm 1: Hug Kernel (one iteration)

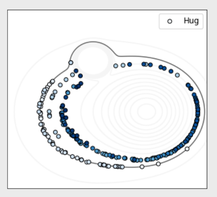
Given x_0, sample auxiliary velocity variable $v_0 \sim \mathcal{N}(0, \mathrm{I})$.
for $b = 0, \ldots, B-1$ **do**
  $x_{b+\delta/2} = x_b + (\delta/2) v_b$
  $v_{b+1} = v_b - 2 \widehat{\nabla} f(x_{b+\delta/2}) \widehat{\nabla} f(x_{b+\delta/2})^\top v_b$
  $x_{b+1} = x_{b+\delta/2} + (\delta/2) v_{b+1}$
end
With probability $\widehat{p}_X(x_B)/\widehat{p}_X(x_0)$ accept x_B, otherwise stay at x_0.

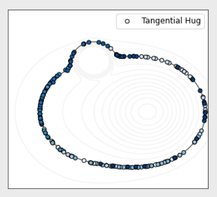
At each step it uses the gradient at the midpoint $x_{b+\delta/2}$ to reflect the velocity back towards $\mathcal{M}$ using the bouncing mechanism of the Bouncy Particle Sampler. The bounce mechanism approximates the curvature information and as a result the samples "hug" $\mathcal{M}$. Gradients can be computed because the smoothness of f allows to disintegrate $\mathbb{R}^n$ into $\{f^{-1}(y)\}_{y \in \mathbb{R}}$. In other words, for each $x \in \mathbb{R}^n$, $x \in f^{-1}(\{f(x)\})$.



Tangential Hug kernel for Filamentary Distributions

Sampling $v \sim \mathcal{N}(0, \mathrm{I})$ can lead to unstable trajectories when $\nabla_x^2 f(x)$ is large, and can still lead to low acceptance probabilities if $\widehat{p}_X(x)$ is very highly concentrated around $\mathcal{M}$. The reason is that if the velocity is reflected using a midpoint too far away in the direction orthogonal to $\mathcal{M}$, then the sampler can lose track of $\mathcal{M}$, as shown in the left figure. We propose to sample $v \sim \mathcal{N}(0, \mathrm{I})$ and then "squeeze" it towards the tangent plane before integrating the dynamics. To ensure reversibility of the involution at the end of the integration path, we need to "unsqueeze" the velocity. The result is Tangential Hug, a family of kernels parametrized by $\alpha \in [0, 1)$ that determines the degree of the compression of v towards the tangent plane. For $\alpha = 0$ we recover Hug, and for $\alpha \approx 1$ our initial velocity approximately lies on the tangent plane.





Algorithm 2: Thug Kernel (one iteration)

Sample auxiliary velocity variable $v_0 \sim \mathcal{N}(0, \mathrm{I})$.
Squeeze velocity $w_0 = v_0 - \alpha \widehat{\nabla} f(x_0) \widehat{\nabla} f(x_0)^\top v_0$.
for $b = 0, \ldots, B-1$ **do**
  $x_{b+\delta/2} = x_b + (\delta/2) w_b$
  $w_{b+1} = w_b - 2 \widehat{\nabla} f(x_{b+\delta/2}) \widehat{\nabla} f(x_{b+\delta/2})^\top w_b$
  $x_{b+1} = x_{b+\delta/2} + (\delta/2) w_{b+1}$
end
Unsqueeze velocity $v_B = w_B + (\alpha/(1-\alpha)) \widehat{\nabla} f(x_B) \widehat{\nabla} f(x_B)^\top w_B$.
With probability $\exp(\ell(x_B) - \ell(x_0) - \|v_B\|^2/2 + \|v_0\|^2/2)$ accept x_B, otherwise stay at x_0.

As shown in the right figure above, this leads to a more stable sampler that doesn't lose track of $\mathcal{M}$. There is no free lunch, however: while $\alpha > 0$ allows us to stay closer to the manifold, it also introduces a new term in the acceptance ratio, which reduces the acceptance probability. There is a trade-off between the decrease in acceptance probability due to $\alpha > 0$, and the increase in acceptance probability due to the increased ability to track $\mathcal{M}$. When $\widehat{p}_X(x)$ is highly concentrated, we found THUG to outperform HUG.

Adaptive SMC-THUG for Filamentary Distributions

Choosing α can be difficult in practice. While we know that larger values should be used for more concentrated targets, we don't have a systematic way of determining the optimal value of α. However, it is natural to use THUG as the transition kernel for an SMC sampler targeting a filamentary distribution and adapting α at each iteration, based on the estimated acceptance probability $\hat{a}_i$ and a target acceptance probability a^*

$$\tau_{i+1} = \tau_i + \gamma_{i+1}(\hat{a}_i - a^*) \qquad \text{where} \qquad \tau_i = \log\left(\frac{\alpha_i}{1 - \alpha_i}\right).$$

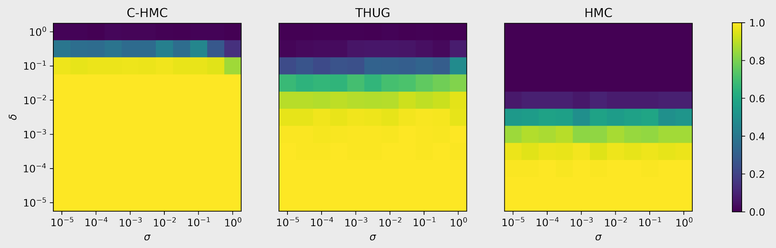
Consider posterior Bayesian inference of $p(\theta \mid y)$ given observation $y \in \mathbb{R}$ generated as

$$y = F(\theta) + \sigma\eta \qquad \text{where } \sigma > 0 \quad \text{and } \eta \sim \mathcal{N}(0, 1),$$

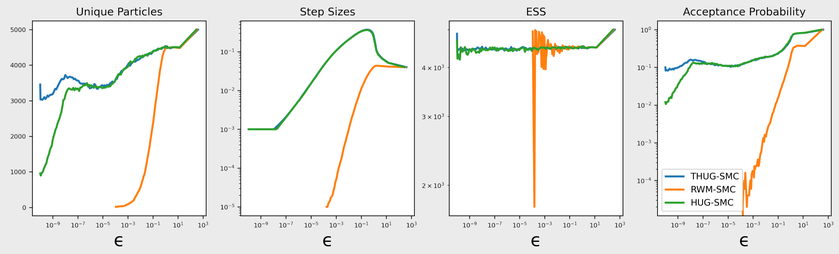
where $F : \mathbb{R}^2 \to \mathbb{R}$ is defined as $F(\theta) = \theta_2^2 + 3\theta_1^2(\theta_1^2 - 1)$. The posterior for different noise scales is shown in Figure 1. Given fixed $\sigma > 0$ it is possible to define the manifold of parameters and noise variables exactly reconstructing the observed data

$$\mathcal{M} = \left\{ (\theta, \eta) \in \mathbb{R}^3 \,:\, F(\theta) + \sigma\eta = y \right\}.$$

Using Theorem 0.1 one can find the posterior density $\overline{p}_\sigma(\theta, \eta \mid y)$ on $\mathcal{M}$ and sample from it with C-HMC (left figure). Similarly, one can introduce a smoothing kernel (e.g. Epanechnikov) and define a filamentary distribution $\widehat{p}_{\epsilon,\sigma}(\theta, \eta \mid y) \propto p_\sigma(\theta, \eta \mid y) k_\epsilon^{\mathcal{M}}(\theta, \eta)$ that tends to $\overline{p}_\sigma$ as $\epsilon \to 0$. We use THUG and HMC to sample from $\widehat{p}$ (middle and right figure) for $\sigma = 1 \times 10^{-8}$. We can see that THUG is able to use a step size almost two orders of magnitude larger than HMC while keeping the same acceptance probability.



Below we compare SMC with Random Walk, HUG and THUG kernels. Both HUG and THUG outperform RWM and squeezing the velocity towards the tangent plane brings benefits when $\widehat{p}$ is very highly concentrated.



Bibliography

[1] Lelievre, T., Rousset, M., and Stoltz, G. Hybrid monte carlo methods for sampling probability measures on submanifolds, 2019.

[2] Au, K. X., Graham, M. M., and Thiery, A. H. Manifold lifting: scaling mcmc to the vanishing noise regime, 2021

[3] Ludkin, M. and Sherlock, C. Hug and hop: a discrete-time, non-reversible markov chain monte-carlo algorithm, 2021.