ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Markov Chain Monte Carlo Sampling on the Stiefel Manifold

Master Thesis

Maxime Scali

July 24, 2024

Advisors: Prof. S. Bacallado, Prof. A. Kraus
Department of Mathematics, ETH Zürich

Abstract

This Master thesis studies Markov Chain Monte Carlo (MCMC) methods on the Stiefel manifold, a compact Riemannian manifold of orthonormal matrices. We present novel algorithms that efficiently explore this space, addressing the challenges arising from its non Euclidean geometry. We discuss and analyse the different classes of invariant kernel we introduce, and we compare our methods with traditional MCMC algorithms. The proposed algorithms show improved computational efficiency for similar sampling performance to existing methods, particularly for high-dimensional problems.

.

Contents

Contents ii

Introduction 1

1 Stiefel Manifold 3
- 1.1 Definition and tangent space . . . 3
  - 1.1.1 Closed form of the tangent space from the Euclidean inner product . . . 4
  - 1.1.2 Derivation of the tangent space from quotient definition . . . 5
- 1.2 Canonical metric . . . 6
- 1.3 Geodesics in $O(n)$ and $\mathcal{V}_{k,n}$ . . . 7
- 1.4 Givens matrices . . . 7
- 1.5 Invariant measures on $\mathcal{V}_{k,n}$ . . . 10

2 MCMC 12
- 2.1 Markov Kernel . . . 12
- 2.2 Convergence . . . 13
- 2.3 Metropolis-Hastings kernels . . . 16
- 2.4 A general perspective on Metropolis-Hastings kernels . . . 18
- 2.5 Non reversible Markov kernels . . . 19
- 2.6 HMC . . . 21
- 2.7 NUTS . . . 22

3 MCMC on the Stiefel manifold 23
- 3.1 Polar expansion . . . 23
- 3.2 MultiG kernel . . . 25
- 3.3 Nonreversible kernel . . . 30
- 3.4 Improvement directions . . . 34

Conclusion 35

A Appendix **36**
A.1 Radon-Nikodym derivative 36
A.2 Invariant kernels with partial refreshment 37
A.3 Polar decomposition 38
A.4 MCMC kernel acceptance ratios 39

Bibliography **41**

Introduction

In 1953, [15] Metropolis et al. noticed that we can simulate the equilibrium distribution of a dynamical system of interest instead of the system itself This marks the start of Markov Chain Monte Carlo (MCMC) methods. In 1970 and built on this work, Hastings introduced in [9] a formal framework to simulate distributions, and the resulting algorithms introduced are called Metropolis-Hastings algorithms. Rapidly, MCMC became a fundamental tool in machine learning and computational statistics when one needs to sample from a complex distribution known up to a multiplicative constant. Such problems appear in Bayesian inference or generative modeling.

The main challenge in designing an MCMC algorithm is how one can accelerate the mixing time, i.e. how one can adapt the algorithm to the state space $\mathcal{X}$ and to the target distribution π so that the Markov chain $X_n \sim pi$ with the fewest number of simulation steps. A lot of work has been done to assess the quality of the samples [7], but also to find upper bounds and reduce the mixing time of the generated Markov chain. A common direction of work reverse the original creation process of MCMC and identifies dynamical processes such as PDE/SPDEs which admits the desired invariant measure, and then simulate those processes to form tractable chains (e.g [18]). On the other hand, Power et al. [20] in the discrete case make use of the symmetries and algebraic structure that are usually available in the state space to accelerate the sampling. One can wonder how this can be extended to some continuous cases.

In this work, we use the specific properties of the Stiefel manifold to define MCMC invariant kernels. The Stiefel manifold $\mathcal{V}_{k,n}$ consists of all orthonormal matrices of the size $n \times k$ for $k \leq n$. This manifold is used in statistical modeling for a long time, and has gained popularity recently as orthonormal constraints on deep learning parameters provides significant advantages, as explained in [14]. As the Stiefel manifold is a submanifold of $\mathbb{R}^{n\times k}$, it is possible to sample MCMC in $\mathbb{R}^{n\times k}$ and constrain the samples to be on the Stiefel manifold, but usually this constraining step is computationally expensive with respect to the manifold dimensions. In this

work we develop MCMC kernels that achieve state of the art performance with respect to the computational time. In section 1, we introduce the Stiefel manifold and study its properties. In section 2, we introduce the MCMC setup and some important results, and we define Metropolis-Hastings from different perspectives. In section 3 we define some MCMC kernels and analyse their performance.

Chapter 1

Stiefel Manifold

This section introduces the Stiefel manifold and its tangent space. We also introduce Givens matrices in 1.4 and explain how they are related to the Stiefel manifold. Givens matrices will play a central role in 3. Lastly, we briefly introduce in 1.5 how to construct the invariant measure on the Stiefel manifold, which will be useful to properly define probability distributions on the Stiefel manifold.
We assume basic knowledge of differential geometry, we refer to [13] for a complete and formal introduction to differential geometry.

1.1 Definition and tangent space

Let $k, n \in \mathbb{N}^*, k \leq n$. The Stiefel manifold, denoted by $\mathcal{V}_{k,n}$, is the space of k-frames in $\mathbb{R}^n$ where a k-frame is identified with k orthogonal vectors in $\mathbb{R}^{n\times n}$. Formally,

$$\mathcal{V}_{k,n} = \left\{ X \in \mathbb{R}^{n\times k} \mid X^T X = I_k \right\}$$

Equation $X^T X = I_k$ can be seen as introducing a total of $\frac{k(k+1)}{2}$ independent conditions on the elements of X. We deduce that the Stiefel manifold $\mathcal{V}_{k,n}$ has dimension

$$\dim(\mathcal{V}_{k,n}) = nk - \frac{k(k+1)}{2} \tag{1.1}$$

Some special cases of the Stiefel manifold $\mathcal{V}_{k,n}$ include the group $O(m)$ of orthogonal matrices when $k = m$, or the unit sphere $\mathbb{S}^m$ when $k = 1$. It is compact Riemannian manifold with different choice of metric as discussed in Section 1.2.

From the equation $X^T X = I_k$, we get that an element Δ_X of the tangent space $T_X\mathcal{V}_{k,n}$ at X must verify

$$\Delta_X^T X + X^T \Delta_X = 0 \tag{1.2}$$

i.e. $\Delta^T X$ is skew-symmetric. Again, equation 1.2 imposes $\frac{k(k+1)}{2}$ independent conditions on Δ, so we can check that the dimension of $T_X\mathcal{V}_{k,n}$ for every $X \in \mathcal{V}_{k,n}$

has dimension

$$nk - \frac{k(k+1)}{2}$$

as expected.

1.1.1 Closed form of the tangent space from the Euclidean inner product

As we will see, there is no unique natural choice of inner product on the Stiefel manifold. An inner product defined on the tangent space is necessary to see the Stiefel manifold as a Riemannian manifold and to define a notion of distance on $\mathcal{V}_{k,n}$. In this section we follow the argument of Section 2.2.1 of [4] to determine a closed form of the tangent space.

We view the Stiefel manifold as an embedded manifold of the Euclidean space $\mathbb{R}^{n\times k}$, so we choose the standard np-dimensional euclidean inner product

$$\langle \Delta_1, \Delta_2 \rangle_e = \operatorname{tr} \Delta_1^T \Delta_2$$

for $\Delta_1, \Delta_2 \in \mathbb{R}^{n\times k}$

Let $\mathcal{S}_n$ and $\mathcal{SK}_n$ denote the space of $\mathbb{R}^{n\times n}$ symmetric and skew-symmetric matrices respectively. These spaces have dimensions

$$\dim(\mathcal{S}_n) = \frac{k(k+1)}{2} \qquad \dim(\mathcal{SK}_n) = \frac{k(k-1)}{2}$$

Definition 1.1.1. *The normal space $N_X\mathcal{V}_{k,n}$ at $X \in \mathcal{V}_{k,n}$ is the space of matrices $N \in \mathbb{R}^{n\times k}$ such that*

$$\langle \Delta, N \rangle_e = 0 \tag{1.3}$$

for all Δ in $T_X\mathcal{V}_{k,n}$.

The dimension of the normal space at X is therefore

$$\dim(\mathbb{R}^{n\times k}) - \dim(T_X\mathcal{V}_{k,n}) = \frac{k(k+1)}{2}$$

Here is a characterisation of the normal space :

Lemma 1.1.2. *Let $X \in \mathcal{V}_{k,n}$, we have*

$$N_X\mathcal{V}_{k,n} = XS$$

where $S \in \mathcal{S}_k$.

Proof. Let $N = XS$ with $S \in \mathcal{S}_k$. Since the trace of the product of a skew-symmetric and a symmetric matrix is equal to 0, using 1.2 we see that $N \in N_X\mathcal{V}_{k,n}$. Furthermore, matrices of the form $N = XS$ with $S \in \mathcal{S}_k$ define a space of dimension $\dim(\mathcal{S}_k) = \frac{k(k+1)}{2} =$, which concludes the proof. □

Let the symmetric and skew-symmetric projections be

$$\begin{aligned} &\text{sym} : \mathbb{R}^{n\times n} \to \mathcal{S}_n && \text{skew} : \mathbb{R}^{n\times n} \to \mathcal{SK}_n \\ &\text{sym}(A) = \frac{A + A^T}{2} && \text{skew}(A) = \frac{A - A^T}{2} \end{aligned}$$

.
The map

$$\pi_N(Y) = X\text{sym}(X^T Y)$$

defines a projection on $N_X\mathcal{V}_{k,n}$. Similarly, the map

$$\pi_T(Y) = X\ \text{skew}(X^T Y + (I - XX^T)Y) \tag{1.4}$$

defines a projection onto $T_X\mathcal{V}_{k,n}$, (it can be checked that $\langle \pi_N(Y), \pi_T Y\rangle = 0$). From 1.4 we see that tangent directions $\Delta \in T_X\mathcal{V}_{k,n}$ have the general form

$$\Delta = XK + (I_n - XX^T)Y \tag{1.5}$$

where $K \in \mathbb{R}^{k\times k}$ is skew-symmetric and $Y \in \mathbb{R}^{n\times k}$ arbitrary. The term $(I_n - XX^T)Y$ can be replaced by $X_\perp Y'$, where $Y' \in \mathbb{R}^{n-k\times k}$ is arbitrary and $X_\perp$ verifies

$$XX^T + X_\perp X_\perp^T = I_n$$

We indeed see that $\frac{k(k-1)}{2}$ dimensions of the tangent space are parameterised by K and $k(n-k)$ by Y'.

1.1.2 Derivation of the tangent space from quotient definition

The Stiefel manifold is an embedded submanifold of the orthogonal group $O(n) \subset \mathbb{R}^{n\times n}$. In the following sections we we will extensively use this fact to derive properties of $\mathcal{V}_{k,n}$ from $O(n)$. We now present $\mathcal{V}_{k,n}$ as an embedding of $O(n)$ and derive its tangent space using this new characterisation.
Let $k < n$, we consider the action of $O(n-k)$ on $O(n)$ given by the right multiplication

$$(H, G) \mapsto G \begin{pmatrix} I_k & 0 \\ 0 & H \end{pmatrix}$$

where $G \in O(n)$ and $H \in O(n-k)$. We see that there is a one-to-one correspondence between the orbits of this action and $\mathcal{V}_{k,n}$, hence

$$\mathcal{V}_{k,n} \cong O(n)/O(n-k)$$

In the rest of this section we will therefore refer to an element of $\mathcal{V}_{k,n}$ as $[G] = G \begin{pmatrix} I_k & 0 \\ 0 & O(n-k) \end{pmatrix}$.

For $G \in O(n)$, let

$$\mathcal{H}_G = G \begin{pmatrix} A & -B^t \\ B & 0 \end{pmatrix}$$

$$\mathcal{V}_G = G \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$$

with $A \in \mathbb{R}^{k\times k}$ and $C \in \mathbb{R}^{(n-k)\times(n-k)}$ skew-symmetric and $B \in \mathbb{R}^{(n-k)\times k}$ arbitrary. We easily see that $\mathcal{H}_G, \mathcal{V}_G \subset T_G O(n)$ and

$$T_G O(n) = \mathcal{H}_G \oplus \mathcal{V}_G$$

$\mathcal{H}_G, \mathcal{V}_G$ are called the horizontal and vertical space respectively. We choose specifically this decomposition of the tangent space $T_G O(n)$ because movements in the vertical space make no change in the quotient space $O(n)/O(n-k)$. Therefore our analysis of $\mathcal{V}_{k,n}$ can be restricted to the horizontal space and by counting the dimension of each space we deduce that

$$T_{[G]}\mathcal{V}_{k,n} = \mathcal{H}_G = G \begin{pmatrix} A & -B^t \\ B & 0 \end{pmatrix} \tag{1.6}$$

We write

$$\mathcal{D} = \left\{ \begin{pmatrix} A & -B^t \\ B & 0 \end{pmatrix} \,\middle|\, A \in \mathbb{R}^{k\times k} \text{ skew-symmetric}, B \in \mathbb{R}^{(n-k)\times k} \right\}$$

1.2 Canonical metric

In Section 1.1.1 we determined the tangent space of $\mathcal{V}_{k,n}$ using the Euclidean inner product. There exists a more natural metric on the Stiefel manifold, as the Euclidean metric weights the independent dimensions of the tangent space unequally : let $\Delta = XK + (I_n - XX^T)Y \in T_X\mathcal{V}_{k,n}$. We get

$$\langle \Delta, \Delta \rangle_e = \operatorname{tr} K^T K + \operatorname{tr} Y^T Y \tag{1.7}$$

$$= 2\sum_{i<j} k_{ij}^2 + \sum_{ij} y_{ij}^2 \tag{1.8}$$

$$\tag{1.9}$$

where in 1.8 we used that K is skew-symmetric. A desirable metric would uniformly weights the coordinates k_{ij} and y_{ij}. This metric can be calculated explicitly from the description of the tangent space derived in 1.6 : let $\Delta_i = G \begin{pmatrix} A_i & -B_i^t \\ B_i & 0 \end{pmatrix} \in \mathcal{H}_G$, $i = 1, 2$, then we define the canonical inner product as

$$\langle \Delta_1, \Delta_2 \rangle = \frac{1}{2} \langle \Delta_1, \Delta_2 \rangle_e = \frac{1}{2} \operatorname{tr} A_1^t A_2 + \operatorname{tr} B_1^t B_2 \tag{1.10}$$

Using Equation 1.7 we see that the factor $\frac{1}{2}$ makes the weights of the independent dimensions uniform. Interestingly, the natural metric defined on $\mathcal{V}_{k,n}$ from its quotient definition does not coincide with the Euclidean one.

1.3 Geodesics in $O(n)$ and $\mathcal{V}_{k,n}$

Given a manifold $\mathcal{M}$, geodesics in $\mathcal{M}$ can be defined as curves $\gamma : [0,1] \to \mathcal{M}$ of shortest path between 2 points on $\mathcal{M}$. The formal definition of geodesics can be found in [21]. Informally, geodesics in $O(n)$ should always go to the same direction in $O(n)$ to minimize the path distance, and this is always possible given the group structure of $O(n)$. More precisely, let

$$\mathfrak{g} = \{\Delta \in \mathbb{R}^{n\times n} \mid \Delta = -\Delta^T\}$$

denote the Lie algebra of $O(n)$. Given a tangent direction $\Delta \in \mathfrak{g}$, a geodesic $G(t)$ in $O(n)$ should satisfy

$$\frac{d}{dt}G(t) = G(t)\Delta$$

It can be checked that the solution of this equation is given by

$$G(t) = G(0)\exp(\Delta t) \tag{1.11}$$

where $\exp : \mathfrak{g} \to O(m)$ is the Lie exponential map. In the case of linear Lie groups, it coincides with the matrix exponential map $\exp(A) = \sum_{i=0}^{\infty} \frac{1}{i!}A^i$. A formal computation of the geodesics of $O(n)$ can be found in [4]. Equation 1.11 holds not only for $O(n)$ but more generally for every compact Lie groups. To derive an explicit form of the geodesics in $\mathcal{V}_{k,n}$, we cannot use the Lie group framework (there is no well defined exponential map $\mathcal{D} \to \mathcal{V}_{k,n}$), but we can again use the definition of $\mathcal{V}_{k,n}$ as a quotient of $O(n)$. By taking Δ such that $G\Delta \in \mathcal{H}_G$ we can observe that 1.11 also defines a curve of shortest path in the quotient $O(n)/O(n-k) \cong \mathcal{V}_{k,n}$ (see Section 1.1.2). It is indeed shown in [4] that geodesics in $\mathcal{V}_{k,n}$ are of the form

$$X(t) = \tilde{X}\exp(\Delta t)I_{n,k} \tag{1.12}$$

for $\tilde{X} \in O(n)$ such that $\tilde{X}_{1:k,:} = X$, $\Delta \in \mathcal{D}$.

1.4 Givens matrices

Definition 1.4.1 (Givens matrix). *Let $n \in \mathbb{N}^*$, $(i,j) \in [\![1,n]\!]$, the Givens matrix (or Givens rotation) of axis (i,j) is the rotation matrix $G(i,j,\theta) \in \mathbb{R}^{n\times n}$ of angle θ in the 2*

dimensional subspace given by axis i and j

$$G(i,j,\theta) = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \cdots & \vdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cos(\theta) & \vdots & -\sin(\theta) & \cdots & \vdots \\ \vdots & \cdots & \cdots & \ddots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \sin(\theta) & \vdots & \cos(\theta) & \cdots & \vdots \\ \vdots & \cdots & & \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & \cdots & \cdots & 1 \end{pmatrix} \tag{1.13}$$

By convention, we let $G(i,j,\theta) = G(j,i,\theta)$, so that the order of the axis does not matter. With this convention, we set $G(i,j,\theta)_{\min(i,j),\max(i,j)} = -\sin(\theta)$ and $G(i,j,\theta)_{\max(i,j),\min(i,j)} = \sin(\theta)$.

Let

$$\mathbb{G}_n = \left\{ G(i,j,\theta) \in \mathbb{R}^{n\times n} \mid i,j \in [\![1,n]\!], -\pi < \theta \leq \pi \right\} \tag{1.14}$$

$$\mathbb{G}_{n,k} = \left\{ G(i,j,\theta) \in \mathbb{R}^{n\times n} \mid i,j \in [\![1,n]\!], \min(i,j) \leq k, -\pi < \theta \leq \pi \right\} \tag{1.15}$$

$\mathbb{G}(n)$ and $\mathbb{G}(n,k)$ both have monoid structures : $G(i,j,\theta)$ verifies $G(i,j,\theta)^{-1} = G(i,j,-\theta)$, however the product of two non trivial Givens rotation is not a Givens rotation if the rotation subspaces are not the same.

Givens rotations belong to the special orthogonal group $SO(n) \subset O(n)$, that is the group of orthogonal matrices whose determinant is 1. Given $X \in \mathcal{V}_{k,n}$, we therefore have

$$G(i,j,\theta)X \in \mathcal{V}_{k,n} \tag{1.16}$$

since

$$\begin{aligned} (G(i,j,\theta)X)^T G(i,j,\theta)X &= X^T G(i,j,\theta)^T G(i,j,\theta)X \\ &= X^T X \\ &= I_k \end{aligned}$$

Using Givens rotation as a MH exploration step is not something new. Interestingly, one of the pioneer paper to introduce MCMC sampling uses Givens matrix multiplication to define a MH algorithm reversible with respect to the uniform distribution on $O(n)$, see [9].

We now illustrate how Givens rotations are related to the Stiefel manifold. Let $\Delta_{i,j} \in \mathbb{R}^{n\times n}$ be the standard basis of $\mathcal{D}$

$$\Delta_{i,j} = e_j e_i^T - e_i e_j^T \tag{1.17}$$

for $i, j \in [\![1, n]\!]$, $j \leq \min(n, k)$, where e_i is the i^{th} vector of the canonical basis of $\mathbb{R}^n$. We recall that with respect to the canonical inner product, the geodesic starting at $X \in \mathcal{V}_{k,n}$ associated to $\Delta \in \mathcal{D}$ is given by $X(t) = \tilde{X}e^{\Delta t}I_{n,k}$.

Below is the key observation that links $\mathbb{G}(n, k)$ to coordinate descent on $\mathcal{V}_{k,n}$.

Proposition 1.4.2. *Let $i \in [\![1, n]\!]$, $j \leq \min(n, k)$ and let $\Delta_{i,j}$ be an element of the tangent space basis 1.17. Then,*

$$\exp(\alpha\Delta_{i,j}) = G(i, j, \alpha)$$

Proof. It is easily seen that $(\alpha\Delta_{i,j})^2 = -\alpha^2(e_ie_i^T + e_je_j^T)$, and $(\alpha\Delta)^3_{i,j} = -\alpha^3\Delta_{i,j}$. On the other hand, $\cos\alpha = \sum_{m=0}^{\infty}(-1)^m\frac{(\alpha)^{2m}}{(2m)!}$ and $\sin\alpha = \sum_{m=0}^{\infty}(-1)^m\frac{(\alpha)^{2m+1}}{(2m+1)!}$. Using these identities, we obtain that

$$\exp(\alpha\Delta_{i,j}) = I_n + \sum_{m=1}^{\infty}\frac{1}{m!}(\alpha\Delta_{i,j})^m = G(i, j, \alpha)$$

for all $1 \leq i \leq n, 1 \leq j \leq k$ and $-\pi < \alpha \leq \pi$. □

From this result it is now quite intuitive to see that from any point on $\mathcal{V}_{k,n}$ we can go anywhere only by travelling on the paths $\theta \mapsto G(i, j, \theta)$. We illustrate this by introducing an algorithm that given $X \in \mathcal{V}_{k,n}$, finds a maximum of $m = nk - \frac{k(k+1)}{2}$ Givens rotations $G_1, ..., G_m$ such that $G_1...G_mX = I_{n,k}$. Alternatively, this can be seen as finding a Givens decomposition of any matrix $X \in \mathcal{V}_{k,n}$ as $X = G_1...G_m$.

Lemma 1.4.3. *Let $v \in \mathbb{R}^2$ a vector such that $\|v\| = 1$. Then, there exists $\theta \in (-\pi, \pi]$ such that*

$$G(1, 2, \theta)v = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Based on this result, we can simply eliminate coordinate wise all non diagonal entries of X that are not 0.

$$\begin{pmatrix} . & . & . & . \\ n-1 & . & . & . \\ \vdots & 2n-3 & . & . \\ 2 & \vdots & \cdots & . \\ 1 & n & \cdots & nk - \frac{k(k+1)}{2} \end{pmatrix} \tag{1.18}$$

1.18 shows the order in which one could eliminate each coordinate. We only need to act in the lower triangular part of the matrix since at each step of the algorithm

the matrix stays in $\mathcal{V}_{k,n}$, so once we eliminated coordinate $(i+1,i)$ as in 1.18, we know that coordinate (i,i) is 1 and coordinates (j,i) are 0 for $j<i$.

Lastly, we discuss briefly the special case $n=k$. We pointed out in 1.16 that $GX \in \mathcal{V}_{k,n}$ for $G \in \mathbb{G}_n$ and $X \in \mathcal{V}_{k,n}$. If $k=n$, we have $\mathcal{V}_{k,n} = O(n)$ and $GX \in SO(n)X \subsetneq O(n)$ where $SO(n)X$ denotes the orbit of X by the action of $SO(n)$. In other words, X is constrained to stay in its orbit and cannot travel across the whole space $O(n)$ simply with Givens multiplication. This would prevent us from sampling from the target distribution if the target and initial distribution do not coincide on the orbit classes (e.g. we cannot approximate the uniform distribution on $\mathcal{V}_{k,n}$ if we start at $X_0 = I_{n,k}$ with probability 1). One possibility to overcome this would be to permute two rows or to multiply a row by -1, but for simplicity we choose to consider the case $k<n$ in Section 3. In this case, it is instructive to see that for $a \in \{-1,1\}$, from $X \in \mathcal{V}_{k,n}$, it is always possible to complete X with $n-k \times k$ orthogonal completion Z such that the concatenation $(X|Z)$ has determinant a. Indeed, if $\det(X|Z) = -a$ we can multiply a column of Z by -1 so $\det(X|Z) = a$ and one can check that this still defines an orthogonal matrix.

1.5 Invariant measures on $\mathcal{V}_{k,n}$

In this section we define the invariant measure on the Stiefel manifold. Let $X \in \mathbb{R}^{n\times k}$ and $\mathrm{Leb}_{n,k}(dX) = \bigwedge_{i=1}^{n}\bigwedge_{j=1}^{k} dX_{ij}$ denote the ordinary Lebesgue measure. Similarly to $\mathbb{R}^{n\times k}$, one can define the Lebesgue measure on $\mathcal{S}_n$ as

$$\mathrm{Leb}_n^{\mathcal{S}}(dS) = \bigwedge_{i\le j} dS_{ij} \tag{1.19}$$

Roughly speaking, we want to define a measure on $\mathcal{V}_{k,n}$ that would represent a notion of volume, as the Lebesgue measure on $\mathbb{R}^{n\times k}$ that is translation invariant. In the case of $\mathcal{V}_{k,n}$, translation would corresponds to transformation of the form $X \to GX$ and $X \to XG'$ for $G \in O(n)$ and $G' \in O(k)$.
We recall that the Stiefel manifold $\mathcal{V}_{k,n}$ can be defined as the quotient $O(n)/O(n-k)$. To define an invariant measure on $\mathcal{V}_{k,n}$, we will first determine the invariant measure on $O(n)$ and use the quotient definition of $\mathcal{V}_{k,n}$ to get the desired measure. As $O(n)$ is a Lie group, the existence of an invariant measure on $O(n)$ holds more generally :

Theorem 1.5.1. *Let $\mathcal{X}$ be a topological space and let G be a transitive compact topological group of transformations of X into itself such that the mapping $X \to GX$ is continuous with respect to g and X. Then there exists a unique (up to multiplicative constant) invariant measure on X under transformations of G.*

The corresponding measure is called the Haar measure. From the general theory of Lie group, it is also possible to show that the quotient $O(n)/O(n-k)$ admits a

unique invariant measure.

Let dX be the matrix of differential, i.e. $(dX)_{ij} = dX_{ij}$. By definition, differentiating $X^TX = I_k$ leads $X^TdX + dX^TX = 0$, i.e. X^TdX is a skew symmetric matrix of differential forms. We then introduce the differential form that takes the exterior product of all the upper triangular part of X^TdX

$$dX = \bigwedge_{i<j} X_i^T dX_j \tag{1.20}$$

where X_i is the i^{th} column of X. The following lemma from [2] proves that the differential form 1.20 is indeed invariant :

Lemma 1.5.2. *Let $H \in O(n)$, then $d(HX) = dX$. If $H \in O(k)$, $d(XH) = dX$.*

Now, given $X \in \mathcal{V}_{k,n}$, let $Z \in \mathbb{R}^{n\times n-k}$ such that the concatenation $(X|Z) \in O(n)$. The differential form

$$dX = \bigwedge_{i<j} X_j^T dX_i \bigwedge_{i=1}^{m-k} \bigwedge_{j=1}^{k} Z_i^T dX_j \tag{1.21}$$

is invariant on $\mathcal{V}_{k,n}$ for left and right multiplication by $O(n)$ and $O(k)$ respectively ([2]). Moreover, one can show that the differential form 1.21 is independent of the choice of orthogonal completion Z.

From the invariant differential form 1.21, we can construct the (normalised) invariant measure $\mathcal{I}_{n,k}$ on $\mathcal{V}_{k,n}$

$$\mathcal{I}_{n,k}(dX) = \frac{dX}{\int_{\mathcal{V}_{k,n}} dX} \tag{1.22}$$

where $\int_{\mathcal{V}_{k,n}} dX$ is given in [2]

$$\int_{\mathcal{V}_{k,n}} dX = \frac{\pi^{\frac{nk}{2}}}{\Gamma_k(\frac{n}{2})}$$

The invariant measure $\mathcal{I}_{n,k}$ allows use to define other distributions via densities. Let $X \in \mathbb{R}^{n\times k}$ and $\pi(dX) = \tilde{\pi}(X)\,\mathrm{Leb}_{n,k}(dX)$ a probability distribution. Since $\mathcal{V}_{k,n} \subset \mathbb{R}^{n\times k}$, when the context is clear we will slightly abuse of notation and also use $\tilde{\pi}$ to refer to the function $\mathcal{V}_{k,n} \xrightarrow{i} \mathbb{R}^{n\times k} \xrightarrow{\tilde{\pi}} \mathbb{R}$ where $i : \mathcal{V}_{k,n} \to \mathbb{R}^{n\times k}$ is the natural inclusion. $\tilde{\pi}$ needs to be normalized to define a density on $\mathcal{V}_{k,n}$, and one can check that the distribution

$$\pi_{\mathcal{V}}(dX) = \frac{\tilde{\pi}(X)\mathcal{I}_{n,k}(dX)}{\int_{\mathcal{V}_{k,n}} \tilde{\pi}(X)\mathcal{I}_{n,k}(dX)}$$

defines a probability measure on $\mathcal{V}_{k,n}$.

Chapter 2

MCMC

We formally introduce the concept of Markov Chain Monte Carlo in a general state space, as well as some of the properties of the underlying invariant kernels. We then introduce the famous Metropolis-Hastings class of algorithms as it is usually defined in $\mathbb{R}^n$ and extend this definition so that we could define it in more complex spaces, such as the Stiefel manifold. Some background on the Radon-Nykodym derivative is given in Appendix A.1.

2.1 Markov Kernel

Let $(\mathcal{X}, \mathfrak{A})$ be a σ-algebra.

Definition 2.1.1 ((Markov) Kernel)**.** a Markov kernel on $\mathcal{X}$ is a function $K : \mathcal{X} \times \mathfrak{A} \to [0,1]$ such that for all $x \in \mathcal{X}$, $K(x,.)$ is a probability measure and such that for all $F \in \mathfrak{A}$, $x \to K(x,F)$ is measurable function with respect to $\mathfrak{A}$.

Together with an initial distribution μ on $\mathcal{X}$, a Markov kernel defines a Markov chain $(X_n)_{n\geq 0}$ such that $X_0 \sim \mu$, and the distribution of X_n is defined recursively defined for $n \geq 1$ as

$$X_n \,|\, X_{n-1} = x \sim K(x,.)$$

If the initial probability distribution does not matter, we will identify a Markov chain by its corresponding kernel K.

Definition 2.1.2 (Composition Kernel)**.** Let K_1, K_2 be two kernels on $\mathcal{X}$, we define the composition kernel K_1K_2 as

$$K_1K_2(x,A) = \int_{\mathcal{X}} K_2(x,dy)K_1(y,A)$$

For $k \in \mathbb{N}^*$, the kernel K^k (called product kernel or k-step kernel) is defined recursively as $K^1 := K$ and $K^k := KK^{k-1}$

We illustrate the name k-step kernel by the following fact : if $(X_n)_{n\geq 0}$ is a Markov chain with underlying kernel K, the law of X_n verifies

$$X_n \mid X_0 = x \sim K^n(x, \cdot)$$

.

Let $L^2(\pi)$ be the space of π-integrable functions $f : \mathcal{X} \to \mathcal{X}$ such that

$$\int_{\mathcal{X}} f^2(x)\pi(dx) < \infty$$

A kernel K acts on $L^2(\pi)$ via the operator

$$Kf(y) = \int_{\mathcal{X}} f(x)K(y, dx)$$

Let $\mathbf{1}$ be the constant function with value 1, it is straightforward to verify that $K\mathbf{1} = \mathbf{1}$, i.e. the constant function is an eigenfunction of K with eigenvalue 1.

K also acts on the right on measures as

$$\mu K(A) = \int_{\mathcal{X}} K(x, A)\mu(dx)$$

For the trivial kernel $K(x, dy) = \delta_{y=x}(dy)$, we see that $\mu K = \mu$.

The relation between these two operators is the following :

Proposition 2.1.3. *Let μ a measure such that $\mu \ll \pi$ and $\frac{d\mu}{d\pi} = f \in L^2(\pi)$. Then $\frac{d\mu P}{d\pi} = Kf$*

2.2 Convergence

It is reasonable to ask if the distribution π_n of a Markov chain X_n converges to a target distribution π as $n \to \infty$.

Definition 2.2.1 (Stationary measure)**.** A probability measure π on $(\Omega, \mathfrak{A})$ is a stationary distribution for the Markov chain $(X_n)_{n\geq 0}$ if for all $A \in \mathfrak{A}$,

$$\int_{\mathcal{X}} K(x, A)d\pi(x) = \pi(A)$$

where K is the underlying kernel of X. We call K (resp. X) a π-invariant kernel (resp. chain).

By a recursive argument, This equality implies that if $X_0 \sim \pi$, $X_n \sim \pi$ for every $n \geq 1$.

If the state space $\mathcal{X}$ is finite, then every Markov chain admits a stationary distribution. If $\mathcal{X}$ is countable, and if X admits a stationary distribution π, is irreducible and aperiodic, then $\lim_{n\to\infty} X_n \sim \pi$. Moreover, if a Markov chain X converges to a distribution π, then X must be π-invariant. For the continuous case, we refer to [16] for general convergence results.

We now relate MCMC algorithms with the theory introduced so far. Let π a distribution we would like to sample from. A Markov chain Monte Carlo (MCMC) algorithm consists of simulating a Markov chain such that

$$\lim_{n\to\infty} X_n \sim \pi \tag{2.1}$$

A necessary condition for 2.1 to hold is K being π-invariant. It is in general quite hard to construct π-invariant kernels directly from the definition, we instead use the stronger notion of π-reversibility.

Definition 2.2.2 (Reversible measure)**.** A probability measure π on $(\Omega, \mathcal{X})$ is *reversible* for K if

$$\int_B \pi(dx)K(x,A) = \int_A \pi(dx)K(x,B) \tag{2.2}$$

This means that $(X_n)_{n\geq 0}$ steps from A to B as often as from B to A. It is easy to see that a time-reversible distribution π is stationary. We call K a π-reversible kernel. An equivalent definition is the following :

Definition 2.2.3 (Reversible measure 2)**.** Let $F, G \in L^2(\pi)$. A probability measure π on $(\mathcal{X}, \mathfrak{A})$ is reversible if

$$\int_{\mathcal{X}^2} F(x)G(x')\pi(dx)K(x,dx') = \int_{\mathcal{X}^2} F(x')G(x)\pi(dx)K(x,dx') \tag{2.3}$$

for all $F, G \in L^2(\pi)$.

Equation 2.2 is the special case of Equation 2.3 with $F(x) = \mathbb{1}_A(x), G(x) = \mathbb{1}_B(x)$. On the other hand, the first definition also implies the second one by using the fact that every measurable function can be approximated by a sum of indicator functions.

A desirable property one would like to know when sampling from a π-invariant kernel is how fast does it converge to the invariant distribution π. Proposition 2.2.5 gives an answer to this question.

Definition 2.2.4. A π-reversible kernel K has a spectral gap if there exists $\rho < 1$ such that for all $f \in L^2(\pi)$,

$$\|Kf\|_{L^2(\pi)} \leq \rho \|f\|_{L^2(\pi)} \tag{2.4}$$

The spectral gap of K is the value $1 - \rho > 0$. We say that a π-reversible kernel K is geometrically ergodic if it has a spectral gap.

Spectral gap ensures exponential convergence, as the following proposition from [22] says :

Proposition 2.2.5. *If K is a π-reversible with spectral gap $1-\rho > 0$, then*

$$\|K^n f\|_{L^2(\pi)} \leq \rho^n \|f\|_{L^2(\pi)}$$

for any $f \in L^2(\pi)$ orthogonal to the constant function (i.e. $\mathbb{E}_\pi(f) = 0$). Furthermore, for any signed measure ν such that $\nu \ll \mu$ and $C_\nu := \|\frac{d\nu}{d\mu}\|_{L^2(\pi)} < \infty$,

$$\|\nu K^n - \pi\|_{TV} \leq C_\nu \rho^n$$

The second important implication of geometric ergodicity is a Central Limit Theorem (CLT) for Markov Chain, that can be found in [22] :

Theorem 2.2.6. *Let $(X_n)_{n\geq 0}$ be a π-reversible and geometrically ergodic Markov chain. Let $\psi : \mathcal{X} \to \mathbb{R}$ a function of interest,*

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=0}^{n} \psi(X_i)$$

and

$$\sigma_\infty^2 = \lim_{n\to+\infty} \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=0}^{n}\left(\psi(X_i) - \mathbb{E}_\pi(\psi)\right)\right)^2\right]$$

Then,

$$\sqrt{n}(\hat{\mu}_n - \mathbb{E}_\pi(\psi)) \underset{n\to+\infty}{\longrightarrow} \mathcal{N}(0, \sigma_\infty^2)$$

in distribution.

Different CLT results for MCMC can be found in [16]. We now introduce an upper bound for σ_∞^2.

Corollary 2.2.7. *Let $(X_n)_{n\geq 0}$ be a π-reversible Markov chain with spectral gap $1-\rho$. Then,*

$$\sigma_\infty^2 \leq \frac{1+\rho}{1-\rho}\mathrm{Var}_\pi(\psi(X))$$

Let

$$ESS = \frac{1-\rho}{1+\rho}n,$$

we directly obtain from Corollary 2.2.7 that

$$\frac{\sigma_\infty^2}{n} \leq \frac{\mathrm{Var}_\pi(\psi(X))}{ESS} \tag{2.5}$$

ESS is called the effective sample size. Inequality 2.5 precisely tells us that asymptotically, the variance of $\hat{\mu}_n$ is smaller than the variance of ESS iid samples. Therefore having ESS close to n indicates sampling performance close to iid samples. In practice, ESS is difficult to compute, but we can estimate it empirically using

$$\rho = \sum_{k=-\infty}^{\infty} \hat{\rho}_k = 1 + 2\sum_{k=1}^{\infty} \hat{\rho}_k$$

where $\hat{\rho}$ denotes the autocorrelation function

$$\hat{\rho}_k = \frac{1}{(n-k)\hat{\sigma}^2} \sum_{i=1}^{n-k} (\psi(X_i) - \hat{\mu}_n)(\psi(X_{i+k}) - \hat{\mu}_n),$$

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (\psi(X_i) - \hat{\mu}_n)^2$$

The resulting estimator ESS is given by

$$\hat{ESS} = \frac{n}{1 + 2\sum_{k=1}^{N} \hat{\rho}_k}$$

where N is a suitable truncation index. The variance of $\hat{\rho}_k$ for k close to n is usually large, as the autocorrelation estimate is an average of very few terms. Choosing $N = n$ would produce a very noisy estimate of ρ. One practice for choosing N is given by [8] which proves that the sequence

$$m \mapsto \hat{\rho}'_{2m} + \hat{\rho}'_{2m+1}$$

is strictly positive and strictly decreasing if the chain is reversible, where $\hat{\rho}'_k$ is the biased estimator $\frac{1}{n\hat{\sigma}^2} \sum_{i=1}^{n-k} (\psi(X_i) - \hat{\mu}_n)(\psi(X_{i+k}) - \hat{\mu}_n)$. N is then chosen as the first index i such that $\hat{\rho}_{2i} + \rho_{2i+1} < 0$.
An other alternative is to choose

$$N = \min\{k \mid \hat{\rho}_k < c\}$$

with $c > 0$ close to 0.

2.3 Metropolis-Hastings kernels

Metropolis-Hastings kernels [15], [9] refers to a flexible class of π-reversible kernels for which simulation of the corresponding Markov chain can be simulated quite easily. These kernels can therefore be used to define MCMC algorithms. The Markov chain combines an exploration step using a candidate kernel with an accept-reject step. The strength (and also weakness) of Metropolis-Hastings algorithm comes from the freedom one has for the choice of the exploration kernel, and the accept-reject step is here to turn the exploration kernel into a π-reversible kernel.

Let q be a Markov kernel on $\mathcal{X}$. given a current state $x \in \mathcal{X}$, we sample x' as follows : let $\tilde{x} \sim q(x, \cdot)$, with probability $\min(1, r(x, \tilde{x}))$, we set $x' = \tilde{x}$, otherwise we set $x' = x$. Here, $r(x, \tilde{x})$ is the Radon-Nikodym derivative of $\pi(d\tilde{x})q(\tilde{x}, x)$ with respect to $\pi(dx)q(x, d\tilde{x})$ (we suppose it exists), i.e. the unique function $r : \mathcal{X}^2 \to \mathbb{R}$ which satisfies

$$\int_{\mathcal{X}^2} h(x,y)r(x,y)\pi(dx)q(x,dy) = \int_{\mathcal{X}^2} h(x,y)\pi(dy)q(y,dx)$$

In the case where $q(x, \cdot)$ and π admit densities with respect to an other measure ν, i.e. $\pi(dx) = \tilde{\pi}(x)\nu(dx)$ and $q(x, dy) = \tilde{q}(x, y)\nu(dy)$, we then have

$$r(x, \tilde{x}) = \frac{\tilde{\pi}(d\tilde{x})\tilde{q}(\tilde{x}, dx)}{\tilde{\pi}(dx)\tilde{q}(x, d\tilde{x})}$$

Definition 2.3.1. Let q be a proposal kernel and π a probability measure, the Metropolis-Hastings kernel P defined by q, π is

$$P(x, dy) = r(x, y)q(x, dy) + s(x)\delta_x(dy)$$

with $s(x) = 1 - \int_{\mathcal{X}} r(x, dy)q(x, dy)$

As a first remark, we observe that only the ratio of the distribution evaluated at the current state and the proposal is needed to compute the kernel. Therefore one can sample from unnormalised distributions such as posterior distributions as the normalising constants cancel each other.

The simplest and most studied example of MH algorithm is $\mathcal{X} = \mathbb{R}^d$, $\pi(dx) = \hat{\pi}(x)\,\mathrm{Leb}_n(dx)$ with Leb_n the Lebesgue measure in $\mathbb{R}^d$ and $q(x, \cdot) \sim \mathcal{N}(x, \Sigma)$ for a carefully chosen Σ.
The proof of the following theorem is quite known and straightforward, it can be found in [1].

Theorem 2.3.2. *Let P be a Metropolis-Hastings kernel with target distribution π. Then P is π-reversible.*

The main strength (but also weakness) of MH kernels comes from the proposal kernel $q(x, \cdot)$. It provides a great degree of freedom in the exploration step, and the kernel remains π-reversible whatever our exploration function is. MH algorithms can be seen as a way of turning a general Markov chain whose kernel is the proposal kernel to a π-reversible chain by appropriately accepting / rejecting steps. This freedom can also become a burden as there is no golden rule on how to choose the exploration function appropriately. A measure of performance of MH algorithms is the acceptance ratio, where a ratio of 1 means that the chain never rejects steps (hence the kernel q is π-reversible) and a ratio of 0 means that the chain does not move.

The acceptance probability r is not obvious to determine, for instance when π or $q(x,\cdot)$ does not have a density with respect to a base measure. Before defining MH algorithms on the Stiefel manifold, we will therefore view this class of MCMC algorithm from a more general point of view.

2.4 A general perspective on Metropolis-Hastings kernels

In this section, let $(\mathcal{X},\mathfrak{A})$ be a σ-algebra. We start with some definitions.

Definition 2.4.1 (Pushforward measure)**.** Let λ be a measure on $\mathcal{X}$ and let $\phi : \mathcal{X} \to \mathcal{X}$ be a measurable function. We define the pushforward measure λ^ϕ as

$$\lambda^\phi(A) = \lambda(\phi^{-1}(A))$$

for all $A \in \mathfrak{A}$.

Definition 2.4.2 (Involution)**.** Let $\phi : \mathcal{X} \to \mathcal{X}$ be a measurable function. ϕ is called an involution if

$$\phi \circ \phi = \mathrm{Id}$$

We present the slightly abstract result at first sight from [1] that we will use to define π-reversible kernels on $\mathcal{V}_{k,n}$ in Chapter 3.

Theorem 2.4.3. *Let $\phi : \mathcal{X} \to \mathcal{X}$ be an involution and μ be a probability measure on $(\mathcal{X},\mathfrak{A})$. Let $\mathscr{S} \subset \mathfrak{A}$ such that $\mu_{\mathscr{S}} \equiv \mu^\phi_{\mathscr{S}}$, where $\mu_{\mathscr{S}}(A) := \mu(A \cap \mathscr{S})$ denotes the restriction of μ to $\mathscr{S}$. For $\xi \in \mathcal{X}$, let*

$$r(\xi) = \begin{cases} \frac{d\mu^\phi_{\mathscr{S}}}{d\mu_{\mathscr{S}}}(\xi) & \xi \in \mathscr{S} \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

and let $a(\xi) = \min(1, r(\xi))$. Then the kernel on $\mathcal{X}$ defined by

$$\Pi(\xi, d\xi') = a(\xi)\delta_{\phi(\xi)}(d\xi') + (1 - a(\xi))\delta_\xi(d\xi')$$

is μ-reversible.

One can think that we could use this μ-reversible kernel to define a MCMC algorithm. However, if we start applying this kernel at ξ, applying it successively leads to a state sequence that only alternates between ξ and $\phi(\xi)$, and therefore produces a very poor exploration. This is why we usually couple a μ-reversible kernel with a refreshment of some parameters to ensure some exploration of the state space. If one is interested in sampling from π supported on $\mathcal{X}_0$, we can define $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_1$ where $\mathcal{X}_1$ is a parameter space endowed with a distribution easy to sample from. The following proposition shows that we can use a reversible kernel on $\mathcal{X}$ together with some refreshments of the state on $\mathcal{X}_1$ to get a π-reversible kernel defined on $\mathcal{X}_0$.

Proposition 2.4.4. *let $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_1$, $\xi = (\xi_0, \xi_1)$, and let π be a probability measure on $(\mathcal{X}_0, \mathfrak{B})$. If $\mu(d\xi) = \pi(d\xi_0)\mu_{\xi_0}(d\epsilon_1)$, with μ_{ξ_0} being the conditional distribution of ξ_1 given ξ_0 under μ, then the kernel $P : \mathcal{X}_0 \times \mathfrak{B} \to [0,1]$ defined by*

$$P(\xi_0, B) = \int_{\mathcal{X}_1 \times \mathcal{X}} \mathbb{1}_{\{\xi_0' \in B\}} \mu_{\xi_0}(d\xi_1)\Pi(\xi, d\xi')$$

is π-reversible.

Proof. Apply Proposition 2.5.2 with $\mathfrak{G}(\xi, \xi') = \delta_{\xi}(\xi')$. In this case, the kernel $\mathfrak{L}$ of Proposition 2.5.2 coincides with Π and we obtain the desired result. □

Algorithm 1 Sample from P

Input: initial value $\xi_0 \in \mathcal{X}_0$
  Sample $\xi_1 \sim \mu_{\xi_0}$
  Compute $r = \frac{d\mu^{\phi}_{\mathscr{L}}}{d\mu_{\mathscr{L}}}((\xi_0, \xi_1))$
  Sample $U \sim \text{Uniform}(0,1)$
  if $U \leq r$ **then**
    Return ξ_0'
  else
    Return ξ_0
  end if

This theoretical framework for defining MH kernels can be extended to a broader class of invariant kernels.

2.5 Non reversible Markov kernels

As pointed out in Section 2.1, the essential property of a chain is being invariant, not reversible. In some settings, invariant but non reversible kernels can converge faster [3] and have smaller asymptotic variance than their reversible counterparts [23], [17]. In this section we introduce a class of non reversible invariant kernels, first introduced in [24]. Instead of exploring the state space the same way at each step, this class of kernel adds additional flexibility by introducing a non trivial rejection step. It can be expressed as the composition of two MH kernels, as the following theorem from [1] says :

Theorem 2.5.1. *Let $\phi, \sigma : \mathcal{X} \to \mathcal{X}$ be involutions and let μ, λ be probability measures on $\mathcal{X}$ such that $\mu^{\sigma} = \mu$, $\lambda^{\phi} = \lambda^{\sigma} = \lambda$ and $\mu \ll \lambda$ with $\frac{d\mu}{d\lambda} = \rho$. Let Π be the μ-reversible kernel defined in Theorem 2.4.3, and let $\mathfrak{G}$ be the kernel $\mathfrak{G}(\xi, d\xi') = \delta_{\sigma(\xi)}(d\xi')$. Let $\psi = \phi \circ \sigma$, and let the corresponding kernel be $\Psi(\xi, d\xi') = \delta_{\psi(\xi)}(d\xi')$.*

Let $\mathscr{S} = \{\xi \in \mathfrak{A} \mid \min(\rho(\xi), \rho \circ \psi(\xi)) > 0\}$. *With*

$$r(\xi) = \begin{cases} \frac{d\mu^{\psi}}{d\mu(\xi)} & \xi \in \mathscr{S} \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

the kernel

$$\mathfrak{L}(\xi, \xi') = a(\xi)\Psi(\xi, d\xi') + (1 - a(\xi))\mathfrak{G}(\xi, d\xi')$$

is μ*-invariant. More precisely,* $\mathfrak{L}$ *satisfies the skew detailed balance*

$$\mu(d\xi)\mathfrak{L}(\xi, d\xi') = \mu(d\xi')\mathfrak{G}\mathfrak{L}\mathfrak{G}(\xi', d\xi)$$

Theorem 2.5.1 is a generalisation of Theorem 2.4.3, as setting $\sigma = Id$ gives a reversible MH kernel. As in the reversible case, we are interested in coupling the non reversible kernel with some partial refreshment of the state from their marginal density.

Proposition 2.5.2. *let* $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_1$, $\xi = (\xi_0, \xi_1)$, *and let* π *be a probability measure on* $(\mathcal{X}_0, \mathfrak{B})$ *such that* $\mu(d\xi) = \pi(d\xi_0)\mu_{\xi_0}(d\epsilon_1)$, *with* μ_{ξ_0} *being the conditional distribution of* ξ_1 *given* ξ_0 *under* μ. *We assume that* $\mathfrak{G}(\xi_0, \xi_1) = (\mathfrak{G}_0(\xi_0), \xi_1)$. *Then, the kernel* $Q : \mathcal{X}_0 \times \mathfrak{B} \to [0, 1]$ *defined by*

$$Q(\xi_0, B) = \int_{\mathcal{X}_1 \times \mathcal{X}} \mathbb{1}_{\{\xi_0' \in B\}} \mu_{\xi_0}(d\xi_1)\mathfrak{L}(\xi, d\xi')$$

is π*-invariant, and verifies*

$$\pi(d\xi_0)Q(\xi_0, d\xi_0') = \pi(d\xi_0')\mathfrak{G}_0 Q \mathfrak{G}_0(\xi', d\xi)$$

The proof is in Appendix A.2

The assumptions $\mu^{\sigma} = \mu$ can be satisfied only if we have some knowledge of the distribution μ, which does not always occur in practice. If the distribution of interest is μ_0 defined on $\mathcal{X}_0$, we can however extend the state space as $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_1$ and define as target distribution $\mu = \mu_0 \otimes \mu_1$ and let σ be of the form $\sigma(X_0, X_1) = (X_0, \sigma'(X_1))$ for $(X_0, X_1) \in \mathcal{X}$ and $\sigma' : \mathcal{X}_1 \to \mathcal{X}_1$. In this setting $\mathcal{X}_1$ contains some exploration parameters, and the function σ' should be chosen to optimize exploration after a rejection.

In what follow we will present an instructive example of both a reversible and nonreversible kernels.

2.6 HMC

Hamiltonian Monte Carlo (HMC) [18] is a MH kernel that aims at approximating a distribution π with density $\tilde{\pi}$ supported on $\mathbb{R}^n$, but it can be extended to $\mathbb{R}^{n\times k}$ without any difficulty. We will present it in the framework of Theorem 2.4.3.

Let $\mathcal{X}$ be a state space $\mathcal{X} \subset \mathbb{R}^{n\times k} \times \mathbb{R}^{n\times k}$, and let μ be the target distribution given by the density $\tilde{\mu}(\xi) = \tilde{\mu}_0(\xi_0)\tilde{\mu}_1(\xi_1)$ with respect to the Lebesgue measure such that μ has full support on $\mathcal{X}$. Let $\epsilon > 0$ and σ be the involution $\sigma(\xi_0, \xi_1) = (\xi_0, -\xi_1)$. Let

$$l_0(\xi_0, \xi_1) = (\xi_0 + \epsilon \log \nabla \tilde{\mu}_1(\xi_1), \xi_1) \qquad l_1(\xi_0, \xi_1) = (\xi_0, \xi_1 + \frac{\epsilon}{2} \nabla \log \tilde{\mu}_0(\xi_0))$$

. We call $l = l_1 \circ l_0 \circ l_1$ the Leapfrog step.

Lemma 2.6.1. *Let $\sigma(\xi_0, \xi_1) = (\xi_0, -\xi_1)$ and $\psi = l^k$ for $k > 0$. The kernel $H(\xi, \xi') = a(\xi)\delta_{\psi(\xi)}(d\xi') + (1 - a(\xi))\delta_{\sigma(\xi)}(d\xi')$ is μ-invariant.*

Proof. We have

$$\begin{aligned} l^{-1} &= l_1^{-1} \circ l_0^{-1} \circ l_1^{-1} && (2.8)\\ &= l_1^{-1} \circ (\sigma \circ l_0 \circ \sigma) \circ l_1^{-1} && (2.9)\\ &= \sigma \circ l_1 \circ l_0 \circ l_1 \circ \sigma && (2.10)\\ &= \sigma \circ l \circ \sigma && (2.11) \end{aligned}$$

From 2.11 we get that that $\sigma = l \circ \sigma \circ l$ and more generally that $\sigma = l^k \circ \sigma \circ l^k$. We have therefore that $\phi = \sigma \circ \psi$ is an involution. Moreover, since μ has full support, $\min(\tilde{\mu}(\xi), \mu \circ \psi(\xi)) > 0$, and the Lebesgue measure is invariant by σ since the Jacobian is 1. We conclude by applying Theorem 2.5.1. □

Remark : In the above calculation we only used in 2.9 the fact that $\nabla \log \tilde{\mu}_1(-\xi_1) = -\nabla \log \tilde{\mu}_1(\xi_1)$. This equality therefore holds more generally by replacing $\nabla \log \tilde{\mu}_0(\xi_0)$ by an arbitrary function $f_0(\xi_0)$ and $\nabla \log \tilde{\mu}_1(\xi_1)$ by $f_1(\xi_1)$ such that $f_1(-\xi_1) = -f_1(\xi_1)$.

The Leapfrog step has a physical meaning : ξ_0 can be interpreted as a particle position in $\mathbb{R}^{n\times k}$ and ξ_1 is its momentum. The leapfrog step simulates the trajectory of the particle with ϵ being the precision of the time discretization of the simulation. The Leapfrog step is a discretization of the (continuous) Hamiltonian dynamics and it is shown in [18] that if one was able to simulate them exactly, the MH acceptance ratio would always be 1, i.e. simulating exactly the Hamiltonian dynamics define a reversible kernel.

Moreover, when combining this step with a refreshment of the velocity ξ_1 at each step, it can be shown that the kernel obtain on $\mathbb{R}^{n\times k}$ is π_0-reversible. Algo

2.7 NUTS

The performance of HMC depends strongly on the choice of ϵ and k. The parameter k determines the length of the trajectory simulated, hence low values would lead to poor exploration and large values can lead to trajectories that loop back on previous steps. ϵ determines the precision of the Hamiltonian dynamics, hence too large would leads to inaccurate trajectories, implying acceptance ratios far from 1. A too small ϵ is also problematic since in this case k must be very large to generate efficient trajectory length.
No-U-Turn Sampler (NUTS) [11] simulates the same Hamiltonian trajectories as HMC but adaptively selects the parameters ϵ and k. More precisely, let (x_t, v_t) the current state of the Hamiltonian dynamics, NUTS aims at running the simulation until the condition

$$\frac{d}{dt}\|x_t - x_0\|_2 \leq 0 \tag{2.12}$$

is met. This condition is met when the particle starts going back to its initial position (i.e. it makes a U-turn). Unfortunately, choosing the next state as soon as condition 2.12 is met does not produce a reversible kernel. The main idea behind NUTS to overcome this issue is to extend the state space with a variable $\mathcal{C}$ that stores some states encountered during the exploration and which are candidates for the next step. This set grows as we keep generating trajectories, and once condition 2.12 is met by a trajectory, the exploration stops and the new state is chosen uniformly in $\mathcal{C}$. Trajectories are generated following a binary tree structure, and if condition 2.12 is not met for explored states at step $j-1$, step j runs a trajectory of length 2^j in a direction. Therefore NUTS needs to compute $O(2^j)$ gradient evaluations where j is the depth of the recursive procedure for the trajectory. The other parameter ϵ that needs to be manually selected in HMC is adapted at each warmup iteration using dual averaging [19] : ϵ is chosen at each step so that the average acceptance ratio among the visited states of the last tree doubling phase converges to an optimal acceptance rate, see Section 3.2 of [11] for further details.

NUTS performs on average better than HMC and most of the common MCMC samplers, which makes it the common MCMC algorithm to use for when sampling in a real vector space. In fact, the effective sample size per gradient evaluation of NUTS is larger than HMC kernels with different values of L and ϵ, see [11] for further comparison of NUTS and HMC.

Chapter 3

MCMC on the Stiefel manifold

In Chapter 2 we viewed different MCMC algorithms from a general perspective. In this chapter we are interested in defining π-invariant kernels on the Stiefel manifold. The main difficulty in restricting MCMC algorithms to an embbeded manifold like $\mathcal{V}_{k,n}$ is how can we ensure that each sample stays on the manifold while still exploring efficiently. We begin by reviewing an existing method to project any MCMC algorithm defined on $\mathbb{R}^{n\times k}$ onto $\mathcal{V}_{k,n}$.
In the rest of this section we fix n, k such that $k < n$.

3.1 Polar expansion

Polar expansion [12] is a method to perform MCMC on the Stiefel manifold by projecting at each step our sample from $\mathbb{R}^{n\times k}$ into $\mathcal{V}_{k,n}$.
Let $X \in \mathbb{R}^{n\times k}$ be a full rank matrix and let

$$X = Q_X S_X^{1/2}$$

be its polar decomposition, where $Q_X = X(X^TX)^{-\frac{1}{2}}$ and $S_X = X^TX$. Alternatively, let $X = UDV^T$ be the singular value decomposition (with D being a squared $k \times k$ diagonal matrix) , we have $Q_X = UV^T$ and $S^{1/2} = VDV^T$.
Q_X verify $QX = \underset{Q \in \mathcal{V}_{k,n}}{\operatorname{argmin}} \|X - Q\|_F$, i.e. we can interpret Q_X as the closest matrix in $\mathcal{V}_{k,n}$ to X in the Frobenius norm. Polar expansion defines a distribution on $\mathcal{V}_{k,n}$ using the diffeomorphism

$$\Psi : \mathbb{R}^{n\times k} \mapsto \mathcal{V}_{k,n} \times \mathcal{S}_k$$

$$X \mapsto (Q_X, S_X)$$

where $\mathcal{S}_k$ denotes the space of positive semi-definite k x k matrices. As it turns out the Lebesgue measure on $\mathbb{R}^{n\times k}$ can be decomposed onto $\mathcal{V}_{k,n}$ and $\mathcal{S}_k$. The lemma below comes from [10].

Lemma 3.1.1. *Let $X \in \mathbb{R}^{n\times k}$, we have*

$$\mathrm{Leb}_{n,k}(dX) = \underbrace{\frac{\pi^{\frac{nk}{2}}}{\Gamma_k(\frac{n}{2})}|S_X|^{\frac{n-k-1}{2}}}_{:=J(X)} \mathrm{Leb}_k^{\mathcal{S}}(dS_X)\mathcal{I}_{n,k}(dQ_X)$$

where Γ_k is the multivariate gamma function.

Corollary 3.1.2. *Let π be a probability measure on $\mathbb{R}^{n\times k}$ such that $\pi \equiv \mathrm{Leb}_{n,k}$ and $\pi(dX) = \tilde{\pi}(X)\,\mathrm{Leb}_{n,k}(dX)$. Let (Q_X, S_X) denote the polar decomposition of X and let $\pi_Q, \pi_{S|Q}$ be distributions on $\mathcal{S}_k$ and $\mathcal{V}_{k,n}$ respectively, such that*

$$\pi_{S|Q}(dS) = \tilde{\pi}_{S|Q}(S_X|Q_X)\,\mathrm{Leb}_n^{\mathcal{S}}(dS) \quad \textit{and} \quad \pi_Q(dQ) = \tilde{\pi}_Q(Q)\mathcal{I}_{n,k}(dQ)$$

Then,

$$\tilde{\pi}(X) = \tilde{\pi}_{S|Q}(S_X|Q_X)\tilde{\pi}_Q(Q_X)J^{-1}(X)$$

if and only if $Q_X \sim \pi_Q$.

The proof is in Appendix A.3

Corollary 3.1.2 tells us that we can control the distribution of Q_X while sampling X with the appropriate kernel. To sample from $\tilde{\pi}_Q$ on $\mathcal{V}_{k,n}$, we therefore need a kernel K defined on $\mathbb{R}^{n\times k}$ whose invariant distribution is $\pi(dX) = \tilde{\pi}_{S|Q}(S_X|Q_X)\tilde{\pi}_Q(Q_X)\,\mathrm{Leb}_{n,k}(dX)$.

Algorithm 2 Polar expansion

Input: Conditional density $\tilde{\pi}_{S|Q}$ on $\mathcal{S}_k$, target density $\tilde{\pi}_Q$ on $\mathcal{V}_{k,n}$, full rank matrix $X_0 \in \mathbb{R}^{n\times k}$
$Q_0 = X_0(X_0^TX_0)^{-1/2}$
for $t = 0$ to $T-1$ **do**
  Sample $X_{t+1} \sim K(X_t, .)$
  $Q_{t+1} = X_{t+1}(X_{t+1}^TX_{t+1})^{-1/2}$
end for

We follow recommendations of [12] and set $S \sim W_n(I_k)$, i.e. S is Wishart distributed. S is therefore independent of Q with density

$$\tilde{\pi}_{S|Q}(S \mid Q) = \tilde{\pi}_{S|Q}(S) = \frac{1}{2^{nk}\Gamma_k(\frac{n}{2})}\exp\big(-\mathrm{Tr}(S)/2\big) \tag{3.1}$$

and using Corrolary 3.1.2, the density $\tilde{\pi}$ on X is

$$\tilde{\pi}(X) = \frac{1}{2\pi^{nk}}\exp\big(-\mathrm{Tr}(X^TX)/2\big)\tilde{\pi}_Q(Q_X) \tag{3.2}$$

where we used the equality $S_X = X^T X$. Equation 3.2 gives us a nice way to sample from the uniform distribution $\mathcal{I}_{n,k}$ on $\mathcal{V}_{k,n}$. We first introduce the Gaussian distribution on matrices. Section 1.5 explains how one can extend a distribution defined via a density on $\mathbb{R}^{n\times k}$ to $\mathcal{V}_{k,n}$.

Definition 3.1.3. *The density of a Gaussian distribution on $\mathcal{V}_{k,n}$ with mean $M \in \mathcal{V}_{k,n}$, row covariance $\Sigma \in \mathbb{R}^{n\times n}$ and column covariance $\Gamma \in \mathbb{R}^{k\times k}$ is given at X by*

$$f(X) = \frac{1}{(2\pi)^{\frac{nk}{2}}|\Sigma|^{\frac{k}{2}}|\Gamma|^{\frac{n}{2}}} \exp\left(\frac{-1}{2}\left(\Gamma^{-1}(X-M)^T\Sigma^{-1}(X-M)\right)\right) \tag{3.3}$$

Corollary 3.1.4. *Let $X \in \mathcal{V}_{k,n}$ such that $X \sim \mathcal{N}(I_{n,k}, I_k, I_n)$. Then, $Q_X \sim \mathcal{I}_{n,k}$*

Proof. If $Q_X \sim \mathcal{I}_{n,k}$, then $\tilde{\pi}_Q(Q_X) \propto 1$, and from 3.2 we see that $\pi(X) \sim \mathcal{N}(I_{n,k}, I_k, I_n)$. □

We observe that to evaluate $\tilde{\pi}(X)$ we first need to compute the polar decomposition of X, which consists in inverting a $k \times k$ matrix. Therefore polar expansion can be hard to perform for very large values of k.

3.2 MultiG kernel

Introductory example

In the framework of Theorem 2.4.3, let $\mathcal{X} = \mathcal{V}_{k,n} \times \mathbb{G}_n$, and let ϕ be the involution

$$\phi(X, G) = (GX, G^{-1})$$

Let π be a distribution on $\mathcal{V}_{k,n}$ with density $\tilde{\pi}$ with respect to $\mathcal{I}_{n,k}$. Moreover, we assume that $\tilde{\pi}(X) > 0$ for all $X \in \mathcal{V}_{k,n}$, i.e. $\pi \equiv \mathcal{I}_{n,k}$.In this case, we also have $\pi^\phi \equiv \mathcal{I}_{n,k}$, and by transitivity $\pi^\phi \equiv \pi$.

By fixing i and j, $G(i,j,\theta)_{-\pi<\theta\leq\pi} \cong SO(2) \cong \mathbb{S}^1$ where $\cong$ is a group isomorphism and $SO(2)$ is the special orthogonal group in dimension 2. It is easily seen that we can identify $\mathbb{G}_n$ with a wedge of $\frac{n(n-1)}{2}$ copies of $\mathbb{S}^1$ whose intersection is at the identity I_n. Therefore let $\Delta_{\mathbb{G}_n}$ be the uniform measure on $\mathbb{G}_n$.

We define the target distribution μ on $\mathcal{X}$ such that

$$d\mu(dX) = \tilde{\pi}(X)\mathcal{I}_{k,n}(dX)\Delta_{\mathbb{G}_n}(dG)$$

i.e. the distribution on $\mathbb{G}_n$ is uniform. It is easily checked that assumptions of Theorem 2.4.3 are verified.

Lemma 3.2.1. *Let μ, π, ϕ be as defined above. Then,*

$$\frac{d\mu^\phi}{d\mu} = \frac{\tilde{\pi}(GX)}{\tilde{\pi}(X)}$$

Algorithm 3 Basic μ-reversible kernel with refreshment

Input: $X \in \mathcal{V}_{k,n}$
  Sample $G \sim \Delta_{\mathbb{G}_n}$
  Sample $U \sim \text{Uniform}(0,1)$
  if $U \leq \frac{\tilde{\pi}(GX)}{\tilde{\pi}(X)}$ **then**
    Return GX
  else
    Return X
  end if

Proof. See the proof of Lemma 3.2.3 □

In this simple kernel (and in the following ones), we use the fact that multiplication by a Givens matrix is relatively easy to do : multiplication by a Givens rotation only acts on 2 rows of $X \in \mathcal{V}_{k,n}$, we only need to recompute these rows so the cost of this operation is $O(k)$.
The main issue with the kernel defined in 3 is its poor scaling of exploration with respect to dimension : ϕ changes only 2 rows of X, hence as the dimension n increases, from the Coupon Collector lemma we would need $\Theta(n\log(n))$ on average to explore across each row of X. We solve this issue by increasing the number of Givens multiplication in each step.

Uniform MultiG

Let $m := m(n)$ be a function of n, $\mathcal{X} = \mathcal{V}_{k,n} \times \mathbb{G}_n^m$ be the extended state space and let $\phi : \mathcal{X} \to \mathcal{X}$ be the involution

$$\phi(X, G_1, ..., G_m) = (G_m ... G_1 X, G_m^{-1}, ..., G_1^{-1})$$

We again assume that $\pi(dX) \equiv \mathcal{I}_{n,k}$, let $\tilde{\pi}$ such that $\pi(dX) = \tilde{\pi}(X)\mathcal{I}_{n,k}(dX)$, and for $\xi = (X, G_1, ..., G_m)$, let

$$\mu(d\xi) = \tilde{\pi}(X)\mathcal{I}_{n,k}(dX)\prod_{i=1}^{m}\Delta_{\mathbb{G}_n}(dG_i)$$

Lemma 3.2.2. *Let μ, π, ϕ be as defined above. Then,*

$$\frac{d\mu^{\phi}}{d\mu}(\xi) = \frac{\tilde{\pi}\left(G_m ... G_1 X\right)}{\tilde{\pi}(X)} \tag{3.4}$$

Proof. See the proof of Lemma 3.2.3 □

The special case $m = 1$ corresponds to 3.

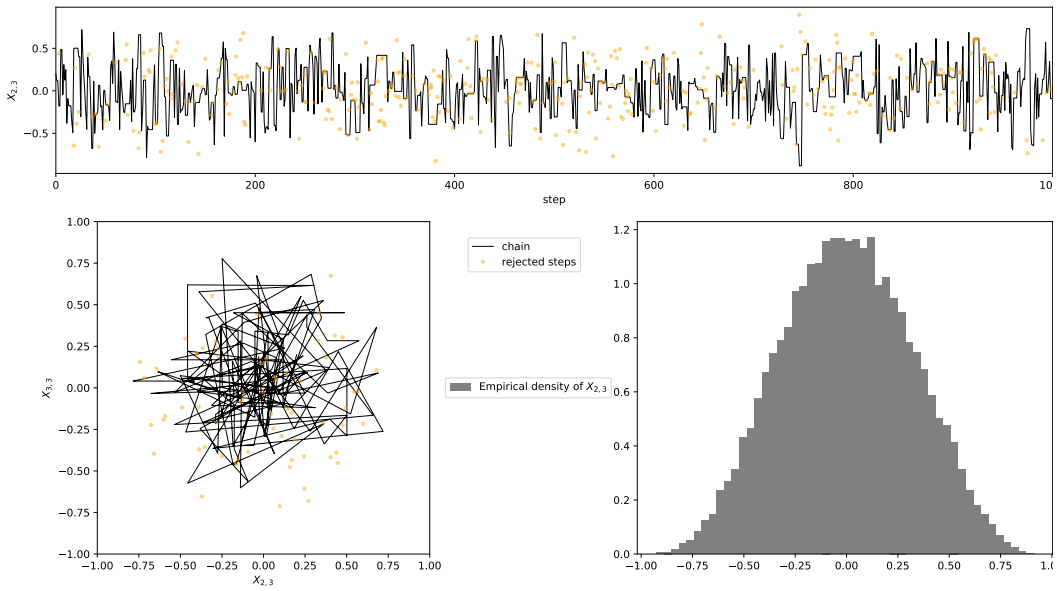


Figure 3.1: MultiG algorithm in $\mathcal{V}_{10,5}$ with target distribution $\mathcal{N}(I_{n,k}, I_k, I_n)$. 50 chains are sampled for 1000 steps with no warmup. Top : coordinate $(2,3)$ of the first chain X. Bottom left : coordinate $(3,3)$ of X with respect to coordinate $(2,3)$. For clarity, only the first 200 steps are displayed. Bottom right : Histogram of the empirical density of coordinate $(2,3)$ across all chains.

Choosing m is not obvious. A first remark is that we should choose m smaller than $nk - \frac{k(k+1)}{2}$ as this already enables us to travel to the whole space $\mathcal{V}_{k,n}$ (see Section 1.4. Moreover, from the Coupon Collector's Lemma, the product of $n\log(n)$ independently uniformly sampled Givens rotation is dense with high probability [5]. Therefore we also restrained $m \leq n\log(n)$.

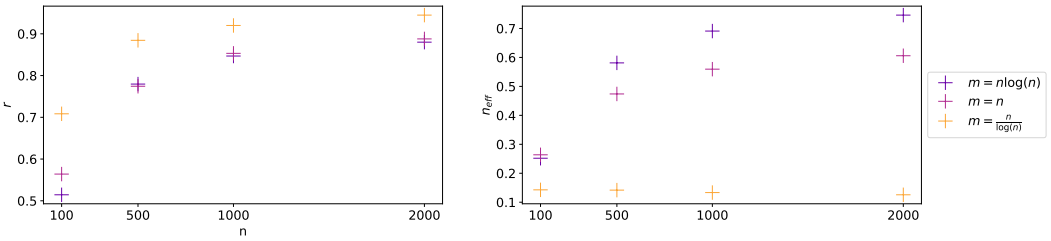


Figure 3.2: Acceptance ratio and relative effective sample size of multiG algorithm on $\mathcal{V}(n,80)$ with target distribution $\mathcal{N}(I_{n,k}, I_k, I_n)$ and uniform sampling on the m Givens rotations. Each plot is the average of 5 independent chains of 1000 steps with no warmup.

Figure 3.2 shows how the acceptance ratio and the effective sample size behave with respect to the dimension for different functions $m(n)$. We observe that we should keep $m \geq n$ for a good exploration, as the number $m = \frac{n}{n\log(n)}$ has very poor performance. Interestingly, there is an improvement for $m = n\log(n)$ over $m = n$,

which gives similar conclusion to the fact that the product of n Givens matrix is not dense in $O(n)$ [5]. However from Figure 3.6 the effective sample size improvement does not compensate the longer computation time of the $m = n\log(n)$ case. From Figure 3.2 one may also wonder if there is an optimal acceptance ratio, independent of the dimension, for a broad class of target distributions. A future line of work could tackle this question as [6] did for MH kernels in $\mathbb{R}^n$.

General MultiG

It is natural to think about other choices of distribution over the Givens rotation. We first analyse the case where the axis of the Givens samples are still uniform but now the angle is constrained by a value $\vartheta < \pi$. Formally, we now set

$$\mu(d\xi) \propto \tilde{\pi}(X)\mathcal{I}_{n,k}(dX)\prod_{i=1}^{m}\mathbb{1}_{-\vartheta\leq\theta_i\leq\vartheta}\Delta_{\mathbb{G}_n}(dG_i)$$

where θ_i denote the angle of G_i. Again, we observe that $\mu^\phi \equiv \mu$, and the acceptance ratio is still given by 3.4.

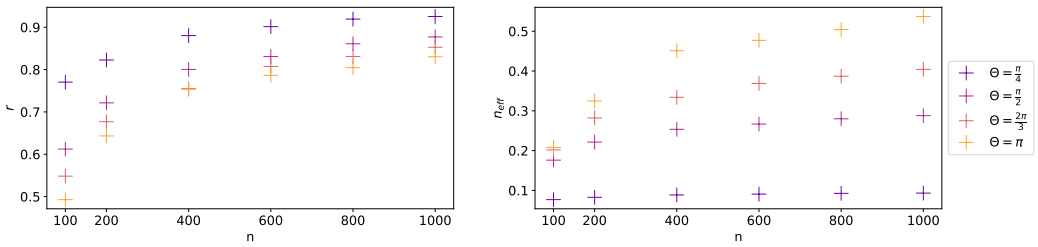


Figure 3.3: Acceptance ratio and relative effective sample size of multiG algorithm on $\mathcal{V}(n, 100)$ with target distribution $\mathcal{N}(I_{n,k}, I_k, I_n)$ and with n Givens rotations sampled with angle constraint Θ. Each plot is the average of 10 independent chains of 1000 steps with no warmup.

From figure 3.3, we see that the best effective sample size (and thus the less autocorrelation) is obtained in the unconstrained case for a standard Gaussian. For sharper distributions, a large angle could lead to a small accpetance ratio as the steps would jump over the modes of the target. It is however still unclear how one could choose the angle from the target from a theoretical perspective, and we leave this question for future work.

We now move to the more general case where the Givens rotations are not sampled from the same distribution. Let

$$\mu(d\xi) = \pi(dX)\prod_{i=1}^{m} g_i(G_i)\Delta_{\mathbb{G}_n}(dG_i) \tag{3.5}$$

where g_i are densities with respect to $\Delta_{\mathbb{G}_n}$ To define a μ-reversible kernel from Theorem 2.4.3, we need to have $\mu^\phi \equiv \mu$ and ideally, the acceptance ratio should be

easy to compute and should encapsulate information only from the space we are interested in, i.e. $\mathcal{V}_{k,n}$. The following lemma gives some sufficient conditions to satisfy these conditions.

Lemma 3.2.3. *Let π be a distribution such that $\pi \equiv \mathcal{I}_{n,k}$ and $\frac{d\pi}{d\mathcal{I}_{n,k}} = \tilde{\pi}$. Let μ as defined in 3.5, and let g_i be densities on $\mathbb{G}_n$ such that $g_i(G) = g_{n-i}(G)$ and $g_i(G) = g_i(G^{-1})$ for all $G \in \mathbb{G}_n$. Then, $\mu^\phi \equiv \mu$ and*

$$\frac{d\mu^\phi}{d\mu}(\xi) = \frac{\tilde{\pi}(G_m ... G_1 X)}{\tilde{\pi}(X)} \tag{3.6}$$

and this acceptance ratio defines a π-reversible MH kernel.

Proof. See Appendix A.4 □

We will define a non trivial distribution on $\mathbb{G}_n$ that satisfies the assumptions of Lemma 3.2.3. Intuitively, the first axis of G_i is sampled uniformly between i and $n-i$, and the second axis is sampled uniformly over the remaining axis. Let

$$\mathbb{G}_n(i, j, [\vartheta_1, \vartheta_2]) = \{G(i, j, \theta) \mid \vartheta_1 < \theta < \vartheta_2\} \subset \mathbb{G}_n$$

Given a fixed angle constraint ϑ,we let

$$g_i(G) \propto \sum_{j \in A^{\times i}} \mathbb{1}_{\mathbb{G}_n(i,j,[-\vartheta,\vartheta])}(G) + \sum_{j \in A^{\times n-i}} \mathbb{1}_{\mathbb{G}_n(n-i,j,[-\vartheta,\vartheta])}(G) \tag{3.7}$$

where $A^{\times i} = \{0, ..., i-1, i+1, ..., n\}$ and $\vartheta \in (0, \pi]$ is the angle constraint.

Algorithm 4 multiG π-reversible kernel on $\mathcal{V}_{k,n}$

Input: initial value $X_0 \in \mathcal{V}_{k,n}$, $m \in \mathbb{N}^*$ and densities g_i, $i = 1, ..., m$ as in Lemma 3.2.3

```
for t = 1 to T do
    for i = 1 to m do
        Sample G_i ~ g_i
    end for
    Compute Y_t = G_m ... G_1 X_t
    Sample U ~ Uniform([0, 1])
    Compute r = π(Y_t) / π(X_t)
    if U ≤ r then
        X_{t+1} = Y_t
    else
        X_{t+1} = X_t
    end if
end for
```

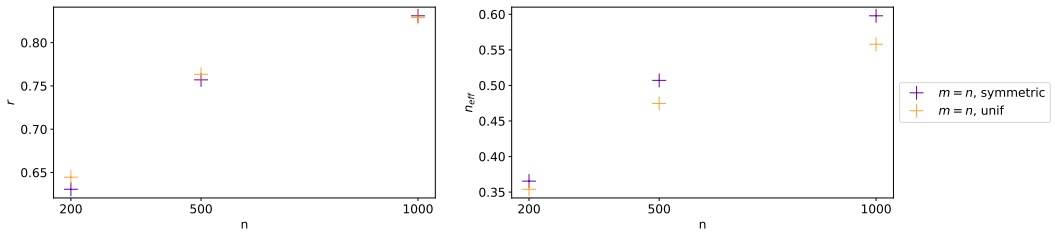


Figure 3.4: Acceptance ratio and relative effective sample size of multiG algorithm on $\mathcal{V}(n, 100)$ with target distribution $\mathcal{N}(I_{n,k}, I_k, I_n)$ and with n Givens rotations sampled with angle constraint $\vartheta = \pi$. Each plot is the average of 10 independent chains of 1000 steps with no warmup.

Figure 3.4 illustrates the performance of selecting the axis on the Givens rotations from 3.7 compared to the uniform distribution, we see that selecting axes non uniformly accelerate the mixing of the chain. One could try to extend this idea to more complex settings while respecting the assumptions of Lemma 3.2.3, but usually the drawback would be a longer computation time. This is the same trade off one has to think about when choosing the number m of Givens rotation per step, and a measure of this performence is shown in Figure 3.6. For simple distributions, MultiG achieve correct effective sample size per time unit. However, one is generally interested in more complex distribution, this primarily why one uses MCMC algorithms. We will now extend our intuition of the MultiG kernels to the nonreversible setting, allowing us to adapt the sampling process more efficiently.

3.3 Nonreversible kernel

We now aim at extending the kernel 4 to the non reversible setting 2.5.1. A non reversible kernel makes a step in case of rejection, which can be interpreted in our case as changing the exploration direction.

Let g_i be are arbitrary densities with support $U_i \subset \mathbb{G}_n$ for $i \in [\![1, m]\!]$. Let $\mathcal{X} = \mathcal{V}_{k,n} \times \prod_{i=1}^{m} U_i \times \{-1; 1\}$ and let $\xi = (X, G_1, ..., G_m, \tau) \in \mathcal{X}$. We define

$$\phi(\xi) = ((G_m ... G_1)^\tau X, G_1, ..., G_m, -\tau) \tag{3.8}$$
$$\sigma(\xi) = (X, G_1, ..., G_m, -\tau) \tag{3.9}$$
$$\psi(\xi) := \phi \circ \sigma(\xi) = ((G_m ... G_1)^\tau X, G_1, ..., G_m, \tau) \tag{3.10}$$

Let μ be the target distribution

$$\mu(d\xi) = \frac{1}{2}\pi(dX) \prod_{i=1}^{m} g_i(G_i) \Delta_{\mathbb{G}_n}(dG_i)$$

where as before $\pi(dX) \equiv \mathcal{I}_{n,k}$ and $\pi(dX) = \tilde{\pi}(X)\mathcal{I}_{n,k}(dX)$.

Lemma 3.3.1. *Let ψ, σ and μ as defined above. Then*

$$\frac{d\mu^{\psi}}{d\mu}(\xi) = \frac{\tilde{\pi}(G_m ... G_1 X)}{\tilde{\pi}(X)} \tag{3.11}$$

and the kernel

$$P(\xi, \xi') = a(\xi)\delta_{\psi(\xi)}(d\xi') + (1 - a(\xi))\delta_{\sigma(\xi)}(d\xi')$$

is μ-invariant, and satisfies

$$\mu(d\xi)P(\xi, d\xi') = \mu(d\xi')\delta_{\sigma(\xi)}P(\xi, d\xi')\delta_{\sigma(\xi)}(\xi, d\xi')$$

for all $\xi, \xi' \in \mathcal{X}$.

Proof. See Appendix A.4 □

We observe that this time, we do not need to invert the order of the Givens rotations in 3.8 to make ϕ an involution. This means that we do not need to have g_i and g_{n-i} supported on the same space anymore, and we can control the distribution on each G_i more freely.

Let $(a_l, b_l)_{l \in [\![1,m]\!]}$ be an ordering of the pairs (i, j) such that $i, j \in [\![1, n]\!]$ and $\min(i, j) \leq k$. A desirable choice for m would be $m = nk - \frac{k(k+1)}{2} = \dim(\mathcal{V}_{k,n})$, so we could set

$$g_l \sim \text{Uniform}(\mathbb{G}_n(a_l, b_l, [0, \vartheta_l]))$$

With the densities g_i defined like this, From Proposition 1.4.2 the distribution of $G_m ... G_1 X$ would have full support on $\mathcal{V}_{k,n}$, for every fixed $X \in \mathcal{V}_{k,n}$, allowing us to explore $\mathcal{V}_{k,n}$ with more freedom.

However, the number $m = \dim(\mathcal{V}_{k,n})$ is in practice too large to make this algorithm efficient. Instead, we will use $m = n$ Givens rotation whose distributions are for $i \in [\![1, m]\!]$

$$g_i(G) \propto \sum_{j \in A^{\times i}} \mathbb{1}_{\mathbb{G}_n(i,j,[0,\vartheta_i])}(G)$$

Figure 3.5 illustrates the choice of the densities g_i and the role of τ : each orange cross is a rejection step, meaning that τ flips sign. We observe that the direction of X_{00} flips after the rejection step. We observe that most of the rejection steps occur when X_{00} go towards low values. In the example, $\pi \sim \mathcal{N}(I_{n,k}, I_k, I_n)$ so the marginal density of π on coordinate $(0, 0)$ has highest density at 1. We also observe that once $X_{0,0}$ goes close to 1 the chain goes down even when no rejection occurs. This behaviour is due to the orthogonality constraint of staying on $\mathcal{V}_{k,n}$: To satisfy $\|X_0\| = 1$ and $X_{0,0}$ close to 1 the other coordinates $X_{0,j}, j > 0$ must be close to 0. Since a Givens rotation $G(0, i, \theta)$ acts on $X_{0,0}$ as

$$(G(0, i, \theta)X)_{0,0} = \cos(\theta)X_{0,0} - \sin(\theta)X_{i,0}$$

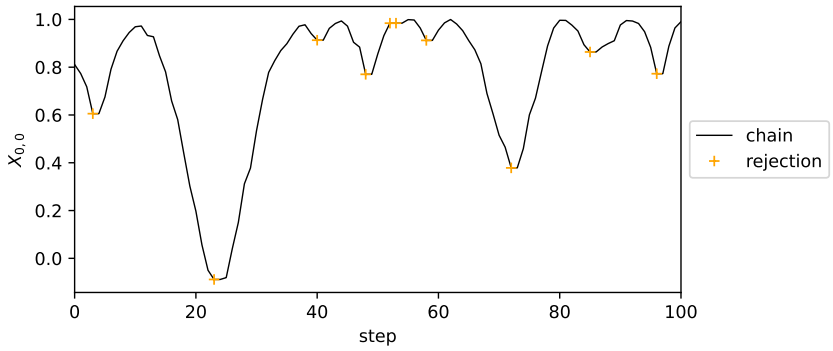


Figure 3.5: Example of the nonreversible kernel dynamics for coordinate $(0,0)$ of $X \in \mathcal{V}_{k,n}(3,2)$, with angle constraint $\vartheta_i = \pi/16$ for $i = 1,2,3$. The target distribution on $\mathcal{V}_{k,n}$ is $\mathcal{N}(I_{n,k}, I_k, I_n)$

Algorithm 5 Non reversible kernel step with partial refreshment

Input: $X_0 \in \mathcal{V}_{k,n}$, $\tau \in \{-1,1\}$, $\vartheta \in (0,\pi]^m$, $g_0, ..., g_m$
  for $t = 1$ to T **do**
    for $i = 1$ to m **do**
      Sample $G_i \sim g_i$
      Compute $Y_t = G_1 ... G_m X_t$
    end for
    Sample $U \sim \text{Uniform}([0,1])$
    Compute $r = \frac{\tilde{\pi}(Y_t)}{\tilde{\pi}(X_t)}$
    if $r \leq U$ **then**
      $X_{t+1} = Y_t$
    else
      $X_{t+1} = X_t$
      $\tau = -\tau$
    end if
  end for

we observe that in this case $X_{0,0}$ decrease by a factor of $\cos(\theta)$ approximately. More generally, each coordinate is not independant of the others as it is the case in $\mathbb{R}^{n\times k}$.

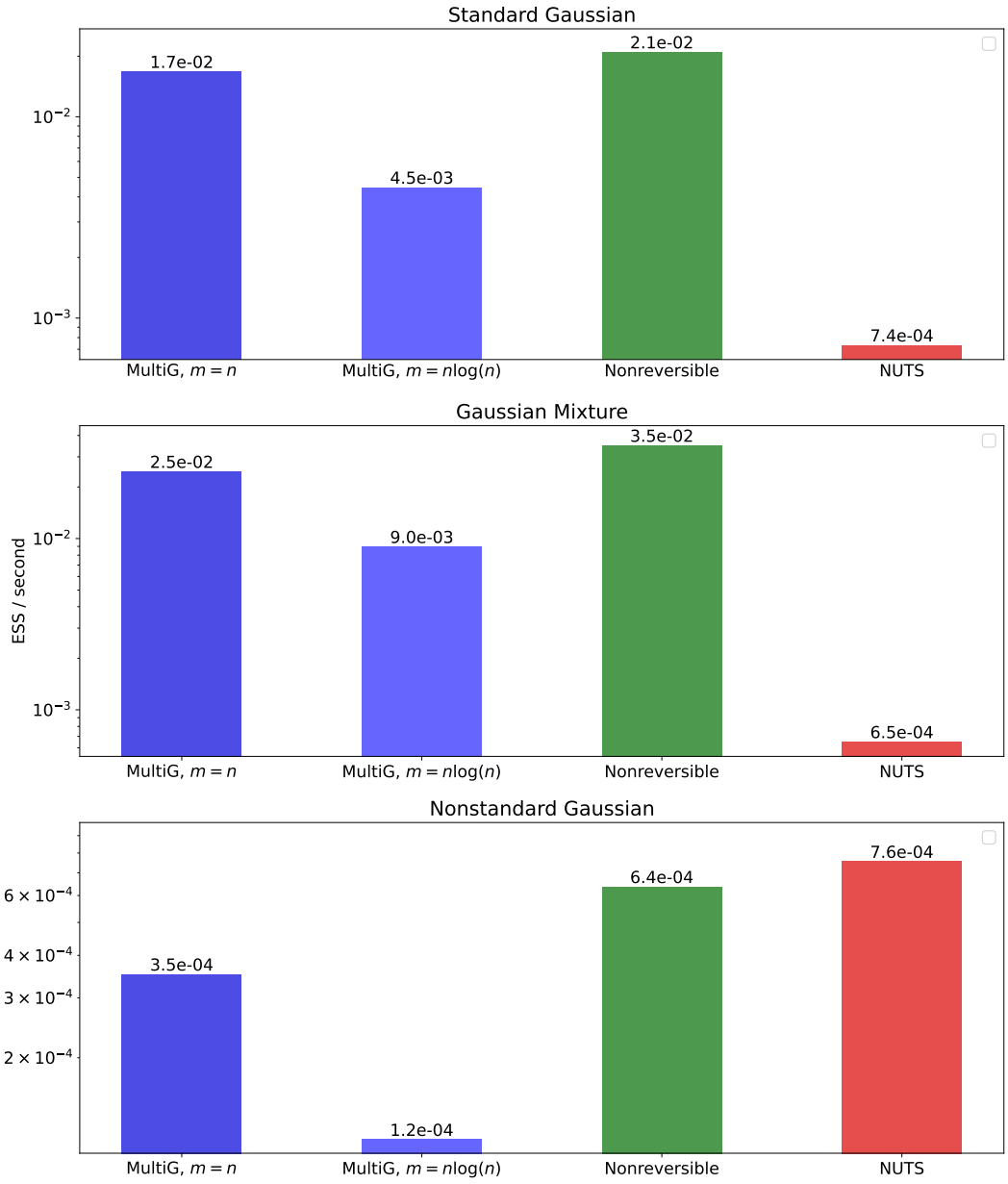


Figure 3.6: Effective sample size per computation time of different MCMC algorithms on $\mathcal{V}_{150,80}$. 5 chains of 1000 samples are used for each bar. For each target distribution, NUTS has 20 warmup steps, and the $m = n\log(n)$ Givens rotations of the first MultiG algorithms are sampled uniformly on $\mathbb{G}_n$, while the $m = n$ Givens rotations of the second MultiG algorithm are sampled according to 3.7 with angle constraint π. Top plot : the target distribution is $\mathcal{N}(I_{n,k}, I_k, I_n)$. The non reversible kernel has angle constraint $\vartheta = \pi$ for all $l \in [\![1, n]\!]$. Middle plot : The target density is a mixture of 2 standard Gaussians $\frac{1}{2}\mathcal{N}(I_{n,k}, I_k, I_n) + \frac{1}{2}\mathcal{N}(-I_{n,k}, I_k, I_n)$ centered at the opposite from eachother. The non reversible algorithm is defined similarly as before, and the function of interest to compute the effective sampled size in this case is $X \mapsto \mathbb{1}_{\|X - I_{n,k}\| < \|X + I_{n,k}\|}(X)$ for all chains. Bottom plot : The target density is a Gaussian with column covariance I_n and diagonal row covariance $\sigma_i^2 = 0.1$ for $i \leq \lfloor \frac{n}{2} \rfloor$ and $\sigma_i^2 = 1$ otherwise. This time, the non reversible algorithms has angle constraints $\vartheta_i = \lfloor \frac{\pi}{2} \rfloor$ for $i \leq \lfloor \frac{k}{2} \rfloor$ and π otherwise.

Figure 3.6 shows the measure ESS/s for different algorithms and targets, where s stands for the computation time in seconds. A first remark is the good result of the non reversible algorithm overall. It can achieve correct effective sample size, while taking much less time than NUTS and a bit less time than multiG with $m = n\log(n)$. It is also important to say that NUTS performs better in terms of effective sample size almost every time, which makes sense since it uses the gradient of the logdensity several time at each step, but its computation time is much more important which is partly because of the polar expansion computation of the sample at each step. As already said, polar expansion needs to invert a $k \times k$ matrix, therefore sampling with larger dimension k would have produced lower result for NUTS than what we observe. Also, every Gaussian density considered is implemented such that no matrix inversion is necessary for evaluation (which is possible given the simple structures of the covariance matrices), which speeds up the density evaluation at each step.

3.4 Improvement directions

Figure 3.6 illustrates the good results of the non reversible kernelover the different targets. The distribution on which we tested the MCMC algorithms were known in advance, that is why we were able to manually choose the angle constraint ϑ over the axes. Driven by the drastic improvement of adaptive ϵ in NUTS over static ϵ, we conjecture that adapting the parameter ϑ at each step would allow us to extend to extend the non reversible kernel to unknown distribution. Lastly, implementing angle self-tuning alone may not be a desirable property in certain case, as the exploration on some axes could be very poor : if the distribution over some axes is very sharpe, the angle constraint could converge to very low values, leading to a correct acceptance ratio but with very small steps, leading to poor effective sample size. We think that this approach needs to be coupled with an exploration stopping time process, i.e. the acceptance ratio is computed at different locations and the next step is first evaluated step that satisfy a criterion chosen usually to enhance exploration. We refer to Section 5 of [1] for more details on this class of algorithms.

Conclusion

In this thesis, we have presented and developed a novel approach to Markov Chain Monte Carlo sampling on the Stiefel manifold. We introduce some key knowledge about the Stiefel manifold such as closed forms of its tangent space and a closed form of its geodesics which are essential to justify the use of Givens rotations, the central piece of the kernels presented. We introduced the broad field of MCMC, and more specifically we discussed different approaches to define Metropolis-Hastings kernels and non reversible kernels. We observed different behaviors of the simple MultiG kernel depending on the number of Givens rotations and the angle constraint, and future lines of work could analyse in more detail the relation of these parameters with the sampling performance with respect to the dimension, or finding the asymptotically optimal acceptance rate of these kernels. MultiG algorithms as well as the non reversible kernel introduced in section 3.3 produce chains with a better effective sample size with respect to time compared to NUTS. However, the distributions considered in this work were relatively simple, and one could extend this comparison to more complex distributions like posteriors, where NUTS would have an advantage as it does not need parameter selection from the user. The non reversible kernel shows promising results, but we conjecture that adaptively selecting the parameter constraint ϑ, as well as introducing an exploration stopping criterion as in NUTS would greatly improve its performance. Indeed, for more complex distribution, ϑ would quickly become non trivial and a stopping criterion would allow the algorithm to explore some regions without taking any step.

Appendix A

Appendix

A.1 Radon-Nikodym derivative

Let $(\mathcal{X}, \mathfrak{A})$ be a σ-algebra.

Definition A.1.1 (Absolute continuity)**.** Let μ, λ be two σ-finite measures, we say that μ is absolutely continuous with respect to λ (we write $\mu \ll \lambda$) if for all $A \in \mathfrak{A}$, $\lambda(A) = 0 \Rightarrow \mu(A) = 0$. If $\mu \ll \lambda$ and $\lambda \ll \mu$, we write $\lambda \equiv \mu$.

Theorem A.1.2 (Radon–Nikodym theorem)**.** *Let $(\mathcal{X}, \mathfrak{A})$ be a σ-algebra. Let μ, λ be two σ-finite measures defined on $(\mathcal{X}, \mathfrak{A})$ such that $\mu \ll \lambda$. Then there exists a measurable function $f : \mathcal{X} \to [0, \infty)$ such that $\lambda(A) = \int_A f(x)\mu(dx)$. f is unique up to a μ-null set, and we write $\frac{d\lambda}{d\mu} := f$ or equivalently $\lambda(dx) = f(x)\mu(dx)$.*

A.2 Invariant kernels with partial refreshment

Proof of lemma 2.5.2. We will use the skew detailed balance of $\mathfrak{L}$. Let $f, g \in L^2(\mu)$ such that $f(\xi) = f(\xi_0), g(\xi) = g(\xi_0)$. In this case, $f, g \in L^2(\pi)$. We observe that every function $h : \mathcal{X}_0 \to \mathbb{R}$ can be extended to a function $h' : \mathcal{X} \to \mathbb{R}$. For $\xi, \xi' \in \mathcal{X}$,

$$\begin{aligned}
\int f(\xi_0) g(\xi_0') \pi(d\xi_0) Q(\xi_0, d\xi_0') &= \int f(\xi_0) g(\xi_0') \pi(d\xi_0) \mu_{\xi_0}(d\xi_1) \mathfrak{L}(\xi, d\xi') \\
&= \int f(\xi_0) g(\xi_0') \mu(d\xi) \mathfrak{L}(\xi, d\xi') \\
&= \int f(\xi_0) g(\xi_0') \mu(\xi') \mathfrak{G}\mathfrak{L}\mathfrak{G}(\xi', d\xi) \\
&= \int f(\xi_0) g(\xi_0') \pi(d\xi_0') \mu_{\xi_0'}(d\xi_1') \mathfrak{G}\mathfrak{L}\mathfrak{G}(\xi', d\xi)
\end{aligned}$$

Since $\mu^\sigma = \mu$ by definition of $\mathfrak{L}$, we have $\mu_{\xi_0} = \mu_{\sigma(\xi_0)}$. Moreover, $\mathfrak{G}$ only acts on ξ_0, and from these arguments one can check that

$$\int f(\xi_0) g(\xi_0') \pi(d\xi_0') \mu_{\xi_0'}(d\xi_1') \mathfrak{G}\mathfrak{L}\mathfrak{G}(\xi', d\xi) = \int f(\xi_0) g(\xi_0') \pi(d\xi_0') \mu_{\xi_0'}(d\xi_1') \mathfrak{G}_0 Q \mathfrak{G}_0(\xi_0', d\xi_0)$$

This gives the skew reversibility of Q, which implies that Q is π-invariant. □

A.3 Polar decomposition

Proof of Lemma 3.1.2. Let $(Q, S) \in \mathcal{V}_{k,n} \times \mathcal{S}_k$. From the assumption that $\pi \equiv \text{Leb}_{n,k}$ and by Lemma 3.1.1, π defines a distribution β on $\mathcal{V}_{k,n} \times \mathcal{S}_k$ such that $\beta \equiv \text{Leb}_k^{\mathcal{S}} \otimes \mathcal{I}_{n,k}$. Its density $\tilde{\beta}$ is given by

$$\tilde{\beta}(S, Q) = \tilde{\pi}(S^{1/2}Q)J(S^{1/2}Q) \tag{A.1}$$

We decompose the density of β as $\tilde{\beta}(S, Q) = \tilde{\beta}_S(S \mid Q)\tilde{\beta}_Q(Q)$ where $\tilde{\beta}_S(S \mid Q)$ is the conditional density of S given Q. If we assume that

$$\tilde{\pi}(X) = \tilde{\pi}(S_X^{1/2}Q_X) = \tilde{\pi}_S(S_X \mid Q_X)\tilde{\pi}_Q(Q_X)J^{-1}(S_X^{1/2}Q_X) \tag{A.2}$$

by replacing A.2 in A.1 we get

$$\tilde{\beta}_S(S \mid Q)\tilde{\beta}_Q(Q) = \tilde{\beta}(S, Q) = \tilde{\pi}_S(S \mid Q)\tilde{\pi}_Q(Q) \tag{A.3}$$

For all $(Q, S) \in \mathcal{V}_{k,n} \times \mathcal{S}_k$. Hence, $\tilde{\beta}_Q = \tilde{\pi}_Q$.

On the other hand if $Q \sim \gamma_Q$, from A.1 and by taking the conditional decomposition of π we obtain that π must have marginal density γ_Q. □

A.4 MCMC kernel acceptance ratios

Proof of Lemma 3.2.3. Let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function, and let $\xi = (X, G_1, ..., G_m)$, $Y = G_m ... G_1 X$. We will show that $\mu^\phi \equiv \mu$ by direct computation. Since ϕ is an involution $\phi = \phi \circ \phi \circ \phi^{-1} = \phi^{-1}$. By definition of μ and ϕ, we get

$$\int_{\mathcal{X}} f(X, G_1, ..., G_m) d\mu^\phi(\xi) = \int_{\mathcal{X}} f(X, G_1, ..., G_m) d\pi(Y) \prod_{i=1}^{m} g_i(G_{m-i}^{-1}) \Delta_{\mathbb{G}_n}(dG_{m-i}^{-1}) \tag{A.4}$$

By using assumptions on g_i and by rearanging terms, and since $\Delta_{\mathbb{G}_n}$ is the uniform measure on $\mathcal{V}_{k,n}$, the right hand side of A.4 equals

$$\int_{\mathcal{X}} f(X, G_1, ..., G_m) d\pi(Y) \prod_{i=1}^{m} g_i(G_i) \Delta_{\mathbb{G}_n}(dG_i) \tag{A.5}$$

Now, suppose that there exists a measurable set $A \in \mathfrak{A}$ such that $\mathcal{I}_{n,k}(A) > 0$ and $\tilde{\pi}(X) = 0$ for all $X \in A$. Since $\pi \equiv \mathcal{I}_{n,k}$ this would mean that $\pi(A) > 0$, and we obtain

$$0 < \pi(A) = \int_A \tilde{\pi}(X) \mathcal{I}_{n,k}(dX) = 0$$

which is a contradiction. Therefore, $\tilde{\pi} > 0$ π-almost surely. The right hand side of A.5 can be rewritten as

$$\begin{aligned}
\int_{\mathcal{X}} f(X, G_1, ..., G_m) d\mu^\phi(\xi) &= \int_{\mathcal{X}} f(X, G_1, ..., G_m) \tilde{\pi}(Y) d\mathcal{I}_{n,k}(Y) \prod_{i=1}^{m} g_i(G_i) \Delta_{\mathbb{G}_n}(dG_i) \\
&= \int_{\mathcal{X}} f(X, G_1, ..., G_m) \frac{\tilde{\pi}(Y)}{\tilde{\pi}(X)} \tilde{\pi}(X) d\mathcal{I}_{n,k}(X) \prod_{i=1}^{m} g_i(G_i) \Delta_{\mathbb{G}_n}(dG_i) \\
&= \int_{\mathcal{X}} f(X, G_1, ..., G_m) \frac{\tilde{\pi}(Y)}{\tilde{\pi}(X)} d\mu(\xi)
\end{aligned}$$

and therefore

$$\frac{d\mu^\phi}{d\mu}(\xi) = \frac{\tilde{\pi}(Y)}{\tilde{\pi}(X)} = \frac{\tilde{\pi}(G_m ... G_1 X)}{\tilde{\pi}(X)}$$

We can set $\mathscr{S} = \mathfrak{A}$ in Theorem 2.4.3 and since, $\mu_{\mathfrak{A}} = \mu$, we conclude from Theorem 2.4.3 that the ratio 3.6 defines a π-reversible kernel. □

Proof of Lemma 3.3.1. We can choose λ to be the measure on $\mathcal{X}$

$$\lambda(d\xi) = \frac{1}{2} \mathcal{I}_{n,k}(dX) \prod_{i=1}^{m} \Delta_{\mathbb{G}_n}(dG_i) \tag{A.6}$$

Since μ and λ admit a density with respect to the same measure, we easily obtain the Radon-Nikodym derivative

$$\frac{d\mu}{d\lambda}, (X, G_1, ..., G_m, \tau) := \rho(X, G_1, ..., G_m, \tau) = \frac{1}{2} \tilde{\pi}(X) \prod_{i=1}^{m} g_i(G_i)$$

Since π has full support, for all $\xi \in \mathcal{X}$ we have

$$\min(\mu(\xi), \mu \circ \psi(\xi)) > 0$$

and by definition of the invariant measure $\mathcal{I}_{n,k}$ on $\mathcal{V}_{k,n}$, we have $\lambda^{\psi} = \lambda$. The assumptions of Theorem 2.5.1 are satisfied and the acceptance ratio A.6 defines a nonreversible kernel. □

Bibliography

[1] Christophe Andrieu, Anthony Lee, and Sam Livingstone. "A general perspective on the Metropolis-Hastings kernel". In: *arXiv preprint arXiv:2012.14881* (2020).

[2] Yasuko Chikuse. "Statistics on Special Manifolds". In: *Annalen der Physik* 17.10 (1993), pp. 891–921.

[3] Persi Diaconis, Susan Holmes, and Radford M Neal. "Analysis of a non-reversible Markov chain sampler". In: *Annals of Applied Probability* (2000), pp. 726–752.

[4] Alan Edelman, Tomás A Arias, and Steven T Smith. "The geometry of algorithms with orthogonality constraints". In: *SIAM journal on Matrix Analysis and Applications* 20.2 (1998), pp. 303–353.

[5] Thomas Frerix and Joan Bruna. "Approximating orthogonal matrices with effective Givens factorization". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 1993–2001.

[6] Andrew Gelman, Walter R Gilks, and Gareth O Roberts. "Weak convergence and optimal scaling of random walk Metropolis algorithms". In: *The annals of applied probability* 7.1 (1997), pp. 110–120.

[7] Charles J Geyer. "Introduction to markov chain monte carlo". In: *Handbook of markov chain monte carlo* 20116022 (2011), p. 45.

[8] Charles J Geyer. "Practical markov chain monte carlo". In: *Statistical science* (1992), pp. 473–483.

[9] W Keith Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: (1970).

[10] Carl S Herz. "Bessel functions of matrix argument". In: *Annals of Mathematics* 61.3 (1955), pp. 474–523.

[11] Matthew D Hoffman, Andrew Gelman, et al. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1593–1623.

[12] Michael Jauch, Peter D Hoff, and David B Dunson. "Monte Carlo simulation on the Stiefel manifold via polar expansion". In: *Journal of Computational and Graphical Statistics* 30.3 (2021), pp. 622–631.

[13] Serge Lang. *Fundamentals of differential geometry*. Vol. 191. Springer Science & Business Media, 2012.

[14] Jun Li, Li Fuxin, and Sinisa Todorovic. "Efficient riemannian optimization on the stiefel manifold via the cayley transform". In: *arXiv preprint arXiv:2002.01113* (2020).

[15] Nicholas Metropolis et al. "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[16] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[17] Radford M Neal. "Improving asymptotic variance of MCMC estimators: Non-reversible chains are better". In: *arXiv preprint math/0407281* (2004).

[18] Radford M Neal et al. "MCMC using Hamiltonian dynamics". In: *Handbook of markov chain monte carlo* 2.11 (2011), p. 2.

[19] Yurii Nesterov. "Primal-dual subgradient methods for convex problems". In: *Mathematical programming* 120.1 (2009), pp. 221–259.

[20] Samuel Power and Jacob Vorstrup Goldman. *Accelerated Sampling on Discrete Spaces with Non-Reversible Markov Processes*. 2019. arXiv: 1912.04681 [stat.CO].

[21] Quentin Rentmeesters et al. "Algorithms for data fitting on some common homogeneous spaces". PhD thesis. Ph. D. thesis, Université Catholique de Louvain, Louvain, Belgium, 2013.

[22] Gareth Roberts and Jeffrey Rosenthal. "Geometric ergodicity and hybrid Markov chains". In: (1997).

[23] Yi Sun, Jürgen Schmidhuber, and Faustino Gomez. "Improving the asymptotic performance of Markov chain Monte-Carlo by inserting vortices". In: *Advances in Neural Information Processing Systems* 23 (2010).

[24] Akiva Moiseevich Yaglom. "On the statistical reversibility of Brownian motion". In: *Matematicheskii Sbornik* 66.3 (1949), pp. 457–492.