MATHÉMATIQUES
VISION
APPRENTISSAGE

Riemanniann Manifold Hamiltonian Monte Carlo

Samuel Bazaz, Mathis Hardion

Plan

1. Context

2. Experiments

3. Conclusion

HMC and RHMC : a better exploration!

Geodesic flows

- Hamiltonian equations : modeling motion dynamics
- Hamiltonian = kinetic energy + potential energy
- RHMC: generalization with non-constant curvature

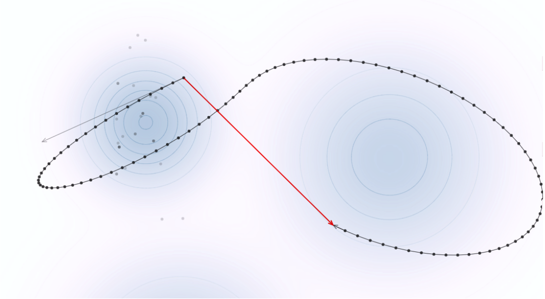


FIGURE 1: HMC from Chi Feng demo

MATHÉMATIQUES VISION APPRENTISSAGE

Questions raised by the document studied

Girolami and Calderhead's paper areas of improvements [7]

- A better integration scheme ?
- A better metric ?
- A numerical estimation of the curvature?

Theorem: two theoretical results

- Verlet's leapfrog is sympletic and assures time reversibility which implies detailed balance [4]
- The geodesic flow for negatively curved compact Riemannian manifolds is ergodic [1]

Three improvements already made

Three interesting papers

- Autodifferentiation for Bayesian neural network[6]
- No-U-Turn (NUTS) : The No-U-Turn Sampler [8]
- Softabs metrics[3] : " maintain the desirable behavior of the Hessian in convex neighborhood"

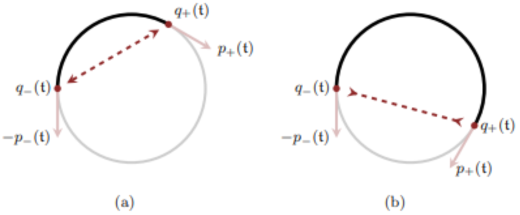


FIGURE 2: NUTS from 'A Conceptual Introduction to Hamiltonian Monte Carlo' [2]

MATHÉMATIQUES VISION APPRENTISSAGE

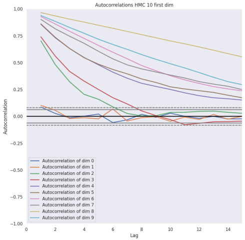
Plan

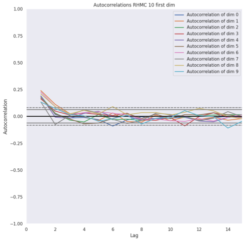
Automatic differentiation and HMC vs RHMC

A trade-off between speed and precision

- Autodifferentiation makes RHMC slower and needs explicit solutions (x50 to x1000).
- Autocorrelations are far better for RHMC (no small oscillations).
- RHMC sampling has smaller KL divergence ($0.3 < 1670$)

HMC vs RHMC 200D Gaussian distribution



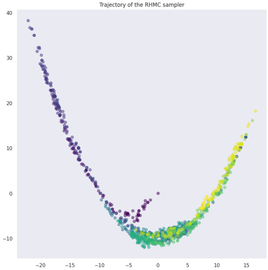


Softabs

More stable but needs better implementation

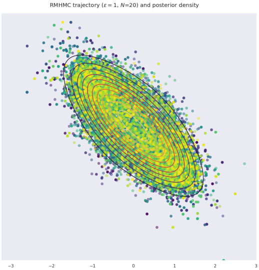
- Too slow to be used with implicite integration.
- Assure SDP stability.
- Should use a specialised framework, STAN ?

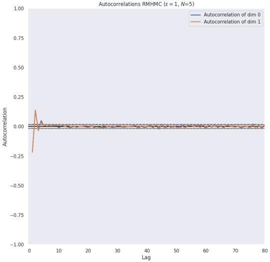
HMC vs RHMC Softabs

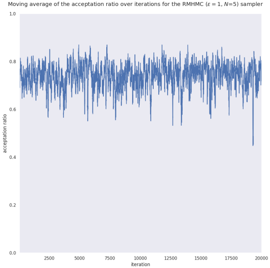


Hyperparameter influences

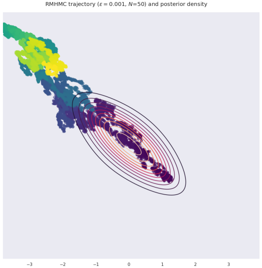
$\varepsilon = 1,$
$N = 5$

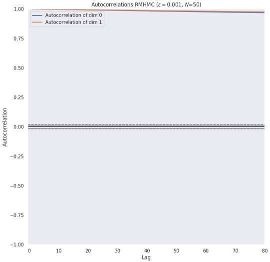


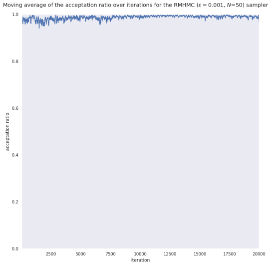




$\varepsilon = 10^{-3},$
$N = 50$

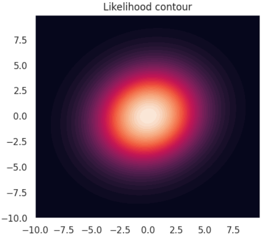


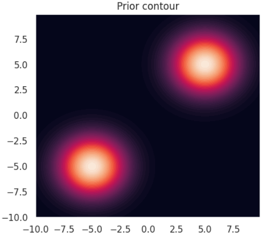




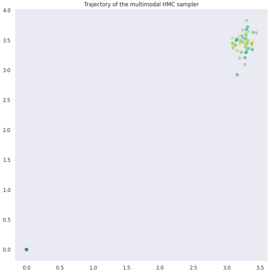
Multimodal distributions for RHMC

Gaussian mixture (RHMC)



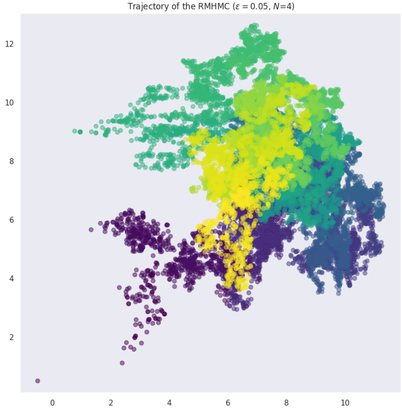


HMC



Adding the likelihood to the potential

- $\mathcal{L}(\theta) \rightarrow \mathcal{L}(\theta, Y) = \mathcal{L}(\theta) + \mathcal{L}(Y|\theta)$



- Does not help explore other modes in this case ...

Plan

Conclusion & extensions

- Easily implemented and can work really well, but hyperparameter tuning is not simple, done manually in the paper

- Derivation of the (expected) Rao metric is intractable in some cases (e.g. mixture models)
  - Approximation must be used or other metrics chosen
  - The problem of the right choice of matrix is not entirely solved although it gives very good directives

- Could be coupled with parallel tempering methods to better handle the multimodal case

- Alternative geometries could be investigated, for instance with MMD [5] or Wasserstein [9] distances

References

[1] D V Anosov. *Geodesic flows on closed Riemann manifolds with negative curvature*. Providence: American Mathematical Society Providence, 1969, iv, 235 pages 26 cm.

[2] Michael Betancourt. *A Conceptual Introduction to Hamiltonian Monte Carlo*. 2018. arXiv: 1701.02434 [stat.ME].

[3] Michael Betancourt. "A General Metric for Riemannian Manifold Hamiltonian Monte Carlo". In: *Geometric Science of Information*. Springer Berlin Heidelberg, 2013, pp. 327–334.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006, pp. 548–554.

[5] Francois-Xavier Briol et al. *Statistical Inference for Generative Models with Maximum Mean Discrepancy*. 2019. arXiv: 1906.05944 [stat.ME].

[6] Adam D Cobb and Brian Jalaian. "Scaling Hamiltonian Monte Carlo Inference for Bayesian Neural Networks with Symmetric Splitting". In: *Uncertainty in Artificial Intelligence* (2021).

[7] Mark A. Girolami and Ben Calderhead. "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (2011).

[8] Matthew D. Hoffman and Andrew Gelman. *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*. 2011. arXiv: 1111.4246 [stat.CO].

[9] Wuchen Li and Jiaxi Zhao. *Wasserstein information matrix*. 2020. arXiv: 1910.11248 [math.ST].

The curse of dimensionality

Measure concentration

- Motivates Bayesian inference
  - Integration instead of Optimization
  - Monte Carlo estimation
- Motivates manifold hypothesis

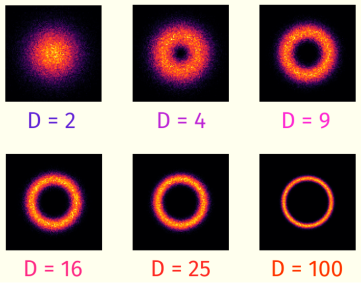


FIGURE 3: Gaussian sampling projections from Jean Feydy course

NUTS HMC using STAN

STAN a hard but powerful framework

- Good performance but bad documentation
- Autograd and RHMC are not upgraded in python yet.
- Needs to create the cython script by yourself for now.

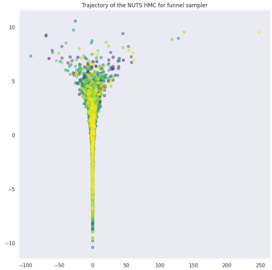


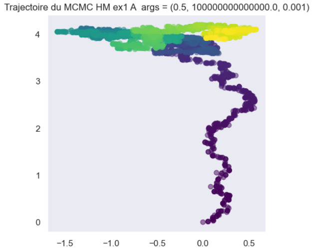
FIGURE 4: NUTS HMC for funnel's distribution

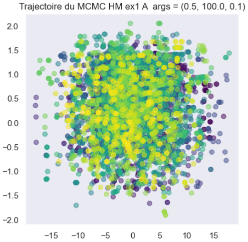
What is an efficient sampling ?

Sampling techniques

- Use of dependencies : MCMC
- Exploration-exploitation dilemma
    - Exploration to have all the ”typical” states
    - Exploitation to obtain the right proportions

Exploration Exploitation Metropolis Hastings TP4

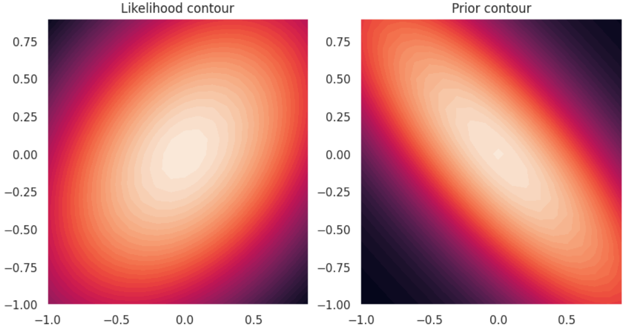




MATHÉMATIQUES VISION APPRENTISSAGE

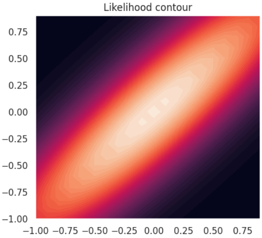
Toy problem

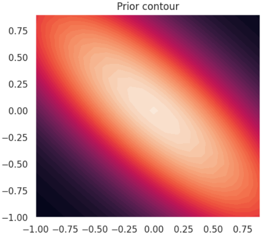
- $Y|\theta \sim \mathcal{N}\left(\frac{1}{2}\theta^{\odot 2}, \Sigma_0\right)$ ($\odot$: Hadamard product)
- Prior $\theta \sim \mathcal{N}(0, \Sigma)$
- Fisher-Rao metric tensor $G(\theta) = \mathrm{diag}(\theta)\Sigma_0^{-1}\mathrm{diag}(\theta) + \Sigma^{-1}$

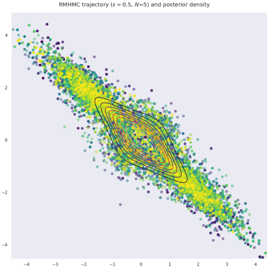


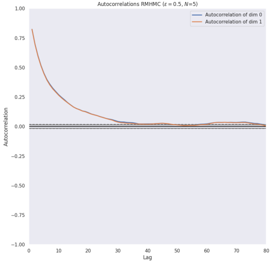
MATHÉMATIQUES VISION APPRENTISSAGE

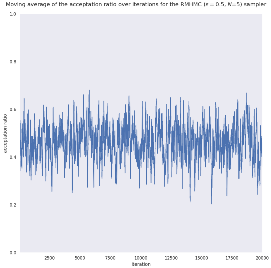
Experiments using our own implementation











MATHÉMATIQUES VISION APPRENTISSAGE

Multimodal distributions for HMC

HMC

- HMC can be computed but doesn't explore enough.

HMC 2D Gaussian Mixture and Banana-shape