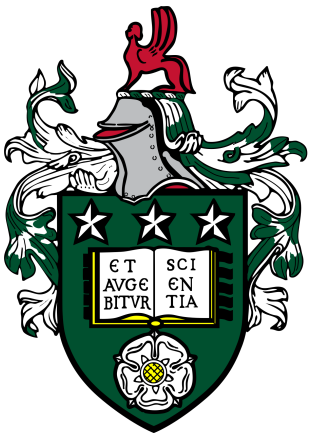
# Theory, Analysis and Implementation of Wavelet Monte Carlo




**Lukas Čironis**

Submitted in accordance with the requirements for the degree of

*Doctor of Philosophy*

The University of Leeds

Department of Statistics

May 2019

# Intellectual Property and Publication Statements

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Lukas Čironis to be identified as Author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

# Acknowledgements

I would like to thank my supervisors Walter R. Gilks and Stuart Barber for their continuous support, constant feedback and patience throughout my research. The knowledge that I have gained during our meetings is immeasurable and I am extremely grateful to both of you for providing me with the opportunity of working with both of you and sharing mathematics related ideas. Meetings and discussions that we had will be greatly missed.

I also wish to express my gratitude to the Engineering and Physical Sciences Research Council (EPSRC) for providing me with the financial support without which this PhD would not be possible.

Most importantly I would like to thank my Parents - Jolanta and Raimondas, for their love and care. I could not have achieved anything so significant in my life without Your help.

# Abstract

Theory of *Wavelet Monte Carlo* (WMC) - a novel sampling algorithm is presented and analysed. It is shown how *Wavelet theory* and *Survival analysis* can be combined together, producing a method that is able to generate independent samples from a non-standard multimodal distribution when a direct sampling approach is not viable. It is demonstrated that due to the way the algorithm is constructed it could be easily parallelised, to boost the execution time. Several issues regarding the implementation of WMC are presented and discussed. In particular, the choice of the wavelet family, curse of dimensionality and computation of wavelet coefficients is investigated in detail revealing critical problems with certain wavelet families. Two possible modifications to the original WMC are outlined with their strengths and weaknesses highlighted. Finally, an important connection between *Besov spaces* and WMC theory is established, revealing intriguing implications of the implicit assumptions made in WMC theory.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 History of sampling methods - from Metropolis-Hastings to Wavelet Monte Carlo

To be able to efficiently calculate the probabilities of specific events and moments given a probability distribution of interest is essential for the statistical community. During the rise of Bayesian statistics in the 1970s, it was especially important to tackle this task, as it is known that the key component in Bayesian inference is a posterior distribution that encapsulates all the required information from a probability model being analysed. Given the usual complexity of high-dimensional integrals involved in the computation of a normalisation constant and a non-standard nature of an associated probability density function (pdf), direct inferences about moments and event probabilities were restricted. However, with improvements in computational power and motivation to analyse complicated posterior distributions in Bayesian analysis, sampling algorithms were introduced to produce realisations from desired probability distributions, either exactly or approximately, using random variate generating procedures.

In the 1990s the boost in computing power opened the door for the Metropolis-Hastings (M-H) algorithm (Metropolis et al. 1953, Hastings 1970) to be used efficiently in practice. M-H utilises a Markov Chain, that in theory, should converge to a required target distribution and with each single jump in the Markov Chain produce a sample from a target probability distribution. The outstanding feature of the M-H algorithm is the requirement that one should only be able to evaluate function $f(\theta)$, which is proportional to the true density of a target distribution. This condition allows a user to bypass the normalisation constant of the target distribution and using an unnormalised version $f(\theta)$ still produce good quality samples from the target. This feature is especially useful in a high-dimensional setting. In Roberts et al. (1997), the product form structure is discussed for the target distribution and a more general setting for target is analysed in Beskos et al. (2009), ensuring the efficiency of M-H in multidimensional problems. A few decades later, after the publication of M-H algorithm, a special version of it was introduced which allows one to produce approximate samples from a target distribution by sampling from full conditional distributions rather than the joint target itself. By essentially setting an acceptance probability equal to 1 and exploring the conditional structure of the posterior distribution, the M-H algorithm could be transformed into what is now known as the Gibbs sampler (Geman & Geman 1984). These two algorithms serve as roots for a majority of the algorithms stemming from the Markov Chain Monte Carlo (MCMC) family of methods (Smith & Roberts 1993).

Aside from MCMC methods, there are several other forms of 'black box' algorithms that utilise random number generating nature and produce samples from a target distribution of interest. Key examples of such algorithms include rejection sampling (RS) (Devroye 1986), adaptive rejection sampling (ARS) (Gilks & Wild 1992) for log-concave densities and the ratio-of-uniforms method (Kinderman & Monahan 1977). Time has shown that many of the sampling algorithms could be used jointly in the same problem to significantly improve results. In particular, if in the Gibbs sampler

setting, certain conditionals are non-conjugate or simply non-standard, an ARS step could be utilised to produce samples from a conditional distribution efficiently, similarly, M-H could be used to deal with the non-log-concave situation in ARS implementations as shown in Gilks, Best & Tan (1995).

It is quite clear that all sampling algorithm families mentioned above have their strengths and weaknesses and tend to work best under certain specific conditions. In particular, MCMC methods tend to be popular if the dimensionality of a problem is quite high; however, they suffer from the inherent Markov Chain nature, leading to dependence of samples, the difficulty of tuning the proposal density (Gilks, Richardson & Spiegelhalter 1995) and inefficient exploration of any multimodal structure in the posterior (Neal 1993, Celeux et al. 2000, Sminchisescu & Welling 2011). In the RS set-up poor choice of an envelope function leads to a high number of samples being rejected in a process producing an inefficient algorithm, ARS deals with this situation much better by adapting the envelope with each realisation, nevertheless the curse of a dimensionality is relevant as the number of low probability regions increase rapidly with the dimension of a problem, leading to a significant decrease in the acceptance probability.

The most recent sampling algorithms that show big potential of being able to deal with high dimensionality are modifications of Hamiltonian Monte Carlo (HMC) (Mark & Ben 2011), an algorithm that utilises the geometry of the sample space and augments the posterior by introducing an additional momentum parameter. In particular, Wormhole Hamiltonian Monte Carlo (WHMC) (Lan et al. 2014) and Generalized Darting Monte Carlo (Sminchisescu & Welling 2011) demonstrate that by introducing additional jumping rules multimodality could also be dealt with under reasonable computational complexity costs. Even though the results are quite promising regarding recent developments in sampling algorithms, the majority of them are still built on an underlying nature of a Markov Chain, which leads to a

dependency structure between samples produced, the necessary tuning of a proposal density and an unpleasant difficulty in parallelising the algorithm for utilisation of several CPU or GPU cores available.

Wavelet Monte Carlo (WMC), discussed in this thesis, will not be treated as a 'panacea' of sampling algorithms, but as a novel approach towards sampling methods that produces independent samples from a known target and entirely circumvents the problem of multimodality by allowing wavelets $\psi_{ji}$ to represent a local information about a sample space at resolution $j$ and location $i$. The hope is that ideas introduced in WMC will be carried over to other families of sampling algorithms to pool the best features of both families and produce a much better alternative. History has shown that the best methods have been produced by exactly following this approach.

WMC is a Monte Carlo type algorithm where, by a repeated procedure, a sample from a target is generated. The most distinct feature of WMC is the utilisation of wavelet theory (Mallat 2008) in a completely new and non-standard manner, which places the algorithm into an entirely new family of sampling methods. Wavelets were mainly developed for image, sound or generally any signal processing techniques (Grossmann & Morlet 1982, Mallat 1989*b*). Significant highlights of the development of wavelet theory include the construction of the Multi-Resolution Analysis (MRA) by Mallat and Meyer (Meyer 1986-1987*a*, Mallat 1989*a*). MRA can be considered as a framework in which functions $f \in L^2(\mathbb{R}^d)$ can be considered as a limit of successive approximations, $f = \lim_{j\to-\infty} P_j f$, where different $P_j f \in \mathbb{Z}$ correspond to smoothed versions of $f$, where the smoothing radius is of order $2^j$. Wavelet coefficients $f^{\psi}_{j,i} = \langle \psi_{j,i}, f \rangle$, where $\psi_{j,i}$ is a wavelet of resolution $j$ and location $i$, correspond to the difference between the two successive approximations $P_{j-1}f$ and $P_j f$. More details regarding MRA will be given in Chapter 2.

Another important point in the history of wavelets was the construction of

Daubechies wavelets (Daubechies 1988), which is an orthogonal wavelet family defining a discrete wavelet transform and characterized by a maximal number of vanishing moments for some given compact support. Due to their compact support structure and characterisation involving the number of vanishing moments, this family will be used quite extensively in our WMC theory and implementation.

## 1.2 Outline of the thesis

After introducing reader to the theory of wavelets (§2), WMC will be outlined and key theorems (3.3.2, 3.2.1) presented. In addition to WMC implementation (§4) and other practical aspects, a lot of attention will be given to theoretical and practical issues (§5) that arise in WMC. A unique connection to Besov spaces (§7) will be established and it will be shown how results from Besov space theory could be directly transferred to WMC theory to explain some important phenomena of WMC (7.3). Finally, possible modifications of WMC will be introduced in §8, suggesting two alternative versions of the original WMC; a theoretical approach to analyse a distribution of jumps will be presented in §6, providing necessary conditions for the validity of WMC theory.

# Chapter 2

# Wavelet theory

This chapter will focus on giving a reader an easy introduction to wavelet theory and in particular the motivation to transition from Fourier to wavelet based methods. Details, derivations and proofs of various statements and theorems will be skipped as these could be referred to in Mallat (2008). The goal of this chapter is to familiarise the reader with the orthonormal wavelet expansion of a function $f \in L^2(\mathbb{R})$.

## 2.1 Short history

The roots of a signal decomposition into separate components go back to 1807 when Fourier presented a memoir to the Institut de France, claiming that any periodic function can be represented as a series of harmonically related sinusoids. Given the outstanding practical implications of this discovery, throughout the next 160 years this theory was improved and generalised. In particular, the Fourier transform (FT) is not able to cope with signals that frequency depends on time, i.e. FT is not able to deal with the time-frequency localisation. To solve this problem, the Windowed-Fourier transform (WFT) (Gabor 1946) was invented by Gabor in 1946. The trick was to use a Gaussian distribution function as a window function that

would localise FT, and, by shifting the window, one would extract the information about the signal at separate time steps. FT was cemented as one of the most useful and widely used algorithms in 1965 when Cooley and Tukey created the Fast Fourier Algorithm (FFA) (Cooley & Tukey 1965).

However, even with such great success, these Fourier analysis (FA) based algorithms faced one big issue - they were using the same window function for an entire signal. In the late 1970s, Morlet was faced with the problem of analysing signals that had very high-frequency components with short time spans, and low-frequency components with long time spans. WFT was able to analyse either high-frequency components using narrow windows, or low-frequency components using wide windows, but not both. This led to a discovery of windows that are localised both in time and frequency domains. Morlet used the same Gaussian function but by dilating and compressing it he was able to precisely analyse different frequency levels. These basis functions were named as 'wavelets of constant shape'. Noticing the importance of this wavelet transform Alex Grossman came up with the exact inversion formula for it. Given the similar interest, Morlet and Grossman started working jointly and contributed to the continuous wavelet transform (CWT) and its applicability.

French mathematician Yves Meyer quickly noticed an underlying connection between the constructed Morlet-Grossman wavelet transform and Calderon formula in Harmonic analysis. Using the knowledge of Calderon-Zygmund operators and Littlewood-Paley theory, Meyer produced a mathematical foundation for wavelet theory (Meyer 1986-1987*b*). Although Meyer was one of the first who started laying the basis for wavelet theory, the first orthonormal wavelet basis was constructed by German mathematician Alfred Haar back in 1909 (Haar 1910).

Following Meyer's formalisation of wavelet theory, the next significant contribution came from Daubechies (Daubechies et al. 1986) with the development of wavelet frames for the discretisation of time and scale parameters in wavelet transform.

Daubechies together with Mallat initiated the transition from analysing continuous signals to discrete ones. In 1986 Mallat developed the idea of Multi-Resolution Analysis (MRA) which later became his PhD thesis in 1988. Important details of MRA will be discussed in later section 2.3. Furthermore, in 1988 Daubechies created her celebrated and still most widely used compactly supported, orthonormal wavelet basis (Daubechies 1988), which also allowed for the control of the wavelet smoothness. The latter two discoveries by Mallat and Daubechies could be treated as a marker for the modern theory of wavelets.

### 2.1.1 Applications

To put wavelet theory (WT) into a context of applications, several fields will be mentioned here where wavelets have shown great potential and performed really well.

**Data compression** (Rao & Bopardikar 1998): Due to the resolution level nature and sparsity of wavelet coefficients many signals could be easily compressed using the discrete wavelet transform (DWT). Given that energy of a signal is mainly concentrated in few wavelets, tiny coefficients could be discarded without introducing large errors into the approximation.

**Denoising**: WT could be also applied for denoising problems. This was explored by Donoho & Johnstone (1994) and finalised by Donoho (1995) which led to the construction of wavelet shrinkage denoising (WSD). Similarly, as in data compression, noise is usually detected at finer scales, therefore coefficients associated with those levels could be set to zero to remove the noise from a signal.

**Genomic sequences:** Saini & Dewan (2016) showed that based on the calculation of the energy of wavelet decomposition coefficients of complete genomic sequences, the similarity between different sequences of *Mycobacterium tuberculosis* could be

determined without the use of conventional methods such as the Basic Local Alignment Search Tool (BLAST).

**Numerical Solution of partial differential equations**: Cohen et al. (2001) showed strong convergence results for wavelet-based algorithms for solving PDEs. This particular discovery led to new methods in finite element analysis.

**Fractals** (Rao & Bopardikar 1998): Some types of wavelets, such as Daubechies wavelets have a self-similar structure, and when combined with Multi-Resolution Analysis, they provide a very natural way of analysing fractals. Farge (Farge 1992), Wornell and Oppenheim (Wornell & Oppenheim 1992) have successfully applied wavelets to fractal analysis.

These are just a few, but there are many more other fields to which WT has and could be applied. In particular, analysis of financial data (Gallegati 2012), analysis of turbulent flows of low viscosity fluids (Camussi & Guj 1997), neural networks (Zhang & Benveniste 1992), analysis of distant universes (Bijaoui et al. 1996), biomedical engineering (Akay 1995), etc.

As it will be apparent throughout this thesis, wavelets will potentially add a further field of applications - random variate generation.

## 2.2 Fourier bases versus wavelet bases

Here, we briefly describe and discuss key differences between Fourier and wavelet type transforms and bases associated with them.

### 2.2.1 Fourier transform

Using Fourier analysis (FA), any function $f(t)$ with finite energy:

$$\|f\|_{L^2} = \sqrt{\int_{\mathbb{R}} |f(t)|^2 \mathrm{d}t} < \infty$$

can be represented as a sum of trigonometric waves $e^{i\omega t}$:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega)\, \mathrm{d}\omega,$$

where $i = \sqrt{-1}$. The amplitude $\hat{f}(\omega)$ is equal to the inner product between the function being analysed $f(t)$ and the trigonometric wave $e^{i\omega t}$, where $\omega$ is the frequency of the wave. This inner product is known as the Fourier transform (FT):

$$\hat{f}(t) = \int_{-\infty}^{+\infty} f(t) e^{i\omega t}\, \mathrm{d}t.$$

The decay properties of the amplitude $|\hat{f}(\omega)|$ are characterised by the regularity of $f(t)$. The smoother $f(t)$, the faster the decay as $\omega$ increases.

Furthermore, $\{e^{i2\pi mt}\}_{m\in\mathbb{Z}}$ forms a Fourier orthonormal basis (FOB) of $L^2[0,1]$. Therefore, if function $f(t)$ lives on this interval, it can be decomposed using FOB. High differentiability of $f(t)$ also implies the rapid decay of Fourier coefficients with the increase of frequency $2\pi m$; therefore, FT defines a sparse representation of uniformly regular functions.

As long as one is analysing uniformly regular signals, FA provides a sufficient set of tools to solve most of the problems. However, FA is not able to cope with transient features — events in a signal where the frequency changes rapidly over time. In short, time-frequency localisation in FT is poor, and other types of analyses need to be used to solve these problems.

As one can see, $\mathrm{supp}\{e^{i\omega t}\} = \mathbb{R}$, so $\hat{f}(\omega)$ combines all the frequency information extracted from $f(t)$ at all times $t \in \mathbb{R}$. Therefore, $\hat{f}(\omega)$ does not represent any local information about the signal $f(t)$.

### 2.2.2 Windowed Fourier transform

An attempt to overcome certain issues of FT can be made by introducing a window function that would localise waveforms in both time and frequency. Let $g(t)$ be a time window centred at $t = 0$ with unit norm $\|g\| = 1$, then one can define a windowed Fourier dictionary of waveforms:

$$\mathcal{D} = \left\{ g_{u,\xi}(t) = g(t-u)e^{i\xi t} \right\}_{(u,\xi)\in\mathbb{R}^2}.$$

The windowed Fourier transform (WFT) is performed by projecting $f(t)$ onto each $g_{u,\xi}$:

$$Sf(u,\xi) = \langle f, g_{u,\xi} \rangle = \int_{-\infty}^{+\infty} f(t)g(t-u)e^{-i\xi t}\,\mathrm{d}t.$$

Paying the cost of disrupting the basis structure one creates an atom $g_{u,\xi}$ that has a good localisation in time and frequency domains. However, it can be shown that the time-frequency variation of $g_{u,\xi}$ is independent of $u$ and $\xi$ and the window is always of fixed size and frequency, i.e. the WFT decomposes signals over waveforms that have the same time and frequency localisation. Therefore, WFT is only useful at analysing signals that do not have structures having different time-frequency resolutions, i.e. some being very localised in time and others very localised in frequency. Unfortunately, the majority of signals in fact contain structures that vary in the time-frequency domain.

Wavelets address this problem by introducing atoms that change in both time and frequency resolution.

### 2.2.3 Wavelet transform

To address the fact that signals incorporate structures of very different sizes, it is essential to use time-frequency atoms of different time support. The wavelet transform (WT) decomposes a signal over dilated and translated wavelets. A wavelet

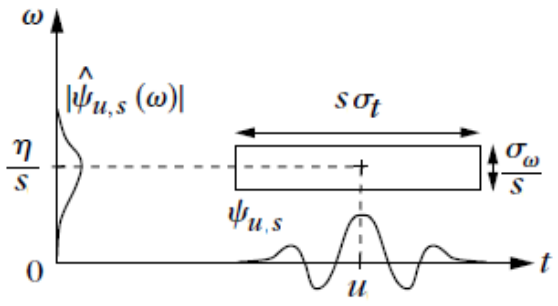



Figure 2.1: *Heisenberg box.*

is a function $\psi \in L^2(\mathbb{R})$ that satisfies certain specific conditions. It is normalised $\|\psi\| = 1$, centred around $t = 0$ and has zero average:

$$\int_{-\infty}^{+\infty} \psi(t)\,\mathrm{d}t = 0.$$

By translating $\psi$ by $u$ and scaling by $s$, one obtains a dictionary of time-frequency atoms:

$$\mathcal{D} = \left\{\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\Big(\frac{t-u}{s}\Big)\right\}_{u\in\mathbb{R},s\in\mathbb{R}^+}. \tag{2.2.1}$$

These translated and scaled versions $\psi_{u,s}$ maintain the same norm as $\psi$, $\|\psi_{u,s}\| = 1$. So, we have that the WT of $f \in L^2(\mathbb{R})$ at time $u$ and scale $s$ is

$$Wf(u,s) = \langle f, \psi_{u,s}\rangle = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\psi\Big(\frac{t-u}{s}\Big)\,\mathrm{d}t.$$

It is clear from the construction that the time localisation of $\psi_{u,s}$ depends on the scale $s$. An increase in $s$ is associated with coarser wavelets, and, for small values of $s$, we get highly-concentrated wavelets in time. It can be shown using FT, that the energy spread of a wavelet atom $\psi_{u,s}$ corresponds to a Heisenberg box (Figure 2.1) centred at $(u, \eta/s)$, of size $s\sigma_t$ along time and $\sigma_\omega/s$ along frequency. Therefore, by controlling $s$, one is able to get clearer localisation in frequency or time. Wavelets with coarser scales are responsible in detecting average behaviour of a signal and with finer scales — sharper transitions within it.

However, given the redundancy of dictionary $\mathcal{D}$ in (2.2.1), we look for a sub-dictionary of $\mathcal{D}$ that forms a basis for $L^2(\mathbb{R})$ and allows for sparse representations

of function of interest.

## 2.3 Multi-Resolution Analysis

Given that a function of interest $f \in L^2(\mathbb{R})$, a *Multi-Resolution Analysis* (MRA) (Mallat 2008) can be performed and function $f$ can be decomposed into a series of orthonormal, compactly supported wavelets.

A MRA is an increasing sequence of closed subspaces $\{\mathbf{V}_j\}_{j\in\mathbb{Z}}$ that approximate $L^2(\mathbb{R})$. The construction of a MRA starts with a smart choice of a *scaling function* $\phi$. It is chosen to satisfy some regularity conditions, these will not be covered here in detail, but most importantly it is chosen such that a family $\{\phi(x-i)\}_{i\in\mathbb{Z}}$ forms an orthonormal basis for the reference space $\mathbf{V}_0$. The following relations describe the analysis.

0. $\{0\}... \subset \mathbf{V}_{-1} \subset \mathbf{V}_0 \subset \mathbf{V}_1 \subset ... \subset L^2(\mathbb{R})$,
1. $\forall (j,i) \in \mathbb{Z}^2,\ f(t) \in \mathbf{V}_j \iff f(t-2^j i) \in \mathbf{V}_j$,
2. $\forall j \in \mathbb{Z},\ f(t) \in \mathbf{V}_{j+1} \iff f(\frac{t}{2}) \in \mathbf{V}_j$,
3. $\lim\limits_{j\to-\infty} \mathbf{V}_j = \bigcap\limits_{j=-\infty}^{+\infty} \mathbf{V}_j = \{0\}$,
4. $\lim\limits_{j\to+\infty} \mathbf{V}_j = \text{closure}\Big\{ \bigcup\limits_{j=-\infty}^{+\infty} \mathbf{V}_j \Big\} = L^2(\mathbb{R})$,
5. $\exists \phi \in \mathbf{V}_0$ a *scaling function*, such that $\{\phi(t-i)\}_{i\in\mathbb{Z}}$ forms a Riesz basis of $\mathbf{V}_0$.

Assuming the subspaces $\mathbf{V}_j$ in (0), relation (1) describes that, if $f(t) \in \mathbf{V}_j$, then translated versions of the original function still belong to the space $\mathbf{V}_j$. Relation (2) shows that function can climb the resolution ladder by being scaled, i.e. if

$f(t) \in \mathbf{V}_{j+1}$ then by scaling it down by dyadic factor function jumps down to coarser approximation space $\mathbf{V}_j$. Relations (3) and (4) show that limiting coarsest approximation is basically a constant function space and finest approximation space coincides with the space of interest $L^2(\mathbb{R})$. Finally, (5) is a technical part that requires the existence of the basis on a reference space $\mathbf{V}_0$, from which basis for $L^2(\mathbb{R})$ can be constructed.

**Definition 2.3.1.** A countable set $\{f_n\}$ of a Hilbert space is a Riesz basis if every element $f$ of the space can be written uniquely as $f = \sum_n c_n f_n$, and positive constants $A$ and $B$ exist such that

$$A\|f\|^2 \leq \sum_n |c_n|^2 \leq B\|f\|^2. \tag{2.3.2}$$

Since $\phi \in \mathbf{V}_0 \subset \mathbf{V}_1$, a sequence $(h_i) \in l^2(\mathbb{Z})$ exists such that the scaling function satisfies the *refinement equation*

$$\phi(x) = \sum_i h_i \phi(2x - i). \tag{2.3.3}$$

The collection of coefficients $h = \{h_i\}_{i\in\mathbb{Z}}$ is called a *conjugate mirror filter* and it is responsible for characterising the scaling function. It is quite clear that the collection of functions $\{\phi_{j,i}\}_{i\in\mathbb{Z}}$, with $\phi_{j,i}(x) = 2^{j/2}\phi(2^j - i)$, is a Riesz basis of $\mathbf{V}_j$. By integrating both sides of (2.3.3) and normalising by the integral of $\phi$ we get

$$\sum_i h_i = 2. \tag{2.3.4}$$

Now, let $\mathbf{W}_j$ denote the orthogonal complement of the space $\mathbf{V}_j$ in the space $\mathbf{V}_{j+1}$, so that $\mathbf{V}_{j+1} = \mathbf{V}_j \oplus \mathbf{W}_j$. Relations (3) and (4) above for $\mathbf{V}_j$ spaces imply that

i. $\oplus_{j\in\mathbb{Z}} \mathbf{W}_j = L^2(\mathbb{R})$,

and similarly it can be shown that

ii. $\forall (j,i) \in \mathbb{Z}^2,\ f(t) \in \mathbf{W}_j \iff f(t - 2^j i) \in \mathbf{W}_j,$

iii. $\forall j \in \mathbb{Z},\ f(t) \in \mathbf{W}_{j+1} \iff f(\frac{t}{2}) \in \mathbf{W}_j.$

Now any $f \in L^2(\mathbb{R})$ has a sequence of orthogonal decompositions

$$f = v_k + \sum_{j=k+1}^{+\infty} w_j = \sum_{k \in \mathbb{Z}} w_k \tag{2.3.5}$$

where $v_k \in \mathbf{V}_k$ and $w_k \in \mathbf{W}_k$.

**Theorem 2.3.1** (Mallat, Meyer)**.** *Let $\phi$ be a scaling function and $h$ the corresponding conjugate mirror filter. Let $\psi$ be the function having a Fourier transform*

$$\hat{\psi}(\omega) = \frac{1}{\sqrt{2}} \hat{g}\Big(\frac{\omega}{2}\Big) \hat{\phi}\Big(\frac{\omega}{2}\Big), \tag{2.3.6}$$

*with*

$$\hat{g}(\omega) = e^{-i\omega} \hat{h}^*(\omega + \pi), \tag{2.3.7}$$

*where $\hat{h}^*$ denotes a conjugate of discrete Fourier transform of $h$ coefficients. Let us denote*

$$\psi_{j,i}(t) = 2^{j/2} \psi(2^j - i). \tag{2.3.8}$$

*For any scale $2^j$, $\{\psi_{j,i}\}_{i \in \mathbb{Z}}$ is an orthonormal basis of $\boldsymbol{W}_j$. For all scales, $\{\psi_{j,i}\}_{(j,i) \in \mathbb{Z}^2}$ is an orthonormal basis for $L^2(\mathbb{R})$.*

Functions $\psi_{j,i}$ will be called wavelets at resolution $j$ and location $i$, and with $\psi$ we will denote a standard *mother wavelet*. The scaling function $\phi$ is sometimes also called a *father wavelet*.

The theorem (2.3.1) by Mallat and Meyer gives specific conditions for the construction of the orthonormal basis $\{\psi_{j,i}\}$. It is clear that there is no unique basis $\{\psi_{j,i}\}$; the next section will give a brief introduction to possible families of wavelets $\{\psi_{j,i}\}$.

From (2.3.5), we can see that there are mainly two ways of representing functions using wavelets - with the scaling function or without it:

$$f(x) = \sum_{i\in\mathbb{Z}} \langle f, \phi_{j_0,i}\rangle \phi_{j_0,i}(x) + \sum_{j=j_0}^{+\infty} \sum_{i\in\mathbb{Z}} \langle f, \psi_{j,i}\rangle \psi_{j,i}(x) \tag{2.3.9}$$

gives the representation of a function using father wavelets at reference resolution $j_0$, while

$$f(x) = \sum_{j,i\in\mathbb{Z}} \langle f, \psi_{j,i}\rangle \psi_{j,i}(x) \tag{2.3.10}$$

gives representation of a function using mother wavelets $\psi_{j,i}$ only. Throughout this dissertation we will mainly focus on the later form.

We will also denote mother wavelet coefficients (later we will refer to these as just wavelet coefficients) by

$$f^{\psi}_{j,i} := \langle f, \psi_{j,i}\rangle = \int_{-\infty}^{+\infty} f(x)\psi_{j,i}(x)\,\mathrm{d}x, \tag{2.3.11}$$

and similarly we denote father wavelet coefficients by

$$f^{\phi}_{j,i} := \langle f, \phi_{j,i}\rangle = \int_{-\infty}^{+\infty} f(x)\phi_{j,i}(x)\,\mathrm{d}x. \tag{2.3.12}$$

## 2.4 Wavelet families

Here a few wavelet families will be presented. All wavelets form an orthonormal basis for $L^2(\mathbb{R})$ and satisfy

$$\int_{-\infty}^{\infty} \psi_{j,i}(t)\psi_{\ell,k}(t)\,\mathrm{d}t = \begin{cases} 1, & i = k, j = \ell \\ 0, & \text{otherwise} \end{cases}. \tag{2.4.13}$$

### 2.4.1 Haar wavelets

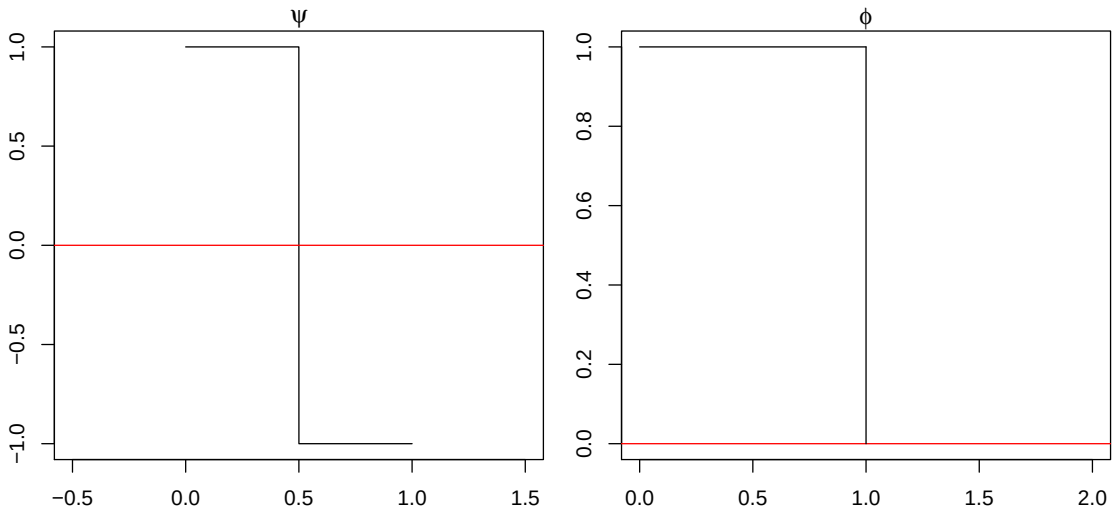



Figure 2.2: *Mother and father wavelets of the Haar basis.*

**Definition 2.4.1** (Haar wavelets)**.** For every pair $j$, $i$ of integers in $\mathbb{Z}$, the Haar mother wavelet $\psi_{j,i}(t)$ is defined on the real line $\mathbb{R}$ by the function

$$\psi_{j,i}(t) = 2^{j/2}\psi(2^j t - i) \quad t \in \mathbb{R} \tag{2.4.14}$$

with $\text{supp}\{\psi_{j,i}(t)\} = I_{j,i} = [i2^{-j}, (i+1)2^{-j})$, where

$$\psi(t) = \begin{cases} 1 & 0 \leq t < 0.5, \\ -1 & 0.5 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, we define the Haar father wavelet $\phi(t)$ to be

$$\phi(t) = \begin{cases} 1 & 0 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

This is the most simple form of a wavelet (Figure 2.2), first constructed by Haar in 1909 (Haar 1910). Despite the simplistic step-like nature, it still does form a basis

for the $L^2(\mathbb{R})$ space and allows for the approximation of functions. However, one big issue with Haar wavelets is that they are discontinuous and hence not differentiable everywhere, which could be a desired property in the analysis of more complicated and specific signals. From a statistical point of view Haar wavelets might seem very attractive as they are scaled combinations of uniform distributions along finite intervals. It will be demonstrated later in the future chapters that there is other important reason why Haar wavelets cannot be used in a general WMC setting.

### 2.4.2 Daubechies wavelets

Probably the most significant wavelet family ever constructed was created by I. Daubechies in 1988 (Daubechies 1988). These are orthogonal wavelets referred to simply as Daubechies wavelets (Figure 2.3), characterized by a maximal number of vanishing moments for some given compact support.

**Definition 2.4.2** (Vanishing moments)**.** A wavelet $\psi(x)$ has $K$ vanishing moments if

$$\int_{\mathbb{R}} x^k \psi(x) \mathrm{d}x = 0 \quad \text{for} \quad 0 \leq k < K.$$

The vanishing moment is a criterion about how a function decays toward infinity. A theorem in Mallat (2008) on page 288 shows that if $\psi(x)$ has $K$ vanishing moments, then

$$|\psi(x)| = \mathcal{O}\big((1+x^2)^{-K/2-1}\big) \tag{2.4.15}$$

Hence, with the increase in a number of vanishing moments functions could be approximated more sparsely (using less wavelet coefficients).

Daubechies wavelets are not defined in terms of scaling and wavelet functions; in fact, they cannot be written down in closed form. Daubechies wavelets have a support of minimum size for any given number $K$ of vanishing moments, and the size of the support is $2K-1$.

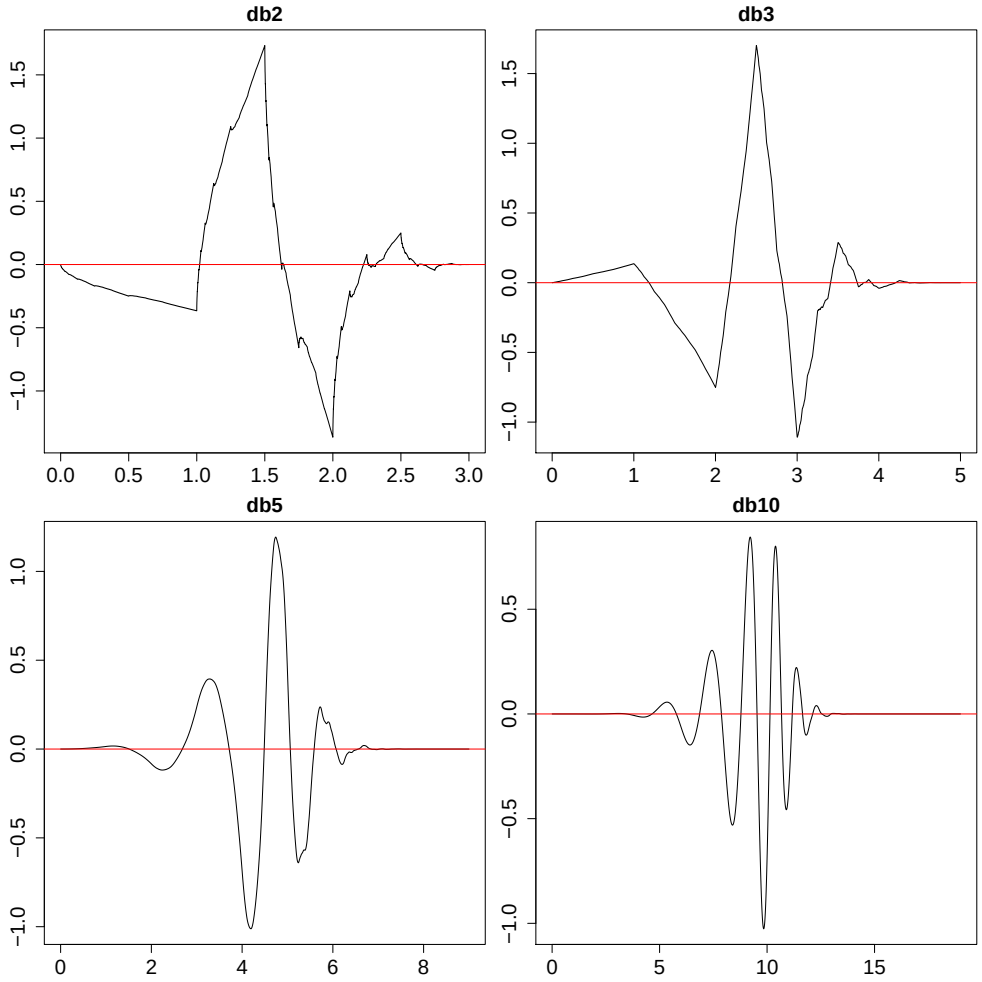



Figure 2.3: *Examples of Daubechies mother wavelets with a different number of vanishing moments ($K = 2, 3, 5, 10$). One can observe that wavelets get smoother as the number of vanishing moments increases.*

It also turns out that Daubechies wavelets with $K = 1$ produce the Haar family. So Haar wavelets are orthonormal wavelets of the worst possible smoothness but shortest compact support.

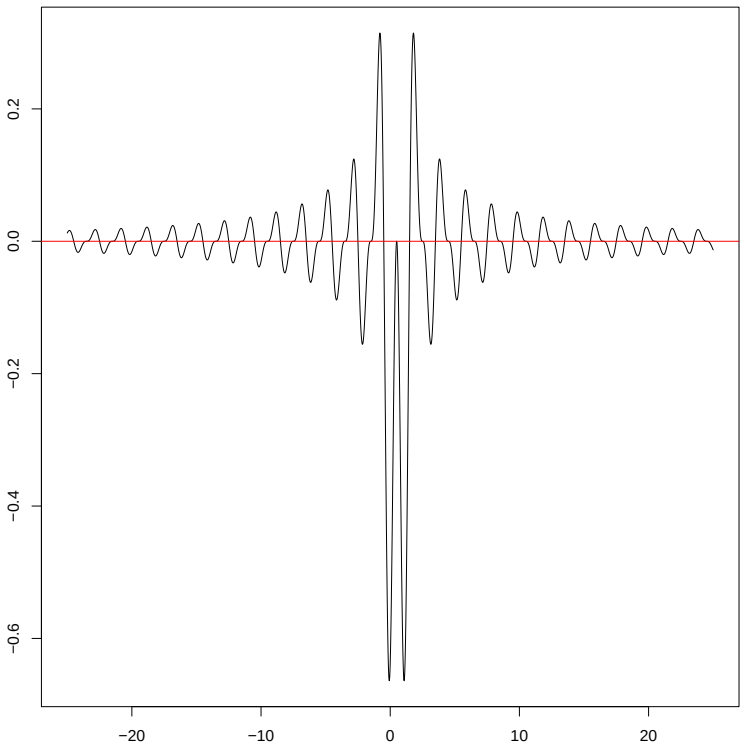



Figure 2.4: *Shannon wavelet as defined in (2.4.16).*

### 2.4.3 Shannon wavelets

The Shannon wavelet (Figure 2.4) is constructed by taking $\hat{\phi}(\omega) = \mathbb{1}_{[-\pi,\pi]}$ and $\hat{h}(\omega) = \sqrt{2}\mathbb{1}_{[-\pi/2,\pi/2]}(\omega)$ for $\omega \in [\pi, \pi]$. Using (2.3.7), one can derive that

$$\hat{\psi}(\omega) = \begin{cases} e^{-i\omega/2} & \text{if } \omega \in [-2\pi, -\pi] \cup [\pi, 2\pi] \,, \\ 0 & \text{otherwise.} \end{cases}$$

and thus,

$$\psi(t) = \frac{\sin 2\pi(t - 1/2)}{2\pi(t - 1/2)} - \frac{\sin \pi(t - 1/2)}{\pi(t - 1/2)}. \tag{2.4.16}$$

The constructed wavelet belongs to the $\mathbb{C}^\infty$ space, but decays very slowly as $t \to \pm\infty$. In addition to that, it has an infinite number of vanishing moments.

### 2.4.4 Coiflets

Coiflets are special types of wavelets which were constructed by Daubechies following a request from Coifman for applications in numerical analysis. In addition to the wavelets having $K$ vanishing moments, coiflets were constructed such that scaling functions $\phi$ also satisfy conditions for a number of vanishing moments

$$\int_{-\infty}^{+\infty} \phi(t)\,\mathrm{d}t = 1 \quad \text{and} \quad \int_{-\infty}^{+\infty} t^k \phi(t)\,\mathrm{d}t = 0 \quad \text{for } 1 \leq k < K. \tag{2.4.17}$$

Apparently, such properties of scaling functions allow for the construction of accurate quadrature formulas. Also, at fine resolutions scaling coefficients can be approximated as samples from a signal itself:

$$2^{-J/2}\langle f, \phi_{J,n}\rangle \approx f(2^J n) + \mathcal{O}(2^{(k+1)J}), \quad k < K. \tag{2.4.18}$$

Here only a few wavelet families have been mentioned, however, there are many others, in particular: Symmlets, Morlet wavelets, Meyer wavelets, Ricker (Mexican hat) wavelets, Beta wavelets, which involves the beta distribution in their construction, and several others. Throughout this thesis, the focus will be mainly given to Daubechies wavelets, due to their good energy localisation features, possible control of smoothness, but, most importantly, their compact support.

# Chapter 3

# Theory of Wavelet Monte Carlo

This chapter will focus on describing the theory of WMC and the first three sections will closely follow material from Gilks (2017), the last Section 3.4 and everything onward is the original work of the author of this thesis. A non-standard notation will be introduced first, together with a provisional-WMC (pWMC) algorithm. The pWMC algorithm will then be used as a prerequisite to construct our main WMC algorithm.

## 3.1 Notation and set-up of a framework

We wish to produce samples from a non-standard probability distribution with density proportional to $g(\cdot)$. The non-standard distribution should be interpreted as one from which direct sampling is not possible. Through a sequence of steps, WMC transforms samples from a starting distribution $f(\cdot)$ to samples from the target $g(\cdot)$. Ideally, $f(\cdot)$ is chosen such that it is as similar to $g(\cdot)$ as possible and the user is able to directly sample from $f(\cdot)$. However, even if $f(\cdot)$ is a substantially different distribution from $g(\cdot)$, as long as direct sampling from $f(\cdot)$ is available, WMC will produce samples from the target. We will be working with wavelet expansions of

the densities $f(\cdot)$ and $g(\cdot)$:

$$f(x) = \sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}} f^{\psi}_{j,i}\psi_{j,i}(x), \qquad g(x) = \sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}} g^{\psi}_{j,i}\psi_{j,i}(x), \tag{3.1.1}$$

where we denote the associated mother wavelet coefficients by

$$f^{\psi}_{j,i} = \int_{-\infty}^{+\infty} f(x)\psi_{j,i}(x)\,\mathrm{d}x \qquad \text{and} \qquad g^{\psi}_{j,i} = \int_{-\infty}^{+\infty} g(x)\psi_{j,i}(x)\,\mathrm{d}x. \tag{3.1.2}$$

Throughout this chapter, we will be working extensively with positive and negative parts of certain values. Therefore, we introduce a notation that allows us to conveniently operate with these parts.

**Definition 3.1.1.** For any scalar $a \in \mathbb{R}$, let

$$a^{+} = \begin{cases} a, & a \geq 0 \\ 0, & a < 0 \end{cases}, \qquad a^{-} = \begin{cases} 0, & a \geq 0 \\ -a, & a < 0. \end{cases}$$

Following the notation described in Definition 3.1.1, we have the equalities:

$$a = a^{+} - a^{-} \qquad \text{and} \quad |a| = a^{+} + a^{-}. \tag{3.1.3}$$

Given that each mother wavelet $\psi_{j,i}$ integrates to zero, we make a trivial observation:

$$A_j = \int_{-\infty}^{+\infty} \psi^{+}_{j,i}(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} \psi^{-}_{j,i}(x)\,\mathrm{d}x. \tag{3.1.4}$$

The value $A_j$ will be known as a normalisation constant of the positive and negative part of the wavelet $\psi^{+}_{j,i}$ and $\psi^{-}_{j,i}$. As location shifts do not affect the value of the integral, the constant $A_j$ depends only on the resolution level $j$ and the choice of the wavelet.

Let us also define $r$ to be the ratio of normalising constants

$$r = \frac{\int g(x)\,\mathrm{d}x}{\int f(x)\,\mathrm{d}x}. \tag{3.1.5}$$

The methodology below assumes that a functional form of both $f(\cdot)$ and $g(\cdot)$ is known up to a normalisation constant and there is a way to accurately estimate the ratio of the normalisation constants $r$.

We also define a difference function between $g(x)$ and $f(x)$:

$$d(x) = g(x) - rf(x). \tag{3.1.6}$$

We have a wavelet expansion of $d(x)$:

$$d(x) = \sum_{j\in\mathbb{Z}} \sum_{i\in\mathbb{Z}} d^{\psi}_{j,i}\psi_{j,i}(x). \tag{3.1.7}$$

From equation (3.1.6), it follows that

$$d^{\psi}_{j,i} = g^{\psi}_{j,i} - rf^{\psi}_{j,i}. \tag{3.1.8}$$

The WMC algorithm uses positive and negative parts of the wavelet $\psi^{+}_{j,i}$ and $\psi^{-}_{j,i}$ to construct probability distributions, from which points will be sampled, to update a sample point $x \sim f(\cdot)$ to produce $y \sim g(\cdot)$. In addition to this, a single wavelet $\psi_{j,i}$ itself will be sampled from the sub-collection of wavelets $\{\psi_{j,i}(x)\}_{(j,i)\in\mathbb{Z}^2}$, which are supported at the sampled point $x$, to determine which wavelet will be used to do the updating process. Wavelet coefficients $d^{\psi}_{j,i}$ together with the value $\psi_{j,i}(x)$ will be used to construct weights to sample $\psi_{j,i}$. The next section will describe in detail a method that incorporates the notation described here to produce a novel sampling algorithm.

## 3.2 Provisional Wavelet Monte Carlo

In this section, we present an algorithm which will in a single step transform a sample $x \sim f(\cdot)$ to produce a sample $y \sim g(\cdot)$. However, two rather strong assumptions need to be satisfied:

**A1.** $r$ is known,

**A2.** $\sum_{j\in\mathbb{Z}} \sum_{i\in\mathbb{Z}} [d^{\psi}_{j,i}\psi_{j,i}(x)]^{-} \leq rf(x) \quad \forall x \in \mathbb{R}.$

The reason for the inequality in **A2** will be apparent soon.

**pWMC** method. Let $x \sim f(x)$.

**Step 1.** Sample a pair $(j, i)$ with probability

$$p_{j,i}(x) = \frac{[d^{\psi}_{j,i}\psi_{j,i}(x)]^{-}}{rf(x)}, \tag{3.2.9}$$

where $(j, i) \in \mathbb{Z}^2$. **A2** ensures that $\sum_{j,i} p_{j,i}(x) \leq 1$. With probability

$$1 - \sum_{j\in\mathbb{Z}} \sum_{i\in\mathbb{Z}} p_{j,i}(x) \tag{3.2.10}$$

no pair $(j, i)$ is selected.

**Step 2.** If pair $(j, i)$ is selected at Step 1, sample

$$y \sim \begin{cases} \psi^{+}_{ji}(y)/A_j & , \text{ if } d^{\psi}_{ji} \geq 0 \\ \psi^{-}_{ji}(y)/A_j & , \text{ if } d^{\psi}_{ji} < 0 \end{cases},$$

otherwise set

$$y = x. \tag{3.2.11}$$

**END**.

**Proposition 3.2.1.** *The pWMC algorithm above is guaranteed to produce* $y \sim g(y)$.

Next, a proof will be given to show that the correct normalised marginal density $g(y)$ is obtained after applying the pWMC algorithm to $x \sim f(x)$. For convenience, we will use $\sum_{j,i}$ notation as a shorthand for $\sum_{j\in\mathbb{Z}} \sum_{i\in\mathbb{Z}}$, and similarly integrals with no limits should be interpreted as integrals over the full support of the integrand.

*Proof of Proposition 3.2.1.* By construction, the goal is to work out the marginal distribution $p(y)$. By $\delta$ we denote a Dirac measure and by $\mathbb{1}(\cdot)$ an indicator function.

$$
\begin{aligned}
p(y) = \int_{x\in\mathbb{R}} \frac{f(x)}{\int f(z)\,\mathrm{d}z}\Big\{ \big(1-\sum_{j,i} p_{j,i}(x)\big)\delta(x-y) \\
+ \sum_{j,i} p_{j,i}(x)\big[\mathbb{1}(d^{\psi}_{j,i}\geq 0)\frac{\psi^{+}_{j,i}(y)}{A_j} + \mathbb{1}(d^{\psi}_{j,i}<0)\frac{\psi^{-}_{j,i}(y)}{A_j}\big]\Big\}\,\mathrm{d}x \\
= \int_{x\in\mathbb{R}} \frac{f(x)}{\int f(z)\,\mathrm{d}z}\Big\{ \big(1-\sum_{j,i} \frac{[d^{\psi}_{j,i}\psi_{j,i}(x)]^{-}}{rf(x)}\big)\delta(x-y) \\
+ \sum_{j,i} \frac{[d^{\psi}_{j,i}\psi_{j,i}(x)]^{-}}{rf(x)}\frac{1}{A_j}\big[\mathbb{1}(d^{\psi}_{j,i}\geq 0)\psi^{+}_{j,i}(y) + \mathbb{1}(d^{\psi}_{j,i}<0)\psi^{-}_{j,i}(y)\big]\Big\}\,\mathrm{d}x
\end{aligned}
$$

substituting $p_{ji}(x)$ as in (3.2.9),

$$
\begin{aligned}
= \frac{1}{\int f(z)\,\mathrm{d}z}\Big\{ f(y) - \frac{1}{r}\sum_{j,i}[d^{\psi}_{j,i}\psi_{j,i}(y)]^{-} \\
+ \sum_{j,i}\frac{1}{A_j r}\int_{x\in\mathbb{R}} \big[d^{\psi^-}_{j,i}\psi^{+}_{j,i}(x) + d^{\psi^+}_{j,i}\psi^{-}_{j,i}(x)\big]\big[\mathbb{1}(d^{\psi}_{j,i}\geq 0)\psi^{+}_{j,i}(y) + \mathbb{1}(d^{\psi}_{j,i}<0)\psi^{-}_{j,i}(y)\big]\,\mathrm{d}x\Big\} \\
= \frac{1}{\int f(z)\,\mathrm{d}z}\Big\{ f(y) - \frac{1}{r}\sum_{j,i}\big[d^{\psi^-}_{j,i}\psi^{+}_{j,i}(y) + d^{\psi^+}_{j,i}\psi^{-}_{j,i}(y)\big] \\
+ \sum_{j,i}\frac{1}{r}\big[d^{\psi^-}_{j,i} + d^{\psi^+}_{j,i}\big]\big[\mathbb{1}(d^{\psi}_{j,i}\geq 0)\psi^{+}_{j,i}(y) + \mathbb{1}(d^{\psi}_{j,i}<0)\psi^{-}_{j,i}(y)\big]\Big\}
\end{aligned}
$$

integrating over $x$ and using (3.1.4), now we will expand the brackets and apply the indicator function property

$$
\begin{aligned}
= \frac{1}{\int f(z)\,\mathrm{d}z}\Big\{ f(y) - \frac{1}{r}\sum_{j,i}\big[d^{\psi^-}_{j,i}\psi^{+}_{j,i}(y) + d^{\psi^+}_{j,i}\psi^{-}_{j,i}(y) - d^{\psi^+}_{j,i}\psi^{+}_{j,i}(y) - d^{\psi^-}_{j,i}\psi^{-}_{j,i}(y)\big]\Big\} \\
= \frac{1}{\int f(z)\,\mathrm{d}z}\Big\{ f(y) - \frac{1}{r}\sum_{j,i}\big[d^{\psi^+}_{j,i} - d^{\psi^-}_{j,i}\big]\big[\psi^{-}_{j,i}(y) - \psi^{+}_{j,i}(y)\big]\Big\} \\
= \frac{1}{\int f(z)\,\mathrm{d}z}\Big\{ f(y) + \frac{1}{r}\sum_{j,i} d^{\psi}_{j,i}\psi_{j,i}(y)\Big\} \qquad (3.2.12) \\
= \frac{1}{\int f(z)\,\mathrm{d}z}\Big\{ f(y) + \frac{1}{r}d(y)\Big\}
\end{aligned}
$$

using (3.1.6) and (3.1.7),

$$= \frac{1}{\int f(z)\,\mathrm{d}z}\frac{1}{r}g(y)$$
$$= \frac{g(y)}{\int g(z)\,\mathrm{d}z}$$

from (3.1.5).

We see that the marginal distribution is equal to the target $g(y)$. □

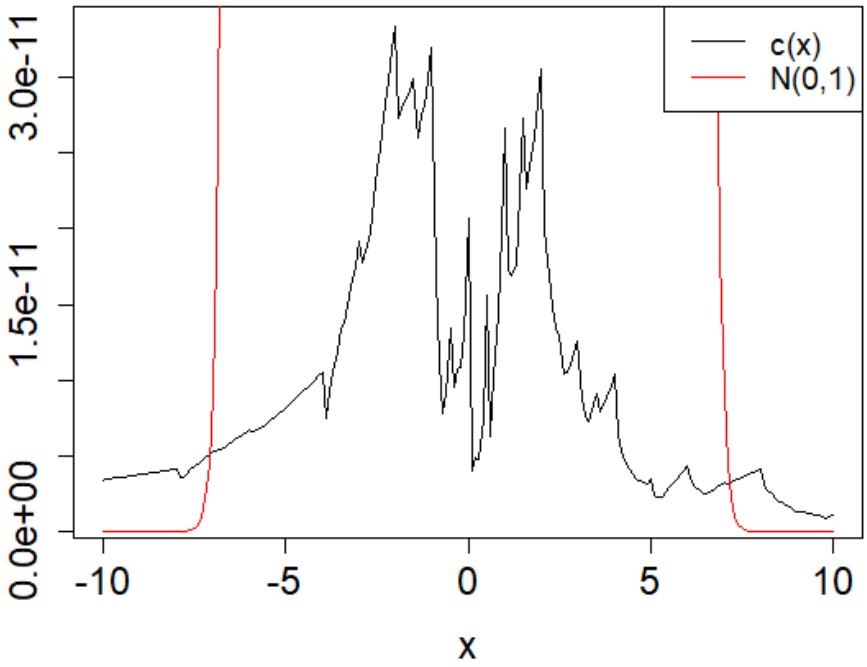



Figure 3.1: *Visual comparison between a starting density of a standard normal distribution* $\mathcal{N}(0,1)$ *and* $c(x) = \sum_{j,i}[d^{\psi}_{j,i}\psi_{j,i}(x)]^{-}$ *using Daubechies wavelets with two vanishing moments. The target distribution was chosen to be* $\mathcal{N}(0, 1 + 10^{-10})$.

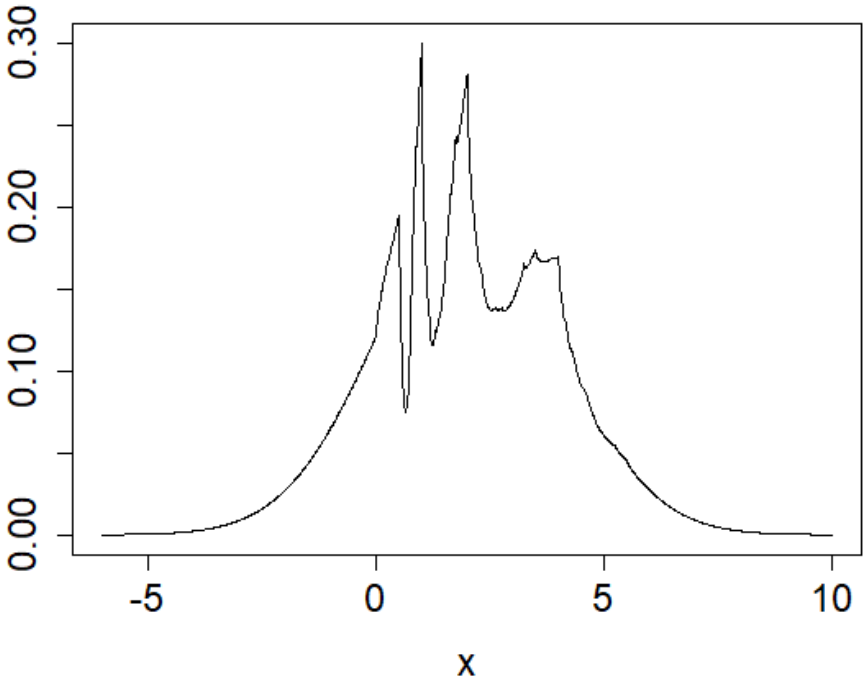



Figure 3.2: *Shape of the target density (3.2.13) with* $\delta = 0.05$ *for which pWMC is applicable.*

Although this algorithm does work theoretically, it is highly dependent on the strong assumption **A2**. If the inequality is not satisfied, a negative no-pair probability could be encountered in Step 1.

As one can see in Figure 3.1, even by choosing target distribution extremely close to the starting one, we have clear regions where $c(x) > f(x)$ and therefore assumption **A2** on page 43 is not satisfied. Now we present an example of a starting distribution and the target for which pWMC would be applicable and assumption **A2** would be satisfied. Let the starting distribution be $\mathcal{N}(2, 4)$ with density $f(x)$ and the target

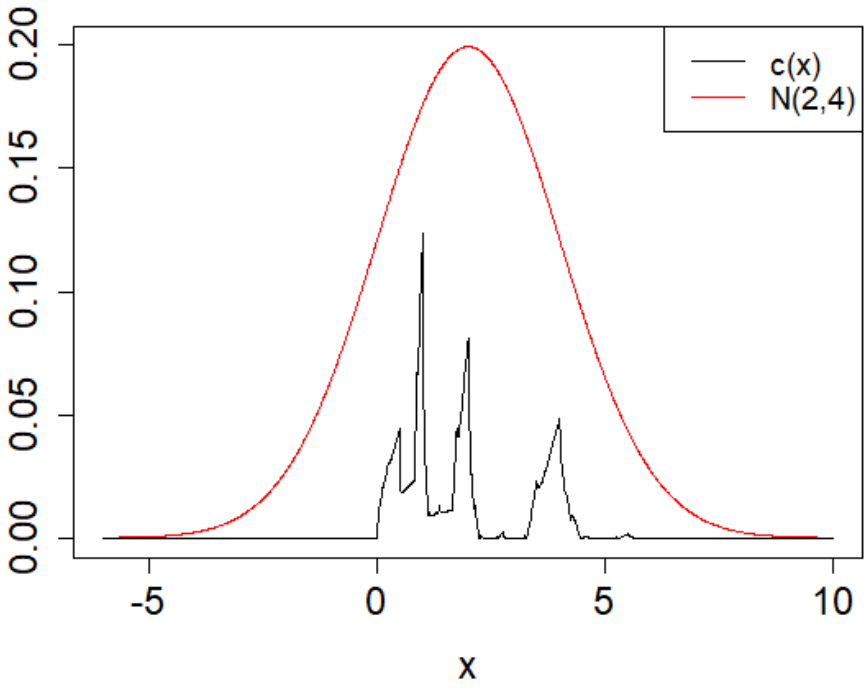



Figure 3.3: *For the starting distribution* $\mathcal{N}(2,4)$ *and the target density as in (3.2.13) with* $\delta = 0.05$ *the assumption A2 is always satisfied, as* $c(x) \leq f(x)\ \forall x \in \mathbb{R}$.

be a distribution with density

$$g(x) = f(x) + \delta \sum_{j=-1}^{1} \psi_{j,0}(x), \tag{3.2.13}$$

where $\psi_{j,0}(x)$ for $j = \{-1, 0, 1\}$ are Daubechies wavelets with two vanishing moments. Compared to the starting normal distribution the shape of the target could be inspected in Figure 3.2.

Using the limited number of compactly supported Daubechies wavelets and scaling parameter $\delta$, we are able to artificially construct a target density for which **A2** assumption always holds (Figure 3.3). We could have chosen Daubechies wavelet with a different number of vanishing moments to construct a valid example. For

different choices of wavelets the acceptable ranges for $\delta$ would change, but due to a finite support feature of Daubechies wavelets, the acceptable ranges for $\delta$ would always exist.

As it could be seen from examples above, pWMC is highly impractical and requires specific conditions for the target density. Fortunately, it is possible to mitigate restrictions of the assumption **A2** by discretising pWMC and approaching $g(x)$ from $f(x)$ in a large number of small steps.

## 3.3 Wavelet Monte Carlo

Let us define a target density at time $t$:

$$f_t(x) = rf(x) + td(x), \tag{3.3.14}$$

where $t$ is an artificial time parameter that indexes all intermediate distributions between $f(x)$ and $g(x)$ in the linear form. We can see that for $t = 0$ and for $t = 1$ we recover the starting and target distributions:

$$f_0(x) = f(x), \qquad f_1(x) = g(x).$$

Suppose we were to apply pWMC at each time $t \in \{\mathrm{d}t, 2\,\mathrm{d}t, 3\,\mathrm{d}t, ..., 1\}$, where $\mathrm{d}t > 0$ is an arbitrarily small value. At each time $n\,\mathrm{d}t$, $n \in \mathbb{N}$, pWMC would be applied to decide whether to stay at our current point or to sample a new point according to the sampling rules defined in the pWMC algorithm.

Let $x_t$ denote the $x$-value current at time $t$. Then, according to the pWMC method, a transition intensity for moving via wavelet $(j, i)$ from $x = x_t$ to $y = x_{t+\mathrm{d}t}$, where $y \neq x$, is:

$$\lambda_{t,j,i}(y|x) = \frac{[d^{\psi}_{j,i}\psi_{j,i}(x)]^{-}}{f_t(x)}\Big\{\mathbb{1}(d^{\psi}_{j,i} \geq 0)\frac{\psi^{+}_{j,i}(y)}{A_j} + \mathbb{1}(d^{\psi}_{j,i} < 0)\frac{\psi^{-}_{j,i}(y)}{A_j}\Big\} \tag{3.3.15}$$

which could be written in a slightly simplified form, avoiding indicator functions,

$$\lambda_{t,j,i}(y|x) = \frac{1}{A_j f_t(x)}\Big\{ d_{j,i}^{\psi+}\psi_{j,i}^{-}(x)\psi_{j,i}^{+}(y) + d_{j,i}^{\psi-}\psi_{j,i}^{+}(x)\psi_{j,i}^{-}(y)\Big\}. \tag{3.3.16}$$

Using equation (3.3.16) we are able to write down a total transition intensity for moving from $x = x_t$ to $y = x_{t+\mathrm{d}t}$:

$$\lambda_t(y|x) = \begin{cases} \sum_{j,i}\lambda_{t,j,i}(y|x), & x \neq y \\ -\int_{z\neq y}\sum_{j,i}\lambda_{t,j,i}(\mathrm{d}z|x), & x = y \end{cases} \tag{3.3.17}$$

**Lemma 3.3.1** (Kolmogoroff, 1931)**.** *For a general state-space, and general continuous-time Markov process,*

$$\frac{\mathrm{d}}{\mathrm{d}t}f_t(y) = \int_{-\infty}^{+\infty} f_t(x)\lambda_t(y|x)\,\mathrm{d}x, \tag{3.3.18}$$

*where $f_t(y)$ is the marginal probability density function of the event $y$ at time $t$ and $\lambda_t(y|x)$ is the transition intensity from $x$ to $y$.*

**Theorem 3.3.2** (Gilks, 2017)**.** *Assume that at time $t = 0$ we draw a sample $x_0 \sim f(x)$, and that transition intensities at each time $t \geq 0$ are defined by (3.3.15, 3.3.17), where $f_t(x)$ is defined by (3.3.14) and $d_{j,i}^{\psi}$ is defined by (3.1.8). Furthermore, assume that if $f(x) = f_s(x)$ and $g(x) = f_{s+\delta t}(x)$, then $\exists\delta t > 0$ such that assumption A2 on page 43 holds true always. Given these assumptions, the marginal distribution of the state $x_t$ at any time $t \geq 0$ is given by (3.3.14).*

*Proof of Theorem 3.3.2 Gilks (2017).* From (3.3.14), $f_0(x) = rf(x)$. Hence, (3.3.14) holds at $t = 0$. Assume (3.3.14) holds at a given $t \geq 0$. Then the RHS of the general formula (3.3.18) is, upon substituting $\lambda_t(y|x)$ as defined in (3.3.17),

$$\int_{-\infty}^{+\infty} f_t(x)\lambda_t(y|x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} f_t(x)\Big\{ -\delta(x-y)\int_{z\neq y}\sum_{j,i}\lambda_{t,j,i}(z|x)\,\mathrm{d}z + (1-\delta(x-y))\sum_{j,i}\lambda_{t,j,i}(y|x)\Big\}\,\mathrm{d}x$$

as before, we are using Dirac measure $\delta$ to concentrate mass at $y$,

$$= -f_t(y) \int_{z \neq y} \sum_{j,i} \lambda_{t,j,i}(z|y)\, \mathrm{d}z + \int_{x \neq y} f_t(x) \sum_{j,i} \lambda_{t,j,i}(y|x)\, \mathrm{d}x$$

now using the previously expanded form of $\lambda_{t,j,i}$ in (3.3.16),

$$= -f_t(y) \int_{z \neq y} \sum_{j,i} \frac{1}{A_j f_t(y)} \Big\{ d_{j,i}^{\psi+} \psi_{j,i}^{-}(y) \psi_{j,i}^{+}(z) + d_{j,i}^{\psi-} \psi_{j,i}^{+}(y) \psi_{j,i}^{-}(z) \Big\}\, \mathrm{d}z$$
$$+ \int_{x \neq y} f_t(x) \sum_{j,i} \frac{1}{A_j f_t(x)} \Big\{ d_{j,i}^{\psi+} \psi_{j,i}^{-}(x) \psi_{j,i}^{+}(y) + d_{j,i}^{\psi-} \psi_{j,i}^{+}(x) \psi_{j,i}^{-}(y) \Big\}\, \mathrm{d}x$$

now after completing all integrals, which all evaluate to $A_j$ and after canceling $A_j$ terms, we end up with

$$= \sum_{j,i} \Big\{ - d_{j,i}^{\psi+} \psi_{j,i}^{-}(y) - d_{j,i}^{\psi-} \psi_{j,i}^{+}(y) + d_{j,i}^{\psi+} \psi_{j,i}^{+}(y) + d_{j,i}^{\psi-} \psi_{j,i}^{-}(y) \Big\}$$

noticing common factors, we get

$$= \sum_{j,i} \big( d_{j,i}^{\psi+} - d_{j,i}^{\psi-} \big) \big( \psi_{j,i}^{+}(y) - \psi_{j,i}^{-}(y) \big)$$

using (3.1.3),

$$= \sum_{j,i} d_{j,i}^{\psi} \psi_{j,i}(y)$$

from (3.1.7) we finally arrive at

$$= d(y)$$
$$= \frac{\mathrm{d}}{\mathrm{d}t} f_t(y).$$

Hence, at a given time $t$, the Markov process equation (3.3.18) holds with the marginal distribution given by (3.3.14).

Therefore, by induction, $f_t$ given in (3.3.14) is the marginal distribution of $x_t$ for all time $t \geq 0$. □

If we were to apply pWMC scheme with the transition intensity $\lambda_{t,j,i}$ (3.3.15), an algorithm would involve small probabilities of transition at each of a large number of stages, and hence can not be implemented in practice. However, by noting that point $x_s$ sampled at time $s$ will remain unchanged over many stages, the scheme could be simulated exactly by employing survival analysis theory.

### 3.3.1 Survival analysis

When a transition rate is involved in the analysis of states of a process, survival analysis theory can be applied to calculate probabilities of leaving certain states or of hitting them.

Let $s$ be any time at or after time 0. We will evaluate the probability that $x_s$ does not move up to a time $t > s$. At any time $t \geq s$, assuming the current point $x_t = x_s$, the total moving intensity is,

$$\begin{aligned}
\lambda_t(x_s) &= \int_{y \notin \mathrm{d}x_s} \lambda_t(y|x_s)\,\mathrm{d}y = \sum_{j,i} \int_{y \notin \mathrm{d}x_s} \lambda_{t,j,i}(y|x_s)\,\mathrm{d}y \\
&= \sum_{j,i} \frac{1}{A_j f_t(x_s)} \int_{y \notin \mathrm{d}x_s} \left\{ d_{j,i}^{\psi+} \psi_{j,i}^{-}(x_s)\psi_{j,i}^{+}(y) + d_{j,i}^{\psi-} \psi_{j,i}^{+}(x_s)\psi_{j,i}^{-}(y) \right\} \mathrm{d}y \\
&= \sum_{j,i} \frac{1}{f_t(x_s)} \left\{ d_{j,i}^{\psi+} \psi_{j,i}^{-}(x_s) + d_{j,i}^{\psi-} \psi_{j,i}^{+}(x_s) \right\} \\
&= \frac{1}{rf(x_s) + td(x_s)} \sum_{j,i} [d_{j,i}^{\psi} \psi_{j,i}(x_s)]^{-} \\
&= \frac{c(x_s)}{rf(x_s) + td(x_s)},
\end{aligned}$$

where for future convenience we define

$$c(x_s) = \sum_{j,i} [d_{j,i}^{\psi} \psi_{j,i}(x_s)]^{-}. \tag{3.3.19}$$

From survival analysis theory (Kartsonaki 2016), it is known that, the probability

of a particle not leaving a state $x_s$ in the half-open interval $(s,t]$ is

$$S_{(s,t](x_s)} = \exp\Big\{ -\int_s^t \lambda_{t^*}(x_s)\,\mathrm{d}t^* \Big\} = \exp\Big\{ -\int_s^t \frac{c(x_s)}{rf(x_s)+t^*d(x_s)}\,\mathrm{d}t^* \Big\}. \quad (3.3.20)$$

The value of $S_{[s,t)}(x_s)$ can be computed explicitly by considering three separate cases: for $c(x_s) > 0$ when $d(x_s) = 0$, $c(x_s) > 0$ when $d(x_s) \neq 0$ and when $c(x_s) = 0$.

Case $d(x_s) = 0$: Assuming $c(x_s) > 0$,

$$S_{(s,t](x_s)} = \exp\Big\{ -\int_s^t \frac{c(x_s)}{rf(x_s)}\,\mathrm{d}t^* \Big\} = \exp\Big\{ -(t-s)\frac{c(x_s)}{rf(x_s)} \Big\}. \quad (3.3.21)$$

Then the CDF of a survival variable is

$$F_s(t|x_s) = 1 - S_{(s,t]}(x_s) = 1 - \exp\Big\{ -(t-s)\frac{c(x_s)}{rf(x_s)} \Big\}. \quad (3.3.22)$$

By differentiating $F_s(t|x_s)$ we obtain the PDF

$$f_s(t|x_s) = \frac{c(x_s)}{rf(x_s)} \exp\Big\{ -(t-s)\frac{c(x_s)}{rf(x_s)} \Big\}, \quad (3.3.23)$$

which is a shifted exponential distribution with support $t \in [s, +\infty)$ and rate parameter $\gamma(x_s) = \dfrac{c(x_s)}{rf(x_s)}$.

Case $d(x_s) \neq 0$: Again, assuming $c(x_s) > 0$,

$$\begin{aligned} S_{(s,t](x_s)} &= \exp\Big\{ -c(x_s)\Big[\frac{1}{d(x_s)} \ln\big(rf(x_s)+t^*d(x_s)\big)\Big]_s^t \Big\} \\ &= \exp\Big\{ -\frac{c(x_s)}{d(x_s)}\Big[ \ln\big(rf(x_s)+td(x_s)\big) - \ln\big(rf(x_s)+sd(x_s)\big)\Big] \Big\} \\ &= \exp\Big\{ \ln\Big(\frac{rf(x_s)+sd(x_s)}{rf(x_s)+td(x_s)}\Big)\frac{c(x_s)}{d(x_s)} \Big\} \\ &= \left(\frac{rf(x_s)+sd(x_s)}{rf(x_s)+td(x_s)}\right)^{c(x_s)/d(x_s)}. \end{aligned}$$

Similarly, we derive the CDF of the survival variable,

$$F_s(t|x_s) = 1 - \left(\frac{rf(x_s)+sd(x_s)}{rf(x_s)+td(x_s)}\right)^{c(x_s)/d(x_s)}. \quad (3.3.24)$$

Hence, the PDF is

$$f_s(t|x_s) = \frac{c(x_s)}{f(x_s)+td(x_s)}\left(\frac{f(x_s)+sd(x_s)}{f(x_s)+td(x_s)}\right)^{c(x_s)/d(x_s)-1}. \tag{3.3.25}$$

Here, we note that $t$ has a *scale-location shifted generalised Pareto distribution* (SGP). This could be seen by looking at a standard form of a CDF for SGP variable and matching parameters. The CDF is

$$F_\xi(z) = 1-(1+\xi z)^{-1/\xi}, \quad \xi > 0, \tag{3.3.26}$$

which has a support on $[0,\infty)$. If we let

$$\xi_s = \frac{d(x_s)}{c(x_s)} \quad \text{and} \quad z_s = \frac{c(x_s)(t-s)}{f(x_s)+sd(x_s)}$$

after substituting $\xi_s, z_s$ into (3.3.26) we end up with exactly (3.3.24).

Case $c(x_s) = 0$:

Setting $c(x_s) = 0$, we have $S_{(s,t]}(x_s) = 1\ \forall t > s$. This could be interpreted as the scenario in which there is no force of moving to a different value of $x$, and so the survival time $t = \infty$.

### 3.3.2 Sampling a survival time

In the WMC algorithm we will have to be able to sample a survival time $t$ from $f_s(t|x_s)$, where the functional form of $f_s$ depends on the value of $d(x_s)$.

It is straightforward to produce samples from an exponential distribution. This can be achieved by using standard functions in statistical computing software R (R Core Team 2013) or by applying the inverse sampling formula:

$$t = s - \frac{rf(x_s)}{c(x_s)}\log u_s, \tag{3.3.27}$$

where $u_s \sim \mathcal{U}[0,1]$ is a sample from a uniform distribution on the interval $[0,1]$.

In the case of $d(x_s) \neq 0$, the same inverse sampling technique could be applied

$$t = s + \Big(\frac{rf(x_s)}{d(x_s)} + s\Big)\Big(u_s^{-c(x_s)/d(x_s)} - 1\Big). \tag{3.3.28}$$

### 3.3.3 WMC scheme

Assume that $r$ is known and that we have sampled a point $x_0 \sim f(x)$. We set our initial time $s = 0$ and will perform the steps below repeatedly by replacing the previous time $s$ by $t$ whenever a new point is generated. The process will stop when we have a point $x_t$ whose survival time is $t \geq 1$. At that point, we will set $y := x_t$, and will use it as a sample from $g(x)$.

1. Calculate $c(x_s)$ as in (3.3.19) and $d(x_s)$. If $c(x_s) = 0$, stop and return $y = x_s$. Otherwise, sample $u_s \sim \mathcal{U}[0, 1]$ and set $t$ as in (3.3.27) or (3.3.28), depending on whether $d(x_s) = 0$ or not.

2. If $t \geq 1$, stop and return $y = x_s$. Otherwise, sample a pair $(j, i)$ with probability

$$q_{j,i}(x_s) = \frac{[d^{\psi}_{j,i}\psi_{j,i}(x_s)]^-}{c(x_s)}$$

then sample

$$x_t \sim \begin{cases} \psi^+_{j,i}(x)/A_j, & \text{if } d^{\psi}_{ji} \geq 0 \\ \psi^-_{j,i}(x)/A_j, & \text{if } d^{\psi}_{ji} < 0 \end{cases},$$

set $s = t$ and return to step 1.

## 3.4 Comments on WMC

Here we dive into certain peculiarities of the WMC algorithm and try discuss and analyse them.

### 3.4.1 Approximate computation of $\hat{d}_{j,i}$

As long as one is able to implement the sampling scheme above correctly, samples from a non-standard target distribution will be produced.

One of the goals of any sampling method is to avoid repeated computations of certain integrals. The WMC algorithm, at Step 2, depends on the values of $d^{\psi}_{j,i}$ wavelet coefficients, which involves the computation of an integral

$$\int_{-\infty}^{+\infty} d(x)\psi_{j,i}(x)\,\mathrm{d}x. \tag{3.4.29}$$

This problem could be overcome by applying another sampling method to compute an estimated value of a wavelet coefficients $\hat{d}^{\psi}_{j,i}$. Let us rewrite the integral of (3.4.29) in a slightly different form:

$$\int_{-\infty}^{+\infty} d(x)\psi_{j,i}(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} d(x)(\psi^{+}_{j,i}(x) - \psi^{-}_{j,i}(x))\,\mathrm{d}x$$

rewriting $\psi_{j,i}$ as a combination of a positive and negative part

$$= A_j \int_{-\infty}^{+\infty} d(x)(\psi^{+}_{j,i}(x)/A_j)\,\mathrm{d}x - A_j \int_{-\infty}^{+\infty} d(x)(\psi^{-}_{j,i}(x)/A_j)\,\mathrm{d}x$$

after splitting integrals and by multiplying and deviding by a normalisation constant $A_j$, integrals could be reformulated as expectation

$$= A_j\Big(\mathbb{E}_{\psi^{+}_{j,i}}[d(x)] - \mathbb{E}_{\psi^{-}_{j,i}}[d(x)]\Big),$$

where $\mathbb{E}_{\psi^{+}_{j,i}}$ and $\mathbb{E}_{\psi^{-}_{j,i}}$ denote expectations with respect to $\psi^{+}_{j,i}/A_j$ and $\psi^{-}_{j,i}/A_j$ distributions respectively. This reformulation suggests a possible estimate for $d^{\psi}_{j,i}$

$$\hat{d}^{\psi}_{j,i} = A_j\Big(\frac{1}{N}\sum_{l=1}^{N} d(\pi^{j,i}_l) - \frac{1}{M}\sum_{k=1}^{M} d(\nu^{j,i}_k)\Big), \tag{3.4.30}$$

where $\pi^{j,i}_l$ and $\nu^{j,i}_l$ are samples from distributions $\dfrac{\psi^{+}_{j,i}}{A_j}$ and $\dfrac{\psi^{-}_{j,i}}{A_j}$ respectively, i.e.

$$\big\{\pi^{j,i}_l\big\}_{l=1}^{N} \sim \frac{\psi^{+}_{j,i}}{A_j} \qquad \big\{\nu^{j,i}_k\big\}_{k=1}^{M} \sim \frac{\psi^{-}_{j,i}}{A_j}. \tag{3.4.31}$$

Step 2 of the WMC algorithm involves drawing samples $x_t$ from $\psi^-_{j,i}$ or $\psi^+_{j,i}$ depending on the sign of the wavelet coefficient $d^\psi_{j,i}$. A technique for drawing samples from a positive and negative parts of a wavelet could be also applied in computation of $\hat{d}^\psi_{j,i}$, as the value of an estimate only depends on the number of samples drawn from $\psi^-_{j,i}$ and $\psi^+_{j,i}$, and on the values of those samples.

If we were to replace $d^\psi_{j,i}$ with $\hat{d}^\psi_{j,i}$ in the proof of pWMC (page 44), the argument flow would be exactly the same, however at (3.2.12) we would have

$$p(y|\{\hat{d}^\psi_{j,i}\}) = \frac{1}{\int f(z)\,\mathrm{d}z}\Big\{f(y) + \frac{1}{r}\sum_{j,i}\hat{d}^\psi_{j,i}\psi_{j,i}(y)\Big\}. \tag{3.4.32}$$

Now marginally integrating over the estimates $\hat{d}^\psi_{j,i}$,

$$\begin{aligned} p(y) &= \int p(y|\{\hat{d}^\psi_{ji}\})p(\{\hat{d}^\psi_{j,i}\})\mathrm{d}\hat{d}^\psi_{j,i} \\ &= \mathbb{E}[p(y|\{\hat{d}^\psi_{j,i}\})] \\ &= f(y) + \sum_{j,i}\mathbb{E}[\hat{d}^\psi_{j,i}]\psi_{j,i}(y), \end{aligned}$$

assuming that estimate $\hat{d}^\psi_{j,i}$ is unbiased ($\mathbb{E}[\hat{d}^\psi_{j,i}] = d^\psi_{j,i}$), we get,

$$= f(y) + \sum_{j,i} d^\psi_{j,i}\psi_{j,i}(y).$$

So, all we require is that $\hat{d}^\psi_{j,i}$ is unbiased. By construction, our estimate (3.4.30) is indeed unbiased.

Clearly, using an estimator will affect the total probability of leaving a state $x_t$

$$\sum_{j,i}\frac{[\hat{d}^\psi_{j,i}\psi_{j,i}(x_t)]^-}{rf(x_t)} \tag{3.4.33}$$

and, in turn, will affect the total number of $\hat{d}^\psi_{j,i}$ that need to be estimated in order to achieve a sample from the target.

### 3.4.2 Implications of Theorem 3.3.2

Theorem 3.3.2 proves that, under certain conditions using the transition rate $\lambda_{t,ji}$ defined by (3.3.15) the Kolmogorov forward equation holds, and, hence, $f_t(\cdot)$ is the correct marginal distribution for all times $t \in [0, 1]$. The Markov process induced by $\lambda_{t,ji}$ is essentially applying pWMC algorithm (Section 3.2) at infinitely small time steps. To avoid applying this algorithm at infinitely small increments of time, a survival analysis theory was applied to make the algorithm practical. Although the goal of WMC is to produce samples from the target density $g(\cdot)$, there are points $x_s$ being sampled from intermediate distributions $f_s(\cdot)$, $s \in [0, 1)$ which have an associated survival time $t > s$. What exactly does it mean for a point $x_s$ to have survived for $\delta s = t - s$ amount of time?

At the core of WMC, the transition intensity density $\lambda_{t,ji}$ dictates how the process will unfold and $\lambda_{t,ji}$ is constructed based on pWMC. So, looking from the pWMC perspective, pWMC was applied on point $x_s$ sequentially between times $s \geq 0$ and $t > s$, for $\delta s = t - s$ amount of time. As time was evolving from $s$ to $t > s$, at all instances the point $x_s$ was never 'rejected' because the event of 'no pair $(j, i)$ is selected' was always occurring with probability

$$1 - \sum_{j \in \mathbb{Z}} \sum_{i \in \mathbb{Z}} p_{j,i}(x_s), 0 \leq s < t. \tag{3.4.34}$$

Except at the very last point, at time $t$, a pair $(j, i)$ was sampled indicating that point $x_s$ has survived for $\delta s = t - s$ and now a new point needs to be sampled.

As time was evolving, under the pWMC algorithm the starting sample point $x_s \sim f_s(\cdot)$ was 'accepted' as a sample from all intermediate distribution $f_l(\cdot)$, where $s \leq l < t$. So, as a consequence of Theorem 3.3.2 a sampled survival time $t > s$ for a point $x_s$ with $s \geq 0$ indicates that $x_s \sim f_l(\cdot)$ for $s \leq l < t$. However, we make an observation that we can only make the claim that point $x_s$ is a representative sample from all distributions $f_l(\cdot)$ for $s \leq l < t$ if we do not condition the point $x_s$

on the fact that it did not move for the $t - s$ amount of time, or in other words, $x_s$ point's history from $s$ to $t$, $H_s^t(x_s)$.

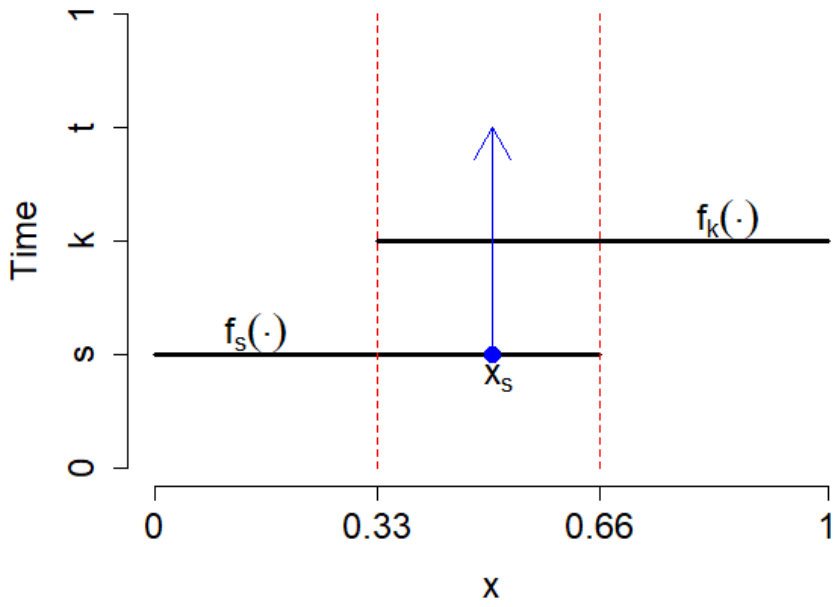



Figure 3.4: *Illustrative example for (3.4.35). Both, $f_s(\cdot)$ and $f_k(\cdot)$ are densities of uniform distributions $\mathcal{U}(0, 0.66)$ and $\mathcal{U}(0.33, 1)$ respectively. Although $x_s$ survives until point in time $t$ and under the standard WMC if we do not condition on the history of the point $x_s$ we would also conclude that $x_s \sim f_k(\cdot)$. However, if we do condition on the history of the point $x_s$ at time $k$, $H_s^k(x_s)$, it is very clear that $x_s \not\sim f_k(\cdot)$, due to the limited range of support it passes through.*

**Definition 3.4.1.** Given any point $x_s$ with $0 \leq s \leq 1$ and time interval $I = (t_1, t_2)$, we denote the history $of x_s$ over the interval $I$ as $H_{t_1}^{t_2}(x_s)$.

In general, we have that if $x_s$ survives until some point in time $t > s$, then

$$x_s | H_s^t(x_s) \not\sim f_l(\cdot), \quad s < l < t. \tag{3.4.35}$$

Figure 3.4 presents an example in which the conditioning issue is rather clearly demonstrated.

On the subject of conditioning, the final target density $g(\cdot)$ could be interpreted as an infinite mixture of distributions, each corresponding to a particular particle history,

$$g(x) = \sum_{H(x)} f\big(x|H(x)\big)p\big(H(x)\big), \tag{3.4.36}$$

where $H(x)$ is a full history of a point $x$ up to a point in time $t = 1$ and $f\big(x|H(x)\big)$ is the conditional density of a point $x$.

### 3.4.3 Finite range of resolution levels

The WMC theorems are proved to hold if one has access to infinite range of resolution levels $j \in (-\infty, +\infty)$ as for example in the decomposition of the difference function $d(x)$ in 3.1.7, it is clear that, in practice, we will restrict ourselves to coarsest $j_{\min}$ and finest $j_{\max}$ resolution levels when implementing WMC. How does this restriction affect samples produced from the target and in particular given this restriction from which exactly target samples are being produced?

The moment the restriction is made, we no longer have access to $j > j_{\max}$ and $j < j_{\min}$ levels and it is clear that samples produced by WMC using a limited range of resolution levels cannot be from the target $\sum_{j,i} g^{\psi}_{j,i}\psi_{j,i}(x)$. As we start with samples from $f(x) = \sum_{j,i} f^{\psi}_{j,i}\psi_{j,i}(x)$ all we really doing in WMC is changing coefficients from $f^{\psi}_{j,i}$ to $g^{\psi}_{j,i}$. If we have access to an infinite range of resolution levels, eventually all coefficients could be changed; however, working with a limited range certain levels are restricted and therefore some $f^{\psi}_{j,i}$ coefficients stay the same. For this reason, our actual distribution at $t = 1$ in practice becomes,

$$\hat{g}(x) = \sum_{j=j_{\min}}^{j_{\max}} \sum_{i} g^{\psi}_{j,i}\psi_{j,i}(x) + \sum_{j<j_{\min}} \sum_{i} f^{\psi}_{j,i}\psi_{j,i}(x) + \sum_{j>j_{\max}} \sum_{i} f^{\psi}_{j,i}\psi_{j,i}(x). \tag{3.4.37}$$

Assuming that $\hat{g}(\cdot)$ above satisfies the probability density properties, if we were to replace $g(\cdot)$ with $\hat{g}(\cdot)$ in Theorem 3.3.2, the proof would still hold and in addition our resolution range across which WMC would be performed would be limited to $j \in [j_{\min}, j_{\max}]$. This would mean that algorithm could be implemented exactly.

Generally, $\hat{g}(x)$ will not satisfy density properties, specifically non-negativity everywhere; for this reason, $g(\cdot)$ will be used as the target in practice, but WMC samples will be treated as though they are from $\hat{g}(x)$.

### 3.4.4 WMC visually

Given a rather complicated nature of the WMC algorithm it might be at first tricky to visualise the process of WMC in action. In Figure 3.5, we present a scheme of what is happening at each step of the algorithm. The example is presented in a situation when we are interested in producing samples from some concentrated normal distribution (red) using samples from a more shallow normal distribution (blue). Samples from blue are being propagated in time until the point they die, and a random wavelet is used to sample a new point to continue a process to the target.

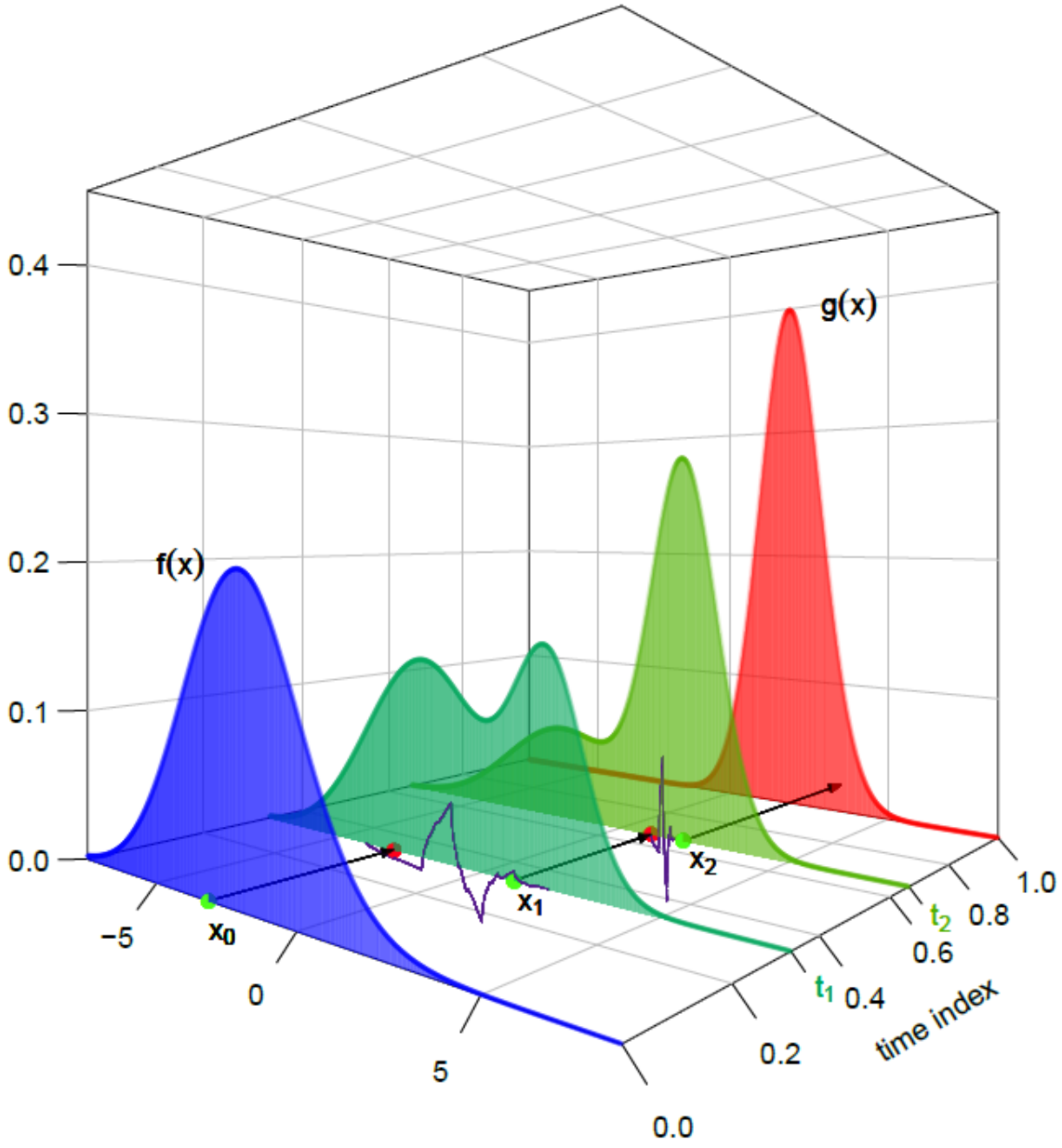



Figure 3.5: *A visual representation of the WMC algorithm. Starting with a sample $x_0$ from $f(x)$ a point survives until time $t_1$, when a new point $x_1$ needs to be sampled according to Step 2 in the WMC scheme. The process is repeated until a point $x_2$ which survival time is $t \geq 1$ at which point the algorithm ceases, producing a sample $y := x_2 \sim g(x)$.*

# Chapter 4

# Implementation of WMC

In this chapter, we will demonstrate how the WMC algorithm could be implemented in practice. The statistical software environment R (R Core Team 2013) will be used as a platform to test and benchmark the algorithm. WMC could be implemented in much faster programming languages like C or Python; however, for purposes of fast coding and easy access to other statistical packages that will be used in the the further analysis, only R will be used here.
A few one-dimensional and two-dimensional examples will be presented here to fully demonstrate some key features of WMC. Given the exponential growth of computational cost with the dimensionality of a problem, three and higher dimensions will not be explored.

## 4.1 Examples

### 4.1.1 1D

In the first example, we will focus on a one-dimensional problem. We will be interested in producing samples from a mixture of standard distributions. Usually

one would try to chose a starting distribution similar to the target; however, here, we will choose the starting distribution to be a uniform defined outside the effective support of the target $g(\cdot)$. The idea of the following example is to demonstrate that even by picking a starting distribution which is substantially different from the target, correct samples can still be obtained using WMC. Throughout these examples, target distributions will be picked such that their normalisation constant is known, and we have an access to a ratio of normalising constants $r$. Clearly, perfect knowledge of $r$ is unrealistic in practice and this will be covered in Chapter 5. Figure 4.1 presents 1000 samples produced by the WMC

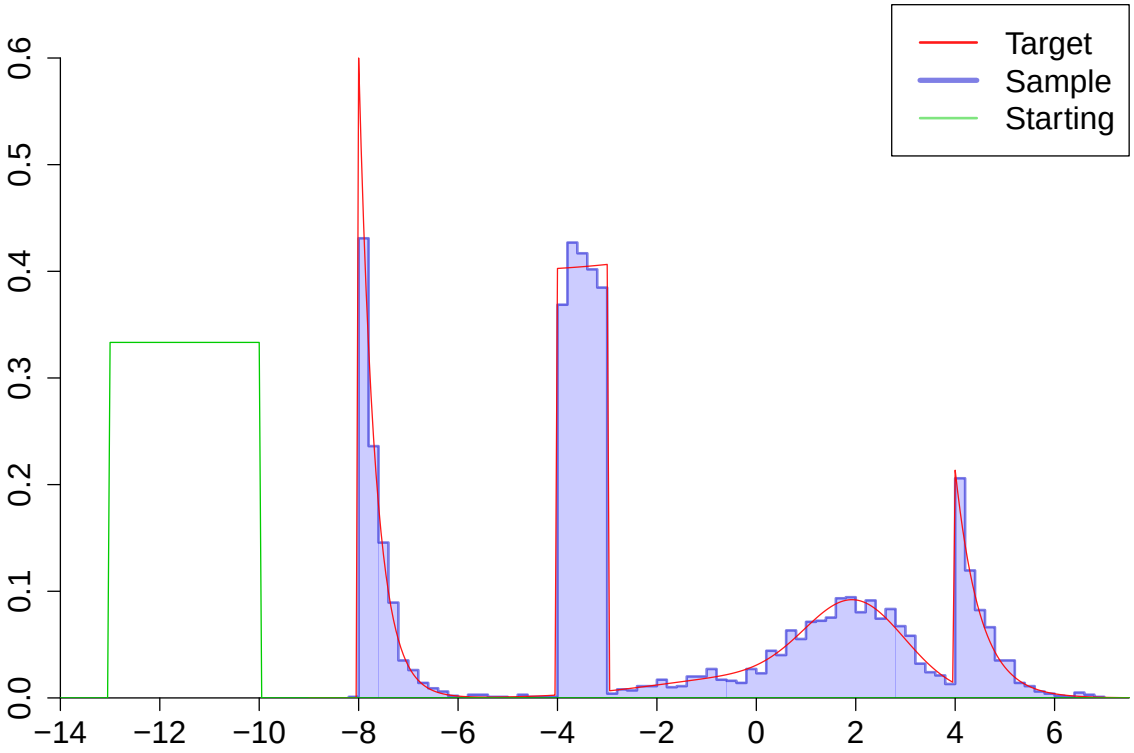



Figure 4.1: *Output of the WMC algorithm produced from using starting distribution* $\mathcal{U}[-13,-10]$ *and the target distributions defined by (4.1.1). The blue histogram depicts a sample of 1000 points from the WMC algorithm.*

algorithm performed using $\mathcal{U}[-13,-10]$ as a starting distribution and target being

a mixture of standard distributions $g(y) = \sum_{k=1}^{5} \omega_k g_k(y)$. Weights were picked to be $\{\omega_k\}_{k=1}^{5} = \{0.2, 0.1, 0.4, 0.1, 0.2\}$ and the mixture components are:

$$
\begin{aligned}
g_1(y) &= \frac{1}{\sqrt{2\pi}} \exp\Big\{ -\frac{(y-2)^2}{2} \Big\}, \quad y \in \mathbb{R}; \\
g_2(y) &= \frac{1}{\sqrt{8\pi}} \exp\Big\{ -\frac{y^2}{8} \Big\}, \quad y \in \mathbb{R}; \\
g_3(y) &= \begin{cases} 1 & y \in [-4, -3], \\ 0 & \text{otherwise}; \end{cases} \\
g_4(y) &= 2 \exp\big\{ -2(y-4) \big\}, \quad y \in [4, +\infty); \text{ and} \\
g_5(y) &= 3 \exp\big\{ -3(y+8) \big\}, \quad y \in [-8, +\infty).
\end{aligned}
\tag{4.1.1}
$$

The Daubechies wavelet with 5 vanishing moments was used to produce results in Figure 4.1. The coarsest resolution level was set to be $j_{\min} = -7$ and the finest one to $j_{\max} = 12$. Sparsity of the wavelet coefficients allowed for the transition from an infinite sum $\sum_{j=-\infty}^{+\infty}$ to a finite one $\sum_{j=-7}^{12}$ by still capturing the information about the most relevant differences between the target $g(\cdot)$ and a starting distribution $f(\cdot)$.

The WMC algorithm seems to be performing quite well even with a starting distribution being chosen from the outside of the effective support of the target. The Kolmogorov-Smirnov test was performed to test the difference between the direct sample form the target $g(y)$ and the WMC one. A p-value of 0.4 was obtained suggesting no significant difference between both. In addition to this, the choice of the multi-modal target did not seem to influence the quality of the samples produced.

In the next one-dimensional example we will consider starting from a standard normal and will try to produce samples from a target that has disjoint probability masses. We again, as before, construct the target distribution as the mixture of standard ones — $g(y) = \sum_{k=1}^{3} \omega_k g_k(y)$. We picked weights to be $\{\omega_k\}_{k=1}^{3} =$

$\{2/3, 1/6, 1/6\}$ and the mixture components as:

$$
\begin{aligned}
g_1(y) &= \frac{1}{9\sqrt{2\pi}} \exp\Big\{ -\frac{(y-30)^2}{2 \times 9^2} \Big\}, \quad y \in \mathbb{R}; \\
g_2(y) &= \frac{1}{0.5\sqrt{2\pi}} \exp\Big\{ -\frac{(y+20)^2}{2 \times 0.5^2} \Big\}, \quad y \in \mathbb{R}; \\
g_3(y) &= \begin{cases} 1 & y \in [40, 41], \\ 0 & \text{otherwise}; \end{cases}
\end{aligned}
\tag{4.1.2}
$$

Results of this example could be observed in Figure 4.2. Wavelet families and

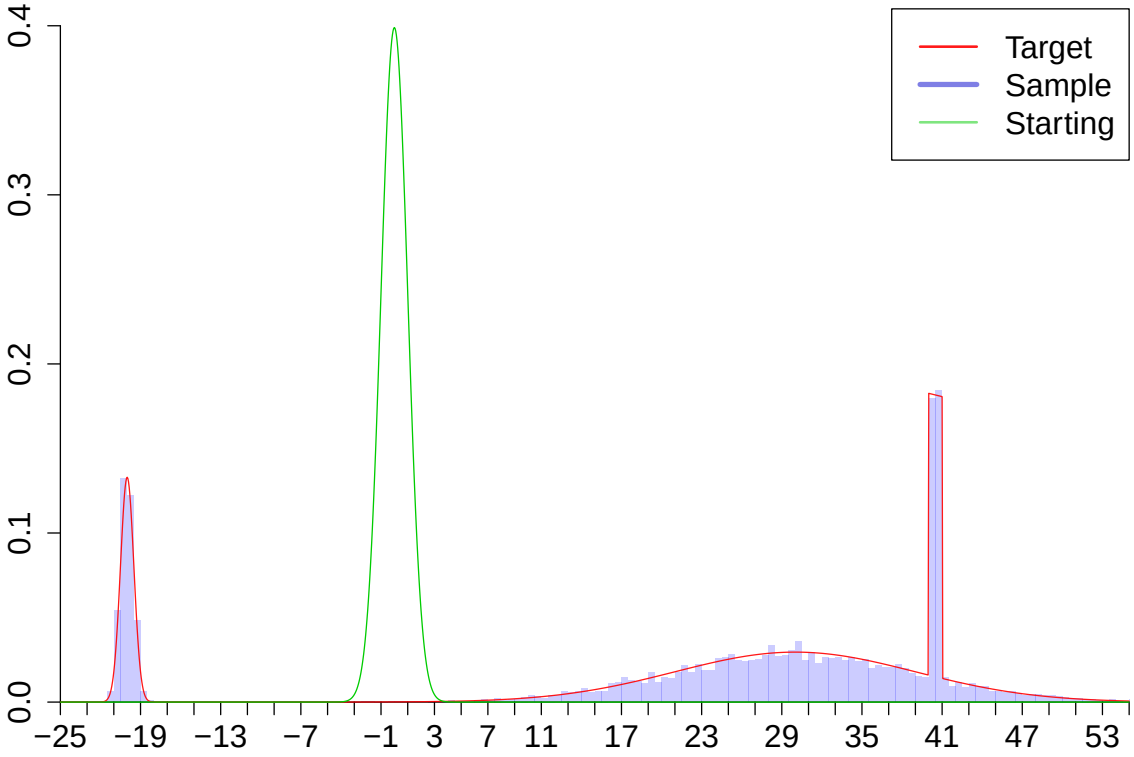



Figure 4.2: *Output of the WMC algorithm produced from using starting distribution* $\mathcal{N}(0, 1)$ *and the target distributions defined by (4.1.2). The blue histogram depicts a sample of 10000 points from the WMC algorithm.*

resolution parameters were kept to be the same as in the previous example, the only difference now being that we produced 10000 points. Visual inspection again suggests satisfactory results; however, this time the p-value associated with the K-S

test is 0.01, so according to the hypothesis test, the sample produced via WMC is not from the target. The associated mean of the target distribution is 23.42 and the standard deviation is 21.13, the WMC sample mean is 22.37 and sample standard deviation is 21.92. As we can see, there is a slight discrepancy in the location parameter and it was high enough for K-S test to find WMC sample to be significantly different from the target. Given the limited range of resolution levels used in the WMC, it is expected to observe a slight discrepancy between WMC sample and the target distribution statistics. Nevertheless, WMC algorithm was able to produce satisfactory samples from the target distribution that had disjoint probability masses across its support. The next step is to investigate how WMC performs in two dimensional space.

### 4.1.2 2D

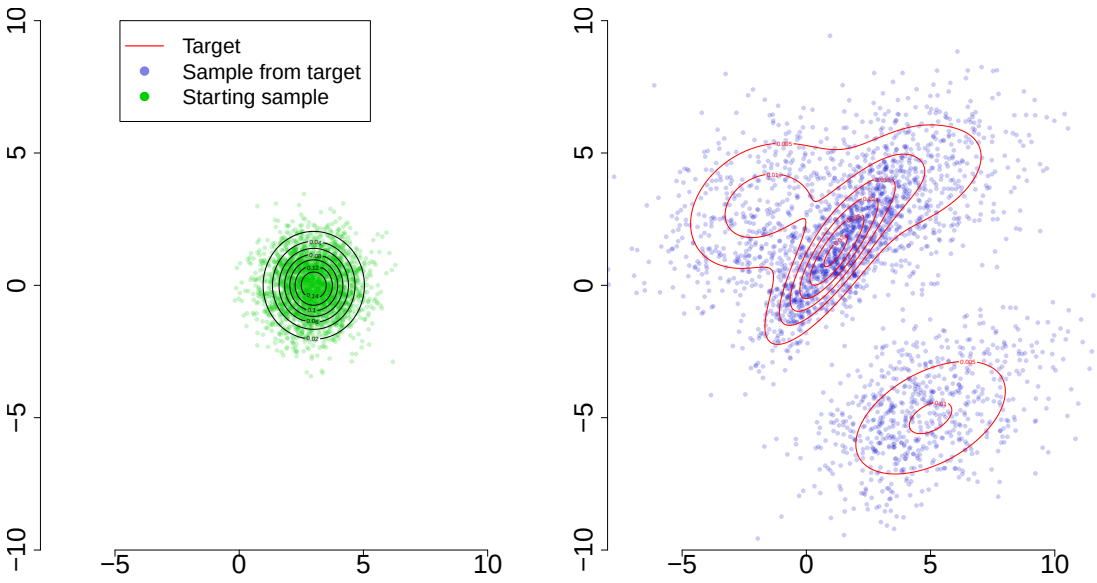



Figure 4.3: *First 2-D WMC example. Although the majority of points are located in the target regions there are some points rather too far from the target, these outlier points will be discussed in Chapter 5.*

In a first two-dimensional example, we will implement WMC to sample from a mixture of normal distributions. The starting density $f(\cdot)$ is going to be that of a two dimensional normal distribution with parameters

$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{4.1.3}$$

and for the target mixture we will have $g(\mathbf{y}) = \sum_{k=1}^{4} \omega_k g_k(\mathbf{y})$, where $\omega_k = 0.25$ for $k = 1, 2, ..., 4$. Next, we will list parameters of normal distributions $k = 1$ to $k = 4$ associated with densities $g_k(\mathbf{y})$:

$$\begin{aligned}
\boldsymbol{\mu} &= \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 2 \\ 2 & 3 \end{pmatrix}, \text{ for } k = 1; \\
\boldsymbol{\mu} &= \begin{bmatrix} 4 \\ 4 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 7 & 2 \\ 2 & 3 \end{pmatrix}, \text{ for } k = 2; \\
\boldsymbol{\mu} &= \begin{bmatrix} -2 \\ 3 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}, \text{ for } k = 3; \\
\boldsymbol{\mu} &= \begin{bmatrix} 5 \\ -5 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 6 & 2 \\ 2 & 3 \end{pmatrix}, \text{ for } k = 4.
\end{aligned}$$

The wavelet used in this example was the two dimensional Daubechies wavelet with $K = 3$:

$$\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x}) = \psi_{j_1,i_1}(x_1)\psi_{j_2,i_2}(x_2), \tag{4.1.4}$$

where $\psi_{j_1,i_1}(x_1)$ and $\psi_{j_2,i_2}(x_2)$ both have 3 vanishing moments.

The coarsest resolution levels were set to be $\mathbf{j}_{\min} = (-2, -2)^T$ and the finest ones to $\mathbf{j}_{\max} = (12, 12)^T$. In this particular example resolution levels were set symmetrically in both directions, however in general one is able to choose different coarsest and finest levels for each direction in $\mathbb{R}^d$ space. Results of the first example in two dimensions can be seen in Figure 4.3.

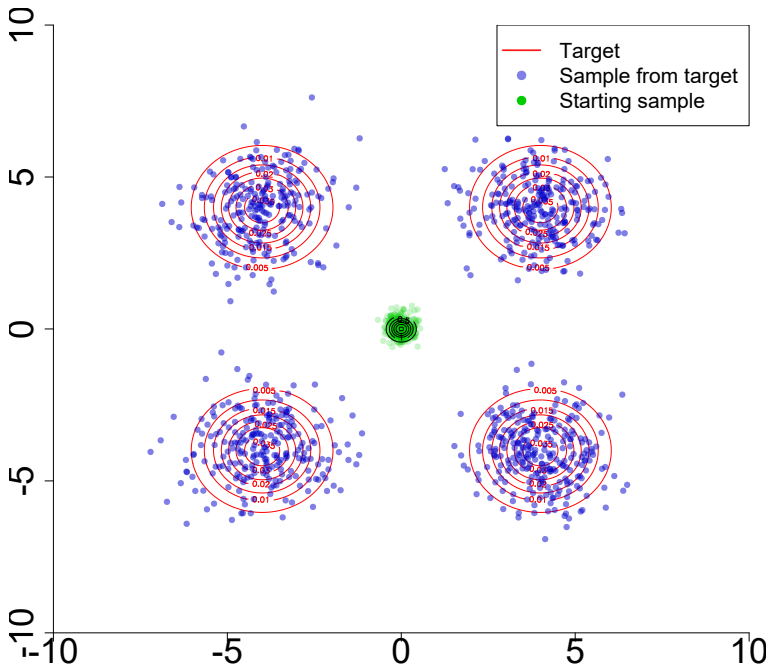



Figure 4.4: *Second 2-D WMC example. After picking starting distribution in the extremely low probability region points are still being sampled appropriately from the multi-modal target distribution.*

For the second example we construct our target using four normal distributions in such a way that there are extremely low probability regions in between the modes. The starting distribution will be chosen to be a highly concentrated normal centred at $(0, 0)$. The target distribution will be again of the form $g(\mathbf{y}) = \sum_{k=1}^{4} \omega_k g_k(\mathbf{y})$, with $\omega_k = 0.25\ \forall k$. Each $g_k(\mathbf{y})$ will have $\mathbf{\Sigma} = \text{diag}(1, 1)$ (a covariance matrix with diagonal components being equal to 1, and off-diagonal components 0). Means of distributions associated with each density $g_k(\mathbf{y})$ will be $\boldsymbol{\mu_1} = (4, 4)^T$, $\boldsymbol{\mu_2} = (4, -4)^T$, $\boldsymbol{\mu_3} = (-4, -4)^T$, $\boldsymbol{\mu_4} = (-4, 4)^T$. The resolution levels were chosen to be $\mathbf{j}_{\text{min}} = (-2, -2)^T$ and $\mathbf{j}_{\text{max}} = (8, 8)^T$, and wavelet of choice was Daubechies with $K = 3$. Results of 1000 WMC samples for this particular case could be observed in Figure 4.4.

## 4.2 Discrete inverse sampling (DIS)

### 4.2.1 Sampling from the discrete density approximation

Given the non-probabilistic nature of wavelets $\psi_{j,i}$, treating them as probability distributions might sound a bizarre idea. However, as demonstrated in the theory of WMC, after normalisation of $\psi^+_{j,i}$ and $\psi^-_{j,i}$ new samples need to be drawn from these parts in order to proceed through the WMC algorithm. In this section a method for producing samples from $\psi^+_{j,i}$ and $\psi^-_{j,i}$ will be covered.

Let us discretise both $\psi^+_{j,i}(\cdot)$ and $\psi^-_{j,i}(\cdot)$. Denote the support $I_{j,i} = [a, b] = \text{supp}\{\psi_{j,i}(\cdot)\}$. We will denote the discretised version of the interval $I_{j,i}$ by

$$\mathbf{x} = \left(x_1 = a, x_2 = a + \frac{(b-a)}{n}, x_3 = a + \frac{2(b-a)}{n}, ..., x_n = a + \frac{(n-1)(b-a)}{n}, x_{n+1} = b\right). \tag{4.2.5}$$

We will also define vectors $\Psi^+$ and $\Psi^-$, which contain the evaluations of $\psi^+_{j,i}(\cdot)$ and $\psi^-_{j,i}(\cdot)$ at points of $\mathbf{x}$:

$$\Psi^+ = \left(\psi^+_{j,i}(x_1), \psi^+_{j,i}(x_2), ..., \psi^+_{j,i}(x_{n+1})\right), \tag{4.2.6}$$

$$\Psi^- = \left(\psi^-_{j,i}(x_1), \psi^-_{j,i}(x_2), ..., \psi^-_{j,i}(x_{n+1})\right). \tag{4.2.7}$$

By setting a large value of $n$, numerical integration could be performed using $\Psi^+$, $\Psi^-$, $\mathbf{x}$ and $\delta x = \frac{b-a}{n}$ as a finite differential to get a value of the normalisation constant $A_j$ of the functions $\psi^+_{j,i}(\cdot)$ and $\psi^-_{j,i}(\cdot)$. By applying cumulative sums to elements of vectors $\Psi^+$ and $\Psi^-$, a discretised version of the cumulative distribution functions $\mathbf{P}^+$ and $\mathbf{P}^-$ for densities $\psi^+_{j,i}(\cdot)/A_j$ and $\psi^-_{j,i}(\cdot)/A_j$ can be obtained:

$$\mathbf{P}^- = \left(p^-_l = \sum_{k=1}^{l} \psi^-_{j,i}(x_k)\delta x\right)_{l=1}^{n+1}, \tag{4.2.8}$$

$$\mathbf{P}^+ = \left(p^+_l = \sum_{k=1}^{l} \psi^+_{j,i}(x_k)\delta x\right)_{l=1}^{n+1}. \tag{4.2.9}$$

We shall now assemble a discrete version of an inverse sampling algorithm (DIS) which can be performed on vectors $\mathbf{P}^+$ and $\mathbf{P}^-$ to produce samples from $\psi^+_{j,i}(\cdot)/A_j$ and $\psi^-_{j,i}(\cdot)/A_j$.

### 4.2.2 Pseudo code

Steps for producing samples from positive and negative parts of the wavelets will be presented here.

0. Obtain values of $\mathbf{x}$, $\mathbf{P}^+$ and $\mathbf{P}^+$.

1. Sample $u \sim \mathcal{U}[0,1]$ and compute $k^+_{\min}$ for producing a sample from $\psi^+_{j,i}(\cdot)/A_j$

$$k^+_{\min} = \underset{k\in\{1,2,\ldots,n+1\}}{\arg\min} |p^+_k - u|, \tag{4.2.10}$$

   or compute $k^-_{\min}$ for producing a sample from $\psi^-_{j,i}(\cdot)/A_j$

$$k^-_{\min} = \underset{k\in\{1,2,\ldots,n+1\}}{\arg\min} |p^-_k - u|. \tag{4.2.11}$$

2. Having obtained $k^+_{\min}$ or $k^-_{\min}$, we report samples from positive and negative parts of the wavelet $\psi_{j,i}$ to be

$$x_{k^+_{\min}} \sim \psi^+_{j,i}(\cdot)/A_j, \qquad x_{k^-_{\min}} \sim \psi^-_{j,i}(\cdot)/A_j, \tag{4.2.12}$$

   where values $x_{k^+_{\min}}$ and $x_{k^-_{\min}}$ are the appropriate entries of vector $\mathbf{x}$.

### 4.2.3 DIS in $d$ dimensions

Here we will demonstrate how the DIS algorithm can be applied for sampling from a multidimensional wavelet $\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x})$. Here $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{j} = \{j_1, j_2, ..., j_d\}$ and $\mathbf{i} = \{i_1, i_2, ..., i_d\}$, where $\mathbf{j}$ and $\mathbf{i}$ are resolution and location vectors in $\mathbb{R}^d$. Recall

that the construction of a multidimensional wavelet involves taking a product of wavelets $\{\psi_{j_k,i_k}(x_k)\}_{k=1}^{d}$:

$$\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x}) = \psi_{j_1,i_1}(x_1)\psi_{j_2,i_2}(x_2)\cdots\psi_{j_d,i_d}(x_d). \tag{4.2.13}$$

Now we are interested in finding the normalisation constant

$$A_{\mathbf{j}} = \int_{\mathbf{x}\in\mathbb{R}^d} \psi^{+}_{\mathbf{j},\mathbf{i}}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\mathbf{x}\in\mathbb{R}^d} \psi^{-}_{\mathbf{j},\mathbf{i}}(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{4.2.14}$$

Given that

$$A_{j_k} = \int_{-\infty}^{+\infty} \psi^{+}_{j_k,i_k}(x_k)\,\mathrm{d}x_k = \int_{-\infty}^{+\infty} \psi^{-}_{j_k,i_k}(x_k)\,\mathrm{d}x_k, \tag{4.2.15}$$

we only need to work out how many terms there are in the expanded version of $[\psi_{j_1,i_1}(x_1)\psi_{j_2,i_2}(x_2)...\psi_{j_d,i_d}(x_d)]^{+}$ and $[\psi_{j_1,i_1}(x_1)\psi_{j_2,i_2}(x_2)...\psi_{j_d,i_d}(x_d)]^{-}$ to get the expression of $A_{\mathbf{j}}$.

We observe that given a product $\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x})$ of length $d$, where each term could have a sign of $+1$ or $-1$, there are $2^d$ possible combinations of signs in a product. The value of a sign sub-product of the first $(d-1)$ terms is either $+1$ or $-1$, therefore only the last term in a product determines whether $\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x}) > 0$ or $\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x}) < 0$. For this reason $[\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x})]^{+}$ and $[\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x})]^{-}$ will both have $2^{d-1}$ terms in the expanded form. Combining this observation with (4.2.15), we get that

$$A_{\mathbf{j}} = 2^{d-1}\prod_{k=1}^{d} A_{j_k} = 2^{d-1-\frac{1}{2}\sum_{k=1}^{d} j_k} A_0^d, \tag{4.2.16}$$

where we have used $A_{j_k} = 2^{-j_k/2}A_0$. Now, given the normalisation constant, how do we produce samples from $[\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x})]^{+}/A_{\mathbf{j}}$ and $[\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x})]^{-}/A_{\mathbf{j}}$? We sample a $d-1$ dimensional vector of signs $s_k$ from a Bernoulli distribution,

$$\{s_k\}_{k=1}^{d-1} \sim 0.5\mathbb{1}(s=-1) + 0.5\mathbb{1}(s=+1). \tag{4.2.17}$$

We sample

$$x_k \sim [\psi_{j_k,i_k}(x_k)]^{s_k}/A_{j_k}, \quad \text{for} \quad k = 1,2,...,d-1, \tag{4.2.18}$$

where $s_k$ denotes a sign. Now, if we are interested in producing a sample from $[\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x})]^+/A_{\mathbf{j}}$, sample

$$x_d \sim \begin{cases} [\psi_{j_d,i_d}(x_d)]^+/A_{j_d} & \text{if} \quad \prod_{k=1}^{d-1} s_k = 1; \\ [\psi_{j_d,i_d}(x_d)]^-/A_{j_d} & \text{otherwise,} \end{cases} \tag{4.2.19}$$

and let $\mathbf{x} = (x_1, x_2, ..., x_d)$ be a sample from $[\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x})]^+/A_{\mathbf{j}}$. However, if we are interested in producing a sample from $[\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x})]^-/A_{\mathbf{j}}$, then

$$x_d \sim \begin{cases} [\psi_{j_d,i_d}(x_d)]^-/A_{j_d} & \text{if} \quad \prod_{k=1}^{d-1} s_k = 1; \\ [\psi_{j_d,i_d}(x_d)]^+/A_{j_d} & \text{otherwise,} \end{cases} \tag{4.2.20}$$

and $\mathbf{x} = (x_1, x_2, ..., x_d)$ will be a sample from $[\psi_{\mathbf{j},\mathbf{i}}(\mathbf{x})]^-/A_{\mathbf{j}}$.

As we can see, the independent product structure of a multidimensional wavelet allows for quite convenient sampling procedures. This method for producing samples from positive and negative parts of wavelets could be used in the future to implement WMC in a multidimensional setting.

## 4.3 Parallelisation

Given the independent structure in the WMC between separate realisations $y$ from the target density $g(\cdot)$, the algorithm could be easily parallelised to utilise several central processing units (CPUs) or graphics processing units (GPUs) to speed up the computation.

In our setting R packages 'doParallel' and 'foreach' will be used to parallelise WMC. A sample code is presented bellow, demonstrating the procedures that should be taken to initiate a 'for' loop using several cores. In this scenario 6 CPU cores are used and 'WMC1d_approx' function is used within a 'foreach' loop to produce $N$ samples from the target of interest.

```
library(foreach)    # for parallel comp
library(doParallel) # foreach backend

cl = makeCluster(6) #set the number of CPU cores

registerDoParallel(cl) #register cores

start = Sys.time()
message('Start ',start)

g.wmc.nona <- foreach(i=1:N, .combine = c) %dopar% {
  WMC1d_approx(x = f[i], filtnr = fnr, res = 4096, lowrez = -8,
               maxrez = 11, d_of_x_FUN = d_of_x, dFUNg = FUNg,
               dFUNf= dunif, parFUNg = NULL,
               parFUNf = list(min = -13, max = -10),
               PosA = PosA, posx = posx, posy = posy, negx = negx,
               negy = negy, xy =xy, time = 1, reps = 200)
}

finish = Sys.time()
message('End ',Sys.time())
message('Total running time:',finish - start)

stopCluster(cl) #detach cores
```

We can investigate the relative gain in the computational speed by comparing two identical WMC set-ups, but one being run in a standard for loop and the other in a parallelised one. For this particular benchmarking experiment Daubechies wavelets with $K = 5$ will be used, the starting density and the target one will be chosen as in the 1-D examples subsection 4.1.1. This particular set-up will be run by varying

$N_d$ parameter that is responsible for the accuracy of the estimate $\hat{d}^{\psi}_{j,i}$ used in the computation,

$$\hat{d}^{\psi}_{j,i} = A_j\Big(\frac{1}{N_d}\sum_{l=1}^{N_d} d(\pi_l^{j,i}) - \frac{1}{N_d}\sum_{k=1}^{N_d} d(\nu_k^{j,i})\Big). \tag{4.3.21}$$

The only difference of the above estimate (4.3.21) to (3.4.30) is that this time we are sampling an equal amount of $N_d$ points from the positive and negative parts of a wavelet. Finally, code will be run for $N = 100$ and $N = 200$, where $N$ is the number of samples, to investigate the dependence of time taken to execute code and the number of samples being generated.

As we can see in Figure 4.5 with parallel computing utilised we can get twice as

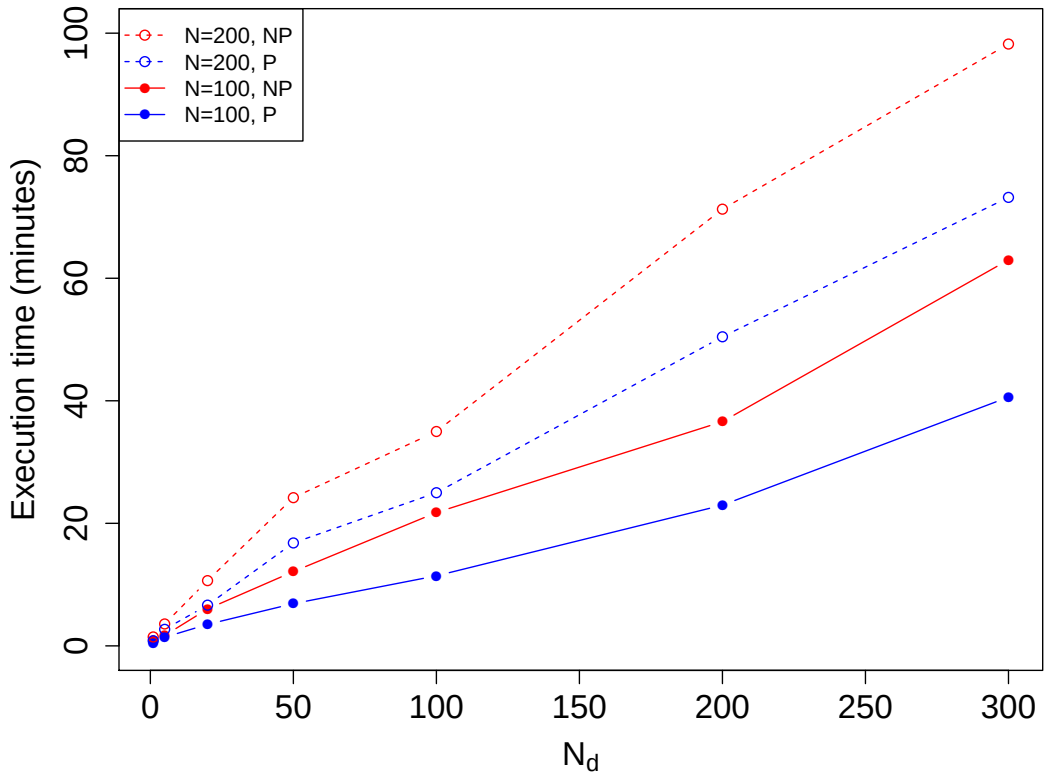



Figure 4.5: *Benchmarking results of 1D WMC for different choices of $N_d$, $N$ and parallelisation option. Where NP stands for 'not parallel' and P for 'parallel' in the legend.*

many samples for approximately the same waiting time. Furthermore, the biggest computational cost savings occur when we decide to increase the value $N_d$ which is responsible for the accuracy in estimating $\hat{d}^{\psi}_{j,i}$ coefficients. On the other hand, time execution savings for $N_d = 1$ seem to be very marginal; nevertheless, parallel computation always seems to dominate the non-parallel one and, hence, is always highly recommended.

## 4.4 Computational cost

Given any sampling method, the computational cost of executing an algorithm is one of the key factors that determines the quality of a method. In this section we will explore the computational cost of WMC algorithm and will analyse how this cost scales with a dimension of a problem.

Most of the computational power used is spent on the estimation of wavelet coefficients $\hat{d}^{\psi}_{j,i}$. We can approximate the total computational load of a single WMC run by the total number of wavelet coefficients that need to be estimated. Here we will focus on Daubechies wavelets, due to the dependence of their support length on the number of vanishing moments $K$ (see section 2.4.2). Given that a chosen Daubechies wavelet has $K$ vanishing moments, the length of support $\text{supp}(\psi_{0,\cdot}(x))$ of a standard mother wavelet in one dimension is $2K-1$, which means that for a given $x \in \mathbb{R}$ and fixed resolution level $j \in \mathbb{Z}$ there are exactly $2K-1$ wavelets that include point $x$ in their support. This means that at each resolution level $2K-1$ wavelet coefficients need to be estimated every time a point is being sampled.

Theoretically, WMC would need to include all resolution levels $j \in \mathbb{Z}$, however as our computation power is finite we will have to pick coarsest and finest resolution levels, they will be denoted $j_{\min}$ and $j_{\max}$. Naturally, by restricting the resolution range we specify the total number of resolution levels $j_{\max} - j_{\min}$ at which coefficients

need to be estimated.

We also observe that a total number of coefficients that needs to be estimated in a single WMC run depends on the total number of jumps $J \in \mathbb{N}_0$ that need to be performed to reach a target $y \sim g(\cdot)$. Taking all factors mentioned above together we can get an expression for the total number of wavelet coefficients that need to be estimated every time WMC is run on a $x_0 \sim f(\cdot)$. In particular,

$$(2K-1)(j_{\max} - j_{\min})(J+1) \tag{4.4.22}$$

turns out to be the total number of coefficients $h_{j,i}^{\psi}$ that need to be estimated. In a $d$-dimensional setting, when $g(x) \in \mathbb{R}^d$, this expression generalises to

$$(2K-1)^d \prod_{k=1}^{d} (j_{k,\max} - j_{k,\min})(J+1), \tag{4.4.23}$$

where $j_{k,\max}$, $j_{k,\min}$ are the coarsest and the finest resolution levels of the associated direction $x_k$.

We can clearly see that the number of coefficients required for WMC grows geometrically with the dimension $d$, which is unfortunately an unfavourable feature. Furthermore, smoother Daubechies wavelets with more vanishing moments would improve the quality of transition from a starting point $x_0 \sim f(\cdot)$ to $y \sim g(\cdot)$, as there are more wavelets to choose from which could perform a transition, hence more precision in performing a single jump. However it does increase the total number of coefficients that need to be estimated. So, although fewer wavelets with high number of vanishing moments are required to approximate a signal accurately, in a WMC setting more vanishing moments mean more computational load.

It is clear that $J$ (total number of jumps) is not deterministic and does follow some probability distribution $\pi(\cdot)$. Given $f(\cdot)$ and $g(\cdot)$, we are interested in drawing some inference about

$$J \sim \pi(\cdot | x_0, K), \tag{4.4.24}$$

where we assume the number of jumps is dependent on the choice of Daubechies wavelet only through the number of vanishing moments $K$. The analysis of $\pi(\cdot|x_0, K)$ will be presented in Chapter 6. However, here we only make an observation that given the target $g(\cdot)$, the expected computational cost of computing a collection of coefficients $\hat{d}^{\psi}_{j,i}$ purely depends on our choice of a starting distribution $f(\cdot)$, the number of vanishing moments $K$ and the total range of resolution levels being used at each direction $x_k$.

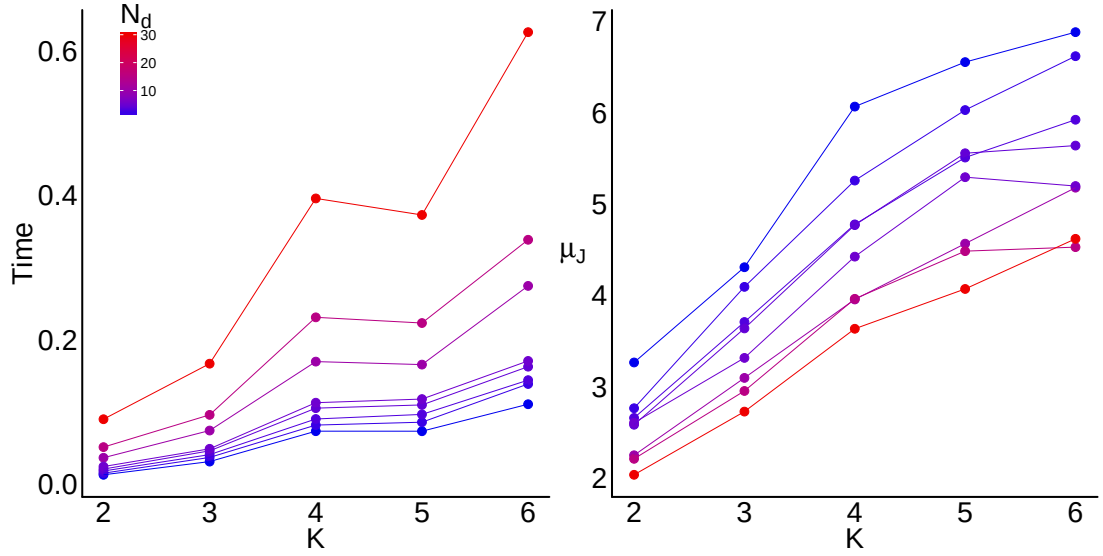



Figure 4.6: *Results of the time (in hours) taken to execute one-dimensional WMC and the average number of jumps made per each sample $x_0 \sim f(\cdot)$ with respect to the choice of Daubechies wavelet and accuracy of estimating $\hat{d}^{\psi}_{j,i}$ . Functions were taken to be the same as in Figure 4.1. As we can clearly see, the choice of wavelets with more vanishing moments increase the average number of jumps required to reach a target and in turn increases the total execution time required to perform WMC. Although more accurate estimation of wavelet coefficients does decrease the average number of jumps, it significantly increases the execution time. Due to the high computation costs it seems that one should stick to Daubechies wavelets with low number of vanishing moments.*

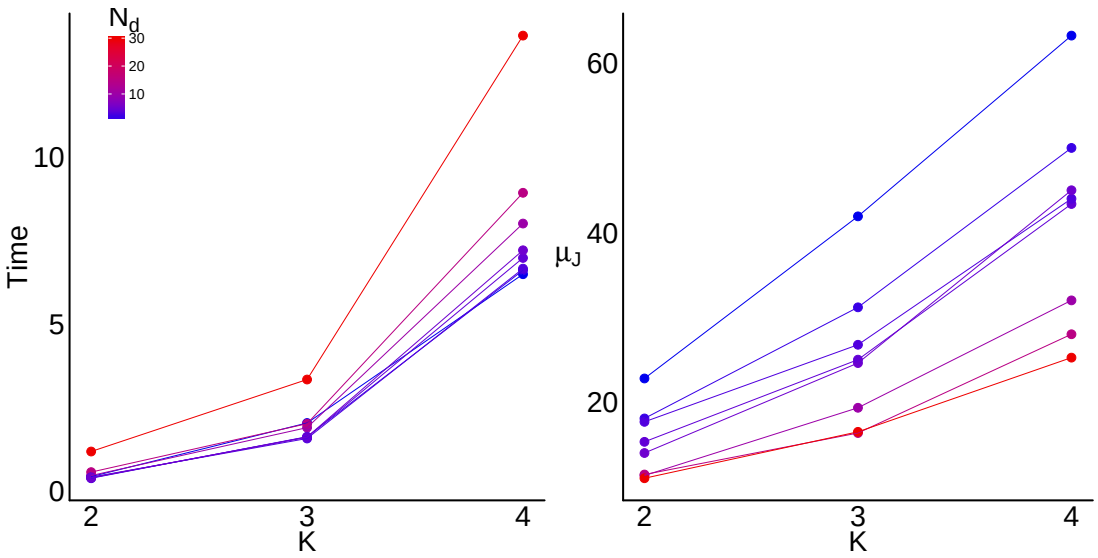



Figure 4.7: *Similarly as in one-dimensional case example the relations between parameters analysed are identically the same in two-dimensional case. Execution time increases even more drastically, greatly supporting an idea of avoiding wavelets with large supports.*

### 4.4.1 Empirical analysis

Here we analyse the relationship between the choice of the number of vanishing moments $K$, the execution time of WMC and the average number of jumps $\mu_J$ required to reach a target:

$$\mathbb{E}[J] = \mu_J. \tag{4.4.25}$$

Given that $d^{\psi}_{j,i}$ coefficients need to be estimated, $N_d$ will denote the number of samples being drawn from $\psi^{+}_{j,i}$ and $\psi^{-}_{j,i}$ for computation of $\hat{d}^{\psi}_{j,i}$ as in (4.3.21).

These results should only be used for purposes of analysing relative computational load but not benchmarking execution speed of WMC itself as computations were not parallelised, choices of $f(\cdot)$ were not optimal with regards to target $g(\cdot)$ and the number of samples $N$ drawn in 1D (Figure 4.6) and 2D (Figure 4.7) examples differed.

## 4.5 Choice of the wavelet family and overall set-up

In this section, we will focus on describing what values should be chosen for various parameters of WMC initialisation. In particular, the choice of the number of vanishing moments $K$ and the coarsest and the finest resolution levels $\mathbf{j}_{\min}$ and $\mathbf{j}_{\max}$. As it was presented in the previous section, the choice of large $K$ values does

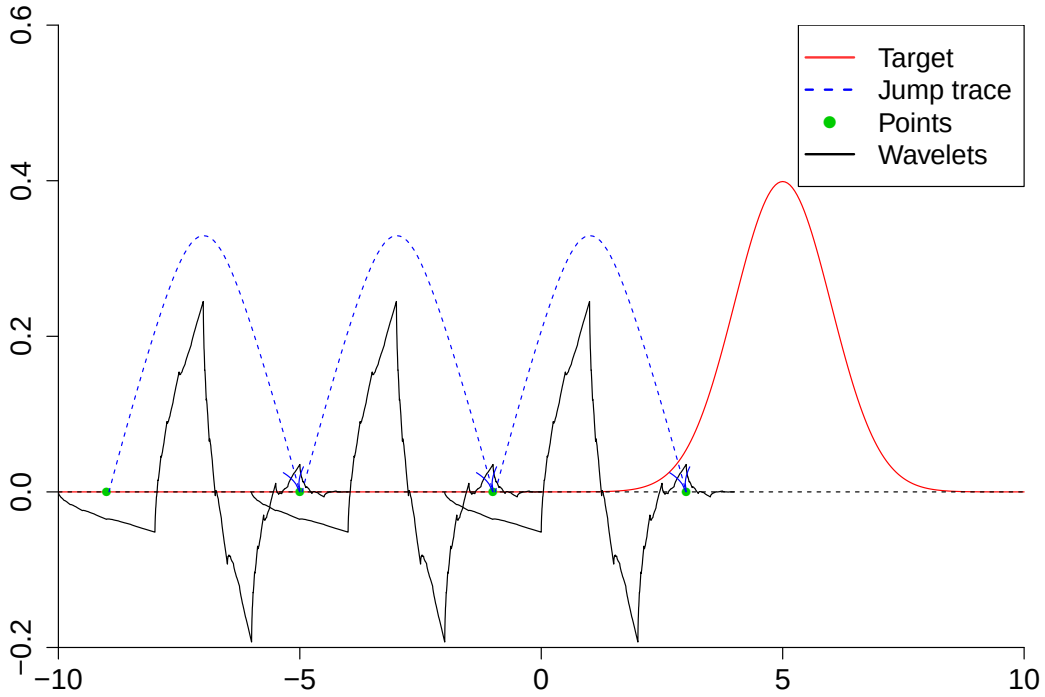



Figure 4.8: *The coarsest resolution level is not coarse enough, leading to the unnecessarily high number of jumps required to move a point to a high density region.*

not seem to bring any positive contributions towards the overall quality of WMC execution. Therefore, $K$ should be kept to the minimum of 2 for the most optimal performance, as this would guarantee the smallest number of jumps required to reach a target in the shortest execution time.

Wavelets in WMC act as tools that transition probability mass from one location to another, for this reason the coarsest and the finest resolution levels need to be

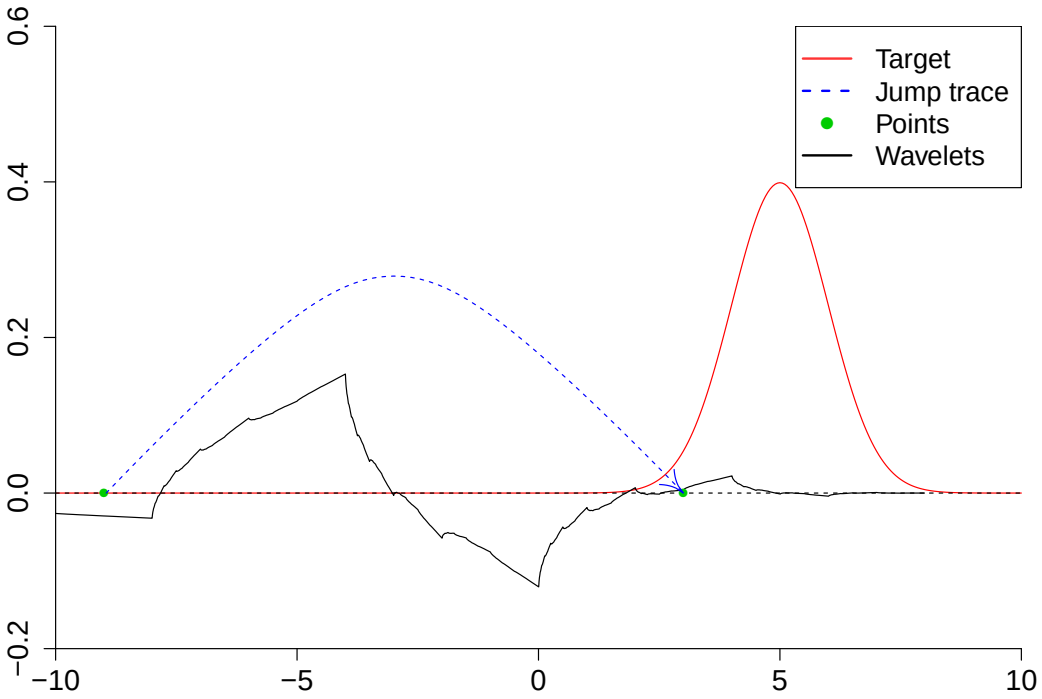



Figure 4.9: *The chosen coarsest resolution level is good enough, providing ability for a point to reach a high density region in a single jump.*

picked carefully to guarantee that every sample point with a starting density $f(\cdot)$ could reach the full range of the effective support of a target $g(\cdot)$ via a single wavelet. This would guarantee that only a single jump is required for a starting point to be moved to the region of high density. If the coarsest resolution level is chosen to be not coarse enough, computational power is going to be wasted for the first few jumps in each WMC run. This issue is illustrated in Figures 4.8 and 4.9.

Choice of the finest resolution level is equally as important. The finest resolution level is going to be responsible for picking up fine details and modes of the target density. If the finest resolution level is too coarse, a sample from WMC is not going to be a representative enough sample from the density $g(\cdot)$. To avoid this, we must make sure that points can be moved freely within the high density regions; this is illustrated in Figure 4.10.

In summary, we want to have $\mathbf{j}_{\max}$ as high as possible, given the computational power limits, as this resolution level will be responsible for extremely fine details of a target. However, choice of $\mathbf{j}_{\min}$ purely depends on where $f(\cdot)$ is located relatively from $g(\cdot)$. The coarsest resolution level $\mathbf{j}_{\min}$ should be chosen such that $\exists \psi_{\mathbf{j}_{\min},\mathbf{i}}$, where $\operatorname{supp}\{\psi_{\mathbf{j}_{\min},\mathbf{i}}\}$ for some $\mathbf{i} \in \mathbb{Z}^d$ contains all high density regions of both $f(\cdot)$ and $g(\cdot)$.

**Definition 4.5.1** (High density regions)**.** Let $f(\cdot)$ be a density, then we will call $H_r$ a high density region if

$$\int_{\mathbf{x}\in H_r} f(\mathbf{x})\,\mathrm{d}\mathbf{x} \approx 1-\epsilon, \tag{4.5.26}$$

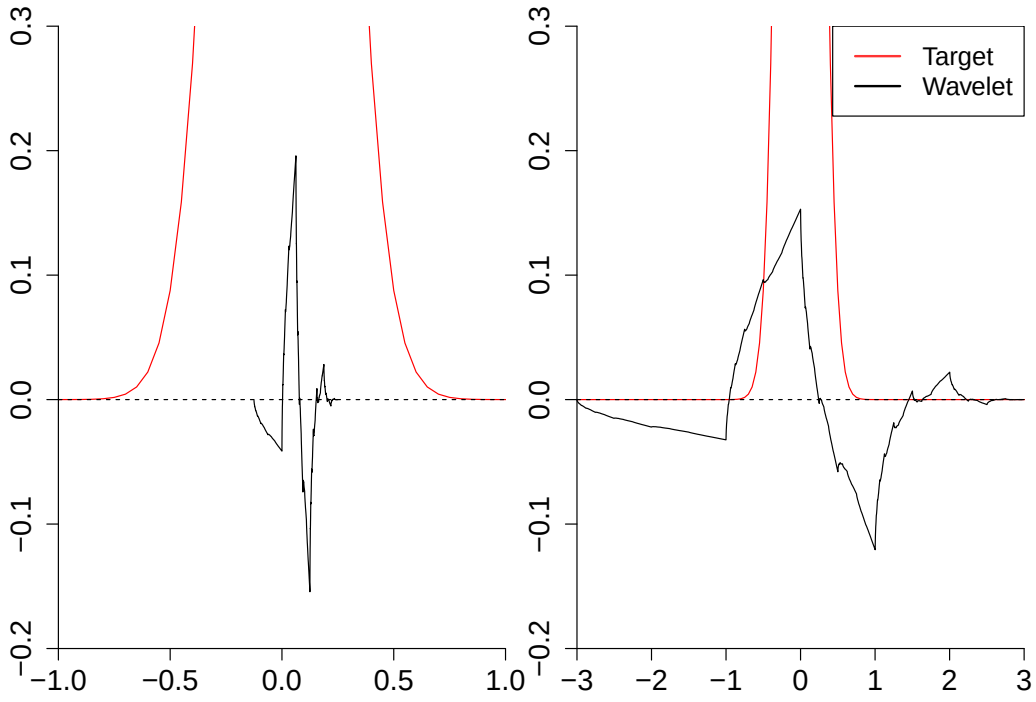



Figure 4.10: *The chosen finest resolution level on the left is good enough, allowing for points to be moved within the high density region and allowing fine wavelets to pick gradual changes in density, however the one on the right side is too coarse, and points will be jumping in and out of the high density region, leading to faulty samples being produced by WMC.*

where $0 < \epsilon \ll 1$.

For example, for a standard normal density $\mathcal{N}(\mu = 0, \sigma^2 = 1)$, if we define $H_r := \{x \in [-2.5, 2.5]\}$, this would ensure that $\approx 99\%$ of data falls to $H_r$,

$$\int_{x \in H_r} \mathcal{N}(x; \mu = 0, \sigma^2 = 1)\, \mathrm{d}x \approx 0.99.$$

In the one dimensional example in Subsection 4.1.1, the coarsest resolution level is chosen to be $j_{\min} = -7$, which actually could be reduced to $j_{\min} = -2$. This would still ensure that $\exists \psi_{j_{\min},i}$ for some $i \in \mathbb{Z}$ that contains $f(\cdot)$ and $g(\cdot)$ (Figure 4.11).

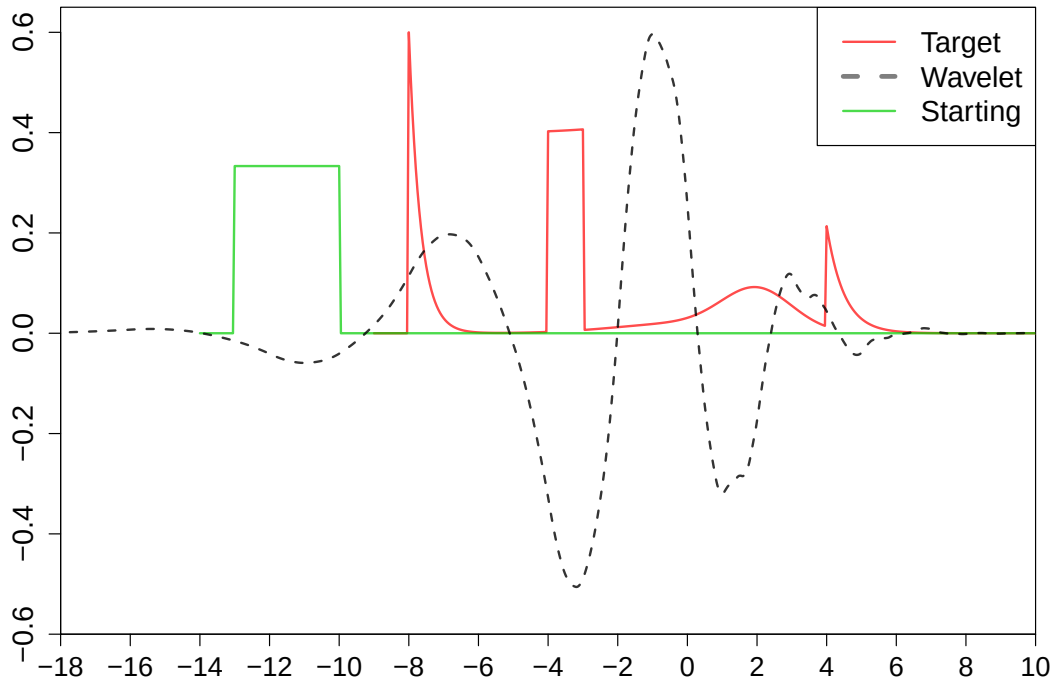



Figure 4.11: *Coarsest wavelet $\psi_{j_{\min},i}$ covers both $H_r$ regions of $f(\cdot)$ and $g(\cdot)$. In this example $K = 5$ and $j_{\min} = -2$.*

## 4.6 Comparison to other MCMC methods

In this section, we compare quality of 1000 samples obtained via WMC to the ones obtained using Metropolis-Hastings (M-H) and Adaptive Rejection Metropolis Sampling (ARMS). We will mainly focus on the out-of-the-box performance of these three algorithms in one-dimensional setting. Kolmogorov-Smirnov test statistic will be used to determine any significant departures from the target distribution. In addition to this, mean and variance together with autocorrelation will be compared across methods. The target density will be chosen to be the one in one-dimensional example in Figure 4.2 (§4.1.1).

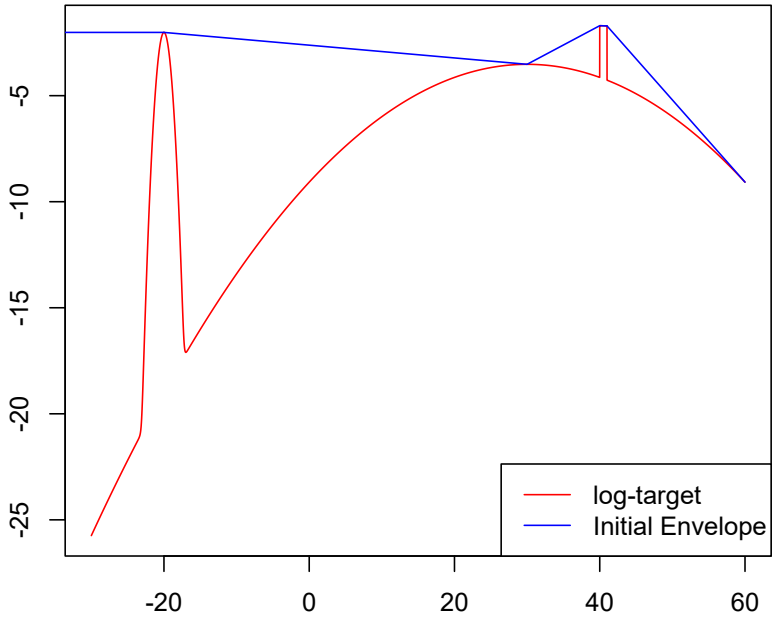



Figure 4.12: *Logarithm of the target density together with a starting envelope for ARMS algorithm.*

Given the nature of one dimension and the choice of a target density, out-of-the-box M-H is doomed to fail as good mixing conditions can not be be achieved with

non-adaptive proposal density in the M-H. Simple Rejection Sampling (RS) is not ideal here, as a huge number of samples is going to be rejected due to poor choices of possible envelopes. ARMS should be able to cope with the target density, as long as we provide a good starting piece-wise envelope for a logarithm of the target density. Starting density for WMC will be $\mathcal{N}(0, 1)$, which is on purpose picked as a poor choice of a starting density, wavelet family was chosen to be Daubechies with 4 vanishing moments and coarsest and finest resolution levels were set to $j_{min} = -7$ and $j_{max} = 12$. For the ARMS algorithm, local and global modes of the target density where provided as initial construction points for the log-envelope (Figure 4.12). A proposal density for M-H algorithm was chosen to be $\mathcal{N}(x_i, 45^2)$, where $x_i$ is a current point in the Markov Chain.

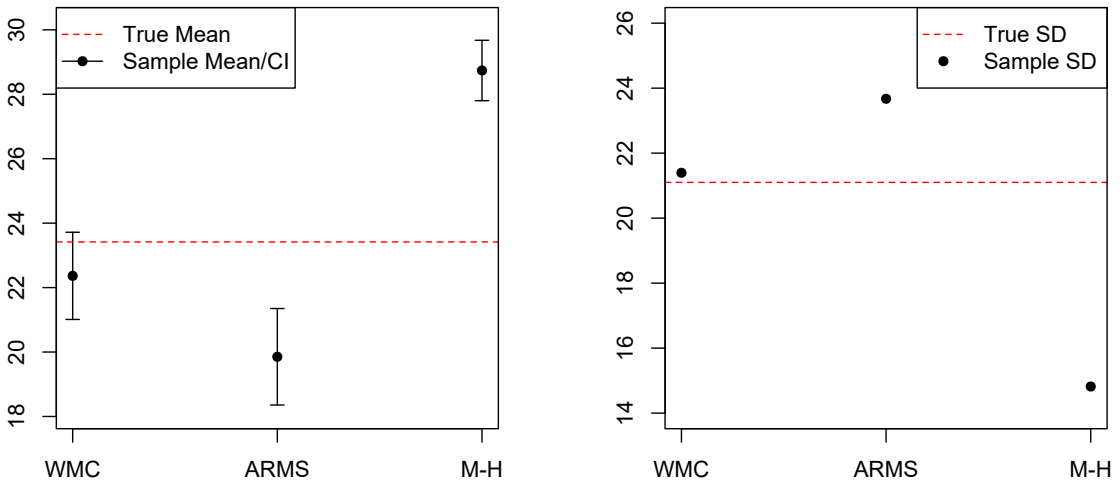



Figure 4.13: *Comparison of means and associated confidence intervals of different sampling methods together with differences in sample standard deviations.*

Inspecting results in Figure 4.14, we can clearly see that both samples produced by

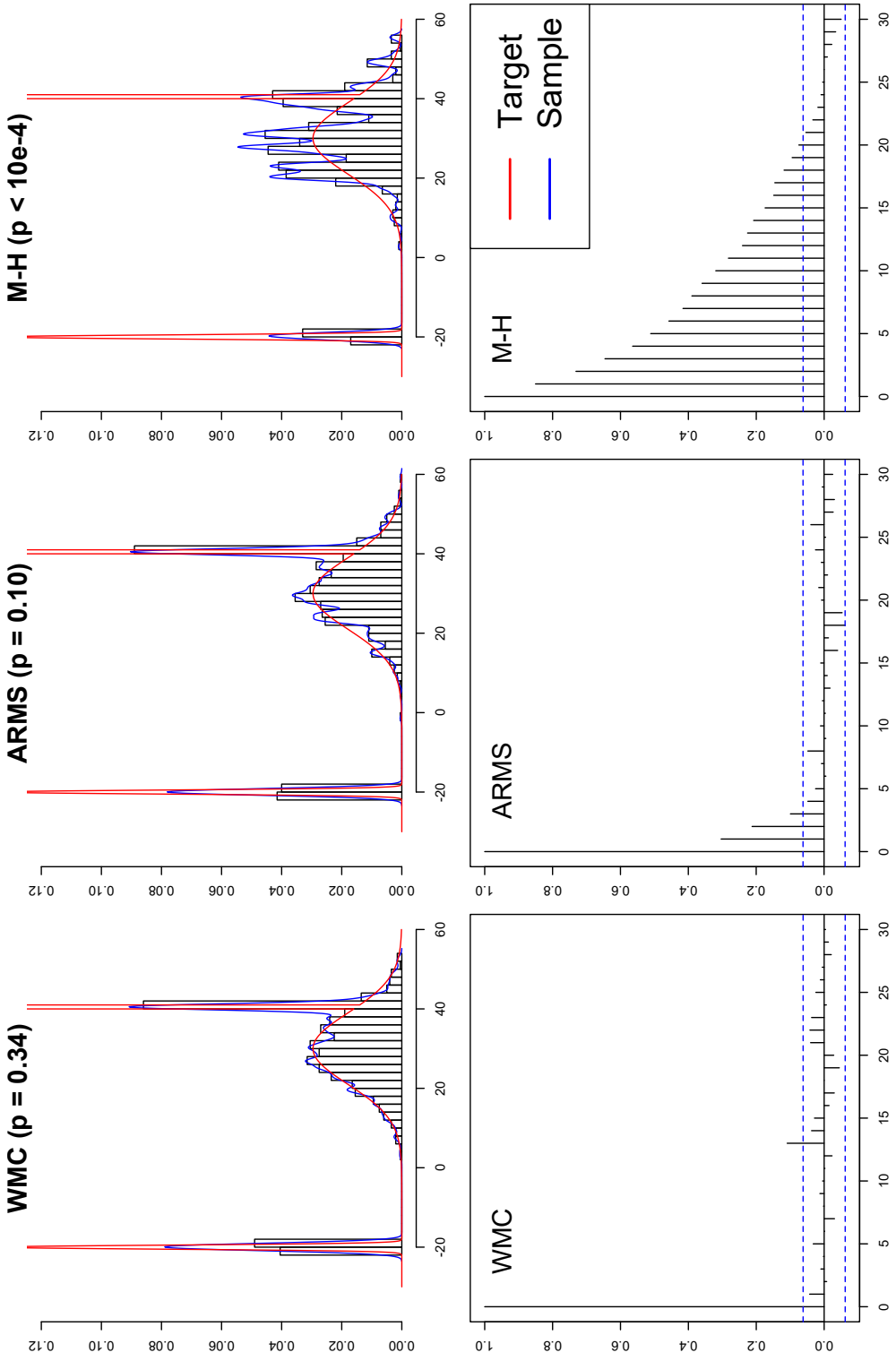



Figure 4.14: *Comparison of results for WMC, ARMS and M-H samples. Associated K-S test statistic is attached to each plot together with autocorrelation function below for in-sample dependence comparison.*

WMC and ARMS did pass the K-S test for the sample comparison with the target density, while M-H algorithm failed as expected. In addition to this, from the autocorrelation function plots, we can observe that samples produced by WMC indicate no dependences within samples, while in ARMS there is a clear correlation present with first two lags, as expected again for M-H case, the algorithm indicates high correlation for all lags up to 20. In addition to these results, we can compare sample means and standard deviations of the methods to the ones of the target density in Figure 4.13. As we can observe in WMC case the confidence interval of the sample mean includes the truth, which could not be observed in other two cases. Similarly for standard deviations, WMC seems to be producing samples with spread that is the closest to the target distribution of all three used methods.

All in all, the out-of-the-box performance of the WMC in the one dimensional setting, using non-optimal choice of a starting distribution, looks to be a very satisfactory one. Compared to a couple of more classic choices of ARMS and M-H, the dependence of samples is eliminated. Furthermore, the ease of implementation of WMC does not require a user to tinker much with starting envelopes, adaptive proposal distributions to ensure the optimal acceptance rates and mixing.

# Chapter 5

# Practical issues of WMC

In this chapter, we will analyse issues surrounding WMC. Some of the issues discussed here will be related to problems encountered more generally in stochastic simulation, such as computing normalising constants and ratios of those (§5.1) and practicality of the algorithm in a high dimensional setting (§5.2). However, some of the problems that we discuss here are unique to WMC: attractor regions (§5.3), ghost points (§5.4) and outliers (§5.5).

## 5.1 Ratio of normalising constants

### 5.1.1 Background

Access to the normalisation constant of the unnormalised probability density $\pi(x)$ of interest usually also implies access to the perfect knowledge about moments of the density itself. The ability to efficiently integrate a density and compute

$$K_\pi = \int \pi(y)\,\mathrm{d}y, \tag{5.1.1}$$

means that $\mathbb{E}[Y]$, $\mathrm{Var}[Y]$ and higher moments could be known explicitly by performing similar integration procedures. In the scenario where this type of integration can be performed analytically, sampling algorithms are usually redundant. Given a high dimensional problem, numerical integration techniques give way to sampling methods. In particular, MCMC methods such as Metropolis-Hastings (M-H) avoid computation of the normalisation constant. In the M-H algorithm, the acceptance probability $\alpha$, only depends on the unnormalised target density $\pi(\cdot)$ and transition kernels $q(\cdot|\cdot)$;

$$\alpha = \min\left(1, \frac{\pi(x^*)q(x^*|x)}{\pi(x)q(x|x^*)}\right). \tag{5.1.2}$$

The normalisation constant $K_\pi$ cancels out and hence does not need to be known explicitly to guarantee that correct samples are being generated from the target. This particular feature is one of the most useful qualities of the MCMC approach, that allows us to completely disregard complex integrals when the dimensionality of a problem is high.

On the other hand, WMC seems to be highly dependent on the value of ratio of normalising constants $r$ (equation (5.1.3)). The intermediate density

$$f_t(x) = rf(x) + td(x)$$

must integrate to 1 for the WMC algorithm to work. Therefore, poor estimates of $r$ will produce samples that are not from the target distribution. It seems that before even implementing the WMC scheme, a quite complicated task of estimating the ratio of normalising constants needs to be performed. The literature regarding estimation of normalising constants and their ratios is extensive (Meng & Wong 1996) and deserves separate investigation. In this thesis, this issue will not be addressed fully; however, a possible method of estimating $r$ will be given later in this section.

The WMC theory is proved to hold (proofs on pages 44 and 50) under the assumption that there exists access to the ratio of normalising constants

$$r = \frac{\int g(y)\,\mathrm{d}y}{\int f(x)\,\mathrm{d}x}. \tag{5.1.3}$$

In practice, $K_f = \int f(x)\,\mathrm{d}x$ is usually known to be $K_f = 1$ because we tend to choose convenient density from which we can sample directly and it is already normalised. In this case, although we are only interested in the ratio $r$, we implicitly need to estimate the normalisation constant of the target density $g(\cdot)$. In a more general setting, the density $f(\cdot)$ of a starting distribution could be unnormalised and in fact the sampling procedure with which samples are obtained from $f(\cdot)$ is not direct, leading to a situation where both $K_f$ and $K_g$ are unknown. However, even with both normalisation constants unknown, only the ratio itself needs to be estimated, leaving us with only one unknown quantity to be estimated rather than two.

### 5.1.2 Estimation of normalisation constant

Recall (2.3.9) that any function $g(\cdot) \in L^2(\mathbb{R})$, in our case $g(\cdot)$ is a density, can be also written in terms of both $\phi_{j_0,i}$ and $\psi_{j,i}$ wavelets rather than only using the mother wavelet $\psi_{j,i}$,

$$g(x) = \sum_i g^{\phi}_{j_0,i}\phi_{j_0,i}(x) + \sum_{j\geq j_0}^{J}\sum_i g^{\psi}_{j,i}\psi_{j,i}(x). \tag{5.1.4}$$

Integrating both sides of (5.1.4) we obtain the representation of the normalisation constant in terms of wavelets and their coefficients,

$$K_g = \int g(y)\,\mathrm{d}y = c_{j_0}\sum_i g^{\phi}_{j_0,i}, \tag{5.1.5}$$

using $\int \psi_{j,i}(x)\,\mathrm{d}x = 0\ \forall\ j,i$ and noting that $\int \phi_{j_0,i}(x)\,\mathrm{d}x = c_{j_0}\ \forall\ i$. In the previous step, we have assumed the exchangeability in the order of infinite sums and infinite integrals, however the validity of this action cannot be guaranteed.

Knowing that we are able to produce samples from a starting distribution, we will use this fact to estimate $g^{\phi}_{j_0,i}$, which can be written as

$$g^{\phi}_{j_0,i} = \int g(x)\phi_{j_0,i}(x)\mathrm{d}x = \int \frac{g(x)}{f(x)}\phi_{j_0,i}(x)f(x)\mathrm{d}x. \tag{5.1.6}$$

From (5.1.6), we can rewrite the coefficient $g^{\phi}_{j_0,i}$ in terms of the expectation with respect to a starting density $f(\cdot)$,

$$g^{\phi}_{j_0,i} = \mathbb{E}_f\left[\frac{g(x)}{f(x)}\phi_{j_0,i}(x)\right]. \tag{5.1.7}$$

Let $\{x_{k,i}\}_{k=1}^{n}$ be a sample from a probability distribution with density $f(\cdot)$, where the subscript $i$ denotes that this is the sample for the estimation of the coefficient $g^{\phi}_{j_0,i}$. Then

$$\hat{g}^{\phi}_{j_0,i} = \frac{1}{n}\sum_{k=1}^{n}\frac{g(x_{k,i})}{f(x_{k,i})}\phi_{j_0,i}(x_{k,i}) \tag{5.1.8}$$

forms an estimate for the father wavelet coefficient $g^{\phi}_{j_0,i}$. Therefore, using this estimate directly in Equation (5.1.5), we get

$$K_g \approx \frac{c_{j_0}}{n}\sum_{i}\sum_{k=1}^{n}\frac{g(x_{k,i})}{f(x_{k,i})}\phi_{j_0,i}(x_{k,i}). \tag{5.1.9}$$

This particular method could be seen as a direct application of the Importance sampling (IS) algorithm (Kahn & Harris 1951). The estimation of normalisation constants and ratios of constants has been a relevant topic and advanced methods have been developed to tackle this issue. Efficient MCMC methods could be employed to estimate ratio of normalising constants (Neal 2005, de Valpine 2008). However, given that MCMC itself needs to be used just to get accurate estimate of $r$, it raises a question why not stick to MCMC sampling directly, skipping the step of estimating the ratio $r$ beforehand. The answer to this difficult question will be apparent in the future chapters of this thesis.

### 5.1.3 Results under the misspecification of the ratio of normalising constants

Here, we present the analysis of how sensitive results of WMC are to the misspecification of the ratio of normalising constants,

$$r = \frac{K_g}{K_f}.$$

In practice, we would normally start with $f(\cdot)$ that is normalised, $K_f = 1$, so usually we would face a situation when $r = K_g$, and the misspecification in $K_g$ is equivalent to the misspecification in $r$. We will use the univariate example presented in §4.1.1 to perform the analysis. We will also stick to the same choice of parameters as before; however, this time we will on purpose misspecify the normalisation constant of the target $g(\cdot)$, in §4.1.1 we had it set to $K_g = 1$. In a perfect scenario, when both $f(\cdot)$ and $g(\cdot)$ are normalised, we end up with $r = 1$. Here, we will run simulations of WMC with,

$$0.5 \leq r \leq 1.5.$$

In each WMC simulation run, we will produce 10000 samples and will compute metrics, to measure the discrepancy from the ideal target. The metrics that we will look at, will be — mean, median, standard deviation, and a p-value from the Kolmogorov–Smirnov test for a two-sample comparison. Results of simulations could be inspected in Figure 5.1.

From the results, we can clearly see that for $r = 1$, across all metrics, we produce results that are in satisfactory ranges from the target. In particular, the target mean is within the expected range of the error bars for the sample mean and the standard deviation together with the median are relatively close to the target as well. Given that we are able to produce samples from $g(\cdot)$, with $K_g = 1$, directly (due to a construction) we can perform a two sample K-S test to test the hypothesis, if two samples were produced from the same distribution. For $r = 1$, with p-value $> 0.6$

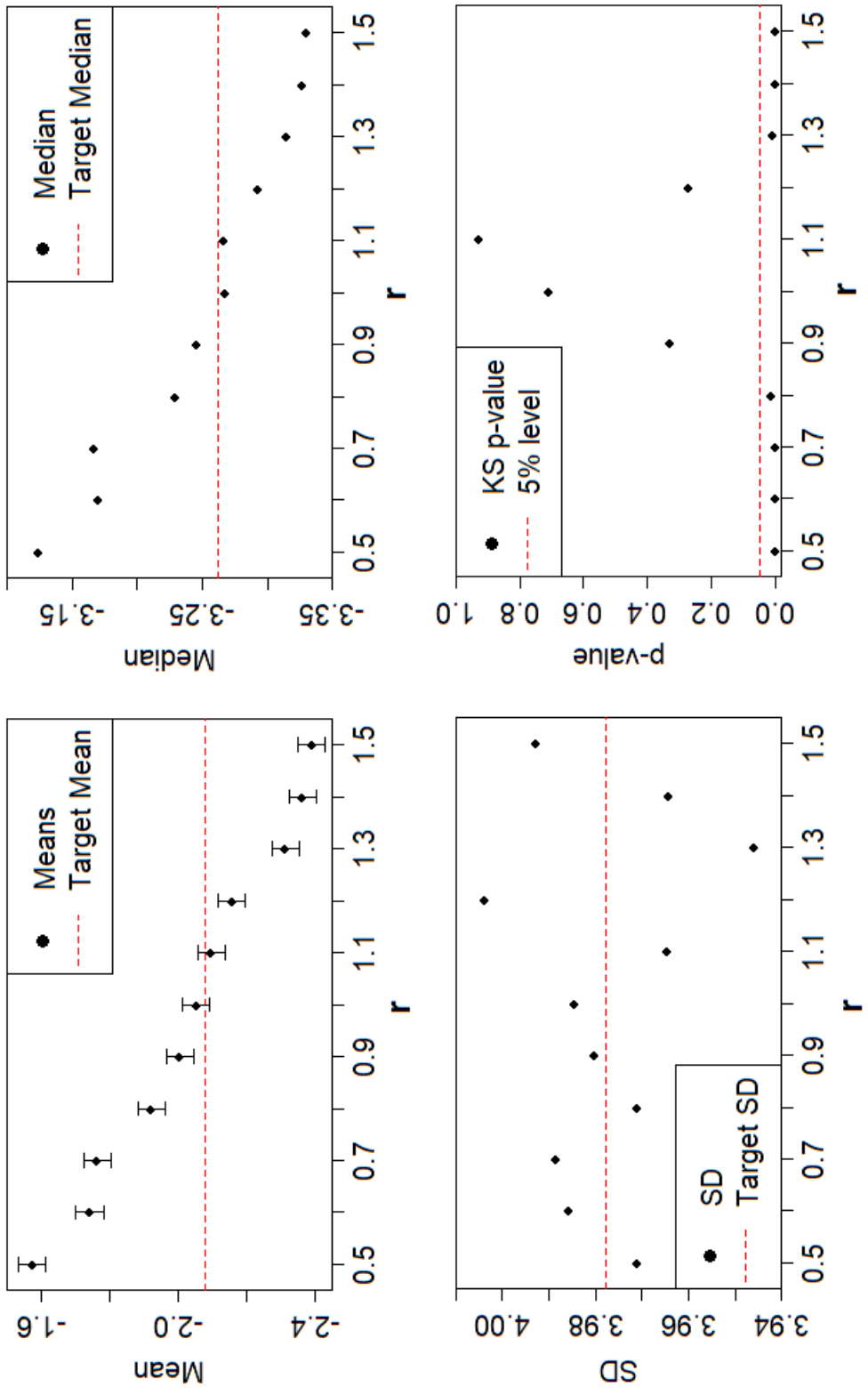



Figure 5.1: *Four plots of four different metrics; mean, median, standard deviation and p-value of K-S test — plotted against the ratio of normalising constants r. Each value of the metric for a given r was calculated over a sample of 10000 points produced via WMC.*

we fail to reject the hypothesis that samples came from different distributions. So, for $r = 1$ the WMC algorithm is indeed performing satisfactory. Unfortunately, in practice we must face cases where we do not have a perfect information about the normalisation constant of our target density $g(\cdot)$. At best, we are able to workout an estimate for the normalisation constant $K_g$. Results show that for sample mean we get a linear decrease in mean as $r$ ranges from 0.5 to 1.5. From this experiment we can see that the penalty for underestimating $r$ seems to be greater compared to overestimating — tests would conclude that WMC samples were produced from the desired target for all cases $0.9 \leq r \leq 1.2$.

We can also observe that for this particular choice of $f(\cdot)$ and $g(\cdot)$ the misspecification in $r$ mainly affects the location metrics and has relatively small effect on the variation. In particular, the maximum departures from the target for sample standard deviation seem to be less than 1%.

In summary, we can see that having a slight misspecification in a ratio $r$ can still lead to surprisingly positive results. Clearly, 'slight misspecification' is a case-dependent statement. A more analytic approach towards the analysis of how the error in $r$ translates to errors in samples at this moment seems intractable, therefore conclusions can only be drawn on the case-specific level.

## 5.2 Curse of dimensionality

As was already discussed in the Section 4.4, the number of wavelet coefficients $d^{\psi}_{j,i}$ that needs to be computed grows geometrically with the dimension of the space $d$ that target density $g : \mathbb{R}^d \mapsto \mathbb{R}$ is defined on.

Given a point $x_t$ at time $t$, we require

$$(2K-1)^d \prod_{k=1}^{d} (j_{k,\max} - j_{k,\min}) \tag{5.2.10}$$

coefficients to be estimated at every intermediate step of the WMC run. By calculating these coefficients, we are essentially locally integrating

$$\int d(x)\psi_{j,i}(x)\,\mathrm{d}x \tag{5.2.11}$$

for all values of $j \in \bigotimes_{k=1}^{d}[j_{k,\min}, j_{k,\max}]$ and $\{i : x_t \in \mathrm{supp}(\psi_{j,i}(x)\}$. As the dimensionality of the space grows, the space gets more complex and the total number of coefficients required to capture details of the space grows geometrically. This feature could not be avoided, which poses serious efficiency issues for the implementation of WMC in high dimensional settings.

## 5.3 Haar wavelets and attractor region

So far, we have only focused on working with wavelets from Daubechies family and more precisely with those wavelets whose number of vanishing moments is $K \geq 2$. The question arises — what is the problem of choosing wavelets with $K = 1$? It turns out that Daubechies wavelets with only one vanishing moment are Haar wavelets. The major problem with Haar wavelets is that, due to their construction, they are not able to transition probability mass across the origin. Figure 5.2 demonstrates the origin crossing problem associated with Haar family.

So, we would like to avoid families of wavelets $\psi_{j,i}(x)$ whose integer shifts do not overlap. In particular, we must pick $\psi_{j,i}(x)$, such that

$$\mathrm{supp}\{\psi_{j,i}(x)\} \cap \mathrm{supp}\{\psi_{j,i+1}(x)\} \neq \emptyset. \tag{5.3.12}$$

If this condition is not satisfied wavelets are not able to transition a probability mass across the origin, in addition to this, *attractor regions* will be created. If wavelets used in the WMC algorithm do not meet the condition (5.3.12) of overlapping supports, then it is guaranteed that *attractor regions* will be created in which points

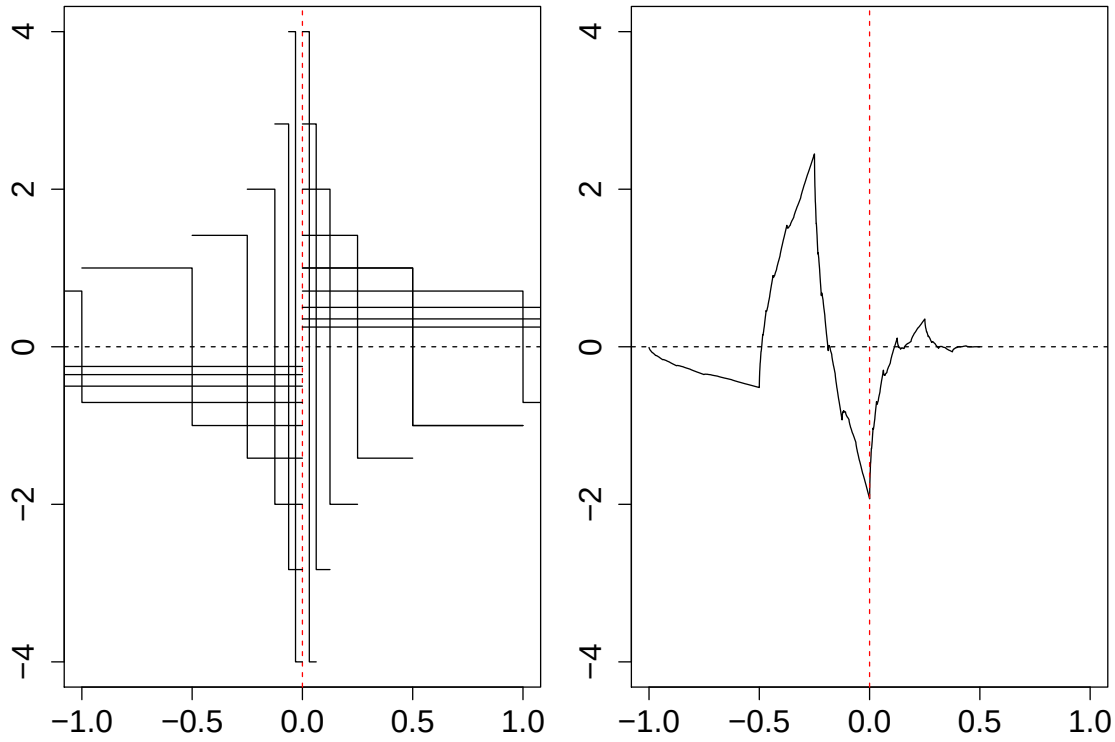



Figure 5.2: *On the left side have been plotted Haar wavelets ranging from resolution levels* $j_{\min} = -4$ *to* $j_{\max} = 4$ *and location* $i \in \{-1, 0\}$. *It could be observed that not a single Haar wavelet plotted contains the origin* $x = 0$ *(red vertical line) inside its support, such that it is not a boundary point of a support region. On the other hand, Daubechies wavelet with* $K = 2$ *on the right at each resolution level contains 3 wavelets that envelope the origin. For demonstration purposes the plotted wavelet is* $\psi_{1,-2}$ *and it clearly contains the origin inside its support, allowing for probability mass transfer across* $x = 0$.

will be stuck during the WMC run and will no longer have any chance of reaching a target.

**Definition 5.3.1.** Let $I_A = (a, b)$ be an interval for some $a, b \in \mathbb{R}$, $a < b$, such that,

$$\sum_{j=j_{\min}}^{j_{\max}} \sum_{i \in \mathbb{Z}} [d_{ji}^{\psi} \psi_{ji}(x)]^{-} = 0, \quad \forall x \in I_A, \tag{5.3.13}$$

then $I_A$ is an attractor region.

**Proposition 5.3.1.** *Let $I_A = (a, b)$ for some $a, b \in \mathbb{R}$ be an attractor region, then $\forall x \in I_A$ the associated survival time is $t = \infty$.*

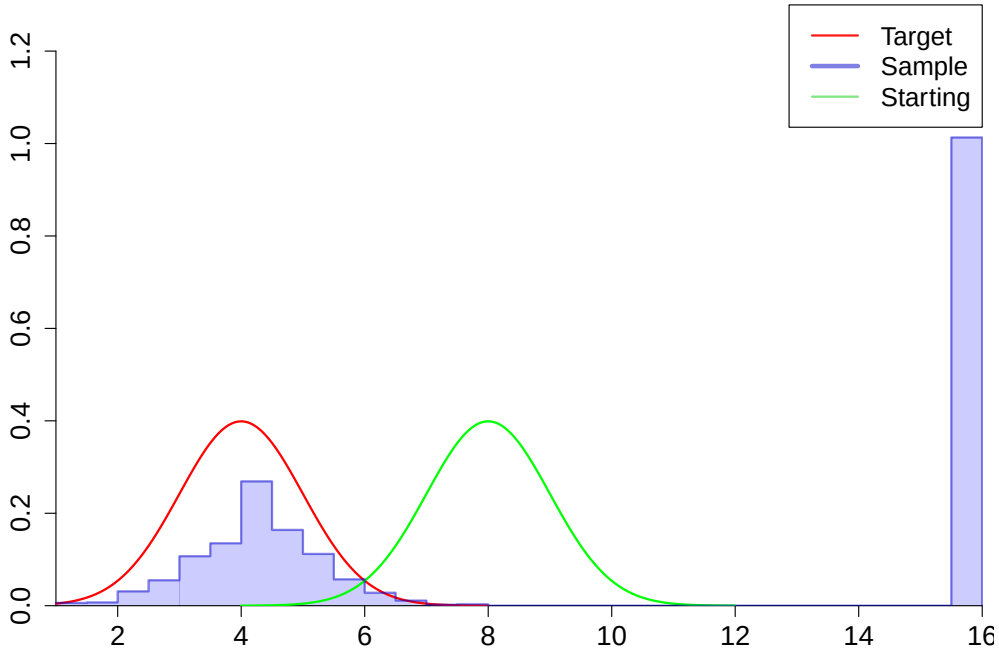



Figure 5.3: *A starting and the target distribution were both chosen to be normal ones with the same variance but different location parameter. Wavelets used in the WMC were set to be Haar with the coarsest resolution level $j_{\min} = -3$ and the finest one $j_{\max} = 8$. As one can notice, the attractor region was formed around the point $x = 16$, which is a support boundary that is being shared across all resolution levels between $j_{\min}$ and $j_{\max}$ for the Haar wavelet.*

From the proposition above it follows that points $x \in I_A$ will never move and automatically will be accepted as samples from the target. The only known wavelet family that exhibits the attractor region phenomena is Haar, Figure 5.3 demonstrates this problem. Given that $I_A$ exists, it is quite clear that condition (5.3.13) leads to an infinite survival time being sampled.

*Proof of Proposition 5.3.1.* For each $x_t$ on which the WMC is being performed, the survival time is going to be sampled via the inverse transform from the Exponential distribution or the Generalized Pareto distribution (3.3.2),

$$t = s - \frac{f(x_s)}{c(x_s)} \log u_s$$

for the exponential case, where $u_s \sim \mathcal{U}(0,1)$ and

$$t = s + \Big(\frac{f(x_s)}{d(x_s)} + s\Big)\Big(u_s^{-d(x_s)/c(x_s)} - 1\Big)$$

for the GPD case, where as before we denote

$$c(x_s) = \sum_{j=j_{\min}}^{j_{\max}} \sum_{i\in\mathbb{Z}} [d_{j,i}^{\psi}\psi_{j,i}(x)]^{-},$$

with the only difference that the sum over resolution levels is restricted, accounting for the limited computational power. If $c(x_s) = 0$, in both cases time becomes infinite $t = \infty$. □

However, a more important part is to prove that attractor regions do exist. Before going into a technical proof, it is not difficult to convince oneself of the existence of these regions by inspecting Figure 5.4. Although, only Haar wavelets are doomed to experience the attractor region problems, in practice one would like to pick a wavelet family with many vanishing moments to avoid the possibility of introducing regions where points are likely to stay much longer than needed.

**Proposition 5.3.2** (Existence of attractors in 1-D)**.** *Let the coarsest and the finest resolution levels be $j_{\min} \in \mathbb{Z}$ and $j_{\max} \in \mathbb{Z}$ respectively, with $j_{\max} > j_{\min}$.*

*(a) Let $H$=$\{\psi_{j,i}(x)\}$ for $j \in \{j_{\min}, ..., j_{\max}\}$ and $i \in \mathbb{Z}$ be a set of Haar wavelets.*

*(b) Let $d(x) \in L^2(\mathbb{R})$ be a difference function, as in (3.1.6), with an infinite support, such that for $|x| > N$, $N \in \mathbb{R}$, it decays monotonically to 0 as $|x| \to \infty$.*

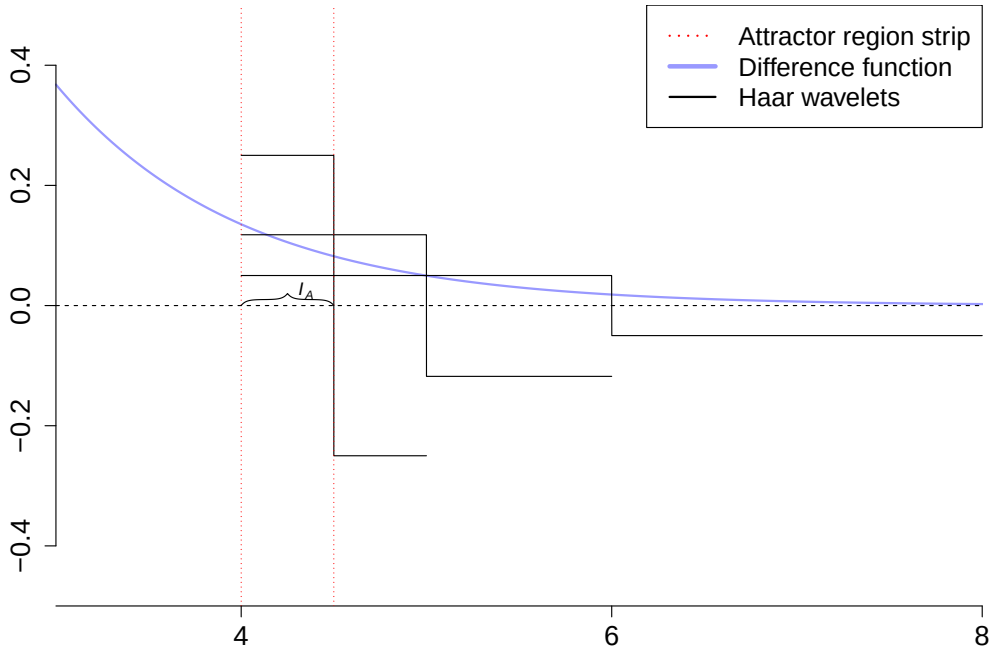



Figure 5.4: *For Haar wavelets there will always be a dyadic point which is going to be shared by exactly one wavelet from each resolution level as a boundary of the wavelet support. In this illustration the restricted range of the resolution levels is* $j \in \{0, -1, -2\}$ *and as we can see point* $x = 4$ *is the common support boundary point for exactly one wavelet from each resolution level. Due to a monotonic decrease in the difference function the associated wavelet coefficients* $d_{ji}^{\psi}$ *with wavelets depicted in this plot will be strictly positive. Furthermore, the attractor region strip is the only region where for all values of* $x$ *in the strip we have* $\psi_{j,i}(x) > 0$*. If we denote the attractor region strip* $I_A$*, then* $\forall x \in I_A$*,* $\sum_{j=-2}^{0} \sum_i [d_{j,i}^{\psi} \psi_{j,i}(x)]^- = 0$*, which would lead to the* $t = \infty$ *survival time associated with all points in the* $I_A$*. Magnitudes of the wavelets in the plot were scaled down for the illustration purposes.*

*If the above conditions (a) and (b) hold, then there exist an infinite number of attractor regions which take the form of* $I_A = (a, b)$*, each with different values of* $a \in \mathbb{R}$ *and* $b \in \mathbb{R}$*.*

*Proof of Proposition 5.3.2.* Let us split the proof into three parts.

1. Given any point $x \in \mathbb{R}$ and the set of Haar family wavelets $\psi_{j,i}$, $j, i \in \mathbb{Z}$, $|j| < \infty$, we have that at each resolution level $j$ there is only a single Haar wavelet $\psi_{j,i}$ for some specific $i$, which envelopes the point $x$, i.e. $\text{supp}(\psi_{j,i}) \ni x$. For each Haar wavelet $\psi_{j,i}$ the support is of the form

$$I_{j,i} = [i2^{-j}, (i+1)2^{-j}).$$

So, for the coarsest resolution level $j_{\min}$, locations of the left boundary of supports are located at points $i2^{-j_{\min}}$. We make a key observation — the left boundary of the support $i2^{-j_{\min}}$ of the wavelet $\psi_{j_{\min},i}$ is going to be shared by exactly one wavelet $\psi_{n,k}$ for each finer resolution level $j_{\min} < n \leq j_{\max}$ and some specific value $k(n) \in \mathbb{Z}$. The value of $k$ is dependent on the resolution level $n$ and here could be seen as a function of $n$. By sharing a support point we mean that if $\text{supp}\big(\psi_{j,i}\big) = [a, b)$ and $\text{supp}\big(\psi_{n,k}\big) = [a, c)$, then wavelets $\psi_{j,i}$ and $\psi_{n,k}$ share a common support boundary $a$.

Let our coarsest wavelet be $\psi_{j_{\min},i_1}$ for some value $i_1$. Then we have $i_1 2^{-j_{\min}}$ for the left boundary of the support. Now, let $j_f > j_{\min}$ be a finer resolution level and so a wavelet $\psi_{j_f,i}$ at the finer resolution $j_f$ has the support of the form $[i2^{-j_f}, (i+1)2^{-j_f})$. For these two wavelets $\psi_{j_{\min},i_1}$ and $\psi_{j_f,i}$ to share left boundary of the support we require

$$i2^{-j_f} = i_1 2^{-j_{\min}} \tag{5.3.14}$$

i.e. $i = i_1 2^{j_f - j_{\min}}$. So, the value of $i$ depends on the resolution levels and the reference location $i_1$. Therefore, for given resolution levels $j_{\min}$ and $j_{\max}$, and the reference location $i_1$, there will always exist a set of wavelets

$$\{\psi_{j_{\min},i_1}, \psi_{j_{\min}+1,i_2}, ..., \psi_{j_{\max},i_{j_{\max}-j_{\min}+1}}\}, \tag{5.3.15}$$

with some specific computable values $\{i_2, ..., i_{j_{\max}-j_{\min}+1}\}$, that depend on the reference location $i_1$, whose supports will satisfy the following inclusion principle,

$$[C, (i_1+1)2^{-j_{\min}}) \supset [C, (i_2+1)2^{-(j_{\min}+1)}) \supset \cdots \supset [C, (i_{j_{\max}-j_{\min}+1}+1)2^{-j_{\max}}). \tag{5.3.16}$$

$C \in \mathbb{R}$ highlights the idea that the left part of the support is always the same, $C \equiv i_1 2^{-j_{\min}}$ in this case.

If $d(x) \in \mathbb{R}$ is a difference function that $\forall |x| > N$, $N \in \mathbb{R}$, decays to 0 monotonically as $|x| \to \infty$, then $\exists M \in \mathbb{R}$, $M > N$, and $i_1 \in \mathbb{Z}$ such that $\max\limits_{|x|>M} \psi_{j_{\min},i_1}(x) > d(x)$. From which follows,

$$\max_x \psi_{j_{\max},i_{j_{\max}-j_{\min}+1}}(x) > \cdots > \max_x \psi_{j_{\min},i_1}(x) > d(x). \tag{5.3.17}$$

By the construction of Haar wavelets, maximum and minimum values of each $\psi_{j,i}(x)$ are attained on intervals $[i2^{-j}, (i+1)2^{-j-1})$ and $[(i+1)2^{-j-1}, (i+1)2^{-j})$ respectively. Given that we have a set of wavelets as described in (5.3.15), and both (5.3.16) and (5.3.17) hold, we can conclude that the part of the support

$$I_A = [i_{j_{\max}-j_{\min}+1}2^{-j_{\max}},\ (i_{j_{\max}-j_{\min}+1}+1)2^{-j_{\max}-1})$$

of the finest resolution wavelet $\psi_{j_{\max},i_{j_{\max}-j_{\min}+1}}$ on which this wavelet takes its maximum value, will also be contained by all coarser wavelets and those coarser wavelets will attain their maximums at this part of their support as well. Hence, we have found explicitly for a given coarsest resolution level and location pair $(j_{\min}, i_1)$, and the finest resolution level $j_{\max}$ a corresponding interval $I_A = (a, b)$, such that if $x \in I_A$, then

$$\psi_{ji}(x) > d(x) > 0, \quad \forall \psi_{j,i}(x) \in \Big\{ \psi_{j,i}(\cdot) \in H \mid \operatorname{supp}\big(\psi_{j,i}(\cdot)\big) \ni x \Big\}. \tag{5.3.18}$$

2. For the second part, let us write down the definition of a mother wavelet coefficient explicitly,

$$d^{\psi}_{j,i} = \int_{x\in\mathbb{R}} d(x)\psi_{j,i}(x)\mathrm{d}x. \tag{5.3.19}$$

If we are working with Haar wavelets, then the Equation (5.3.19) could be written as

$$d^{\psi}_{j,i} = 2^{j/2} \int_{x\in[i2^{-j},(i+1)2^{-j-1})} d(x)\mathrm{d}x - 2^{j/2} \int_{x\in[(i+1)2^{-j-1},(i+1)2^{-j})} d(x)\mathrm{d}x.$$

Given that $(b)$ in the proposition holds, we can set $N = i_1 2^{-j_{\min}}$, for $i_1 \gg 0$ (i.e. when monotonic decay property of $d(x)$ starts to apply), and now $\forall x > N$ it implies that

$$2^{j/2} \int_{x\in[i_1 2^{-j},(i_1+1)2^{-j-1})} d(x)\mathrm{d}x > 2^{j/2} \int_{x\in[(i_1+1)2^{-j-1},(i_1+1)2^{-j})} d(x)\mathrm{d}x, \quad (5.3.20)$$

which in turn implies $d_{ji} > 0$ for all wavelets $\psi_{j,i} \in H$ in the region $x > N$. Therefore, combining the result in part 1 with this observation we conclude that if

$$x \in I_A = \left[i_1 2^{-j_{\min}}, (i_{j_{\max}-j_{\min}+1} + 1)2^{-j_{\max}-1}\right)$$

for some given $j_{\min}, j_{\max}$, and the reference location $i_1$, then $\psi_{ji}(x) > 0$ and $d_{ji} > 0$ for $j \in \{j_{\min}, ..., j_{\max}\}$ and $i \in \mathbb{Z}$. Therefore, $\forall x \in I_A$

$$[\psi_{ji}(x)d_{ji}]^{-} \equiv 0, \forall j \in \{j_{\min}, ..., j_{\max}\}, \forall i \in \mathbb{Z}.$$

Hence,

$$\sum_{j=j_{\min}}^{j_{\max}} \sum_{i\in\mathbb{Z}} [d^{\psi}_{j,i}\psi_{ji}(x)]^{-} = 0, \quad \forall x \in I_A. \quad (5.3.21)$$

3. Finally, given that (5.3.18) and (5.3.21) are proved to hold under the required conditions and for the arbitrary choice of a reference location $i_1$, the generalisation of an infinitely many attractor regions $I_A$ is straightforward. Let the reference location be $i_1 + l$ for $l \in \mathbb{N}$, a new attractor region will be found following steps described above. Hence, this completes the proof of the proposition. □

## 5.4 Ghost points

Compared to points being stuck in attractor regions permanently, we now describe 'ghost points', which could be seen as the complete counter part to attractor regions.

**Definition 5.4.1.** A point $x_g$ is called a *ghost point* if, after being sampled via some wavelet $\psi_{j,i}$, it has an associated survival time $t_g \equiv 0$.

From an inverse sampling algorithm for $d(x_s) = 0$, where $s$ is a current point in time and $t \geq s$ is a survival point in time of $x_s$,

$$t = s - \frac{rf(x_s)}{c(x_s)} \log u_s, \tag{5.4.22}$$

we can clearly see that for $c(x_s) > 0$ and $f(x_s) = 0$ the associated survival time is equal to the previous one $t = s$, indicating that a point $x_s$ has not advanced the process in time at all. So ghost points could be seen as intermediate sample points $x_g$ that were sampled from the zero probability region, $x_g \notin \text{supp}(f) \cup \text{supp}(g)$.

A sampled ghost point $x_g$ exists for a zero amount of time and although it is sampled, the WMC process leaves that point immediately. Clearly, these points are not desired, as computational power is wasted on sampling them in the first place only to find out that this survival time is zero and hence WMC has not advanced in time towards the target sample. The most important question here is, why are these points being sampled in a first place?

This phenomena could be demonstrated using two uniform distributions. In Figure 5.5 intermediate points were sampled in regions of zero probability. This is a very serious issue as it contradicts the claim that all intermediate points $x_t$ come from intermediate distributions $f_t(\cdot)$, $t \in [0, 1]$. The only way that a point sampled from $f(\cdot)$ is moved to the region of $g(\cdot)$ is if chosen wavelet $\psi_{j,i}$ envelopes at least partially supports of a starting and the target distribution. Figure 5.6 shows an

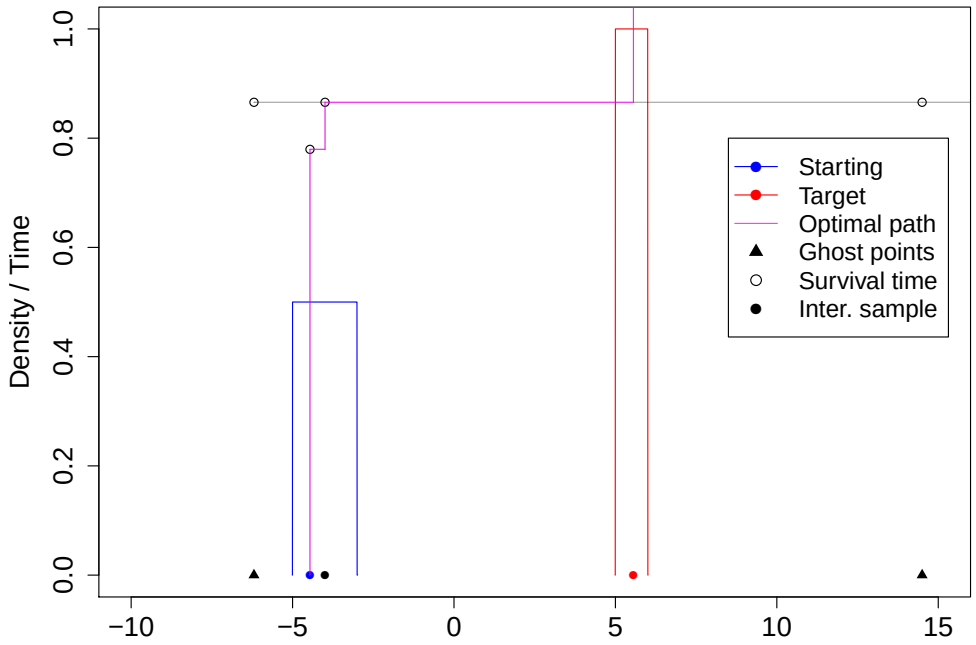



Figure 5.5: *Demonstration of the existence of ghost points using two uniform distribution with* $K = 2$ *and* $N_d = 200$. *Given that PDF and survival time takes values between 0 and 1, the vertical axis corresponds to both. Given that* $f(x_g) = 0$ *and* $g(x_g) = 0$, *the survival time of ghost points is 0.*

example where a chosen wavelet includes a zero density region in its support. This means that there exists a non-zero probability that a sampled intermediate point will fall in the zero density region of $f_t(\cdot)$. This is exactly what happens in practice, leading to many points being sampled from regions of zero density. As intermediate points $x_t$ generated by the WMC process do not necessarily come from a distribution with density $f_t(\cdot)$, Theorem 3.3.2 is put into question, requiring one to update and reformulate assumptions of Proof 3.3.

A simulation was performed using two uniform distributions identical to as in Figure 5.5. The idea was to produce 100 samples from the target distribution $\mathcal{U}[5, 6]$ using samples from $\mathcal{U}[-5, -3]$ and to monitor how many intermediate points were produced from zero density regions, i.e. how many points were generated that did

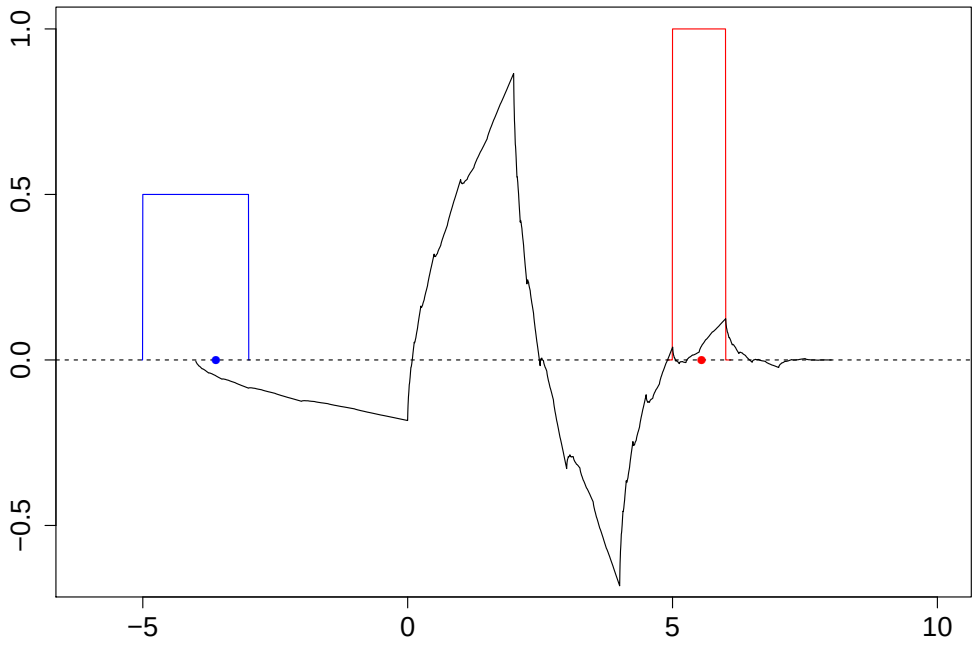



Figure 5.6: *Daubechies $K = 2$ transition wavelet $\psi_{-2,-1}$ partially envelopes $f(\cdot)$ and fully envelope $g(\cdot)$, however it also covers the zero density region in between. The selection of such a wavelet would potentially lead to points being sampled from the zero-density region.*

not follow $f_t(\cdot)$ at a given particular time. Daubechies wavelets with $K = 2$ and $N_d = 200$ were used. Out of 910 sample points generated, 772 were sampled from the zero-density region. That means that around 84% of the computing power was wasted on points that should not have been generated in the first place. On average, there were 7.72 ghost points generated to produce one sample from the target $g(\cdot)$.

In a situation when the choice of a starting and a target distribution creates zero density regions, points that fall into them are immediately classified as ghost points. However, the situation is less clear when the starting and target densities have infinite support but there are regions of near zero-density. For instance, if we choose the starting distribution to be $\mathcal{N}(-5, 1)$ and the target to be $\mathcal{N}(5, 1)$, then all intermediate densities $f_t(\cdot)$ formed will have a near zero-density region in between

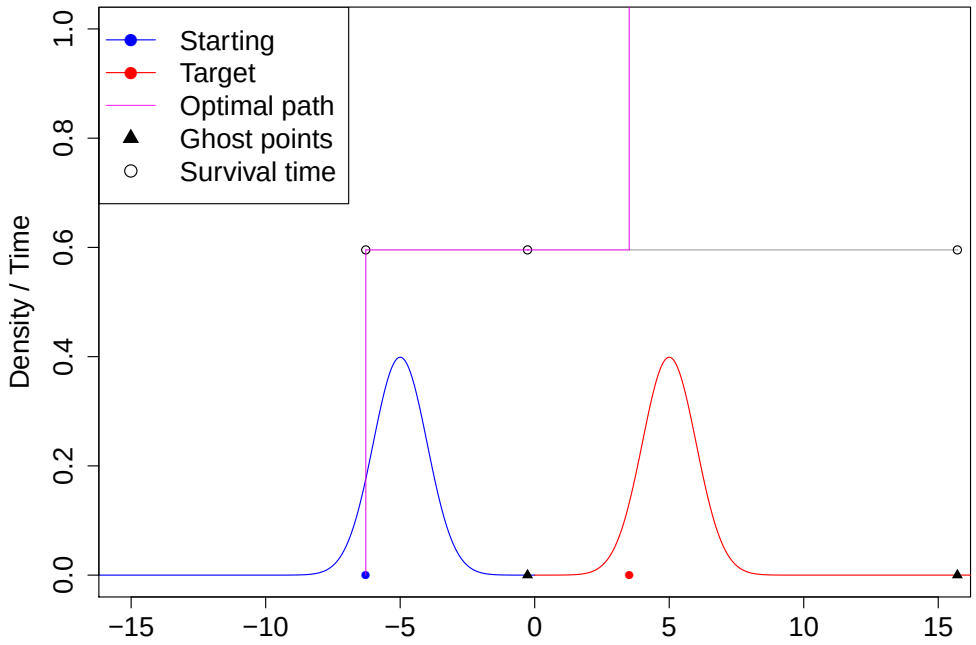



Figure 5.7: *Example of semi-ghost points being generated in a situation when the support is infinite, supp*$(f_t) = \mathbb{R}$*, and there are regions of very low density.*

$x = -5$ and $x = 5$ and as $|x| \to \infty$ regions. Although intermediate points sampled via WMC do belong to the support of the density $f_t(\cdot)$, it should be quite unlikely that points are sampled from regions of low density. Unfortunately, for the same exact reason as before, points are being generated from low density regions and Figure 5.7 illustrates this problem quite clearly. Instead of taking an optimal path, semi-ghost points were generated leading to an inefficient algorithm.

**Definition 5.4.2.** A point $x_{sg}$ is called a *semi-ghost point* if after being sampled via some wavelet $\psi_{j,i}$ it has an associated survival time $0 < t_{sg} \ll 1$.

The examples demonstrated in Figure 5.5 and 5.7 use Daubechies wavelets with $K = 2$, $j_{\min} = -8$ and $j_{\max} = 11$, which is more than enough to cover all the details of the difference function. One might speculate that the ghost point phenomena could be tied to the finite computing power nature and imprecise implementation,

however it is not the case. For example, what if wavelet coefficients $d_{j,i}$ could be computed exactly and one had an access to levels $j_{\min} = -\infty$ and $j_{\max} = +\infty$? Even in this perfect scenario, there would always exist a positive probability $q_{-2,-1} > 0$ for the wavelet $\psi_{-2,-1}$ to be selected as in Figure 5.6, therefore there would always exist a possibility that an intermediate point would be sampled from a region that does not belong to the support of the intermediate density $f_t(\cdot)$.

## 5.5 Outliers

In the WMC setting, we call a point an *outlier* if it was falsely assigned a survival time $t = \infty$. Previously, we have discussed attractor regions, places where points could be stuck forever simply due to WMC having a finite range of resolution levels and using Haar wavelets. The only way that a point $x$ could be assigned a survival time of $t = \infty$ is if

$$\sum_{j=j_{\min}}^{j_{\max}} \sum_{i \in \mathbb{Z}} [d^{\psi}_{j,i} \psi_{ji}(x)]^{-} = 0, \tag{5.5.23}$$

as discussed previously. However, in practice, we work with estimates of wavelet coefficients $\hat{d}^{\psi}_{j,i}$. This means that one bad estimate of $d^{\psi}_{j,i}$ could determine whether a point will be assigned a survival time $t = \infty$ or not. If for some $j, i \in \mathbb{Z}$, $[d^{\psi}_{j,i} \psi_{ji}(x)]^{-} > 0$ with $d^{\psi}_{j,i} < 0$ and $\psi_{ji}(x) > 0$, then a bad estimate with the opposite sign $\hat{d}^{\psi}_{j,i} > 0$ would make $[\hat{d}^{\psi}_{j,i} \psi_{ji}(x)]^{-} = 0$, which potentially could lead to the total sum (5.5.23) being equal to 0. The idea to use estimates $\hat{d}^{\psi}_{j,i}$ instead of true values was to avoid complex integrals involved in computation of $d^{\psi}_{j,i}$. Unfortunately, by doing so we introduce randomness in the estimates of wavelet coefficients that could lead to incorrect samples being produced via WMC. The only possible solution to get rid of outliers is to increase the value of $N_d$ that is responsible for how many values are sampled from a positive $\psi^{+}_{j,i}$ and negative part $\psi^{-}_{j,i}$ of the wavelet in computation of the estimate $\hat{d}^{\psi}_{j,i}$. However, as discussed in Section 4.4, the execution time is highly

dependent on the choice of $N_d$ and large values slow down WMC significantly.

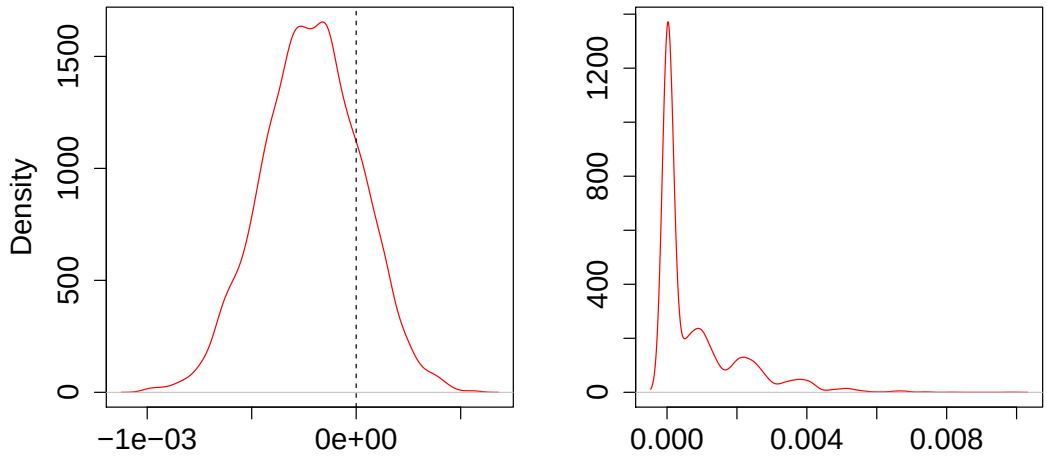



Figure 5.8: *Left: kernel density (KD) estimate of the distribution associated with estimate of Daubechies* $K = 2$ *wavelet coefficient* $\hat{d}^{\psi}_{j,i}$ *for* $j = 2$, $i = 1$, *being sampled with* $N_d = 50$. *KD estimate was based on 5000 realisations of* $\hat{d}^{\psi}_{j,i}$. *The difference function* $d(\cdot)$ *was constructed using* $f(\cdot)$ *and* $g(\cdot)$ *from the example in §4.1.1. The distribution resembles a normal with* $\mu = -0.0002$ *and* $\sigma = 0.0002$, *which includes both positive and negative values of* $\hat{d}^{\psi}_{j,i}$. *Right: KD estimate of* $\hat{c}(x) = \sum_{j=j_{\min}}^{j_{\max}} \sum_{i \in \mathbb{Z}} [\hat{d}^{\psi}_{j,i} \psi_{ji}(x)]^{-}$ *for* $x = -256$ *and* $j_{\max} = 11$, $j_{\min} = -8$. *The density is concentrated around value 0 meaning that most of the time point* $x = -256$ *would be assigned* $t = \infty$. *Similarly KD estimate of* $\hat{c}(x)$ *was based on 5000 realisations.*

In Figure 5.8 we can examine the consequences of using $\hat{d}^{\psi}_{j,i}$ estimates instead of true values. As we go far away from the high density regions of a starting and a target density, wavelets that cover high density regions of both $f(\cdot)$ and $g(\cdot)$ become extremely stretched and require many more samples $N_d$ to accurately estimate the associated wavelet coefficients. In Figure 5.8, kernel density estimation was used to approximate the density function of the $\hat{c}(x = -256)$ estimate, where as before for

we use

$$\hat{c}(x) = \sum_{j=j_{\min}}^{j_{\max}} \sum_{i \in \mathbb{Z}} [\hat{d}_{j,i}^{\psi} \psi_{ji}(x)]^{-},$$

but instead of $c(x)$ we have $\hat{c}(x)$, which highlights the fact that we are working with estimates of the wavelet coefficients and our resolution range is finite. It is quite clear that there is a positive probability that the point $x = -256$ will be accepted as a sample from the target distribution even though it is far away from the high density region and has practically zero probability of being a realistic sample from the target $g(\cdot)$.

To see the effect of using estimates of wavelet coefficients, we can investigate how the value of $\hat{c}(x)$ changes around high and low density regions of $f_t(\cdot)$ (Figure 5.9 and 5.10), where $f_t(\cdot)$ is as in the 1-D example from §4.1.1.

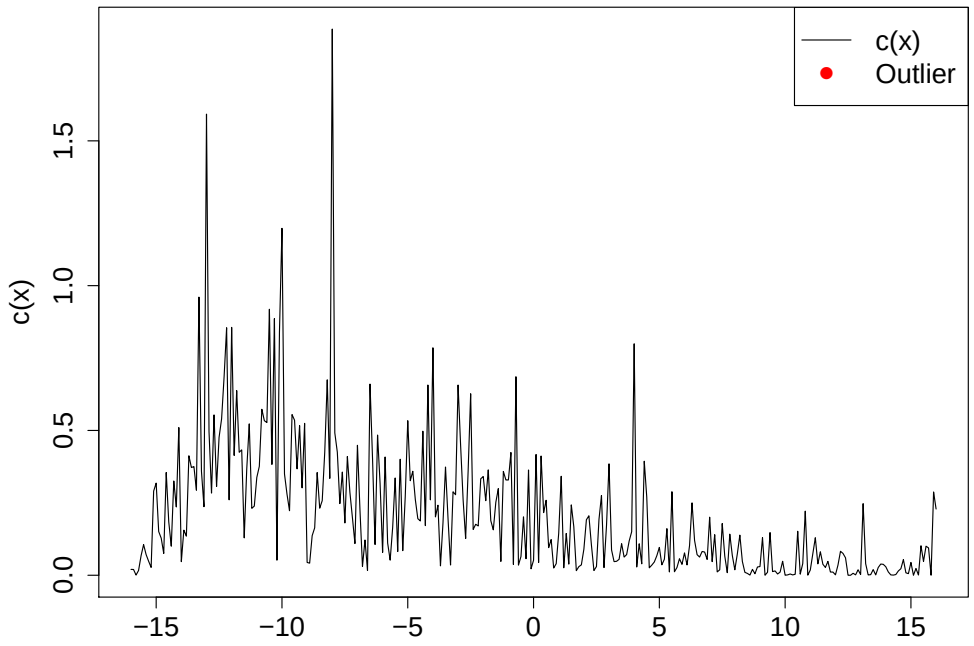



Figure 5.9: *The sampled version $\hat{c}(x)$, computed around the high density region $x \in (-16, 16)$, using $N_d = 1$. Even when estimating $\hat{d}_{j,i}^{\psi}$ by using a single value from the positive and negative part of the wavelets, not a single outlier was detected.*

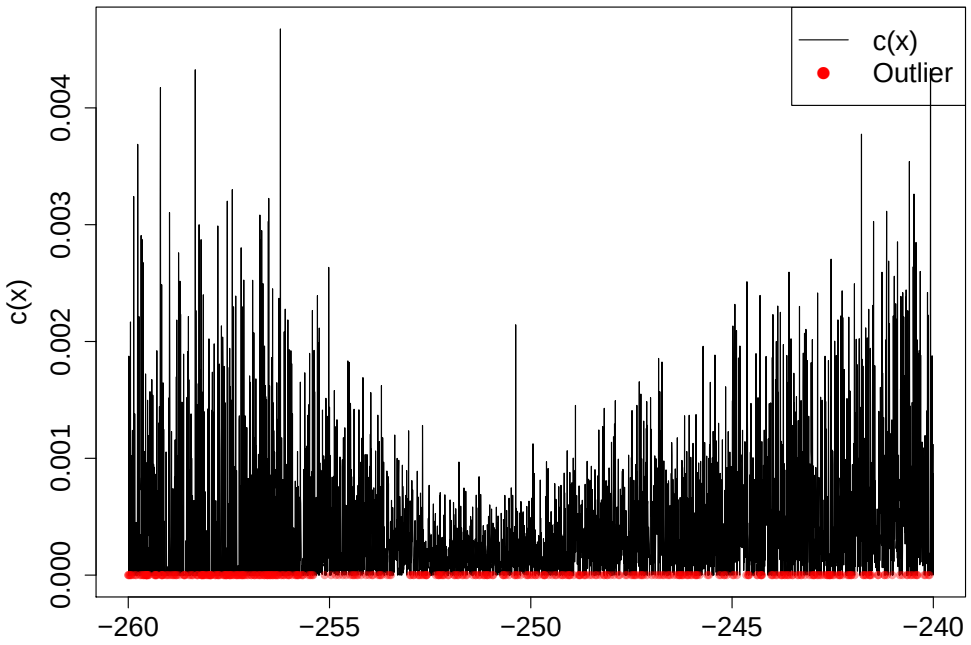



Figure 5.10: *The sampled version $\hat{c}(x)$, computed around the low density region $x \in (-260, -240)$, using $N_d = 100$. As one can see, there are many values for which $c(x) = 0$, which would lead to outliers being produced. The only way to avoid this is to lower the coarsest and increase the finest resolution levels in addition to boosting the value of $N_d$. For regions far away from the target, accurate computation needs to be performed to get good quality estimates of wavelet coefficients. In this case, raising value to $N_d = 100$ has not helped at all, which in high density region would be more than enough.*

Outlier points are purely a consequence of using $\hat{d}^{\psi}_{j,i}$ instead of true values $d^{\psi}_{j,i}$. Introduced variance around the estimate does not always guarantee that

$$\operatorname{sgn}(d^{\psi}_{j,i}) = \operatorname{sgn}(\hat{d}^{\psi}_{j,i}),$$

where

$$\operatorname{sgn}(x) = \begin{cases} 1 & , x > 0 \\ 0 & , x = 0 \\ -1 & , x < 0 \end{cases}$$

This leads to situations where for certain values of $x$, $\hat{c}(x) = 0$, which in turn implies an infinite survival time for the point. Unfortunately, this issue could only be suppressed but not tackled completely. The only possible solution is to increase the value of $N_d$ for very coarse wavelets, in a sense to make $N_d$ adaptive to the resolution level being used. Essentially, $N_d(j) \in \mathbb{N}$ becomes a function of a resolution level $j$. This sort of set up requires separate analysis and potentially will be investigated in the future.

## 5.6 Summary

In conclusion, there are a couple of key issues surrounding this topic that could be dealt with and there are others that at this point are bound to the theory of WMC and require further theoretical development.

In particular, the number of outlier points could be minimised or even potentially reduced down to zero if careful analysis of $N_d$ reveals an appropriate method for how $N_d$ should be controlled with respect to resolution levels. Ideally, making $N_d(j)$ as a function of the current resolution level should solve the outlier problem, as more care would be given for coarse wavelet coefficients.

The problem of attractor regions is tackled by not choosing the Haar wavelet. Although attractor regions were proved to exist only for the choice of Haar wavelet family, one might speculate the existence of these regions for wavelets with $K > 1$. So far, simulation has not revealed any clues for the existence of $I_A$ when wavelets are much smoother and their supports overlap. In addition to this, given the complex theoretical nature of Daubechies wavelets, the proof for the existence/non-existence of attractor regions for $K > 1$ seems to be intractable.

Furthermore, ghost points seem to be tied to the theory of the WMC itself and

eradication of these points requires a theoretical fix. The good point is that a solution to this issue could potentially decrease the execution time of the algorithm substantially, as demonstrated with the example of two uniform distributions.

Finally, the ratio of normalising constants $r$ and the curse of dimensionality are likely to be the two major problems that need to be addressed directly before trying to solve the outlier and ghost point issues. As was shown, the normalisation ratio $r$ stands as a separate difficult problem that needs to be tackled before the execution of WMC. Therefore, from this perspective WMC looks unattractive compared to other sampling methods, for example MCMC family methods that completely ignore any type of integrals. Lastly, the curse of dimensionality also is an intrinsic issue for WMC. As the dimension grows, the total number of coefficients $\hat{d}^{\psi}_{j,i}$ required for the algorithm grows geometrically. Therefore, at this stage of the development of WMC theory, the algorithm is not able to efficiently tackle problems of a dimensionality $d > 2$.

# Chapter 6

# Probability distribution of jumps

## 6.1 Motivation

In this chapter, we will focus on investigating the probability distribution of the total number of jumps performed in a single WMC run. When a survival time $t < 1$ is sampled for any point $x_s$, a new point needs to be sampled to replace the old point. The sampling of a new point represents a jump from a previous point. The total number of jumps performed in WMC is related to the efficiency of the algorithm as was discussed briefly in Section 4.4. It is important to be able to analyse the probability distribution properties of the total number of jumps in WMC given a starting distribution $f(\cdot)$ and some target $g(\cdot)$. Ideally, information on the average number of jumps and the variance could be used as a tuning parameter for the choice of a starting distribution, wavelet family and resolution levels.

We will investigate jump distribution in steps. Firstly, we will focus on a no-jump probability, by asking a question — given a point $x_s$, at time $s$, what is the probability that no jumps will be performed and that point will be accepted right away? Secondly, we will focus on a one jump probability and will try to generalise

results to the $n$-jump case. Before going into the analysis, we brief the reader with the most relevant notation that will be used throughout this chapter.

## 6.2 Notation and set-up

At first we will be interested in a probability distribution $p(J = n|x_s, s)$, where $n \in \mathbb{N}_0$ is the number of jumps required to reach a target sample $y \sim g(\cdot)$, given we are at the point $x_s$ at a time $0 \leq s < 1$. In particular, we would like know the expectation of the total number of jumps to the target from the point $x_s$ at time $s$,

$$\mathbb{E}[J|x_s, s] = \sum_{n=0}^{\infty} np(J = k|x_s, s). \tag{6.2.1}$$

Here we will recap on the notation that will be used extensively in this chapter, the list could be used to assist a reader following derivations in the next section:

- $d(x) := g(x) - f(x)$: a difference function with $r = 1$,
- $c(x) = \sum_{ji}[d_{ji}^{\psi}\psi_{ji}(x)]^{-}$: as defined in (3.3.19),
- $\rho(x) := c(x)/d(x)$: we define new function $\rho(\cdot)$ to simplify notation,
- $f_s(t|x_s)$: the survival time density at a time $s$ for a point $x_s$ as in (3.3.25),
- $F_s(t|x_s)$: the CDF of $f_s(t|x_s)$, identical to (3.3.24).

## 6.3 Probability of zero jumps

Let the PDF and CDF of the survival time density for a point $x_s$ at a time $s$ be denoted $f_s(t|x_s)$ and $F_s(t|x_s)$ respectively. In particular, they take form of a

*Generalised Pareto Distribution*, first introduced in Section 3.3.1,

$$f_s(t|x_s) = \frac{c(x_s)}{f(x_s)+td(x_s)}\left(\frac{f(x_s)+sd(x_s)}{f(x_s)+td(x_s)}\right)^{\rho(x_s)-1}, \quad t \in [s,\infty),$$

and

$$F_s(t|x_s) = 1 - \left(\frac{f(x_s)+sd(x_s)}{f(x_s)+td(x_s)}\right)^{\rho(x_s)}, \quad t \in [s,\infty).$$

Then, the probability that we are interested in is

$$1 - p(\text{survival time for the point } x_s \text{ is } t \text{ s.t. } s \leq t \leq 1|x_s, s),$$

which is,

$$\begin{aligned} p(J=0|x_s,s) &= 1 - F_s(t=1|x_s) \\ &= \left(\frac{f(x_s)+sd(x_s)}{f(x_s)+d(x_s)}\right)^{\rho(x_s)} \\ &= \left(\frac{f(x_s)+sd(x_s)}{g(x_s)}\right)^{\rho(x_s)}. \end{aligned} \tag{6.3.2}$$

Now that we have a functional form for $p(J=0|x_s,s)$, we can investigate it in more detail. This probability approaches 1 as $s \to 1$,

$$\lim_{s\to 1} p(J=0|x_s,s) = 1.$$

Given that we are at a starting point $x_0 \sim f(\cdot)$, we know that the probability for this point to perform zero jumps and be accepted as a sample from the target $g(\cdot)$ is

$$p(J=0|x_0,s=0) = \left(\frac{f(x_0)}{g(x_0)}\right)^{\rho(x_0)} \tag{6.3.3}$$

and from the expression of $p(J=0|x_0,s=0)$ we can clearly see that probability of making zero jumps approaches 1 as $f(x)$ becomes more similar to $g(x)$. Having chosen the starting distribution $f(\cdot)$, the exponent $\rho(x)$ remains the only object that could influence the value of the probability as the value of $\rho(x)$ depends on the choice of the wavelet family. After exploring Figure 6.1, where Daubechies wavelets with

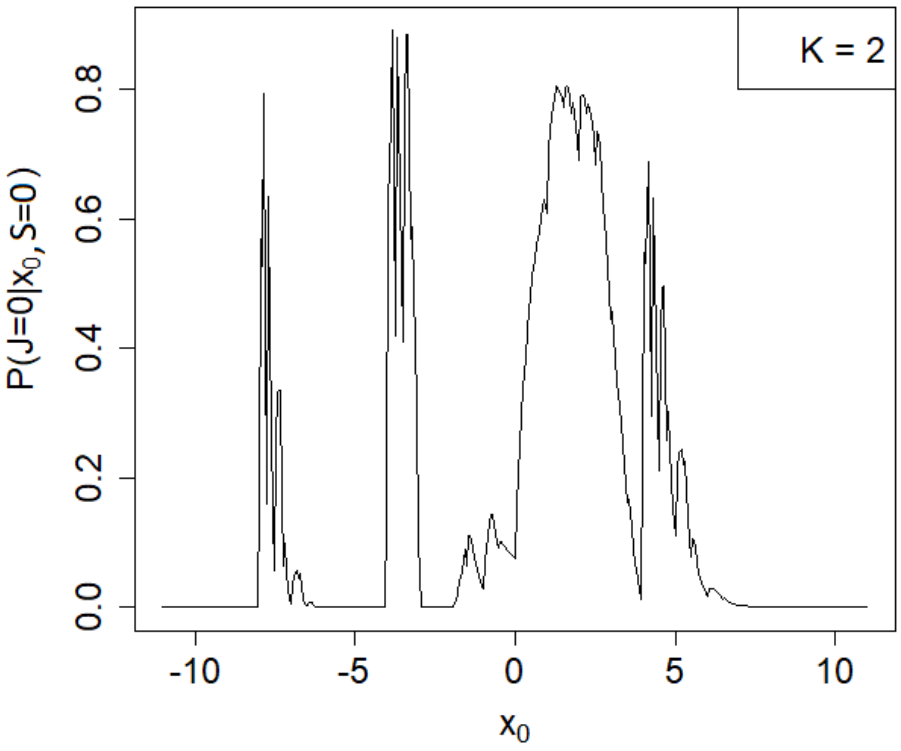



Figure 6.1: *Plot of* $p(J = 0|x_0, s = 0)$ *for the starting distribution* $\mathcal{U}(-10, 10)$ *and the target same as in equation (4.1.1), using Daubechies wavelets with* $K = 2$.

$K = 2$ vanishing moments were used, we are interested in comparing how significant the difference is between $p_{K=6}(J = 0|x_0, s = 0)$ and $p_{K=2}(J = 0|x_0, s = 0)$, where by $K = 6$ and $K = 2$ subscripts we are referring to the Daubechies wavelet family used. From Figure 6.2, it is quite clear that in this scenario the choice of the Daubechies wavelet family does not impact the zero-jump probability significantly, where the maximum difference is approximately around $\pm 1\%$. Given the dependence of $p(J = 0|x_0)$ on both $f(\cdot)$ and $g(\cdot)$, a more theoretical investigation of jump probabilities becomes highly dependent on the assumptions of the families of both distributions, for this reason we will try to simplify the problem as much as possible to begin with.

We investigate jump probabilities when the target distribution $g(\cdot)$ is close to the

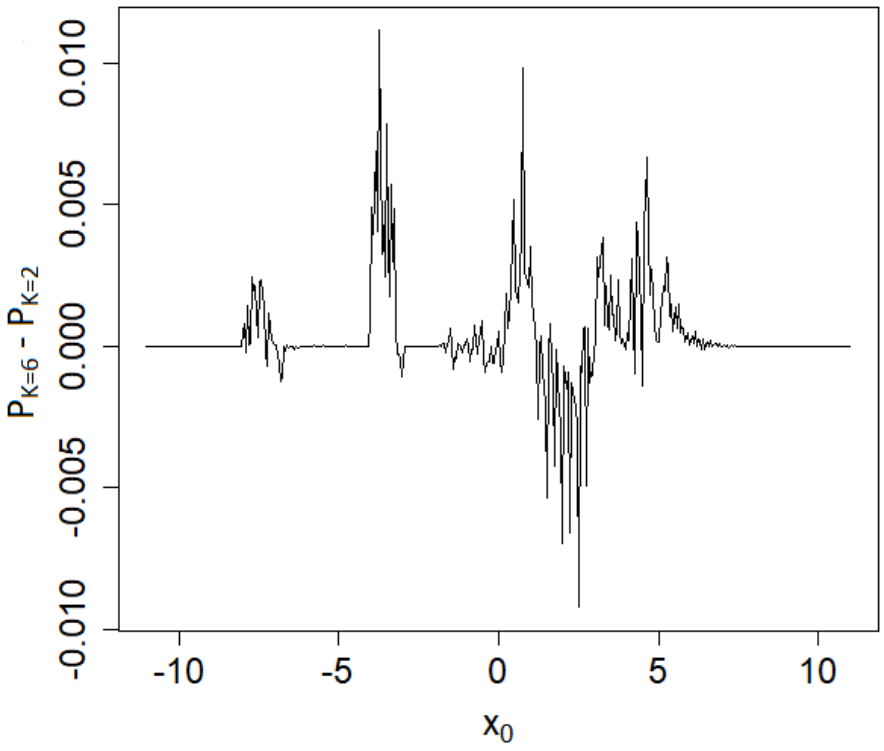



Figure 6.2: *Comparison of the zero jump probabilities for a starting point* $x_0 \sim f(\cdot)$ *between* $K = 6$ *and* $K = 2$ *Daubechies wavelets. Plotted is difference* $p_{K=2}(J = 0|x_0, s = 0) - p_{K=6}(J = 0|x_0, s = 0)$.

starting distribution $f(\cdot)$. For this reason, we define our target density to be

$$g(x) = f(x) + \delta h(x), \quad 0 < \delta \ll 1, \tag{6.3.4}$$

where we have assumed $r = 1$ for simplicity and where $h(x)$ is some reasonably well behaved function in $L^2(\mathbb{R})$, with

$$\int_{-\infty}^{+\infty} h(x)\,\mathrm{d}x = 0, \tag{6.3.5}$$

such that,

$$\int_{-\infty}^{+\infty} g(x)\,\mathrm{d}x = 1, \quad g(x) \geq 0,\ \forall x \in \mathbb{R}. \tag{6.3.6}$$

So, using (6.3.4), we have

$$d(x) = \delta h(x).$$

With $\delta$ parameter in (6.3.4), we can control how close the target is to the starting distribution (and for $\delta = 0$ we recover our starting distribution). Therefore, we are interested in investigating the behaviour of $p(J = 0|x_0, s = 0)$ for small values of $\delta$.

We note that, if $g(x)$ is of the form (6.3.4), then $\rho(x)$ is independent of $\delta$,

$$\rho(x) = \frac{\sum_{j,i}[\delta h^{\psi}_{j,i}\psi_{j,i}(x)]^-}{\delta h(x)} = \frac{\sum_{j,i}[h^{\psi}_{j,i}\psi_{j,i}(x)]^-}{h(x)}.$$

The first-order Taylor expansion of equation (6.3.2) at $\delta = 0$ is,

$$\begin{aligned}
&\left(\frac{f(x) + s\delta h(x)}{f(x) + \delta h(x)}\right)^{\rho(x)} \\
&= 1 + \delta\left[\rho(x)\left(\frac{f(x) + s\delta h(x)}{f(x) + \delta h(x)}\right)^{\rho(x)-1}\frac{sh(x)(f(x) + \delta h(x)) - h(x)(f(x) + s\delta h(x))}{(f(x) + \delta h(x))^2}\right]_{\delta=0} \\
&\qquad + \mathcal{O}(\delta^2), \\
&= 1 + \delta\left(\rho(x)\frac{sh(x)f(x) - h(x)f(x)}{f(x)^2}\right) + \mathcal{O}(\delta^2), \\
&= 1 + \delta(s-1)\frac{\rho(x)h(x)}{f(x)} + \mathcal{O}(\delta^2),
\end{aligned}$$

and from the definition of $\rho(x)$ we obtain,

$$= 1 + \delta(s-1)\frac{\sum_{j,i}[h^{\psi}_{j,i}\psi_{j,i}(x)]^-}{f(x)} + \mathcal{O}(\delta^2).$$

Substituting this Taylor expansion into (6.3.2), we obtain,

$$p(J = 0|x_s) = 1 + \delta(s-1)\frac{\sum_{j,i}[h^{\psi}_{j,i}\psi_{j,i}(x_s)]^-}{f(x_s)} + \mathcal{O}(\delta^2). \tag{6.3.7}$$

We also note that the second term in (6.3.7) is closely related to the underlying assumption A2 of the pWMC method on p.43,

$$\text{A2}: \quad \sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}[d^{\psi}_{j,i}\psi_{j,i}(x)]^- \leq rf(x) \quad \forall x \in \mathbb{R}. \tag{6.3.8}$$

For convenience, we define

$$\sigma(x) := \sum_{j,i}[h^{\psi}_{j,i}\psi_{j,i}(x)]^-. \tag{6.3.9}$$

Then, using (6.3.9), expansion (6.3.7) becomes

$$p(J = 0|x_s, s) = 1 + \delta(s - 1)\frac{\sigma(x_s)}{f(x_s)} + \mathcal{O}(\delta^2). \tag{6.3.10}$$

As expected, the probability of zero jumps in WMC approaches 1 as $\delta \to 0$. Furthermore, as time parameter $s \to 1$, this probability also approaches 1.

## 6.4 Probability of one jump

To start the analysis of the probability of one jump given we are at $x_{s_0}$ at time $s_0$, we will derive the first-order Taylor expansion of the joint probability $p(J = 1, x_{s_1}|x_{s_0}, s_0)$, where $x_{s_1}$ denotes the point to which we jump next at time $s_1$. Based on the expression (6.3.10) for $p(J = 0|x_s, s)$, we would expect the first leading term to be proportional to

$$1 - p(J = 0|x_{s_0}, s_0) \approx \delta(1 - s_0)\frac{\sigma(x_{s_0})}{f(x_{s_0})}.$$

The joint probability that we are interested in is

$$p(J = 1, x_{s_1}|x_{s_0}, s_0) = \int_{s_0}^{1} f_{s_0}(s_1|x_{s_0})\{1 - F_{s_1}(t = 1|x_{s_1})\}\,\mathrm{d}s_1, \tag{6.4.11}$$

where $f_{s_0}(s_1|x_{s_0})$ and $F_{s_1}(t = 1|x_{s_1})$ are as described in recap on notation in Section 6.2. The integral (6.4.11) comes from the observation that we sample a survival time $s_1$, $s_0 \leq s_1 < 1$, for point $x_{s_0}$ and then point $x_{(s_1)}$ survives past $t = 1$. Given that $s_0 \leq s_1 < 1$, we need to integrate $s_1$ to get the probability $p(J = 1, x_{s_1}|x_{s_0}, s_0)$.

It is clear, that for general $f(\cdot)$ and $g(\cdot)$, computations of the integral above become intractable. Therefore, this motivates using a similar approach to simplification as in (6.3.4).

We will first derive the Taylor expansion of the integrand of (6.4.11) and then will focus on its integration. In the previous section we derived the Taylor expansion of

$1 - F_s(t = 1|x_s)$ in (6.3.10), so we now need only to expand $f_{s_0}(s_1|x_{s_0})$ up to the first order:

$$f_{s_0}(s_1|x_{s_0}) = \left(\frac{c(x_{s_0})}{f(x_{s_0}) + td(x_{s_0})}\right)\left(\frac{f(x_{s_0}) + s_1 d(x_{s_0})}{f(x_{s_0}) + td(x_{s_0})}\right)^{c(x_{s_0})/d(x_{s_0})-1}$$

after substituting $c(x_{s_0}) = \delta\sigma(x_{s_0})$ and $d(x_{s_0}) = \delta h(x_{s_0})$ we obtain

$$= \left(\frac{\delta\sigma(x_{s_0})}{f(x_{s_0}) + t\delta h(x_{s_0})}\right)\left(\frac{f(x_{s_0}) + s_1\delta h(x_{s_0})}{f(x_{s_0}) + t\delta h(x_{s_0})}\right)^{\rho(x_{s_0})-1}$$

expanding the first and the second part of the product separetely we get

$$= \left(\delta\frac{\sigma(x_{s_0})}{f(x_{s_0})} + \mathcal{O}(\delta^2)\right)\left(1 - \delta s_1\frac{\sigma(x_{s_0})}{f(x_{s_0})} + \mathcal{O}(\delta^2)\right).$$

We are only interested in the first leading term, therefore after multiplying terms in two brackets above we end up with

$$f_{s_0}(s_1|x_{s_0}) = \delta\frac{\sigma(x_{s_0})}{f(x_{s_0})} + \mathcal{O}(\delta^2). \tag{6.4.12}$$

Now taking a product of (6.3.10) and (6.4.12), and plugging values into the integral (6.4.11), we get the Taylor expansion that we were aiming for:

$$p(J = 1, x_{s_1}|x_{s_0}, s_0) = \delta\frac{\sigma(x_{s_0})}{f(x_{s_0})}\int_{s_0}^{1} \mathrm{d}s_1 + \mathcal{O}(\delta^2).$$

After evaluating the integral we obtain the final form

$$p(J = 1, x_{s_1}|x_{s_0}) = \delta(1 - s_0)\frac{\sigma(x_{s_0})}{f(x_{s_0})} + \mathcal{O}(\delta^2), \tag{6.4.13}$$

as predicted at the start of this section. Given the absence of $x_{s_1}$ in the leading term of the expression (6.4.13), we can also conclude that from (6.4.11)

$$p(J = 1|x_{s_0}, s_0) = \int_{-\infty}^{+\infty} p(J = 1, x_{s_1}|x_{s_0}) f_{s_1}(x_{s_1})\,\mathrm{d}x_{s_1} = \delta(1 - s_0)\frac{\sigma(x_{s_0})}{f(x_{s_0})} + \mathcal{O}(\delta^2), \tag{6.4.14}$$

where $f_{s_1}(x_{s_1})$ is the PDF for the point $x_{s_1}$ and we used it to integrate this intermediate point out.

## 6.5 Generalised probability of $n$ jumps

### 6.5.1 Probability of 2 and $n$ jumps

To investigate probability of 2 jumps, we extend expression (6.4.11), to obtain

$$p(J=2, x_{s_1}, x_{s_2}|x_{s_0}, s_0) = \int_{s_0}^{1}\int_{s_1}^{1} f_{s_0}(s_1|x_{s_0}) f_{s_1}(s_2|x_{s_1})\{1 - F_{s_2}(t=1|x_{s_2})\}\,\mathrm{d}s_2\,\mathrm{d}s_1. \tag{6.5.15}$$

Where the logic behind the integral is the same as before, except we have an additional jump to point $x_{s_2}$ which introduces an additional integral for the time point $s_2$. Although integral becomes more complex, the repeating pattern of the integral allows us to solve this problem rather easily. Similarly as before for (6.4.11), after performing first-order Taylor expansion we multiply all necessary integrand parts and obtain

$$p(J=2, x_{s_1}, x_{s_2}|x_{s_0}, s_0) = \delta^2 \frac{\sigma(x_{s_0})}{f(x_{s_0})}\frac{\sigma(x_{s_1})}{f(x_{s_1})}\int_{s_0}^{1}\int_{s_1}^{1}\mathrm{d}s_2\,\mathrm{d}s_1 + \mathcal{O}(\delta^3). \tag{6.5.16}$$

We can observe that expression (6.5.16) could be generalised quite straightforwardly to the $n$-case,

$$p(J=n, x_{s_1}, ..., x_{s_n}|x_{s_0}, s_0) = \delta^n \prod_{i=0}^{n-1}\frac{\sigma(x_{s_i})}{f(x_{s_i})}\int_{s_0}^{1}\int_{s_1}^{1}...\int_{s_{n-1}}^{1}\mathrm{d}s_n...\,\mathrm{d}s_2\,\mathrm{d}s_1 + \mathcal{O}(\delta^{n+1}). \tag{6.5.17}$$

To finalise general formula (6.5.17) we require to solve the integral

$$\int_{s_0}^{1}\int_{s_1}^{1}...\int_{s_{n-1}}^{1}\mathrm{d}s_n...\,\mathrm{d}s_2\,\mathrm{d}s_1. \tag{6.5.18}$$

Given the apparent symmetry of the integral (6.5.18), we will first focus on working out the value of

$$I_{J=n}(s_0) = \int_{s_0}^{1}...\int_{s_{n-1}}^{1}\mathrm{d}s_n...\,\mathrm{d}s_1,$$

for the first few values of $n$:

$$
\begin{aligned}
I_{J=1}(s_0) &= \int_{s_0}^{1} \mathrm{d}s_1 = 1 - s_0, \\
I_{J=2}(s_0) &= \int_{s_0}^{1}\int_{s_1}^{1} \mathrm{d}s_2\,\mathrm{d}s_1 = \frac{1}{2} - s_0 + \frac{s_0^2}{2}, \\
I_{J=3}(s_0) &= \int_{s_0}^{1}\int_{s_1}^{1}\int_{s_2}^{1} \mathrm{d}s_3\,\mathrm{d}s_2\,\mathrm{d}s_1 = \frac{1}{6} - \frac{s_0}{2} + \frac{s_0^2}{2} - \frac{s_0^3}{6}, \\
I_{J=4}(s_0) &= \int_{s_0}^{1}\int_{s_1}^{1}\int_{s_2}^{1}\int_{s_3}^{1} \mathrm{d}s_4\,\mathrm{d}s_3\,\mathrm{d}s_2\,\mathrm{d}s_1 = \frac{1}{24} - \frac{s_0}{6} + \frac{s_0^2}{4} - \frac{s_0^3}{6} + \frac{s_0^4}{24}.
\end{aligned}
$$

At first the pattern might not be so apparent, however rewriting coefficients in a more convenient way might reveal it:

$$
\begin{aligned}
I_{J=1}(s_0) &= \frac{1}{1!}\Big(\frac{s_0^0}{0!}\Big) - \frac{1}{0!}\Big(\frac{s_0^1}{1!}\Big), \\
I_{J=2}(s_0) &= \frac{1}{2!}\Big(\frac{s_0^0}{0!}\Big) - \frac{1}{1!}\Big(\frac{s_0^1}{1!}\Big) + \frac{1}{0!}\Big(\frac{s_0^2}{2!}\Big), \\
I_{J=3}(s_0) &= \frac{1}{3!}\Big(\frac{s_0^0}{0!}\Big) - \frac{1}{2!}\Big(\frac{s_0^1}{1!}\Big) + \frac{1}{1!}\Big(\frac{s_0^2}{2!}\Big) - \frac{1}{0!}\Big(\frac{s_0^3}{6!}\Big), \\
I_{J=4}(s_0) &= \frac{1}{4!}\Big(\frac{s_0^0}{0!}\Big) - \frac{1}{3!}\Big(\frac{s_0^1}{1!}\Big) + \frac{1}{2!}\Big(\frac{s_0^2}{2!}\Big) - \frac{1}{1!}\Big(\frac{s_0^3}{3!}\Big) + \frac{1}{0!}\Big(\frac{s_0^4}{4!}\Big).
\end{aligned}
$$

Putting coefficients in this form the pattern is clear, hence we can write down the formula for $I_{J=n}(s_0)$,

$$
\begin{aligned}
I_{J=n}(s_0) &= \sum_{i=0}^{n} (-1)^i \frac{1}{(n-i)!}\Big(\frac{s_0^i}{i!}\Big) \\
&= \frac{1}{n!}\sum_{i=0}^{n} \binom{n}{i} (-s_0)^i 1^{n-i}
\end{aligned}
$$

observing that this is a binomial expansion we finalise our result

$$
= \frac{(1-s_0)^n}{n!}. \tag{6.5.19}
$$

Having worked out the value of $I_{J=n}(s_0)$ for $n \in \mathbb{N}_0$ we finally arrive at the final form of the joint probability of $n \geq 1$ jumps together with visiting points $x_{s_1}, ..., x_{s_n}$, conditional on a starting point $x_{s_0}$,

$$p(J = n, x_{s_1}, ..., x_{s_n} | x_{s_0}) = \delta^n I_{J=n}(s_0) \prod_{i=0}^{n-1} \frac{\sigma(x_{s_i})}{f(x_{s_i})} + \mathcal{O}(\delta^{n+1}), \quad n \geq 1. \tag{6.5.20}$$

We are interested in $p(J = n|x_{s_0})$, therefore we next proceed to integrate intermediate points $x_{s_1}, ..., x_{s_n}$ to get a functional form for $p(J = n|x_{s_0})$ up to a first leading term. Essentially we need to solve the integral

$$p(J = n|x_{s_0}, s_0) = \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} p(J = n, x_{s_1}, ..., x_{s_n}|x_{s_0}) \prod_{i=1}^{n} f_{s_i}(x_{s_i}) \, \mathrm{d}x_{s_n} ... \, \mathrm{d}x_{s_1}, \tag{6.5.21}$$

up to a first leading term, where each $f_{s_i}(x_{s_i})$ is a p.d.f of a point $x_{s_i}$ of the form as before,

$$f_{s_i}(x_{s_i}) = f(x_{s_i}) + s_i \delta h(x_{s_i}).$$

Due to the product structure in (6.5.20), integrals

$$I_{s_i} = \int_{-\infty}^{+\infty} \frac{\sigma(x_{s_i})}{f(x_{s_i})} f_{s_i}(x_{s_i}) \, \mathrm{d}x_{s_i} \tag{6.5.22}$$

can be solved independently.

$$\begin{aligned}
I_{s_i} &= \int_{-\infty}^{+\infty} \frac{\sigma(x_{s_i})}{f(x_{s_i})} \Big( f(x_{s_i}) + s_i \delta h(x_{s_i}) \Big) \, \mathrm{d}x_{s_i}, \\
&= \int_{-\infty}^{+\infty} \sigma(x_{s_i}) \Big( 1 + s_i \delta \frac{h(x_{s_i})}{f(x_{s_i})} \Big) \, \mathrm{d}x_{s_i}, \\
&= \int_{-\infty}^{+\infty} \sum_{j,i} [h^{\psi}_{j,i} \psi_{j,i}(x_{s_i})]^- \, \mathrm{d}x_{s_i} + \delta s_i \int_{-\infty}^{+\infty} \frac{\sigma(x_{s_i}) h(x_{s_i})}{f(x_{s_i})} \, \mathrm{d}x_{s_i}, \\
&= \int_{-\infty}^{+\infty} \sum_{j,i} [h^{\psi,+}_{j,i} \psi^-_{j,i}(x_{s_i}) + h^{\psi,-}_{j,i} \psi^+_{j,i}(x_{s_i})] \, \mathrm{d}x_{s_i} + \mathcal{O}(\delta), \\
&= \sum_{j,i} [h^{\psi,+}_{j,i} A_j + h^{\psi,-}_{j,i} A_j] + \mathcal{O}(\delta),
\end{aligned}$$

where $A_j$ is the normalisation constant of $\psi^+_{j,i}(x_{s_i})$ and $\psi^-_{j,i}(x_{s_i})$, and we have assumed the exchangeability of infinite sums and infinite integrals,

$$= \sum_{j,i} A_j |h^{\psi}_{j,i}| + \mathcal{O}(\delta), \tag{6.5.23}$$

using $|h^{\psi}_{j,i}| = h^{\psi,+}_{j,i} + h^{\psi,-}_{j,i}$.

Now we finalise our results, from (6.5.20), (6.5.21) and (6.5.23) we have

$$p(J = n|x_{s_0}, s_0) = \delta^n I_{J=n}(s_0) \frac{\sigma(x_{s_0})}{f(x_{s_0})} \left( \sum_{j,i} A_j |h^{\psi}_{j,i}| \right)^{n-1} + \mathcal{O}(\delta^{n+1}), \quad n \geq 1. \tag{6.5.24}$$

Integrating over the intermediate points $x_{s_1}, ..., x_{s_n}$ we have arrived at (6.5.24), which provides the first leading term of the probability $p(J = n|x_{s_0})$. However, we are still able to ask, what is the probability of reaching target in $J$ jumps given we find ourselves at time $s_0$, and avoiding conditioning on a specific point $x_{s_0}$, which we do next.

### 6.5.2 Generalising probability of jumps and expectations

In this subsection we will focus on investigating a more general probability of jumps required to reach a target, also we will be proving the result of Proposition 6.5.1.

**Proposition 6.5.1.** *If we have a starting distribution with density $f(x)$ and the target with density $g(x)$ as defined in equation (6.3.4), then*

$$\mathbb{E}[J|s_0 = 0] = \delta\beta + \mathcal{O}(\delta^2), \quad \text{Var}[J|s_0 = 0] = \delta\beta + \mathcal{O}(\delta^2),$$

*where*

$$\beta = \sum_{j,i} A_j |h^{\psi}_{j,i}|. \tag{6.5.25}$$

In the previous subsection we have derived (6.5.24), which still conditions on the point present at the start of WMC, it is possible to further generalise the jump probability by integrating over the starting point and conditioning only on the time present in WMC:

$$p(J=n|s_0) = \int_{-\infty}^{+\infty} p(J=n|x_{s_0}, s_0) f_{s_0}(x_{s_0})\,\mathrm{d}x_{s_0}$$

using expression (6.5.24) for $p(J = n|x_{s_0}, s_0)$ and performing identical integral calculation as in (6.5.22), we obtain

$$= \delta^n I_{J=n}(s_0)\bigg(\sum_{j,i} A_j |h^{\psi}_{j,i}|\bigg)^n + \mathcal{O}(\delta^{n+1}), \quad n \geq 1.$$

If we are interested in the probability distribution at $p(J = n|s_0)$ for $s_0 = 0$, a starting point of WMC, then

$$p(J=n|s_0=0) = \frac{\delta^n}{n!}\bigg(\sum_{j,i} A_j |h^{\psi}_{j,i}|\bigg)^n + \mathcal{O}(\delta^{n+1}), \quad n \geq 1, \tag{6.5.26}$$

where for $s_0 = 0$ we have $I_{J=n}(s_0 = 0) = \frac{1}{n!}$. We can use this probability to find an expression for the expected number of jumps to the target given that we are at the starting time of WMC, $s_0 = 0$, using (6.5.26)

$$\begin{aligned}\mathbb{E}[J|s_0=0] &= \sum_{n=1}^{\infty} np(J=n|s_0=0) \\ &= \delta\bigg(\sum_{j,i} A_j |h^{\psi}_{j,i}|\bigg)\sum_{n=1}^{\infty}\frac{\delta^{n-1}}{(n-1)!}\bigg(\sum_{j,i} A_j |h^{\psi}_{j,i}|\bigg)^{n-1} + \mathcal{O}(\delta^2)\end{aligned}$$

taking $\delta \sum_{j,i} A_j |h^{\psi}_{j,i}|$ in front of the summation,

$$= \delta\beta e^{\delta\beta} + \mathcal{O}(\delta^2),$$

where we have denoted

$$\beta = \sum_{j,i} A_j |h^{\psi}_{j,i}|,$$

and used the Taylor expansion of the exponential function $e^x$ at $x = 0$. Therefore, up to the first order of $\delta$ we can claim that,

$$\mathbb{E}[J|s_0 = 0] \approx \delta\beta. \tag{6.5.27}$$

We have consistency for $\delta \to 0$ and $h(x) \equiv 0$, where first term becomes 0. Term $\beta$ acts as a slope coefficient that controls how fast first term grows linearly as $\delta$ increases away from 0. Applying the same technique we are also able to workout expression of $\mathrm{Var}[J|s_0 = 0]$,

$$\mathrm{Var}[J|s_0 = 0] = \mathbb{E}[J^2|s_0 = 0] - \mathbb{E}[J|s_0 = 0]^2. \tag{6.5.28}$$

Similarly as before,

$$\begin{aligned}\mathbb{E}[J^2|s_0 = 0] &= \sum_{n=1}^{\infty} n^2 p(J = n|s_0 = 0) \\ &= \sum_{n=1}^{\infty} n\frac{\delta^n}{(n-1)!}\left(\sum_{j,i} A_j|h_{j,i}^{\psi}|\right)^n + \mathcal{O}(\delta^2)\end{aligned}$$

keeping only the first term of the sum that involves $\delta$,

$$= \delta\beta + \mathcal{O}(\delta^2).$$

Therefore, we have

$$\mathrm{Var}[J|s_0 = 0] = \delta\beta + \mathcal{O}(\delta^2), \tag{6.5.29}$$

so the first $\delta$ term of variance scales identically to the one for the expectation. Next we explore how these first order approximations perform in practice. To test them we chose the starting distribution to be $\mathcal{N}(0, 1)$ with density $f(x)$ and we build our target as,

$$g(x) = f(x) + \delta\mathbb{1}_{[-1<x<1]}\sin(9x). \tag{6.5.30}$$

As long as we keep $0 \leq \delta \leq 0.25$, $g(x)$ is a proper density (Figure 6.3).

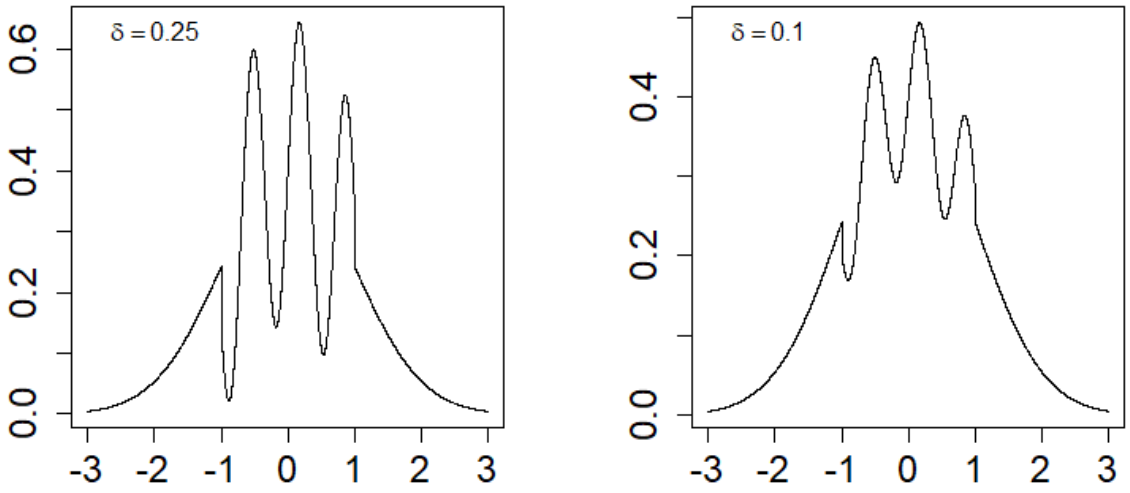



Figure 6.3: *Target density (6.5.30) for two different choices of $\delta$.*

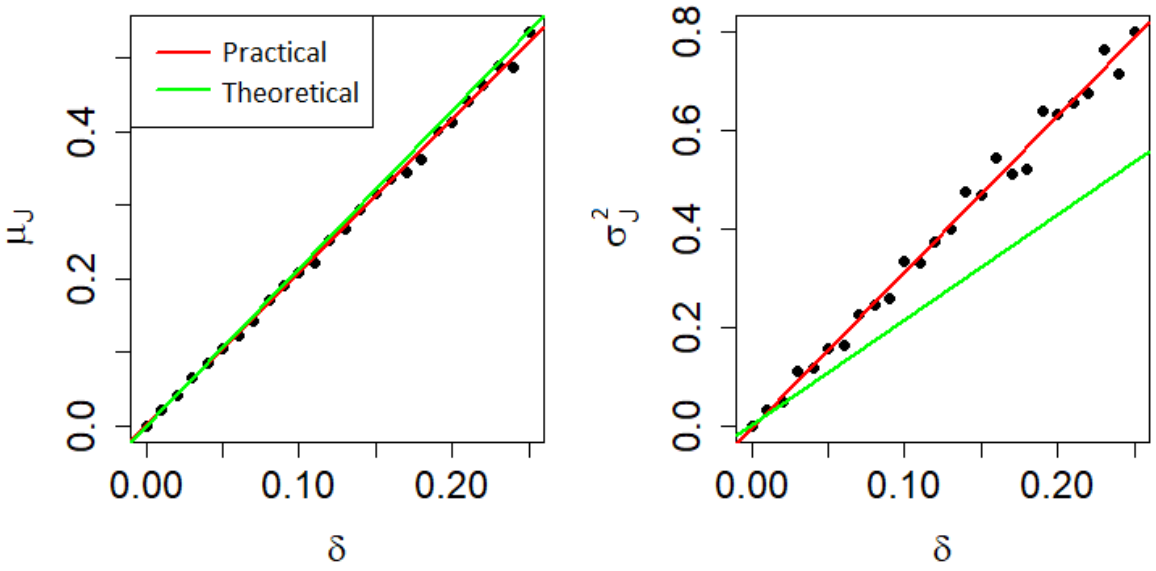



Figure 6.4: *Comparison of practical and theoretical results of the average number of jumps $\mu_J$ and the variance of jumps $\sigma_J^2$. The green line is $\delta\beta$ line with $\beta$ computed beforehand using theoretical results, while the red line is the regression line over simulated points in practical experiment.*

We perform WMC with $N = 10000$ over a set of $\delta$ values ranging from 0 to 0.25 and record the average number of jumps $\mu_J$ required to reach a target and the variance of jumps $\sigma_J^2$ for each simulation.

As we can see in Figure 6.4, the first order approximation for $\mathbb{E}[J|s_0 = 0]$ is a very accurate one, where the relationship between the average number of jumps and $\delta$ is linear and the $\beta = 2.14$ slope coefficient is almost identical to the one predicted from the regression $\hat{\beta} = 2.1$. However, even though the relationship between $\text{Var}[J|s_0 = 0]$ and $\delta$ seems to be linear for the limited range of $\delta$ values in the simulation, the predicted slope coefficient $\hat{\beta} = 3.4$ is significantly different from the theoretical $\beta = 2.14$. We conclude that for small values of $\delta$ first order linear approximation $\delta\beta$ is a good estimator for $\mathbb{E}[J|s_0 = 0]$; however, higher order terms need to be included for approximation of $\text{Var}[J|s_0 = 0]$.

**Side note on $\beta$ coefficient**

In the previous section, we have focused on Proposition 6.5.1. The main results of it being that first order approximations of both $\mathbb{E}[J|s_0 = 0]$ and $\text{Var}[J|s_0 = 0]$ scale linearly in $\delta$ with slope coefficient being $\beta = \sum_{j,i} A_j |h_{j,i}^{\psi}|$. So, $\beta$ controls how small changes in $\delta$ translate to changes in the expected number of jumps in WMC. Throughout Chapter 3, the explicit condition for WMC to produce a sample from the target in a finite number of jumps was not discussed. Here using theoretical results from the expected number of jumps we can deduce one of the necessary conditions rather trivially.

If the slope coefficient $\beta$ is not finite, then for any $\delta > 0$ we end up with $\mathbb{E}[J|s_0 = 0] = \infty$ and WMC is not able to produce a sample from a target in a finite number of jumps. Therefore, we conclude that one of the necessary conditions for WMC to

'converge' in a finite number of jumps is

$$\sum_{j,i} A_j |d^{\psi}_{j,i}| < \infty, \tag{6.5.31}$$

where we have substituted coefficients $h^{\psi}_{j,i}$ in $\beta$ definition with a more general difference function coefficients $d^{\psi}_{j,i}$ without loss of generality. Inequality (6.5.31) is closely related to the norm of Besov spaces and will be discussed in more detail in Chapter 7.

## 6.6 Tuning heuristic for the choice of $f(\cdot)$

From Section 6.3, we will use equation (6.3.3) as a tuning heuristic to find an optimal starting distribution for WMC algorithm. We will use a Monte Carlo estimate

$$\hat{p}(J = 0|s_0 = 0) = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{f(x_i)}{g(x_i)} \right)^{\rho(x_i)}, \tag{6.6.32}$$

where $x_i \sim f(\cdot)$ to aid us in this task. In (6.6.32), we performed a Monte Carlo integration of Equation 6.3.3 to find an approximate probability of performing zero jumps at the start of the algorithm. The idea here is that a better starting distribution for particular choice of the target $g(\cdot)$ should produce the higher probability values $\hat{p}(J = 0|s_0 = 0)$. These could be inspected before implementing WMC, at a much smaller computational cost than running WMC with a blind choice of $f(\cdot)$, to chose a more optimal starting distribution. Figure 6.5 presents an example where the target distribution is $\mathcal{N}(0, 1)$:

$$g(x; \mu = 0, \sigma = 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

and the starting distribution was set to be also normal but with different choices of $\mu$ and $\sigma$. Naturally, the highest zero-jump probability is observed around $\mu = 0$ and $= 1$ and other grid point evaluations suggest a choice of parameters around those values.

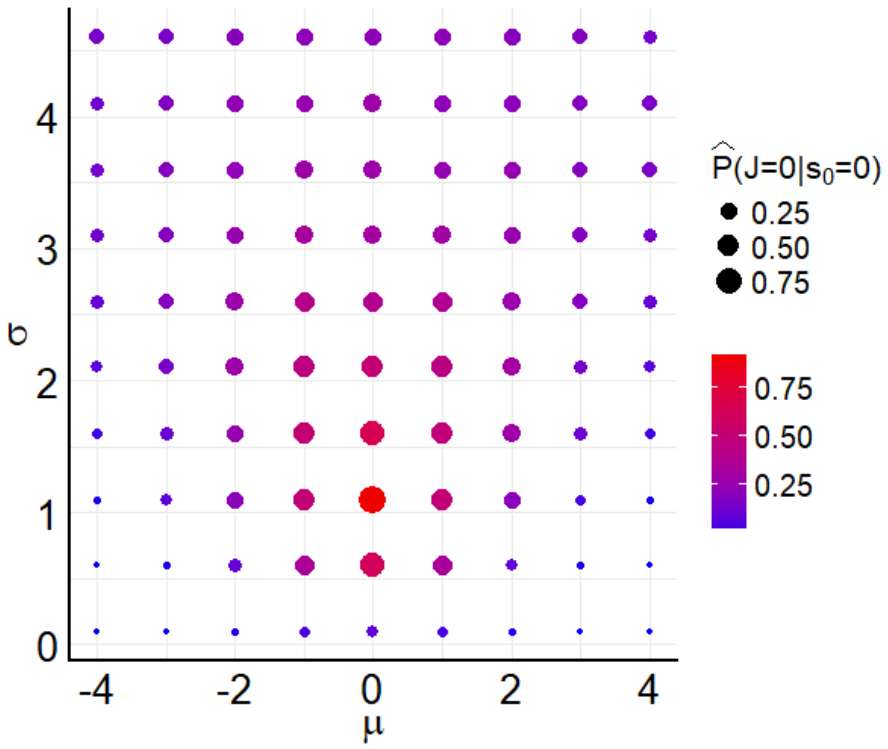



Figure 6.5: *Grid search over $\mu$ and $\sigma$ parameters for the optimal choice of a starting distribution.*

In cases where WMC will have to bet set to run for a long time, it is important to choose the best possible starting distribution to not waste time on unnecessary jumps, this method could be applied apriori as a cost-efficient way of tuning a starting distribution. Performing a grid search over parameters of the starting distribution could reveal the most optimal combination.

In this example, WMC was set to run for $N = 2000$ with $\mathcal{N}(1, 1.5^2)$ as the starting distribution and $\mathcal{N}(0, 1)$ as the target. The run took 2.6 minutes of running time, producing on average 2.2 jumps with a standard deviation of 4. Keeping the same target a starting distribution $\mathcal{N}(-3, 4^2)$ was chosen, the execution time was 4.2 minutes, producing on average 3.6 jumps with a standard deviation of 9.8. The overall difference was a 61% increase in the execution time and 63% increase in the

size of the average jump. An optimal choice of the starting distribution for WMC is one of the key tasks that needs to be completed to ensure the minimal execution cost of the algorithm.

## 6.7 Summary

In this chapter, we have attempted to investigate the probability distribution associated with the total number of jumps in WMC algorithm. At first, the functional form for the $p(J = 0|x_s)$ (6.3.3) was presented and later it was used to construct the MC estimate for $\hat{p}(J = 0|s_0 = 0)$ (6.6.32). The constructed MC estimate could be used as a tuning heuristic for the choice of the more optimal starting distribution in WMC.

Given the theoretical complexity induced by the presence of wavelets $\psi_{j,i}(\cdot)$, intractability of integrals (6.5.15) and general choices of starting and target densities $f(\cdot)$ and $g(\cdot)$, the decision was made to reduce the difficulty of the problem by considering the target density of the form

$$g(x) = f(x) + \delta h(x), \quad 0 < \delta \ll 1.$$

The introduction of the $\delta$ parameter allowed us to apply Taylor expansion techniques to investigate how certain probabilities behave up to certain order of $\delta$. Probabilities $p(J = n|x_s)$ (6.5.24) were worked out up to a first leading term for $n \in \mathbb{N}_0$. Furthermore, probabilities were generalised even further by removing conditioning on a certain starting point and only keeping the conditioning on the time $s_0$ in the algorithm. Finally, $p(J = n|s_0 = 0)$ was used to investigate $\mathbb{E}[J|s_0 = 0]$ and $\mathrm{Var}[J|s_0 = 0]$, it was demonstrated that leading terms for both, expectation and variance were

$$\delta\beta, \tag{6.7.33}$$

with

$$\beta = \sum_{j,i} A_j |h^{\psi}_{j,i}|.$$

The next step involving the analysis of distribution of jumps would be to restrict both $f(\cdot)$ and $g(\cdot)$ to certain families of densities and explore how these relations translate to probabilities of jumps.

# Chapter 7

# Haar wavelets and Besov spaces in WMC

In this chapter, an analysis of the necessary assumption **A2** on page 43 will be presented from a novel perspective, revealing an underlying connection between Besov spaces and the WMC set-up. The smoothness of functions in Besov spaces (Sawano 2018, Triebel 1992) will be discussed, suggesting why Haar wavelets are unable to satisfy assumption **A2** and be used in WMC.

## 7.1 Investigation of the assumption A2

In Section 5.3, it was demonstrated that there is an underlying issue with Haar wavelets, in particular that if Haar wavelets are used in WMC, then no probability mass can be transitioned across the origin. The WMC theory presented in §3.2, §3.3 and proofs of Proposition 3.2.1 and Theorem 3.3.2 involving the validity of the algorithm do not explicitly rule out the usage of Haar wavelets in principle. Therefore, the Haar issue demonstrated in §5.3 could be the consequence of the Haar wavelet failing to satisfy assumptions required to implement WMC. The first

requirement ever imposed on functions $f(\cdot)$, $g(\cdot)$ and wavelets $\psi_{j,i}(\cdot)$, was in pWMC theory (page 43), in particular that

$$\textbf{A2.} \quad \sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}[d^{\psi}_{j,i}\psi_{j,i}(x)]^{-} \leq rf(x) \quad \forall x \in \mathbb{R}.$$

The restriction above was only imposed so that the law of total probability is still satisfied, i.e. if with probability

$$q_{j,i}(x) = \frac{[d^{\psi}_{j,i}\psi_{j,i}(x)]^{-}}{rf(x)}$$

we select a new wavelet in pWMC, then $\sum_{j,i} q_{j,i}(x) \leq 1 \; \forall x \in \mathbb{R}$. Given the strictness of this assumption, the pWMC algorithm was upgraded to WMC by allowing points to jump several times before reaching a target. If the pWMC algorithm is applied to the starting distribution with density

$$f_t(x) = f(x) + td(x)$$

and a target with density

$$f_{t+\epsilon}(x) = f(x) + (t+\epsilon)d(x), \quad 0 < \epsilon \ll 1,$$

then we have that a difference function for this particular case is

$$d^{*}(x) = f_{t+\epsilon}(x) - f_t(x) = \epsilon d(x).$$

Using functions above the assumption A2 takes the form of

$$\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}[d^{\psi}_{j,i}\psi_{j,i}(x)]^{-} \leq \frac{rf_t(x)}{\epsilon} \quad \forall x \in \mathbb{R}, \tag{7.1.1}$$

after dividing both sides by $\epsilon$. Given that, in the WMC case, we work with infinitesimally small time steps, implemented via the survival analysis approach, inequality (7.1.1) should be analysed in the limit as $\epsilon \to 0$. Therefore, the actual restriction that is imposed on the wavelet representation of the difference function in standard WMC is

$$\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}[d^{\psi}_{j,i}\psi_{j,i}(x)]^{-} < \infty \quad \forall x \in \mathbb{R}. \tag{7.1.2}$$

Given the significant relaxation on the condition required to implement WMC successfully, it seems that the only issue which could arise occurs if, for certain choice of wavelet family $\psi_{j,i}(\cdot)$ and/or functions $f(\cdot)$, $g(\cdot)$, the value $\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}[d^{\psi}_{j,i}\psi_{j,i}(x)]^-$ becomes infinite for some $x \in \mathbb{R}$.

Equivalently, if we integrate both sides of (7.1.2), assuming the exchangeability of infinite sums and infinite integrals,

$$\begin{aligned}\int_{-\infty}^{+\infty}\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}[d^{\psi}_{j,i}\psi_{j,i}(x)]^-\,\mathrm{d}x &= \sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}\left(d^{\psi,+}_{j,i}\int_{-\infty}^{+\infty}\psi^-_{j,i}(x)\,\mathrm{d}x + d^{\psi,-}_{j,i}\int_{-\infty}^{+\infty}\psi^+_{j,i}(x)\,\mathrm{d}x\right)\\ &= \sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}A_j(d^{\psi,+}_{j,i}+d^{\psi,-}_{j,i})\\ &= \sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}A_j|d^{\psi}_{j,i}|,\end{aligned}$$

where $A_j = \int_{-\infty}^{+\infty}\psi^+_{j,i}(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty}\psi^-_{j,i}(x)\,\mathrm{d}x$ as before. Using the identity $|d^{\psi}_{j,i}| = d^{\psi,+}_{j,i} + d^{\psi,-}_{j,i}$, then we arrive at the new condition

$$\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}A_j|d^{\psi}_{j,i}| < \infty. \tag{7.1.3}$$

Condition (7.1.3) is not necessarily always true even if (7.1.2) is true, as the integral can still diverge. However, we recall that (7.1.3) is exactly the same condition as the one on the finiteness of the slope coefficient $\beta$ in §6.5.2 where we analysed the expected number of jumps. We showed that $\beta < \infty$ was a necessary condition for WMC, therefore we must have both conditions

$$\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}[d^{\psi}_{j,i}\psi_{j,i}(x)]^- < \infty \quad \forall x \in \mathbb{R} \quad \text{and} \quad \sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}A_j|d^{\psi}_{j,i}| < \infty \tag{7.1.4}$$

satisfied if we want WMC to work. The key question here is whether (7.1.2) is a sufficient condition for WMC which would also imply (7.1.3). To better understand the regularity conditions imposed on the starting distributions and wavelets used in WMC, we will focus our attention on (7.1.3).

For (7.1.3) to be satisfied, we essentially require that there is a fast enough coefficient decay across the resolution levels, i.e., coefficients go to zero when $|j| \to \infty$. Let us define the energy at resolution level $j$ to be

$$E_j := \sum_{i \in \mathbb{Z}} |d^{\psi}_{j,i}|. \tag{7.1.5}$$

What type of energy decay should we have across levels if we want the necessary condition (7.1.3) to be satisfied? Let us for the moment consider that $E_j$ decays geometrically with increasing $|j|$. We also assume that there exists

$$j_{\max} := \arg\max_{j \in \mathbb{Z}} E_j \tag{7.1.6}$$

and

$$E_j \leq C\alpha^{-|j-j_{\max}|}, \quad \forall j \in \mathbb{Z} \text{ and } C, \alpha \in \mathbb{R}. \tag{7.1.7}$$

Is this a reasonable assumption? Let us investigate what is the distribution of $E_j$ across a range of levels in the one-dimensional example of §4.1.1. As we can see in Figure 7.1, the distributions of energy levels for $K \geq 2$ do indeed have a peak at same $j_{\max} = -1$ and they decay rather rapidly. We also make an observation, that distributions of $E_j$ for $K \geq 2$ are almost identical as plots for $K \geq 2$ seem to overlap almost perfectly. However, for $K = 1$ (Haar) the distribution is significantly shifted and does not seem to have a maximum between the resolution levels -5 to 5. We note, that given the limitations of the computational power a finite number of shifts $i$ were taken to compute $E_j$ at each resolution, however given the sparsity of wavelet coefficients it should be the case that at some point for location $i$ significantly far from the high density region, the contribution towards $E_j$ is negligible. We repeat the experiment by increasing the number of resolution levels from $j \in [-5, 5]$ to $j \in [-15, 15]$ and we also increase the effective support size to $x \in [-20, 15]$. From Figure 7.2 we can conclude that for Daubechies $K \geq 2$ wavelets the inclusion of extra locations did not change the differences between energy distributions significantly – energies $E_j$ continue to decay as $j$ increases with no hint of forming another

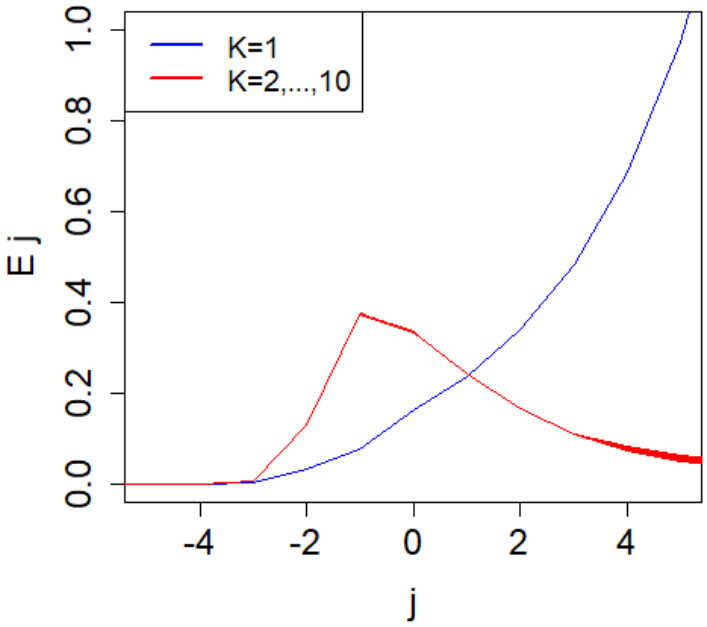



Figure 7.1: *Distribution of energies $E_j$ across resolution levels $j$ for the difference function from the one dimensional example of §4.1.1. Energies were computed using Daubechies wavelets with $K = 1, 2, ..., 10$ vanishing moments. The range of locations used at each resolution level $j$ to compute energies $E_j$ was $-10 \times 2^j \leq i \leq 8 \times 2^j$. Values -10 and 8 were chosen arbitrary but large enough to make sure that the effective support of the difference function is fully covered.*

maximum peak. Furthermore, the energy level for the Haar wavelet is blowing up much more rapidly with no hint towards its $j_{\max}$. These observations suggest for $K \geq 2$ energies $E_j$ seem to have a global maximum and could be modelled by a uni-modal distribution that decays rapidly. At this point we assume that inclusion of more resolution levels and locations will not change overall results dramatically. From these results we conclude that Daubechies wavelets $K \geq 2$ should most likely

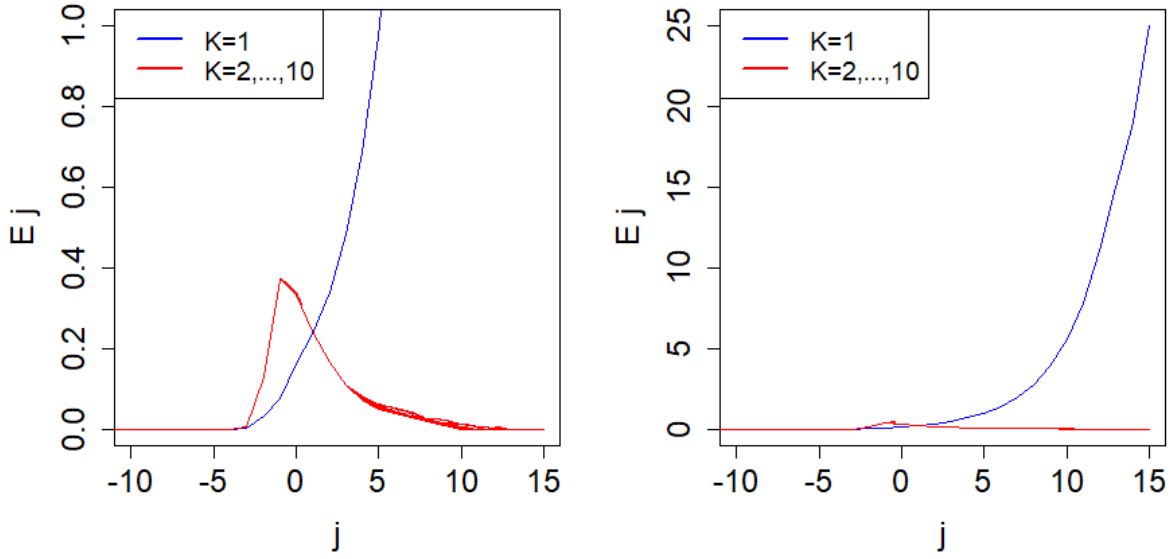



Figure 7.2: *Similar to Figure 7.1, but with larger choice of locations and resolution levels.*

have no problems with the validity of

$$\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}} A_j|d^{\psi}_{j,i}| < \infty,$$

due to the overall decay of $E_j$ for $|j| \to \infty$, however a much more detailed theoretical investigation must be carried out to confirm it. The most concerning issue as we can see from figures is Haar wavelets, where the total energy $\sum_j E_j$ have a tendency to increase very rapidly if more locations and resolutions are added into the computation of $E_j$. This suggests that Haar wavelets could have some serious problems with necessary conditions (7.1.4).

## 7.2 Introduction to Besov spaces

Here, we will give a short introduction to Besov spaces together with required background information and definitions. Most of the concepts describe here come from the theory of functional analysis and at first sight might look alien and unintuitive for a first time reader coming from statistics background. To keep focus on WMC related issues only the necessary definitions will be given together with theorems without proofs. The key goal of this section is to understand the environment of functional analysis surrounding Besov spaces and transfer results of Besov space theory to WMC.

The main result of this section is Theorem 7.2.2 on page 149. The lead up to this result is rather heavy, requiring some technical definitions from the theory of functional analysis. Before going into more theoretical background of functional analysis, we will try to give an intuition on the significance of the results provided in this section. Let $G$ be some general function space and $f \in G$ any function that could be decomposed into a series of orthonormal wavelets,

$$f(\cdot) = \sum_{j,i} f^{\psi}_{j,i} \psi_{j,i}(\cdot).$$

$G$ has an associated space norm $\|\cdot\|_G$ and similarly we can consider coefficients $\{f^{\psi}_{j,i}\} \in \mathcal{G}$ be part of some coefficient space $\mathcal{G}$. Theorem 7.2.2 provides us with the conditions under which we are allowed to approximate norm $\|\cdot\|_G$ using norm of the coefficient space $\|\cdot\|_{\mathcal{G}}$. In other words there exist a norm equivalence and we are able to conclude the characteristics of our functions of interest by analysing the associated wavelet coefficients. Using these results we will be able to draw conclusion on what type of wavelets we are allowed to use in the WMC setting.

So far we have only mentioned and used the space of square integrable functions $L^2(\mathbb{R})$ in one dimension, as this is the requirement for a function to be decomposed

into a series of orthonormal wavelets. In practice however, there are many other interesting functional spaces, for example: $C^k$ - the space of continuous functions with $k$ continuous derivatives, $C^\infty$ - smooth function space, $C_c^\infty$ - space of smooth functions with compact support, $\mathcal{C}^s$ - Holder-Zygmund spaces, $W^{k,p}$ - fractional/non-fractional Sobolev spaces, $\dot{B}^s_{p,q}/B^s_{p,q}$ - homogeneous/inhomogeneous Besov spaces, $\dot{H}^p/H^p$ - homogeneous/inhomogeneous Hardy spaces, $BMO$ - bounded mean oscillation spaces, $\mathcal{S}$ - Schwartz spaces, $\mathcal{O}(\mathbb{C})$ - the space of holomorphic functions. Each of the spaces mentioned deals with different types of regularity, smoothness, differentiability and integrability properties of functions. In this section, we will mainly focus on homogeneous Besov spaces $\dot{B}^s_{p,q}$ (as described in Sawano (2018)) and how they are connected to WMC theory.

Besov spaces were first introduced by O.V. Besov in 1959/60, see Besov (1959, 1961). The idea was to extend and create a more general space that would include the $\mathcal{C}^s$ and $W^{k,p}$ spaces which were extensively studied at that time. Naturally, a space that involves many parameters in its characterisation is able to describe different kind of regularity properties of a function. From the function spaces mentioned above we have that:

- $L^p$, $W^{k,p}$, $\dot{B}^s_{p,q}/B^s_{p,q}$, $\dot{H}^p/H^p$ deal with the size of functions, while
- $C^k$, $C^\infty$, $C_c^\infty$, $\mathcal{O}(\mathbb{C})$, $\mathcal{C}^s$, $W^{k,p}$, $\dot{B}^s_{p,q}/B^s_{p,q}$, deal with the differentiability of functions.

As we can see, Sobolev and Besov spaces are able to describe both size and differentiability. Consequently, the more flexible the space becomes the more intricate its definition becomes. Here we understand, that in both, homogeneous and inhomogeneous Besov spaces $\dot{B}^s_{p,q}$, $B^s_{p,q}$, the parameter $p$ is responsible for describing the size (total energy) of a function, $s$ (also know as index of regularity) is for control of smoothness and $q$ is used for describing additional levels of smoothness

via differences of a function. The parameter $q$ appears via more detailed construction of a Besov space and for the sake of not diving too deep in to the realm of functional analysis it will not be explained in detail. Before moving to definitions of $\dot{B}^s_{p,q}$, $B^s_{p,q}$, we qualitatively state differences between *homogeneous* and *inhomogeneous* spaces.

Essentially, homogeneous spaces are function spaces whose norms are described by a set of partial derivatives of the same order; otherwise the space is inhomogeneous.

Let us define $\alpha = (\alpha_1, ..., \alpha_n)$ with $\alpha_j \in \mathbb{N}_0$ ($\mathbb{N}_0 = \mathbb{N} \cup \{0\}$) to be $n$-dimensional multi-index, $|\alpha| := \sum_{j=1}^n \alpha_j$ and

$$\partial^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}(x).$$

**Example 7.2.1.** Let $m \in \mathbb{N}$ and $1 \leq p \leq \infty$.

1. The homogeneous Sobolev norm is $\|f\|_{\dot{W}^{m,p}} \equiv \sum_{|\alpha|=m} \|\partial^\alpha f\|_{L^p}$.
2. The inhomogeneous Sobolev norm is $\|f\|_{W^{m,p}} \equiv \sum_{|\alpha|\leq m} \|\partial^\alpha f\|_{L^p}$.

We also observe that, since via differentiation we annihilate polynomials or decrease their order, the homogeneous norms lose some information about functions, hence they are not complete. However, despite this observation, homogeneous norms are good at describing certain specific properties of functions, and a typical one would be dilation $f \mapsto f(t\cdot)$, inhomogeneous norms cannot be used in describing this property.

We next proceed by defining some necessary function spaces required for the construction of $\dot{B}^s_{p,q}$ and $B^s_{p,q}$.

**Definition 7.2.1** (*Smooth function space, $C^\infty$*)**.** A function $f$ is said to be smooth and belong to $C^\infty$ if it is differentiable for all degrees of differentiation.

**Definition 7.2.2** (*Schwartz function space, $\mathcal{S}$*)**.** We denote the space

$$\mathcal{S}(\mathbb{R}) = \{f \in C^\infty(\mathbb{R}) : \|f\|_{\alpha,\beta} < \infty, \forall \alpha, \beta \in \mathbb{N} \text{ where } \|f\|_{\alpha,\beta} = \sup_{x\in\mathbb{R}} |x^\alpha f^{(\beta)}(x)|\}$$

as the space of *Schwartz functions*.

The map $\|\cdot\|_{\alpha,\beta}$ is called a semi-norm as it has most properties of the norm, however it does not satisfy the positive definiteness property, i.e. non-zero vectors could be mapped to zero.

**Definition 7.2.3** (*Functional*)**.** We call a map $f$ a functional if

$$f : \Omega \to \mathbb{R}, \text{ where } \Omega \text{ is a function space.}$$

**Example 7.2.2.** Let $I : C(\mathbb{R}) \to \mathbb{R}$ be defined as $I[u] = \int_{-\infty}^{+\infty} u(x)^2 \, \mathrm{d}x$. $I$ is a functional.

From Definition 7.2.2, we can see that the elements in $\mathcal{S}(\mathbb{R})$ are infinitely differentiable and partial derivatives decay rapidly. These functions are very well behaved and most likely too well behaved to be encountered in practice, for this reason we will investigate the space of *tempered distributions* $\mathcal{S}'(\mathbb{R})$. The object *tempered distribution* here has no connection to the standard definition of a probability distribution in the probability theory, therefore these should not be confused together.

**Definition 7.2.4** (*Tempered distributions space*, $\mathcal{S}'(\mathbb{R})$)**.** Let $T : \mathcal{S}(\mathbb{R}) \to \mathbb{R}$ be a functional. We say $T$ is part of $\mathcal{S}'(\mathbb{R})$ and is a *tempered distribution* if it is both linear and continuous. One equips $\mathcal{S}'(\mathbb{R})$ with the weakest topology so that the mapping

$$T_f \in \mathcal{S}'(\mathbb{R}) : f \in \mathcal{S}(\mathbb{R}) \mapsto \langle f, \phi \rangle \in \mathbb{R}$$

is continuous for all test functions $\phi \in \mathcal{S}$.

We note that by $\langle f, \phi \rangle$ we mean a standard inner product $\int f(x)\phi(x) \, \mathrm{d}x$ and test functions $\phi$ should not be confused with father wavelet. We also remark that in general the space of linear and continuous functionals on a vector space is called the continuous *dual* of the space.

**Example 7.2.3.** Let $f$ be a function such that the product $f\phi$ is integrable on $\mathbb{R}$ for all $\phi \in \mathcal{S}(\mathbb{R})$. Then we denote the tempered distribution induced by $f$ as $T_f : \mathcal{S}(\mathbb{R}) \to \mathbb{R}$ and define it as

$$T_f(\phi) = \int_{\mathbb{R}} f(x)\phi(x)\,\mathrm{d}x.$$

It can be shown that $T_f$ is indeed a tempered distribution, by confirming a linearity property (linearity of integrals), homogeneity and finally continuity of $T_f$ by considering an arbitrary sequence of Schwartz functions $\{\phi_n\}_{n=0}^{\infty}$ that converges to $\phi$.

Finally, let us denote $\mathcal{P}$ to be the set of all polynomials and by $\hat{\phi}$ we denote a Fourier transform of a function $\phi(\cdot)$,

$$\hat{\phi}(\xi) = \mathcal{F}\phi(\xi) = \int_{x\in\mathbb{R}} \phi(x)e^{-ix\xi}\,\mathrm{d}x. \tag{7.2.8}$$

**Definition 7.2.5** (*Homogeneous Besov spaces,* $\dot{B}^s_{p,q}$)**.** Let us denote $\phi \in \mathcal{S}$ so that $\mathrm{supp}\{\hat{\phi}\} \subset \{\xi : \frac{1}{2} \leq |\xi| \leq 2\}$ and $|\hat{\phi}(\xi)| \geq c > 0$ if $\frac{3}{5} \leq |\xi| \leq \frac{5}{3}$. For $s \in \mathbb{R}$, $0 < p, q \leq \infty$ and $f : T_f \in \mathcal{S}'/\mathcal{P}$ we define

$$\|f\|_{\dot{B}^s_{p,q}} = \left\{ \sum_{j\in\mathbb{Z}} \left( 2^{js} \|\phi_j * f\|_{L^p} \right)^q \right\}^{1/q}. \tag{7.2.9}$$

We define $\dot{B}^s_{p,q}$ to be the set of all such $f$ for which quasinorm (7.2.9) is finite.

In Definition 7.2.5 above, we have

$$\phi_j(\cdot) = 2^{jn/2}\phi(2^j\cdot), \tag{7.2.10}$$

where $n$ corresponds to the dimension of the space we consider $\dot{B}^s_{p,q}$ over, i.e. $\dot{B}^s_{p,q}(\mathbb{R}^n)$. We also note that although '$*$' is a standard notation for the convolution, in Definition 7.2.5 the convolution is taken between $\phi \in \mathcal{S}$ and $f \in \mathcal{S}'/\mathcal{P}$ and because $f$ is a tempered distribution, a convolution is non-standard and will be defined now. Before doing so we remind the reader of a standard convolution.

**Definition 7.2.6** (*Convolution*)**.** Let $f, g$ be functions. We define their convolution as

$$f * g(x) = \int_{-\infty}^{+\infty} f(x-y)g(y)\,\mathrm{d}y. \tag{7.2.11}$$

**Definition 7.2.7** (*Convolution of a distribution*)**.** Let $\psi, \phi \in \mathcal{S}(\mathbb{R})$ and $T \in \mathcal{S}'(\mathbb{R})$ then the convolution of $\psi$ and $T$ is a distribution and acts on $\phi$ as

$$\psi * T[\phi] := T[\tilde{\psi} * \phi] \tag{7.2.12}$$

where $\tilde{\psi}(x) = \psi(-x)$ is the reflection about 0.

To be more explicit, let us consider $T_f \in \mathcal{S}'(\mathbb{R})$, then we have that

$$\begin{aligned} \psi * T_f[\phi] &= T_f[\tilde{\psi} * \phi] && (7.2.13)\\ &= T_f\left[\int \psi(y-x)\phi(y)\,\mathrm{d}y\right] && (7.2.14)\\ &= \int\int \psi(y-x)\phi(y)\,\mathrm{d}y f(x)\,\mathrm{d}x. && (7.2.15) \end{aligned}$$

For Definition 7.2.7 the fact that convolution of Schwartz function is Schwartz was used, however it will not be proved here.

**Example 7.2.4.** It can be shown that the Dirac $\delta$ function is a tempered distribution and belongs to $\mathcal{S}'$, here we will assume this fact and will show how $\delta$ behaves when interacting with a convolution. Let $\psi, \phi \in \mathcal{S}(\mathbb{R})$ then

$$\begin{aligned} \psi * \delta[\phi] &= \delta[\tilde{\psi} * \phi] \\ &= \delta\left[\int_{-\infty}^{+\infty} \tilde{\psi}(x-y)\phi(y)\,\mathrm{d}y\right] \end{aligned}$$

using reflection $\tilde{\psi}(x) = \psi(-x)$,

$$= \delta\left[\int_{-\infty}^{+\infty} \psi(y-x)\phi(y)\,\mathrm{d}y\right]$$

using $\delta$ property $f(y) = \int f(y-x)\delta(x)\,\mathrm{d}x$,

$$= \int_{-\infty}^{+\infty} \psi(y)\phi(y)\,\mathrm{d}y = T_\psi[\phi].$$

As discussed in Frazier et al. (1991), Kyriazis & Petrushev (2002), Triebel (2004) and Sawano (2018), the space of functions $\dot{B}^s_{p,q}$ also accepts a multi-resolution decomposition, very similarly to $L^2$. Naturally, multi-resolution decomposition leads to the set of coefficients that are associated with 'atoms' used as building blocks of the decomposition. The set of all coefficients used in a decomposition could be considered as a space of coefficients, in particular, given that $f \in \dot{B}^s_{p,q}$, the space of associated decomposition coefficients of function $f$ will be denoted as $\dot{b}^s_{p,q}$.

**Definition 7.2.8** (*Dyadic cubes*)**.** We say that a cube $Q_{j,k} \subset \mathbb{R}^n$ is a dyadic cube if

$$Q_{j,k} = \{x \in \mathbb{R}^n : 2^{-j}k_i \leq x_i \leq 2^{-j}(k_i + 1), i = 1, 2, ..., n\} \tag{7.2.16}$$

for some $j \in \mathbb{Z}$ and $k = (k_1, k_2, ..., k_n) \in \mathbb{Z}^n$. Let $Q = \{Q_{j,k}, j \in \mathbb{Z}, k \in \mathbb{Z}^n\}$, we also denote $Q_j$, $j \in \mathbb{Z}$, for the collection of all cubes $I \in Q$ of side-length $l(I) = 2^{-j}$. For any dyadic cube $I \in Q$, we use $x_I$ for its lower-left corner and $|I|$ for its volume.

**Definition 7.2.9** (*Test functions*)**.** We will call a function $\phi : \mathbb{R}^n \mapsto \mathbb{R}$ a *test function* and say that it belongs to set $\mathcal{D}(\mathbb{R}^n)$ if it is *smooth* and has *compact support*.

**Definition 7.2.10** (*Smooth K-atoms*)**.** A function $a_{j,k} \in \mathcal{D}(\mathbb{R}^n)$ is a smooth $K$-atom for $Q_{j,k}$ if and only if

(1) $\text{supp}\{a_{j,k}\} \subset 3Q_{j,k}$,

(2) $\int_{x \in \mathbb{R}^n} x^\gamma a_{j,k}(x)\,\mathrm{d}x = 0$ for $|\gamma| \leq K$,

(3) $|\partial^\gamma a_{j,k}(x)| \leq c_\gamma l(Q_{j,k})^{-|\gamma|-n/2} \quad \forall \gamma \in \mathbb{N}^n$.

**Theorem 7.2.1** (*Atomic Decomposition Theorem*, Frazier et al. (1991))**.** *Suppose $s \in \mathbb{R}$, $0 < p, q < \infty$ and $K \in \mathbb{N}$. If $f \in \dot{B}^s_{p,q}$ then there exists a sequence $d = \{d_{j,k}\} \in \dot{b}^s_{p,q}$ and smooth $K$-atoms $\{a_{j,k}\}$ such that $f = \sum_{j,k} d_{j,k} a_{j,k}$ and*

$$\|d\|_{\dot{b}^s_{p,q}} \leq C \|f\|_{\dot{B}^s_{p,q}}. \tag{7.2.17}$$

Atomic decomposition theorem gives a relation between norm of functions and sequences and in turn provides an equivalence relation between $\dot{B}^s_{p,q}$ and $\dot{b}^s_{p,q}$. We next proceed in giving a proper definition of Besov sequence space $\dot{b}^s_{p,q}$ and its associated norm $\|\cdot\|_{\dot{b}^s_{p,q}}$.

**Definition 7.2.11** (*Homogenerous Besov sequence space,* $\dot{b}^s_{p,q}$)**.** For $s \in \mathbb{R}$ and $0 < p, q \leq \infty$ the space $\dot{b}^s_{p,q}$ consists of all sequences $s := \{s_I\}_{I \in Q}$, such that

$$\|s\|_{\dot{b}^s_{p,q}} := \left\{ \sum_{j \in \mathbb{Z}} \left( \sum_{I \in Q_j} \left[ |I|^{-s/n+1/p-1/2} |s_I| \right]^p \right)^{q/p} \right\}^{1/q} < \infty. \tag{7.2.18}$$

One of the key features of the wavelet representations is the fact that the wavelet coefficients implicitly contain valuable information about the size and the smoothness of the function being decomposed. In other words, if we have

$$f(x) = \sum_{j,i} f^{\psi}_{j,i} \psi_{j,i}(x)$$

one can determine from $f^{\psi}_{j,i}$ coefficients whether $f$ is contained in certain smoothness spaces, such as Besov or Sobolev spaces. Let us now proceed to the main result of this section.

We recall that multivariate wavelet bases are constructed as tensor products of a univariate scaling function $\psi^0 := \phi$ and associated wavelet $\psi$. Namely, let $E$ ($|E| = 2^n - 1$) denote the set of nonzero vertices of the unit cube in $\mathbb{R}^n$. For each vertex $e = (e_1, ..., e_n) \in E$ we let

$$\psi^e(x) := \psi^{e_1}(x_1) \cdots \psi^{e_n}(x_n) \tag{7.2.19}$$

and define $\Psi := \{\psi^e : e \in E\}$. So each $e_i$, $i \in \{1, ..., n\}$, takes value 0 or 1 and for $e_i = 0$ we recover $\phi$ and for $e_i = 1$ we get $\psi$, ensuring we get all the mixtures of possible types of wavelets. Then the collection

$$W := \{\psi^e_I : I \in Q, e \in E\} \tag{7.2.20}$$

forms an orthonormal basis for the space $L^2(\mathbb{R}^n)$.

Now, let $\Psi := \{\psi^e : e \in E\}$ be a set of orthonormal wavelets for $L^2(\mathbb{R}^n)$ which satisfy the following two conditions:

**C1.** $\Psi \subset C^K$ and

$$|\partial^{|\alpha|}\psi^e(x)| \leq \rho(1+|x|)^{-M}, \quad |\alpha| \leq K, e \in E, \tag{7.2.21}$$

**C2.**

$$\int_{x\in\mathbb{R}^n} x^\alpha \psi^e(x)\,\mathrm{d}x = 0, \quad |\alpha| \leq K, e \in E. \tag{7.2.22}$$

Then the following Theorem 7.2.2 holds (Kyriazis 2003).

**Theorem 7.2.2.** *Let $s \in \mathbb{R}$, $0 < p, q \leq \infty$, $\mathcal{T} := n/\min\{1,p\}$, $K > \max\{\mathcal{T} - n - s, s\}$ and $M > \max\{\mathcal{T}, n+K\}$. For every $f \in \dot{B}^s_{p,q}$ there exist unique coefficients $c_{I,e}(f)$, $(I,e) \in Q \times E$, such that*

$$f = \sum_{I\in Q}\sum_{e\in E} c_{I,e}(f)\psi^e_I \quad \textit{with} \quad c_{I,e}(f) := \langle f, \psi^e_I\rangle. \tag{7.2.23}$$

*Moreover,*

$$\|f\|_{\dot{B}^s_{p,q}} \approx \left\{\sum_{e\in E}\sum_{j\in\mathbb{Z}}\left(\sum_{I\in Q_j}\left[|I|^{-s/n+1/p-1/2}|c_{I,e}(f)|\right]^p\right)^{q/p}\right\}^{1/q}, \tag{7.2.24}$$

*where by $\approx$ we mean 'could be approximated by'.*

Theorem 7.2.2 provides us with exact conditions for the existence of a wavelet system which decomposes a function $f \in \dot{B}^s_{p,q}$ into a series of orthonormal wavelets. Furthermore, theorems 7.2.2 and 7.2.1 give us the equivalence relation between norms of the function space $\dot{B}^s_{p,q}$ and the corresponding sequence space $\dot{b}^s_{p,q}$,

$$\|\cdot\|_{\dot{B}^s_{p,q}} \asymp \|\cdot\|_{\dot{b}^s_{p,q}}. \tag{7.2.25}$$

This norm equivalence allows us to make critical deductions about functions we analyse given we have information about the corresponding wavelet coefficients. In the next section, we connect key results about Besov function spaces and sequence spaces with WMC theory.

## 7.3 Connecting Besov spaces and WMC

In Section 7.1, we deduced that one of the necessary conditions for the validity of the WMC algorithm is (7.1.3):

$$\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}} A_j|d^{\psi}_{j,i}| < \infty.$$

Let us rewrite this inequality in a form that will be convenient for us later. Using $A_j = 2^{-j/2}A_0$, where as before $A_0 = \int \psi^+(x)\,\mathrm{d}x = \int \psi^-(x)\,\mathrm{d}x$, we obtain

$$\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}} 2^{-j/2}|d^{\psi}_{j,i}| < \infty. \tag{7.3.26}$$

Inequality (7.3.26) gives a specific restriction on the wavelet coefficients $d^{\psi}_{j,i}$ of the difference function $d$. Furthermore, if we consider the set of coefficients $\{d^{\psi}_{j,i}\}$ over all $j, i$, (7.3.26) has a pseudo-norm resemblance on some coefficient space. Noticing this similarity we make the first connection with homogeneous Besov sequence spaces $\dot{b}^s_{p,q}$. From Definition (7.2.11), we know that the norm finiteness condition for $\dot{b}^s_{p,q}$ sequence space is,from 7.2.11,

$$\left\{\sum_{j\in\mathbb{Z}}\left(\sum_{I\in Q_j}\left[|I|^{-s/n+1/p-1/2}|s_I|\right]^p\right)^{q/p}\right\}^{1/q} < \infty. \tag{7.3.27}$$

If we set space parameters to be $p = q = 1$ and $s = 0$ with dimensionality $n = 1$, Inequality (7.3.27) becomes

$$\sum_{j\in\mathbb{Z}}\sum_{I\in Q_j}|I|^{1/2}|s_I| < \infty.$$

Instead of using the dyadic cube notation, we can rewrite indices in terms of resolution levels and locations,

$$\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}} 2^{-j/2}|s_{j,i}| < \infty,$$

which is identical in form to (7.3.26). From this we conclude that the necessary condition (7.3.26) can be guaranteed if wavelet coefficients $\{d^{\psi}_{j,i}\}$ live in the homogeneous Besov sequence space $\dot{b}^0_{1,1}$. Under the norm equivalence and Theorem 7.2.2, we have that

$$\{d^{\psi}_{j,i}\} \in \dot{b}^0_{1,1} \iff d(x) \in \dot{B}^0_{1,1}(\mathbb{R}),$$

where $\{d^{\psi}_{j,i}\}$ are wavelet coefficients and $d(x)$ a difference function as defined in (3.1.6). We can also see that for Theorem 7.2.2 to hold certain smoothness and decay conditions need to be satisfied on the wavelet system we are using. In $\dot{B}^0_{1,1}(\mathbb{R})$ we have that $\mathcal{T} = 1$, $K > 0$ and $M > \max\{1, 1+K\}$. So, if we want the necessary WMC condition (7.3.26) to hold, we must have a wavelet system that at least has wavelets $\psi \in C^1(\mathbb{R})$, i.e. wavelets are continuous and at least one time differentiable.

As it was already pointed out in Section 5.3, Haar wavelets are not able to transition probability mass across the origin. However, the WMC theory and the proofs regarding it, given in Chapter 3, do not explicitly state conditions on the wavelet system being used, so it is not really clear where the implicit assumption on the wavelets is made and why Haar wavelets fail at theoretical level.

Via the necessary condition (7.3.26) we are able to connect that the norm restriction on the wavelet coefficients $\{d^{\psi}_{j,i}\}$ also restricts our function space and most importantly wavelets that we are allowed to use in the decomposition. We next provide an example in which condition (7.3.26) fails to hold.

**Example 7.3.1.** Let a starting density be $f(x)$ we then define our target density to be

$$g(x) = f(x) + a\{\mathbb{1}(-1 \le x < 0) - \mathbb{1}(0 \le x < 1)\}. \tag{7.3.28}$$

So the shape of the difference function is a similar to an ordinary Haar wavelet; however, it has a scaling coefficient $a$ and positioned such that it overlaps the origin, also a support is twice as big. Just a reminder, that due to the integer shifts in a typical wavelet decomposition, there is no Haar wavelet that overlaps the origin.

Assuming that we are working with Haar wavelet family let us compute the difference function wavelet coefficients $\{d^{\psi}_{j,i}\}$,

$$d^{\psi}_{j,i} = \int_{-\infty}^{+\infty} d(x)\psi_{j,i}(x)\,\mathrm{d}x = a\int_{-1}^{0} \psi_{j,i}(x)\,\mathrm{d}x - a\int_{0}^{1} \psi_{j,i}(x)\,\mathrm{d}x. \tag{7.3.29}$$

First, we investigate levels $j \geq 0$. Given the nature of Haar wavelets, with $I_{j,i} := \mathrm{supp}\{\psi_{j,i}(x)\} = [i2^{-j}, (i+1)2^{-j})$, where

$$\psi(x) = \begin{cases} 1 & 0 \leq x < 0.5, \\ -1 & 0.5 \leq x < 1, \\ 0 & \text{otherwise}, \end{cases}$$

we have that $\forall j \geq 0$ and $\forall i \in \mathbb{Z}$ there are only 3 possible possibilities:

1. $I_{j,i} \subset [-1, 0)$,
2. $I_{j,i} \subset [0, 1)$,
3. $I_{j,i} \cap [-1, 0) = \emptyset$ and $I_{j,i} \cap [0, 1) = \emptyset$,

where $\emptyset$ denotes an empty set. The first two options mean that integrals of wavelets over their full support will be equal to zero and the third option implies that wavelets defined outside the integral limits will evaluate to zero and integrals will be zero too. All in all, this leads to the fact that for $j \geq 0$,

$$d^{\psi}_{j,i} = 0. \tag{7.3.30}$$

Now we investigate resolution levels $j < 0$. We again observe that there are only three possibilities:

1. $I_{j,i} \cap [-1, 0) \neq \emptyset$ and $I_{j,i} \cap [0, 1) = \emptyset$, for $i = -1$,
2. $I_{j,i} \cap [-1, 0) = \emptyset$ and $I_{j,i} \cap [0, 1) \neq \emptyset$, for $i = 0$,
3. $I_{j,i} \cap [-1, 0) = \emptyset$ and $I_{j,i} \cap [0, 1) = \emptyset$, for $i \neq \{-1, 0\}$.

So at each resolution level $j < 0$ there will be two non-zero integrals involving wavelets with locations $i = -1$ and $i = 0$. Given the shape of a difference function and a Haar wavelet, all integrals turn out to be just areas of rectangles with fixed width of 1 and varying height $a2^{j/2}$. Therefore, integrals for all wavelets with $j < 0$ and $i = -1$ or $i = 0$ will evaluate to $-a2^{j/2}$, where the negative sign comes in from the construction of a difference function in one integral and from the negative limit in the other one. So, we have that for $j < 0$,

$$d^{\psi}_{j,i} = -a2^{j/2} \text{ for } i = -1, 0 \quad \text{and} \quad d^{\psi}_{j,i} = 0 \text{ otherwise.} \tag{7.3.31}$$

To sum up, we have,

$$d^{\psi}_{j,i} = \begin{cases} -a2^{j/2} & \text{if } j < 0 \text{ and } i = \{-1, 0\}, \\ 0 & j \geq 0. \end{cases}$$

Now let us check if the necessary condition (7.3.26) for this particular example holds. Here, we have,

$$\sum_{j \in \mathbb{Z}} \sum_{i \in \mathbb{Z}} 2^{-j/2} |d^{\psi}_{j,i}| = a \sum_{j<0} 2^{-j/2+1} |2^{j/2}|$$

given that there are only two locations per each resolution level that are non-zero,

$$= a \sum_{j<0} 2$$
$$= \infty, \text{ for } a \in \mathbb{R} \backslash \{0\}.$$

As we can see, for this particular choice of the target density we can explicitly evaluate coefficients and find out that the necessary condition is not satisfied, indicating that the choice of Haar wavelets is not allowed.

## 7.4 Implications of Besov theory on WMC

Given the results on the existence of decomposition and assumptions about wavelets of Theorem 7.2.2, what types of wavelets are we allowed to use in practice and, what are the implications on the distributions that we are allowed to use in WMC? In particular what sort of wavelet systems satisfy conditions **C1** and **C2** on page 149?

### 7.4.1 Wavelets

It was shown in Almeida (2005) and Triebel (2004) that Daubechies wavelets $\psi \in C^K(\mathbb{R}^n)$ form an orthonormal basis for the inhomogeneous Besov space $B^s_{p,q}(\mathbb{R}^n)$ with $K > |s|$. Without giving a reader a full definition of $B^s_{p,q}$ space we give a taste for the norm of this space,

$$\|f\|_{B^s_{p,q}} = \|f\|_{L^p} + \left\{ \sum_{j\geq 0} \left( 2^{js} \|\phi_j * f\|_{L^p} \right)^q \right\}^{1/q}. \tag{7.4.32}$$

As we can tell, by inhomogeneous spaces it is meant that a different norm is applied to the coarser resolution levels. However, norm-wise spaces $B^s_{p,q}$ and $\dot{B}^s_{p,q}$ are not totally different, suggesting a possibility of usage of Daubechies wavelets in the decomposition of functions $f \in \dot{B}^s_{p,q}$. From conditions **C1** and **C2**, we know we need to pick continuous and at least one time differentiable wavelets due to restriction $K > 0$, leading to the minimum requirenment of $C^1$ space. From the construction of Daubechies wavelets in Daubechies (1988) (Proposition 4.7), it is known that if Daubechies wavelet $\psi \in C^{\alpha_K}$, then it has $K$ vanishing moments and relation between $\alpha_K$ and $K$ is linear,

$$\alpha_K = \mu_K K, \tag{7.4.33}$$

with a the proportionality factor limited by $\mu_K > 0.2$. In the same paper it was shown that for $K \geq 3$, $\alpha_k > 1$, which means that Daubechies wavelets with more

that 3 vanishing moments are in fact continuously differentiable. So in our $\dot{B}^0_{1,1}(\mathbb{R})$ case, if we pick any Daubechies wavelet with $K \geq 3$ conditions **C1** and **C2** will be satisfied, where decay condition **C1** is trivially satisfied due to compactness.

We also observe that from the formulation of Theorem 7.2.2 it seems that conditions for wavelets to be a valid system for the decomposition depend on the dimensionality parameter $n$. This might imply that regularity of wavelets needs to be adapted as $n$ gets larger suggesting that smoother wavelets need to be used in WMC in high dimensions. Given that our Besov space of interest is $\dot{B}^0_{1,1}$, we have that for $n \in \mathbb{N}$, $\mathcal{T} \equiv n$, $K > s$ and $M > n + K$. So, as dimension increases we require wavelets with a more rapid decay conditions, however if we limit ourselves to working with compactly supported Daubechies wavelets, these requirements do not affect us as due to compactness the condition **C1** is always satisfied.

### 7.4.2 Densities $f$ and $g$

The convergence of a decomposition in $\dot{B}^s_{p,q}$ is considered in $\mathcal{S}'/\mathcal{P}$. It could be shown (Kyriazis 2003) that in the $\dot{B}^0_{1,1}$ setting the convergence is actually considered in $\mathcal{S}'$ and not $\mathcal{S}'/\mathcal{P}$. Even under this convergence we require our starting distribution and target to be a *smooth* function, not mentioning that derivatives have to be bounded with a rapid decay. It is clear that we have a generally unrestricted choice for a starting density and given that density of a normal distribution belongs to $C^\infty$ we have an option for a density of mixture of normals as a starting $f$, although densities of distributions like $Beta(a, b)$ and Cauchy could be chosen too. A much bigger issue comes when we are not able to determine the regularity of our target density $g$. However, given the one dimensional example in §4.1.1 we can see that even if the target is not part of $C^\infty$, WMC still performs well.

## 7.5 Summary

In this chapter we, presented how Besov spaces are connected to the necessary condition (7.3.26) in WMC. Observing that (7.3.26) is the norm of $\dot{b}^0_{1,1}$ coefficient space in disguise, we were able to relate the restriction of coefficients to the restriction of functions and wavelets being used in WMC. In particular, we presented that Haar wavelets are not a valid wavelet system for the decomposition of functions in a WMC setting via the theory of Besov spaces, uncovering the hidden assumption in WMC. Furthermore, we also showed that one of the optimal choices of the wavelet system is Daubechies with at least $K = 3$ vanishing moments, as in this case assumptions of Theorem 7.2.2 are satisfied and decomposition is possible. It was observed that for the decomposition in $\dot{B}^0_{1,1}$ to be possible we require function of interest to be in $\mathcal{S}'$, which is a strong assumption; however, we have no choice, since via the norm equivalence we know that

$$\{d^{\psi}_{j,i}\} \in \dot{b}^0_{1,1} \iff d(x) \in \dot{B}^0_{1,1}(\mathbb{R}),$$

and we must have $\{d^{\psi}_{j,i}\} \in \dot{b}^0_{1,1}$ for the necessary condition (7.3.26) to hold. Although, theoretically assumption on a difference function being part of $\mathcal{S}'$ is strong, we practically saw in §4.1.1 that even having a not continuously differentiable function everywhere can lead to good WMC results, suggesting that restrictions on $d(\cdot)$ could be relaxed. However, this claim requires additional research and could be a topic of interest in the future.

# Chapter 8

# Modified WMC

In this chapter, two possible improvements to the WMC algorithm will be presented. The first one, described in §8.1, will be the *Multiple Importance Sampling WMC* (MIS-WMC), where the goal is to not discard intermediate samples $x_t \sim f_t(\cdot)$, $t \in [0, 1)$, produced by the WMC, but to save them for use in the future computation of moments of the target density $g(\cdot)$.

In Section 8.2, we discuss a second modification of a standard WMC, a *Level WMC* algorithm (LWMC), where the goal is to approach the target sample $y \sim g(\cdot)$ by sequentially moving samples up the resolution ladder.

## 8.1 Multiple Importance Sampling WMC

### 8.1.1 Motivation

During a typical run of the WMC algorithm, for each starting $x_0 \sim f(\cdot)$, there will generally be several intermediate points $x_t \sim f_t(\cdot)$, $t \in [0, 1)$, sampled before reaching a target $y \sim g(\cdot)$. In the standard WMC, these intermediate points act only as a pit-

stop points to recalculate parameters necessary to continue WMC, sample a survival time $s \geq t$ and potentially a new point $x_s$. After a new point is sampled, previous values and calculations of wavelet coefficients $\hat{d}^{\psi}_{j,i}$ and parameters of the generalised Pareto survival distribution are discarded. Due to the high computational load required to reach target samples, as discussed in Chapter 5, it is important to efficiently utilise every computation performed during each WMC run.

The key idea in MIS-WMC is not to discard intermediate values $x_t$ but store them for the computation of moments of the density $g(\cdot)$ later. Given the generally costly production of target samples via WMC, intermediate points with a survival time close to $t = 1$ could be involved in the estimation of moments with an appropriate weighting scheme applied. This would partially mitigate the issue of small samples at the cost of error and variance introduced in estimation of moments.

In the conventional IS algorithm, it is crucial to pick a good covering distribution from which sampling will be performed, similar to picking a suitable envelope distribution in rejection sampling. If, in WMC, we are able pick a starting density $f(\cdot)$ that closely resembles the target, then all intermediate densities $f_t(\cdot)$ will be good approximations of the target $g(\cdot)$, leading to a large number of valuable intermediate samples that could be used in the analysis of the target density.

Due to the fact that samples from several different intermediate distribution will be used in the construction of estimators of moments, the *Multiple Importance Sampling* (MIS) method (Veach & Guibas 1995, Veach 1997) will be adopted to accommodate this. Before going into more details of how MIS and WMC could be used together, an overview of the MIS methodology will be presented, mainly focusing on the construction of MIS estimators.

### 8.1.2 MIS estimator

Here the construction of the MIS estimator will be outlined. In the MIS setting, samples are produced from several distributions rather than from a single one as in IS. The change from one to several distributions leads to the extra layer of complexity when samples need to be combined and even potentially weighted.

We denote $r$ to be the number of densities $f_k(\cdot)$ used, i.e. $k \in \{1, 2, ..., r\}$, i.e. where $k$ indexes intermediate distributions and is not related to time parameter $t$. Also $r$ here is used only locally and should not be confused with ratio of normalising constants. Let $N_k$ be the total number of samples $\{x_{n,k}\}_{n=1}^{N_k}$ produced from a distribution $f_k(\cdot)$, and let $N = \sum_{k=1}^{r} N_k$ be the total number of samples across all distributions. Now we define a MIS estimator for $\int g(x)\,\mathrm{d}x$, which is useful if $g(x)$ is an unormalised density,

$$G = \frac{1}{N}\sum_{k=1}^{r}\sum_{n=1}^{N_k}\frac{g(x_{n,k})}{f_k(x_{n,k})}. \tag{8.1.1}$$

We show that $G$ is unbiased,

$$\begin{aligned}\mathbb{E}[G] &= \frac{1}{N}\sum_{k=1}^{r}\sum_{n=1}^{N_k}\int \frac{g(x)}{f_k(x)} f_k(x)\,\mathrm{d}x \\ &= \frac{1}{N}\sum_{k=1}^{r} N_k \int g(x)\,\mathrm{d}x \\ &= \int g(x)\,\mathrm{d}x,\end{aligned}$$

using $N = \sum_{k=1}^{r} N_k$ and assuming $f_k(x) \neq 0, \forall x,\ \forall k \in \{1, 2, ..., r\}$. In a very straightforward way G might be used to estimate moments of $g(x)$, since

$$\mathbb{E}_g[x^m] \approx \frac{1}{GN}\sum_{k=1}^{r}\sum_{n=1}^{N_k} x_{n,k}^m \frac{g(x_{n,k})}{f(x_{n,k})}. \tag{8.1.2}$$

The key issue with this estimator is that samples are not being weighted and therefore this form of the estimator is not immediately applicable to WMC as it

does not address the importance of samples from the intermediate distribution, $f_k(\cdot)$. Due to the fact that we are not able to control the sampling procedure of the WMC algorithm, it is important to try to weigh samples correctly. Furthermore, samples that are produced from intermediate distributions $f_t(\cdot)$ when $t \approx 1$ are much more important than those with $t \approx 0$. Therefore, in the next section, a weighted form of the estimator will be presented that addresses the importance of samples that are closer to the target $g(\cdot)$.

### 8.1.3 Weighted MIS

Due to the nature of the process that generates WMC we are not able to control how many samples from which intermediate distributions are going to be produced. Therefore, it is important to build an estimator that prioritises samples that were drawn from distributions closer to the target $g(\cdot)$. Let $w_k(\cdot)$ be a weighting function that gives a weight to samples $\{x_{n,k}\}_{n=1}^{N_k} \sim f_k(\cdot)$. Our weighted estimator of $\int g(x)\,\mathrm{d}x$ is parametrised by a set of functions $w_1(\cdot), ..., w_r(\cdot)$; in particular,

$$G_w = \sum_{k=1}^{r} \frac{1}{N_k} \sum_{n=1}^{N_k} w_k(x_{n,k}) \frac{g(x_{n,k})}{f_k(x_{n,k})}. \tag{8.1.3}$$

If we assume that $\sum_{k=1}^{r} w_k(x) = 1$ and $w_k(x) = 0$ whenever $f_k(x) = 0$, then estimator $G_w$ becomes unbiased.

Following ideas from Veach & Guibas (1995), consider the weight function

$$\hat{w}_k(x) = \frac{c_k f_k(x)}{\sum_k c_k f_k(x)}, \tag{8.1.4}$$

with $c_k = N_k/N$. Then it can be proved that this estimator is "almost" optimal in the sense that one cannot improve much on the variance of $G$ if one chooses other $\hat{w}_k(x)$.

**Theorem 8.1.1** (Veach and Guibas, 1995)**.** *Let $w_1(x), ..., w_r(x)$ be any non-negative functions with $\sum_k w_k(x) = 1$, $\forall x \in \mathbb{R}$ and let $\hat{w}_1(x), ..., \hat{w}_r(x)$ be the weight functions*

*defined in (8.1.4). Let $G_w$ be an estimator of the form given in (8.1.3) and $\hat{G}_w$ be the estimator (8.1.3) using $\hat{w}_k(x)$ described in (8.1.4). Then*

$$Var[\hat{G}_w] \leq Var[G_w] + \Big(\frac{1}{\min_k N_k} - \frac{1}{\sum_k N_k}\Big)\mathcal{G}^2,$$

*where*

$$\mathcal{G} = \int_{-\infty}^{+\infty} g(x)dx.$$

This theorem says that no choice of $w_k(x)$ can improve upon the variance of the estimator defined by (8.1.4) by more than $\big(1/\min_k N_k - 1/\sum_k N_k\big)\mathcal{G}^2$. This variance difference is very small relative to the variance caused by a poorly chosen sampling distribution.

We can modify this estimator proposed by Veach and Guibas (1995), to include the time parameter to prioritise samples coming from distributions that are closer to the target. This type of modification would address WMC directly. To be more clear while explaining MIS-WMC, we will deviate from our standard notation of $f_t(\cdot)$ for intermediate distributions and will use $f_{t_k}(\cdot)$, where $t_k \in [0,1]$, however now we are able to index intermediate densities $f_{t_k}(\cdot)$ with $k \in \mathbb{N}_0$. We will also have that $f_{t_0}(\cdot) \equiv f(\cdot)$ and $f_{t_{r+1}}(\cdot) \equiv g(\cdot)$. It will be demonstrated later in §8.1.5 that in MIS-WMC we have the case that $N_k = N - 1\ \forall k$, for this reason we have $c_k \equiv \frac{N-1}{N}$ and this will simplify the form of the estimator. Using this new notation we introduce a weighting scheme adapted for MIS-WMC,

$$\tilde{w}_k(x) = \frac{t_k f_{t_k}(x)}{\sum_l t_l f_{t_l}(x)}. \tag{8.1.5}$$

Weighting scheme (8.1.5) assigns more weight to samples that are produced from distributions with greater $t_k$ value, i.e. distributions that are closer to the target. Using this weighting method, the estimator remains unbiased as the unity criteria $\sum_k \tilde{w}_k(x) = 1$ still holds together with $\tilde{w}_k(x) = 0$ when $f_{t_k}(x) = 0$. We will refer to estimator $G_w$ with the weighting scheme $\tilde{w}_k(x)$ as $\tilde{G}_w$.

The dependence of intermediate samples of a single WMC run complicates the variance estimation problem. Although samples across different WMC runs are independent, intermediate samples within the same WMC run are dependent due to the inherent Markov chain structure. In particular, if a fine scale wavelet is selected to generate new intermediate point, it will have a positive correlation with an old one, however if the coarse scale wavelet is selected it is very likely that the sampled new point will be far from the old one and the correlation will be negative. The overall correlation structure between intermediate points is very complicated and will not be addressed here in detail. However, even assuming the overall independence across all samples, the closed form for the variance of $\tilde{G}_w$ is still intractable. For convenience, let us define

$$\mu_k = \int \tilde{w}_k(x) g(x)\, \mathrm{d}x. \tag{8.1.6}$$

Now let us try to get a closed form for the variance of the estimator, conditioning on the set of intermediate distributions $\{t_k\}$,

$$\mathrm{Var}[\tilde{G}_w|\{t_k\}] = \sum_{k=1}^{r} \frac{1}{N-1} \mathrm{Var}\Big[\tilde{w}_k(x_{n,k}) \frac{g(x_{n,k})}{f_{t_k}(x_{n,k})}\Big]$$

assuming the independence over all intermediate samples,

$$\begin{aligned}
&= \sum_{k=1}^{r} \frac{1}{N-1} \int \frac{\tilde{w}_k(x)^2 g(x)^2}{f_{t_k}(x)}\, \mathrm{d}x - \sum_{k=1}^{r} \frac{1}{N-1} \mu_k^2 \\
&= \sum_{k=1}^{r} \frac{1}{N-1} \int \frac{t_k^2 f_{t_k}(x) g(x)^2}{\big(\sum_l t_l f_{t_l}(x)\big)^2}\, \mathrm{d}x - \sum_{k=1}^{r} \frac{1}{N-1} \mu_k^2 \\
&= \frac{1}{N-1} \Big( \int \frac{\sum_k t_k^2 f_{t_k}(x)}{\big(\sum_l t_l f_{t_l}(x)\big)^2} g(x)^2\, \mathrm{d}x - \sum_{k=1}^{r} \mu_k^2 \Big).
\end{aligned}$$

The involvement of $t_k$ restricts the obvious simplification that would be possible otherwise in the expression of the variance above. The intractability of the functional form of the variance means that we require a numerical approach which we shall consider in §8.1.7.

### 8.1.4 Other types of weightings

Weighting scheme (8.1.5) could be modified further to suit particular WMC settings. The following types of weighting are appropriate for scenarios when even more weight should be put towards samples closer to the target density. These methods are relevant when the starting $f(\cdot)$ density is significantly different in shape and/or location from the target.

1. *Cut-off* method. Discard samples with low weight:

$$w_k(x) = \begin{cases} 0 & \text{if } t_k f_{t_k}(x) < \alpha f_{\max}(x), \\ \frac{t_k f_{t_k}(x)}{\sum_l \{t_l f_{t_l}(x) | t_l f_{t_l}(x) \geq \alpha f_{\max}(x)\}} & \text{otherwise.} \end{cases}$$

   where $f_{\max}(x) = \max_k t_k f_{t_k}(x)$. The constant $\alpha \in [0, 1]$ determines how small $t_k f_{t_k}(x)$ must be compared to $f_{\max}(x)$ before we assign it a zero weight.

2. *Power* method. Raise all weights to a power, $\beta > 1$, and then normalise:

$$w_k(x; \beta) = \frac{(t_k f_{t_k}(x))^\beta}{\sum_l (t_l f_{t_l}(x))^\beta}.$$

3. *Time threshold* method.

$$w_k(x; \lambda) = \begin{cases} 0 & \text{if } t_k < \lambda, \\ \frac{t_k f_{t_k}(x)}{\sum_l t_l f_{t_l}(x)} & \text{otherwise.} \end{cases}$$

With the cutoff method, initial samples and samples from intermediate distributions that have barely moved from the starting density will be completely discarded from the estimation of moments, similar to the 'burn-in' process of MCMC methods.

The power method allows for the precise control of weights. Not only are sample points closer to the target given more weight but through the choice of $\beta \in \mathbb{R}$ we can control how much more weight is assigned to samples. In the limit, as $\beta \to \infty$, we essentially restrict ourselves to only using samples from the target $g(x)$ in the estimation of moments,

$$\lim_{\beta\to\infty} w_k(x;\beta) = \begin{cases} 1 & \text{if } t_k = 1, \\ 0 & \text{otherwise.} \end{cases}$$

However relaxing this parameter we allow for the inclusion of samples that come from distributions that are relatively close to the target.

Similar to the Cutoff method, the Time threshold one disregards all intermediate distributions $f_{t_k}(x)$ and samples associated with them if $t_k < \lambda$, where $\lambda \in (0, 1]$ is the time threshold parameter. Given the usually large number of intermediate distributions created throughout MIS-WMC process it is important to have an option to focus on only distributions that are closer to the target.

Furthermore, these heuristics could be modified even further to reflect particular WMC scenarios, for example it would also be possible to combine the Cutoff or Time threshold with the Power one.

### 8.1.5 Controlling samples from intermediate distributions

In a conventional MIS set up, we have several importance densities picked in advance from which sampling is going to be performed directly to estimate moments of a target distribution. However, in the WMC scenario, we are not able to pick intermediate distributions a priori and sample from them directly; the sampling procedure from intermediate distributions is uncontrolled and determined by a random process. Nonetheless, from the WMC theory (§3.4.2) we know that if a given point $x_s$ at a time $t = s$ has an associated survival time $t = t^\star$, then $x_s$ could be treated as a sample from any distribution between $f_s(\cdot)$ and $f_{t^\star}(\cdot)$ excluding the density at $t = t^\star$,

$$x \sim f_l(\cdot), \quad s \leq l < t^\star, \tag{8.1.7}$$

which means that sample $x$ is a representative sample from all intermediate distributions between $f_s(\cdot)$ and $f_t(\cdot)$, $t > s$, excluding the density at time $t$. Figure

8.1 demonstrates this process for a starting point $x_0$. Firstly we sample $x_0 \sim f(\cdot)$ at time $t = 0$, secondly we sample a survival time $t^\star$ after which we would sample a new point $x_{t^\star} \sim \psi_{j,i}(x)$ if $t^\star < 1$. Having observed that point $x_0$ existed at all times $0 \leq s < t^\star$, we conclude that $x_0 \sim f_s(x)$ for any $0 \leq s < t^\star$.

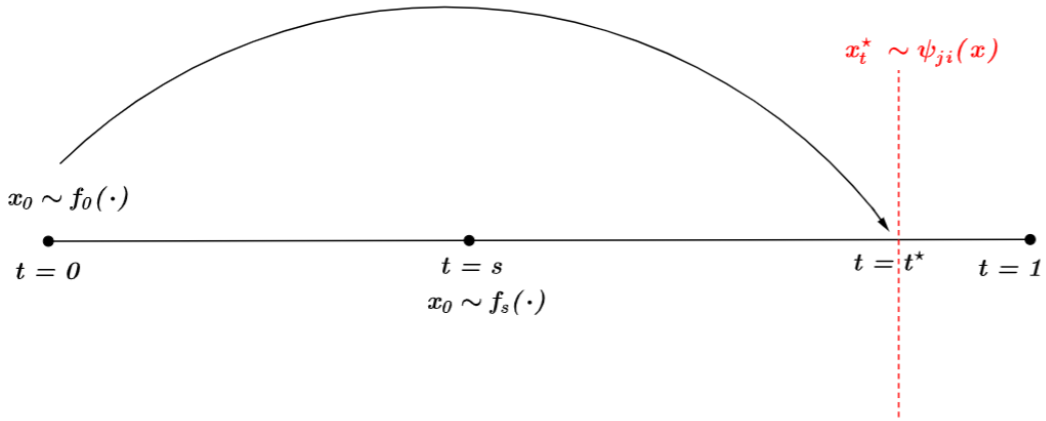



Figure 8.1: *Diagram showing how randomly sampled intermediate points in a WMC are going to be assigned to a distribution. Point* $x_0$ *had a survival time* $t = t^\star$*, where* $0 \leq s < t^\star$*, hence we conclude* $x_0 \sim f_l(\cdot)$*,* $0 \leq l < t^\star$*.*

The question remains, how to decide to which $f_t(\cdot)$ distribution intermediate sample points should be assigned to during the full run of WMC for a starting sample size of $N$ points from $f(\cdot)$.

The idea is to create checkpoints $t_k$ with each single WMC run, which will indicate the intermediate distributions $f_{t_k}$ to which points $x_s$, $s \in [0, 1)$, should be assigned to. For the first sample $x \sim f(\cdot)$ a survival time $t$ is sampled and if $t < 1$ a new point $x^\star \sim \psi_{j,i}(\cdot)$ is sampled according to the WMC algorithm. The sampled survival time $t$ becomes a checkpoint created by the initial point from a starting distribution $f(\cdot)$, after this a survival time $t^\star$ for the point $x^\star$ is sampled and if $t^\star < 1$ we record $t^\star$ as another checkpoint and carry on until we sample a survival time greater than one.

So, each starting point $x_k \sim f(\cdot)$ and its associated intermediate points will create a set of checkpoints $t_{k,l^{(k)}}$, where $k \in \{1, 2, ..., N\}$ indicates at which run the checkpoint was created and $l^{(k)} \in \mathbb{N}$ indicates the $l$-th checkpoint in $k$-th WMC run. Therefore, after the total of $N$ runs we will end up with a pooled collection of checkpoints $\{t_{k,l^{(k)}}\}$, where $k \in \{1, 2, ..., N\}$ and $l^{(k)} = 1, ..., l^{(k)}_{\max}$. It could be the case that no checkpoints are created in the $k$-th run, in that case we would have $l^{(k)}_{\max} = \emptyset$ and $t_{k,l^{(k)}} = \emptyset$. Checkpoint creation procedure could be inspected in Figure 8.2.

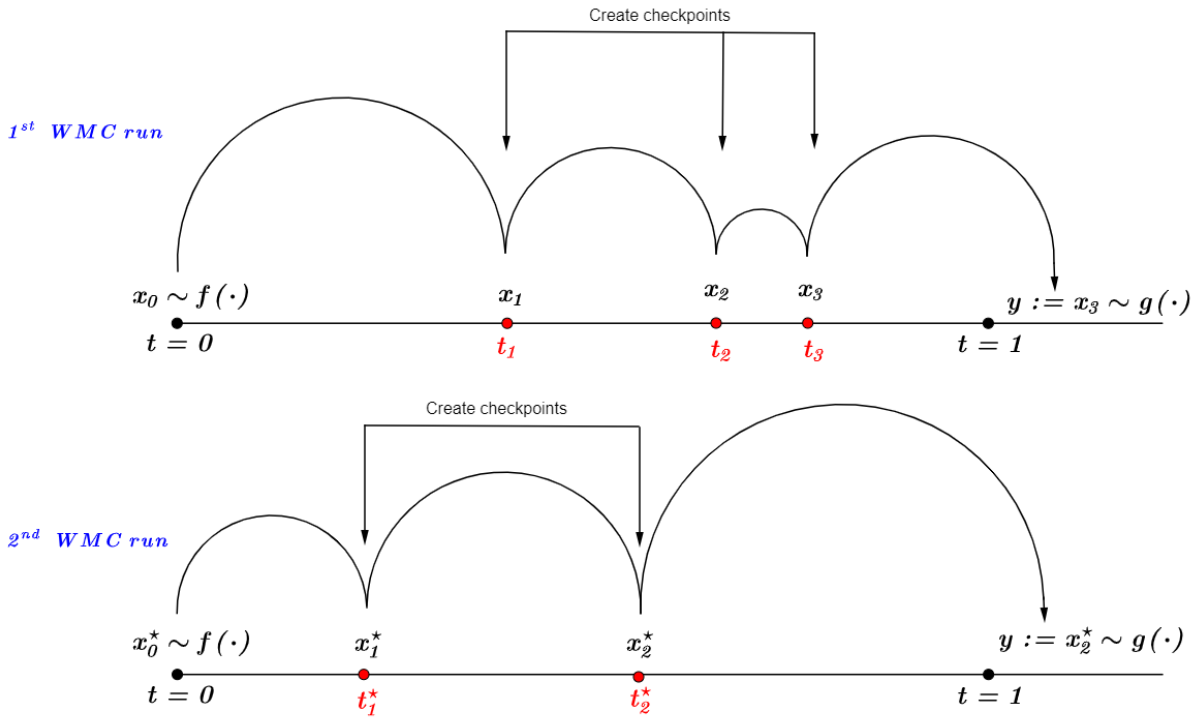



Figure 8.2: *Illustrating how checkpoints are created over several WMC runs. With each run new checkpoints are created then pooled into a single collection.*

Having created all the checkpoints, we next allocate points to intermediate distributions. Given any starting point $x_n \sim f(\cdot)$, where $n \in \{1, ..., N\}$, and its associated intermediate points that were created in $n$-th run, the allocation process is as follows:

1. Given a point, observe its initial time $t_I$, so for $x_n \sim f(\cdot)$ we have $t_I = 0$, we also take note of a survival time of $x_n$ which let us say is $t_{n,1} < 1$.

2. From the full collection of checkpoints $\{t_{k,l^{(k)}}\}$, $k \in \{1, 2, ..., N\}$, $l^{(k)} = 1, ..., l^{(k)}_{\max}$ we discard all checkpoints created by $n$-th run to create a new sub-collection of checkpoints with $k \neq n$, $\{t_{k,l^{(k)}}\}_{k \neq n}$

3. We allocate point $x_n$ to all intermediate distributions $f_{t_{k,l^{(k)}}}(\cdot)$ for which the inequality $t_I \leq t_{k,l^{(k)}} < t$ is satisfied, where $t_{k,l^{(k)}} \in \{t_{k,l^{(k)}}\}_{k \neq n}$.

The same exact steps above are taken in allocating intermediate points $x \sim \psi_{j,i}(x)$.

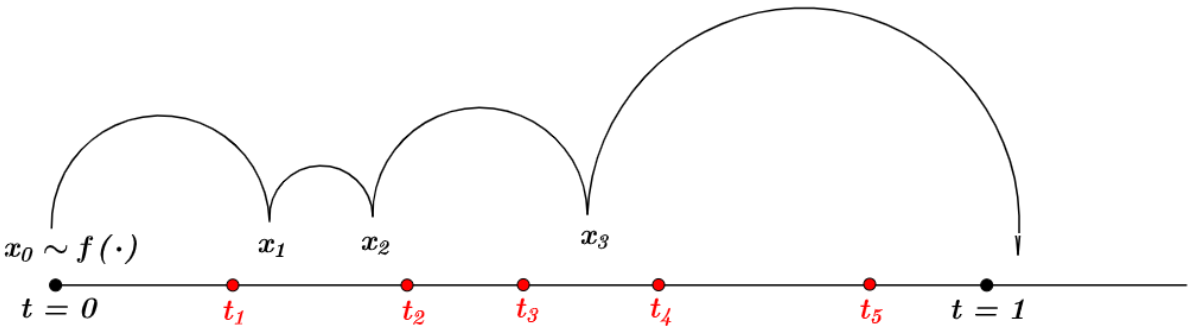



Figure 8.3: *After creating a full collection of checkpoints after $N$ runs, each starting point $x_0 \sim f(\cdot)$ and associated intermediate points $x \sim \psi_{j,i}(\cdot)$ are allocated to intermediate distribution based on those checkpoints that the point has survived through. The point $x_0$ has survived past the time $t_1$ and hence is assigned to $f_{t_1}(\cdot)$. On the other hand, the point $x_1$ is not assigned to any intermediate distribution because there are no checkpoints in between initial time and survival time to which this point could be allocated. Furthermore, points could be allocated to several intermediate distributions at the same time, points $x_2$ and $x_3$ both survive through two checkpoints and hence are assigned to both intermediate distributions.*

Using the method described above, after a full WMC run of $N$ sample points from

| Samples in Figure 8.3 | |
|---|---|
| $x_0 \sim f(\cdot)$ | $x_3 \sim f_{t_4}(\cdot)$ |
| $x_0 \sim f_{t_1}(\cdot)$ | $x_3 \sim f_{t_5}(\cdot)$ |
| $x_2 \sim f_{t_2}(\cdot)$ | $y := x_3 \sim g(\cdot)$ |
| $x_2 \sim f_{t_3}(\cdot)$ | |

Table 8.1: *Table summarising the samples produced in Figure 8.3. In addition to a starting sample $x_0 \sim f(\cdot)$ and a target sample $y \sim g(\cdot)$, there was exactly one point assigned to every intermediate distribution.*

a starting distribution we end up with:

1. $\{x_i\}_{i=1}^{N} \sim f(x)$
2. $\{y_i\}_{i=1}^{N} \sim g(x)$
3. $\{x_{n,k}\}_{n=1}^{N-1} \sim f_{t_k}(x)$, for $k = \{1, ..., r\}$, where as before, $r$ is the total number of intermediate distributions used (checkpoints created) and $x_{n,k}$ is the $n^{th}$ sample from the distribution $f_{t_k}(x)$.

As we can see in Figure 8.3, due to a continuity of the time parameter $t$ each checkpoint needs to be passed exactly one time in each WMC run; this means that if we start with $N$ samples from a starting distribution, there are going to be $N-1$ points assigned to every intermediate distribution that was defined by a checkpoint. There are going to be $N-1$ samples because as described in the allocation process above, when allocating intermediate sample point to intermediate distributions, checkpoints that were created from that particular WMC run are not being used, hence leaving us with $N-1$ samples for each intermediate distribution.

There also exists a possibility to predefine checkpoints in advance, manually. The manual grid selection of checkpoints would significantly reduce the total number

of intermediate distributions used in construction of the estimator $\tilde{G}_w$ and would reduce the correlation present across samples from $f_t(\cdot)$ and $f_s(\cdot)$ where $t \approx s$, i.e. $s$ and $t$ are almost equal. On the other hand, manual selection of checkpoints assumes that user has knowledge of distribution of survival points and can select checkpoints in a meaningful manner. The dynamic allocation of checkpoints presented in this section is not uniform and is highly influenced by the discrepancy present between starting distribution $f(\cdot)$ and the target $g(\cdot)$. If $f(\cdot)$ and $g(\cdot)$ are highly similar it is expected that checkpoints could be more concentrated towards $t = 1$ and therefore a uniform grid would not be a meaningful way of creating checkpoints as a lot of information would be wasted and not directed towards more accurate computation of $\tilde{G}_w$.

A thinned out, informative grid could be constructed after checkpoints have been collected and analysed. The idea would be to reduce the number of checkpoints on the original grid but still maintain the overall distribution and structure created on the original grid. In this way the grid would still represent patterns where points usually tend to get extinct but also it would be coarse enough to mitigate the present correlation between points that were assigned to several intermediate distributions.

### 8.1.6 Ghost points in MIS-WMC

Taking issues described in Section 5.4 into consideration, how one would deal with the inevitable presence of ghost points in a MIS-WMC scheme? At first glance, the creation of ghost points $x_g$ might seem a severe problem that would contaminate intermediate samples $x_t \sim f_t(\cdot), t \in [0, 1)$, and would ruin the possibility of including them in the estimation of moments of the target density. Fortunately, the associated survival time 0 for a ghost point $x_g$ at time $t = s$ essentially means that

$$x_g \sim f_l(\cdot), \quad s \leq l < s, \tag{8.1.8}$$

but $\{l : s \leq l < s\} = \emptyset$. Slightly abusing the mathematical notation in (8.1.8), we demonstrate the importance of the survival time associated with the sampled intermediate point. A sampled intermediate point $x_t$ with a survival time zero does not belong to any intermediate distribution and is automatically discarded. Therefore, the ghost point problem is almost surely not an issue in the MIS-WMC set up.

Nevertheless, there might still be some problems surrounding semi-ghost points, points that have a small but non-zero survival time (§5.4). The allocation of an intermediate point to an intermediate distribution involves the criterion (§8.1.5) that checks if there is a checkpoint between an initial time $t_I$ of a point and a final time $t$ when it dies out, if a checkpoint or several checkpoints have been detected the point is assigned to all intermediate distributions associated with those checkpoints. However, if no checkpoints are found, the point is deemed to be a ghost point/semi-ghost point of no value and is discarded. Given the extremely short survival time of a semi-ghost point, the probability of assigning those points to any intermediate distribution is very small. A probability of assigning a semi-ghost point to a distribution could be viewed as trying to sample two identical points from $\mathcal{U}(0, 1)$. Although the analogy is not perfect as the probability of assigning a semi-ghost point to an intermediate distribution is positive, practically this event never happens and if it does it could be easily detected and dealt with.

### 8.1.7 Numerical analysis

To investigate properties of the estimator produced by MIS-WMC algorithm, the starting distribution was chosen to be $\mathcal{U}(-10, 10)$ and the target was set to be a mixture of standard distributions as in the one-dimensional example of §4.1.1, this set-up is visualised in Figure 8.4.

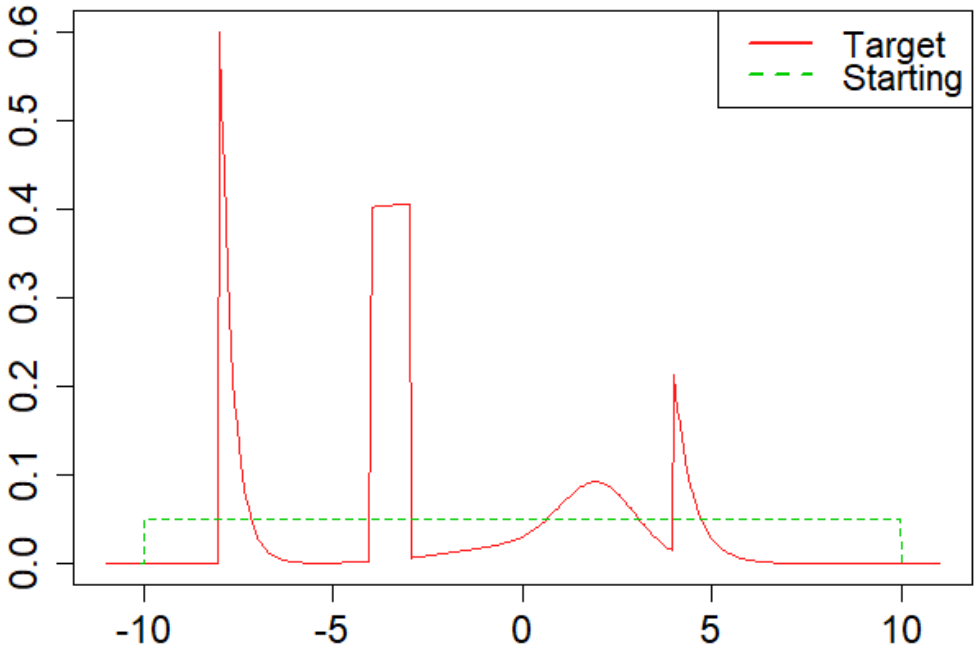



Figure 8.4: *Starting and target densities for the MIS-WMC numerical analysis.*

The starting density $f(\cdot)$ was chosen such that it covers the target density $g(\cdot)$ and is similar in location. The main idea for doing this is that intermediate densities $f_{t_k}(\cdot)$ will be covering the high density areas of $g(\cdot)$ and produced samples from these intermediate densities will be of more value. If $f(\cdot)$ is chosen to be significantly different in location from the target $g(\cdot)$, then a significant amount of time is required for a starting distribution to transform into something of a similar shape and location as the target. This would imply that samples produced from intermediate distributions that are closer to the starting one will be of substantially lesser value.

In this particular example, after MIS-WMC is performed we end with a sample of size $N = 500$ from the target $g(\cdot)$ and with 2276 intermediate distributions between $t = 1$ and $t = 1$ each with 499 samples. Given that throughout 500 WMC runs, 2276 checkpoints were created it is important to investigate the difference between samples. The reason for doing this is, if there is a cluster of checkpoints,

an intermediate point will survive through all of them and will be assigned to each intermediate distribution of each of those corresponding checkpoints. This means that majority of samples from intermediate distributions are sharing the same sample points, making an effective sample size smaller.

We examine the *sample similarity* by constructing a percentage based index $S_\beta(\alpha)$, $\alpha, \beta \in (0, 1)$ and $\alpha \leq \beta$ that measures what percentages of samples of $f_\alpha$ distribution is identical to those of $f_\beta$.

$$S_\beta(\alpha) = \frac{\gamma(t_k = \beta, t_l = \alpha)}{N - 1} \times 100\%, \tag{8.1.9}$$

where $\gamma(t_k, t_l)$ is the function which returns the number of duplicate samples between $f_{t_k}$ and $f_{t_l}$, with $t_k \geq t_l$. Fixing $\beta$, we can investigate how similarity between samples changes as we keep reducing $\alpha$ to 0. As we can see (Figure 8.5), the sample similarity percentage decays quite slowly with each 'lag', which means that given any distribution $f_{t_k}$ and two neighbouring distributions $f_{t_{k-1}}$ and $f_{t_{k+1}}$, the samples associated with each of those distributions are almost the same, with only few sample points being unique for each distribution. This observation suggests that it is relevant to include thinning options before using all intermediate distributions in the computation of statistics using MIS-WMC.

Furthermore, we can see that approximately 15% of samples from a starting distribution ended up surviving to the target, i.e. no intermediate jumps were required in those cases to generate a sample from the target. The efficiency of the WMC run on $N$ points could be also judged on the amount of starting points that were not required to do any intermediate jumps. For this reason, it is important to pick the best possible starting distribution which would be similar in shape and location.

Figure 8.6 presents samples from couple of intermediate distributions that were created using the checkpoint procedure. Not ignoring the intermediate points

produced by WMC, we ended up with 2276 intermediate distributions each containing 999 samples, which could be used for statistic computation purposes of the target distribution.

To analyse distribution properties of the MIS-WMC mean estimator $M_w$,

$$M_w(\lambda) = \frac{1}{\mathcal{G}} \sum_{k=1}^{r} \frac{1}{N_k} \sum_{n=1}^{N_k} w_k(x_{n,k}; \lambda) x_{n,k} \frac{g(x_{n,k})}{f_k(x_{n,k})}, \tag{8.1.10}$$

we will focus on using the time threshold weight function $w_k(x; \lambda)$. To explore the variance of the estimator, we will set up a simulation that will run 100 times for $N = \{10, 25, 100, 1000\}$ and for each of $N$ we will record the $M_w(\lambda)$ estimator value

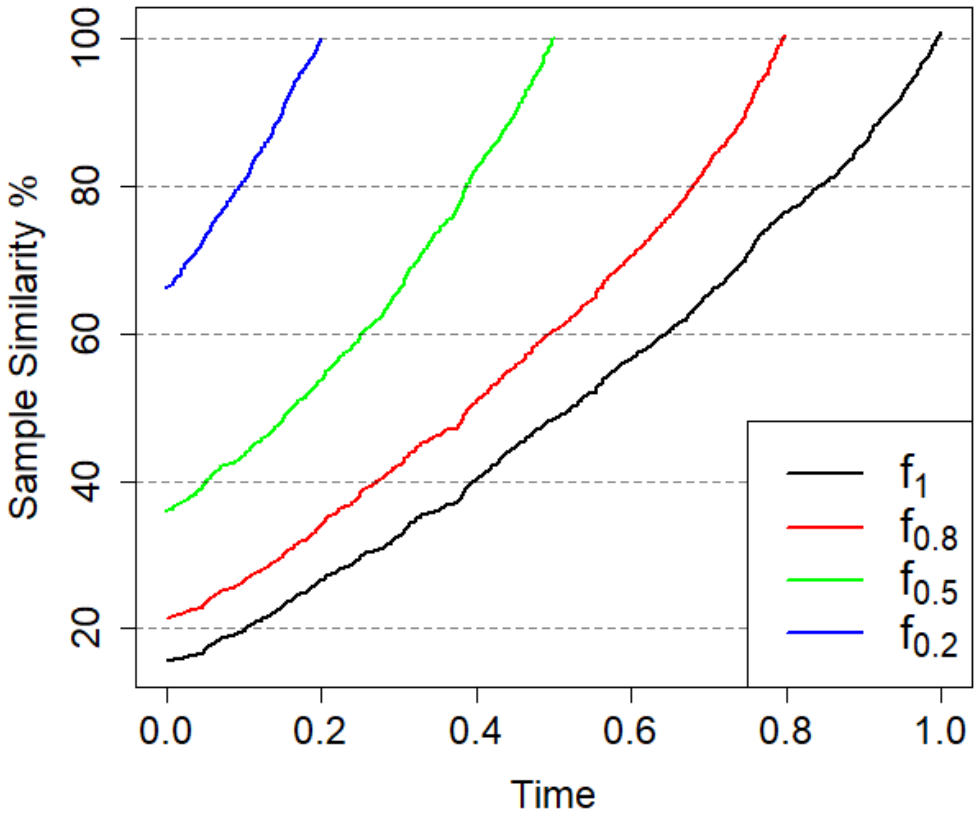



Figure 8.5: *Each trace indicates at what Time (intermediate distribution) what the present sample similarity value is. For example, $f_1$ trace indicates that approximately 75% of samples of $f_{0.8}$ distribution are identical to samples from $f_1$.*

for $\lambda = \{0.8, 0.9, 0.95\}$. The goal of the simulation is to spot what effect the $\lambda$ value together with $N$ has on the variance of the estimator. As expected the empirical mean of the estimator $M_w(\lambda)$ converges to the true value of the target distribution. For $\lambda = 0.95$ the mean of the estimator seems to be the most consistent around the true value, however it has the highest variance for $N = 10$ but lowest one for $N = 25$.

As it could be seen in Figure 8.7, as value of $N$ increases, the benefits of using MIS-

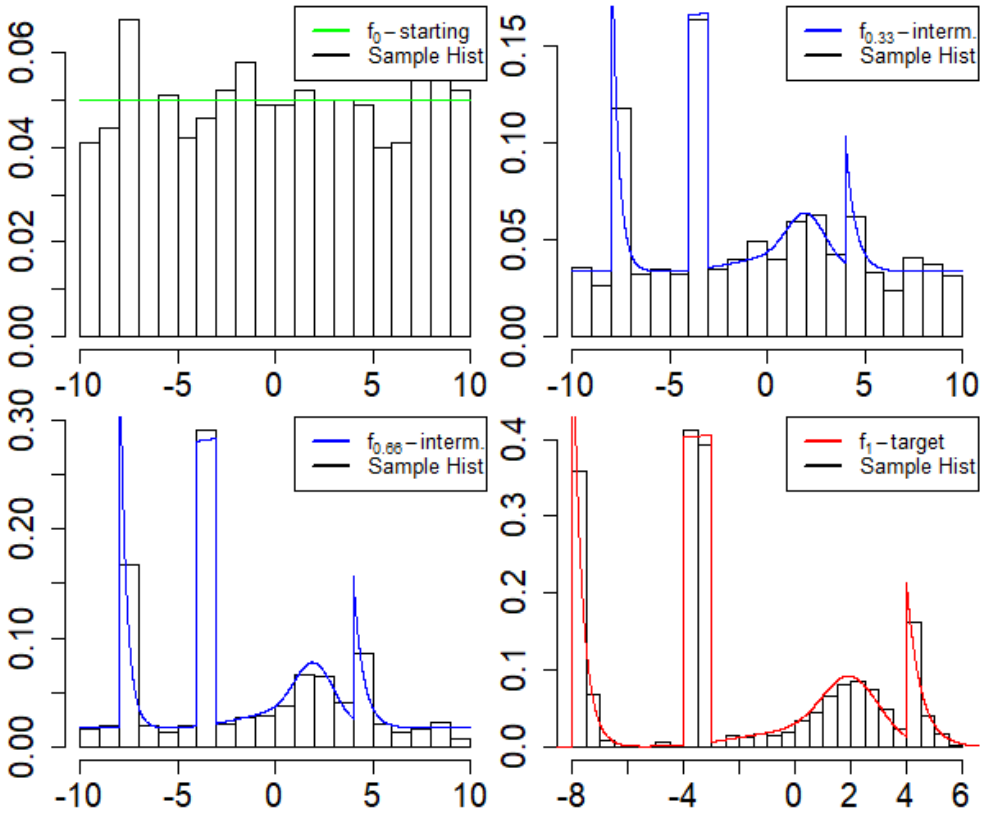



Figure 8.6: *By not discarding intermediate samples and using the intermediate sample allocation procedure described in §8.1.5 we are able to produce samples from intermediate distributions. The figure presents histograms of samples from the starting distribution, two intermediate distributions and the target based on* $N = 1000$ *points from a starting distribution.*

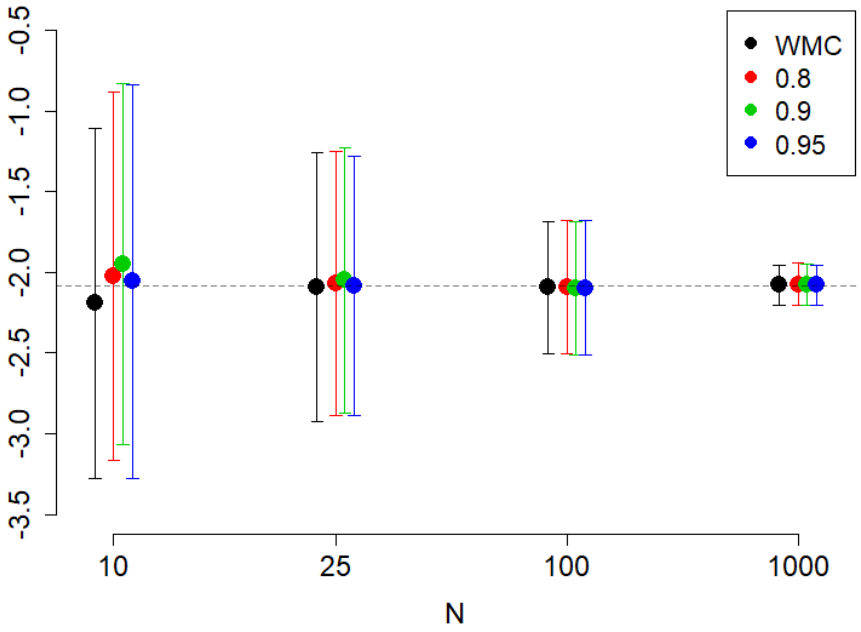



Figure 8.7: *Empirical mean of the estimators is plotted after 100 simulation runs together with error bars. Dashed line indicates the actual mean of the target distribution.*

WMC are not that clear, however MIS-WMC is doing arguably better than a default mean of the pure WMC for lower values of $N$. For the small price of the increase in variance, MIS-WMC should be utilised in the situations when computational cost of producing samples from the target is high. In those cases, intermediate samples could play a big role of producing better estimators at almost no additional cost.

## 8.2 Level WMC

### 8.2.1 Motivation

In this section, we introduce another possible modification to the standard WMC, a Level WMC (LWMC). The key motivation for this sort of algorithm is to minimise

the total number of wavelet coefficients $d^{\psi}_{j,i}$ that are required for estimation at each iteration of WMC run. Furthermore, we are also interested in incorporating the approximate knowledge of the normalisation constant of the target and introducing the possibility of improving the quality of samples even after a target sample has been reached via WMC.

As it was mentioned before in §3.4.3, proofs for the validity of the WMC theory assume that a summation over all resolution levels $j \in (-\infty, +\infty)$ could be performed, in practice we must restrict ourselves to some coarsest $j_{\min}$ and some finest $j_{\max}$ resolution levels. At this point it is clear that the moment this restriction is made the samples produced by the WMC algorithm change from $g(x)$ to being from

$$\hat{g}(x) = \sum_{j=j_{\min}}^{j_{\max}} \sum_{i} g^{\psi}_{j,i}\psi_{j,i}(x) + \sum_{j<j_{\min}} \sum_{i} f^{\psi}_{j,i}\psi_{j,i}(x) + \sum_{j>j_{\max}} \sum_{i} f^{\psi}_{j,i}\psi_{j,i}(x) \qquad (8.2.11)$$

at best. This 'best case' scenario occurs when we have a good estimate of the ratio of normalising constants and good estimates of wavelet coefficients, we will refer to equation (8.2.11) as a practical target of WMC. However, mother wavelets $\psi_{j,i}(\cdot)$ are not able to capture the 'average' behaviour of a function. In particular, if we try to compute the expectation of $\hat{g}(\cdot)$, assuming that $\hat{K}_g$ is the normalisation constant, we get

$$\frac{1}{\hat{K}_g}\int_{-\infty}^{+\infty} x\hat{g}(x)\mathrm{d}x = \int_{-\infty}^{+\infty} \sum_{j<j_{\min}} \sum_{i} f^{\psi}_{j,i} x\psi_{j,i}(x)\,\mathrm{d}x + \int_{-\infty}^{+\infty} \sum_{j>j_{\max}} \sum_{i} f^{\psi}_{j,i} x\psi_{j,i}(x)\,\mathrm{d}x \qquad (8.2.12)$$

where we used,

$$\sum_{j=j_{\min}}^{j_{\max}} \sum_{i} g^{\psi}_{j,i} \int_{-\infty}^{+\infty} x\psi_{j,i}(x)\mathrm{d}x = 0, \qquad (8.2.13)$$

assuming $\psi(\cdot)$ has $K \geq 1$ vanishing moments. Due to the infinite sums and infinite integrals, we are not able to work out the closed form for the expectation, clearly this is not the desired outcome and this issue needs to be addressed. Therefore,

in this 'level-by-level' case we will assume the form of $\hat{g}(x)$ involving the scaling function $\phi_{j_{\min},i}(x)$ as well (see equation (2.3.9) for more detail),

$$\hat{g}(x) = \sum_i f^{\phi}_{j_{\min},i}\phi_{j_{\min},i}(x) + \sum_{j=j_{\min}}^{j_{\max}} \sum_i g^{\psi}_{j,i}\psi_{j,i}(x) + \sum_{j>j_{\max}} \sum_i f^{\psi}_{j,i}\psi_{j,i}(x), \tag{8.2.14}$$

where we have used

$$\sum_{j<j_{\min}} \sum_i f^{\psi}_{j,i}\psi_{j,i}(x) = \sum_i f^{\phi}_{j_{\min},i}\phi_{j_{\min},i}(x). \tag{8.2.15}$$

Given that we are free to chose our starting distribution $f(\cdot)$, we assume that the starting density is behaving regularly, without sharp spikes and discontinuities. Assuming regularity and sparsity of wavelet coefficients,

$$\left| \sum_{j>j_{\max}} \sum_i f^{\psi}_{j,i}\psi_{j,i}(x) \right| \leq \epsilon, \forall x \in \mathbb{R}, \tag{8.2.16}$$

and therefore under these assumptions we will disregard $\mathcal{O}(2^{j_{\max}+1})$ terms. Now we are able to estimate the normalisation constant and higher moments quite straightforwardly,

$$\hat{K}_g = \int_{-\infty}^{+\infty} \hat{g}(x)\mathrm{d}x = \sum_i f^{\phi}_{j_{\min},i} \int_{-\infty}^{+\infty} \phi_{j_{\min},i}(x)\mathrm{d}x = c\sum_i f^{\phi}_{j_{\min},i}, \tag{8.2.17}$$

where $c = \int_{-\infty}^{+\infty} \phi_{j_{\min},i}(x)\mathrm{d}x$, $\forall i \in \mathbb{Z}$. Furthermore, if we decide to add extra resolution levels $j > j_{\max}$ to the approximation $\hat{g}(x)$, the estimate of the normalisation constant (8.2.17) will not change because all mother wavelets $\psi_{j,i}(x)$ integrate to 0, so only the coarsest resolution with a scaling function (the first term of (8.2.19)) will determine the estimate of the normalisation constant.

So, using a standard WMC we are able to produce samples from a distribution with an approximate density

$$\widetilde{g}(x) = \sum_i f^{\phi}_{j_{\min},i}\phi_{j_{\min},i}(x) + \sum_{j=j_{\min}}^{j_{\max}} \sum_i g^{\psi}_{j,i}\psi_{j,i}(x), \tag{8.2.18}$$

assuming that $\widetilde{g}(x)$ satisfies probability density properties. The key issue with $\widetilde{g}(x)$ is the involvement of scaling coefficients of a starting density $f(\cdot)$. By setting the coarsest resolution level parameter $j_{\min}$ to a very low value in a standard WMC we are able to produce approximate samples from the density

$$\widetilde{g}(x) = \sum_i g^{\phi}_{j_{\min},i}\phi_{j_{\min},i}(x) + \sum_{j=j_{\min}}^{j_{\max}} \sum_i g^{\psi}_{j,i}\psi_{j,i}(x). \tag{8.2.19}$$

The reason for this approximation is that we are able to use scaling functions to represent $j < j_{\min}$ levels

$$\sum_{j=-\infty}^{j_{\max}} \sum_i g^{\psi}_{j,i}\psi_{j,i}(x) = \sum_i g^{\phi}_{j_{\min},i}\phi_{j_{\min},i}(x) + \sum_{j=j_{\min}}^{j_{\max}} \sum_i g^{\psi}_{j,i}\psi_{j,i}(x). \tag{8.2.20}$$

Mainly using the assumption that $\hat{g}(x)$ and $\widetilde{g}(x)$ are densities, we are able to produce approximate samples from (8.2.19). We next will will introduce a method which allows to systematically improve samples by applying WMC level-by-level.

### 8.2.2 Set up of the algorithm

1. Begin with

$$f(x) = \frac{1}{\hat{K}_g} \sum_i g^{\phi}_{j_{\min},i}\phi_{j_{\min},i}(x) \tag{8.2.21}$$

with $\hat{K}_g = c\sum_i g^{\phi}_{j_{\min},i}$ . So the starting distribution is going to be the coarsest possible approximation of $g(x)$ and our target is going to be

$$\hat{g}_{j_{\max}}(x) = \frac{1}{\hat{K}_g} \sum_i g^{\phi}_{j_{\min},i}\phi_{j_{\min},i}(x) + \sum_{j=j_{\min}}^{j_{\max}} \sum_i g^{\psi}_{j,i}\psi_{j,i}(x), \tag{8.2.22}$$

where $\hat{g}_{j_{\max}}(x)$ denotes the level $j_{\max}$ estimate of the theoretical target $g(x)$.

2. Assume we are able to produce samples $\{x_n\}_{n=1}^N \sim f(\cdot)$. This could be approximately achieved by running WMC a priori for $j_{\min}$ reasonably low and $j_{\max} = j_{\min^*}$, where $j_{\min^*}$ is the future LWMC $j_{\min}$ value.

In the LWMC case we will try to approach target $\hat{g}_{j_{\max}}(x)$ sequentially, by first applying the WMC to produce samples from $\hat{g}_{j_{\min}+1}(x)$ using $f(x)$ and then using samples from $\hat{g}_{j_{\min}+1}(x)$ to move up to samples from $\hat{g}_{j_{\min}+2}(x)$ etc. until we reach our desired target level $j_{\max}$.

Moving from $f(x) = \hat{g}_{j_{\min}}(x)$ to $\hat{g}_{j_{\min}+1}(x)$:

$$\hat{g}_{j_{\min}+1}(x) = f(x) + \sum_i g^{\psi}_{j_{\min}+1,i}\psi_{j_{\min}+1,i}(x), \tag{8.2.23}$$

so our difference function becomes only a sum of details at level $j_{\min}+1$,

$$d(x) = \hat{g}_{j_{\min}+1}(x) - f(x) = \sum_i g^{\psi}_{j_{\min}+1,i}\psi_{j_{\min}+1,i}(x). \tag{8.2.24}$$

Therefore, applying LWMC algorithm we will have to only worry about coefficients at the next finer level,

$$p_{ji,t}(x) = \begin{cases} \frac{[g^{\psi}_{j_{\min}+1,i}\psi_{j_{\min}+1,i}(x)]^-}{\sum_i [g^{\psi}_{j_{\min}+1,i}\psi_{j_{\min}+1,i}(x)]^-} & \text{if } j = j_{\min}+1, \\ 0 & \text{otherwise,} \end{cases}$$

where $p_{ji,t}(x)$ are the probabilities associated with picking a mother wavelet $\psi_{j,i}$ as in the standard WMC, however this time the only relevant resolution level is $j = j_{\min}+1$. In addition to this, if we are working with Daubechies wavelets with $K$ vanishing moments, we are only required to estimate $2K-1$ wavelet coefficients, because there are only that many locations $i$ at level $j$ for which wavelet $\psi_{j,i}$ envelopes any given point $x$.

After moving to $j_{\min}+1$ from $j_{\min}$ we will end up with a sample from $\hat{g}_{j+1}(x)$ and we will be able to apply the same technique described above again to keep on moving up the resolution levels. The main issue of this algorithm is the precondition of being able to efficiently produce samples from a starting $f(\cdot)$ distribution.

### 8.2.3 Dangers of LWMC

As a concept, LWMC presents a very systematic algorithm that sequentially grows samples being from the coarsest approximation to being from the finest possible approximation to the target. However, given a rather strong assumption that starting density $f(\cdot)$ in (8.2.21) is in fact the probability density that satisfies non-negativity property and integrates to one, LWMC algorithm faces some serious issues.

As it was commented before, WMC practical target is (8.2.11) and it is not guaranteed to be non-negative everywhere, even after choosing $j_{\min}$ and $j_{\max}$ values low and high enough. Nonetheless, samples are still being generated via WMC process even from regions where practical density might be negative. This present discrepancy between what samples are being produced via WMC and what distribution they belong to still remains unexplained.

In step one of LWMC method we start with samples from the chosen coarsest approximation of the target $g(\cdot)$, in particular we define a starting density in LWMC to be of this form,

$$f(x) = \frac{1}{\hat{K}_g} \sum_i g^{\phi}_{j_{\min},i} \phi_{j_{\min},i}(x).$$

If in a standard WMC with a starting density $f(\cdot)$ and target $g(\cdot)$ we set our coarsest level to be $-\infty$ and finest to $j_{\min}$, under present understanding of WMC theory we will be generating samples from the practical target

$$\hat{g}(x) = \sum_{j=-\infty}^{j_{\min}} \sum_i g^{\psi}_{j,i} \psi_{j,i}(x) + \sum_{j>j_{\min}} \sum_i f^{\psi}_{j,i} \psi_{j,i}(x). \tag{8.2.25}$$

Under approximations described in §8.2.1 we claim that

$$\hat{g}(x) \approx \frac{1}{\hat{K}_g} \sum_i g^{\phi}_{j_{\min},i} \phi_{j_{\min},i}(x). \tag{8.2.26}$$

In practice this approximation seems to be very reasonable, Figure 8.8 presents an example how similar

$$g_{-20:0}(x) = \sum_{j=-20}^{0} \sum_{i} g_{j,i}^{\psi} \psi_{j,i}(x) \tag{8.2.27}$$

is to

$$\hat{g}(x) = \sum_{j=-20}^{0} \sum_{i} g_{j,i}^{\psi} \psi_{j,i}(x) + \sum_{j<-20} \sum_{i} f_{j,i}^{\psi} \psi_{j,i}(x) + \sum_{j>0} \sum_{i} f_{j,i}^{\psi} \psi_{j,i}(x). \tag{8.2.28}$$

However, as it could be seen in Figure 8.8 the practical target does not satisfy the

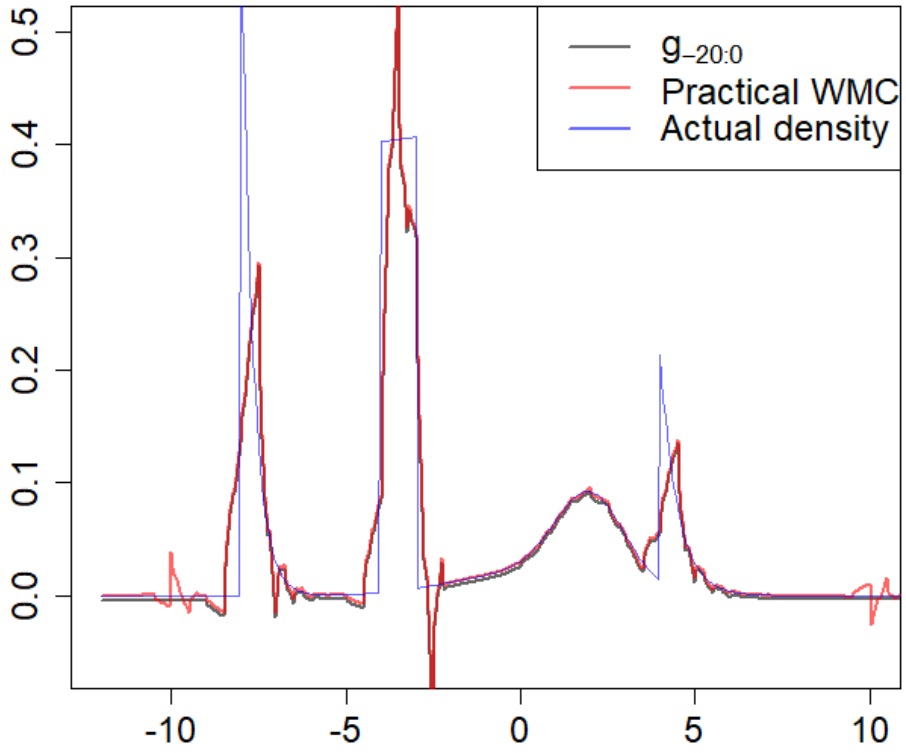



Figure 8.8: *Comparison between practical WMC target density with* $j_{\min} = -20$ *and* $j_{\max} = 0$*, and approximate version* $g_{-20:0}(x)$.

density property of non-negativity and would introduce technical problems if used directly in LWMC as a starting density. Negative density values of a starting density $f(\cdot)$ imply improper survival time densities. The sign of the rate parameter $\frac{c(x_s)}{rf(x_s)}$

in the exponential distribution depends on the value of a starting distribution $f(\cdot)$,

$$f_t(t|x_s) = \frac{c(x_s)}{rf(x_s)} \exp\Big\{ -(t-s)\frac{c(x_s)}{rf(x_s)} \Big\}, \tag{8.2.29}$$

for this reason starting densities that produce negative values cannot be used in LWMC.

Similarly, as in Provisional WMC (Section 3.2), a strong assumption **A2**,

$$\sum_{j\in\mathbb{Z}} \sum_{i\in\mathbb{Z}} [d^{\psi}_{j,i}\psi_{j,i}(x)]^{-} \leq rf(x) \quad \forall x \in \mathbb{R},$$

needs to be satisfied in order to make an algorithm valid; here, we have a strong assumption on the non-negativity of a starting density $f(\cdot)$. Although, LWMC algorithm is far from being practical it does open a new way of thinking about resolution levels in the WMC itself. It might be possible to limit the range of resolution levels at each iteration and sequentially update the quality of samples later without the non-negativity assumption, however it needs further investigation. This particular modification of WMC could easily serve as a future research topic that would also investigate and compare computational costs of LWMC against a standard WMC.

## 8.3 Summary

In this chapter, two possible alternatives to the standard WMC were presented, MIS-WMC and LWMC.

The former, MIS-WMC, motivated by the large amount of intermediate points being computed was constructed to incorporate those intermediate points in the computation of moments of the target density. Together with the method how points should be allocated to intermediate distributions (§8.1.5), ghost points avoided and finally intermediate points weighted when combining into MIS-WMC estimate

(§8.1.4), the algorithm was tested and estimates were numerically analysed (§8.1.7). For the particular example analysed, results revealed that there is not so much difference in utilising MIS-WMC when the number of samples being produced is large, however results were somewhat positive for smaller values of $N$, which is the desired outcome. The ultimate goal is that in a situation when the cost to produce standard $N$ samples from WMC is high, MIS-WMC could be used with a smaller amount of points from the target $M < N$, but still keeping approximately the same accuracy for the estimates computed.

The next algorithm discussed was the LWMC. This algorithm was outlined only in a theoretical manner. Given that in a standard WMC at each iteration, for every intermediate point $x_s$, wavelet coefficients at all resolution levels need to be computed, LWMC was designed in such a way that that samples from a starting distribution are upgraded to samples that incorporate finer resolution levels. Using this level-by-level updated method, the algorithm would be accessing one resolution layer at the time and potentially converging to the desired sample more rapidly. Furthermore, knowing that samples produced by the standard WMC are from

$$\hat{g}(x) = \sum_{j=j_{\min}}^{j_{\max}} \sum_{i} g_{j,i}^{\psi} \psi_{j,i}(x) + \sum_{j<j_{\min}} \sum_{i} f_{j,i}^{\psi} \psi_{j,i}(x) + \sum_{j>j_{\max}} \sum_{i} f_{j,i}^{\psi} \psi_{j,i}(x),$$

The LWMC algorithm could be always applied on top of those standard WMC samples to improve the quality. However, given the strong assumption that expression in Equation 8.2.21 is density, the algorithm requires addition modifications to be viable in practice.

# Chapter 9

# Conclusions

Here we will give a summary of the main results presented throughout the thesis. Where relevant, possible future work will be discussed and potential improvements outlined.

## 9.1 Problems and advantages

In Chapter 5, core problems of WMC were discussed. The ability to properly implement the algorithm seemed to be the main concern, in particular, the assumption of access to the ratio of normalising constants (3.1.5):

$$r = \frac{\int g(x)\,\mathrm{d}x}{\int f(x)\,\mathrm{d}x}.$$

Throughout all chapters, there was not much attention given to the estimation of $r$, and it was mainly assumed to be known. However, in practice we are faced with non-standard target distributions that are not integrable and the normalisation constant is not accessible via standard integration techniques. Clearly, certain estimation methods (Section 5.1) could be used to find good quality estimates of $r$, but in general this is a difficult task, as was discussed in Section 5.1. Furthermore, as was

pointed out in Section 4.4, WMC scales poorly with the dimension of a space, as the number of wavelet coefficients that need to be estimated grows geometrically with the dimension. This is an unwelcome feature that simply can not be avoided due to the nature of the wavelet decomposition. The number of unique building blocks (wavelet type combinations) increases, in turn leading to a more complex decomposition. In addition to the number of coefficients growing geometrically, poor estimation of wavelet coefficients leads to faulty samples being produced via WMC (Section 5.5).

Compared to other sampling algorithms, WMC is very clear to operate and does not require a lot of subjective tuning. For example, the majority of MCMC methods require a user to tinker with a proposal distribution to make an algorithm work properly. The choice of this distribution is not usually a trivial task and it requires careful analysis by specialists. On the other hand, for any starting distribution and wavelet family with compact support, the WMC algorithm will produce independent samples from any target of choice, subject to the computational cost that depends on the range of resolution levels and the dimensionality. Choices of the starting distribution and wavelet family are not critical tasks. A uniform distribution could be always chosen to be a starting one and Daubechies family was shown to be an optimal choice for wavelets. Therefore, at this stage WMC potential could be best utilised by professional programmers who are able to optimise the execution time of the code involved to run the algorithm. Given that decisions made before running WMC do not require much theoretical tuning and a priori analysis, WMC is a very straightforward method and with clear instructions could be handled by many non-professional statisticians.

All in all, it is quite clear that, at this stage of its development, WMC should be approached more as a prototype for future algorithms, rather than a final version itself. From the implementation Chapter 4, we also know that WMC is constructed

in such a way that it ensures the independence of samples and it treats multi-modal distributions in the same way as uni-modal ones, guaranteeing the full exploration of a target distribution and access to convenient parallelisation techniques. In order to make the algorithm fully practical, the future WMC versions will have to find a compromise between the positives (independence and multi-modality) and negatives (curse of dimensionality and normalisation constant).

## 9.2 Theoretical analysis of jumps

Given that the efficiency of the WMC algorithm is highly dependent on the number of jumps taken to produce a sample from the target, a theoretical analysis of a distribution of jumps was carried out in Chapter 6. Unfortunately, due to the complexity introduced by the wavelet decomposition, a simplification (6.3.4) had to be made. It was shown how the expected number of jumps is related to a parameter $\delta$ that controls the discrepancy between the starting and the target density. Simulations showed agreement between the empirical results and theoretical ones for the expected number of jumps, but not for the associated variance (Figure 6.4). Furthermore, the condition (6.5.31),

$$\sum_{j,i} A_j |h^{\psi}_{j,i}| < \infty,$$

was found to be a necessary one for the validity of the WMC theory. Understanding the jump distribution is the key to optimising WMC and making it practical. The functional forms of jump probabilities could be used to fine tune a starting distribution as explained in Section 6.6.

## 9.3 Besov spaces

In Chapter 7, we investigated the WMC assumption **A2** (Section 3.2):

$$\sum_{j\in\mathbb{Z}}\sum_{i\in\mathbb{Z}}[d^{\psi}_{j,i}\psi_{j,i}(x)]^{-} \leq rf(x) \quad \forall x \in \mathbb{R}, \tag{9.3.1}$$

in more detail. In particular, we tried to show how this assumption is related to the issue of Haar wavelets, which are not able to transition probability mass across the origin. The analysis led to the connection of WMC theory with Besov spaces. The assumption **A2** restricts the space that we are allowed to operate on from $L^2$ to $\dot{B}^0_{1,1}$. The theory related to the decomposition of functions using wavelets in $\dot{B}^0_{1,1}$ clearly shows that due to continuity and differentiability reasons Haar wavelets are not a valid system to be used in WMC. The link of Besov space theory and WMC is at the core of the algorithm. This connection should lead to a better understanding of how wavelet systems should be used in the implementation of WMC in higher dimensions.

## 9.4 Algorithm alternatives

In Chapter 8, we have considered modifying the original WMC algorithm and exploiting certain features. Given that a lot of intermediate samples are being produced during a standard WMC run, a Multiple Importance Sampling WMC (Section 8.1) was constructed with a method to allocate intermediate points to intermediate distributions and instead of discarding these points, use them for the estimation of moments of the target distribution. This method showed small improvements in the estimation of mean of the target density, however at the cost of increased variance in the estimate. The method could be explored and fine tuned further by improving the effective utilisation of recycled intermediate points. In

particular, careful analysis could be carried out investigating the appropriate cut-off point of intermediate distributions that can be included in the estimation of moments of the target density.

The second proposed alternative to WMC was Level-WMC (Section 8.2). This new algorithm was only outlined theoretically, suggesting the possibility of approaching samples from the approximate target distribution in levels. Samples would first be produced from a coarser approximation and later upgraded to finer ones, potentially lowering the computational load and saving some computational time. However, as pointed out in Section 8.2, the algorithm is highly dependent on the non-negativity assumption of a starting density and, therefore, the idea requires finer refinements to be fully functional.

If we are interested in producing $N$ samples from the target distribution, we are free to pick whatever starting density we want to achieve this result via WMC. Certain densities will lead to samples being produced in a shorter time and more efficient way. A new modification could be considered in the future that suggests using an adaptive WMC, a method that adjusts the starting distribution adaptively, to better suit the target. The decision to switch to a better choice of starting distribution could be made by monitoring the average number of jumps required to reach a target and using it as a criterion that constantly needs to be minimised.

As with many other sampling algorithms, an original version is usually far from being optimised and in general could be improved. Similarly here, we note several possible modifications to the original WMC, and, hope that in the future, they could be explored even more in detail to produce a superior sampling method.

## 9.5 Future work

Given the significant computational cost attached to the implementation of the WMC, the future work related to WMC should be highly focused on the computational optimisation of the algorithm. In particular, efficient ways how to construct and produce samples from the desired wavelet of interest in real time is a top priority. The same could be said about the computation of wavelet coefficients, required to construct sampling distributions for wavelet resolution levels. If these two goals could be achieved, WMC algorithm has real chances of becoming one of the more popular sampling algorithms in the scientific community.

# Bibliography

Akay, M. (1995), 'Wavelets in biomedical engineering', *Annals of biomedical Engineering* **23**(5), 531–542.

Almeida, A. (2005), 'Wavelet bases in generalized Besov spaces', *Journal of Mathematical Analysis and Applications* **304**(1), 198–211.

Beskos, A., Roberts, G. & Stuart, A. (2009), 'Optimal scalings for local Metropolis - Hastings chains on nonproduct targets in high dimensions', *The Annals of Applied Probability* **19**(3), 863–898.

Besov, O. (1959), On a family of function spaces. embedding theorems and extensions, *in* 'Doklady Akademii Nauk SSSR', Vol. 126, pp. 1163–1165.

Besov, O. V. (1961), 'Investigation of a class of function spaces in connection with imbedding and extension theorems', *Trudy Matematicheskogo Instituta Imeni VA Steklova* **60**, 42–81.

Bijaoui, A., Slezak, E., Rue, F. & Lega, E. (1996), 'Wavelets and the study of the distant universe', *Proceedings of the IEEE* **84**(4), 670–679.

Camussi, R. & Guj, G. (1997), 'Orthonormal wavelet decomposition of turbulent flows: intermittency and coherent structures', *Journal of Fluid Mechanics* **348**, 177–199.

Celeux, G., Hurn, M. & Robert, C. P. (2000), 'Computational and inferential difficulties with mixture posterior distributions', *Journal of the American Statistical Association* **95**(451), 957–970.

Cohen, A., Dahmen, W. & Devore, R. (2001), 'Adaptive wavelet methods for elliptic operator equations: Convergence rates', *Mathematics of Computation* **70**(233), 27–75.

Cooley, J. W. & Tukey, J. W. (1965), 'An algorithm for the machine calculation of complex Fourier series', *Mathematics of Computation* **19**, 297–301.

Daubechies, I. (1988), 'Orthonormal bases of compactly supported wavelets', *Communications on Pure and Applied Mathematics* **41**(7), 909–996.

Daubechies, I., Grossmann, A. & Meyer, Y. (1986), 'Painless nonorthogonal expansions', *Journal of Mathematical Physics* **27**(5), 1271–1283.

de Valpine, P. (2008), 'Improved estimation of normalizing constants from Markov chain Monte Carlo output', *Journal of Computational and Graphical Statistics* **17**(2), 333–351.

Devroye, L. (1986), *Non-Uniform Random Variate Generation*, Springer-Verlag.

Donoho, D. L. (1995), 'De-noising by soft-thresholding', *IEEE Transactions on Information Theory* **41**(3), 613–627.

Donoho, D. L. & Johnstone, J. M. (1994), 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika* **81**(3), 425–455.

Farge, M. (1992), 'Wavelet transforms and their applications to turbulence', *Annual review of fluid mechanics* **24**(1), 395–458.

Frazier, M., Frazier, M. W., Jawerth, B. & Weiss, G. L. (1991), *Littlewood-Paley theory and the study of function spaces*, number 79, American Mathematical Soc.

Gabor, D. (1946), 'Theory of Communication', *Journal of the Institution of Electrical Engineers* **93**(26), 429–457.

Gallegati, M. (2012), 'A wavelet-based approach to test for financial market contagion', *Computational Statistics & Data Analysis* **56**(11), 3491–3497.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6), 721–741.

Gilks, W. R. (2017), 'Notes on Wavelet Monte Carlo', Personal communications.

Gilks, W. R., Best, N. G. & Tan, K. K. C. (1995), 'Adaptive rejection Metropolis sampling within Gibbs sampling', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **44**(4), 455–472.

Gilks, W. R. & Wild, P. (1992), 'Adaptive rejection sampling for Gibbs sampling', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **41**(2), 337–348.

Gilks, W., Richardson, S. & Spiegelhalter, D. (1995), *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC Interdisciplinary Statistics, Taylor & Francis.

Grossmann, A. & Morlet, J. (1982), 'Decomposition of Hardy functions into square integrable wavelets of constant shape', *SIAM Journal on Mathematical Analysis* **15**(4), 723–736.

Haar, A. (1910), 'Zur theorie der orthogonalen funktionensysteme', *Mathematische Annalen* **69**(3), 331–371.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**(1), 97–109.

Kahn, H. & Harris, T. E. (1951), 'Estimation of particle transmission by random sampling', *National Bureau of Standards Applied Mathematics Series* **12**, 27–30.

Kartsonaki, C. (2016), 'Survival analysis', *Diagnostic Histopathology* **22**(7), 263–270.

Kinderman, A. J. & Monahan, J. F. (1977), 'Computer generation of random variables using the ratio of uniform deviates', *ACM Transactions on Mathematical Software* **3**(3), 257–260.

Kyriazis, G. (2003), 'Decomposition systems for function spaces', *Studia Mathematica* **2**(157), 133–169.

Kyriazis, G. & Petrushev, P. (2002), 'New bases for Triebel-Lizorkin and Besov spaces', *Transactions of the American Mathematical Society* **354**(2), 749–776.

Lan, S., Streets, J. & Shahbaba, B. (2014), Wormhole Hamiltonian Monte Carlo, *in* 'Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence', AAAI'14, AAAI Press, pp. 1953–1959.

Mallat, S. (2008), *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*, 3rd edn, Academic Press.

Mallat, S. G. (1989*a*), 'Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$', *Transactions of the American Mathematical Society* **315**(1), 69–87.

Mallat, S. G. (1989*b*), 'A theory for multiresolution signal decomposition: The wavelet representation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**(7), 674–693.

Mark, G. & Ben, C. (2011), 'Riemann manifold Langevin and Hamiltonian Monte Carlo methods', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(2), 123–214.

Meng, X.-L. & Wong, W. H. (1996), 'Simulating ratios of normalizing constants via a simple identity: A theoretical exploration', *Statistica Sinica* **6**(4), 831–860.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equation of state calculations by fast computing machines', *JCP* **21**, 1087–1092.

Meyer, Y. (1986-1987*a*), 'Ondelettes et fonctions splines', *Séminaire Équations aux dérivées partielles (Polytechnique)* pp. 1–18.

Meyer, Y. (1986-1987*b*), 'Ondelettes et fonctions splines', *Séminaire Équations aux dérivées partielles (Polytechnique)* pp. 1–18.

Neal, R. M. (1993), Probabilistic inference using Markov Chain Monte Carlo methods, Technical report, CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

Neal, R. M. (2005), Estimating ratios of normalizing constants using linked importance sampling, Technical report, Department of Statistics and Department of Computer Science, University of Toronto.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
**URL:** *http://www.R-project.org/*

Rao, R. & Bopardikar, A. (1998), *Wavelet Transforms: Introduction to Theory and Applications*, Addison-Wesley.

Roberts, G. O., Gelman, A. & Gilks, W. R. (1997), 'Weak convergence and optimal scaling of random walk Metropolis algorithms', *The Annals of Applied Probability* **7**(1), 110–120.

Saini, S. & Dewan, L. (2016), 'Application of discrete wavelet transform for analysis of genomic sequences of mycobacterium tuberculosis', *SpringerPlus* **5**(1), 64.

Sawano, Y. (2018), *Theory of Besov Spaces*, Developments in Mathematics, 1 edn, Springer Singapore.

Sminchisescu, C. & Welling, M. (2011), 'Generalized darting Monte Carlo', *Pattern Recognition* **44**(10-11), 2738–2748.

Smith, A. F. M. & Roberts, G. O. (1993), 'Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods', *Journal of the Royal Statistical Society. Series B (Methodological)* **55**(1), 3–23.

Triebel, H. (1992), *Theory of Function Spaces II*, Monographs in Mathematics, 1 edn, Birkhäuser Basel.

Triebel, H. (2004), 'A note on wavelet bases in function spaces', *Banach Center Publications* **64**(1), 193–206.

Veach, E. (1997), *Robust Monte Carlo methods for light transport simulation*, Vol. 1610, Stanford University PhD thesis.

Veach, E. & Guibas, L. J. (1995), Optimally combining sampling techniques for Monte Carlo rendering, *in* 'Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques', SIGGRAPH '95, ACM, pp. 419–428.

Wornell, G. W. & Oppenheim, A. V. (1992), 'Estimation of fractal signals from noisy measurements using wavelets', *IEEE Transactions on Signal Processing* **40**(3), 611–623.

Zhang, Q. & Benveniste, A. (1992), 'Wavelet networks', *IEEE transactions on Neural Networks* **3**(6), 889–898.