HyperSphere Markov Chain Monte Carlo

Emanuele Chiarini
Advisor: Giacomo Zanella

December 15, 2023

Outline

Introduction

- Gradient-based Markov Chain Monte Carlo (MCMC) algorithms leverage the gradient information of the target distribution to generate samples from a distribution of interest more efficiently.
- Despite their general effectiveness, Gradient-based MCMC methods tend to be less robust to heterogeneity of scales across dimensions of the target distribution and in cases of exploding gradients.
- The thesis focuses on developing a new Gradient-based MCMC algorithm that decouples the magnitude of the step-size, which remains fixed, from the gradient while still employing the latter to inform the move.

Framework

- Zanella, 2020 proposes a framework for constructing locally balanced proposals in the context of MCMC, as:

$$Q_{g,\sigma}(x, dy) = \frac{g\left(\frac{\pi(y)}{\pi(x)}\right)\mu_\sigma(y-x)dy}{Z_g(x)}$$

  where g is a continuous function mapping $[0, \infty)$ to itself and μ_σ is the uninformed symmetric kernel used to generate proposals in a RWM scheme.

- This framework is able to inherit the topological structure of μ_σ and to incorporate information about the target through the multiplicative term $g\left(\frac{\pi(y)}{\pi(x)}\right)$.

- In continuous state spaces $\mathbb{R}^n$ it is usually not possible to sample efficiently from a locally balanced proposal as the normalizing constant Z_g is typically intractable.
- A solution is to replace the intractable term $\pi(y)$ (the target distribution) in $g\left(\frac{\pi(y)}{\pi(x)}\right)$ with the first order Taylor expansion of $\log \pi(y)$ at x : $e^{\nabla \log \pi(x) \cdot (y-x)}$.
- This leads to a definition of a family of first order locally balanced proposals following:

$$Q_{g,\sigma}(x, dy) \propto g\left(e^{\nabla \log \pi(x) \cdot (y-x)}\right) \mu_\sigma(y-x)dy$$

MALA's proposal can also be defined through this framework.

Derivation

- Our purpose in the context of the locally balanced framework is to find a way to disentangle the direction, given by the gradient, from the step-size.
- To do so, we could define a proposal able to sample from the surface of a hypersphere centered at the current value and with fixed radius σ, in the direction of $\nabla \log \pi(x)$.
- In the context of first order locally balanced proposals we take:

$$g(t) = \sqrt{t}$$

$$\mu_\sigma(x) = \frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{\frac{n}{2}}\sigma^{n-1}} \mathbb{I}(\|\mathbf{x}\| = \sigma)$$

where μ_σ is the uniform distribution over the n-dimensional hypersphere with radius σ, which is symmetric.

- Setting $t = e^{\nabla \log \pi(x)(y-x)}$ yields $g(t) = e^{\frac{\nabla \log \pi(x)(y-x)}{2}}$. The integral $Z(x)$ is thus:

$$Z(x) = \int e^{\frac{\nabla \log \pi(x) z}{2}} \frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{\frac{n}{2}} \sigma^{n-1}} \mathbb{I}(\|\mathbf{z}\| = \sigma) dz$$

- To evaluate the integral we resort to the change of variable $z_i = \sigma w_i$ where each i refers to the dimensions of vector z:

$$\int_{\mathbb{I}(\|\mathbf{z}\|=\sigma)} e^{\frac{\nabla \log \pi(x) z}{2}} dz = \int_{\mathbb{I}(\sigma\|\mathbf{w}\|=\sigma)} e^{\frac{\nabla \log \pi(x) \sigma w}{2}} \sigma^n dw$$

- Thanks to this change of variable, it is possible to recognize in the integral, the kernel of the von Mises-Fisher distribution on the $(n-1)$-hypersphere: $S^{n-1} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$.

- The probability density function of the von Mises-Fisher distribution is given by:

$$f_n(x; \mu, \kappa) = C_n(\kappa) \exp\left(\kappa \mu^T x\right)$$

where $\kappa \geq 0$ is the concentration parameter and $\|\mu\| = 1$ is the direction parameter.

- The normalization constant $C_n(\kappa)$ is equal to the following:

$$C_n(\kappa) = \frac{\kappa^{n/2-1}}{(2\pi)^{n/2} I_{n/2-1}(\kappa)},$$

where $I_{n/2-1}$ denotes the modified Bessel function of the first kind of order $\frac{n}{2} - 1$.

- We need μ to have norm equal to 1 so we multiply and divide by the norm of $\nabla \log \pi(x)$:

$$\int_{\mathbb{I}(\sigma\|\mathbf{w}\|=\sigma)} e^{\frac{||\nabla \log \pi(x)||\sigma}{2} \cdot \frac{\nabla \log \pi(x)}{||\nabla \log \pi(x)||} \cdot w} \sigma^n dw = \sigma^n \int_{\mathbb{I}(\|\mathbf{w}\|=1)} e^{\kappa \mu^T w} dw$$

where $\kappa = \frac{||\nabla \log \pi(x)||\sigma}{2}$ and $\mu = \frac{\nabla \log \pi(x)}{||\nabla \log \pi(x)||}$.

- Being the kernel of a von Mises-Fisher distribution, it integrates to the inverse of the normalizing constant in the distribution pdf:

$$C_n^{-1}(\kappa) = \frac{(2\pi)^{n/2} I_{n/2-1}(\kappa)}{\kappa^{n/2-1}}$$

- Finally, we get a closed-form solution for the normalizing constant of the locally balanced first order informed proposal:

$$Z(x) := \frac{\sigma \Gamma(\frac{n}{2}) I_{n/2-1}(\kappa)}{\kappa^{n/2-1}}$$

- Finally, we obtain the proposal distribution for the HyperSphere algorithm:

$$Q_\sigma^{vMF}(z) = \frac{e^{\frac{\nabla \log \pi(x) z}{2}} \mathbb{I}(\|\mathbf{z}\| = \sigma)}{\sigma^n \frac{(2\pi)^{n/2} I_{n/2-1}(\kappa)}{\kappa^{n/2-1}}}$$

- In the pdf of the proposal distribution that we have identified, we cannot recognize any standard distribution. To solve this problem, we can perform the transformation of variable $W = g(Z) = \frac{Z}{\sigma}$.
- We thus obtain the pdf for random vector W:

$$Q_\sigma^{vMF}(w) = \frac{e^{\frac{||\nabla \log \pi(x)||\sigma}{2} \cdot \frac{\nabla \log \pi(x)}{||\nabla \log \pi(x)||} \cdot w} \mathbb{I}(\|\mathbf{w}\| = 1)}{\frac{(2\pi)^{n/2} I_{n/2-1}(\kappa)}{\kappa^{n/2-1}}}$$

in which we recognize the pdf of a von Mises-Fisher distribution. Therefore, we can sample from the original proposal distribution $Q_\sigma^{vMF}(z)$ by sampling w from $Q_\sigma^{vMF}(w)$ and multiplying by σ.

HyperSphere algorithm

Algorithm 1 HyperSphere algorithm

Require: n iterations, variance σ, x_0, functions $\log \pi(\cdot)$ and $\nabla \log \pi(\cdot)$.

$x \leftarrow x_0$
for $i = 1$ to n **do**
  $\mu_x,\ \kappa_x \leftarrow \frac{\nabla \log \pi(x)}{\|\nabla \log \pi(x)\|},\ \frac{\sigma}{2} \cdot \|\nabla \log \pi(x)\|$
  $y \leftarrow x + \sigma \cdot vMF(\mu_x, \kappa_x)$ ▷ von Mises-Fisher sampling
  $\mu_y,\ \kappa_y \leftarrow \frac{\nabla \log \pi(y)}{\|\nabla \log \pi(y)\|},\ \frac{\sigma}{2} \cdot \|\nabla \log \pi(y)\|$
  $\log(a) \leftarrow \log \pi(y) - \log \pi(x) + LogProp(\sigma, y, x, \nabla \log \pi(\cdot), \kappa_y, \kappa_x)$
  $a \leftarrow \min(1, e^{\log(a)})$
  if $u \sim U(0,1) < a$ **then**
    $x \leftarrow y$
  end if
  $x_n \leftarrow x$
end for

Ensure: $(x_1, \ldots, x_n)$

Algorithm 2 Helper for the Log Proposal

Require: σ, proposal y, current value x, function $\nabla \log \pi(\cdot)$, κ_y, κ_x.

1: $\delta_x \leftarrow \sigma \cdot \frac{x-y}{\|x-y\|}$

2: $\delta_y \leftarrow \sigma \cdot \frac{y-x}{\|y-x\|}$

3: $logQ_1 \leftarrow \frac{1}{2}\Big(\nabla \log \pi(y)\delta_x - \nabla \log \pi(x)\delta_y\Big)$

4: $logQ_2 \leftarrow \left(\frac{n}{2} - 1\right) \log\left(\frac{\kappa_y}{\kappa_x}\right) - \log \frac{I_{n/2-1}(\kappa_y)}{I_{n/2-1}(\kappa_x)}$

Ensure: $logQ_1 + logQ_2$

- The computation of δ_x and δ_y follows from the need to have the difference between proposal and current value to lie on the support of the proposal distribution $Q_\sigma^{vMF}(z)$.

Sampling from a von Mises-Fisher distribution

- Key detail necessary for the Hypersphere MCMC algorithm is a computationally efficient way to sample from a von Mises-Fisher distribution.
- Wood, 1994 proposed an algorithm for sampling from a von Mises-Fisher distribution with our κ of interest but direction $\mu_0 = (0, \dots, 0, 1)$, followed by a rotation step in the direction of interest using a QR decomposition. However, this is computationally inefficient.
- Pinzón and Jung, 2023 leverages the tangent normal decomposition to skip the rotation step, yielding a sampling algorithm that performs in the same order of magnitude of sampling from a $N(0, I)$.

Modified Bessel function of the first kind

- Log-acceptance step involves the computation of the Modified Bessel Function of the First Kind. This function, growing or decaying exponentially, can pose numerical stability problems.
- HyperSphere MCMC Python implementation solves this issue by computing the log of the ratio of $I_\nu(x)$ using the *mpmath* library which affords an higher numerical precision.
- We compute $I_\nu(x)$ through the integral form representation, available for integer ν:

$$I_\nu(x) = \frac{1}{\pi}\int_0^\pi \cos(\nu\theta)\exp(x\cos\theta)d\theta$$

Optimal Step Size

- The Expected Squared Jump Distance (ESJD) evaluates the efficiency of MCMC algorithms, by assessing how well a Markov Chain explores the state space.
- It is computed by calculating the expectation of the Euclidean distance between two consecutive points in the chain: $ESJD = E_T\left[\|x_t - x_{t-1}\|_2^2\right]$
- HyperSphere is characterized by a fixed step-size, while in MALA the gradient affects the amount by which the chain moves.
- However, it is possible to establish a connection between the two, as we would expect that the optimal step-size for both would lead to $ESJD^{MALA} \approx ESJD^{HS}$.

- In the HyperSphere case:

$$E_T\left[\|y-x\|_2^2\right] = E_T\left[\|\sigma_{HS}\cdot vMF(\mu_x,\kappa_x)\|_2^2\right] = \sigma_{HS}^2$$

Thanks to the fact that $vMF(\mu_x,\kappa_x)$, by definition, will have Euclidean norm equal to 1.

- In the MALA case we have:

$$E_T\left[\|y-x\|_2^2\right] = E_T\left[\|\frac{\sigma_{MALA}^2}{2}\nabla\log\pi(x)+\sigma_{MALA}Z\|_2^2\right]$$

Taking $\nabla\log\pi(x)=0$ yields:

$$\sigma_{MALA}^2 E_T\left[\|Z\|_2^2\right] = \sigma_{MALA}^2 E_T\left[z_1^2+\ldots+z_d^2\right] = \sigma_{MALA}^2\cdot d$$

- Assuming that $ESJD^{MALA}\approx ESJD^{HS}$ holds, we can use this heuristic to connect the optimal σ_{MALA} and σ_{HS} as $\sigma_{HS}\approx\sigma_{MALA}\cdot\sqrt{d}$.

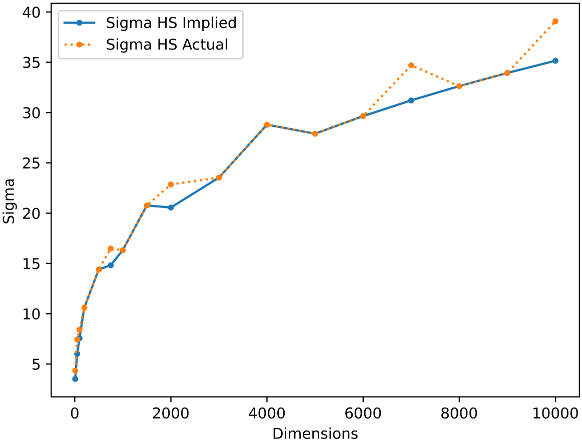
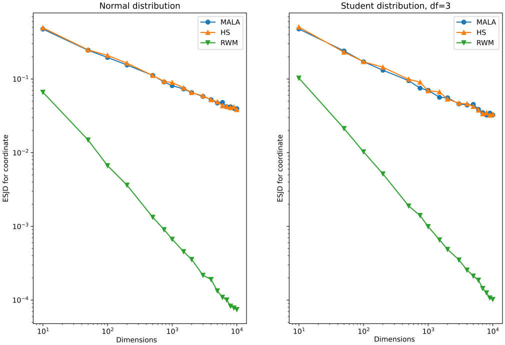


Figure: To verify empirically this heuristic, we leverage simulations of MALA and HyperSphere with a fixed σ, in the context of sampling from a d-dimensional multivariate standard normal, saving the σ maximizing ESJD.

Efficiency on Isotropic Targets

- Comparison of the ESJD for HyperSphere, MALA and RWM algorithms. The target distributions considered are constituted by independent and identically distributed (i.i.d) components.
- We can conclude that the rate of decay is similar to MALA, of order $d^{-1/3}$.



Heterogeneity of Scales

- Adaptive Markov Chain Monte Carlo (MCMC) automatically adjusts the proposal distribution based on the history of the chain to achieve an optimal acceptance rate.
- The target is a multivariate Normal distribution, with a diagonal variance-covariance matrix. The standard deviation of each coordinate increases linearly from 0.01 to 1. Therefore, we have for each dimension $i = 1, ..., 100$, scale $\eta_i = 0.01 \cdot i$.
- MALA and RWM are further enhanced by a pre-conditioning scheme, useful for targets with heterogeneous scales.
- The metrics we use are the Mean Squared Error (MSE) and $D(t)$ that serves as a proxy for the difference between the empirical var-covar matrix and the true variance-covariance matrix.

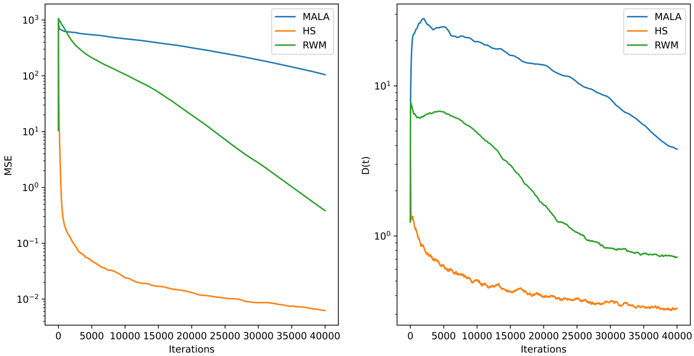


Figure: The evolution of both $D(t)$ and the MSE of the HyperSphere algorithm is superior to the respective results for both MALA and RWM, despite the last two employing a pre-conditioning matrix to speed up adaption to the heterogeneity of each dimension. Therefore, HyperSphere MCMC may be promising for applications involving targets with heterogeneous scales.

Light Tails

- Distributions with light tails can be another area of interest for HyperSphere, as in the tails the gradient often becomes extremely large in magnitude, affecting the stability of gradient-based algorithms.
- Thanks to the decoupling of gradient and step-size in HyperSphere, it can prove superior in this context characterized by exploding gradients over other algorithms such as MALA.
- The target of the simulation is a 100-dimensional Generalized Normal distribution with $\beta = 7$. The pdf follows:

$$p(x; \mu, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|x-\mu|/\alpha)^{\beta}}$$

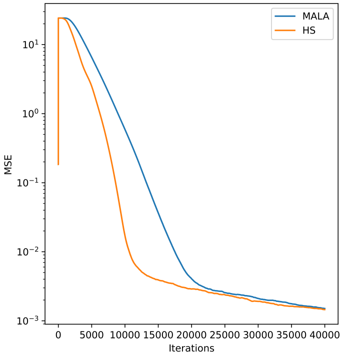


Figure: The results displayed in the figure seem to suggest that HyperSphere is able to achieve convergence to a stationary state faster than MALA for this target distribution. The pattern is similar for both the evolution of the MSE and of $D(t)$.

Conclusion

- The aim of HyperSphere MCMC is to leverage the information of the gradient and decouple the step-size from the direction of the move.
- Through the framework of first order locally balanced proposals, we verify that it is possible to construct a proposal of this kind. We confirm, via numerical simulation, that HS MCMC preserves the efficiency of MALA on targets in high dimensional spaces.
- Furthermore, after empirically verifying the computational cost of the algorithm, we concluded it remains in the same order of magnitude as MALA.
- We highlight two cases, one with heterogeneity of scales and one with light tails, in which the HyperSphere algorithm shows some promise over MALA and RWM.

Thank you!

References I

- Livingstone, Samuel and Giacomo Zanella (2022). "The Barker proposal: Combining robustness and efficiency in gradient-based MCMC". In: *Journal of the Royal Statistical Society. Series B, Statistical methodology* 84.2, pp. 496–523. DOI: 10.1111/rssb.12482. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12482.
- Mardia, Kanti V. and Peter E. Jupp (1999). *Directional Statistics*. 1st. John Wiley & Sons Ltd. ISBN: 978-0-471-95333-3.
- Neal, Radford M. (2010). "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo*. Vol. 54, pp. 113–162.
- Pinzón, Carlos and Kangsoo Jung (2023). *Fast Python sampler for the von Mises Fisher distribution*. Tech. rep. URL: https://hal.science/hal-04004568.

References II

Temme, N.M (1976). "On the numerical evaluation of the ordinary bessel function of the second kind". In: *Journal of Computational Physics* 21.3, pp. 343–350. ISSN: 0021-9991. DOI: https://doi.org/10.1016/0021-9991(76)90032-2. URL: https://www.sciencedirect.com/science/article/pii/0021999176900322.

Wood, Andrew T. A. (1994). "Simulation of the von mises fisher distribution". In: *Communications in statistics. Simulation and computation* 23.1, pp. 157–164. DOI: 10.1080/03610919408813161. URL: https://www.tandfonline.com/doi/abs/10.1080/03610919408813161.

Zanella, Giacomo (2020). "Informed Proposals for Local MCMC in Discrete Spaces". In: *Journal of the American Statistical Association* 115.530, pp. 852–865. DOI: 10.1080/01621459.2019.1585255. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.2019.1585255.