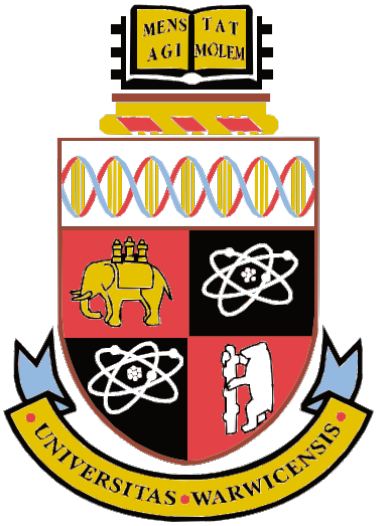

Markov chain Monte Carlo on the sphere

by

David Maxen

MA4K8 Scholarly Report

Submitted to The University of Warwick

Mathematics Institute

April, 2024

Contents

1 Introduction

Firstly, I would like to thank my supervisor Professor Tim J. Sullivan for his time, expertise and patience.

In this report, I will provide an introduction to the field of Markov chain Monte Carlo (MCMC), with both theory and examples. Many of the proofs were from the literature, however there are some which are my own, a couple of which are non-trivial.

We will compare various MCMC algorithms, and explore how their performance varies with dimension. We also explored how the spherical reprojection methods of H. C. Lie, D. Rudolf, B. Sprungk and T. J. Sullivan [Lie+23] compare to standard projection methods on $\mathbb{R}^d$.

I have also made a significant contribution in the form of an open source body of code available on my Github, Genomesh. This provides a pipeline for running MCMC algorithms both in $\mathbb{R}^d$ and on the sphere. It includes the algorithms themselves, as well as chain analysis and plotting software.

2 Markov Chains

2.1 Motivation

We want to be able to sample from a given distribution. One use of this is computing integrals using the Monte Carlo method. Example, calculating π. The need for samples. [Lie+23] and [RS22].

2.2 Notation and preliminaries

For a space $\mathbb{X}$ and set A, we define the indicator function of that set as follows.

$$\mathbb{I}_A(x) := \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

Given a topological space E, we define $\mathcal{P}(E)$ to be the set of probability measures on the Borel σ-algebra of E. We say the random variable $X \sim \mu$ if X has a μ distribution, that is for all A in the corresponding σ-algebra, $\mathbb{P}(X \in A) = \mu(A)$. It is useful to remove the need to repeat myself by defining a single underlying probability space, $(\Omega, \mathcal{E}, \mathbb{P})$, which will be the codomain of all the random variables defined henceforth.

When two functions are equal up to multiplication by a positive scalar A, for example if $f(x) = Ag(x)$ for all x, we say f is proportional to g, written $f \propto g$. This is particularly useful when dealing with probability densities. These always integrate to 1 there is no ambiguity regarding this normalising factor.

2.3 Basic definitions

The theory presented here is a selection of a few results that are particularly relevant to our analysis. These results are from the books by Meyn and Tweedie [MT93], and Robert and Castella [RC04].

We follow the basic definitions as set out in [Lie+23]. Let $\mathbb{X}$ be a Hilbert space. The sequence of $\mathbb{X}$-valued random variables $(X_n)_{n\geq 1}$ is a Markov chain if it has the following property for all $n \in \mathbb{N}$, $A \in \mathcal{B}(\mathbb{X})$ and values $(x_j)_{j=1}^n \subset \mathbb{X}$.

$$\mathbb{P}(X_{n+1} \in A | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} \in A | X_n = x_n)$$

It is useful to think of this property as saying that conditional on the present, $X_n = x_n$, the future of the chain (the values of $X_{n+1}, X_{n+2}, \ldots$) is independent of the past $(X_1, X_2, \ldots, X_{n-1})$.

We define a transition kernel $K : \mathbb{X} \times \mathcal{B}(\mathbb{X}) \to [0, 1]$ to have two properties, firstly that for each $x \in \mathbb{X}$, $K(x, \cdot) \in \mathcal{P}(\mathbb{X})$, that is, $A \mapsto K(x, A)$ is a probability measure. We also require that for all $A \in \mathcal{B}(\mathbb{X})$, $x \mapsto K(x, A)$ is measurable. This transition kernel defines how the Markov chain transitions between states.

$$\mathbb{P}(X_{n+1} \in A | X_n = x) = K(x, A)$$

Crucially, the kernel doesn't depend on the time n; this is because we are only considering time-homogeneous Markov chains.

We now want to define a notion of equilibrium for our Markov chains. Suppose we have a measure $\mu \in \mathcal{P}(\mathbb{X})$, and that $X_n \sim \mu$. Hence $\mathbb{P}(X_n \in A) = \mu(A)$. We then take one step according to K. Now, we calculate the probability the Markov chain is in A at time $N + 1$.

$$\begin{aligned}\mathbb{P}(X_{n+1} \in A) &= \int_{\mathbb{X}} \mathbb{P}(X_{n+1} \in A | X_n = x)\mu(x)dx \\ &= \int_{\mathbb{X}} K(x, A)\mu(dx)\end{aligned}$$

The first equality comes by conditioning over all possible values that $X_n \sim \mu$ can take. For K to be invariant, we want that applying K preserves the distribution; that if $X_n \sim \mu$, then $X_{n+1} \sim \mu$ also. Hence we want $\mathbb{P}(X_{n+1} \in A) = \mu(A)$ also, which motivates the following definition. We say our transition kernel K is μ-invariant if the following holds for all $A \in \mathcal{B}(\mathbb{X})$.

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

Where in μK, we are considering K as an operator, this is a slight abuse of notation. We also say μ is a stationary distribution of the Markov chain. Furthermore, we define a kernel K to be μ-reversible if it satisfies the detailed balance equation.

$$K(x, dy)\mu(dx) = \mu(dy)K(y, dx) \tag{1}$$

There is an equivalent condition for μ-reversibility that we will make use of going forward. K and μ satisfy the detailed balance equations if and only if for all $A, B \in \mathcal{B}(\mathbb{X})$, the following is true.

$$\int_B K(b, A)\mu(db) = \int_A K(a, B)\mu(da)$$

This can be interpreted as the probability mass that K moves from B to A is equal to the mass that it moves from A to B.

Proof. We first recall from the definition of the Lebesgue integral that for each $x \in \mathbb{X}$, $K(x, A) = \int_{\mathbb{X}} \mathbb{I}_A(y)K(x, dy)$. We use Fubini-Tonelli twice, as the densities are non-negative and measurable.

$$\begin{aligned}
\int_B K(b, A)\mu(db) &= \int_B \left(\int_A K(b, da) \right) \mu(db) \\
&= \int_{A \times B} K(b, da)\mu(db) \\
&= \int_{A \times B} \mu(da)K(a, db) \\
&= \int_A \left(\int_B K(a, db) \right) \mu(da) \\
&= \int_A K(a, B)\mu(da)
\end{aligned}$$

The other direction comes from Fubini-Tonelli, and the uniqueness (up to sets of measure 0) of the probability density (the Radon-Nikodym derivative).

Corollary. If a kernel K is μ-reversible, then it is μ-invariant. As $K(x, \cdot)$ is a probability measure on $\mathbb{X}$, we have $K(x, \mathbb{X}) = 1$.

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

At this point we relate these concepts back to Markov chain Monte Carlo (MCMC). In MCMC, our aim is to sample from a target distribution, μ. Suppose that we design a Markov chain with transition kernel K whose stationary distribution is μ. Suppose further that for some n, $X_n \sim \mu$. Then by our stationarity assumption, for all $t \in \mathbb{N}$, $X_{n+t} \sim \mu K^t = \mu$. This means that we can generate as many samples from μ as we like. One problem is that even though these are valid samples, ideally we want independent samples. These samples could be highly correlated, which is a large consideration when designing a MCMC algorithm that we will come back to later. A second problem is that we can't sample $X_n \sim \mu$; the entire reason for using MCMC is that we cannot otherwise sample from μ. There is an entire literature on whether Markov chains converge to a stationary distribution, and if so how fast. We will not cover that here, with the exception of the following from fact 5 of Roberts and Rosenthal, 2004, [RR04].

If $(X_i)_{i\in\mathbb{N}}$ is a μ-irreducible, recurrent $\mathbb{R}^d$-valued Markov chain which admits μ as a stationary distribution, then the following strong law of large numbers holds (convergence is with probability 1) for any integrable function $\phi : \mathbb{R}^d \to \mathbb{R}$:

$$\lim_{t\to\infty} \frac{1}{t} \sum i = 1^t \phi(X_i) = \int \phi(x)\mu(x)dx$$

for μ-almost every starting value $x_0 \in \mathbb{R}^d$. That is, for all x_0 except perhaps for some null set $\mathcal{N}$ which satisfies $\int_{\mathcal{N}} \mu(x)dx = 0$.

3 Markov Chain Monte Carlo on $\mathbb{R}^d$

Here we present the setup for MCMC as done in [Lie+23]. One difference is that they consider Hilbert spaces, however we will often work in $\mathbb{R}^d$, for a large dimension d. This is because at some point we will want to implement these algorithms computationally, at which point we can only consider finitely many dimensions. As all finite dimensional Hilbert spaces are isomorphic to a Euclidean space $\mathbb{R}^d$ with the normal inner product, this restriction to $\mathbb{R}^d$ is justified, and makes no difference to the algorithms provided. That said, where it doesn't result in much longer or more complicated proofs, many of the proofs I provide will still hold in infinite-dimensional Hilbert spaces.

Suppose that we have a target distribution ν that we are trying to sample from, and that we can compute its density with respect to a centred Gaussian prior distribution $\nu_0 \sim N(0, C)$, where C is the covariance matrix.

$$\frac{d\nu}{d\nu_0}(x) \propto \exp(-\Phi(x)), \qquad \nu_0 - \text{a.e.} x \in \mathbb{R}^d \tag{2}$$

Where $\Phi : \mathbb{R}^d \to \mathbb{R}$ is measurable and satisfies

$$\int_{\mathbb{R}^d} \exp(-\Phi(x))\nu_0(dx) < \infty$$

We first define the Metropolis-Hastings transition mechanism. Firstly, a possible new state y_{k+1} is proposed based on the current state x_k by the proposal kernel $Q : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0, 1]$. This proposed move is then accepted with probability $\alpha(x_k, y_{k+1})$, where $\alpha : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$, in which case we set $x_{x+1} = y_{k+1}$. Otherwise, the proposal is rejected and we remain at $x_{k+1} = x_k$. This results in the transition kernel

$$K(x, dy) = \alpha(x, y)Q(x, dy) + r(x)\delta_x(dy)$$

Where δ_x is the Dirac delta measure, and r is the probability that a move from x is rejected.

$$r(x) := \int_{\mathbb{R}^d} (1 - \alpha(x, y))Q(x, dy)$$

One of the simplest examples of this is the symmetric random walk Metropolis-Hastings (SRW-MH) algorithm on $\mathbb{R}^d$, which has a normal proposal kernel density centred at the current value x, making it symmetric.

$$Q(x, y) := N(y; x, sC)$$

C is the proposal covariance and $s \in (0, \infty)$ is a step size parameter which can be tuned. Proposal covariances different to the prior covariance can be used, but we will not consider that here. The acceptance probability is

$$\alpha(x, y) := \min(1, \mu(y)/\mu(x))$$

The SRW-HW algorithm can be implemented using algorithm 1. For completeness, I include a proof that the symmetric random walk Metropolis-Hastings (SRW-MH) algorithm is ν-reversible in $\mathbb{R}^d$. A proof for general state spaces can be found in [RR04]. We show the detailed balance equations 1 are satisfied, writing the densities relative to the Lebesgue measure in $\mathbb{R}^d$. The case $x = y$ is trivial, so we assume that $x \neq y$, and hence $\delta_x(y) = 0$. We also assume without loss of generality that $\nu(x) \geq \nu(y)$, hence $\alpha(x, y) = \nu(y)/\nu(x)$ and $\alpha(y, x) = 1$. We compute

$$\begin{aligned}\nu(x)K(x,y) &= \mu(x)\alpha(x,y)Q(x,y)\\ &= \nu(x)\frac{\nu(y)}{\nu(x)}N(y;x,sC)\\ &= \nu(y)N(x;y,sC)\\ &= \nu(y)\alpha(y,x)Q(y,x) = \nu(y)K(y,x)\end{aligned}$$

Algorithm 1 Symmetric Random Walk Metropolis Hastings algorithm on $\mathbb{R}^d$

Require: prior covariance C, target potential Φ, step size $s \in (0,1]$, initial state $x_0 \in \mathbb{R}^d$
  for $k \in \mathbb{N}_0$ **do**
    Draw a sample $w_k \sim N(0,C)$ and set $y_{k+1} := x_k + sw_k$
    Compute $\alpha := \min(1, \exp(\Phi(x_k) + \frac{1}{2}x_k^\top C^{-1}x_k - \Phi(y_{k+1}) - \frac{1}{2}y_{k+1}^\top C^{-1}y_{k+1}))$
    Draw a sample $u \sim U[0,1]$
    if $u \leq \alpha$ **then**
      Set $x_{k+1} = y_{k+1}$
    else
      Set $x_{k+1} = x_k$
    end if
  end for

We now discuss another Metropolis-Hastings algorithm, the pCN (preconditioned Crank-Nicolson) algorithm [Nea99], [Cot+13]. Here, the proposal kernel Q is asymmetric, with density

$$Q(x,y) := N\left(y; \sqrt{1-s^2}x, s^2C\right)$$

where $s \in (0,1]$ is a step size parameter. The acceptance probability is given by

$$\alpha(x,y) := \min(1, \exp(\Phi(x) - \Phi(y)))$$

The pCN algorithm can be implemented as per algorithm 2. We will also prove that the resulting transition kernel K is ν-invariant. Again, we assume $x \neq y$ and, without loss of generality, that $\Phi(x) \leq \Phi(y)$, so $\alpha(x,y) = \exp(\Phi(x) - \Phi(y))$ and $\alpha(y,x) = 1$.

$$
\begin{aligned}
K(x,y)\nu(x) &= \alpha(x,y)Q(x,y)\nu(x,y) \\
&= \exp(\Phi(x)-\Phi(y))N\left(y;\sqrt{1-s^2}x, sC\right)N(x;0,C)\exp(-\Phi(x)) \\
&= \frac{1}{Z}\exp\left(-\Phi(y)-\frac{1}{2}\left(y-\sqrt{1-s^2}x\right)^\top (s^2C)^{-1}\left(y-\sqrt{1-s^2}x\right)-\frac{1}{2}x^\top C^{-1}x\right) \\
&= \frac{1}{Z}\exp\left(-\Phi(y)-\frac{1}{2s^2}\left(y^\top C^{-1}y-\sqrt{1-s^2}x^\top C^{-1}y-\sqrt{1-s^2}y^\top C^{-1}x+x^\top C^{-1}x\right)\right) \\
&= \frac{1}{Z}\exp\left(-\Phi(y)-\frac{1}{2}\left(x-\sqrt{1-s^2}y\right)^\top (s^2C)^{-1}\left(x-\sqrt{1-s^2}y\right)-\frac{1}{2}y^\top C^{-1}y\right) \\
&= \exp(-\Phi(y))N(y;0,C)N\left(x;\sqrt{1-s^2}y, sC\right) \\
&= \alpha(y,x)Q(y,x)\nu(y,x) \\
&= K(y,x)\nu(y)
\end{aligned}
$$

Where $Z = (2\pi)^d\sqrt{\det(C)}\sqrt{\det(s^2C)}$ is a product of normalisation constants.

Algorithm 2 pCN algorithm on $\mathbb{R}^d$

Require: prior covariance C, target potential Φ, step size $s \in (0,1]$, initial state $x_0 \in \mathbb{R}^d$
  for $k \in \mathbb{N}_0$ **do**
    Draw a sample $w_k \sim N(0,C)$ and set $y_{k+1} := \sqrt{1-s^2}x_k + sw_k$
    Compute $\alpha := \min(1, \exp(\Phi(x_k)-\Phi(y_{k+1})))$
    Draw a sample $u \sim U[0,1]$
    if $u \leq \alpha$ **then**
      Set $x_{k+1} = y_{k+1}$
    else
      Set $x_{k+1} = x_k$
    end if
  end for

3.1 Example in $\mathbb{R}^d$

Now that we've introduced two MCMC algorithms, we will apply them to a problem of nonparametric Bayesian density estimation. Suppose you have some samples from an unknown distribution, and you're aiming to infer the distribution from these samples. A common way to do this is to inspect the data, pick a distribution that fits well (for example a normal distribution), then use a maximum likelihood estimation approach to find the parameters of the distribution (for a normal distribution these would be the mean and variance). However, what if the unknown distribution doesn't follow a known distribution, or you don't want to make an assumption on the distribution? Furthermore, you only have finitely many samples from the distribution, so it would be useful to incorporate your uncertainty over what the unknown distribution is into your model.

One answer to these problems is to use a non-parametric Bayesian approach. In this case, by non-parametric we mean that we do not rely on data belonging to any particular parametric family of probability distributions. We also follow a Bayesian approach, which allows us to incorporate our beliefs about what the distribution should look like. For example, we don't want the distribution to be too spiky - although we do want our model to be able to infer spikes in a distribution if the evidence in the data outweighs our prior assumptions. We will now give some concrete examples of how this is done.

Following the approach of [Hol+20], we have independent samples $y_1, y_2, \ldots, y_n \in [0,1]$, and we want to infer the Lebesgue probability density $p : D \to [0, \infty]$, where we will consider D to be a closed interval of $\mathbb{R}$. We denote by P the set of all possible densities.

$$P := \left\{ p : D \to [0,\infty] \middle| \int_D p(y)dy = 1 \right\}$$

Two difficulties when sampling this distribution are ensuring that the density is non-negative, and ensuring that it integrates to 1. Firstly I applied the approach of [Cot+13]. To ensure normalisation and non-negativity, they write

$$p(x) = \frac{\exp(u(x))}{\int_D \exp(u(s))ds} \tag{3}$$

We then infer the function $u : D \to \mathbb{R}$. I decided to use a basis $t \mapsto \sin(nt)$ for $n \geq 1$, although we will see that this approach has limitations. We write

$$u(t) = \sum_{n=1}^{d} a_n \sin(nt)$$

For the prior, I put a Gaussian process prior μ_0 on u - as we are working in finite dimensions this is equivalent to putting a normal distribution on each of the coefficients. In particular, I took $a_n \sim N(0, 1/n^2)$. Some of the reasoning behind this is firstly that variance decreases for higher frequencies - this has a regularising effect, making spikes less likely. Furthermore, the choice of $1/n^2$ is useful in high/infinite dimensions, as our covariance operator will be trace-class, as $\sum_{n\geq 1} 1/n^2 = \pi^2/6 < \infty$. It is related to the Laplacian operator as $\sin(nx) + \Delta \sin(nx)/n^2 = 0$. Our target distribution, the posterior distribution of u given $(y_i)_{i=1}^n$, $\mu = f_{U|\bar{Y}}$, can be rewritten using Bayes' formula.

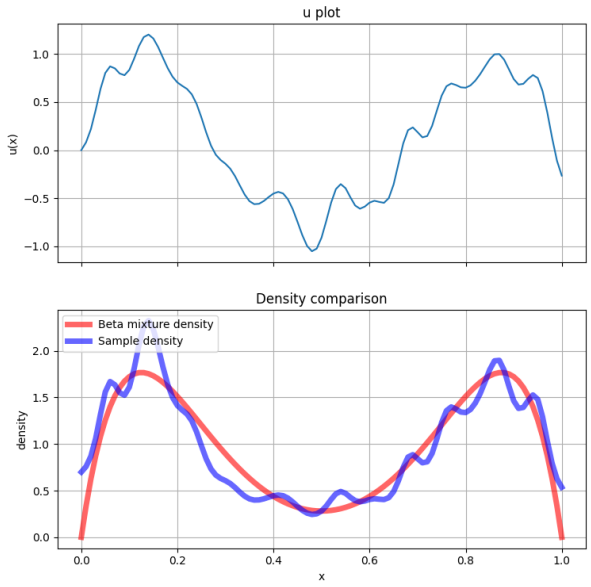

Figure 1: A single MCMC sample for u and the corresponding sample density, plotted against the true density. This using u as in (3).

$$\begin{aligned}\frac{f_{U|\bar{Y}}}{d\mu_0}(u|(y_i)_{i=1}^n) &\propto f_{\bar{Y}|U}((y_i)_{i=1}^n|u) \\ &= \prod_{i=1}^n p(y_i) \\ &= \exp\left(\sum_{i=1}^n \ln p(y_i)\right)\end{aligned}$$

So from our definition of the potential 2 we get $\Phi(u) = -\sum_{i=1}^n \ln p(y_i)$. We are now ready to apply our MCMC techniques. I took 100 samples mixture of two beta distributions, and ran the pCN MCMC algorithm for $d = 20$, so with sin waves up to $sin(20t)$. One sample is shown in figure 1. We can see that the sample fits to the true density fairly well, but we are still seeing lots of spikes. However, this is due to firstly the likelihood dominating the prior when there is a large amount of evidence (100 data points), and due to my naive choice of a basis for u which is struggling to form the shape of the distribution. Another way of solving this is to use a different covariance matrix, increasing the regularisation. One other problem is that the choice of parametrising p using u as in 3 isn't well suited to densities which are zero, or close to zero, at points.

We now explore an alternative for ensuring positivity, using $t \mapsto t^2$ instead of $t \mapsto \exp(t)$. We now write

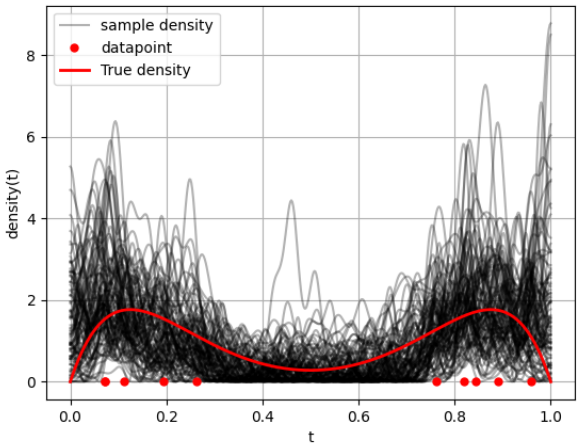

Figure 2: 1000 densities sampled using MCMC for u as in (4). Only 10 data points from the underlying Beta distribution were used.

$$p(t) = \frac{u^2(t)}{\int_D u^2(s)ds} \tag{4}$$

We will also use a different method to parameterise $u : D \to \mathbb{R}$, that used in [Hol+20]. We write

$$u(t) = u(t, \bar{a}) = \sum_{n=1}^{d} a_n \phi_i(t), \qquad \phi_1 \equiv 1, \quad \phi_i(t) = \sqrt{2}\cos(\pi(i-1)t),\ i \geq 2 \tag{5}$$

There is a slightly abuse of notation when we write $u(t) = u(t, \bar{a})$, but is useful for making a distinction between a random function $u(t)$, and a function with random coefficients, $u(t, \bar{a})$. We still have that $\Phi(u) = -\sum_{i=1}^{n} \ln p(y_i)$, and for now we keep the inverse Laplacian covariance, $a_n \sim N(0, 1/n^2)$. I ran our MCMC algorithms with this approach, using only 10 data points from the underlying beta distribution. Some of the density functions sampled are shown in figure 2. Considering we've only given the algorithm 10 samples, it is still able to sample many densities which have a similar overall shape to the beta distribution, while still allowing variance for other underlying distributions. Many of these are more spiky then we would like, so for later examples we change to a prior which has a stronger regularising effect, which helps suppress the spikes in line with our prior beliefs about how smooth a density function should be. Before we explore these results in more detail, we provide some motivation for why we would want to create an MCMC algorithm on the sphere, as opposed to in $\mathbb{R}^d$.

One downside of our target u as in 4 is that, due to the normalisation (that is, the division by the integral of the density), the densities p_n could fail to be a Markov chain despite

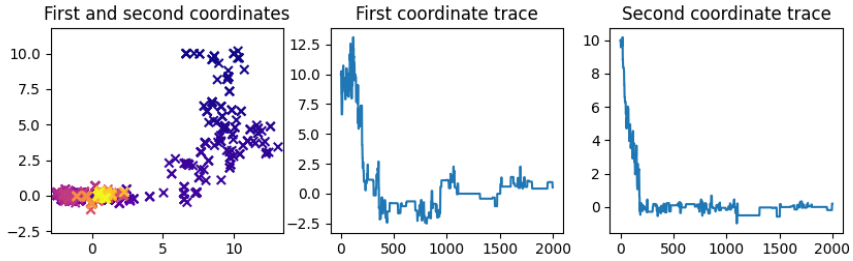

Figure 3: Trace plots for the first 2000 samples for the SRW-MH algorithm with 10 data points from the underlying beta distribution, and dimension 10.

the u_n forming a Markov chain, as per [Lie+23] appendix D. One solution is to perform the MCMC on a manifold where the normalisation integral is one. The overall approach is then to infer the square root g, where $g^2 = p$ and belongs to

$$Q := \left\{ g : D \to \mathbb{R} \middle| \int_D g^2(y)dy = 1 \right\}$$

That is, g is in the unit sphere of $L^2(D)$. One observation we make at this point is that $g^2 = (-g)^2 = p$, so we will need to take this into account choosing our prior and basis for $L^2(D)$.

3.2 MCMC in practice

We recall that our aim when running these Monte Carlo simulations is to approximate an expectation. In this case, our quantity of interest is the probability that a sample from the unknown distribution is less than half (we know this to be 0.5 by construction). That is, we want to approximate the following expectation over $u \sim \mu$ by taking random samples of $u_n \sim \mu$.

$$\mathbb{E}\left[\int_0^{0.5} u(y)dy\right] \approx \frac{1}{N}\sum_{n=1}^{N}\int_0^{0.5} u_n(y)dy$$

Before we return to our example, we discuss some of the practical computational considerations when implementing MCMC. The first is burn-in. When a chain is initiated in practice, it is often the case that the location of the highest density areas of the target distribution is unknown. Hence, there is a period of transience where algorithms move from low to high density areas, see figure 3. We wouldn't expect to get many, if any, samples from these low density regions, so for this reason, the first samples are often discarded. In all the algorithms today, I took 110,000 samples and discarded the first 10,000. There is no real theoretical justification for using burn-in, but from the figure we can appreciate why this is done.

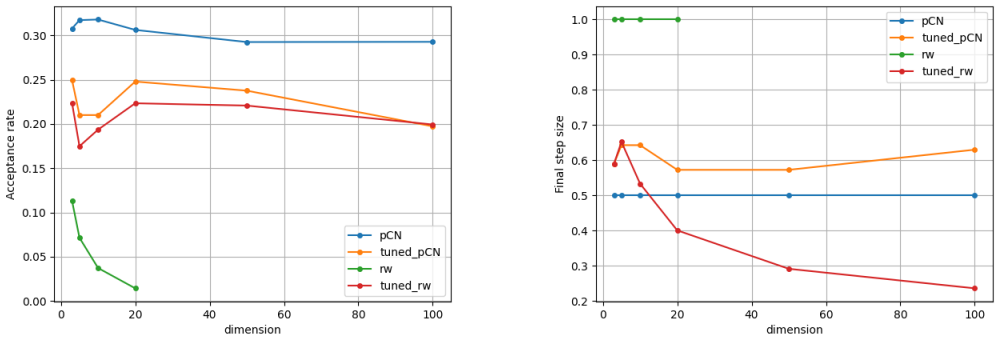

Figure 4: Left: acceptance rate vs dimension for the different algorithms. Right: final step size vs dimension, for untuned algorithms the step size is constant.

At this point we first encounter a weakness of the SRW-MH algorithm. As the dimension increases, the number of proposals that are accepted tends to zero, as is shown in figure 4. This is a problem, as our chain stops mixing well and the samples become highly correlated. One solution to this problem is to tune the step size of the algorithm to ensure that the acceptance rate is kept near an 'optimal' level. A common choice of optimal acceptance rate is 23%, as shown in [GRG96]. There are different ways of tuning the step size, but my method was to check how many proposals were being accepted every 500 samples. If it was less than 18%, I decreased the step size. If it was over 28%, I decreased the step size. This approach can be implemented for both the SRW-MH algorithm and the pCN algorithm, however it is important to note that their step sizes are in the ranges $(0, \infty)$ and $(0, 1]$ respectively. This tuning does influence the kernel, and hence to ensure we still have the correct stationary distribution, tuning stops at the end of the burn-in period.

While this tuning does help with the acceptance rate problem, we now have another issue - the step size is tending to zero as the dimension increases. This also increases the correlation of our samples. Amazingly, the both the pCN and tuned pCN algorithms are suffering from neither of these problems, and hence they have been crowned with the title 'dimension-independent algorithms'. The theoretical justification for this was provided by Martin Hairer, Andrew J. Steward and Sebastian J. Vollmer in [HSV14]. To over summarise, they show that these algorithms have a dimension-independent spectral gap, and hence dimension-independent autocorrelation times and central limit theorem type results. We can see this clearly when looking at the autocorrelation, integrated autocorrelation time (IACT) and effective sample size (ESS) of our algorithms, as shown in figure 5. For reference, for autocorrelation, 0 is for independent samples, and 1 is for perfectly correlated samples. This can be used to ESS, a statistic which tries to tell us the number of independent samples that our 100,000 correlated samples are equivalent to, in terms of information. The IACT is a statistic which relates to the sampling variance. We

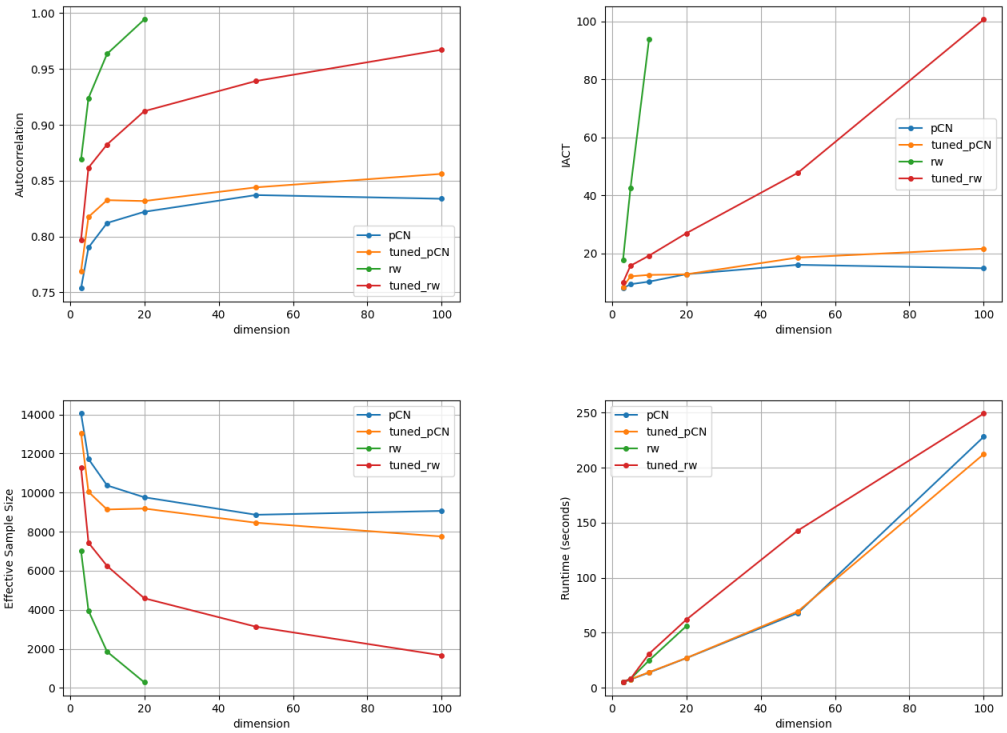

Figure 5: Plots of correlation and runtime vs dimension. Top left: autocorrelation. Top right: IACT. Bottom left: ESS. Bottom right: runtime.

won't cover it here in detail, but it can be thought of as representing the number of steps of the chain needed to 'forget' where it started, so that we can take another 'independent' sample.

Also shown in figure 5 is how the runtime increases with dimension. We briefly mention some techniques used to aid in the computation. Firstly, it is important to compute everything on the logarithm scale, as we often dealing with incredibly small or large densities, especially in high dimensions. For example, for the acceptance of a proposal, instead of taking comparing $u \sim U(0,1)$ to α, we compare $\log(u)$ to $\log(\alpha)$. This prevents overflow in the computer's memory. There are other techniques that we will not mention here, for example the use of Cholesky decompositions for sampling from multivariate normal distributions. These techniques become more necessary in higher dimensions, where some of the algorithms presented here took over 40 minutes to run.

4 MCMC on the Sphere

We now turn our attention to performing MCMC on the sphere, which we define as

$$\mathbb{S}^{d-1} := \left\{ x \in \mathbb{R}^d \middle| \|x\| := \sum_{i=1}^{d} x_i^2 = 1 \right\}$$

Throughout the rest of the dissertation, I will follow the notation in [Lie+23] by denoting elements, sets and kernels on the sphere with a bar, for example $\bar{x} \in \mathbb{S}^{d-1}$, $\bar{A} \in \mathcal{B}(\mathbb{S}^{d-1})$ and $\bar{K} : \mathbb{S}^{d-1} \times \mathcal{B}(\mathbb{S}^{d-1}) \to [0, 1]$. I hope this improves the clarity of the presented content.

4.1 A brief probability theory excursion

We first discuss the Law of the Unconscious Statistician - this will be one of the tools we use the most in this section. Suppose we have a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, a measurable space $(F, \mathcal{F})$, and a random variable (measurable function) $X : \Omega \to F$. Suppose further that we have a measurable function $g : F \to \mathbb{R}$. Hence $g \circ X$ is a real-valued random variable. We recall the definition of the expected value of a random variable.

$$\mathbb{E}[g(X)] := \int_\Omega g(X(\omega))\mathbb{P}(d\omega)$$

The Law of the Unconscious statistician allows us to rewrite this as follows.

$$\int_\Omega g(X(\omega))\mathbb{P}(d\omega) = \int_F g(x) X_\# \mathbb{P}(dx)$$

We recall that the pushforward of a measure is defined by $X_\#\mathbb{P}(A) := \mathbb{P}(X^{-1}(A))$. In the case that X has a density function ν in F, we can replace $X_\#\mathbb{P}(dx)$ with $\nu(dx)$. The reason that this is called the law of the unconscious statistician is that it's often used as if it were the definition of the expectation of $\mathbb{E}[g(X)]$, and not a result of this law. We start the proof with the case that $g = \mathbb{I}_A$, for some measurable $A \in \mathcal{F}$.

$$\int_F g(x) X_\# \mathbb{P}(dx) = \int_F \mathbb{I}_A(x) X_\# \mathbb{P}(dx) = X_\#\mathbb{P}(A) = \mathbb{P}(X^{-1}(A))$$

$$\begin{aligned}
\int_\Omega g(X(\omega))\mathbb{P}(d\omega) &= \int_\Omega \mathbb{I}_A(X(\omega))\mathbb{P}(d\omega) \\
&= \int_\Omega \mathbb{I}_{X^{-1}(A)}(\omega)\mathbb{P}(d\omega) \\
&= \mathbb{P}(X^{-1}(A))
\end{aligned}$$

Now for the case that g is a simple function, that is $g = \sum_{k=1}^{m} a_k \mathbb{I}_{A_k}$, for some $k \in \mathbb{N}$, $m \in \mathbb{N}$, and for all k, $a_k \in \mathbb{R}$ and $A_k \in \mathcal{F}$. By linearity, and using the result for indicator functions, we have the following.

$$\begin{aligned}
\int_\Omega g(X(\omega))\mathbb{P}(d\omega) &= \int_\Omega \sum_{k=1}^{m_k} a_k \mathbb{I}_{A_k}(X(\omega))\mathbb{P}(d\omega) \\
&= \sum_{k=1}^{m_k} a_k \int_\Omega \mathbb{I}_{A_k}(X(\omega))\mathbb{P}(d\omega) \\
&= \sum_{k=1}^{m_k} a_k \int_\Omega \mathbb{I}_{A_k}(x) X_\# \mathbb{P}(dx) \\
&= \int_\Omega \sum_{k=1}^{m_k} a_k \mathbb{I}_{A_k}(x) X_\# \mathbb{P}(dx) \\
&= \int_\Omega g(x) X_\# \mathbb{P}(dx)
\end{aligned}$$

Now we prove the law for non-negative measurable functions g. We take an increasing sequence of non-negative simple functions $g_n \uparrow g$, for example $g_n := (2^{-n}\lfloor 2^n g(x) \rfloor) \wedge n$. In the first equality we use the monotone convergence theorem, noting that $g_n \circ X$ is both measurable and increasing in n.

$$\begin{aligned}
\int_\Omega g(X(\omega))\mathbb{P}(d\omega) &= \lim_{n\to\infty} \int_\Omega g_n(X(\omega))\mathbb{P}(d\omega) \\
&= \lim_{n\to\infty} \int_\Omega g_n(x) X_\# \mathbb{P}(dx) \\
&= \int_\Omega g(x) X_\# \mathbb{P}(dx)
\end{aligned}$$

It is simple to extend this result from non-negative to real-valued functions, but it is not required for our analyses.

4.2 The sphere

At this point we turn our attention to performing MCMC on our manifold of choice, the sphere. This section will in large part follow the work of H. C. Lie, D. Rudolf, B. Sprungk and T. J. Sullivan [Lie+23]. I will endeavour to make it clear throughout whether the content is theirs, or my own.

We use the following set up. Suppose we are looking to modify a Markov chain on $\mathbb{R}^d$, where the dimension d is large, so that it takes values only on the sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, where $||\cdot||$ is the normal Euclidean norm. We take the normal Borel σ-algebra on $\mathbb{R}^d$, and for $\mathbb{S}^{d-1}$ we take the Borel σ-algebra generated by the subspace topology. To project values from $\mathbb{R}^d$ onto $\mathbb{S}^{d-1}$, we define the map $\Pi : \mathbb{R}^d \to \mathbb{S}^{d-1}$ as follows.

$$\Pi(x) := \begin{cases} \frac{x}{||x||} & \text{if } x \neq 0 \\ e_1 & \text{if } x = 0 \end{cases}$$
$$e_1 := (1, 0, 0, \ldots, 0) \in \mathbb{S}^{d-1}$$

We first prove that $\Pi : \mathbb{R}^d \to \mathbb{S}^{d-1}$ is measurable. There is a proof presented in [Lie+23], however this one presented here is my own.

As we have the Borel σ-algebra generated by the subspace topology, it suffices to show that for any open set $A \subset \mathbb{R}^d$, $\Pi^{-1}(A \cap \mathbb{S}^{d-1})$ is in $\mathcal{B}(\mathbb{R}^d)$. If $A \cap \mathbb{S}^{d-1} = \emptyset$, then $\Pi^{-1}(A \cap \mathbb{S}^{d-1}) = \emptyset \in \mathcal{B}(\mathbb{R}^d)$. We first consider the case where $e_1 \notin A$, so $\Pi^{-1}(A \cap \mathbb{S}^{d-1}) = \{x \in \mathbb{R}^d \setminus \{0\} : \Pi(x) \in A\}$. We aim to show this set is open, and hence measurable. Take $y \in \Pi^{-1}(A \cap \mathbb{S}^{d-1})$, so $\Pi(y) \in A$. As A is open, there exists $\delta > 0$ such that the open ball $B_\delta(\Pi(y)) \subset A$. Let $\epsilon = \min(\frac{||y||\delta}{2}, \frac{||y||}{2})$. We claim that $B_\epsilon(y) \subset \Pi^{-1}(A \cap \mathbb{S}^{d-1})$. Let $x \in B_\epsilon(y)$. As $||x|| > ||y|| - ||y - x|| > ||y|| - \epsilon \geq ||y||/2$, we know $x \neq 0$. We aim to show $||\Pi(x) - \Pi(y)|| < \delta$, as then $\Pi(x) \in B_\delta(\Pi(y)) \subset A$, implying $x \in \Pi^{-1}(A \cap \mathbb{S}^{d-1})$. We compute the following, using the triangle and reverse triangle inequalities on the second and fourth lines respectively.

$$\begin{aligned} ||\Pi(x) - \Pi(y)|| &= \left|\left| \frac{x}{||x||} - \frac{y}{||y||} \right|\right| \\ &\leq \left|\left| \frac{x}{||x||} - \frac{x}{||y||} \right|\right| + \left|\left| \frac{x}{||y||} - \frac{y}{||y||} \right|\right| \\ &\leq \frac{||y|| - ||x||}{||y||} + \frac{||x - y||}{||y||} \\ &\leq \frac{2||x - y||}{||y||} \\ &< \frac{2\epsilon}{||y||} \leq \delta \end{aligned}$$

So we have shown $A \subset \mathbb{R}^d$, $\Pi^{-1}(A \cap \mathbb{S}^{d-1})$ is open, and hence measurable, if $e_1 \notin A$. If $e_1 \in A$, then we have $\Pi^{-1}(A \cap \mathbb{S}^{d-1}) = \{x \in \mathbb{R}^d \setminus \{0\} : \Pi(x) \in A\} \cup 0$, which is the union of two measurable sets, and hence itself measurable. So we have shown that Π is measurable.

We now define the notion of a pushforward measure. Suppose we have measurable spaces $(E, \mathcal{E})$, $(F, \mathcal{F})$, a measure ν on $\mathcal{E}$, and a measurable function $T : E \to F$. Then we define the pushforward measure $T_\# \nu := \nu \circ T^{-1}$ on $\mathcal{F}$. That is, for all $A \in \mathcal{F}$,

$$T_\# \nu(A) = \nu(T^{-1}(A))$$

This brings us to a key object of interest, the angular central Gaussian distribution (ACG). Suppose that we have a central (mean 0) Gaussian distribution with covariance matrix C on $\mathbb{R}^d$, call it $\nu_0 = N(0, C)$. Then we define

$$\mu_0 := \Pi_{\#}\nu_0 = \Pi_{\#}N(0, C)$$

to be the angular central Gaussian distribution with covariance C. We will denote this, as in [Lie+23], as $\mu_0 = ACG(C)$. A crucial property of this distribution is that it is antipodally symmetric. That is, $\mu_0(x)dx = \mu_0(-x)dx$. This will be a desirable property for our applications. This is used as a reference measure to define our target distribution, μ, via the Radon-Nikodym derivative.

$$\frac{d\mu}{d\mu_0}(\bar{x}) := \exp(-\bar{\Phi}(\bar{x})) \tag{6}$$

Where $\bar{\Phi} : \mathbb{S}^{d-1} \to \mathbb{R}$ be a measurable function, commonly referred to as a potential. The Bayesian interpretation of this statement is that we have an $\mu_0 = ACG(C)$ prior, a likelihood $\bar{x} \mapsto \exp(-\bar{\Phi}(\bar{x}))$, and a posterior (target) distribution μ. The first key proposition of our spherical MCMC algorithm is that we can lift our sampling problem on $\mathbb{S}^{d-1}$ to $\mathbb{R}^d$ as follows.

Proposition. Suppose $\nu_0 = N(0, C)$, and define $\Phi : \mathbb{R}^d \to \mathbb{R}$ by $\Phi(x) := \bar{\Phi}(\Pi(x))$. Furthermore, we define a measure ν via its Radon-Nikodym derivative.

$$\frac{d\nu}{d\nu_0}(x) := \frac{1}{Z}\exp(-\Phi(x))$$

Where Z is a normalising constant. Then the following equality holds μ_0-almost everywhere.

$$\frac{d\Pi_{\#}\nu}{d\mu_0}(x) = \frac{1}{Z}\exp(-\bar{\Phi}(\bar{x}))$$

This tells us that $\mu = \Pi_{\#}\nu$ μ_0-almost everywhere. Therefore, we can move our problem from sampling μ on $\mathbb{S}^{d-1}$ to sampling ν in $\mathbb{R}^d$, then project our samples back onto the sphere.

Proof. By the uniqueness of the Radon-Nikodym derivative, it suffices to show that for all $A \in \mathcal{B}(\mathbb{S}^{d-1})$, the following holds.

$$\Pi_{\#}\nu(A) = \int_A \frac{1}{Z}\exp(-\bar{\Phi}(\bar{x}))\mu_0(dx)$$

We define random variables (measurable functions) $X \sim \nu_0$, and $\bar{X} = \Pi(X) \sim \mu_0$. We compute the following, using the Law of the Unconscious Statistician (LOTUS) twice.

$$\begin{aligned}
\Pi_{\#}\nu(A) &= \nu(\Pi^{-1}(A)) \\
&= \int_{\Pi^{-1}(A)} \frac{d\nu}{d\nu_0}(x)\nu_0(dx) \\
&= \int_{\mathbb{R}^d} \mathbb{I}_{\Pi^{-1}(A)}(x)\frac{1}{Z}\exp(-\Phi(x))\nu_0(dx) \\
&= \int_{\mathbb{R}^d} \mathbb{I}_{A}(\Pi(x))\frac{1}{Z}\exp(-\bar{\Phi}(\Pi(x)))\nu_0(dx) \\
\text{(LOTUS with } X) \quad &= \mathbb{E}[\mathbb{I}_A(\Pi(X))\frac{1}{Z}\exp(-\bar{\Phi}(\Pi(X)))] \\
&= \mathbb{E}[\mathbb{I}_A(\bar{X})\frac{1}{Z}\exp(-\bar{\Phi}(\bar{X}))] \\
\text{(LOTUS with } \bar{X}) \quad &= \int_{\mathbb{S}^{d-1}} \mathbb{I}_A(\bar{x})\frac{1}{Z}\exp(-\bar{\Phi}(\bar{x}))\mu_0(d\bar{x}) \\
&= \int_A \frac{1}{Z}\exp(-\bar{\Phi}(\bar{(x)}))\mu_0(d\bar{x})
\end{aligned}$$

We define the reprojected transition kernel $\bar{K} : \mathbb{S}^{d-1} \times \mathcal{B}(\mathbb{S}^{d-1}) \to [0,1]$ according to [Lie+23]. Define random variables $X \sim \nu$. For each $\bar{x} \in \mathbb{S}^{d-1}$, $\bar{A} \in \mathcal{B}(\mathbb{S}^{d-1})$,

$$\bar{K}(\bar{x}, \bar{A}) := \mathbb{E}[K(X, \Pi^{-1}(\bar{A}))|\Pi(X) = \bar{x}]$$

One interpretation for this is that you project, according to ν, $\bar{x}$ into the latent space $\mathbb{R}^d$, make a move according to one of our Metropolis-Hastings transition kernels, then reproject back onto the sphere. This process will be made clear when we discuss the MCMC reprojection algorithm later. For a further discussion of the intuition behind this choice, and of some other transition kernels on the sphere, I refer the reader to [Lie+23].

To make this expression more tractable, and to allow us to sample from it, we want to find the regular conditional distribution of $X|T(X)$. That is, we want to find a transition kernel (also known as a probability or Markov kernel) $\nu_{|\Pi} : \mathbb{S}^{d-1} \times \mathcal{B}(\mathbb{R}^d)$ such that

$$\nu_{|\Pi}(T(X), A) = \mathbb{P}(X \in A|T(X)) \qquad \mathbb{P}\text{-almost surely}$$

Theorem 8.5 of [Kal21] tells us that as we are using the Borel σ-algebra, this kernel $\nu_{|\Pi}$ exists and is μ-almost surely unique. Furthermore, we also get, as in [RS22], that we can disintegrate our measure ν with respect to Π to rewrite our transition kernel $\bar{K}$ as follows.

$$\bar{K}(\bar{x}, \bar{A}) := \mathbb{E}[K(X, \Pi^{-1}(\bar{A}))|\Pi(X) = \bar{x}] = \int_{\mathbb{R}^d} K(x, \Pi^{-1}(\bar{A}))\nu_{|\Pi}(\bar{x}, dx)$$

Armed with this, we prove that $\bar{K}$ is indeed a kernel: that $\bar{x} \mapsto \bar{K}(\bar{x}, \bar{A})$ is measurable for all $\bar{A} \in \mathcal{B}(\mathbb{S}^{d-1})$; and that $\bar{A} \mapsto \bar{K}(\bar{x}, \bar{A})$ is a probability measure for all $\bar{x} \in \mathbb{S}^{d-1}$. I believe this work, albeit basic, is my own - I haven't found it stated or referenced in either [Lie+23] or [RS22]. We start by showing $\bar{A} \mapsto \bar{K}(\bar{x}, \bar{A})$ is a probability measure for all $\bar{x} \in \mathbb{S}^{d-1}$. Let $\bar{x} \in \mathbb{S}^{d-1}$. We need to show three properties: non-negativity, countable additivity and that $\bar{K}(\bar{x}, \mathbb{S}^{d-1}) = 1$. These are all inherited from the properties of K.

$$\bar{K}(\bar{x}, \bar{A}) = \mathbb{E}[K(X, \Pi^{-1}(A)|\Pi(X) = \bar{x}] \geq \mathbb{E}[0|\Pi(X) = \bar{x}] = 0$$

We now let $(\bar{A}_k)_{k\geq 1} \subset \mathcal{B}(\mathbb{S}^{d-1})$ be a countable sequence of disjoint measurable sets.

$$\begin{aligned}
\bar{K}(\bar{x}, \cup_{k\geq 1}\bar{A}_k) &= \mathbb{E}[K(X, \Pi^{-1}(\cup_{k\geq 1}\bar{A}_k))|\Pi(X) = \bar{x}] \\
&= \mathbb{E}[K(X, \cup_{k\geq 1}\Pi^{-1}(\bar{A}_k))|\Pi(X) = \bar{x}] \\
&= \mathbb{E}\left[\sum_{k\geq 1} K(X, \Pi^{-1}(\bar{A}_k))|\Pi(X) = \bar{x}\right] \\
\text{(Monotone convergence)} \quad &= \sum_{k\geq 1} \mathbb{E}[K(X, \Pi^{-1}(\bar{A}_k))|\Pi(X) = \bar{x}] \\
&= \sum_{k\geq 1} \bar{K}(\bar{x}, \bar{A}_k)
\end{aligned}$$

$$\begin{aligned}
\bar{K}(\bar{x}, \mathbb{S}^{d-1}) &= \mathbb{E}[K(X, \Pi^{-1}(\mathbb{S}^{d-1}))|\Pi(X) = \bar{x}] \\
&= \mathbb{E}[K(X, \mathbb{R}^d)|\Pi(X) = \bar{x}] \\
&= \mathbb{E}[1|\Pi(X) = \bar{x}] = 1
\end{aligned}$$

We now move onto proving $\bar{x} \mapsto \bar{K}(\bar{x}, \bar{A})$ is measurable for all $\bar{A} \in \mathcal{B}(\mathbb{S}^{d-1})$. We take an increasing sequence of non-negative simple functions

$$f_n(\cdot) := \sum_{k=1}^{m_n} a_{n,k}\mathbb{I}_{A_{n,k}}(\cdot) \uparrow K(\cdot, \Pi^{-1}(\bar{A}))$$

For some $m_n \in \mathbb{N}$, $a_{n,k} \in \mathbb{R}_{\geq 0}$ and $A_{n,k} \in \mathcal{B}(\mathbb{R}^d)$. We compute,

$$\begin{aligned}
\bar{K}(\bar{x}, \bar{A}) &= \int_{\mathbb{R}^d} K(x, \Pi^{-1}(A)) \nu_{|\Pi}(\bar{x}, dx) \\
&= \lim_{n\to\infty} \int_{\mathbb{R}^d} f_n(x) \nu_{|\Pi}(\bar{x}, dx) \\
&= \lim_{n\to\infty} \sum_{k=1}^{m_n} a_{n,k} \int_{\mathbb{R}^d} \mathbb{I}_{A_{n,k}}(x) \nu_{|\Pi}(\bar{x}, dx) \\
&= \lim_{n\to\infty} \sum_{k=1}^{m_n} a_{n,k} \nu_{|\Pi}(\bar{x}, A_{n,k})
\end{aligned}$$

At which point we are done, as $\bar{x} \mapsto \nu_{|\Pi}(\bar{x}, A_{n,k})$ is measurable, linear combinations of measurable functions are measurable, and limits of measurable functions are measurable.

We now prove some of the most important results for this method, provided by [RS22]. The proofs shown here are based on those by Rudolf and Sprungk's. However, a key contribution of mine has been filling in many of the details to make the proofs accessible to the target audience of this dissertation, Master's students.

Firstly, we prove the following. Suppose K is ν-reversible, and that $\mu = \Pi_{\#}\nu$. Then $\bar{K}$ is μ-reversible. This is equivalent to proving that for any $\bar{A}, \bar{B} \in \mathcal{B}(\mathbb{S}^{d-1})$, the following holds.

$$\int_{\bar{A}} \bar{K}(\bar{x}, \bar{B}) \mu(d\bar{x}) = \int_{\bar{B}} \bar{K}(\bar{x}, \bar{A}) \mu(d\bar{x})$$

We start by defining the random variables $X \sim \nu$, $\bar{X} = \Pi(X) \sim \Pi_{\#}\nu = \mu$. We compute the following. (1) follows using the Law of the Unconscious Statistician (see appendix) with the random variable $\bar{X}$ and measurable non-negative function $\bar{x} \mapsto \mathbb{I}_{\bar{A}}(\bar{x}) \bar{K}(\bar{x}, \Pi^{-1}(\bar{B}))$. (2) follows as $\mathbb{I}_{\bar{A}}(\Pi(X))$ is $\Pi(X)$-measurable, hence we can move it inside the expectation. (3) is an application of the law of total expectation, also known as the tower property. (4) is another application of the law of the unconscious statistician, this time using the random variable X, and measurable function $x \mapsto \mathbb{I}_{\Pi^{-1}(\bar{A})}(x) K(x, \Pi^{-1}(\bar{B}))$.

$$
\begin{aligned}
\int_{\bar{A}} \bar{K}(\bar{x}, \Pi^{-1}(\bar{B}))\mu(d\bar{x}) &= \int_{\mathbb{S}^{d-1}} \mathbb{I}_{\bar{A}}(\bar{x})\bar{K}(\bar{x}, \Pi^{-1}(\bar{B}))\mu(d\bar{x}) \\
(1) \quad &= \int_{\Omega} \mathbb{I}_{\bar{A}}(\bar{X}(\omega))\bar{K}(\bar{X}(\omega), \Pi^{-1}(\bar{B}))\mathbb{P}(d\omega) \\
&= \mathbb{E}\left[\mathbb{I}_{\bar{A}}(\bar{X})\bar{K}(\bar{X}, \Pi^{-1}(\bar{B}))\right] \\
&= \mathbb{E}\left[\mathbb{I}_{\bar{A}}(\Pi(X))\mathbb{E}[K(X, \Pi^{-1}(\bar{B}))|\Pi(X)]\right] \\
(2) \quad &= \mathbb{E}\left[\mathbb{E}[\mathbb{I}_{\bar{A}}(\Pi(X))K(X, \Pi^{-1}(\bar{B}))|\Pi(X)]\right] \\
(3) \quad &= \mathbb{E}[\mathbb{I}_{\bar{A}}(\Pi(X))K(X, \Pi^{-1}(\bar{B}))] \\
(4) \quad &= \int_{\mathbb{R}^d} \mathbb{I}_{\Pi^{-1}(\bar{A})}(x)K(x, \Pi^{-1}(\bar{B}))\nu(dx) \\
&= \int_{\Pi^{-1}(\bar{A})} K(x, \Pi^{-1}(\bar{B}))\nu(dx)
\end{aligned}
$$

An analogous result holds when $\bar{A}$ and $\bar{B}$ are interchanged, hence by using the fact that K is ν-invariant, we have

$$
\begin{aligned}
\int_{\bar{A}} \bar{K}(\bar{x}, \bar{B})\mu(d\bar{x}) &= \int_{\Pi^{-1}(\bar{A})} K(x, \Pi^{-1}(\bar{B}))\nu(dx) \\
&= \int_{\Pi^{-1}(\bar{B})} K(x, \Pi^{-1}(\bar{A}))\nu(dx) \\
&= \int_{\bar{B}} \bar{K}(\bar{x}, \bar{A})\mu(d\bar{x})
\end{aligned}
$$

We now know that our pushforward kernel has the correct invariant distribution. However, before we implement this algorithm we need to be able to sample from our regular conditional distribution $\nu_{|\Pi}$. The following results from [Lie+23] (Prop 3.4 and 3.5) allow us to do this. Firstly, we recall the setup to our MCMC problem. We have $\nu_0 = N(0, C)$, a measurable functions $\bar{\Phi}$ and $\Phi = \bar{\Phi} \circ \Pi$. We also have our target measure ν in the latent space $\mathbb{R}^d$, defined via its Radon-Nikodym derivative as

$$
\frac{d\nu}{d\nu_0}(x) := \frac{1}{Z}\exp(-\Phi(x))
$$

Suppose we have $X_0 \sim \nu_0$ and $X \sim \nu$. Using the disintegration theorem (theorem 8.5 of [Kal21]), there exist regular conditional probability kernels $\nu_{0|\Pi} : \mathcal{B}(\mathbb{S}^{d-1}) \times \mathbb{R}^d \to [0, 1]$ and $\nu_{|\Pi} : \mathcal{B}(\mathbb{S}^{d-1}) \times \mathbb{R}^d \to [0, 1]$ for the random variables $X_0|\Pi(X_0)$ and $X|\Pi(X)$ respectively. In particular, from [al04], we know $\nu_{|\Pi}$ is the unique (up to ν_0-almost every $x \in \mathbb{R}^d$) probability kernel such that the following equality holds for all $A \in \mathcal{B}(\mathbb{R}^d)$, $B \in \mathcal{B}(\mathbb{S}^{d-1})$.

$$\mathbb{P}(X \in A, \Pi(X) \in B) = \int_B \nu_{|\Pi}(\bar{x}, A)\mu(d\bar{x}) = \int_{\Pi^{-1}(B)} \nu_{|\Pi}(\Pi(x), A)\nu(dx) \tag{7}$$

Where the second equality comes from applying the LOTUS to lift the problem to the latent space $\mathbb{R}^d$. We also have that the following equality holds for all measurable $g : \mathbb{R}^d \to \mathbb{R}$.

$$\mathbb{E}[g(X_0)|\Pi(X_0)] = \int_{\mathbb{R}^d} g(x)\nu_{0|\Pi}(\Pi(X_0), dx)$$

In particular,

$$\mathbb{E}[\mathbb{I}_A(X_0)|\Pi(X_0)] = \int_{\mathbb{R}^d} \mathbb{I}_A(x)\nu_{0|\Pi}(\Pi(X_0), dx) = \nu_{0|\Pi}(\Pi(X_0), A) \tag{8}$$

With these results, we are ready to prove that $\nu_{|\Pi} = \nu_{0|\Pi}$ ν_0-almost everywhere. The following is based on the proof in [Lie+23], however some extra details are included. We use the ν_0-almost everywhere uniqueness of $\nu_{|\Pi}$ in 7. In particular, we will show the following equation holds.

$$\mathbb{P}(X \in A, \Pi(X) \in B) = \int_{\Pi^{-1}(B)} \nu_{0|\Pi}(\Pi(x), A)\nu(dx)$$

We compute the following, first using that $X \sim \nu$.

$$
\begin{aligned}
\mathbb{P}(X \in A, \Pi(X) \in B) &= \mathbb{P}(X \in A \cap \Pi^{-1}(B)) \\
&= \nu(A \cap \Pi^{-1}(B)) \\
&= \int_{\mathbb{R}^d} \mathbb{I}_A(x)\mathbb{I}_{\Pi^{-1}(B)}(x)\frac{d\nu}{d\nu_0}(x)\nu_0(dx) \\
&= \int_{\mathbb{R}^d} \mathbb{I}_A(x)\mathbb{I}_B(\Pi(x))\frac{1}{Z}\exp(-\Phi(x))\nu_0(dx) \\
\text{(LOTUS with } X_0) \quad &= \mathbb{E}[\mathbb{I}_A(X_0)\mathbb{I}_B(\Pi(X_0))\frac{1}{Z}\exp(-\Phi(X_0))] \\
\text{(Law of total expectation)} \quad &= \mathbb{E}[\mathbb{E}[\mathbb{I}_A(X_0)\mathbb{I}_B(\Pi(X_0))\frac{1}{Z}\exp(-\bar{\Phi}(\Pi(X_0)))|\Pi(X_0)]] \\
&= \mathbb{E}[\mathbb{I}_B(\Pi(X_0))\frac{1}{Z}\exp(-\bar{\Phi}(\Pi(X_0)))\mathbb{E}[\mathbb{I}_A(X_0)|\Pi(X_0)]] \\
\text{(By equation 8)} \quad &= \mathbb{E}[\mathbb{I}_B(\Pi(X_0))\frac{1}{Z}\exp(-\bar{\Phi}(\Pi(X_0)))\nu_{0|\Pi}(\Pi(X_0), A)] \\
\text{(LOTUS with } X_0) \quad &= \int_{\Pi^{-1}(B)} \frac{1}{Z}\exp(-\Phi(x))\nu_{0|\Pi}(\Pi(x), A)\nu_0(dx) \\
&= \int_{\Pi^{-1}(B)} \nu_{0|\Pi}(\Pi(x), A)\nu(dx)
\end{aligned}
$$

With this result we've simplified our problem of sampling $\nu_{|\Pi}$ to one of sampling $\nu_{0|\Pi}$. Our next proposition, 3.5 from [Lie+23], tells us how to do this. Beforehand, we briefly discuss the co-ordinate system for $\mathbb{R}^d$ and its effect on the probability densities.

Usually we think of the density of a normal distribution $N(0, C)$ in d dimensions with respect to the Lebesgue measure, and the volume form dx on $\mathbb{R}^d$, as

$$
f_X(x)dx = \frac{1}{(2\pi)^{d/2}\sqrt{||C||}}\exp(-\frac{1}{2}x^\top C^{-1}x)dx
$$

However, as we want to parameterise our point $x \in \mathbb{R}^d$ as $(r, \bar{x}) := (||x||, x/||x||) \in [0, \infty) \times \mathbb{S}^{d-1}$, we convert our volume form dx in $\mathbb{R}^d$ to $r^{d-1}drd\bar{x}$, where dr is a 1-form for the radius, and $d\bar{x}$ is a $(d-1)$-form on $\mathbb{S}^{d-1}$ for the direction. The presence of the r^{d-1} term is familiar from changes to polar coordinates and spherical co-ordinates, where we have $dA = rdrd\theta$ and $dV = r^2\sin(\phi)drd\phi d\theta$. In general, it comes from the determinant of the Jacobian for the change of coordinates map $x \mapsto (r, \bar{x}) = (||x||, x/||x||)$. So, with respect to this coordinate system our $N(0, C)$ distribution has density

$$
f_X(r, \bar{x})drd\bar{x} = r^{d-1}\frac{1}{(2\pi)^{d/2}\sqrt{||C||}}\exp(-\frac{1}{2}r^2\bar{x}^\top C^{-1}\bar{x})drd\bar{x}
$$

With this we can calculate the density of our prior, $\mu_0 = ACG(C) = \Pi_\# N(0, C)$. We use the substitution $r_2 := r^2$, $dr_2 = 2rdr$, along with the identity

$$\int_0^\infty r^{\alpha-1}\exp(-\beta r) = \frac{\Gamma(\alpha)}{\beta^\alpha}dr$$

$$\begin{aligned}
\mu_0(\bar{x})d\bar{x} &= \int_0^\infty r^{d-1}\frac{1}{(2\pi)^{d/2}\sqrt{||C||}}\exp(-\frac{1}{2}r^2\bar{x}^\top C^{-1}\bar{x})drd\bar{x} \\
&= \frac{1}{(2\pi)^{d/2}\sqrt{||C||}}\frac{1}{2}\int_0^\infty r_2^{\frac{d}{2}-1}\exp(-\frac{1}{2}r_2\bar{x}^\top C^{-1}\bar{x})dr_2d\bar{x} \\
&= \frac{1}{(2\pi)^{d/2}\sqrt{||C||}} \times \frac{1}{2} \times \frac{\Gamma(\frac{d}{2})}{(\frac{1}{2}\bar{x}^\top C^{-1}\bar{x})^{d/2}}d\bar{x} \\
&= \frac{\Gamma(\frac{d}{2})}{2\pi^{d/2}\sqrt{||C||}(\bar{x}^\top C^{-1}\bar{x})^{d/2}}d\bar{x}
\end{aligned}$$

We now prove the following proposition (3.5 from [Lie+23]). Let $\bar{x} \in \mathbb{S}^{d-1}$. We denote the conditional distribution of $X_0 \sim \nu_0$ given $\Pi(X_0) = \bar{x}$ by $\nu_{0|\Pi}(\bar{x}, \cdot)$. Let R be a non-negative random variable with distribution $R^2 \sim \text{Gam}(\frac{d}{2}, \frac{1}{2}\bar{x}^\top C^{-1}\bar{x})$, where $\text{Gam}(\alpha, \beta)$ represents a gamma distribution with shape parameter α and rate (inverse scale) parameter β. Then

$$\bar{x}R \sim \nu_{0|\Pi}(\bar{x}, \cdot)$$

Proof. Let $\bar{X}_0 = \Pi(X_0) \sim \mu_0$, and $R = ||X_0||$. We can assume $X_0 \neq 0$ as this event has probability 0, so $X_0 = R\bar{X}_0 \sim N(0, C)$. We write the joint density of R, $\bar{X}_0$ as

$$f_{R,\bar{X}_0}(r, \bar{x}) = r^{d-1}\frac{1}{(2\pi)^{d/2}\sqrt{||C||}}\exp(-\frac{1}{2}r^2\bar{x}^\top C^{-1}\bar{x})$$

So

$$f_{R|\bar{X}_0=\bar{x}}(r) = \frac{f_{R,\bar{X}_0}(r, \bar{x})}{\int_0^\infty f_{R,\bar{X}_0}(l, \bar{x})dl} \propto r^{d-1}\exp(-\frac{1}{2}r^2\bar{x}^\top C^{-1}\bar{x})$$

Hence using the monotonic transformation $r \mapsto r^2 =: r_2$, we can find the density of R^2 given $\Pi(X_0) = \bar{x}$ using the change of variables formula,

$$
\begin{aligned}
f_{R^2|\bar{X}_0=\bar{x}}(r_2) &= f_{R|\bar{X}_0=\bar{x}}(\sqrt{r_2}) \left| \frac{d}{dr_2}\sqrt{r_2} \right| \\
&\propto r_2^{\frac{d-1}{2}} \exp(-\frac{1}{2} r_2 \bar{x}^\top C^{-1} \bar{x}) \frac{1}{2\sqrt{r_2}} \\
&\propto r_2^{\frac{d}{2}-1} \exp(-\frac{1}{2} r_2 \bar{x}^\top C^{-1} \bar{x}) \\
&\propto \mathrm{Gam}(\frac{d}{2}, \frac{1}{2}\bar{x}^\top C^{-1}\bar{x})
\end{aligned}
$$

Hence, we have simplified our problem of sampling $\nu_{|\Pi}(\bar{x}, \cdot)$ to one of sampling from a gamma distribution. Armed with this, we can now create an algorithm for pCN on the sphere.

Algorithm 3 Reprojected pCN algorithm on $\mathbb{S}^{d-1}$

Require: prior covariance C, target potential $\bar{\Phi}$, step size $s \in (0, 1]$
  Initialise $\bar{x_0} = e_1 = (1, 0, 0, \dots) \in \mathbb{S}^{d-1}$
  for $k \in \mathbb{N}_0$ **do**
    Draw a sample $r_k^2 \sim \mathrm{Gam}(d/2, \frac{1}{2}\bar{x}_k^\top C^{-1}\bar{x}_k)$ and set $x_k = r_k \bar{x}_k$
    Draw a sample $w_k \sim N(0, C)$ and set $y_{k+1} := \sqrt{1-s^2}x_k + s w_k$
    Set $\bar{y}_{k+1} := y_{k+1} / \|y_{k+1}\|$
    Compute $\alpha := \min(1, \exp(\bar{\Phi}(\bar{x}_k) - \bar{\Phi}(\bar{y}_{k+1})))$
    Draw a sample $u \sim U[0, 1]$
    if $u \leq \alpha$ **then**
      Set $\bar{x}_{k+1} = \bar{y}_{k+1}$
    else
      Set $\bar{x}_{k+1} = \bar{x}_k$
    end if
  end for

Now, we want to make sure that all our hard work analysing $\bar{K}$ was worthwhile. That is, we need to check that the following equality holds for all $\bar{A} \in \mathcal{B}(\mathbb{S}^{d-1})$ and $\bar{x} \in \mathbb{S}^{d-1}$.

$$
\mathbb{P}(\bar{X}_{k+1} \in \bar{A} | \bar{X}_k = \bar{x}) = \bar{K}(\bar{x}, \bar{A}) = \mathbb{E}[K(X, \Pi^{-1}(\bar{A})) | \Pi(X) = \bar{x}]
$$

There is a proof for this in [RS22], however it is more general, complicated, and requires many prerequisite results. The proof I present here is my own, and is still general, in that it holds for any MH algorithm, not just the pCN algorithm. This is my most significant theoretical contribution. We define the following random variables corresponding to the algorithm above.

$$
(X_k | \bar{X}_k = \bar{x}) \sim \nu_{|\Pi}(\bar{x}, \cdot) \qquad (Y_{k+1} | X_k) \sim Q(X_k, \cdot)
$$

$$\bar{X}_{k+1}|\bar{Y}_{k+1},\bar{X}_k = \begin{cases} \bar{Y}_{k+1} & \text{with probability } \bar{\alpha}(\bar{X}_k,\bar{Y}_{k+1}) \\ \bar{X}_k & \text{with probability } 1-\bar{\alpha}(\bar{X}_k,\bar{Y}_{k+1}) \end{cases}$$

We start by taking the law of total probability over $\bar{Y}_{k+1}$.

$$\begin{aligned}\mathbb{P}(\bar{X}_{k+1}\in\bar{A}|\bar{X}_k=\bar{x}) &= \mathbb{E}[\mathbb{P}(\bar{X}_{k+1}\in\bar{A}|\bar{Y}_{k+1},\bar{X}_k)|\bar{X}_k=\bar{x}] \\ &= \mathbb{E}[\bar{\alpha}(\bar{X}_k,\bar{Y}_{k+1})\mathbb{I}_{\bar{A}}(\bar{Y}_{k+1})+(1-\bar{\alpha}(\bar{X}_k,\bar{Y}_{k+1}))\mathbb{I}_{\bar{A}}(\bar{X}_k)|\bar{X}_k=\bar{x}]\end{aligned}$$

We now use linearity to split this into a sum of two expectations, and use the law of total expectation on the value of X_k.

$$\begin{aligned}\mathbb{E}[\bar{\alpha}(\bar{X}_k,\bar{Y}_{k+1})\mathbb{I}_{\bar{A}}(\bar{Y}_{k+1})|\bar{X}_k=\bar{x}] &= \mathbb{E}[\mathbb{E}[\bar{\alpha}(\bar{X}_k,\Pi(Y_{k+1}))\mathbb{I}_{\bar{A}}(\Pi(Y_{k+1}))|X_k]|\bar{X}_k=\bar{x}] \\ \mathbb{E}[\bar{\alpha}(\bar{X}_k,\Pi(Y_{k+1}))\mathbb{I}_{\bar{A}}(\Pi(Y_{k+1}))|X_k] &= \int_{\mathbb{R}^d}\bar{\alpha}(\Pi(X_k),\Pi(y))\mathbb{I}_{\bar{A}}(\Pi(Y))Q(X_k,dy) \\ &= \int_{\Pi^{-1}(\bar{A})}\alpha(X_k,y)Q(X_k,dy)\end{aligned}$$

Similarly we find

$$\begin{aligned}\mathbb{E}[(1-\bar{\alpha}(\bar{X}_k,\bar{Y}_{k+1}))\mathbb{I}_{\bar{A}}(\bar{X}_k)|\bar{X}_k=\bar{x}] &= \mathbb{E}[\mathbb{E}[(1-\bar{\alpha}(\bar{X}_k,\bar{Y}_{k+1}))\mathbb{I}_A(\bar{X}_k)|X_k]|\bar{X}_k=\bar{x}] \\ \mathbb{E}[(1-\bar{\alpha}(\bar{X}_k,\bar{Y}_{k+1}))\mathbb{I}_{\bar{A}}(\bar{X}_k)|X_k] &= \mathbb{I}_{\Pi^{-1}(\bar{A})}(X_k)\int_{\mathbb{R}^d}(1-\bar{\alpha}(\Pi(X_k),\Pi(y)))Q(X_k,dy) \\ &= \mathbb{I}_{\Pi^{-1}(\bar{A})}(X_k)\int_{\mathbb{R}^d}(1-\alpha(X_k,y))Q(X_k,dy) \\ &= r(X_k)\mathbb{I}_{\Pi^{-1}(\bar{A})}(X_k)\end{aligned}$$

At this stage, recalling the definition of the Metropolis-Hastings kernel K, we observe

$$K(X_k,\Pi^{-1}(\bar{A})) = r(X_k)\mathbb{I}_{\Pi^{-1}(\bar{A})}(X_k)+\int_{\Pi^{-1}(\bar{A})}\alpha(X_k,y)Q(X_k,dy)$$

Hence we conclude

$$\begin{aligned}\mathbb{P}(\bar{X}_{k+1} \in \bar{A} | \bar{X}_k = \bar{x}) &= \mathbb{E}[K(X_k, \Pi^{-1}(\bar{A}))|\bar{X}_k = \bar{x}] \\ &= \int_{\Pi^{-1}(x)} K(x, \Pi^{-1}(\bar{A}))\nu_{|\Pi}(\bar{x}, dx) \\ &= \mathbb{E}[K(X, \Pi^{-1}(\bar{A}))|\Pi(X) = \bar{x}] \\ &= \bar{K}(\bar{x}, \bar{A})\end{aligned}$$

4.3 Example on $\mathbb{S}^{d-1}$

We illustrate this new algorithm on the sphere by an example used by Holbrook et al. [Hol+20] and Lie et al. [Lie+23]. The set up is the same as our last example, we are estimating an unknown density $p : [0, 1] \to [0, \infty)$ using data points $y_1, y_2, \ldots, y_n \in [0, 1]$. We use the British coal mine disaster data set, which provides the dates of 191 disasters recorded between 1850 and 1965. For simplicity, these are mapped onto the interval $[0, 1]$.

For the pCN, tuned pCN and tuned SRW-MH algorithms on $\mathbb{R}^d$, we have the same approach as in the last example, where we infer u as defined in 4 and 5. However, we now change to using a Whittle-Matérn C, where $C = \text{diag}(\lambda_i : i \in \mathbb{N})$ with

$$\lambda_i = \frac{1}{0.4 + 4\pi^2(i-1)^2}$$

For our spherical pCN algorithm, we aim to infer the square root of the density g, where $g^2 = p$ and belongs to

$$Q := \left\{ g : [0, 1] \to \mathbb{R} \middle| \int_0^1 g^2(y)dy = 1 \right\}$$

That is, g is in the unit sphere of $L^2([0, 1])$. At this point we discuss how we will map between our function $g \in L^2([0, 1])$ and our sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$. Given a point $\bar{x} = (\bar{x}_1, \cdots, \bar{x}_d) \in \mathbb{S}^{d-1}$, we get the unique function g given by

$$g(\bar{x}) = \sum_{i=1}^{d} \bar{x}_i \phi_i, \qquad \phi_1 \equiv 1, \quad \phi_i(t) = \sqrt{2}\cos(\pi(i-1)t),\ i \geq 2$$

It is easy to check using elementary trigonometric identities that the ϕ_i are orthonormal. Crucially, this is indeed a function in Q.

$$
\begin{aligned}
\int_0^1 g^2(y)dy &= \int_0^1 \left(\sum_{i=1}^d \bar{x}_i \phi_i(y)\right)^2 dy \\
&= \int_0^1 \sum_{i,j=1}^d \bar{x}_i \bar{x}_j \phi_i(y)\phi_j(y)dy \\
\text{(By Fubini-Tonelli)} \quad &= \sum_{i,j=1}^d \bar{x}_i \bar{x}_j \int_0^1 \phi_i(y)\phi_j(y)dy \\
\text{(By orthonormality)} \quad &= \sum_{i,j=1}^d \bar{x}_i \bar{x}_j \delta_i(j) \\
(\bar{x} \in \mathbb{S}^{d-1}) \quad &= \sum_{i=1}^d \bar{x}_i^2 = 1
\end{aligned}
$$

Note that as we are not in infinite-dimensional space, we don't need to use Fubini-Tonelli. Hence by taking a random element $\bar{X} \in \mathbb{S}^{d-1}$, we get a random function $y \mapsto g(y, \bar{X}) \in Q$. We observe that $g^2 = (-g)^2 = p$. This works perfectly with our angular central Gaussian prior, where antipodal points have the same density. Hence, we place a $\mu_0 = ACG(C)$ prior on $\bar{X}$. Following [Lie+23], for a given $g(\cdot, \bar{x}) \in Q$, the likelihood of observing the data is

$$
L((y_j)_{j=1}^n; \bar{x}) = \prod_{j=1}^n g^2(y_j) = \exp\left(\sum_{j=1}^n \ln(g^2(y_j))\right) = \exp\left(2\sum_{j=1}^n \ln(|g(y_j)|)\right)
$$

Which gives rise to our potential

$$
\bar{\Phi}(\bar{x}) := -2\sum_{j=1}^n \ln\left(\left|\sum_{i=1}^d \bar{x}_i \phi_i(y_j)\right|\right)
$$

Hence, we can now write the distribution of our coefficients $\bar{X} \sim \mu$ where μ is our target posterior distribution as defined in 6. We take the same quantity of interest as [Lie+23], namely the posterior expectation for the probability mass between 0.435 and 0.574, the probability of disaster between 1900 and 1916. That is, we approximate

$$
\begin{aligned}
\mathbb{E}\left[f(\bar{X})\right] &\approx \frac{1}{N}\sum_{n=1}^{100000} f(\bar{X}_n) \\
f(\bar{x}) &:= \int_{0.435}^{0.574} g^2(y, \bar{x})dy
\end{aligned}
$$

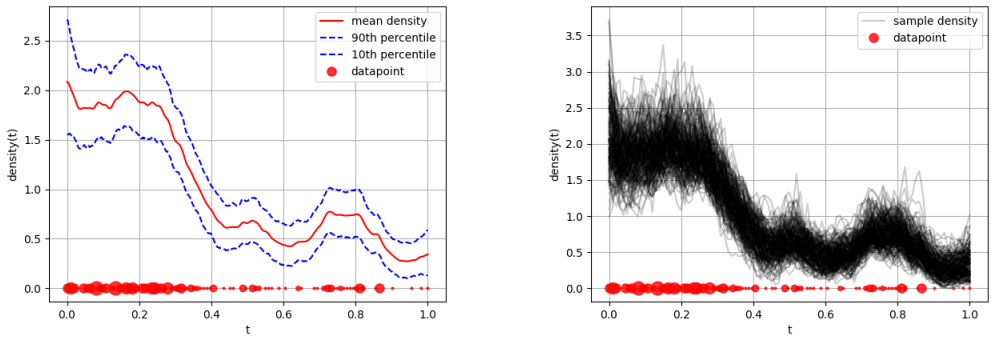

Figure 6: Plots of 1000 sampled densities, 1 sample taken every 1000 iterations. Left: mean density and percentiles. Right: sample densities.

For the simulations, I took dimensions $d = 10, 20, 30, 40, 50, 100$. Due to long run times on my computer, I was unable to test higher dimensions. I ran the pCN, tuned pCN and tuned SRW-MH algorithms on $\mathbb{R}^d$, and the pCN sphere algorithm. Due to the similarity in performance with the pCN and tuned pCN algorithm, the pCN algorithm was only run for dimensions 10 and 20. Due to the performance drop off, the SRW-MH algorithm was not ran for dimension 100. For each I ran 110,000 iterations, discarding the first 10,000 as burn-in resulting in 100,000 samples. Results are shown in figures 6 and 7.

Key features of these plots are as follows. Firstly, we see that the tuned SRW-MH algorithm is suffering from the same performance drop-off with increasing dimension that we observed for the first example in $\mathbb{R}^d$. One thing to note is that it was not able to successfully tune itself in 10,000 samples; we see that the acceptance rate at dimension 50 is less than 10% despite the tuning aiming for 18%-28%. I could've rerun all the tests with a longer burn in, however as can be seen in the runtime plot of figure 7, some of the tests take 40 minutes to run. With that said, we see that both the pCN, tuned pCN and spherical pCN all have dimension-independent performance, with the exception of runtime, as expected.

The key finding here is that the spherical pCN algorithm and the tuned pCN algorithm in $\mathbb{R}^d$ have very similar performance, at least in this setting. As stated in [Lie+23], one downside of the tuned pCN algorithm is that although the sampled coefficients $X_n \in \mathbb{R}^d$ form a Markov chain, the probability distributions they correspond to, $p(\cdot, X_n)$, do not. However, in practice we are trying to approximate the a posterior expectation of a quantity of interest, say $f : \mathbb{S}^{d-1} \to [0, \infty)$. Hence, even in the spherical setting, despite the coefficients $\bar{X}_n \in \mathbb{S}^{d-1}$ forming a Markov chain, $f(\cdot, \bar{X}_n)$ may not form a Markov chain. The approximation of this equivalent to us approximating $f \circ \Pi : \mathbb{R}^d \to [0, \infty)$ in $\mathbb{R}^d$.

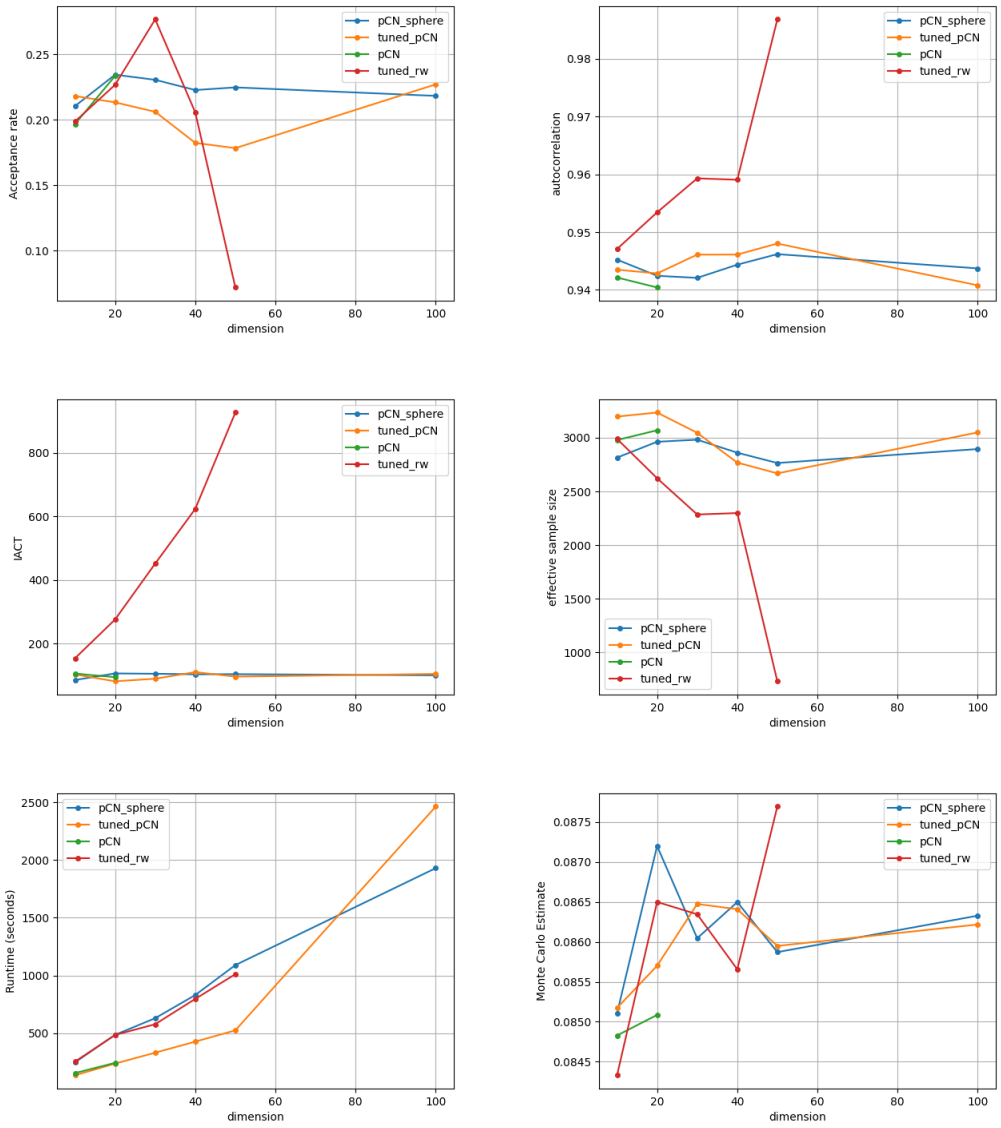

Figure 7: Plots displaying various statistics for algorithm performance against dimension for each algorithm. Bottom right: Monte Carlo estimate for quantity of interest.

5 Conclusion

In this report I have provided an introduction to the field of Markov chain Monte Carlo, with both theory and examples. Many of the proofs were from the literature, however there are some which are my own, a couple of which are non-trivial.

I also performed significant computational work. One key contribution is the development of an open source body of code available on my Github, Genomesh, which provides a pipeline for running MCMC algorithms both in $\mathbb{R}^d$ and on the sphere. We compared pCN algorithms to SRW-MH algorithms, and explored how their performance varies with dimension. We also explored how the spherical reprojection methods of H. C. Lie, D. Rudolf, B. Sprungk and T. J. Sullivan [Lie+23] compare to the standard projection methods on $\mathbb{R}^d$. Here, we found very similar performance between these two techniques, although we only applied this to one example.

However, the work presented in [Lie+23] is still very promising. As shown in their paper, it outperforms many other techniques on manifolds. There could also be many other cases where the reprojection method is superior to the standard projection method. Furthermore, combined with [RS22], they've laid a theoretical foundation for MCMC techniques on other manifolds which are potentially less amenable to standard techniques.

Finally, I would like to thank my supervisor Professor Tim J. Sullivan again. Without you this report wouldn't have been possible. Thank you.

Bibliography

References

[al04] D. Leao Jr. et al. "Regular conditional probability, disintegration of probability and Radon spaces". In: *Proyecciones* (2004). URL: https://www.scielo.cl/pdf/proy/v23n1/art02.pdf.

[Cot+13] S. L. Cotter et al. "MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster". In: *Statistical Science* 28.3 (2013). ISSN: 0883-4237. DOI: 10.1214/13-sts421. URL: http://dx.doi.org/10.1214/13-STS421.

[GRG96] A Gelman, G O Roberts, and W R Gilks. "Efficient Metropolis Jumping Rules". In: *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*. Oxford University Press, 1996. ISBN: 9780198523567. DOI: 10.1093/oso/9780198523567.003.0038.

[Hol+20] Andrew Holbrook et al. "Nonparametric Fisher Geometry with Application to Density Estimation". In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. Ed. by Jonas Peters and David Sontag. Vol. 124. Proceedings of Machine Learning Research. PMLR, 2020, pp. 101–110.

[HSV14] Martin Hairer, Andrew J. Steward, and Sebastian J. Vollmer. "Spectral Gaps for a Metropolis-Hastings Algorithm in Infinite Dimensions". In: *The Annals of Applied Probability* (2014). DOI: 10.1214/13-AAP982.

[Kal21] Olav Kallenberg. *Foundations of Modern Probability 3rd Edition*. Springer Cham, 2021. DOI: 10.1007/978-3-030-61871-1.

[Lie+23] H. C. Lie et al. "Dimension-independent Markov chain Monte Carlo on the sphere". In: *Scandinavian Journal of Statistics* 50.4 (2023). DOI: https://doi.org/10.1111/sjos.12653. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12653.

[MT93] S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. Springer-Verlag, London, 1993. URL: probability.ca/MT.

[Nea99] Radford M Neal. "Regression and Classification Using Gaussian Process Priors". In: *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting June 6-10, 1998*. Oxford University Press, 1999. ISBN: 9780198504856. DOI: 10.1093/oso/9780198504856.003.0021.

[RC04] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Secaucus, NJ, USA: Springer, New York, Inc., 2004.

[RR04] Gareth O. Roberts and Jeffrey S. Rosenthal. "General state space Markov chains and MCMC algorithms". In: *Probability Surveys* (2004). ISSN: 1549-5787. DOI: 10.1214/154957804100000024. URL: http://dx.doi.org/10.1214/154957804100000024.

[RS22] Daniel Rudolf and Björn Sprungk. *Robust random walk-like Metropolis-Hastings algorithms for concentrating posteriors*. 2022. arXiv: 2202.12127 [stat.CO].