Approximate Manifold Sampling

Mauro Camara Escudero, Christophe Andrieu, Mark Beaumont

University of Bristol

Motivation

Overview

- Probability distributions on lower-dimensional submanifolds:
  - Bayesian inverse problems (Au et al., 2021)
  - Approximate Bayesian Computation (Graham & Storkey, 2017)
  - Molecular Dynamics (Lelièvre et al., 2010)
  - Topological Statistics (Diaconis et al., 2012)
  - Diffusion models (Graham et al., 2019)

Overview

- Probability distributions on lower-dimensional submanifolds:
  - Bayesian inverse problems (Au et al., 2021)
  - Approximate Bayesian Computation (Graham & Storkey, 2017)
  - Molecular Dynamics (Lelièvre et al., 2010)
  - Topological Statistics (Diaconis et al., 2012)
  - Diffusion models (Graham et al., 2019)
- Constrained samplers such as C-HMC and C-RWM (Lelièvre et al., 2019; Zappa et al., 2018) are very expensive: require 2 calls to optimization routines per sample.

Overview

- Probability distributions on lower-dimensional submanifolds:
  - Bayesian inverse problems (Au et al., 2021)
  - Approximate Bayesian Computation (Graham & Storkey, 2017)
  - Molecular Dynamics (Lelièvre et al., 2010)
  - Topological Statistics (Diaconis et al., 2012)
  - Diffusion models (Graham et al., 2019)
- Constrained samplers such as C-HMC and C-RWM (Lelièvre et al., 2019; Zappa et al., 2018) are very expensive: require 2 calls to optimization routines per sample.
- Contribution: avoid costly operations by developing an efficient sampler (THUG) for a relaxation of the problem.

Application: Bayesian Inverse Problems

- Observational model with data-generating mechanism

$$y = F(\theta) + \upsilon \qquad \upsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}) \qquad F \text{ smooth.}$$

- Observe y^* and perform inference on

$$p_\sigma(\theta \mid y^*) \propto p(\theta)\mathcal{N}(F(\theta) \mid y^*, \sigma^2 \mathrm{I})$$

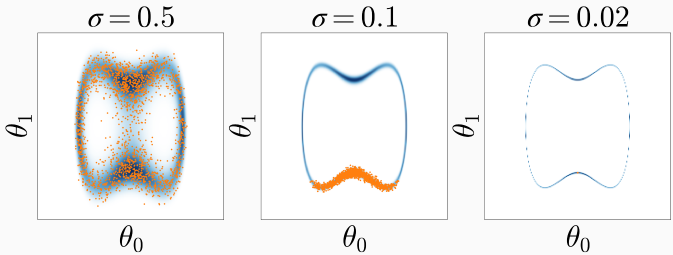
- For $\sigma > 0$ small the posterior is concentrated around

$$\mathcal{M} = \{\theta \in \Theta \,:\, F(\theta) = y^*\}\,.$$

- For $\sigma \to 0$ the posterior $p_\sigma(\theta \mid y^*)$ is supported on $\mathcal{M}$.

Application: Bayesian Inverse Problem[1]

- Let $F(\theta_0, \theta_1) = \theta_1^2 + 3\theta_0^2(\theta_0^2 - 1)$ and observe $y^* = 1$.
- Posterior for 3 values of noise scale. Samples via HMC.



[1] Au et al. (2021)

Tools and Background

Assumptions and Notation

Assumptions

- $n > m$.

Facts and Notation

Assumptions and Notation

Assumptions

- $n > m$.
- $\pi(x)$ prior density on $\mathbb{R}^n$ wrt Lebesgue measure.

Facts and Notation

Assumptions and Notation

Assumptions

- $n > m$.
- $\pi(x)$ prior density on $\mathbb{R}^n$ wrt Lebesgue measure.
- $f : \mathbb{R}^n \to \mathbb{R}^m$ smooth

Facts and Notation

Assumptions and Notation

Assumptions

- $n > m$.
- $\pi(x)$ prior density on $\mathbb{R}^n$ wrt Lebesgue measure.
- $f : \mathbb{R}^n \to \mathbb{R}^m$ smooth

Facts and Notation

- $f^{-1}(y)$ manifold for almost every $y \in \mathbb{R}^m$.

Assumptions and Notation

Assumptions

- $n > m$.
- $\pi(x)$ prior density on $\mathbb{R}^n$ wrt Lebesgue measure.
- $f : \mathbb{R}^n \to \mathbb{R}^m$ smooth

Facts and Notation

- $f^{-1}(y)$ manifold for almost every $y \in \mathbb{R}^m$.
- At each $x \in f^{-1}(y)$ tangent $\mathcal{T}_x$ and normal $\mathcal{N}_x$ spaces are defined.

Assumptions and Notation

Assumptions

- $n > m$.
- $\pi(x)$ prior density on $\mathbb{R}^n$ wrt Lebesgue measure.
- $f : \mathbb{R}^n \to \mathbb{R}^m$ smooth

Facts and Notation

- $f^{-1}(y)$ manifold for almost every $y \in \mathbb{R}^m$.
- At each $x \in f^{-1}(y)$ tangent $\mathcal{T}_x$ and normal $\mathcal{N}_x$ spaces are defined.
- J full row-rank and $\mathcal{J}_m f := |\det \mathrm{J}(x)\mathrm{J}(x)^\top|^{1/2} > 0$ a.e.

Assumptions and Notation

Assumptions

- $n > m$.
- $\pi(x)$ prior density on $\mathbb{R}^n$ wrt Lebesgue measure.
- $f : \mathbb{R}^n \to \mathbb{R}^m$ smooth

Facts and Notation

- $f^{-1}(y)$ manifold for almost every $y \in \mathbb{R}^m$.
- At each $x \in f^{-1}(y)$ tangent $\mathcal{T}_x$ and normal $\mathcal{N}_x$ spaces are defined.
- J full row-rank and $\mathcal{J}_m f := |\det \mathrm{J}(x)\mathrm{J}(x)^\top|^{1/2} > 0$ a.e.
- T_x and N_x projection matrices well-defined a.e.

Assumptions and Notation

Assumptions

- $n > m$.
- $\pi(x)$ prior density on $\mathbb{R}^n$ wrt Lebesgue measure.
- $f : \mathbb{R}^n \to \mathbb{R}^m$ smooth

Facts and Notation

- $f^{-1}(y)$ manifold for almost every $y \in \mathbb{R}^m$.
- At each $x \in f^{-1}(y)$ tangent $\mathcal{T}_x$ and normal $\mathcal{N}_x$ spaces are defined.
- J full row-rank and $\mathcal{J}_m f := |\det \mathrm{J}(x)\mathrm{J}(x)^\top|^{1/2} > 0$ a.e.
- T_x and N_x projection matrices well-defined a.e.
- $\mathcal{H}^{n-m}(dx)$ Hausdorff measure on $f^{-1}(y)$.

General Setup

Observe $y^* \in \mathbb{R}^m$.

- Exact Manifold Sampling:

General Setup

Observe $y^* \in \mathbb{R}^m$.

- Exact Manifold Sampling:
  - Posterior: $\pi(x)$ restricted on $f^{-1}(y^*)$ (Manifold Distribution)

$$\eta(x) = \pi(x)\mathcal{J}_m f^{-1}(x)$$

General Setup

Observe $y^* \in \mathbb{R}^m$.

- Exact Manifold Sampling:
  - Posterior: $\pi(x)$ restricted on $f^{-1}(y^*)$ (Manifold Distribution)

$$\eta(x) = \pi(x)\mathcal{J}_m f^{-1}(x)$$

- Approximate Manifold Sampling

General Setup

Observe $y^* \in \mathbb{R}^m$.

- Exact Manifold Sampling:
  - Posterior: $\pi(x)$ restricted on $f^{-1}(y^*)$ (Manifold Distribution)

$$\eta(x) = \pi(x)\mathcal{J}_m f^{-1}(x)$$

- Approximate Manifold Sampling
  - Posterior: $\pi(x)$ concentrated around $f^{-1}(y^*)$ (Filamentary Distribution)

$$\eta_\epsilon(x) = \pi(x)k_\epsilon(\|y^* - f(x)\|)$$

where k_ϵ is a kernel (approximation to the identity).

General Setup

Observe $y^* \in \mathbb{R}^m$.

- Exact Manifold Sampling:
  - Posterior: $\pi(x)$ restricted on $f^{-1}(y^*)$ (Manifold Distribution)

$$\eta(x) = \pi(x)\mathcal{J}_m f^{-1}(x)$$

- Approximate Manifold Sampling
  - Posterior: $\pi(x)$ concentrated around $f^{-1}(y^*)$ (Filamentary Distribution)

$$\eta_\epsilon(x) = \pi(x)k_\epsilon(\|y^* - f(x)\|)$$

where k_ϵ is a kernel (approximation to the identity).

When is this relaxation sensible?

$\mathbb{E}_{\eta_\epsilon}[\psi] \to \mathbb{E}_\eta[\psi]$ as $\epsilon \to 0^+$

What smoothing kernels are allowed?

Definition (Approximation to the identity (ATI))

A sequence $\{k_\epsilon : \mathbb{R}^m \to \mathbb{R}\}_{\epsilon>0}$ *of integrable functions is an approximation to the identity if there exists a constant* A *such that*

$$\int_{\mathbb{R}^m} k_\epsilon(y)dy = 1 \qquad \forall\, \epsilon > 0$$

$$|k_\epsilon(y)| \leq \frac{A}{\epsilon^m} \qquad \forall\, \epsilon > 0,\ \forall y \in \mathbb{R}^m$$

$$|k_\epsilon(y)| \leq \frac{A\epsilon}{\|y\|^{m+1}} \qquad \forall\, \epsilon > 0,\ \forall\, y \in \mathbb{R}^m \backslash \{0\}$$

Convergence of Filamentary Distributions

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ Lipschitz, with J full row-rank almost everywhere. Let $\psi : \mathbb{R}^n \to \mathbb{R}$ be π-integrable, and $\{k_\epsilon\}_{\epsilon>0}$ be an ATI. Then for almost every $y^* \in \mathbb{R}^m$

$$\lim_{\epsilon \to 0^+} \int_{\mathbb{R}^n} \psi(x)\eta_\epsilon(x)dx = \int_{f^{-1}(y^*)} \psi(x)\eta(x)\mathcal{H}^{n-m}(dx)$$

Alternative Theorems

Weaker conditions on ψ are possible (see paper).

Exact Manifold Sampling

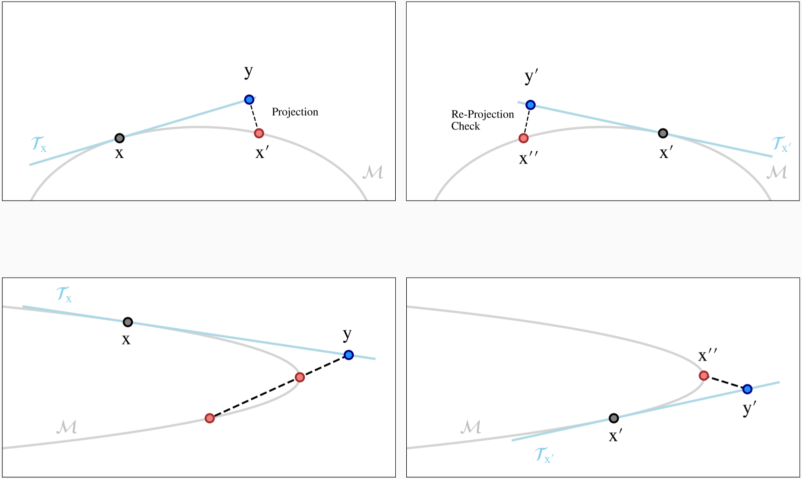
Constrained Random Walk Metropolis (C-RWM)

- *Proposal Step* Given $x \in \mathcal{M}$, sample a Gaussian perturbation on the tangent space[2] $v \in \mathcal{T}_x$ and move to $y = x + v$. Typically $y \notin \mathcal{M}$ so a non-linear projection is required: find $\lambda \in \mathbb{R}^m$ such that $x' = y + \mathrm{J}_x^\top \lambda$ lies on $\mathcal{M}$ via e.g. Newton method (Au et al., 2021).
- *Reversibility Check* Multiple such λ might exist, but not all might satisfy detailed balance. Need to check that running the algorithm backwards from x' one would get to x with tolerance $\rho > 0$.
- *Acceptance Step* Metropolis-Hastings

$$a(x, x') = \min\left\{1, \frac{\eta(x')\mathcal{N}(v' \mid 0, \mathrm{I})}{\eta(x)\mathcal{N}(v \mid 0, \mathrm{I})}\right\}.$$

[2] Sample $\nu \sim \mathcal{N}(0, \mathrm{I}_n)$ and project $v = \mathrm{T}_x \nu$.

Illustration of C-RWM



Approximate Manifold Sampling

The concept of a bounce

Imagine a billiard ball hitting the cushion of a pool table. A bounce is the composition of three operations: straight line movement, a reflection of the direction of motion, and another straight line movement in this new direction.

Bounce

For any orthogonal matrix R, and step size $\delta > 0$ the bounce

$$\mathbb{B}_{\mathrm{R},\delta}(x,v) = \left(x + \frac{\delta}{2}v + \frac{\delta}{2}\mathrm{R}v, \mathrm{R}v\right)$$

is time-reversible and volume-preserving, i.e.

- $\phi \circ \mathbb{B}_{\mathrm{R},\delta}(x,v)$ is an involution (here $\phi(x,v) = \phi(x,-v)$).
- has unit absolute determinant Jacobian $|\det(\mathrm{J}_{\mathbb{B}_{\mathrm{R},\delta}})| = 1$.

THUG Bounce

- Tangential Hug (THUG) uses a particular reflection matrix

$$\mathrm{R} = \mathrm{I}_n - 2\mathrm{N}_{x+(\delta/2)v}.$$

THUG Bounce

- Tangential Hug (THUG) uses a particular reflection matrix

$$\mathrm{R} = \mathrm{I}_n - 2\mathrm{N}_{x+(\delta/2)v}.$$

- Velocity reflection off $\mathcal{T}_{x+(\delta/2)v}$, the tangent space at the bounce point.

THUG Bounce

- Tangential Hug (THUG) uses a particular reflection matrix

$$\mathrm{R} = \mathrm{I}_n - 2\mathrm{N}_{x+(\delta/2)v}.$$

- Velocity reflection off $\mathcal{T}_{x+(\delta/2)v}$, the tangent space at the bounce point.
- Intuition: Moving along $\mathcal{N}_x$ would lead to largest change in f, so by going in the opposite direction we are trying to minimize this change.

THUG Bounce

- Tangential Hug (THUG) uses a particular reflection matrix

$$\mathrm{R} = \mathrm{I}_n - 2\mathrm{N}_{x+(\delta/2)v}.$$

- Velocity reflection off $\mathcal{T}_{x+(\delta/2)v}$, the tangent space at the bounce point.
- Intuition: Moving along $\mathcal{N}_x$ would lead to largest change in f, so by going in the opposite direction we are trying to minimize this change.

THUG Bounce Precision (inspired by Ludkin & Sherlock (2019))

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be smooth, and let J_x and $\mathrm{H}[x]$ be its Jacobian matrix and Hessian tensor respectively. If H is bounded by $\beta \in (0, \infty)$ and γ-Lipschitz, then applying the THUG bounce $B \in \mathbb{Z}_+$ times starting from $x_0, v_0 \in \mathbb{R}^n$ gives

$$\|f(x_B) - f(x_0)\| \leq \frac{\delta^2 \|v_0\|^2}{8} \left(2\beta + \gamma \|T v_0\|\right) =: \mathcal{B}_0$$

where $T = B\delta$ is the total integration time.

THUG bounce as an integrator I

- THUG bounce is an explicit second-order integrator for the dynamics of a particle with constant speed and centripetal acceleration on $\mathcal{M}$

$$\begin{aligned}\dot{x} &= v \\ \dot{v} &= -\mathrm{J}_x^\top (\mathrm{J}_x \mathrm{J}_x^\top)^{-1} \mathrm{H}[x](v, v)\end{aligned}$$

THUG bounce as an integrator I

- THUG bounce is an explicit second-order integrator for the dynamics of a particle with constant speed and centripetal acceleration on $\mathcal{M}$

$$\dot{x} = v$$
$$\dot{v} = -\mathrm{J}_x^\top (\mathrm{J}_x \mathrm{J}_x^\top)^{-1} \mathrm{H}[x](v, v)$$

- Although dynamic requires velocity to be on tangent space at all times, all properties above are still satisfied even if that's not the case.

THUG bounce as an integrator I

- THUG bounce is an explicit second-order integrator for the dynamics of a particle with constant speed and centripetal acceleration on $\mathcal{M}$

$$\begin{aligned} \dot{x} &= v \\ \dot{v} &= -\mathrm{J}_x^\top (\mathrm{J}_x \mathrm{J}_x^\top)^{-1} \mathrm{H}[x](v, v) \end{aligned}$$

- Although dynamic requires velocity to be on tangent space at all times, all properties above are still satisfied even if that's not the case.
- However, one expects THUG bounce to be more precise if initial velocity has smaller normal component.

THUG bounce as an integrator I

- THUG bounce is an explicit second-order integrator for the dynamics of a particle with constant speed and centripetal acceleration on $\mathcal{M}$

$$\begin{aligned}\dot{x} &= v \\ \dot{v} &= -\mathrm{J}_x^\top(\mathrm{J}_x\mathrm{J}_x^\top)^{-1}\mathrm{H}[x](v, v)\end{aligned}$$

- Although dynamic requires velocity to be on tangent space at all times, all properties above are still satisfied even if that's not the case.
- However, one expects THUG bounce to be more precise if initial velocity has smaller normal component.
- Introduce squeezing matrix and operator for $\alpha \in [0, 1)$

$$\mathrm{T}_{x,\alpha} = \mathrm{I}_n - \alpha\mathrm{N}_x \qquad \text{and} \qquad \mathbb{T}_\alpha(x, v) = (x, \mathrm{T}_{x,\alpha}v)$$

THUG bounce as an integrator II

Squeezed THUG Bounce Precision

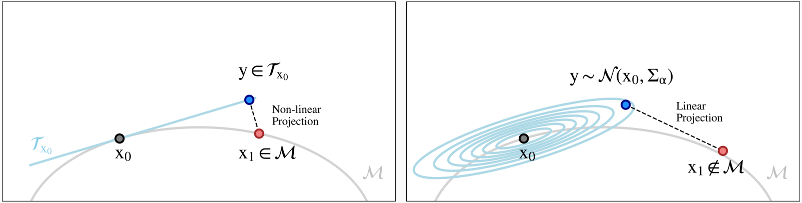
Applying $\mathbb{T}_\alpha$ with $\alpha > 0$ and $\|v_0^\perp\| > 0$ before using the THUG bounce allows one to improve the previous constant $\mathcal{B}_0$

$$\|f(x'_B) - f(x_0)\| \leq \mathcal{B}_0 - \frac{\alpha(2-\alpha)\delta^2\|v_0^\perp\|^2}{8}(2\beta + \gamma\|Tv_0\|) =: \mathcal{B}_\alpha,$$

where $\mathcal{B}_\alpha < \mathcal{B}_0$.

In practice we found that using $\alpha > 0$ can lead to important performance improvements when η_ϵ is particularly tight around $\mathcal{M}$.

Illustration of C-RWM and THUG



Tangential Hug (THUG)

- $\mathbb{B}_{\text{THUG}} \circ \mathbb{T}_\alpha$ is not time-reversible.

Tangential Hug (THUG)

- $\mathbb{B}_{\text{THUG}} \circ \mathbb{T}_\alpha$ is not time-reversible.
- Need to unsqueeze the velocity at the end via

$$\mathrm{T}_{x,\alpha}^{-1} = \mathrm{I}_n + \frac{\alpha}{1-\alpha}\mathrm{N}_x \qquad \text{and} \qquad \mathbb{T}_\alpha^{-1}(x, v) = (x, \mathrm{T}_{x,\alpha}^{-1} v).$$

Tangential Hug (THUG)

- $\mathbb{B}_{\text{THUG}} \circ \mathbb{T}_\alpha$ is not time-reversible.
- Need to unsqueeze the velocity at the end via

$$\mathrm{T}_{x,\alpha}^{-1} = \mathrm{I}_n + \frac{\alpha}{1-\alpha}\mathrm{N}_x \qquad \text{and} \qquad \mathbb{T}_\alpha^{-1}(x, v) = (x, \mathrm{T}_{x,\alpha}^{-1} v).$$

- The full THUG proposal mechanism is $\mathbb{T}_\alpha^{-1} \circ \mathbb{B}_{\text{THUG}}^B \circ \mathbb{T}_\alpha$.

Tangential Hug (THUG)

- $\mathbb{B}_{\text{THUG}} \circ \mathbb{T}_\alpha$ is not time-reversible.
- Need to unsqueeze the velocity at the end via

$$\mathrm{T}_{x,\alpha}^{-1} = \mathrm{I}_n + \frac{\alpha}{1-\alpha}\mathrm{N}_x \qquad \text{and} \qquad \mathbb{T}_\alpha^{-1}(x, v) = (x, \mathrm{T}_{x,\alpha}^{-1} v).$$

- The full THUG proposal mechanism is $\mathbb{T}_\alpha^{-1} \circ \mathbb{B}_{\text{THUG}}^{B} \circ \mathbb{T}_\alpha$.
- Since the squeezing and unsqueezing operations happen at different positions, $\|v_B\|^2 - \|v_0\|^2$ will appear in the acceptance ratio.

Tangential Hug (THUG)

- $\mathbb{B}_{\text{THUG}} \circ \mathbb{T}_\alpha$ is not time-reversible.
- Need to unsqueeze the velocity at the end via

$$\mathrm{T}_{x,\alpha}^{-1} = \mathrm{I}_n + \frac{\alpha}{1-\alpha}\mathrm{N}_x \qquad \text{and} \qquad \mathbb{T}_\alpha^{-1}(x, v) = (x, \mathrm{T}_{x,\alpha}^{-1} v).$$

- The full THUG proposal mechanism is $\mathbb{T}_\alpha^{-1} \circ \mathbb{B}_{\text{THUG}}^B \circ \mathbb{T}_\alpha$.
- Since the squeezing and unsqueezing operations happen at different positions, $\|v_B\|^2 - \|v_0\|^2$ will appear in the acceptance ratio.

No-free Lunch

The change in norm squared after using THUG with B steps with $\alpha \in [0, 1)$ is

$$\|v_B\|^2 - \|v_0\|^2 = \mathcal{O}\left(\delta \frac{\alpha(2-\alpha)}{(1-\alpha)^2}\right).$$

Tangential Hug (THUG) Algorithm

Algorithm 1: Tangential Hug (One Iteration)

Sample auxiliary: $v_0 \sim \mathcal{N}(0, \mathrm{I})$. Set $(x, v) = (x_0, v_0)$.
Squeeze: $v \longleftarrow v - \alpha \text{LinearProjection}(\mathrm{J}(x), v)$
for $b = 1, \ldots, B$ **do**
    Move: $x \longleftarrow x + (\delta/2)v$
    Bounce: $v \longleftarrow v - 2\text{LinearProjection}(\mathrm{J}(x), v)$
    Move: $x \longleftarrow x + (\delta/2)v$
end
Unsqueeze: $v \longleftarrow v + (\alpha/(1-\alpha))\text{LinearProjection}(\mathrm{J}(x), v)$
MH: Accept with prob $a = \exp(\ell(x) - \ell(x_0) - \|v\|^2/2 + \|v_0\|^2/2)$.

Experiments

Bayesian Inverse Problem - Acceptance Probability

- THUG/HMC/RM-HMC target $p_\sigma(\theta \mid y^*)$
- C-RWM target $p_\sigma(\theta, \upsilon \mid y^*)$ on

$$\mathcal{M}_\sigma = \{(\theta, \upsilon) \,:\, F(\theta) + \upsilon = y^*\}\,.$$

Notice $p_\sigma(\theta, \upsilon \mid y^*)$ remains diffuse for $\sigma \to 0$, unlike $p_\sigma(\theta \mid y^*)$.

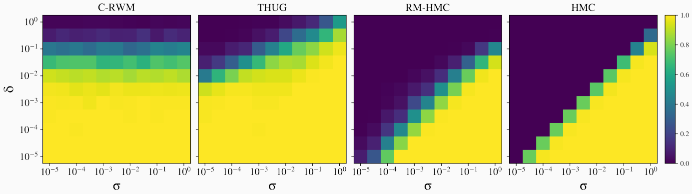


Figure 1: Average Acceptance Probability for a grid of $\sigma > 0$ and $\delta > 0$. Results averaged over 10 runs of 50 samples each, keeping $B = 20$ fixed.

Bayesian Inverse Problem - Computational Cost

- Run 12 chains of 2500 samples keeping $B = 20$ and $\delta = 0.1$ fixed.
- Phase one: σ large then posterior is not filamentary and HMC is better.
- Phase two: σ small and THUG superior.
- Phase three: σ very small and C-HMC is more advantageous.

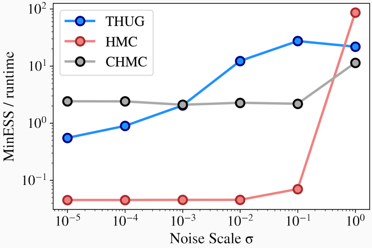


Figure 2: minESS over total runtime (in seconds).

ABC - G and K distribution - Computational Cost

- min-bulk-ESS across 4 chains of 1000 samples each for increasing dimensionality $m \in \{50, 100, 200\}$.
- Run algorithms for $B \in \{1, 10, 50\}$, $\epsilon \in \{10^0, \ldots, 10^{-8}\}$ and $\alpha \in \{0, 0.9, 0.99\}$.

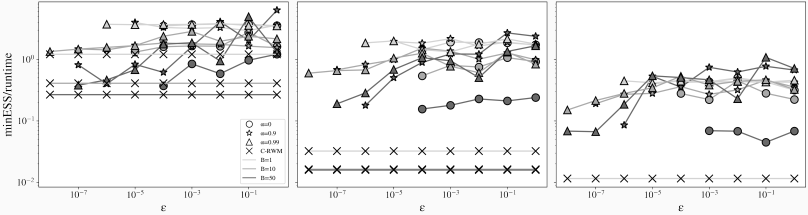
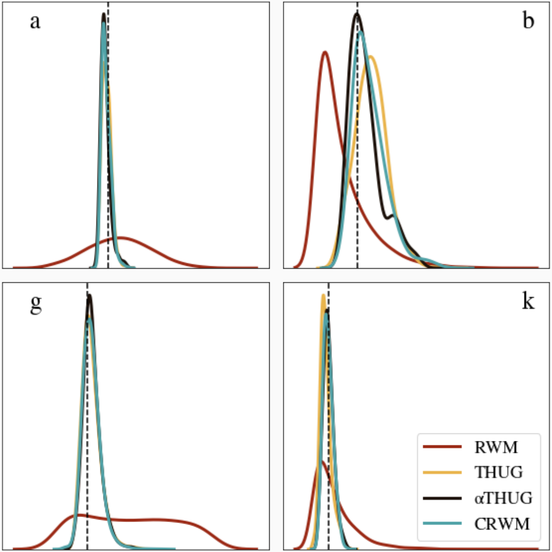


Figure 3: minESS over total runtime (in seconds).

ABC - G and K distribution - Density Estimation

$10k$ samples after an initial warmup. Each algorithm run at their best ϵ.



Conclusion and Future Work

- Real-world applications.
- Comparing manifold and filamentary distributions.
- Develop a suitable notion of ESS of these problems.

Thank you

References

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

Diaconis, P., Holmes, S., and Shahshahani, M. Sampling from a manifold, 2012.

Federer, H. *Geometric Measure Theory*. Classics in Mathematics. Springer Berlin Heidelberg, 2014. ISBN 9783642620102. URL `https://books.google.co.uk/books?id=jld-BgAAQBAJ`.

Graham, M. M. and Storkey, A. J. Asymptotically exact inference in differentiable generative models, 2017.

References ii

Graham, M. M., Thiery, A. H., and Beskos, A. Manifold markov chain monte carlo methods for bayesian inference in diffusion models, 2019. URL *https://arxiv.org/abs/1912.02982*.

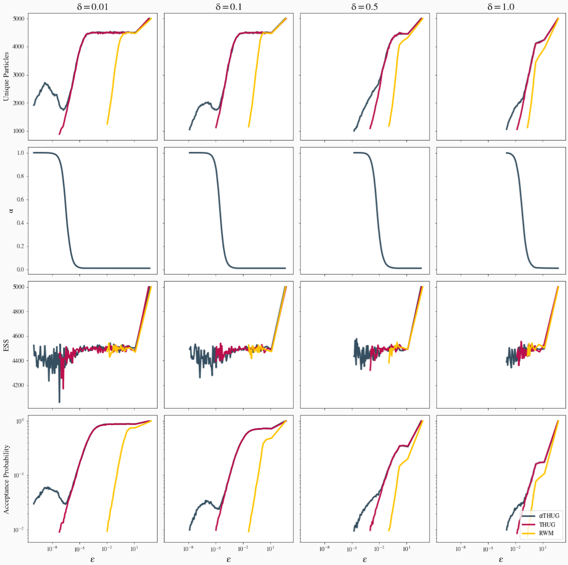
Lelièvre, T., Rousset, M., and Stoltz, G. *Free Energy Computations*. IMPERIAL COLLEGE PRESS, 2010. doi: 10.1142/p579. URL *https://www.worldscientific.com/doi/abs/10.1142/p579*.

Lelièvre, T., Rousset, M., and Stoltz, G. Hybrid monte carlo methods for sampling probability measures on submanifolds, 2019.

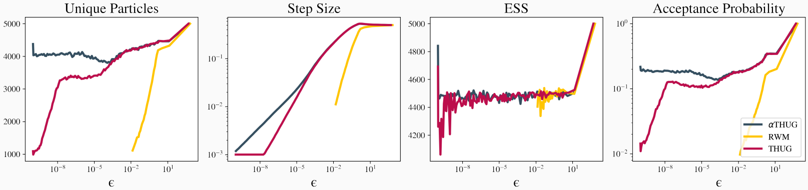
Ludkin, M. and Sherlock, C. Hug and hop: a discrete-time, non-reversible markov chain monte-carlo algorithm, 2019. URL *https://arxiv.org/abs/1907.13570*.

Zappa, E., Holmes-Cerfon, M., and Goodman, J. Monte carlo on manifolds: Sampling densities and integrating functions. *Communications on Pure and Applied Mathematics*, 71(12):2609–2647, 2018. doi: https://doi.org/10.1002/cpa.21783. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21783`.

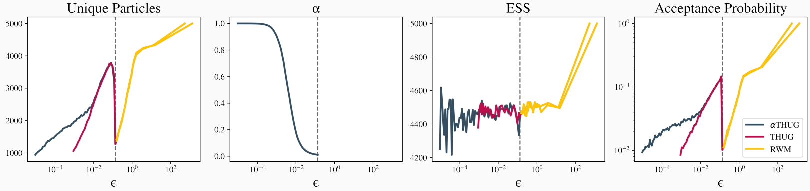
SMC Results - Fixed Step Size



SMC Results - Adapting Both



SMC Results - RWM followed by THUG



THUG MCMC III

Aim of Experiment

Does the acceptance probability of Hug/Thug deteriorate at slower rate than HMC/RM-HMC with respect to step size?

- Run Thug, Hug, RM-HMC and HMC to target filamentary posterior

$$p_\sigma(\theta \mid y) \propto p(\theta)\mathcal{N}(h(\theta), \sigma^2 \mathrm{I}).$$

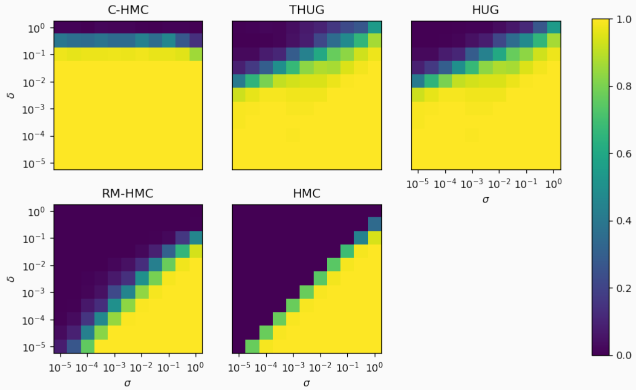
  and C-HMC to target lifted manifold posterior

$$\overline{p}(\theta, \eta \mid y) \propto p(\theta)p(\eta)|J_{f_\sigma}(\theta, \eta)J_{f_\sigma}(\theta, \eta)^\top|^{-1/2}$$

- Run across a grid of noise scale $\sigma \in (1 \times 10^{-5}, 1.0)$ and step-sizes $\delta \in (1 \times 10^{-5}, 1.0)$, keeping number of steps/bounces per iteration $B = L = 20$ fixed.
- Average acceptance probability across 10 runs of 50 samples.

Thug MCMC IV

HMC and RM-HMC need $\mathcal{O}(\delta) = \mathcal{O}(\sigma)$ for a good acceptance probability. Hug and Thug can achieve the same acceptance probability with 3 order of magnitude larger step-size.



Acceptance Probability vs Discretization Order

- Ludkin & Sherlock (2019) showed that when $f = \ell$, H is γ-Lipschitz and bounded above by $\beta > 0$ Hug satisfies

$$|\ell_B - \ell_0| \leq \frac{\delta^2}{8}\|v_0\|^2(2\beta + \gamma T\|v_0\|) =: \mathcal{B}_{\text{HUG}}$$

  Thug satisfies a tighter bound when $\alpha > 0$ and $\hat{g}_0^\top v_0 \neq 0$

$$|\ell_B - \ell_0| \leq \mathcal{B}_{\text{HUG}} - \frac{\alpha(2-\alpha)\delta^2(\hat{g}_0^\top v_0)^2}{8}(2\beta + \gamma T\|v_0\|) =: \mathcal{B}_{\text{THUG}}.$$

- When $f \neq \ell$ will require assumptions on relationship between f and ℓ

$$p_\epsilon(x \mid y) \propto p(x)k_\epsilon(\|y - f(x)\|)$$

GPV for Partitioned Systems

For a Partitioned ODE

$$\begin{aligned}\dot{x} &= F_1(x, v)\\ \dot{v} &= F_2(x, v)\end{aligned}$$

the Generalized Position Verlet (GPV) integrator

$$\begin{aligned}x_{n+1/2} &= x_n + \frac{\delta}{2} F_1(x_{n+1/2}, v_n)\\ v_{n+1} &= v_n + \frac{\delta}{2}\left[F_2(x_{n+1/2}, v_n) + F_2(x_{n+1/2}, v_{n+1})\right]\\ x_{n+1} &= x_{n+1/2} + \frac{\delta}{2} F_1(x_{n+1/2}, v_{n+1})\end{aligned}$$

is **implicit**, second-order, symmetric and symplectic.

GPV for Separable Systems

For a Separable ODE

$$\dot{x} = F_1(v)$$
$$\dot{v} = F_2(x)$$

the GPV integrator

$$x_{n+1/2} = x_n + \frac{\delta}{2} F_1(v_n)$$
$$v_{n+1} = v_n + \delta F_2(x_{n+1/2})$$
$$x_{n+1} = x_{n+1/2} + \frac{\delta}{2} F_1(v_{n+1})$$

is **explicit**, second-order, symmetric and symplectic.

Alternative Integrator I

Although in general the Generalized Position Verlet for a non-separable system is implicit, it turns out that one can actually solve explicitly for v_{n+1} in the velocity update.

$$v_{n+1} = \underbrace{v_n - \frac{\delta}{2} \frac{v_n^\top H_F(x_{n+1/2} v_n}{\|\nabla_x F(x_{n+1/2})\|} \widehat{\nabla_x F}(x_{n+1/2})}_{:=a}$$

$$\underbrace{-\frac{\delta}{2} \frac{\widehat{\nabla_x F}(x_{n+1/2})}{\|\nabla_x F(x_{n+1/2})\|}}_{=:b} v_{n+1}^\top H_F(x_{n+1/2}) v_{n+1},$$

then the expression has the form (we write $H_{n+1/2} = H(x_{n+1/2})$)

$$v_{n+1} = a + b v_{n+1}^\top H_{n+1/2} v_{n+1}.$$

This can be solved by solving a simple quadratic equation for ϑ

$$c_1 \vartheta^2 + (2c_2 - 1)\vartheta + c_3 = 0$$

Alternative Integrator II

where

$$
\begin{aligned}
c_1 &= b^\top H_{n+1/2} b \\
c_2 &= a^\top H_{n+1/2} b \\
c_3 &= a^\top H_{n+1/2} a.
\end{aligned}
$$

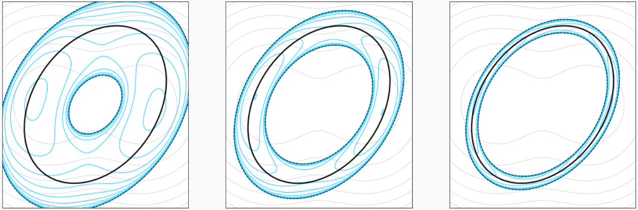
Interestingly, this discretization works well for sampling from filamentary distributions only when the initial velocity is perpendicular to the gradient at the initial position $v_0 \perp \hat{g}_0$, otherwise it quickly blows up. This is in contrast with the generalised Hug algorithm which remains stable thanks to the BPS reflection mechanism.

A General Approximate Manifold Sampling Framework

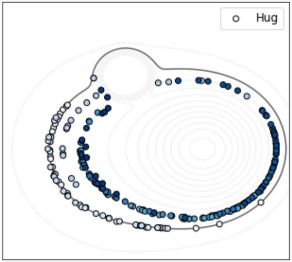
Let π be a filamentary distribution whose limiting manifold distribution is $\overline{\pi}$. A general approximate manifold sampling algorithm consists of a triplet (H_p, Φ, H_a) where

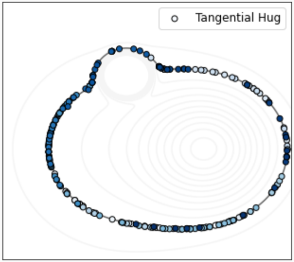
- H_p is a Hamiltonian system that forms the base of our proposal mechanism. A good H_p would follow/stay close to $\mathcal{M}$ and perhaps be a good Hamiltonian system for $\overline{\pi}$.
- Φ is a reversible (or skew-reversible) integrator for H_p of suitably high order and preferably with $|\det J_\Phi| = 1$, symplecticity is desired but not needed.
- H_a is a Hamiltonian that determines which samples get accepted or rejected. This should include π for the algorithm to be correct.

Contours of Filamentary Distribution



Tangential Hug Stays closer





Manifold Distributions I

Transformation of Random Variable by Diffeomorphism

Let X be an $\mathbb{R}^n$-valued random vector with density p_X. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a diffeomorphism and $Y = f(X)$. Then

$$p_Y(y) dy = p_X(f^{-1}(y)) |\det J_{f^{-1}}(y)| dy$$

The Co-Area formula for Lipschitz functions generalizes the above results to non-injective functions (see Theorem 5.3.9 in Federer (2014)).

Conditional Density of Random Variable on Submanifold

Let X be an $\mathbb{R}^n$-valued random vector with density p_X. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a smooth function with $n > m$, and let $y \in \mathbb{R}^m$. Then on the sub-manifold $f^{-1}(y)$

$$p(x \mid f(x) = y) \mathcal{H}^{n-m}(dx) \propto p_X(x) |\det(J_f(x) J_f(x)^\top)|^{-1/2} \mathcal{H}^{n-m}(dx)$$

Manifold Distributions II

Assumption 2

π admits a density with respect to the Hausdorff measure on $\mathcal{M}$.

Manifold Distribution

Let $X : \Omega \to \mathbb{R}^n$ be a vector-valued random variable with distribution π and finite covariance matrix $\mathbb{V}_\pi[X]$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth function. Consider $y \in \mathbb{R}$ fixed, then at any point $\xi \in f^{-1}(y)$ we denote by $\hat{g}(\xi)$ the normalized gradient of f and by $\mathsf{T}(\xi) = \{\hat{t}_1(\xi), \ldots, \hat{t}_{n-1}(\xi)\}$ a basis for the tangent space at ξ. Then π is a manifold distribution if $\forall \xi \in \mathbb{R}^n$ and $\forall \hat{t}_i(\xi) \in \mathsf{T}(\xi)$

$$\hat{g}(\xi)^\top \mathbb{V}_\pi[X] \hat{g}(\xi) = 0 \qquad \text{and} \qquad \hat{t}_i(\xi)^\top \mathbb{V}_\pi[X] \hat{t}_i(\xi) > 0.$$

Typically obtained as limiting posterior density as some scale parameter goes to zero.

Manifold Distribution III

Manifold Distribution

Let U be an $\mathbb{R}^n$-valued random variable, and $f : \mathbb{R}^n \to \mathbb{R}^m$ be smooth. Let $\mathrm{K}(y, du)$ be a regular conditional distribution of U given $\sigma(f(U))$ and let $\mathcal{H}_y^{n-m}$ be the Hausdorff measure on $f^{-1}(y)$. If $\mathrm{K}(y, \cdot) \ll \mathcal{H}_y^{n-m}$ then $\pi = \mathrm{K}(y, \cdot)$ is a manifold distribution.

Manifold Distribution IV

Graham's Theorem Revisited

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $U : \Omega \to \mathbb{R}^n$ be a random vector with distribution P_U and density p_U with respect to λ^n, the Lebesgue measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Let $n >$m and $f : \mathbb{R}^n \to \mathbb{R}^m$ be a smooth function with Jacobian matrix $J_f(u)$ having full row-rank λ^n-almost everywhere, and let $f \circ U$ have distribution $P_f = P_U \circ f^{-1}$. Let $\sigma(f)$ be the sigma-algebra generated by $f \circ U$, and let $\phi : \mathbb{R}^n \to \mathbb{R}$ be a $\mathcal{B}(\mathbb{R}^n)$-measurable test function. Let $\mathrm{K}(y, du)$ be a RCD of U given $\sigma(f)$ from $(\mathbb{R}^m, \sigma(f))$ to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, such that $\mathrm{K}(y, \cdot) \ll \mathcal{H}^{n-m}$. Then expectations with respect to K can be written as

$$\mathbb{E}[\phi(U) \mid f(U) = y] = \int_{f^{-1}(\{y\})} \phi(u) k_y(u) \mathcal{H}^{n-m}(du)$$

where $k_y(u)$ is the density of $\mathrm{K}(y, \cdot)$ on $f^{-1}(\{y\})$ with respect to $\mathcal{H}^{n-m}$, given by

$$k_y(u) \propto p_U(u) \left|\det J_f(u) J_f(u)^\top\right|^{-1/2}.$$

When is $f^{-1}(y)$ a submanifold?

Let $\mathcal{X}$ and $\mathcal{Y}$ be manifolds and $f : \mathcal{X} \to \mathcal{Y}$ be smooth.

Regular Value

Then $y \in \mathcal{Y}$ is a **regular value** for f if for all $x \in f^{-1}(y)$ the differential $df_x : \mathcal{T}_x\mathcal{X} \to \mathcal{T}_y\mathcal{Y}$ is surjective. (alternatively, f is a submersion at every $x \in f^{-1}(y)$).

Preimage Theorem

If $y \in \mathcal{Y}$ is a regular value of f then $f^{-1}(y)$ is a submanifold of $\mathcal{X}$.

Filamentary Distributions

Assumption 1

Manifold of interest has co-dimension 1.

Filamentary Distribution

Let $X : \Omega \to \mathbb{R}^n$ be a vector-valued random variable with distribution π and finite covariance matrix $\mathbb{V}_\pi[X]$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth function. Consider $y \in \mathbb{R}$ fixed, then at any point $\xi \in f^{-1}(y)$ we denote by $\hat{g}(\xi)$ the normalized gradient of f and by $\mathsf{T}(\xi) = \{\hat{t}_1(\xi), \ldots, \hat{t}_{n-1}(\xi)\}$ a basis for the tangent space at ξ. We say that π is a filamentary distribution if

$$\forall \xi \in \mathbb{R}^n, \quad \forall \hat{t}_i(\xi) \in \mathsf{T}(\xi) \qquad 0 < \hat{g}(\xi)^\top \mathbb{V}_\pi[X]\hat{g}(\xi) \ll \hat{t}_i(\xi)^\top \mathbb{V}_\pi[X]\hat{t}_i(\xi).$$

In practice one doesn't need to check the definition, it will be clear if the posterior has a filamentary structure.

- Filamentary distributions are highly concentrated around a submanifold.
- Orthogonal scaling $\ll$ tangential scaling.