Scaling and the Structure of Neural Networks

An Investigation into Possible Connections between Deep Learning and the Renormalisation Group

Siphelele Danisa

Supervisor: Dr Jonathan Shock

Department of Pure and Applied Mathematics
University of Cape Town

This dissertation is submitted for the degree of
Honours in Mathematics

November 2020

I would like to dedicate this thesis to my loving family and friends …

Declaration

Plagiarism Declaration:

1. I know that plagiarism is a serious form of academic dishonesty.
2. I have read the UCT document Avoiding Plagiarism: A guide for students, am familiar with its contents and have avoided all forms of plagiarism mentioned there.
3. Where I have used the words of others, I have indicated this by the use of quotation marks.
4. I have referenced all quotations and properly acknowledged other ideas borrowed from others.
5. I have not and shall not allow others to plagiarise my work.
6. I declare that this is my own work.

Signature:

Siphelele Danisa
November 2020

Acknowledgements

I am profoundly grateful to Dr Jonathan Shock for all the opportunities of research that he has given me since I started my undergraduate degree, and the immense support he has given me this year. This research journey has been life changing, to say the least, and I appreciate all the time that he invested into it, as well as the other smaller journeys before this one. I have learnt a lot, and I would like to think that I have become a much better Mathematician through these opportunities.

Table of contents

List of figures

Preface

Over the last couple of years, we have seen rapid growth in machine learning as a research field. Most of this growth, however, has been in the form of algorithms and procedures that have been established to carry out certain tasks. While there is no doubt that the applications have been impressive, there is a still a lot to learn as far as the theory of learning, in context, is concerned. There are numerous papers that have been published with attempts, and we mention a few in the following to give some grounding. The first example is [1] wherein it is shown that we can discuss the functionality of aspects of deep learning in the context of group theory. Similarly [2] "show how the success of deep learning could depend not only on mathematics but also on physics: although well-known mathematical theorems guarantee that neural networks can approximate arbitrary functions well, the class of functions of practical interest can frequently be approximated through “cheap learning” with exponentially fewer parameters than generic ones." More has been done to improve our understanding, including asking questions about how well we understand generalisation [3], attempts to establish connections between dynamical systems and deep learning [4], etc. but more remains to be done. This thesis focuses on the connection between what is known as the renormalisation group and deep learning via restricted Boltzmann machines. The conventional framework for this is in the context of measure spaces, and we start the journey by introducing this theory. It is then followed by an introduction on spin systems, after which we see the concepts behind renormalisation. At this point, one gets a brief treatment of restricted Boltzmann machines, then the last two chapters then discuss results from two relevant papers.

Chapter 1

Preliminary Probability Theory

1.1 Introduction to Measure Theory

In this chapter we shall introduce the measure theory that is crucial for discussing our analysis on spin models, the context on which this thesis is focused. In what follows, we shall introduce measure spaces, discuss measurable functions and, lastly, discuss integration in this context. This content is adapted from [5], [6] and [7].

The context of measure spaces shows up quite implicitly when we talk about spin systems– in particular, we might want to talk about taking expectations of relevant quantities in this context, to make simplifications to our models by making assumptions that can be translated into constraints on the measures, etc. We also see it in the third chapter when we talk about the renormalisation group. In this section it is quite explicit in how it appears, as opposed to being a subtlety. Lastly, we see it again as we talk about Restricted Boltzmann Machines. It shows up in all of the core areas of our discussion, and is hence very important for our discussion.

There are two objectives for this section:

1. Defining measure spaces.
2. Defining what it means to integrate functions in this context.

Definition 1.1.1. *Let Ω be a set. We define a σ-algebra or σ-field on Ω as a collection, $\mathcal{A}$, of subsets of Ω such that:*

1. $\emptyset \in \mathcal{A}$.

2. *If $A \in \mathcal{A}$ then $\Omega \backslash A \in \mathcal{A}$.*

3. *If $A_n \in \mathcal{A}$ for $n = 1, 2, \cdots$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$.*

Example 1.1.2.
1. Given any Ω, then $\{\emptyset, \Omega\}$ is a σ-field.
2. Given any Ω, the set of all subsets of Ω, $\mathcal{P}(\Omega)$, is a σ-field.
3. This example is more interesting. Given a space Ω and a collection of subsets of Ω, S, we define the σ-algebra generated by S on Ω to be $\mathcal{A}_S = \bigcap\{\mathcal{A} | \mathcal{A}$ is a σ-algebra on Ω and $\mathcal{A} \supset S\}$. If Ω is a topological space and S is the topology, then $\mathcal{A}_S$ is known as the Borel σ-field, denoted $\mathcal{B}$.
4. This example makes 3. more concrete. If we consider $\mathbb{R}$ with the usual topology, then the Borel σ-field is the smallest sigma algebra that is generated by the usual topology.

We call a pair $(\Omega, \mathcal{A})$ a measurable space. The elements of $\mathcal{A}$ are referred to as the measurable sets in Ω. Examples of measurable spaces follow from what is given above.

Given two measure spaces, we can define structure preserving maps between them which we call measurable functions.

Definition 1.1.3. *Let $(X, \mathcal{A}_X)$ and $(Y, \mathcal{A}_Y)$ be measurable spaces. A function $f : X \to Y$ is measurable if $f^{-1}(\mathcal{A}_Y) \subseteq \mathcal{A}_{\mathcal{X}}$.*

In what follows we give examples in the form of characterisations. More concrete examples can be obtained by simply selecting a function that satisfies any of the relevant statements.

Example 1.1.4.
1. If f is a continuous map of topological spaces considered with Borel σ-fields, then f is a (Borel-) measurable function.
2. Let $f : X \to [0, +\infty]$ then f is measurable if and only if the set $\{x \in X | f(x) > \lambda\}$ is Borel measurable for any real number $\lambda \geq 0$.
3. If $1_A : \Omega \to \mathbb{R}$, where $A \subseteq \Omega$, such that $1_A(x) = 1$ for any $x \in A$ and $1_A(x) = 0$ otherwise, then this function (known as the indicator function) is measurable if and only the subset A is measurable .

In this context of measure spaces, a function $f : X \to \mathbb{C}$ is said to be simple if its range has finitely many points. If we let $(a_i)_{i=1}^n$ be the set of distinct points that a simple function, f, attains, and if $\beta_i = \{x \in X | f(x) = a_i\}$, then we can write $f(x) = \sum_{i=1}^n a_i 1_{\beta_i}(x)$.

Moreover, if $f : X \to [0,\infty]$ is measurable, then it can be shown that there exists a sequence of simple measurable $(f_n)_{n=1}^{\infty}$ on X such that:
1. $0 \leq f_n \leq f_{n+1} \leq f$,
2. $f_n \to f$ as $n \to \infty$ in pointwise fashion.

This key insight is very useful for our purposes to define integration, but before that we need to define what a measure is.

Definition 1.1.5. *A measure, μ, on a measurable space $(\Omega, \mathcal{A})$ is a function*

$\mu : \mathcal{A} \to [0,+\infty]$, *such that:*
1. $\mu(\emptyset) = 0$.
2. For any set of mutually disjoint $(A_i)_{i\geq 1} \subset \mathcal{A}$, we have that $\mu(\cup_{i\geq 1} A_i) = \sum_{i\geq 1} \mu(A_i)$,where $i = 1,2,\cdots$.

A triple $(\Omega, \mathcal{A}, \mu)$ is called a measure space. Here Ω is a set, $\mathcal{A}$ is a σ-algebra on Ω and μ is a measure on $(\Omega, \mathcal{A})$.

Example 1.1.6.
1. The first example of a measure that one can consider is the trivial measure, which maps every element of the σ-algebra to zero.
2. The second example is the Dirac measure δ_x for some $x \in \Omega$ which is defined by $\delta_x(A) = 1_A(x)$.
3. Extending the example above, we may consider for any countable set $A = \{a_1, a_2, \cdots\}$ the associated counting measure defined by $\nu = \sum_{n\geq 1} \delta_{a_n}$.

When we have measure spaces, we can talk about integration. We define integrals using the concept of limits of step functions, which we saw above.

We begin by defining integrals for measurable simple functions.

Definition 1.1.7. *Let $f : X \to [0,\infty)$ be a measurable simple function. Consider the representation $f = \sum_{i=1}^{n} a_i 1_{\beta_i}$ where $(a_i)_{i=1}^{n}$ are the distinct values of the range of f. If $A \in \mathcal{A}$ then we define $\int_A f d\mu = \sum_{i=1}^{n} a_i \mu(\beta_i \cap A)$*

Remark 1.1.8. *The convention $0 \cdot \infty = 0$ is used whenever this is necessary in the above definition since the evaluation of the measure on some $\beta_i \cap A$ might not be finite while the value of the function is zero.*

At this point, we would like to discuss integration of measurable functions. The definition relies on the statement that given any measurable function, we can attain

a sequence of simple measurable functions that converge pointwise to this given function.

Definition 1.1.9. *Let $f : X \to [0,\infty]$ be a measurable function, $g : X \to [0,\infty)$ be a measurable step function, and $A \in \mathcal{A}$, then define $\int_A f d\mu = \sup \int_A g d\mu$ where the supremum is taken over the simple measurable functions g which satisfy $0 \leq g \leq f$.*

Although we have talked about functions with values on the interval $[0,\infty]$, we can extend this argument to $\overline{\mathbb{R}}$, the extended real numbers, as well as to $\mathbb{C}$. This discussion establishes the objectives for this section.

1.2 Introduction to Probability Spaces

We would like to talk about probability spaces at this point, which are simply normalised measure spaces. We are more interested in defining distributions, which we shall see quite early in the discussion about spin systems. The content for this subsection is adapted from [5]. When we talk about spin systems and Restricted Boltzmann Machines the concept of a distribution becomes prominent, and so we wish to define what a distribution is formally, since this knowledge is important to our study.

Definition 1.2.1. *A probability space is a normalised measure space. In particular, it is a triple $(\Omega, \mathcal{A}, P)$ where Ω is still a set, $\mathcal{A}$ is the σ-algebra, and P is a measure such that $P(\Omega) = 1$. P is called a probability measure.*

In the context of probability spaces, the elements of the σ-algebra are known as events, and for some $A \in \mathcal{A}$ we say that $P(A)$ the probability of A.

A mapping $\phi : X \to S'$ where $X = (\Omega, \mathcal{A})$ and $S' = (S, \mathcal{S})$, say, are measure spaces is referred to as a random element in S. Where the target measure space is the set of real numbers with the Borel measure, we call this a random variable. If $O \in \mathcal{S}$, then $\{\phi \in O\} = \phi^{-1}O \in \mathcal{A}$, and we may consider the associated probabilities $P\{\phi \in O\} = P(\phi^{-1}O) = (P \circ \phi^{-1})B$.

Definition 1.2.2. *We define the function $\mathcal{L}(\phi) = P \circ \phi^{-1}$ as a probability distribution on the range of spaces of ϕ.*

There are many examples of distributions (of random variables) that one usually finds in introductory probability courses, namely, the Bernouli, Binomial, Poisson distributions, etc. We shall not be too concerned about these distributions, and we

shall, instead, give examples that will be relevant to our discussions where there is a need.

Chapter 2

Spin Systems

2.1 Ising Model

We now introduce the Ising model, which is the context in which we shall discuss our applications of concepts in the following sections. The Ising model is of interest because it is a simple model that exhibits non-trivial properties and behaviour, but at the same time its theory well-understood. The content of this chapter is inspired by the discussion of spin models in [8].

Definition 2.1.1. *A spin system is a collection of random variables, called spins, that we shall denote* $(\phi_x)_{x\in\Lambda}$ *or* $(\sigma_x)_{x\in\Lambda}$ *for some indexing set* Λ.

We would like to consider, for now, an indexing set Λ that is finite but large.

Definition 2.1.2. *Let* $\Lambda \subset \mathbb{Z}^d$. *We define an Ising configuration as* $\sigma = (\sigma_x)_{x\in\Lambda}$ *such that* $\sigma_x \in \{1, -1\}$.

An example can be seen on 2.1 for $d = 2$ where we have sites on a 2D-lattice and each site is associated with a spin that can either be up or down (i.e binary valued).

We would like to be able to talk about the energy of this system, and to be able to talk about associated spin probability distributions, because this will allow us to study the properties of this model. In order to get to this point, we need the following definitions.

Definition 2.1.3. *Let* v *be a unit vector in* $\mathbb{Z}^d$. *By the discrete gradient of a function* $f : \mathbb{Z}^d \to \mathbb{C}$ *we shall refer to*

$$(\nabla^v f)_x = f_{x+v} - f_x.$$

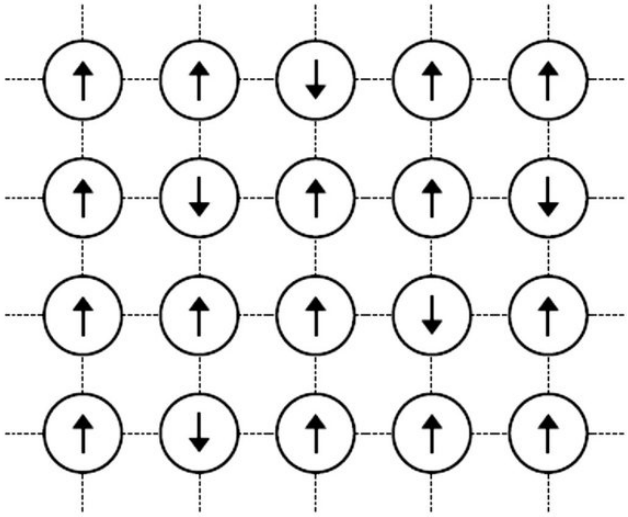


Fig. 2.1 2D Ising Configuration Example [9]

Definition 2.1.4. *Let v be a unit vector as above, once again. The Laplacian of $f : \mathbb{Z}^d \to \mathbb{C}$ is*

$$(\Delta f)_x = -\frac{1}{2} \sum_{v:|v|=1} \nabla^{-v} \nabla^v f_x = \sum_{v:|v|=1} \nabla f_x$$

Finally, we can define the hamiltonian associated with the model as follows.

Definition 2.1.5. *Let σ be an Ising spin configuration, given Λ, then an energy is associated to each such configuration by*

$$H_{0,\Lambda}(\sigma) = \frac{1}{4} \sum_{v:|v|=1} \sum_{x \in \Lambda} (\nabla^v \sigma)_x^2,$$

together with boundary condition terms for the boundary, $\partial\Lambda$, of Λ.

If we let $E^{(2)}$ be the set of nearest neighbours on the lattice, then up to an additive constant we can rewrite the definition as $-\sum_{\{x,y\} \in E^{(2)}} \sigma_x \sigma_y$, which is the common definition in texts. Finally, we can talk about probabilities of configurations.

Definition 2.1.6. *Let σ be an Ising configuration. The probability of such a configuration is given by the finite volume Gibbs measure defined as*

$$P_{T,\Lambda}(\sigma) = Z e^{-H_{0,\Lambda}(\sigma)/T} \prod_{x \in \Lambda} \left(\delta_{\sigma_x,+1} + \delta_{\sigma_x,-1} \right),$$

where T is referred to as temperature.

The constant, Z, normalises the $P_{T,\Lambda}$, making it a probability measure.

Remark 2.1.7. *The model we have introduced is said to be Ferromagnetic in the sense that configurations with aligned $\{x,y\} \in E^{(2)}$ are energetically favourable, and have a high probability of occurrence. The lowest energy state has all spins either equal to 1 or all of them equal to -1. In high energies there is a greater number of configurations that realise a specific energy. We say that high energies lead to a greater entropy of these realisations in the probability measure. What one sees is that there is a competition between energy and entropy whose relative weight is controlled by temperature. There is a temperature T_c such that for $T < T_c$ there is a dominant energy minimising mechanism, while for $T > T_c$ there is dominance of entropy. It is then said that the model exhibits as second order phase transition at $T = T_c$.*

We can generalise the model above by considering an external magnetic field that interacts with each spin variable. In this case, we proceed as follows.

Definition 2.1.8. *Let $h \in \mathbb{R}$. We define the Ising model with an external magnetic field by specifying the energy, $H_{h,\Lambda}(\sigma)$, as*

$$H_{h,\Lambda}(\sigma) = H_{0,\Lambda} - h \sum_{x \in \Lambda} \sigma_x$$

The probabilities are defined via a finite volume Gibbs measure with

$$P_{h,T,\Lambda} = Z' e^{-H_{h,\Lambda}(\sigma)/T} \prod_{x \in \Lambda} \left(\delta_{\sigma_x,+1} + \delta_{\sigma_x,-1} \right),$$

similar to what we had before. The infinite volume limit $P_{h,T}$ is the limit of $P_{h,T,\Lambda}$ as $\Lambda \uparrow \mathbb{Z}^d$ when this exits. We shall use the notation $\langle \cdot \rangle_{h,T}$ to denote the expectation given the probability measure $P_{h,T}$.

Now, we turn to the discussion about physical quantities which allow us to understand the model better, and we start this by defining magnetisation, spontaneous magnetisation, and magnetic susceptibility.

Definition 2.1.9. *We define*

$$M(h,T) = \langle \sigma_0 \rangle_{h,T},$$

and call this the magnetisation associated with the Ising model. Similarly,

$$M_+(T) = lim_{h\downarrow 0} M(h,T)$$

is said to be the spontaneous magnetisation. Lastly, slope of the magnetisation at $h = 0$ is called the magnetic susceptibility.

We further proceed by defining the following quantities in order to probe and understand the model behaviour near the critical temperature, where a phase transition occurs–we say that the model is critical here. The question we ask is: *"How do the physical quantities behave around the criticality?"*

Definition 2.1.10. *Let $T \geq T_c$, then define:*

1. *The two-point function as*
$$\tau_{0x}(T) = \langle \sigma_0 \sigma_x \rangle_{0,T}.$$

2. *The correlation length as*
$$\xi(T)^{-1} = -\lim_{n\to\infty} n^{-1} \log \tau_{0,nv_1}(T),$$
where $v_1 = (1,0,\cdots,0)$, a $\mathbb{Z}^d$ unit vector

3. *Susceptibility as*
$$\chi(T) = \sum_{x\in\mathbb{Z}^d} \tau_{0x}(T) = \frac{\partial}{\partial h} M(h,T)\big|_{h=0}.$$

Now we can talk about behaviour near the criticality. In this regime, we find that the spins develop strong non-trivial correlations. These scale into the the following critical exponents $(y,\nu,\eta,\delta,\beta) \in \mathbb{R}^5$:

1.
$$\chi(T) \sim A_1 (T - T_c)^{-y},$$
as T descends to T_c.

2.
$$\xi(T) \sim A_2 (T - T_c)^{-\nu},$$
with the same behaviour of T.

3.
$$\tau_{0x}(T_c) \sim A_3 |x|^{-(d-2+\eta)},$$

as $|x| \to \infty$.

4. $$M(h, T_c) \sim A_4 h^{1/\delta},$$
as h descends to 0.

5. $$M_+(T) \sim A_5 (T_c - T)^{\beta},$$
as T ascends to T_c.

The critical exponents are conjectured to obey certain scaling relations. For example, one such relation is Fisher's relation, which is $y = (2 - \eta)\nu$. They are also predicted to depend primarily on the dimension d, given an ambient space $\mathbb{Z}^d$, and not on the model's *fine print*. We say that they are universal in this sense (or are at least predicted to be). Determining the exponents for a model allows one to describe the behaviour of the model around the critical point of interest.

We now provide examples of what critical exponents can be.

Example 2.1.11.

1. *For $d = 2$, it has been shown that $T_c^{-1} = \frac{1}{2}\log(1 + \sqrt{2})$, and that y, β, δ, η and ν exist. They take the values $y = 7/4$, $\beta = 1/8$, $\delta = 15$, $\eta = 1/4$ and $\nu = 1$.*
2. *For $d > 4$, one finds that y, β, δ and η exist, and that they assume the values $y = 1$, $\beta = 1/2$, $\delta = 3$ and $\eta = 0$.*

Similar results for dimensions 3 and 4 are largely open questions that we shall not attempt to tackle here.

2.2 Universality of Spin Models

Now we discuss the universality of spin models, which is an interesting property. The Ising model is only an example of a general class of models defined as follows.

Definition 2.2.1. *Let Λ be a finite set, and let $\beta_{xy} = \beta_{yx}$ be non-negative spin-spin coupling constants that we shall index by $\Lambda \times \Lambda$. A spin configuration, ϕ, constitutes a spin $\phi_x \in \mathbb{R}^n$ for all $x \in \Lambda$, and $\phi : \Lambda \to \mathbb{R}^n$, which we shall write as $\phi \in \mathbb{R}^{n\Lambda}$.*

Definition 2.2.2. *Define the total energy associated with ϕ as*

$$H(\phi) = \frac{1}{4}\sum_{x,y\in\Lambda}\beta_{xy}|\phi_x - \phi_y|^2 + \sum_{x\in\Lambda} h\cdot\phi_x,\ h\in\mathbb{R}.$$

Definition 2.2.3. *For a given reference measure, μ, on $\mathbb{R}^n$, referred to as a single-spin distribution, a probability measure on the spin configuration is defined by*

$$\langle F\rangle \propto \int_{\mathbb{R}^{n\Lambda}} F(\phi)e^{-H(\phi)}\prod_{x\in\Lambda}\mu(d\phi_x).$$

If the measure μ is absolutely continuous, which is to say that if λ denotes the Lebesgue measure ([6], page 49) then $\lambda(A) = 0$ (for some measurable set A) implies that $\mu(A) = 0$, then it is convenient to take μ as the Lebesgue measure, and equivalently add a potential term to the total energy to obtain a form

$$H(\phi) = \frac{1}{4}\sum_{x,y\in\Lambda}\beta_{xy}|\phi_x - \phi_y|^2 + \sum_{x\in\Lambda} h\cdot\phi_x + \sum_{x\in\Lambda} w(\phi_x)$$

Definition 2.2.4. *Let β be the matrix of couplings, $f:\Lambda\to\mathbb{R}$. We define the Laplacian matrix as Δ_β such that*

$$(\Delta_\beta f)_x = \sum_{y\in\Lambda}\beta_{xy}(f_y - f_x).$$

One can see from here that if $\beta_{xy} = \mathbb{1}_{x\sim y}$ for x and y as nearest neighbours in $\mathbb{Z}^d$ we get the Laplacian we defined previously.

Let $f = (f_1,\cdots,f_n)$ then

$$(\Delta_\beta f)^i = (\Delta_\beta f^i).$$

Thus we can write, for these functions, the total energy as

$$H(\phi) = \frac{1}{2}\sum_{x\in\Lambda}\phi_x(-\Delta_\beta)\phi_y + \sum_{x\in\Lambda} h\cdot\phi_x + \sum_{x\in\Lambda} w(\phi_x),$$

and one may include boundary terms.

We get a variety of spin models by varying the way in which we define μ, w and the form of β_{xy}. In the appropriate infinite volume limits $|\Lambda|\to\infty$ the models in this

class typically undergo phase transitions as parameters are varied. The universality conjecture for critical behaviour predicts that this behaviour is the same within very general symmetry classes, where these symmetries are defined by the number of components, n, corresponding to the symmetry group $O(n)$ and the class of coupling constants.

Chapter 3

The Renormalisation Group

3.1 Theory

In the last part of the previous chapter, we mention that in the infinite volume limit, the models of interest typically undergo phase transitions as one varies their parameters. In this section we offer a more rigorous framework on how this is done in the context of hierarchical models. This chapter is largely derived from the treatment of the topic in [8] and [10]. The latter is the main source.

The renormalisation group is a procedure that is used to integrate out short distance degrees of freedom in systems that usually have infinite degrees of freedom. This procedure allows one to obtain an effective description of the system which tends to be described a finite and small number of operators. This theory is well-established, and has been used in Physics to solve problems that have earned people the Nobel Prize. We give a simple outline for this procedure in this chapter.

Even with our intentions of keeping things simple, this chapter is quite technical– in the sense of requiring a lot of definitions and results stated to achieve our objectives– but we facilitate this journey by establishing a roadmap for the exposition. We take the following approach:

1. First, we define Gaussian measures and the concept of covariance in the context.
2. Secondly, we consider a specific decomposition of the covariance under which we define the renormalisation group as a map on what we shall define as global functions.

3. Thirdly, we show that we can consider local maps, also to be defined, and establish an equivalence to 2. This is known as the global to local programme.
4. Lastly, we spend the remaining time talking about the renormalisation group in the local context and discuss the infinite volume limit problem.

The advantages of going from the global scale to the local scale are that we can work with perturbative tools, which tend to be insightful given tough problems, and the idea is that we can then obtain equivalence by taking the infinite volume limit.

We shall start by laying out the context in which we would like to work. We assume that $\Lambda \subset \mathbb{Z}^d$ is finite, and consider $\mathbb{R}^\Lambda$ as a probability space. An element ϕ of $\mathbb{R}^\Lambda$ is called a field, and we have that $\phi : \Lambda \to \mathbb{R}$. The assumption that we make is that we are given a map $S : \mathbb{R}^\Lambda \to \mathbb{R}$, which we shall call an action. If we let $d\phi$ be the Lebesgue measure on the probability space $\mathbb{R}^\Lambda$, then the action induces what is known as the finite volume Gibbs measure on the same space, which is defined as

$$d\mu_\Lambda(\phi) = \Theta^{-1} e^{-S} d\phi,$$

where $\Theta = \Theta(\Lambda)$ is a normalisation constant, called the partition function.

Definition 3.1.1. *Let $\Lambda \subset \mathbb{Z}^d$ be finite, then a measure on $\mathbb{R}^\Lambda$ is said to be Gaussian (with mean zero) if*

$$d\mu(\phi) = \mathcal{G} d\phi e^{-\frac{1}{2}Q(\phi,\phi)},$$

where $Q : \mathbb{R}^\Lambda \times \mathbb{R}^\Lambda \to \mathbb{R}$ is a quadratic form such that $Q(\phi,\phi) > 0$[1] for $\phi \neq 0$.

Remark 3.1.2.
1. Recall that a quadratic form on $\mathbb{R}^n \times \mathbb{R}^n$ is a map $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that

$f(x,x) = x^T A x$ where A is a unique symmetric $n \times n$ matrix.
2. For our case, we shall denote this matrix as $A = (A(x,y)_{x,y\in\Lambda})$, and write $Q(\phi,\phi) = (\phi, A\phi) = (A\phi,\phi)$, where $(f,g) = \sum_{x\in\Lambda} f(x)g(x)$. Note that since the matrix A is positive definite (i.e the associated quadratic form is positive definite), we have that its eigenvalues are strictly positive, and so it is invertible. The inverse is also symmetric and positive definite.
3. Given a positive-definite symmetric matrix $C = (C(x,y)_{x,y\in\Lambda})$, one can define a Gaussian measure on $\mathbb{R}^\Lambda$ by considering $A = C^{-1}$.

[1] We refer to this condition as (strict) positivity of the quadratic form.

In light of this remark, we can then consider a parametrisation of Gaussian measures by positive-definite matrices, which we express as $\mu \in N(C)$ or $\phi \sim N(C)$ where the field ϕ is distributed according to the definition above with $A = C^{-1}$.

For $\mu \in N(C)$, one can show that C is the covariance of $d\mu$ ([8], Proposition 2.1.9).

Definition 3.1.3. *Consider Λ as usual. A function F is said to be a local function if it assigns to each site $x \in \Lambda$ an interaction $F(\{x\})$. $F(\{x\})$ is a function of $(\phi(y), y \in \{x\}^*)$, where $\{x\}^*$ is a neighbourhood of $\{x\}$. A function F^Λ, which we define as $F^\Lambda = \prod_{x\in\Lambda} F(\{x\})$, is called a global function on Λ since it depends on all fields on Λ.*

Remark 3.1.4. *A translation T_y by $y \in \mathbb{Z}^d$ acting on a local function F yields $T_y F(\{x\}, \phi) = F(\{x+y\}, T_y\phi)$, where $T_y\phi(x) = \phi(x-y)$. Then a local function is said to be translation invariant if $T_y F = F$ for all $y \in \mathbb{Z}^d$.*

In what's coming we will be interested in the limit of

$$\int d\mu e^{i\phi(a)} e^{i\phi(b)} F^\Lambda \Big/ \int d\mu F^\Lambda$$

as $\Lambda \uparrow \mathbb{Z}^d$. We will conveniently write this integral ratio as

$$\int d\mu F^\Lambda_{a,b} \Big/ \int d\mu F^\Lambda,$$

where $F_{a,b}(\{x\}) = F(\{x\}) e^{i\phi(1_{x=1}+1_{x=b})}$, and this function differs from F only at the points $\{a,b\}$[2]. This limit is the formal realisation of the infinite volume limit that we previously stated to be of interest.

Definition 3.1.5. *Let $\Lambda = \Lambda_N$ be a cube of side L^N with $L \in 2\mathbb{N}+1$[3] and $N \in \mathbb{N}$. That is,*

$$\Lambda = \{x \in \mathbb{Z}^d | ||x||_\infty \leq \frac{1}{2}(L^N - 1)\},$$

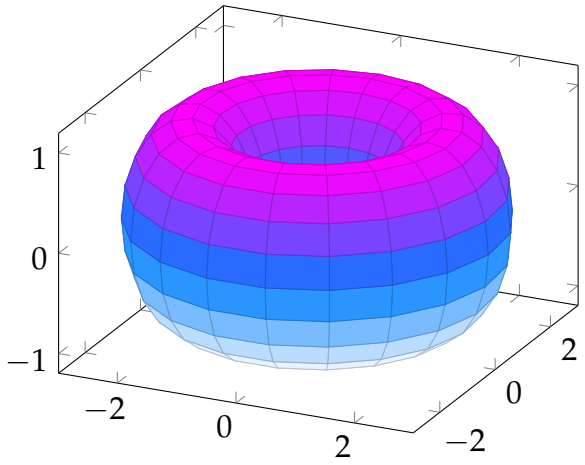
where

$$||x||_\infty = \max_i |x_i|.$$

Remark 3.1.6. *The following should be kept in mind:*
1. Unless stated otherwise, we shall assume that x can be identified with $x + L^N\hat{v}$, where $\hat{v}$ is in the standard basis of $\mathbb{Z}^d$, which we shall call $\mathcal{E}$.
2. For $x, y \in \Lambda$, the addition and subtraction are done componentwise modulo L^N.

[2] In general, one considers any finite number of points.
[3] We adopt this notation for positive odd numbers that are greater than or equal to 3.

An example of the structure we would like to have in mind is a torus, which is a result of the periodic boundary conditions. For example on this diagram below we could imagine that each intersection of the longitude lines as well as the lattitude lines will be assigned a spin, but recall that we are not limiting ourselves to $d = 3$ in general.



Definition 3.1.7. *A matrix $C = (C(x,y)_{x,y\in\Lambda})$ is translation invariant if it is a function of $x - y$.*

Definition 3.1.8. *We say that a positive definite, translation invariant matrix C, as above, admits a finite range decomposition if there exists a sequence of matrices $(C_j)_{j=1}^N$, $N \in \mathbb{N}$, such that $C_j = (C_j(x,y))_{x,y\in\Lambda}$,*

$$C = \sum_{j=1}^{N} C_j,$$

and $C_j(x,y) = 0$ if $|x - y| \geq \frac{1}{2}L^j$.

Suppose that $\phi \sim N(C)$ where C has a finite range decomposition, then it is possible to show that there exist independent $\xi_j \sim N(C_j)$ such that

$$\phi = \sum_{j=1}^{N} \xi_j,$$

with the equality in distribution. The function ξ_j is referred to as a fluctuation fields on scale j. $\xi_j(x)$ and $\xi_j(y)$ are independent if $|x - y| \geq \frac{1}{2}L^j$ because they are Gaussian, and their covariance is zero.

Let μ be a Gaussian measure on $\mathbb{R}^\Lambda$ such that its covariance admits a finite range decomposition $C = \sum_j C_j$, and μ_j be a Gaussian measure whose covariance is C_j. Then

$$\int d\mu F^\Lambda = \int d\mu_N \int d\mu_{N-1} \cdots \int d\mu F^\Lambda.$$

On the left of the equality $F = F(\phi)$, while on the right one inserts $\sum_j \xi_j$ instead of ϕ. This is a consequence of Proposition 2.1.11. in [8].

This can be rewritten by defining $\mathbb{E}_j = \int d\mu_j$, and then setting $Z_0 = F^\Lambda$, $Z_j = \mathbb{E}_j Z_{j-1}$ for $j \in \{1, \cdots, N\}$ and $Z_N = \int d\mu Z_0$. Here $\mathbb{E}_j$ is perceived as a map on global functions, and it is referred to as a renormalisation group transformation or $\tilde{RG}$ map.

Definition 3.1.9. *The renormalisation group is defined as the set* $\{\tilde{RG} = \mathbb{E}_j | j \in \{1, \cdots, N\}\}$.

Definition 3.1.10 (DCMPSTN)**.** *We shall assume the following relationship between fields and fluctuations, which is associated with the Wilsonian procedure of the renormalisation group–the approach that we are interested in.*

$$\phi_j = \sum_{k>j}^{N} \xi_k, \phi_j = \phi_{j+1} + \xi_{j+1}, \text{ for } j \in \{0, \cdots, N-1\}, \begin{cases} \phi_0 = \phi \\ \phi_N = 0 \end{cases}$$

What we have so far is an action on global functions, and we would like to show that this action is equivalent to an action (RG) on local fields F. This is known as the global to local program. We shall do this in the context of hierarchical models, since this is the important context for us, but these ideas can be extended to much more complicated models, like what are known as Euclidean models.

Definition 3.1.11 (Blocks)**.** *For each $j \in \{0, \cdots, N\}$ the torus Λ, still assuming periodic boundary conditions, can be paved in natural way by L^{N-j} cubes which are disjoint. These are cubes of side length L^j with $L \in 2\mathbb{N}+1$ as before. The cube which contains the origin has the form $\{x \in \Lambda | |x| \leq \frac{1}{2}(L^j - 1)\}$, and the rest are translations of this by vectors in $L^j\mathbb{Z}^d$. These cubes are called j-blocks (or just blocks). The notation to denote the set of j-blocks is $\mathcal{B}_j = \mathcal{B}_j(\Lambda)$. When $j = 0$, the blocks are singletons, and $B = \{x\}$ with $x \in \Lambda$.*

Definition 3.1.12 (Polymers)**.** *A union of j-blocks is called a (j-) polymer, and similarly $\mathcal{P}_j = \mathcal{P}_j(\Lambda)$ is the set of j-polymers. Furthermore $\emptyset \in P_j$. For $X \in \mathcal{P}_j$, the set of $j-$ blocks in X is $\mathcal{B}_j(X)$ and the number of j-blocks in X is $|X|_j = |\mathcal{B}_j(X)|$. Lastly, for $X, Y \in \mathcal{P}_j$ the difference $X \backslash Y = \cup_{B \in X, B \notin Y} B$, and this is an element of $\mathcal{P}_j$.*

Definition 3.1.13. *The hierarchical distance between $x,y \in \Lambda$ is side length of the smallest cube in $\cup_j \mathcal{B}_j$ which contains these points. This distance is a metric that we shall denote $dist_h(\cdot,\cdot)$.*

Remark 3.1.14. *In fact, $dist_h$ is an ultrametric, which means that for $x,y,z \in \Lambda$ it is the case that $dist_h(x,y) \leq dist_h(x,z) \vee dist_h(z,y)$*

Definition 3.1.15. *A covariance is said to be hierarchical if $C = \sum_{j=1}^{N} C_j$ where C_j is positive semi-definite and $C_j(x,y) = 0$ if $dist_h(x,y) > L^j$.*

Remark 3.1.16. *Hierarchical covariance is important since the fluctuation fields $\xi_j(x)$ and $\xi_j(y)$ become independent if x and y are not in the same block. This is the basis of the global to local argument for this case.*

We give contextual machinery once again to pave way for results. Let $X \subset \Lambda$ be given, and $\mathcal{N}_j(X)$ be the algebra of functions that are measurable with respect to the σ-algebra generated by $\{\phi_j(x) | x \in X\}$.

Remark 3.1.17. *An element of $\mathcal{N}_j(X)$ is a function only of fields with evaluated at sites $x \in X$. By the DCMPSTN assumption, given on a definition above, $\mathcal{N}_j(X)$ are functions of ϕ_{j+1} and ξ_{j+1} but through only $\phi_{j+1} + \xi_{j+1}$.*

We extend this algebra by defining $\tilde{\mathcal{N}}_j(X)$ as the algebra generated by

$\{\xi_{j+1}, \phi_{j+1} | x \in X\}$. We write $\mathcal{N}_j = \mathcal{N}_j(\Lambda)$, and we do similarly for the extension.

We will consider $\mathcal{N}_j^{\mathcal{B}_j}$ as the set of maps such that $F(B) \in \mathcal{N}_j(B)$[4], where B is a j-block.

Definition 3.1.18. *Let $X \in \mathcal{P}_j$ and $F \in \mathcal{N}_j^{\mathcal{B}_j}$. Define*

$$F^X = \prod_{B \in \mathcal{B}_j(X)} F(B).$$

Remark 3.1.19. *We consider the following conventions and remark:*
1. $F^{\emptyset} = 1$
2. We take sums over null indices as zero.
3. We make the same definitions with $\mathcal{N}_j$ replaced with $\tilde{\mathcal{N}}_j$.

[4]Generally, models require $\mathcal{N}_j(B^*)$ where B^* is the neighbourhood of B.

Theorem 3.1.20. *Let μ be a Gaussian measure whose covariance is hierarchical and let $F \in \tilde{\mathcal{N}}_j^{\mathcal{B}_j}$ be integrable. Then*

$$\mathbb{E}_{j+1} F^{\Lambda} = (F')^{\Lambda},$$

with $F' \in \mathcal{N}_{j+1}^{\mathcal{B}_{j+1}}$ defined for $B' \in \mathcal{B}_{j+1}$ by

$$F'(B') = \mathbb{E}_{j+1} F^{B'},$$

and $j \in \{0, \cdots, N-1\}$.

Definition 3.1.21. *Define $(RG) : \tilde{\mathcal{N}}_j^{\mathcal{B}_j} \to \mathcal{N}_{j+1}^{\mathcal{B}_{j+1}} \subset \tilde{\mathcal{N}}_{j+1}^{\mathcal{B}_{j+1}}$*

Proof. We start by recalling that $\xi_{j+1}(x)$ and $\xi_{j+1}(y)$ are independent if x and y are in different $j+1$-blocks. Hence,

$$\begin{aligned}
\mathbb{E}_{j+1} F^{\Lambda} &= \mathbb{E}_{j+1} \prod_{B \in \mathcal{B}_j} F(B) \\
&= \mathbb{E}_{j+1} \prod_{B' \in \mathcal{B}_{j+1}} \prod_{B \in \mathcal{B}_j(B')} F(B) \\
&= \prod_{B' \in \mathcal{B}_{j+1}} \mathbb{E}_{j+1} \prod_{B \in \mathcal{B}_j(B')} F(B) \\
&= \prod_{B' \in \mathcal{B}_{j+1}} \mathbb{E}_{j+1} F^{B'} \\
&= (RG)(F)(B')
\end{aligned}$$

□

This last point ends the global to local programme, and we have the desired equivalence of actions.

The infinite volume limit, at this point, can be written as

$$(RG)^N F_{a,b} / (RG)^N F.$$

We will consider infinite iterations of (RG) that bring the local function to the fixed point $F(B)$ for all B. In these cases, the iteration on $F_{a,b}$ may converge to the same point up to a constant, and in this case we will call the constant the formal infinite

volume limit. In order to make identification with

$$\lim_{N\to\infty} \int d\mu F_{a,b} / \int d\mu F^{\Lambda_N},$$

one needs to show that the final (RG) iteration is continuous near 1. We would like to say more about how one goes about studying this limit.

The last context required is to set up domains on which (RG) can be studied. We consider the setting such that for each scale j a norm exists on $\tilde{\mathcal{N}}_j$ such that

$$||F^X|| \leq ||F||^X$$

where $||F||^X = \prod_{B\in\mathcal{B}_j(X)} ||F(B)||$ for $F \in \tilde{\mathcal{N}}_j$ and

$$\mathbb{E}_{j+1} : \tilde{\mathcal{N}}_j \to \mathcal{N}_{j+1} \subset \tilde{\mathcal{N}}_{j+1}$$

where $||\mathbb{E}_{j+1}Z|| \leq ||Z||$, and this norm is complete. This completeness ensures that the finite norm elements on $\tilde{\mathcal{N}}_j$ are a Banach space, which we denote with the same notation, $\tilde{\mathcal{N}}_j$. The norm must be such that $\mathcal{N}_j$ is a closed subspace. Other spaces get their norms as subspaces of products (cartesian), for example $F \in \tilde{\mathcal{N}}_j^{\mathcal{B}_j}$ the norm is $\max\{||F(B)|||B \in \mathcal{B}_j\}$.

Definition 3.1.22. *Let B_X denote an open ball centered at the origin on a Banach space X. We say that a function defined on such a ball, with values in another Banach space, is smooth (near the origin) if it is $\mathcal{C}^2$, in the sense of having two Frechet derivatives which are defined and continuous on the ball.*

Theorem 3.1.23. *The map $(RG) : \tilde{\mathcal{N}}_j^{\mathcal{B}_j} \to \mathcal{N}_{j+1}^{\mathcal{B}_{j+1}}$ is a smooth map of Banach spaces and the derivative $D(RG)_F$ of (RG) at F in the direction $\dot{F}$ is*

$$D(RG)_F\dot{F}(B') = \sum_{B\in\mathcal{B}_j(B')} \mathbb{E}_{j+1}F^{B'\setminus B}\dot{F}(B)$$

Proof. See ([10], Lemma 2.12). □

An important result in the proof is that

$$||D(RG)_F\dot{F}|| \leq L^d(||F||)^{L^d-1}||\dot{F}||.$$

The L^d is a sign of what is called expanding direction in (RG), the most obvious of which is shown by

$$(RG)(e^{\lambda}F) = e^{L^d\lambda}(RG)(F).$$

We call 1 a relevant operator in the Wilsonian paradigm. Here the "1" is the function of fields which doesn't depend on them, and λ is its coefficient. The notion of expanding directions is what makes it hard to study the convergence of the infinite volume limit.

The strategy we take to study the action of (RG) on $F \in \mathcal{N}_j^{\mathcal{B}_j}$ is to decompose the local function to get $F = I + K$, where we can compute the action $\mathbb{E}_{j+1}$ for $I \in \mathcal{N}_j^{\mathcal{B}_j}$. This part carries the part of F that expands. The part $K \in \mathcal{N}_j^{\mathcal{B}_j}$ is an error that contains the contract parts as well as parts that are small in comparison to I.
The directions that expand and contract are changing along the orbit of the action, so we have to make a change of coordinates of the two parts above each time we have an action. This is the role of $\tilde{I}$ in what follows. We state the following theorem which we shall not prove.

Theorem 3.1.24. *For any integrable $\tilde{I} \in \mathcal{N}_{j+1}^{\mathcal{B}_j}$,*

$$(RG)(I + K) = I' + K',$$

where for $B' \in \mathcal{B}_j$,

$$I'(B') = \tilde{I}^{B'},$$

and

$$K'(B') = \sum_{B \in \mathcal{B}_j(B')} \tilde{I}^{B' \setminus B} \mathbb{E}_{j+1}(K + I - \tilde{I})(B) + O(||K + I - \tilde{I}||^2),$$

where $O(||K + I - \tilde{I}||^2)$ is a smooth function of $(K, I, \tilde{I}) \in \mathcal{N}_j^{\mathcal{B}_j} \times \mathcal{N}_j^{\mathcal{B}_j} \times \mathcal{N}_{j+1}^{\mathcal{B}_j}$ whose norm is bounded as indicated.

Proof. See ([10], Lemma 2.14) □

Lastly, for each scale j we define an element $F = 1$ which lives in $\mathcal{N}_j^{\mathcal{B}_j}$ as $F(B) = 1$ for all $B \in \mathcal{B}_j$. (RG) takes $1 \in \mathcal{N}_j^{\mathcal{B}_j}$ to $1 \in \mathcal{N}_{j+1}^{\mathcal{B}_{j+1}}$. Even though these Banach spaces are not the space, this is referred to as the trivial fixed point.

One can note at this point that

$$D(RG)_1\dot{F}(B') = \sum_{B\in\mathcal{B}_j(B')} \mathbb{E}_{j+1}\dot{F}(B),$$

and to make the infinite volume argument more precise, we give the following definition.

Definition 3.1.25. *If* $\lim_{N\to\infty} ||(RG)^N F - 1|| = 0$ *and there exists a constant* A *such that* $\lim_{N\to\infty} ||(RG)^N F - A1|| = 0$ *then one calls* A *the formal infinite volume limit.*

At this point the plan is to consider (RG) as a map on pairs (I,K), in which K will be an element of a Banach space that may depend on the scale j and I will be determined explicitly by parameters $\lambda \in \mathbb{R}^\gamma$ for some $\gamma \in \mathbb{N}$. These parameters are called coupling constants–these are coordinates for the non-contracting directions, relevant and marginal operators in Wilsonian terminology. In addition to this, I will be such that the trivial fixed point is $(\lambda,K) = (0,0)$. Using the *stable manifold theorem* ([10], Theorem 2.16), it is possible to show that (λ,K) ends up on the fixed point if λ is correctly chosen, a process called tuning. Theories defined on these fixed points are well-studied, and this is where we get possible effective descriptions of the systems at large scales.

3.2 Example Application

The content of this section will conceal the machinery in the previous chapter, but we go over it because it gives a simple way to concretise the important ideas in a context that is much closer to what we are concerned with for the rest of the thesis. We encourage the reader to see the example with a perturbative element, similar to what we in the theoretical coverage, from Chapter 3 of [10] and a different one on the discussion of [11]. A QFT approach to this section may also be found at [12]. What follows is derived from [11] and [13].

Consider the following picture for intuition in the few paragraphs below.

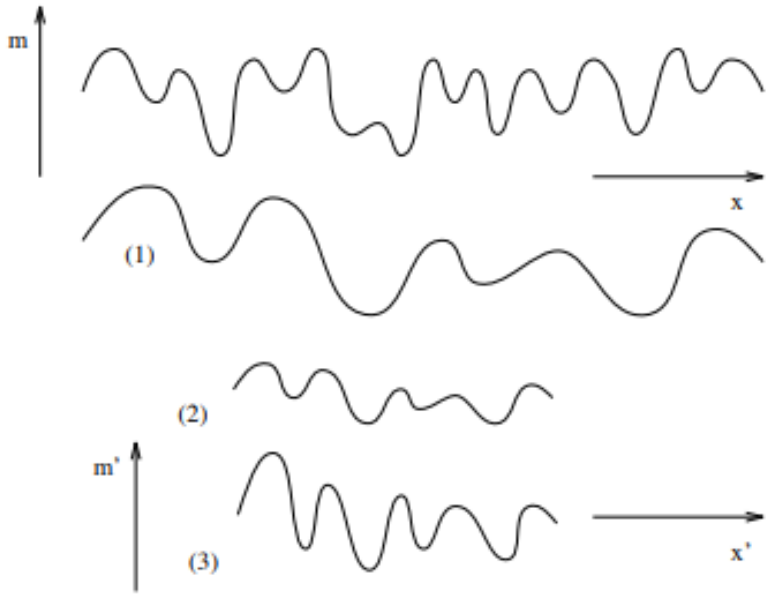


Fig. 3.1 A Rough Schematic of the Renormalisation Procedure [11]

Concerning the mentioned approach of interest, we usually consider a hamiltonian H dependent on fields m, for example this could be the magnetisation field, and define three operations:

1. Changing the minimum length scale for allowed variations of the field (fluctuations) from $a \in \mathbb{R}$ to ba where $b \in \mathbb{R}$ and $b > 1$. That is, consider

$$\tilde{m}(x) = \frac{1}{(ba)^d} \int_{c(x)} m(y)dy,$$

where $c(x)$ is a cell centred at x. This is known as coarse-graining.

2. The action above amounts to changing the resolution of the picture given above. This decrease in resolution makes this picture grainier, and the original resolution can be restored by considering

$$x' = \frac{x}{b}.$$

   This is known as rescaling.

3. In the same way, there is a need to correct for the variations in fluctuations in the rescaled image, which is done by introducing some renormalisation factor ζ, such that

$$\tilde{m'}(x') = \frac{1}{\zeta}\tilde{m}(x').$$

The key insight is that on length scales that are smaller than the correlation length, a quantity that we saw in the context of the Ising model, but that we can define similarly in other contexts, then the renormalised configurations are statistically similar to the initial configurations. This is a statement that says that they may be distributed by similar hamiltonians, with similarity defined in a distributional sense. What this tells us is that if the original hamiltonian is driven to a fixed point by parameter tuning (to zero), then the second will achieve the same state with the scaled parameters. The correlation length scales as ξ/b.

From this we get RG as a map that takes parameters associated with the initial configuration to the parameters associated with the renormalised configurations. In general this mapping is non-linear. Hamiltonians that correspond to statistically self-similar configurations correspond to fixed points of this map, since the transformations themselves describe dilation effects on the hamiltonian of the system. The correlation length scales as mentioned above, and as such the correlation at fixed points has to be either zero or infinity. In the former case, the description is that of independent fluctuations at each point, which corresponds to complete disorder (infinite temperature) or complete order (zero temperature[5]), and the latter describe critical points. We can then linearise the RG transformation on the parameters, which paves the way to discussing flows and as above, but we omit this explicit discussion, and consider a practical settting below.

We shall consider the renormalisation procedure applied to the 1-dimensional Ising model following discussions in [14] and [15].

[5]The lowest temperature essentially.

Consider the Ising model context. The renormalisation group here is the grouping of spins into blocks using their couplings so that there are fewer sites, but the structure of the lattice is retained. Each block gets an associated spin that we will assume can be obtained by choosing the most frequent spin, by letting one spin determine what the block spin will be (which works better in low temperatures), or by averaging. In order to determine block spin, we define let $f(\sigma_i)$ be a function for σ_i inside a chosen block. The new renormalised lattice will have a Hamiltonian $H'(\sigma')$ where σ' is the block spin. The old partition function is then $Z = Tr_\sigma e^{-\beta H(\sigma)}$, while the new one is $Z' = Tr_{\sigma'} e^{-\beta H'(\sigma')}$. Using the formal definition of the renormalised partition,

$$e^{-\beta H'(\sigma'_J)} = Tr_\sigma \prod_J (\delta(\sigma'_J - f(\sigma_i))) e^{-\beta H(\sigma)},$$

one can show that the two partition functions are the same.
The Hamiltonian for the 1-dimensional system is

$$H = -K \sum_{\langle i \rangle} \sigma_i \sigma_{i+1} - h \sum_i \sigma_i,$$

with a magnetic field. If we remove all odd sites then the remaining spins can be perceived as the blocks. The energy contribution of each removed atom, i, is

$$H_i = -K\sigma_i(\sigma_{i-1} + \sigma_{i+1}) - h\sigma_i - \frac{h(\sigma_{i-1} + \sigma_{i+1})}{2},$$

so that the partition function is the sum over these, which is

$$Z = 2\cosh(K(\sigma_{i-1} + \sigma_{i+1}) + h) e^{\frac{h(\sigma_{i-1} + \sigma_{i+1})}{2}}.$$

One the other hand, we write the renormalised Hamiltonian as

$$H' = -K' \sum_{\langle i \rangle} \sigma_{i-1}\sigma_{i+1} - h' \sum_i \sigma_i - \sum_i g'(K),$$

where the sum of $g'(K)$ anticipates site energy constants.

Using the equality of partition functions we obtain

$$2\cosh(K(\sigma_{i-1}+\sigma_{i+1})+h)e^{\frac{h(\sigma_{i-1}+\sigma_{i+1})}{2}} = e^{K'\sum_{\langle i\rangle}\sigma_{i-1}\sigma_{i+1}+h'\sum_i\sigma_i+\sum_i g'(K)}$$

Finally by considering the cases where both spins have the same orientation in the positive and negative direction as well as when they are opposite they get relations that yield the following solutions to the constants.

$$h' = \frac{1}{2}\ln\left(\frac{\cosh(2K+h)}{\cosh(-2K+h)}\right)+h,$$

$$K' = \frac{1}{4}\ln\left(\frac{\cosh(2K+h)\cosh(-2K+h)}{\cosh^2 h}\right),$$

and

$$g'(K) = \frac{1}{4}\ln(\cosh(2K+h)\cosh(-2K+h)\cosh^2 h)+\ln 2.$$

At this point, we can talk about fixed points of this system. We consider the simple case with $h=0$.

Considering K', we find that this yields $e^{2K'} = \frac{e^{2K}+e^{-2K}}{2}$.

Unless $K=\infty$, K' is a decreasing function in K[6], and so there are two fixed points, with the second being $K=0$. K is a constant that is proportional to the inverse temperature. Hence, as K' decreases the system flows towards $T=\infty$. If $K=\infty$, then the system stays at $T=0$.

We can visualise the flows for this example as follows.

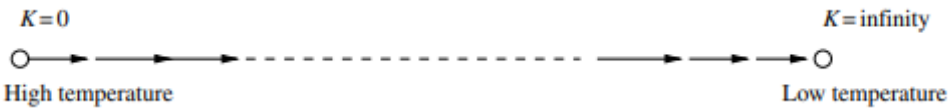


Fig. 3.2 Flow Visualisation for the 1D Ising [15]

In this example, there are no critical fixed points. Such fixed points lead to phase transitions, and emergence of interesting properties. The 2-dimensional model

[6]We think of the equation above as being recursive.

exhibits a phase transition, as suggested by examples in Chapter 2, and we shall say more about this in the chapters that are coming up.

Chapter 4

Energy Based Models

4.1 Theory

4.1.1 Restricted Boltzmann Machines' Theory

In this chapter, we shall discuss Restricted Boltzmann Machines (RBMs), which fall under the umbrella of a class of architectures known as energy based models. This is the context in which we shall see the ideas behind possible connections between deep learning and the renormalisation group being explored. One of the pioneering papers for this architecture is [16], but for this discussion we follow closely what is on [17]. The paper [18] offers further insights on the structure of RBMs from an algebraic geometric perspective, but we shall not go into that much detail.

In order to get to the definition of RBMs, we need to introduce some graph structure, the idea of conditional independence given random variables, as well as define what is known as the global Markov property. We do this below.

Definition 4.1.1. *A graph is a pair $G = (V, E)$ such that V is a finite set of vertices, and E is the set of edges between them.*

Remark 4.1.2. *We shall only consider undirected graphs where the edges have no direction information associated with them.*

Definition 4.1.3. *Let $u, v \in V$, and $S \subset V$ such that any path from u to v contains a node from S, then we say that S separates u and v or is a $u - v$ separating set.*

Definition 4.1.4. *Consider a measure space $(\Omega, \mathcal{A}, p)$. Let X_1, X_2 and X_3 be random variables. We say that X_1 and X_2 are conditionally independent given X_3 if*

$$p(X_1, X_2|X_3) = p(X_1|X_3)p(X_2|X_3).$$

Definition 4.1.5. *Let X_v be a random variable with values in Λ_v for $v \in V$ given*

$G = (V, E)$. Let p be an associated measure. The random variables $\boldsymbol{X} = (X_v)_{v\in V}$ are called a Markov random field (MRF) if for any partition (A, B, S) of V with all vertices of A separated from B by vertices S we have that $(X_a)_{a\in A}$ and $(X_b)_{b\in B}$ are conditionally independent given $(X_s)_{s\in S}$. We say that p satisfies the global Markov property.

Finally, a RBM is defined as follows.

Definition 4.1.6. *A restricted Boltzmann machine (RBM) is defined as a MRF associated with a bipartite graph $K_{m,n}$. We refer to the $m-$vertex set as the set of visible units, which we shall denote $\boldsymbol{V} = (V_1, \cdots, V_m)$. Similarly, the hidden units are the other set of disjoint vertices, which we denote $\boldsymbol{H} = (H_1, \cdots, H_n)$. We consider $\boldsymbol{V}$ and $\boldsymbol{H}$ as random variables.*

Remark 4.1.7. *We shall assume, unless stated otherwise, that the random variables $(\boldsymbol{V}, \boldsymbol{H})$ take values $(\boldsymbol{v}, \boldsymbol{h}) \in \{0,1\}^{m+n}$.*

The following definition further establishes the structure of the RBM context that we are considering.

Definition 4.1.8. *We define the joint probability under the model as a Gibbs distribution*

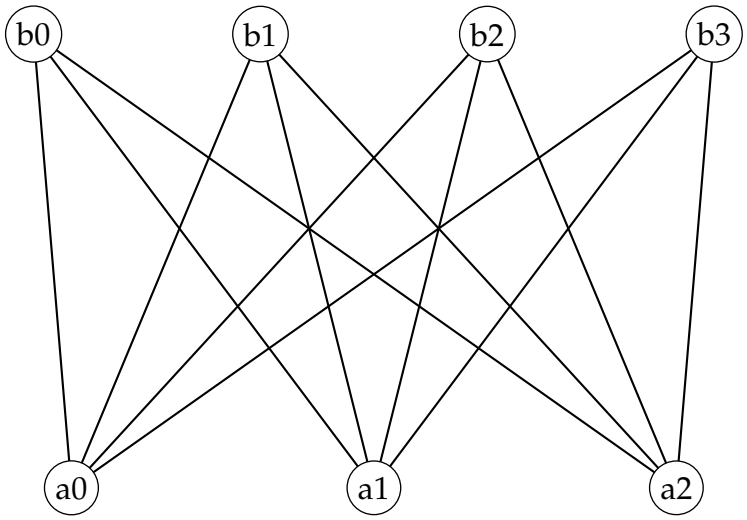
$$p(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} e^{-E(\boldsymbol{v}, \boldsymbol{h})},$$

where

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} h_i v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i,$$

where w_{ij} are real weights between units V_j and H_i, and $b_i, c_i \in \mathbb{R}$ are bias terms correspond to the variables v_j and h_i.

Hence, the rough picture we want to keep in mind is the following.



where label $\mathbf{V} = (b0, b1, b2, b3)$ and $\mathbf{H} = (a0, a1, a2)$, and we have an assignment of weights to the edges of the graph which are the w_{ij}.

The assumed task is that one would like to model an m-dimensioal unknown probability distribution μ, and usually it is not true that all variables $\boldsymbol{X} = (X_v)_{v \in V}$ in an MRF need to have correspondence to some observed component. This split between visible and hidden units falls out from this situation. The hidden variables allow the description of complex distributions over the visible part via conditional probabilities–they introduce dependencies between the visible variables, which correspond to components of observation.

Remark 4.1.9. *The bipartite structure of the associated graph means that the hidden variables are independent given the state of the visible variables and vice-versa. This is to say that, considering an associated measure p,*

$$p(\boldsymbol{h}|\boldsymbol{v}) = \prod_{i=1}^{n} p(h_i|\boldsymbol{v}),$$

and

$$p(\boldsymbol{v}|\boldsymbol{h}) = \prod_{i=1}^{m} p(v_i|\boldsymbol{h}).$$

The following result gives as an analytical expression for the probability distribution over the visible units, and we can do similar for the hidden units.

Claim 4.1.10. *In this context,*

$$p(\boldsymbol{v}) = \frac{1}{Z}\prod_{j=1}^{m} e^{b_j v_j} \prod_{i=1}^{n}\left(1 + e^{c_i + \sum_{j=1}^{m} w_{ij} v_j}\right).$$

Proof. Observe that

$$\begin{aligned} p(\boldsymbol{v}) &= \sum_{\boldsymbol{h}} p(\boldsymbol{v}, \boldsymbol{h}) \\ &= \frac{1}{Z}\sum_{\boldsymbol{h}} e^{-E(v,h)} \\ &= \frac{1}{Z}\sum_{h_1}\cdots\sum_{h_n} e^{\sum_{j=1}^{m} b_j v_j} \prod_{i=1}^{n} e^{h_i\left(c_i + \sum_{j=1}^{m} w_{ij} v_j\right)} \\ &= \frac{1}{Z} e^{\sum_{j=1}^{m} b_j v_j} \prod_{i=1}^{n}\sum_{h_i} e^{h_i\left(c_i + \sum_{j=1}^{m} w_{ij} v_j\right)}. \end{aligned}$$

At this point, the expression we want pops out via properties of the exponential and the remark on the values that $\boldsymbol{h}$ can take. □

In what follows, we shall assume that the context is as above, unless otherwise stated. The following results tells us about the modelling abilities of RBMs in the context of recovering distributions mentioned above. We do not prove this result, but a reference is given.

Theorem 4.1.11. *Any probability distribution can be modelled arbitrarily well by an RBM with m visible and $k+1$ hidden units where k denotes the number of input elements from $\{0,1\}^m$ with non-zero probability of being observed, which is also known as the cardinality of the support set of the target distribution.*

Proof. See ([19], Theorem 2.4). □

In what follows the main task is to talk about gradients of the likelihood function that we will define. The process of recovering a distribution amounts to *skilfully* tuning the weights and biases in order to get a model that reproduces the correct underlying description. We can go about this by defining a likelihood function whose gradients we then use to achieve this effective tuning of parameters. What

follows below is a discussion of this. First, we discuss some conditional dependences between the hidden units and the visible units.

The sigmoid function is defined

$$\sigma(x) = 1/(1 + e^{-x}).$$

Using Bayes' rule of conditional probabilities, it can be shown that

$$p(H_i = 1|\boldsymbol{v}) = \sigma\left(\sum_{j=1}^{m} w_{ij}v_j + c_i\right),$$

and

$$p(V_j = 1|\boldsymbol{h}) = \sigma\left(\sum_{i=1}^{m} w_{ij}h_i + b_j\right).$$

Proof. We prove the second statement. Let $\boldsymbol{v}_{-l}$ denote state of visible units except the lth unit. Let $\alpha_l(\boldsymbol{h}) \equiv -\sum_{i=1}^{n} w_{il}h_i - b_l$, and

$\beta(\boldsymbol{v}_{-l}, \boldsymbol{h}) \equiv -\sum_{i=1}^{n}\sum_{j=1,j\neq l}^{m} w_{ij}h_i v_j - \sum_{j=1,j\neq l}^{m} b_i v_i - \sum_{i=1}^{n} c_i h_i$. Notice that

$E(\boldsymbol{v}, \boldsymbol{h}) = \beta(\boldsymbol{v}_{-l}, \boldsymbol{h}) + v_l\alpha_l(\boldsymbol{h})$, where the latter term collects all v_l terms.

$$\begin{aligned}
p(V_l = 1|\boldsymbol{h}) &= p(V_l = 1|\boldsymbol{v}_{-l}, \boldsymbol{h}) \\
&= \frac{p(V_l = 1, \boldsymbol{v}_{-l}, \boldsymbol{h})}{p(\boldsymbol{v}_{-l}, \boldsymbol{h})} \\
&= \frac{e^{-E(v_l=1,\boldsymbol{v}_{-l},\boldsymbol{h})}}{e^{-E(v_l=1,\boldsymbol{v}_{-l},\boldsymbol{h})} + e^{-E(v_l=0,\boldsymbol{v}_{-l},\boldsymbol{h})}} \\
&= \frac{e^{-\beta(\boldsymbol{v}_{-l},\boldsymbol{h})} \cdot e^{-\alpha_l(\boldsymbol{h})}}{e^{-\beta(\boldsymbol{v}_{-l},\boldsymbol{h})} \cdot (e^{-\alpha_l(\boldsymbol{v}_{-l},\boldsymbol{h})} + 1)} \\
&= \frac{1}{1 + e^{\alpha_l(\boldsymbol{v}_{-l},\boldsymbol{h})}} \\
&= \sigma(-\alpha_l(\boldsymbol{h}))
\end{aligned}$$

□

We can do similar for $p(H_i = 1|\boldsymbol{v})$, but we omit this as the procedure is similar. Now, we would like to define the likelihood function.

Definition 4.1.12. *Define the likelihood function of an MRF given a set of independent and identically distributed variables, S, called the training set, as a function $\mathcal{L} : \Theta \to \mathbb{R}$ such that*

$$\mathcal{L}(\theta|S) = \prod_{i=1}^{l} p(\boldsymbol{x}_i|\boldsymbol{\theta}),$$

for some parameters θ.

The problem of recovering an unknown distribution given samples applied to RBMs amounts to finding parameters that maximise the likelihood. In practice, people usually talk about the logarithm of the likelihood (the log-likelihood). The following derivatives can be obtained with relative ease from the usual methods from calculus, and are worth mentioning as part of the theory.

1. $$\frac{\partial \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{v})}{\partial w_{ij}} = p(H_i = 1|\boldsymbol{v})v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v})p(H_i = 1|\boldsymbol{v})v_j.$$

2. $$\frac{\partial \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{v})}{\partial b_j} = v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v})v_j.$$

3. $$\frac{\partial \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{v})}{\partial c_i} = p(H_i = 1|\boldsymbol{v}) - \sum_{\boldsymbol{v}} p(\boldsymbol{v})p(H_i = 1|\boldsymbol{v}).$$

We now discuss briefly how one comes to these equations, taking the first example. Looking at the first equation,

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{v})}{\partial w_{ij}} &= -\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})\frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial w_{ij}} + \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h})\frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial w_{ij}} \\ &= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i v_j,\end{aligned}$$

which then gives us the right hand side of 1.

Remark 4.1.13. *The procedure to get the first line is straightforward differentiation, noting that*

$$\begin{aligned}
\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) \frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial w_{ij}} &= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j \\
&= \sum_{\boldsymbol{h}} \prod_{k=1}^{n} p(h_k|\boldsymbol{v}) h_i v_j \\
&= \sum_{h_i} \sum_{\boldsymbol{h}_{-i}} p(h_i|\boldsymbol{v}) p(\boldsymbol{h}_{-i}|\boldsymbol{v}) h_i v_j \\
&= \sum_{h_i} p(h_i|\boldsymbol{v}) h_i v_j \sum_{h_{-i}} p(\boldsymbol{h}_{-i}|\boldsymbol{v}) \\
&= p(H_i = 1|\boldsymbol{v}) v_j \\
&= \sigma(\sum_{j=1}^{m} w_{ij} v_j + c_i) v_j
\end{aligned}$$

For the mean of the derivative in question over a training set $S = \{v_1, \cdots, v_l\}$, we often write

$$\begin{aligned}
\frac{1}{l} \sum_{\boldsymbol{v} \in S} \frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{v})}{\partial w_{ij}} &= \frac{1}{l} \sum_{\boldsymbol{s} \in S} \left(- \mathbb{E}_{p(\boldsymbol{h}|\boldsymbol{v})} \left(\frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial w_{ij}} \right) + \mathbb{E}_{p(\boldsymbol{h},\boldsymbol{v})} \left(\frac{\partial E(\boldsymbol{v},\boldsymbol{h})}{\partial w_{ij}} \right) \right) \\
&= \frac{1}{l} \sum_{\boldsymbol{v} \in S} \left(\mathbb{E}_{p(\boldsymbol{h}|\boldsymbol{v})}[v_i h_j] - \mathbb{E}_{p(\boldsymbol{h},\boldsymbol{v})}[v_i h_j] \right) \\
&= \langle v_i h_j \rangle_{p(\boldsymbol{h}|\boldsymbol{v}) q(\boldsymbol{v})} - \langle v_i h_j \rangle_{p(\boldsymbol{h},\boldsymbol{v})},
\end{aligned}$$

where q is the empirical distribution. This leads to the rule

$$\sum_{\boldsymbol{v} \in S} \frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{v})}{\partial w_{ij}} \propto \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model},$$

for updating the weights.

Chapter 5

Relating Aspects of Deep Learning and the Renormalisation Group

We saw in chapter 3 that the renormalisation group extracts relevant features by marginalising over short distance degrees of freedom. The procedure starts at a microscopic scale and, through iterations, moves towards large scale fluctuations. The major draw back is that it is often impossible to do the renormalisation group procedure exactly, and approximate procedures include a class of variational real-space renormalisation schemes introduced by Khadanoff on the relevant papers [20], [21] and [22]. This is the idea of variational renormalisation group. This chapter is derived from [23].

Consider $\Lambda \subseteq \mathbb{Z}^d$ with spins $\{\sigma_x\}$ for $x \in \Lambda$, such that $\sigma_x \in \{-1,1\}$ and $|\Lambda| = N$. Assume that the spins Gibbs distributed and let σ denote a configuration of spins corresponding to the sites. We write $p(\sigma) = \frac{e^{-H(\sigma)}}{Z}$[1], as in Chapter 2, where $H(\sigma)$ is the hamiltonian and Z is the partition function. In this context, when we write $f(\sigma,\sigma')$, for some function f, where σ' is another configuration of interest, then we mean $f(\{\sigma_x,\sigma'_y\})$.

The hamiltonian is typically dependent on a set of couplings, $\boldsymbol{K} = \{K_s\}$, that parametrises the set of all possible hamiltonians. An example with binary spins is the following

$$H(\sigma) = -\sum_x K_x\sigma_x - \sum_{x,y} K_{xy}\sigma_x\sigma_y - \sum_{x,y,z} K_{xyz}\sigma_x\sigma_y\sigma_z + \cdots,$$

[1]Here we have set the temperature to 1, but we do this without loss of generality in the discussion.

where $x, y, z \in \Lambda$. Here the coupling constants qualify spin interactions of different orders.

Definition 5.0.1. *Consider the spin system above. We define the free energy of the spin system as* $F^{\sigma} = -\log Z$.

Let σ' denote a field over spins on $\Lambda' \subset \Lambda$ with $|\Lambda'| = M < N$. We view these as a coarse-graining of the first system. Each of the $x' \in \Lambda'$ is represents a collection of spins on the initial lattice whose small scale fluctuations have been integrated out.

An example is illustrated in the following.

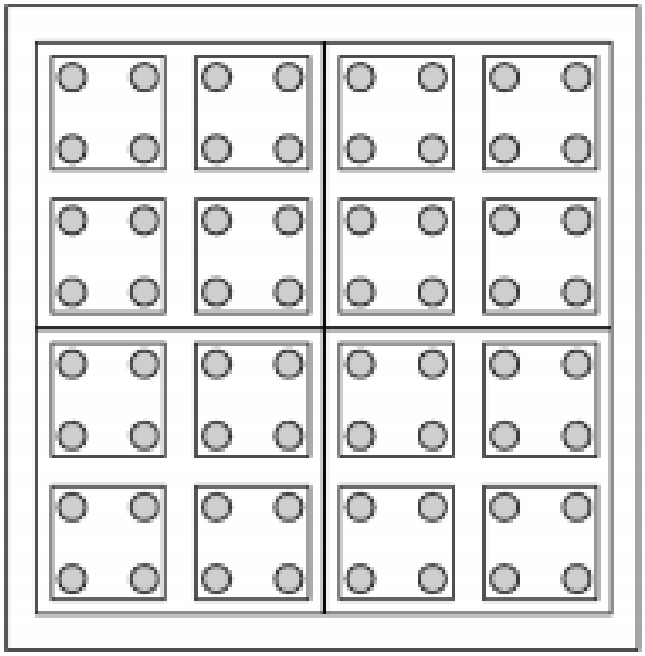


Fig. 5.1 A Coarse-Graining Example for d=2[24]

In this picture, the grey points act as the original system, and the clusters of four can be regarded as coarse-grained spins in the new picture.

We define a hamiltonian over these spins as $H'(\sigma') = -\sum_x \tilde{K}_x \sigma'_x - \sum_{x,y} \tilde{K}_{xy} \sigma'_x \sigma'_y - \sum_{x,y,z} \tilde{K}_{xyz} \sigma'_x \sigma'_y \sigma'_z + \cdots$.

We then define the renormalisation group map (RG) as the a map from $\{\boldsymbol{K}\} \to \{\tilde{\boldsymbol{K}}\}$.

Definition 5.0.2. *Define an operator* $T_\lambda(\sigma, \sigma')$ *such that this operator depends on parameters* $\{\lambda\}$ *and*

$$e^{-H'_\lambda(\sigma)} \equiv Tr_{\sigma_x} e^{T_\lambda(\sigma,\sigma') - H(\sigma)}.$$

Remark 5.0.3. *We think of $T_\lambda(\sigma,\sigma')$ as an operator that encodes pairwise interactions between physical and coarse-grained degrees of freedom.*

In the same way above, we define the free energy for the coarse-grained system as $F_\lambda^{\sigma'} = -\log Z'_\lambda$ with Z'_λ defined from the system defined by H'_λ.

With Khadanoff's procedure, the goal is to ensure that long-distance physical observables are invariant to the course graining process by choosing parameters $\{\lambda\}$ such that we minimise the function

$$\Delta F = F_\lambda^{\sigma'} - F^\sigma.$$

Note that

$$\Delta F = 0 \iff Tr_{h_j} e^{T_\lambda(\sigma,\sigma')} = 1,$$

whenever this holds we say that the renormalisation group transformation is exact. In general, however, this is not possible to do.

We shall now show that this variation has a natural interpretation as deep scheme based on RBMs.

Consider the RBM model again with a set of N visible units σ, and M hidden nodes σ' with the associated measure p_λ, where $\lambda = \{h_j, w_{ij}, c_i\}$ is a set of real parameters. We define the energy as in the previous chapter and write

$$E(\sigma,\sigma') = -\sum_j b_j h_j - \sum_{i,j} v_i w_{ij} h_j - \sum_i c_i v_i,$$

with joint probability

$$p_\lambda(\sigma,\sigma') = e^{-E(\sigma,\sigma')}/Z.$$

We shall define a variational RBM hamiltonian for $p_\lambda(\sigma)$ which we define as

$$p_\lambda(\sigma) \equiv e^{-H''_\lambda(\sigma)}/Z,$$

and similarly $p_\lambda(\sigma')$.

Observe that $T_\lambda(\sigma,\sigma')$ in variational renormalisation group plays the role of $E(\sigma,\sigma')$ in RBM theory. We shall show that these quantities are related through the following equation

$$T(\sigma,\sigma') = -E(\sigma,\sigma') + H(\sigma).$$

This map is a one-one mapping between the variational RG scheme and RBMs.

Claim 5.0.4. *In this context,*

$$H''_\lambda = H'_\lambda$$

Proof. Recall that

$$e^{-H'_\lambda(\sigma)} \equiv Tr_{\sigma_x} e^{T_\lambda(\sigma,\sigma')-H(\sigma)}.$$

Now

$$\begin{aligned} e^{-H'_\lambda(\sigma')}/Z &= Tr_{v_i} e^{T_\lambda(\sigma,\sigma')-H(\sigma)}/Z \\ &= Tr_{v_i} e^{-E(\sigma,\sigma')}/Z \\ &= p_\lambda(\sigma') \\ &= e^{-H''_\lambda(\sigma')}/Z. \end{aligned}$$

□

This equivalence turns out to hold if and only if we have that

$$p_\lambda(\sigma) = \sum_{\{\sigma_i\}} p_\lambda(\sigma,\sigma') = Tr_{\sigma_i} p_\lambda(\sigma,\sigma'),$$

and similar for $p_\lambda(\sigma')$. If this is not the case then there are counter examples [25]. Lastly, one may observe that

$$\begin{aligned} e^{T(\sigma,\sigma')} &= e^{-E(\sigma,\sigma')+H(\sigma)} \\ &= \frac{P_\lambda(\sigma,\sigma')}{P_\lambda(\sigma)} e^{H(\sigma)-H''_\lambda(\sigma)} \\ &= P_\lambda(\sigma|\sigma') e^{H(\sigma)-H''_\lambda(\sigma)}. \end{aligned}$$

This implies that when RG is exact then $H(\sigma) = H''_\lambda(\sigma)$ and $T_\lambda(\sigma,\sigma')$ is the conditional probability.

Chapter 6

Insights into Computational Investigations and Recent Results

The goal of this chapter is to discuss how one would go about investigating this mentioned connection numerically, and progress that has been made with this regard. We derive the content from one of the recent papers in the field [25]. The nature of this will be largely discursive.

Definition 6.0.1. *Consider the RBM context in Chapter 5. Define the Kullback-Leibler divergence as*

$$D_{KL}(p(\sigma)||p_\lambda(\sigma)) = \sum_\sigma p(\sigma) \log\left(\frac{p(\sigma)}{p_\lambda(\sigma)}\right).$$

Minimising this quantity is equivalent to maximising the log-likelihood [26]. When this is done in training then the RBM gives exactly the data distribution.

We consider a KL-divergence approach moving forward, and as we did with the log-likelihood, we can write the gradients as

$$\frac{\partial D_{KL}(q||p)}{\partial W_{ia}} = \langle v_i h_a \rangle_{data} - \langle v_i h_a \rangle_{model},$$

$$\frac{\partial D_{KL}(q||p)}{\partial b_i^{(v)}} = \langle v_i \rangle_{data} - \langle v_i \rangle_{model},$$

and

$$\frac{\partial D_{KL}(q||p)}{\partial b_a^{(h)}} = \langle h_a \rangle_{data} - \langle h_a \rangle_{model}.$$

[27] shows an approach to computing these gradients, if one is interested. Computing the gradients then proceeds by an approximation strategy, because the exact computations are computationally expensive. We proceed by considering a basic sampling strategy:

Assuming that one is a given a vector of visible vectors, then one can sample hidden vectors by setting $h_a = 1$ *with probability*

$$p(h_a = 1|v) = \frac{1}{2}\left(1 + \tanh(\sum_i W_{ia} v_i + b_a^{(h)})\right),$$

and given the vectors in the reversed order, then we similarly set $v_i = 1$ *with probability*

$$p(v_i = 1|h) = \frac{1}{2}\left(1 + \tanh(\sum_a W_{ia} h_a + b_i^{(v)})\right).$$

Notice that this is the same what we saw in Chapter 4 since $\tanh(x) = 2\sigma(x) - 1$.

Expectations for the data are then obtained using $\hat{v}$, which is obtained from the training data, as well as using $\hat{h}$, which we generate using the outline of setting each $h_a = 1$ with the specified probability.

This is to say that order to determine expectation values, one considers a sample of visible vectors $\{\tilde{v}\}$, and a sample of hidden vectors $\{\tilde{h}\}$ in this iterative sampling process known as the *Gibbs sampling* wherefrom we get that the Ath vectors are

$$\hat{h}_a^{(A)} = \tanh(\sum_i W_{ia} \hat{v}_i^{(A)} + b_a^{(h)}),$$

$$\tilde{v}_i^{(A)} = \tanh(\sum_a W_{ia} \hat{h}_a^{(A)} + b_i^{(v)}),$$

and

$$\tilde{h}_a^{(A)} = \tanh(\sum_i W_{ia} \tilde{v}_i{}^{(A)} + b_a^{(h)})$$

Then the expressions used in the training of the RBM are

$$\langle v_i h_a \rangle_{data} = \frac{1}{N_s} \sum_A \hat{v}_i^{(A)} \hat{h}_a^{(A)}$$

$$\langle v_i h_a \rangle_{model} = \frac{1}{N_s} \sum_A \tilde{v}_i^{(A)} \tilde{h}_a^{(A)}$$

$$\langle v_i \rangle_{data} = \frac{1}{N_s} \sum_A \hat{v}_i^{(A)}$$

$$\langle v_i \rangle_{model} = \frac{1}{N_s} \sum_A \tilde{v}_i^{(A)}$$

$$\langle h_a \rangle_{data} = \frac{1}{N_s} \sum_A \hat{h}_a^{(A)}$$

$$\langle h_a \rangle_{model} = \frac{1}{N_s} \sum_A \tilde{h}_a^{(A)},$$

where $A = 1, \cdots, N_s$.

At this point, we can talk about flows from learnt weights. Here, one uses the weights and biases that are obtained from a trained RBM to construct a continuous flow from an initial state to a final state. To proceed, label the data set by an index A, as $\hat{v}^{(A)}$. For each index, this quantity is a collection of spin values–one for each lattice site. We generate an RBM flow using the sampling strategy mentioned previously, and we denote the data set produced after k steps as $\tilde{v}^{(A,k)}$, where $\tilde{v}^{(A,0)}$ denotes the original training set. We obtain a flow of length n as follows

$$\tilde{v}_i^{(A,1)} = \tanh(\sum_a W_{ia} \hat{h}_a^{(A)} + b_i^{(v)})$$

$$\tilde{v}_i^{(A,2)} = \tanh(\sum_a W_{ia} \tilde{h}_a^{(A,1)} + b_i^{(v)}),$$

immediate steps, and

$$\tilde{v}_i^{(A,n)} = \tanh(\sum_a W_{ia} \tilde{h}_a^{(A,n-1)} + b_i^{(v)}).$$

With the hidden flows,

$$\tilde{h_a}^{(A)} = \tanh(\sum_i W_{ia}\tilde{v}_i^{(A)} + b_a^{(h)})$$

$$\tilde{h}_a^{(A,1)} = \tanh(\sum_i W_{ia}\tilde{v}_i^{(A,1)} + b_a^{(h)}),$$

immediate steps, and

$$\tilde{h}_a^{(A,n)} = \tanh(\sum_i W_{ia}\tilde{v}_i^{(A,n-1)} + b_a^{(h)}).$$

We consider an $|\lambda| = N \equiv L_v \times L_v$ lattice, with sites indexed by a vector $\vec{k}$. In the practical implementation, a vector of the same length is considered, and the rows of the array are simply concatenated to obtain a vector whose components constitute the training data input to the visible units of the RBM. The 2-dimensional Ising model is chosen because this model has fixed point, which is described by a well-known theory. In the context of the chapter on renormalisation, the fixed point is unstable, which means that generic flows move away from it. If one seeks flows that move towards then tuning is necessary. In order to study the behaviour around the fixed point, one may use primary operators whose correlators are power-laws of distance on the lattice. This model exhibits a phase transition at the critical temperature, which we can recall as $T_c = 2J/k\ln(1+\sqrt{2})$.

Examples of the two point and three point correlators are respectively,

$$\langle\phi(\vec{x_1})\phi(\vec{x_2})\rangle = \frac{B_1}{|\vec{x_1} - \vec{x_2}|^{2\Delta}},$$

and

$$\langle\phi(\vec{x_1})\phi(\vec{x_2})\phi(\vec{x_3})\rangle = \frac{B_2}{|\vec{x_1} - \vec{x_2}|^{\Delta}|\vec{x_1} - \vec{x_3}|^{\Delta}|\vec{x_2} - \vec{x_3}|^{\Delta}},$$

where $\Delta = 1/8$ is known as the scaling dimension of the field.

If the RBM reproduces this behaviour then we would have reasonable reasons to believe that the RBM is doing something close to RG.

In what follows we discuss the findings from [25] when comparing the RBM mechanism to RG.

i) In a theoretical sense, there seems to be a difference between RBM and RG flows because while RBM appear to drive the configurations to critical temperature, the

RG flow drives them to higher temperatures since temperature corresponds to a relevant operator.

ii) The number of spins decreases along an RG flow while it is an invariant for the case of RBMs.

iii) Numerical experiments show that RBMs generate configurations of the Ising models which are close to what would be given by RG, and these can be used to determine the scaling dimension of the spin variable from spatial statistics of RBM generated patterns.

iv) Another difference that is observed stems from the fact that with regards to correlation functions, the RBM does not reproduce the correct scaling dimension, which implies that these two alternatives are different although they agree on large scales.

v) The paper explores the possibility that deep-learning is an RG flow with each layer performing an RG step. This discussion is done by studying correlation between the visible and hidden units from RBM patterns, but there is no confidence about emergent agreements between the two instruments on large scales.

vi) Another interesting observation is that if there are 3 stacked RBMs, then temperature flows between the first and second layers but appears to be constant from the second to the third layer.

A Comment on Future Research Plans

Concerning future research work, I would like to investigate if there is an analytical way to recover the behaviour observed in chapter 6, and to see what happens if we look into Hopfield networks instead. I would also like to dive into more machine learning research over the next k months ($k \in \mathbb{N}$). I am exceedingly grateful, once again, to my supervisor for this opportunity and more many of its kind to follow, as well as to all the researchers who contributed to this field and whose insight inspired the writing in this document.

References

[1] Paul, A. and Venkatasubramanian, S. *WHY DOES UNSUPERVISED DEEP LEARNING WORK? - A PERSPECTIVE FROM GROUP THEORY*. arXiv:1412.6621v3 [cs.LG]. 2015.

[2] Lin, H. W., Tegmark, M. and Rolnick, D. *Why does deep and cheap learning work so well?*. arXiv:1608.08225v4 [cond-mat.dis-nn]. 2017.

[3] Zhang, C. Bengio, S., Hardt, M., Recht, B. and Vinyals, O. *UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION*. arXiv:1611.03530v2 [cs.LG]. 2017.

[4] Weinan, E. *A Proposal on Machine Learning via Dynamical Systems*. Commun. Math. Stat. 5:1–11. 2017.

[5] Kallenberg, O. *Foundations of Modern Probability: 2nd Edition*. Springer. New York. 2002.

[6] Rudin, W. *Real and Complex Analysis* McGraw-Hill Book Company. Singapore. 1987.

[7] Tao, T. *An Introduction to Measure Theory* `https://terrytao.files.wordpress.com/2011/01/measure-book1.pdf`. 2010.

[8] Bauerschmidt, R. Brydges, D. C. and Slade, G. *Roland Bauerschmidt, David C. Brydges, and Gordon Slade*. arXiv:1907.05474v2 [math-ph]. 2019.

[9] Khudier, D. and Fawaz, N. *Ising Model Phase Transition Calculation for Ferro-Paramagnetic Lattice*. Article in International Letters of Chemistry Physics and Astronomy 201-212. 2013.

[10] Brydges, D. C. *Lectures on the Renormalisation Group*. `https://pdfs.semanticscholar.org/5f7b/579b39814b912bc64e9eadf7b361fbd26a8c.pdf`.

[11] *Renormalisation Group*. Physics Notes, Cambridge University. `https://www.tcm.phy.cam.ac.uk/ bds10/phase/rg.pdf`

[12] Hollowood, J. T. *6 Lectures on QFT, RG and SUSY*. arXiv:0909.0859v1 [hep-th]. 2009.

[13] Kardar, M. *III.D The Renormalization Group (Conceptual)*. Statistical Mechanics II: Statistical Physics of Fields Notes. 2014.

[14] Liden, P. *Renormalization group approach to statistical systems*. http://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A825498&dswid=-9120

[15] McComb, W. D. *Renormalization Methods: A Guide for Beginners*. Clarendon Press, Oxford. 2004.

[16] Freund, Y., and Haussler. *Unsupervised learning of distributions on binary vectors using two layer networks*. Advances in Neural Information Processing Systems 4. 1991.

[17] Fischer, A. and Igel, C. *An Introduction to Restricted Boltzmann Machines*. Proceedings of the 17th Iberoamerican Congress on Pattern Recognition (CIARP). Volume: LNCS 7441. 2012.

[18] Cueto, M. A., Morton, J., and Sturmfels, B. *Geometry of the Restricted Boltzmann Machine*. arXiv:0908.4425v1 [stat.ML]. 2009.

[19] Le Roux, N., Bengio, Y. *Representational power of restricted Boltzmann machines and deep belief networks*. Neural Computation 20(6), 1631–1649 (2008).

[20] Kadanoff, L. P. *Statistical Physics: Statics, Dynamics and Renormalisation*. World Scientific. Singapore. 2000.

[21] Kadanoff, L. P., Houghton, A., and Yalabi, M. C. *Variational Approximations for Renormalisation Group Transformations*. Journal of Statistical Physics, Vol. 14, No. 2. 1976.

[22] Efrati, E., Whang, Z., Kolan A., and Kadanoff, L. P. *Real-space Renormalisation in Statistical Mechanics.*

[23] Mehta, P. and Schwab, D. J. *An exact mapping between the Variational Renormalization Group and Deep Learning*. arXiv:1410.3831v1 [stat.ML]. 2014.

[24] Craigie, J. *Field Theory and Renormalization Group for the Magnetocaloric Effect*. https://www.imperial.ac.uk/media/imperial-college/research-centres-and-groups/theoretical-physics/msc/dissertations/2012/Jacob-Craigie-Dissertation.pdf

[25] De Mello Koch, E., De Mello Koch, R. and Cheng, L. *Is Deep Learning a Renormalisation Group Flow?*. arXiv:1906.05212v2 [cs.LG]. 2020.

[26] Kristiadi, A. *Maximizing likelihood is equivalent to minimizing KL-Divergence*. https://wiseodd.github.io/techblog/2017/01/26/kl-mle/

[27] ter Hoeve, J. *Renormalisation Group Connected to Neural Networks*. https://dspace.library.uu.nl/bitstream/handle/1874/366784/ThesisFinalterHoeve.pdf?sequence=2&isAllowed=y