Université de Montréal

Renormalization Group Theory, Scaling Laws and Deep Learning

par
Parviz Haggi

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en computer science

August, 2022

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Renormalization Group Theory, Scaling Laws and Deep Learning

présenté par:

Parviz Haggi

a été évalué par un jury composé des personnes suivantes:

PrénomPrésident NomPrésident,	président-rapporteur
Irina Rish,	directeur de recherche

Mémoire accepté le: . . . .

ABSTRACT

The question of the possibility of intelligent machines is fundamentally intertwined with the machines' ability to reason. Or not. The developments of the recent years point in a completely different direction : What we need is simple, generic but scalable algorithms that can keep learning on their own. This thesis is an attempt to find theoretical explanations to the findings of recent years where empirical evidence has been presented in support of phase transitions in neural networks, power law behavior of various entities, and even evidence of algorithmic universality, all of which are beautifully explained in the context of statistical physics, quantum field theory and statistical field theory but not necessarily in the context of deep learning where no complete theoretical framework is available.

Inspired by these developments, and as it turns out, with the overly ambitious goal of providing a solid theoretical explanation of the empirically observed power laws in neural networks, we set out to substantiate the claims that renormalization group theory may be the sought-after theory of deep learning which may explain the above, as well as what we call algorithmic universality.

Keywords : Renormalization, Ising Model, Hopfield Networks, Restricted Boltzmann Machines, Energy Models, Phase Transitions, Power laws, Algorithms, Universality, Quantum Mechanics, Statistical Mechanics, Quantum Field Theory, Statistical Field Theory

ABSTRACT

La question de la possibilité de machines intelligentes est intimement liée à leur capacité à raisonner. Ou pas. Les développements des dernières années pointent dans une direction complètement différente : ce dont nous avons besoin, ce sont des algorithmes simples, génériques mais évolutifs qui peuvent continuer à apprendre par eux-mêmes. Cette thèse tente de trouver des explications théoriques aux constatations des dernières années où des éléments de preuve empiriques ont été présentés pour étayer les transitions de phase dans les réseaux de neurones, le comportement en loi de puissance de diverses entités et même la preuve d'une universalité algorithmique, tout cela étant merveilleusement expliqué dans le contexte de la physique statistique, de la théorie quantique des champs et de la théorie statistique des champs, mais pas nécessairement dans le contexte de l'apprentissage profond où aucun cadre théorique complet n'est disponible.

Inspirés par ces développements et, comme il s'avère, avec l'objectif trop ambitieux de fournir une explication théorique solide des lois de puissance empiriquement observées dans les réseaux de neurones, nous avons entrepris de justifier les affirmations selon lesquelles la théorie du groupe de renormalisation pourrait être la théorie recherchée de l'apprentissage profond qui pourrait expliquer ce qui précède, ainsi que ce que nous appelons l'universalité algorithmique.

Mots-Clés : Renormalisation, Modèle d'Ising, Réseaux de Hopfield, Machines de Boltzmann Restreintes, Modèles Ènergétiques, Transitions de Phase, Lois de Puissance, Algorithmes, Universalité, Mécanique Quantique, Mécanique statistique, Théorie Quantique des Champs, Théorie Statistique des Champs

CONTENTS

LIST OF ABBREVIATIONS

CD	Contrastive Divergence
DL	Deep Learning
DNN	Deep Neural Networks
d.o.f	Degrees of freedom
EBM	Energy-Based Models
GD	Gradient Descent
MC	Monte Carlo
MFT	Mean Field Theory
ML	Maximum Likelihood
NLL	Negative Log-Likelihood
NN	Neural Networks
RG	Renormalization Group
RNN	Recurrent Neural Networks
RSMI	Real Space Mutual Information
SFT	Statistical Field Theory
SGD	Stochastic Gradient Descent
QED	Quantum Electro-Dynamics
QFT	Quantum Field Theory

To Sezen, Urartu and Shusha

PRÉFACE

Plato : GPT-3, what does it mean to be human ?

GPT-3 : Humans are the measure of all things[1], of things that are, that they are, of things that are not, that they are ... nothing but a joke.

[1] Humans are the measure of all things. Of things that are, that they are. Of things that are not, that they are not. Protagoras cited in Plato's Theaetetus. Oxford.

CHAPITRE 1

INTRODUCTION

In his reflections over the experiences of the past research [60], R. Sutton concludes that leveraging on computation has proven more efficient than carefully crafted algorithms that supposedly mimic human thinking. He puts forward the examples of computer chess, computer Go, speech recognition and computer vision. Computers winning over humans through deep search i.e. brute force, was at the time considered unsatisfactory and perhaps just a coincidence by human-thinking-oriented researchers who pointed out that human thinking did not involve brute force methods and that better algorithms i.e. closer to human mind, were yet to be developed and prove their superiority over massive computational systems.

It has, however, turned out that the human-thinking-based approaches have been counterproductive and that systems based on massive computation have shown significant success. A reasonable question to ask is how we understand the workings of the mind as it seems far too complex and beyond reach to us to grasp, let alone replicate. The (bitter) lesson, according to Sutton, is that building algorithms based on our understanding of how human minds function [1] either has not worked or has worked only in short term just to plateau with time and hinder further progress. The bitterness lies in the fact that the opposing, counter-intuitive and non-human-centric approaches are showing more and more empirically successful, scale better and are somehow claimed to be reminiscent of an emergent behavior.

The great power of these general purpose methods that on one hand are more successful and on the other hand scale better with increased computation has been demonstrated most recently with the introduction of GPT-3 [7] and other Foundation Models i.e. large-scale neural networks (NNs) that are pre-trained on large diverse data sets. Another rela-

[1] ...how we think we think...

tively recent development is the empirically demonstrated scaling laws [36]. There it is shown that the performance of language models improves smoothly when certain order parameters (model size, data set size, amount of compute for training) are increased. Furthermore, it is shown empirically that the performance exhibits a power-law relationship with each order parameter when not bottle-necked by the other two order parameters. These findings confirm the bitter lesson stated above : crafting sophisticated algorithms to achieve a certain performance may turn out to be less useful in a certain limit of the order parameters where they perform equally well or are outperformed by simpler or computationally less demanding algorithms.

Empirical evidence aside, the current research lacks a thorough theoretical understanding of not only scaling laws and phase transitions in NNs but also Deep Learning(DL) itself. We can even be bold enough to say that the inner workings of NNs has no theoretical backing. This is indeed the ambitious and quite unrealistic motivation behind this thesis. What are the governing laws of NNs ? How can we understand the empirically proven scaling laws ? How are these laws related to specific algorithms ? How can a simple algorithm be better than a more sophisticated one ? How do we explain these laws based on macroscopic variables/order parameters ?

What are the consequences of these laws in the long run and for the way we think about artificial intelligence ? Are these laws related to optimization and the networks' ability to find the minima in the loss landscape ? Is there a more profound theory that explains the empirical findings of the recent years ? If yes, what are the implications of such a theory ? How does it explain the current discoveries and what can we possibly learn from this theory ?

Prior to appearing in the current context, scaling laws and phase transitions have been studied in the framework of both classical and quantum physics. The commonality between this field and DL is that both attempt to extract relevant information from data. However, while the extraction of macroscopically plausible information from microscopic data is done via so-called coarse-graining in Renormalization Group Theory (RG), the mechanisms behind this extraction in NNs is poorly understood. Some researchers [46, 13] claim that NNs too perform sophisticated coarse-graining procedures. If true, since coarse-graining is a fundamental ingredient in RG, the entire field of RG would also be a natural framework for understanding of DL.

Furthermore, what type of coarse-graining do NNs do ? Is it even relevant to try to understand the specifics of coarse-graining in RG ? Or is it perhaps sufficient to establish RG as the theory of NNs and use it to put algorithms into universality classes and compute their universality constants ? We should note some research [13] claim to have found "evidence of presence of RG-like patterns" in the correlation functions between visible and hidden neurons of RBMs.

Assuming that NNs do perform some kind of coarse-graining, the question is in the context of what theory it should be studied. Speaking of consciousness in their Orch OR paper[24] the authors claim that consciousness is the ability to sustain a superposition of quantum states that collapse because they originate from different space-time origins/manifolds that generate the waves. The collapse occurs naturally and is known as Orchestrated object reduction (Orch OR). Although it is not clear why, consciousness is attributed to quantum gravitational effects.

If Penrose's idea [24] is correct, the least we can do is to try to understand the workings of the brain in terms of Quantum Field Theory (QFT). This means that we can assume that our dynamic variable is a field instead of a path and that the path integral can be an integral over all possible field configurations. This is not far-fetched as it is done also in the transition from quantum mechanics (QM) to QFT.

We explore (and attempt to build upon) these ideas in this work with a particular focus on Hopfield Networks. This is motivated by (**a**) the relative mathematical simplicity of these networks and (**b**) the fact that they have been used as analogs for associative memory models i.e. as simplified brain prototypes. Furthermore, based on the fact that there is 14 orders of magnitude's difference in scale [2] between observable macroscopic brain acti-

[2] See section 2.4 for details

vity and synaptic transmission process [48], it seems that applying RG transformations is very meaningful [3].

But, as we will see later, RG transformations are a methodical way of zooming out quantum fluctuations and so we could, in some sense, say that by considering RG theory, we are already in the quantum realm. However, in order to further consolidate the idea that brain activity involves quantum processes, we note that it is possible to design an experiment involving signals transmitted by neurons analogous to the double slit experiment involving photons. Keeping this in mind, the fact that the observed photon interference patterns in the double slit experiment is the strongest motivating factor behind quantum mechanics provides yet another motivation for neuronal quantum processes.

We should also mention that Feynman's path integral approach in physics is designed to include the uncertainty involved in quantum processes and that a consistent description of a theory in terms of path integrals is equivalent to it being a quantum theory. In the current context, this is done for Hopfield Networks [48] where it is shown that a path integral formulation is justified and that NNs are in fact discrete versions of path integrals. [4]

This thesis is organized as follows : We first review the general state of research on the relationship between DL and physics. We then discuss scaling laws, critical exponents and phase transitions in the context of statistical physics. More often than not, we use the historically important Ising model to describe the evolution of the ideas from statistical physics to mean field theory to RG (The latter is also discussed in the context of QFT). This is done with the hope of better applicability in the absence of a lattice structure. We then give a short review of two energy-based models but for the reasons we explained

[3] As we will see in later chapters, RG is a systematic approach to understanding the manifestations of the microscopic interactions in the macroscopic observables i.e. different scales

[4] Without going into details, we just mention that to mimic the "double slit experiment" they consider three sets of neurons : source neurons (representing emitted photons or electrons), interneurons I_k(corresponding to the slits) and targent neurons T_k (corresponding to the locations of the arrival of the photons). The rest of the formulation is simply due to the fact that neuronal signals are treated as waves and so the outcome is a book example of how interference patterns are generated.

above, focus only on Hopfield Networks. As we will see, RG, either viewed as a subfield of statistical physics or as a meta theory, seems to play a more prominent role in whatever relationship physics has with DL. This thesis also includes a chapter on how RG is used in the context of DL with a focus on Information Theory. The idea was to investigate the possible existence of a relationship between RG, Information Theory and Stochastic Gradient Descent but the matter was not pursued and so this chapter is a stand-alone part of this document that has been moved to the Appendix in its entirety.

This work includes four projects, all exploring the ideas put forward in this introduction. Inspired by the transition from classical variables to QM, as formulated in [48], Project 1 is limited to a classical Mean Field Theoretical treatment of Hopfield Networks. Project 2 is moves from Classical Physics to QFT, where we also derive Feynman-like propagators of Hopfield Networks and run into inexplicable divergencies. Admittedly, the Lagrangian was have used is a construction of [48] and so Project 3 seeks to remedy the issues encountered in Project 2 by using a different Lagrangian. Project 4 is done in a Quantum Field Theoretical setting where we derive Feynman-like rules, provide Feynman-graphs of a few basic interactions and the correlation functions of Hopfield Networks. Finally, this work ends with a U-turn, where we return to Statistical Quantum Field Theory and find the answers to the problems encountered in Projects 1-4. Note also that the details of most of the calculations in the four projects have been moved to the Appendix.

CHAPITRE 2

ON THE RELATIONSHIP BETWEEN DEEP LEARNING AND PHYSICS

2.1 The General State of Research on DL and Physics

The broad range of modeling tools and algorithms encompassed by ML has been applied in a vast number of scientific disciplines. Since ML as a tool aims at discerning/recognizing patterns in data, thereby also predicting the behavior of unseen problems, it makes it ideal for fundamental research also in other fields e.g. physics. The two disciplines also share other fundamental characteristics concerning data collection processes, model building, analysis and predictions. However, if physics seeks to understand the complexities of natural mechanisms through human intuition or some basic principles gained through it, ML, lacking the said intuition or intelligence, seeks to extract the intuition i.e. the fundamental concepts from the data. It goes without saying that although ML has been very successful in certain fields, the scope of its algorithms is often limited and our lack of theoretical understanding of how it succeeds or fails is a problem.

The focus of this thesis is not to give a full account of diverse connections between DL and physics. Therefore, we will not delve into the applications of ML techniques in physical sciences in general. Suffice it to say that ML has been quite successfully applied to **cosmology** [37] [1], **quantum computing** [23] [2], **quantum many body problem** [3], **three-body problem** [6], **particle physics** [10] [4],[25] [5],[51] [6], just to mention a few.

[1] Photometric redshift

[2] Quantum state tomography

[3] with a very diverse set of references that we skip

[4] To characterize the landscape of string theories

[5] To understand the AdS/CFT correspondence (the subject of my PhD thesis)

[6] To classify QCD phase transitions

2.2 DL and Statistical Physics

Based on the above, it may be premature to assume that ML can be more than a tool for some sophisticated data analysis. However, the cross-fertilization between ML and physics goes beyond this observation since the conceptual developments in ML are in many ways also rooted in insights gained from the relatively old field theoretical physics. In fact, the large body of evidence of tantalizing parallels between physics and ML [18, 49, 3, 47, 11, 33, 9, 46, 56, 59, 45, 44, 4] seem to indicate a complexly intertwined relationships. While some research merely draw parallels between the observations in the two domains, others go further and refer to the frequently encountered physical, statistical or thermodynamic concepts e.g. **symmetry, scale, locality, compositionality, hierarchy** etc as well as **phase transitions and power laws** in DL, as indications of the existence of a deeper connection, and perhaps also with the allusions to physics (statistical) as the missing theoretical framework for understanding DL.

In what follows, we will describe the part of the research that has laid the foundation for what may very well turn out to be the (future) theory of DL. The research is still in its infancy but there are too many connections/similarities to ignore the possibility of finding a theoretical explanation for DL within physics. [4] predict that these connections will only deepen and that the conceptual insights from many parts of physics and mathematics [7] will contribute towards finding a mature theory of DL. As we will gradually see, this does not seem to be limited to just classical statistical physics. Quantum physics and later on, if we are aiming at creating sentient machines, also such alien theories as supergravity seem to be at play.

We know that the existing learning theory in ML is not able to explain the success of DL. It cannot explain why deep networks have good generalization properties the number of adjustable parameters or dimensions or weights greatly exceeds the number of training examples. Similarly, we are neither able to understand what learning problems are com-

[7] spin glasses, random landscapes, phase transitions, chaos, Riemannian geometry, random matrix theory, free probability, and non-equilibrium statistical mechanics

putationally tractable[11] nor what architectures or hyperparameters are best suited in different situations.

It turns out that within the general field of theoretical physics, it is statistical physics and within its modern versions, **RG** may be the sought-after framework that is closest to the theory of ML/DL. Assuming the reader is familiar with **supervised** and **unsupervised** learning, let's just briefly review some of the contributions of statistical physics to understanding some basic problems in ML.

Supervised Learning : Simply put, linear regression is the most common learning method of supervised learning and uses the **least squares method** to find W, a vector of coefficient whose scalar product with the data points X gives the corresponding label y. The Bayesian version of this approach seeks to establish the relationship $y_i = X_i W + \xi_i$ by assuming Gaussian prior on the weights and Gaussian noise. And the generalized version of this approach assumes a generic priors $p_W(.)$ and generic conditional $p_{out}(y_i|X_i W)$ the weights. When the amount of data is limited [11] state that statistics is not applicable and propose statistical physics (the replica method)[8] as a possible path forward. It is also suggested that the replica method can be used to compute the **mutual information** between X and y and that it is related to the **free energy** in physics[9].

Unsupervised Learning : is a key branch of ML that models the structure of complex data x e.g. the structure of language, images etc. Since we do not know how to model the distribution of any of the above, unsupervised learning aims at adjusting the parameters w of a family of distributions $p(x,w)$ to find the most similar to the empirical data distribution of some samples. This is often done by minimizing the log likelihood of the data with respect to the parameters.[4] point out that this can be thought of as an inverse statistical mechanics problem. While the goal of equilibrium statistical mechanics is to compute bulk statistics of the microstates x from a Boltzmann distribution $p(x,w)$ where w are the couplings, unsupervised learning aims at constructing the appropriate distribution $p(x,w)$ from the microstates samples x.

[8] For the analysis of optimization methods in statistical physics

[9] See [58, 17] for related references

Apart from this reversed perspective and the insights that follow, unsupervised learning relies on a few basic mathematical tools based on **low rank matrix decomposition** [10]. Low rank decomposition is used to uncover a structure in the data by finding a matrix of much lower rank in terms of dimensionality and the number of samples. This can be challenging and is so in the regime of high-dimensional noisy data regime. The relevance of statistical physics becomes apparent if the low rank matrix estimation in this regime is treated as a spin glass with lower dimensional vector variables and a particular planted configuration to be found [11].

Regarding low-rank matrix decomposition, it turns out that cluster detection in **stochastic block model** which was studied in statistical physics [19] was a contributing factor in finding the exact solution and an understanding of the algorithmic limitations of stochastic block model [14]. There are numerous other similar areas where statistical physics has inspired/contributed to machine learning but we leave this subject for now as it is not the objective of this work [11].

Autoencoders and **Variational autoencoders** (VAE)[38] are other types of unsupervised learning with close links to statistical physics. Autoencoders with linear activation functions are related to principle component analysis. VAE use variational inference and are trained using a prior on latent variables. Some VAEs are closely related to dictonary learning which in turn has been studied using the techniques from statistical physics[34].

Boltzmann Machines or **Restricted Boltzmann Machines** which we have used in this work, belong to the **unsupervised learning** category. BMs are sometimes referred to as an **Inverse Ising Model**. Both have strong relationship with statistical physics. In order to remedy the difficulties of an analytic study of these models' learning process through **contrastive divergence algorithm** [26], a statistical physics approach is to replace the **Gibbs sampling** in contrastive divergence with Thouless-Anderson-Palmer equations

[10] This includes matrix completion, independent component analysis (ICA), principle component analysis (PCA), data clustering etc

[11] These include, for instance, understanding approximate message passing algorithms (AMP) [2, 5] for low rank matrix estimation i.e. generalizations of Thouless-Anderson-Palmer equations [61] in spin glasses etc

which in turn stem from statistical physics [20].

2.3 DL and RG

After this short expose on the insights into ML gained through general statistical physics methods/techniques, we now turn our attention towards **RG** i.e. the part of statistical physics that is the focus of this work and that has seems to be a potential candidate for **a future theory of DL**. There are multiple accounts of the appearance of RG or RG-like behavior in ML and consequently questions about the possibly (at least) of it being the correct framework for understanding many long-standing issues in ML/DL.

We should note that this line of research is not necessarily based on utilizing some technical solution to a problem achieved by e.g. replacing a distribution in an ML algorithm by equations from statistical physics, as we described above. In fact, considering some simple but fundamental notions such as scale and locality, [9] compare the relevance of scale in RG with scale and depth in deep neural networks. In their work, the authors conclude that at least some ML algorithms perform some sort of sophisticated RG. This is clearly a bold statement but in what follows we will review the role of RG transformations and what it entails.

[44] examine the question of criticality in language models, and show that the decay of mutual information between two symbols can follow a power law for context-free grammar and that their findings is closely related to the absence of phase transitions in fewer than two dimensions in classical statistical mechanics, a result that can have potential applications on training RNNs. [45] use physics to answer questions about what neural networks can compute, what they compute, why and how they generalize, how they can be taught to come alive i.e. imagine[21], and finally, why DL is so successful if shallow neural networks can approximate any function.

It is known that the classical mathematical guarantees on the ability of neural networks to approximate arbitrary functions do not set a limit for the width of shallow networks and do not explain why and how deep networks work so well. [45] argue that the success of neural networks is at least in part due to the fact that the functions that are frequently approximated have exponentially fewer parameters which in turn is related their physical properties. This, they argue, is not coincidental : real world data and functions satisfy such conditions as locality, symmetry and compositionality, characteristics that reduce the number of parameters exponentially, giving rise to significantly simpler neural networks. Furthermore, it is argued that when the statistical process of information generation is hierarchical, deep neural networks can be made more efficient than shallow networks.

[29, 56] have investigated the concept of **scale invariance** [12] and **scale invariant feature extraction** in deep neural networks. For instance, [29] train an energy model (RBM) on **Monte-Carlo** simulated **Ising Hamiltonian** data samples for different values of temperature and external magnetic field [13], concluding that the features are extracted hierarchically through **coarse-graining**, which in turn is a fundamental ingredient in **RG in statistical physics**.

[46] construct **an exact mapping between RG flows and DL** [14] suggesting that DL may indeed be employing an **RG-like** scheme to learn relevant features from data. Note that this is done in the context of Information Theory. We refer the reader to Appendix V where we have done a large part of the calculations in [46]. At the same time, while investigating this relationship, [13] use the prototypical statistical model of magnets to, again, train an RBM after which the configurations created by the physical model, and those generated by RBM are compared. The focus of their work is a comparison between a single layer of a deep neural network and one step in the RG flow. And so they conclude that at least in this experimental setting, it is possible to confirm the presence of "RG-like" patterns in the correlators computed via RBM.

[12] See also the last part of this work and comments on Conformal Field Theory

[13] Note that this is discussed in later parts of this work as well, both in terms of **Mean Field Theory** and **RG**.

[14] They use what is known as **Variational RG**

Empirical findings of recent years strengthen the links between RG and ML. In their study of scaling laws for language models [36] show that the cross-entropy loss scales as a power-law with model size, dataset size, and the amount of training compute. Interestingly enough, the network width or depth have minimal effects within a wide range and so the findings are architecture agnostic. This is yet another addition to the idea of the possibility of DL performing some sort of RG transformation simply because it was only in the RG framework that many of the observed power laws in nature were explained [15]

A consequence of these findings is that, for instance, in the limit of infinite data, neither architecture nor algorithms seems to matter. This is very similar to the idea of universality in physics and perfectly described in RG. For example, in the same way as seemingly very different types of matter (glass and water) turn out to be in the same universality class, we might be able to talk about **algorithmic universality classes**. Note also that this has some practical implications : Instead of blindly training all kinds of models until convergence, optimally compute-efficient training involves training large models on a modest amount of data and stopping significantly before convergence as other smaller models trained on larger amount of data or with a much bigger compute budget will simply not perform better.

With these expressions of RG in DL, is it then premature to say that the **Theory of DL** is hidden within the **Theory of Statistical Physics** ? To put it differently, if neural networks are capable of identifying different phases of matter [39, 9], proper coarse-graining schemes when the knowledge of microscopic details is insufficient [43] or other physical concepts [30], do they also somehow apply RG transformations to approximate functions i.e. solve particular ML tasks ? In other words, **does RG also provide a theoretical understanding of neural networks ?** Or perhaps, **is DL a renormalization group flow ?**[13]

We cannot possibly ignore the observed scaling laws and phase transitions as they seem to demonstrate something more than just coincidental parallels. But to conclude that

[15] We will discuss this extensively later in this work.

RG is THE THEORY of DL seems premature. The fact is that research has uncovered differences at every stage of these discoveries, which may or may not be irreconcilable with this proposed theoretical framework. For instance, [9] observe that while the trained machine obtains the the flow of spin state configurations and reproduces the obserevables of the physical system e.g. physical phase transitions, this cannot always be put in a one to one relationship with the fixed points of the renormalization group flow.

[45] point out that **Effective Field Theory** and RG, both revolving around the idea of distilling out desired information from undesired noise, have little to do with the idea of **unsupervised learning** or pattern recognition. On the contrary, since RG and Effective Field Theory distill the long wavelength i.e. macroscopic degrees of freedom, they make sense if the features of interest are specified, and so it is only meaningful to discuss RG in the context of **supervised learning**.

So perhaps a direct adaptation of RG to ML or the expectation of its full presence in ML may be too simplistic. The starting point of RG in physics is to reveal the way in which microscopic effects are crystallized on the macroscopic scale of interest, which may coincide with the goals of some supervised ML algorithms. But since all ML algorithms do not aim at distilling long-range properties of statistical distributions, the exact nature of the involvement of RG in ML may be alien at this stage of research.

As we will see later, coarse-graining which is at the heart of RG is related to scale transformation and so perhaps in the context of ML, we may have to replace scale transformation with other transformations before we have an RG-like theory of DL. The downside of this approach is that while we have an understanding of scale, we may lack the right intuition to deal with an abstract form of transformation. Once again, we may have to rethink our approach and to study the symmetries of the problems at hand, as well as understanding what type of manifolds are best suited to describe natural languages or images. There is a long way to go.

2.4 Statistical Mechanics or Statistical Field Theory ?

Most of the research about the relationship between ML and statistical physics is focused on classical arguments describing the links between the two. This relationship is most easily investigated in terms of energy based models in ML [42] with little to no emphasis on the relevance of quantum fluctuations. But it is nearly impossible to discuss RG without considering quantum effects in physics. And so any meaningful relationship between DL and RG, if it truly exists, will have to involve quantum physics. The procedure of coarse-graining in RG is essentially a zooming out of unobserved quantum fluctuations that, depending on the system under investigation, may influence the macroscopic behavior of the system in particular ways [22]. We will see later that the fundamental meaning of this is in terms of operators in the Hamiltonian that may or may not survive RG-transformations. In fact, as we know from the double-slit experiment, quantum mechanics originated in the inability of classical physics to explain natural phenomena solely based on the **determinism** [64].

We could (and should) ask what we mean by **quantum effects** and why we should consider them in e.g. image recognition. Although it may be incomplete, one answer is that since we have observed RG or RG-like behavior in ML, we have already confirmed the existence of quantum effects. Another, perhaps more reasonable or equally incomplete, answer is that whatever RG-like behavior we see in ML, could be the the effect of the unseen and unobserved microscopic variables in the same way as we experience magnetism without observing electrons or their spin.

As usual, there is more to this than the eye can see. It turns out that we can approach this problem from a completely different angle. [48] investigate **Hopfield Networks** as a model of the brain and its activity by proposing a double-slit experiment and the inherent quantum nature of neuronal signals being transmitted in the brain. This is motivated by the fact that there are "at least 14 orders of magnitude difference in scale between brain observable macroscopic brain activity and synaptic transmission process"[48]. The result of their work is in fact a **Schrödinger**-like equation deduced from Hopfield networks.

Note the utmost important word **scale** and its two-fold implications in the current context. First of all, we should mention that quantum effects are present at all scales, but are negligible at macroscopic scales. At the same time, RG theory shows that processes at scales that are much smaller in comparison to the scale at which their effect is measured can influence the latter. Secondly, since the human brain contains roughly 10^{11} neurons and 10^{14} synapses, observations at macroscopic scale of what happens at neuronal or sub-neuronal scale motivates the study of brain activity from the perspective of quantum mechanics.

So assuming that the recent observations in ML point in the direction of RG and reminding ourselves that RG is fundamentally quantum mechanical, we might stop there and somehow try to formulate a theory of ML based on the unknown and unobserved microscopic effects and their manifestations at macroscopic scales (or something equivalent to scale) [16].

But there are recent claims that not only Quantum Mechanics or Statistical Field Theory but also such (at least in the present state of research) distant theories as Supergravity must be taken into account if we want to have a complete understanding of brain activity. At best, this means that our current efforts to do ML in a classical setting is beyond primitive and that ML will not be able to **understand why, how and what** it does as long as the right framework is not used.

This idea is put forward by Penrose et al in what is called **Orch OR Theory** i.e. **Orchestrated Objective Reduction** [1, 24]. This theory claims that quantum effects occur in micro-tubules, the walls of which are composed of "tubulin dimers", which are supposedly the units that encode the quantum effects. It is noteworthy that according to this theory, consciousness, the holy grail of artificial intelligence, consists of the ability of these tubulin dimers to hold superpositions of quantum gravity waves that collapse

[16] Scale refers to symmetries in the system which can also be studied in an abstract form i.e. group theoretically. Note that we previously referred to the question of what kind of a sophisticated RG-like behavior neural networks perform and whether it is based on something other than coarse-graining as it is done in statistical physics.

(reduce as in Reduction in OR) due to an objective factor ascribed to a fundamental characteristic of space time. Here, the reduction produces the classical output of a (quantum)computer.

The Orch Or theory is mentioned here solely as another motivating factor behind the possible relevance of some kind of quantum theory, without which RG theory as we know it, is not meaningful. If RG theory is to be taken seriously as a possible candidate for the theory of DL, it will then be necessary to leave the classical deterministic input-output framework of neural networks as models of brain activity. Indeed, if the human mind is viewed as a metaphor for what we know as information technology, or prior to that, a telegraph switching circuit, the modern metaphor for brain activity is dominated by classical computers which at least in their current form cannot include such fundamental but elusive concepts as consciousness. Regarding the latter, the Orch OR theory refers to it as the collapse of quantum waves in micro-tubules due to an intrinsic property of space-space time geometry [52, 53, 54], which in turn is understood within the framework of **Quantum Gravity**.

Otherwise, since it is not the objective of this thesis, we will not delve into the details of Orch OR theory's claims, particularly due to the fact that they are neither mathematically proven nor substantiated in any other way, albeit put forward by one of the greatest minds of our time.

CHAPITRE 3

STATISTICAL PHYSICS - A CLASSICAL APPROACH

The most beneficial aspect **statistical physics** is that it, by construction, incorporates ensembles of systems. An ensemble of systems is composed of many systems that are constructed as replica of the original system, such that each system in the ensemble represents one of the quantum/**microscopic** states accessible to the system. Since the **microscopic** conditions are immeasurable/inaccessible, this is equivalent to repeating an experiment multiple times under the same **macroscopic** conditions. The idea of ensembles allows for the introduction of ensemble average, i.e. the overall average of a quantity over the ensembles. This is equivalent to averaging over all possible **microscopic** states viewed from a **macroscopic** perspective [1].

Falling short of understanding the inner workings of thermodynamic systems in **classical physics**, **statistical physics** was developed to bridge the gap between **microscopic** interactions and their **macroscopic** manifestations. Without delving into the details of constructing such a theory, we note that the starting point of **statistical physics**(and many other theories) is to write down a generic **Hamiltonian** for a **microscopic** system

$$-\beta H = \sum_n K_n \Theta_n \tag{3.1}$$

Here, $\beta = 1/k_B T$, K_n are the coupling constants (external to the system), Θ_n are so-called local operators (internal) and the sum is over all possible internal operators. Using the Ising model [65] as a simple but descriptive representation of a statistical systems, we note that Θ_n are different spin configurations S_i, $S_i S_j$ etc [2]. In this approach, the

[1] Let's note that already at this point we can discern parallels between this idea and its suitability when considering NNs : If the objective is to determine average performance of a certain type of NN on many data sets, ensemble methods circumvents the tedious and elaborate process of testing the performance of an algorithm on the data sets separately.

[2] See section 3.3 for a treatment of this approach that models ferromagnets as a collection of micro-

probability of a certain spin configuration in the Ising model is given by

$$P(\{S_i\}) = \frac{e^{-\beta H(\{S_i\})}}{Z} \tag{3.2}$$

It turns out that the **partition function** Z, entering here as a normalization constant, is what we will encounter the most and includes all we need to know about the system i.e. its microscopic interactions. It is defined as

$$Z[K] = Tr(e^{-\beta H}) \tag{3.3}$$

where $[K] = [\{K_n\}]$ and Tr (or Tr()) denote trace and refer to the sum over all possible degrees of freedom. In the case of the Ising model with N spins, this means

$$Tr = \sum_{S_1}\sum_{S_2}...\sum_{S_N} \tag{3.4}$$

Crucially, this results in the **partition function** being a function of the coupling constants only. Another important quantity/notion from which all the **macroscopic** observables can be deduced is given by the **free energy**

$$F[K] = -\frac{1}{\beta} log Z[K] \tag{3.5}$$

and its derivatives with respect to the coupling constants.

In the thermodynamic limit (typically when system size or number of particles approach infinity), if the **free energy** exists, we can study the D-dimensional phase space of D coupling constants, where the **free energy** will be analytical except in singular loci that can be points, lines, planes etc. Here, we also encounter the somewhat ambiguous notion of phases and phase boundary. Phases can generally be defined as regions of analyticity of the **free energy**. A **phase transition** is said to occur in two ways, either if the first

scopic particles on a grid that in simplest cases can take on two values up/down or $+1/-1$. Here, spin configurations refer to how the microscopic entities interact with an external field $\sum KS_i$, pairwise $\sum KS_iS_j$ etc derivative of the **free energy** is discontinuous in some direction (**first order phase transition**) or if the derivative is continuous across the phase boundary which is referred to as **second order** or **continuous phase transition**.

In the context of **statistical physics**, a crude argument for the existence of phase transitions is the so-called energy-entropy argument. Entropy, as a measure of disorder, is a link between **microscopic** and **macroscopic** worlds in the sense that it quantifies the probabilities that quantum states of a system can possibly acquire. Given this probability, also known as the Boltzmann distribution (3.2), entropy is related to the multiplicity of the energy states of a system. A system in contact with a reservoir will always show a net increase of entropy hence, the notion being indirectly linked to the forward direction of time. The competition between lowering the internal energy E and increasing entropy S of a system at a temperature T is also encapsulated in the classical definition of the **free energy**

$$F = E - TS \tag{3.6}$$

Here, temperature controls the the competition between energy and the number of available states. At high temperatures, the **free energy** will be lowered by maximizing the entropy. On the other hand, at low temperatures, the first term might dominate the entropy term and so the **free energy** can be lowered by lowering the internal energy. We say that a phase transition must have occurred if these two procedures lead to two different **macroscopic** states.

As a final note, we should briefly mention that the dynamics of a statistical system is reflected in the identification of time averages of any quantity with its ensemble average. **Ergodic Hypothesis** identifies the time average of an observable quantity $A(\eta_i)$, where $\eta_i(t)$ are the dynamical degrees of freedom

$$\langle A \rangle = \lim_{t\to\infty} \frac{1}{t} \int_0^t A(\eta_i(t'))dt' \tag{3.7}$$

with its statistical average

$$\langle A \rangle = \int \prod_i P(\eta_i) A(\eta_i) d\eta_i \tag{3.8}$$

based on the hypothesis that the dynamical degrees of freedom $\eta_i(t)$ come arbitrarily close to all possible configurations of $A(\eta_i)$ (within the limits of conservation **laws**), as time goes to infinity.

We end this short exposé with a comment that relates the **statistical physics** approach to ML : Training a NN can be viewed as a stochastic optimization process. As such, the ensemble approach would allows for the calculation of the typical learning behavior with the weights of the network as its outcome. Similar to above definitions, the probability of observing a certain weight configuration is then given by the Boltzmann density where weights $\mathbf{w}$ replace the spins S_i and $H(\mathbf{w})$ is a cost/energy/**Lyapunov function** to be minimized. We will draw many more parallels between ML and other theories in this thesis. Indeed, the overly ambitious goal of this thesis is(was) to provide a motivation for why and how the black box of NNs should be opened.

3.1 Scaling Laws

In its simplest form **scaling** refers to the dependence between quantities in a power law fashion. **Power laws** are by no means specific to the study of **microscopic** systems. For instance, Kepler's third law (clearly **macroscopic**) states that the ratio of the square of an planet's orbital period T with the cube of its circular orbit R is constant [63]

$$T \propto R^{3/2} \tag{3.9}$$

scaling laws can often be derived from simple dimensional analysis where the power is a **rational number**. There is, however, also a much broader class of phenomena exhibiting

scaling laws[3] where the **exponent** is an **irrational number** hence cannot be derived through dimensional analysis. For example, the relationship between magnetization and temperature in the vicinity of the **critical exponent** T_c in ferromagnets is given by

$$M \propto |T_c - T|^{.311 \pm .005} \tag{3.10}$$

And relationship between the density and temperature for different phases of a particular fluid follows a different **scaling** law[4]

$$|\rho_+ - \rho_-| \propto |T - T_c|^{.327 \pm .006} \tag{3.11}$$

We should mention that even other fluids, despite behaving differently in many other aspects and having, say, a different coexistence curve, demonstrate nearly exactly the same **critical exponent**. Note also that the **critical exponents** are valid only near the **critical points** as the system deviates from this behavior when we move further from the **critical exponents**.

It should be added that the classical study of power **laws** produces incorrect (rational-valued) **exponents**. **Landau's theory of phase transitions** is based on **Mean Field Theory** (MFT) where a physical variable is replaced by its average value. An implication of MFT is to ignore (important, as it turns out) fluctuations near a **critical** point. As a result, even if MFT offers a qualitative understanding of the observed phenomena, it falls short of adequately addressing the divergencies when a variable such as the temperature approaches its **critical** value. This is believed to explain why MFTs do not explain the numerical discrepancies between the theoretically deduced rational **exponents** and the measured irrational **exponents**.

[3] In general, when a quantity approaches its **critical** value, the meaning of $f(t) \sim t^{\lambda}$ is

$$\lambda = \lim_{t \to 0} \frac{log f(t)}{log t}$$

[4] The fact that the two powers are close to each other is a much broader subject related to **universality** which will be discussed in different context in this thesis

3.2 Characterization of Phases

Landau's MFT approach defines phases based on a **symmetry principle**. It states that different phases of matter are distinguished according to the symmetries that are present in them i.e. in the **Hamiltonian** of the system [5]. These symmetries reflect the possibility of transitioning between phases, abruptly when the symmetries change/break or continuously when the symmetries remain unchanged. Again, since the phases of matter are defined based on the symmetries of the Hamiltonian, if two seemingly different phases (e.g. glass and water) are described by the same Hamiltonian the inevitable conclusion is that these "different" phases are essentially the same [6].

Traditionally, but incorrectly, different phases were considered, and ordered, as discontinuities of the derivatives of the **free energy**[31](thereof the expression second order phase transitions, which is nowadays replaced by continuous phase transitions). We will return to this subject in the final chapter of this work on **Statistical Field Theory**.

3.3 Ising model : A Classical Treatment

Studying the Ising model, a prototype in many fields of physics as well as in NNs, information theory etc, sheds light on what is possible/impossible to explain/understand in the MFT approach (classical) versus modern approaches such as **RG Theory**. Here we will review how phases, phase transitions, **correlation functions** and **critical exponents** of the Ising model can be computed and what lessons can be learned from it [7]. Of particular interest is keeping in mind the differences between the predictions made in the two approaches in terms of **scaling laws** and **universality** classes.

As mentioned above, the objective of **statistical physics** is to compute the **partition**

[5] We will return to this in the final chapter of this thesis. In particular, we will explain what is meant by the **Hamiltonian** and what symmetries are used to characterize different or same phases

[6] Glass is in fact considered a very slow moving liquid

[7] Later, for comparison and in the interest of moving beyond the Ising model, we will do explicit calculations of the much more complex model of Hopfield Networks

function for the Ising model for a system with N spins (For simplicity, in 1 dimension) and nearest neighbor interaction

$$H = -K\sum_i S_i S_{i+1} - h\sum_i S_i \tag{3.12}$$

In the absence of an external field ($h = 0$), this is given by

$$Z(N) = 2(2\cos hK)^{N-1} \tag{3.13}$$

The **free energy** for this system in the limit $N \to \infty$ is given by

$$F = -\frac{N}{\beta}[log(2\cos hK) + O(1/N) \tag{3.14}$$

This result does not seem very interesting. The behavior of this system can is better illustrated if we compute the **free energy** through the **transfer matrix method**, which is also applicable in the presence of an external field, and includes elements that are reminiscent of the the study of phase transitions in RG. Using

$$\mathbf{T} = \begin{pmatrix} T_{11} & T_{1-1} \\ T_{-11} & T_{-1-1} \end{pmatrix} = \begin{pmatrix} e^{h+K} & e^{-K} \\ e^{-K} & e^{-h+K} \end{pmatrix} \tag{3.15}$$

the **partition function** (3.3) can be written as

$$Z = \sum_{S_1}\sum_{S_2}\dots\sum_{S_N} T_{S_1S_2}T_{S_2S_3}\dots T_{S_NS_1} = Tr\mathbf{T}^N = \lambda_1^N + \lambda_2^N \tag{3.16}$$

which we arrive at after diagonalizing **T**. Calculating the eigenvalues

$$\lambda_{1,2} = e^K[\cosh h \pm \sqrt{\sinh^2 h + e^{-4K}}] \tag{3.17}$$

it is easily verified that in the thermodynamic limit the largest eigenvalue dominates the **partition function**

$$Z = \lambda_1^N(1 + O(e^{-\alpha N})) \tag{3.18}$$

and that the **free energy** is given by

$$\frac{F}{N} = \frac{1}{\beta} log\{e^{K}[\cosh h \pm \sqrt{\sinh^2 h + e^{-4K}}]\} \tag{3.19}$$

In 1 dimension, in the limit of $T = 0$ or equivalently $K \to \infty$ the largest eigenvalue and the **free energy** can be simplified to

$$\lambda_1 = e^{K+|h|} \tag{3.20}$$

and

$$F = -N(J + |H|) \tag{3.21}$$

In this case, depending on the sign of the external field, the magnetization (defined as the derivative of the **free energy**) will be

$$M = -\frac{1}{N}\frac{\partial F}{\partial H} = \pm 1 \tag{3.22}$$

We conclude that a phase transition occurs at $T = 0$. However, a phase transition is not possible at $T > 0$ in the one dimensional Ising model. Technically, this can be proven by **Perron's theorem** which states that the largest eigenvalue is real, positive, non-degenerate and analytic. Therefore the **free energy** above will be manifestly analytic in this regime and so phase transitions will not occur.

Apart from the above results, this approach is inadequate. In order to venture into the field of phase transitions, we take a look at the **correlation functions** between spins. This helps us calculate an expression for the **correlation length**. Knowledge of the behavior of the **correlation length** i.e. knowing whether the spins are ordered at long or short range, is intimately related to the phases of the matter. The **correlation function** between the spins at different sites is

$$G(i, i + j) = \langle S_i S_{i+j} \rangle = (\tanh K)^j = e^{-j\log(\coth K)} = e^{-j/\xi} \tag{3.23}$$

where ξ is defined as the **correlation length**. Note that at zero temperature i.e. $K \to \infty$, $G(i, i+j) = 1$, which is equivalent to **diverging correlation length** ξ i.e. spins being correlated over long ranges. We can understand the specific **power law** in this case by studying how this divergence happens when the temperature approaches zero. For $T > 0$ i.e. finite but large $K >> 1$, the **correlation length**

$$\xi = e^{J/k_B T} \tag{3.24}$$

As $T \to 0$ the divergence is obviously exponential in the 1-dimensional case. However, this is due to the particularity of the 1-dimensional Ising model. In its most general form the dependence on temperature in the vicinity of the **critical exponents** T_c is more like

$$\xi = (T - T_c)^{\nu} \tag{3.25}$$

3.4 Ising Model : Mean Field Theory

The experimentally motivated **scaling** behavior, at least quantitatively, can be achieved through MFT as well. Here the basic idea is to replace an interacting field[8] by its average value and to ignore fluctuations all together. Although the **critical exponents** will not be correct, MFT gives a better picture of the expected and experimentally proven **scaling** behavior of the Ising model. For example, we can calculate the magnetisation, also known as the **order parameter** by making the plausible assumption that each spin experiences a combination of the external field and an **effective** field originating from all the other spins[9].

Using the new modified Ising **Hamiltonian**, and dropping the last fluctuation term, we

[8] It is too early to use the word field here. We will be able to provide proper motivation for the use of this word in the last chapter of this work. To avoid confusion, we can also use the word quantity instead of field.

[9] Later in this work, we will do this calculation for the Hopfield Network

arrive at the magnetization

$$M = -\frac{1}{N}\frac{\partial F}{\partial H} = ... = \tanh(H/k_BT + 2dJM/K_BT) = \tanh(H/k_BT + M\tau) \tag{3.26}$$

where $\tau = T_c/T$. This expression can be expanded for small values of H and M. The **critical exponents** for this as well as other physical quantities can then be extracted. For instance, we get

$$M^2 \approx 3\frac{T - T_c}{T_c} + ... \tag{3.27}$$

The **critical exponent** $1/2$ above is good but not exact. Nor is it in accordance with the experimental results. What went wrong ? In the above calculations we made two assumptions, the first being that the fluctuations near T_c can be neglected all together (we replaced the "field" by its average). The second assumption is that the magnetization is small near the **critical exponents**.

As it turns out, none of these assumptions are entirely correct as fluctuations are of utmost importance in the vicinity of **critical points** and the order parameter is not necessarily small. Despite these shortcomings, MFT gives a good qualitative picture of **scaling laws** and **universality** classes for different systems. Modifications of MFT, such as inclusion of fluctuations, do not seem to resolve the issues. Hence the next approach i.e. the RG theory.

CHAPITRE 4

RENORMALIZATION GROUP THEORY ON A LATTICE

So what went wrong and why is classical physics not able to calculate the observed **power law exponents** correctly ? Or perhaps the real question is : What are the origins of **scaling laws** ? RG springs from, among other things, Kadanoff's intuitive explanation that near a **criticality**, the system looks the same at all length scales. Put differently, and more precisely, divergencies in **correlation length**, mean that there is a relationship between **coupling constants** and the (**effective**) **Hamiltonian**.

RG is a **meta theory** i.e. a **theory of theories** which has been applied in diverse fields, from forest fires to disease control to spin/lattic models and quantum field theory, just to mention a few. The common thread in all its applications is that it is an attempt to distill large scale structure behavior from complex **microscopic** interactions. As such, it seems absolutely ideal for the purpose of this thesis as we too try to understand observed **scalinglaws** in neural networks and the logical consequences that follow. In what follows, we will describe RG, as applied in statistical physics and quantum field theory. By the time we arrive at the chapter on Statistical field theory, the validity of RG will be taken for granted. As a meta theory, it is simply the language we use to understand what is and is not possible to observe at a given scale and how the **microscopic** world is or is not manifested in the macroscopic world.

4.1 Renormalization Group Theory in Statistical Physics

As we will see, RG offers a consistent framework for modeling **phase transitions, critical phenomena, scaling laws** and **universal constants**, of which the latter describes common behavior in wildly different types of phenomena as they are perceived by us in a low energy regime. It is not too far-fetched to say that RG explains the "nature" of observed phenomena. For instance, it turns out that a **self-avoiding random walk** [57] and an **isolated polymer in a solution** [16] demonstrate very close critical exponents (around $.5880 \pm .0015$). Thanks to RG, the inevitable conclusion is that it is not the specific chemistry but the "nature" of the matter in question that is responsible for this behavior.

The central theme of the theory consists of a so-called **coarse-graining**. This is a process under which degrees of freedom of a given system are grouped together resulting in a **Hamiltonian** with fewer degrees of freedom. For instance, in the case of Ising model, if the distance between spins is a, blocks of spin of linear dimension la where $l > 1$ are grouped together. We are going to see that that this process allows for non-integer values of l, although integer values are helpful in creating the intuitive picture in lattice models.

Coarse-graining results in a new system. If the distances between the blocks of spin are re-scaled in terms of la, the **correlation length** will also be re-scaled, leading to a new system that for all purposes is identical to the original system albeit with a different **Hamiltonian**. If the new system looks like the old system, there is no reason to assume the physics will be different. Therefore, a crucial point here is the reasonable imposition of the new condition that the **new Hamiltonian** should have more or less the **same functional form** after a block spin transformation.

In its most general form, this type of **re-scaling** or RG transformation consists of a re-definition of the **coupling constants** of the system. For instance, starting with the **Hamiltonian** for a d-dimensional spin system (a hypercube with lattice spacing a) with nearest neighbor interaction and an external field h

$$\beta H = ... = -K \sum_{<ij>} S_i S_j - h \sum_i S_i \tag{4.1}$$

If the spins are correlated on lengths of order ξ, then the spins on length scale la can be considered acting as a single unit as long as

$$a \ll la \ll \xi(T) \tag{4.2}$$

While there is no reason to assume that the block spins interact differently the consequence of **coarse-graining** is that the block spins interact with a different **effective** external field at the given scale

$$\beta H_l = ... = -K_l \sum_{<IJ>} S_I S_J - h_l \sum_I S_I$$

where the **effective coupling constants** under successive RG transformation are defined such that $K_1 = K,\ h_1 = h$.

4.2 What are the consequences of RG transformations ?

Since an RG transformations into blocks of length scale l, here denoted by R_l, will include l^d spins S_i, RG reduces the degrees of freedom from N described by a the original **Hamiltonian** H_N to $N' = N/l^d$ where the block spins S_I are described by an **effective Hamiltonian** $H'_{N'}$. Furthermore, even though the actual **correlation length** measured in a is unchanged, the **correlation length** measured in units of la vs a changes to

$$\xi_l = \xi_1 / l \tag{4.3}$$

An important consequence of this is that successive RG transformation push the new block spin **Hamiltonian** further from **criticality**, to longer distances and lower temperatures t_l. Under and RG transformation, the **coupling constants** K_n of a generic **Hamiltonian** (3.1) transform

$$[K'] = R_l[K] \tag{4.4}$$

Here $[K']$ refers to the set of RG transformed **coupling constants**. RG transformations are not reversible as two different systems may give rise to the same block spin Hamiltonian. Therefore, these transformations form a semi-group

$$R_{l_1 l_2}[K] = R_{l_1} R_{l_2}[K] \tag{4.5}$$

Computing R_l is usually neither simple nor unique. Consequently, an equivalent approximation can be found even in abstract scenarios [1].

Despite these difficulties, we can gain certain insights into the behavior of a system subject to **coarse-graining** and subsequent re-scaling. How are the partition function and consequently free energy affected by RG transformations ? As we mentioned above, **coarse-graining** reduces the degrees of freedom by a factor equal to the size of the block l^d i.e. from N to $N' = N/l^d$, giving rise to an **effective Hamiltonian** $H'_{N'}$ of block variables $\{S'_I\}$. A common approach that we will also encounter in the last chapter on Statistical Field Theory is to do a partial trace over the degrees of freedom within $\{S_i\}$ while keeping the block degrees of freedom $\{S'_I\}$ fixed.

$$e^{H'_{N'}\{[K'],S'_I\}} = Tr'_{\{S_i\}} e^{H_N\{[K],S_i\}} = Tr_{\{S_i\}} P(S_i,S'_I) e^{H_N\{[K],S_i\}} \tag{4.6}$$

This seems daunting to look at but in the case of Ising model with an odd number of spins in the blocks, this block spins and the projection operator $P(S_i,S'_I)$ would look like

$$S'_I = sign \sum_{i \in I} S_i = \pm 1$$

and

$$P(S_i,S'_I) = \prod_I \delta(S'_I - sign \sum_{i \in I} S_i)$$

i.e. 1 if in the blocks and zero otherwise. We can use the properties of the projection operator to prove that while the partition function remains invariant under RG transformations

$$Z_{N'}[K'] = ... = Z_N[K]$$

the free energy per degree of freedom (note different notation) will be re-scaled

$$g[K] = \frac{1}{N} \log Z_N[K] = ... = \frac{1}{l^d} g[K'] \tag{4.7}$$

[1] Although it is not clear how these results can be extended to abstract cases, where the notions of space, length, blocks etc can be hard to define, this is good news for what we are trying to achieve in the case of Neural Networks.

An important fact to note is that since at every step of RG transformation we sum over a finite number of degrees of freedom, the local operators are analytic.

4.3 The Critical Manifold in the Theory Space

Consecutive RG transformations, define trajectories in the space of **coupling constants**. Since each iteration reduces the **correlation length**, RG transformations move the theory away from **criticality**. The flows created by RG transformations can either lead to fixed points or infinity[2]. A fixed point of the RG transformation is defined as one that is invariant under the transformation

$$[K^*] = R_l[K^*] \tag{4.8}$$

A consequences of the fact that an R_l reduces the length scale (measured in Ångstrom)[3] by a factor l is that the **correlation length** ξ at the fixed point has to satisfy

$$\xi[K^*] = \xi[K^*]/l \tag{4.9}$$

which in turn results in $\xi = 0$ or $\xi = \infty$, referred to as trivial and critical fixed points, respectively.

The basin of attraction of a fixed point, also known as the **critical manifold**, is defined as the set of points that flow into the fixed point. It can be proven that all points in the basin of attraction of a critical fixed point have infinite **correlation length**. The critical fixed points (with $\xi = \infty$) are associated with singular critical behavior and the trivial fixed points ($\xi = 0$) describe the bulk **phases** of the system. We will see that it is the behavior of the system near particular fixed points that is associated with **scaling**behavior.

[2] Although it is theoretically possible to also have limit cycles too.

[3] The actual **correlation length** measured in $\mathring{A}$ is unchanged but the **correlation length** measured in la vs a is changed :

$$\xi_l = \xi_1/l$$

4.4 What is the Origin of Singular Behavior ?

A system described by the **Hamiltonian** H before an RG transformation and a **Hamiltonian** H_l after an RG transformation is the same apart from the difference in spacing between the spins(a vs la). Therefore, the spin block **Hamiltonian** has lower temperature and is further from **criticality**. Since consecutive **coarse-graining** result from performing a sum over a finite number of degrees of freedom an RG transformation is an analytic process. However, non-analyticities arise from the mere fact that we may have to do infinite RG iterations during which all degrees of freedom are integrated out. This is, for instance, what we do in the thermodynamic limit $N \to \infty$, where an infinite number of RG transformations would be needed in order to eliminate all the degrees of freedom.

This can happen also in much more mundane situations. This is no stranger than the observation that depending on the starting point of a particle in a perfectly analytic two-well potential, its final destination is a discontinuous function of its initial position as it can role into one of the two minima. This non-analyticity is not due to pathologies associated with the potential well but due to the fact that even if the position of the particle is a continuous function of its initial position for a finite time, its final destination need not be [22].

4.5 Understanding Universality

In the current context, the possibility of describing seemingly different types of matter/phenomena in a unified manner i.e. with the same **Hamiltonian**, places them in a class of common **universality**. This way, water and glass are in the same **universality** class and the only reason they appear different is that glass flows much more slowly than water. **Universality** is related to the behavior of the systems close to but not at a **criticality**. To understand this, the system is usually analyzed slightly off the critical manifold. Performing an RG transformation on a **coupling constant** K_n at a point close to the critical manifold $K_n = K_n^* + \delta K_n$, we get

$$\begin{aligned} K_n' &= K_n^* + \sum_m \frac{\partial K_n'}{\partial K_m}|_{K_m = K_m^*} \delta K_m + O((\delta K)^2) \\ &= K_n^* + M_{nm} \delta K_m + O((\delta K)^2) \end{aligned} \tag{4.10}$$

The linearized RG transformation near K^*

$$\delta K_n' = \sum_m M_{nm} \delta K_m \tag{4.11}$$

can help clarify how RG flows behave near a fixed point. Assuming, for clarity, that the above matrix $\mathbf{M}$ is symmetric, it can be expanded in the its eigen-directions.

$$\delta \mathbf{K}' = \mathbf{M} \delta \mathbf{K} = ... = \sum_s a^{(s)} \lambda^{(s)} \mathbf{e}^{(s)} = \sum_s a'^{(s)} \mathbf{e}^{(s)} \tag{4.12}$$

where s enumerates the eigenvectors/eigenvalues and $a'^{(s)}$ is the projection of $\delta \mathbf{K}$ on the eigen-directions. Depending on whether the eigenvalues $|\lambda^{(s)}| > 1$, $|\lambda^{(s)}| < 1$ or $|\lambda^{(s)}| = 1$, the components in the eigen-direction will grow(relevant), shrink(irrelevant) or remain the same (marginal).

4.6 The Origin of Scaling Laws

How do RG transformations account for **scalinglaws** ? Since RG transformations form a semi-group

$$R_{(ll')}[K] = R_{l'} R_l [K] \tag{4.13}$$

we can use the above linearization to write

$$\mathbf{M}^{(l)} \mathbf{M}^{(l')} = \mathbf{M}^{(ll')} \quad => \quad \lambda_l^{(s)} \lambda_{l'}^{(s)} = \lambda_{ll'}^{(s)} \tag{4.14}$$

which has the solution $\lambda_l^{(s)} = l^{y_s}$. This means that some (relevant) eigenvalues drive slightly off-critical systems away from the fixed points regardless of the initial values of the **coupling constants**. We can get this result in a slightly different manner. Suppose we have a system with one **coupling constant**, playfully called T. After an RG transformation from T to T' near a fixed point T^*, we can write

$$T' - T^* = R_l(T) - R_l(T^*) = \frac{\partial R_l}{\partial T}|_{T=T^*}(T - T^*) + O((T - T^*)^2) \tag{4.15}$$

Again, the linearized RG transformation has the solution

$$\frac{\partial R_l}{\partial T}|_{T=T^*} = l^{y_t} \tag{4.16}$$

which after n iterations (and some modifications and re-definitions) can be expressed as

$$t^{(n)} = t(l^{y_t})^n \tag{4.17}$$

Similarly, n-fold RG iterations, transform the **correlation length** to

$$\xi(t) = l^n \xi(t^{(n)}) = l^n \xi(t l^{n y_t}) \tag{4.18}$$

which with a suitable choice of variables, can be written in a more familiar from (in the context of **scaling laws**)

$$\xi(t) = (t/b)^{-1/y_t} \xi(b) \tag{4.19}$$

with b as an arbitrary large positive number. The relevant quantities can be read off

$$\xi \sim t^{-\nu} \sim t^{-1/y_t} \tag{4.20}$$

4.7 RG Applied to The Ising Model in two dimensions (Exact Calculation)

Here, we will put the above in practice by applying RG to the $2D$ Ising model. This is quite standard and illustrates how RG works in practice and what problems may lie ahead if we try to apply RG in other models[22]. Consider a $2D$ Ising model with triangular block spins where the value of the block spin is decided by a majority rule. The **Hamiltonian** for such a system, and all other systems in fact, is largely determined by the symmetries. The simplest such **Hamiltonian** (a modification of (3.12)) with nearest neighbor interaction is given by

$$H = K \sum_{<ij>} S_i S_j + h \sum_i S_i \tag{4.21}$$

4.7.1 Coarse-Graining

In the process of **coarse-graining**, each block spin S_I can arise from four different combination of 3 spins $|\{\sigma_I\}| = 4$ and the distance between the block spins is easily calculated to $\sqrt{3}$. The **effective Hamiltonian** H' after an RG transformation can be approximated using perturbation theory on the original **Hamiltonian** ($H = H_0 + V$) which in this context amounts to separating spin interactions within a block of spins H_0 and the interactions between spins in different blocks V.

$$H_0 = K \sum_I \sum_{i,j \in I} S_i S_j, \qquad V = K \sum_{I \neq J} \sum_{i \in I, j \in J} S_i S_j \tag{4.22}$$

Defining the average of a quantity A with respect to H_0 as

$$\langle A \rangle_0 = \frac{\sum_{\{\sigma_I\}} e^{H_0\{S_I,\sigma_I\}} A(S_I, \sigma_I)}{\sum_{\{\sigma_I\}} e^{H_0\{S_I,\sigma_I\}}} \tag{4.23}$$

we can easily prove

$$e^{H'} = \sum_{\{\sigma_I\}} e^{H\{S_I,\sigma_I\}} = ... = \langle e^V \rangle_0 \sum_{\{\sigma_I\}} e^{H_0\{S_I,\sigma_I\}} \tag{4.24}$$

If there are M blocks and Z_0 is the partition function for one block, we get

$$e^{H'} = \langle e^V \rangle_0 Z_0^M \tag{4.25}$$

Assuming that V is small, we can do the following expansion in terms of cumulants

$$\langle e^V \rangle_0 = e^{\langle V \rangle_0 + \frac{\langle V^2 \rangle_0}{2} - \frac{\langle V \rangle_0^2}{2} + O(V^3)} \tag{4.26}$$

And the **effective Hamiltonian** is given by

$$H'\{S_I\} = MlogZ_0 + \langle V \rangle_0 + \frac{\langle V^2 \rangle_0}{2} - \frac{\langle V \rangle_0^2}{2} + O(V^3) \tag{4.27}$$

Note that the first term(easily calculated by summing over all possible spin configuration in one block) is the contribution from a finite number of blocks and so it does not contribute to the singular behavior. We can write V, the interaction between blocks of spins as

$$V = \sum_{I \neq J} V_{IJ} \tag{4.28}$$

where

$$V_{IJ} = KS_3^J(S_1^I + S_2^I) \tag{4.29}$$

due to the particular nature of interaction between triangular blocks. After some calculation we get

$$\langle V \rangle_0 = 2K\Phi(K)^2 \sum_{<IJ>} S_I S_J \tag{4.30}$$

where

$$\Phi(K) = \frac{e^{3K} + e^{-K}}{e^{3K} + 3e^{-K}} \tag{4.31}$$

And the **effective Hamiltonian** to the first order in V is given by

$$H'\{S_I\} = M log Z_0(K) + K' \sum_{<IJ>} S_I S_J + O(V^2) \tag{4.32}$$

with $K' = 2K\Phi(K)^2$.

4.7.2 The phase space

At this point, the analysis of the phase space is quite trivial. We first find the fixed points of the RG transformation through

$$K^* = 2K^*\Phi(K^*)^2 \tag{4.33}$$

which has the solutions $K^* = 0$, $K^* = \infty$ or $\Phi(K^*) = 1/\sqrt{2}$ where the latter relationship gives the non-trivial fixed point. Inserting this in (4.31), we get

$$K_c = \frac{log(1+\sqrt{2})}{4} \tag{4.34}$$

and the eigenvalue of the linearized RG transformation

$$\lambda_t = \frac{\partial K'}{\partial K}|_{K_c} = 1.62 \tag{4.35}$$

Since we used perturbation theory, this is an approximation. More precise calculations can be done with the inclusion of higher order terms.

CHAPITRE 5

RENORMALIZATION GROUP THEORY IN QUANTUM FIELD THEORY

We have seen the application of RG to lattice models. Although it is possible that this is most suitable approach to **Hopfield Networks** or **Restricted Boltzmann Machines** (RBM), which we are aiming to do, it is not clear whether we should imagine neural networks as lattices. And if we do, it does not seem reasonable to discuss **coarse-graining**, particularly block-spin transformations and **correlation lengths** in the absence of a lattice structure. Yes, some researchers point at the similarities between **coarse-graining**, and the inner workings of RBMs in the context of information theory but as far as we have seen, their goal is to find meaningful **coarse-graining** schemes in physical systems where our knowledge of the **microscopic** structure of the system is limited. Our goal is quite the opposite. We are hoping to prove that the observed scaling laws in neural networks can be studied in the context of RG in the absence of a lattice structure.

In this section, we explore RG from the point of view of **Quantum Field Theory**. It should be noted that RG is used in similar manner in a wide range of fields from Quantum Electrodynamics to Quantum Chromo-dynamics to Statistical Field Theory to which we will return at the end of this work. In all these cases, a **path-integral** formulation is used. Our objective is to understand if we can move beyond prototyping RBMs as "**coarse-graining machines**" and to general NNs. Furthermore, if we make the bold assumption that NNs really do perform some kind of **coarse-graining** beyond RBMs, it is plausible to start with RG as a working hypothesis that will hopefully explain the empirical results of recent few years.

Is it possible to explain the empirical finding that many algorithms seem to behave similarly in certain limits as **algorithmic universality** ? Is it possible that discovered scaling laws in NNs set a limit or constraint on our ability to perform certain calculations ? Is there an algorithm (**M-algorithm**, perhaps) [1] capable of producing the same results as all possible algorithms ? As far as we know, there is no thorough theoretical understanding of why we observe **algorithm-agnostic phase transitions**.

In the **Wilsonian RG**, large momenta above a certain so-called cut-off are integrated out. This is equivalent to removing short-distance behavior/fluctuations of the system as a result of which the remaining theory is a lower energy description of the original system. To this end, a cut-off Λ and a dimensionless parameter b are introduced and the field is separated into low and high energy parts :

$$\phi(x) = \phi_l(x) + \phi_h(x)$$

and integrate out the fields with high-momenta k in the interval $b\Lambda \leq k \leq \Lambda$

$$e^{-S_{eff}[\phi_l]} = e^{-S[\phi_l]} \int \mathscr{D}\phi_h e^{-S[\phi_l,\phi_h]} e^{-S[\phi_h]}, \tag{5.1}$$

where we have put $h = 1$ for simplicity. While the purely high energy part of this is just a normalization factor, the interesting contribution comes from the part of the action that involves both ϕ_l and ϕ_h. The partition function in terms of the **effective action** (or Largrangian) is then given by

$$Z = \int \mathscr{D}[\phi_l] e^{-\frac{1}{h}S_{eff}(\phi_l)} = \int \mathscr{D}[\phi]_{b\Lambda} e^{-\frac{1}{h}\int d^dx \mathscr{L}_{eff}(\phi)}. \tag{5.2}$$

What the **effective action** (Euclidean) looks like, how it changes across energy scales and how it is interpreted depends on the action. For instance, the prototypical ϕ^4-theory above, often used in QFT to understand the RG flow, provides an adequate understanding of RG flows as it leads to the general discovery that operators of RG flow can be classified as relevant, irrelevant and marginal. In general, an RG transformation is defined such that the functional form of the Lagrangian is kept intact while rescaling/redefining other entities.

[1] Similar to M-theory being the origin of all string theories

$$Z=\int\mathscr{D}[\phi_l]e^{-S(\phi_l)}\int\mathscr{D}[\phi_h]e^{-\int d^dx\{\frac{1}{2}(\partial_\mu\phi_h)^2+\frac{1}{2}m^2\phi_h^2+\lambda(\frac{1}{6}\phi_l^3\phi_h+\ldots)\}}. \tag{5.3}$$

where

$$S(\phi_l)=\int d^dx\{\frac{1}{2}(\partial_\mu\phi_l)^2+\frac{1}{2}m^2\phi_l^2+\frac{\lambda}{4!}\phi_l^4\}$$

Note that due to orthogonality condition in the momentum space i.e. Fourier transforming the fields,

$$\phi(x)=\int\frac{d^dk}{(2\pi)^d}e^{-ik.x}\phi(k), \tag{5.4}$$

the mixed terms linear in ϕ_l and ϕ_h and/or their derivatives do not contribute to the above action. Integrating out the high energy modes in the first part of the above action results in the **effective action**

$$Z=\int\mathscr{D}[\phi_l]e^{-S_{eff}(\phi_l)}=\int\mathscr{D}[\phi_l]e^{-\int d^dx\{\frac{1}{2}(\partial_\mu\phi_l)^2+\frac{1}{2}m^2\phi_l^2+\frac{\lambda}{4!}\phi_l^4+O(\lambda)\}} \tag{5.5}$$

Note that apart from integrating out the high energy modes in the mixed terms, which will contribute with λ-correction terms to the effective action, it is also customary to disregard the high energy mass term as due to the fact that $m^2\ll\Lambda$. As such, the largest contribution will come from the kinetic high energy term.

$$Z\sim\int\mathscr{D}[\phi_h]e^{-\int d^dx\frac{1}{2}(\partial_\mu\phi_h)^2}=\int\mathscr{D}[\phi_h]e^{-\int\frac{d^dk}{(2\pi)^d}\frac{1}{2}\phi_h(k)k^2\phi_h(-k)} \tag{5.6}$$

We use the method of external source

$$Z\sim\int\mathscr{D}[\phi_h]e^{-\int\frac{d^dk}{(2\pi)^d}\{\frac{1}{2}\phi_h(k)k^2\phi_h(-k)+J(k)\phi_h(k)\}} \tag{5.7}$$

and demand $\frac{k^2}{(2\pi)^d}$ to be the momentum space propagator by imposing

$$D_F(k+p)=\frac{(2\pi)^d}{k^2}\delta^d(k+p) \tag{5.8}$$

The remaining terms can now be integrated out using Wick's theorem i.e. approximating each of the terms as $e^{-x} = 1 - x + ...$ For instance

$$e^{-\int d^dx\frac{\lambda}{4}\phi_l^2(x)\phi_h(x)\phi_h(x)} = 1 - \frac{\lambda}{4}\int d^dx\phi_l^2(x)\phi_h(x)\phi_h(x) \tag{5.9}$$

which in momentum space results in

$$\begin{aligned}
Z &\sim \int \mathscr{D}[\phi_l]e^{-S[\phi_l]}\phi_l^2(x)\frac{-\lambda}{4}\int \mathscr{D}[\phi_h]e^{-\int\frac{d^dk}{(2\pi)^d}\{\frac{1}{2}\phi_h(k)k^2\phi_h(-k)\}}\phi_h(q)\phi_h(p) && (5.10)\\
&\sim \int \mathscr{D}[\phi_l]e^{-S[\phi_l]}\phi_l^2(x)\frac{-\lambda}{4}\int d^dqd^dpD_F(p+q) && (5.11)\\
&\sim \int \mathscr{D}[\phi_l]e^{-S[\phi_l]}\phi_l^2(x)\frac{-\lambda}{4}\int \frac{d^dq}{(2\pi)^d}\frac{d^dp}{(2\pi)^d}(2\pi)^d\frac{1}{q^2}\delta(p+q) && (5.12)\\
&\sim \int \mathscr{D}[\phi_l]e^{-S[\phi_l]}\phi_l^2(x)\frac{-\lambda}{4}\int_{b\Lambda}^{\Lambda}\frac{d^dq}{(2\pi)^d}\frac{1}{q^2} && (5.13)\\
&\sim \int \mathscr{D}[\phi_l]e^{-S[\phi_l]}\frac{-\mu}{2}\phi_l^2(x) && (5.14)
\end{aligned}$$

which looks like A contribution to the mass of low energy modes if we define

$$\mu = \frac{\lambda}{2}\int_{b\Lambda}^{\Lambda}\frac{d^dq}{(2\pi)^d}\frac{1}{q^2} \tag{5.15}$$

This cut-off and dimension-dependent quantity is interpreted as a shift in the mass term m^2. This is in fact a general conclusion in the **Wilsonian RG** : Whatever IR terms we have in the theory, are the result of integrating out high energy modes. It is quite clear that this process can potentially generate an infinite tower of higher dimensional operators that were either present or absent in the original theory. For instance, the same term in second order in λ

$$(\frac{-\lambda}{4})^2\int d^dx\phi_l^2(x)\phi_h^2(x)\int d^dy\phi_l^2(y)\phi_h^2(y) \tag{5.16}$$

generates a correction to the ϕ^4-term, present in the original theory. But the $\phi_l^3\phi_h$ in $O(\lambda)$ gives a contribution towards ϕ_l^6 that was not there to begin with. Another interesting fact in the RG theory is that while all corrections are cut-off dependent, some are renormalizable but others are not. For instance, due to division by $d-4$, the contribution to the phi^4-term is non-renormalizable in 4 dimensions. Neither is the new $\frac{\lambda^2}{\Lambda^2}\phi_l^6$ operator. Interestingly enough, we also see that this operator was not visible in the low energy theory simply due to being suppressed by the cut-off. This is in fact the definition of a renormalizable theory as one which has a high enough cut-off compared to its typical energy scale.

5.1 RG flow in Quantum Field Theory

In summary, integrating out high energy modes leads to a modification of existing parameters of the theory. These modifications appear as changes in the couplings, masses, fields etc. We also know that integrating out corresponds to a change in perspective i.e. the scale at which we examine the theory. Although there is a natural interpretation of scale in lattices, it is more natural to consider the change of scale as a continuous process. In other words, we are interested in understanding how the parameters of the theory change continuously with scale.

Under a rescaling with $b < 1$, the relationship between high and low momentum modes (k' and k, respectively) or equivalently short and long distances (x' and x) is given by

$$k' = \frac{k}{b}, \quad x' = bx, \quad \Longrightarrow d^dx = b^{-d}d^dx', \quad \partial_\mu = \frac{\partial}{\partial x_\mu} = b\frac{\partial}{\partial x'_\mu} = b\partial'_\mu \tag{5.17}$$

As we saw, integrating out the high energy modes results in an action with modified couplings. Suppose we have done this and that the resulting low energy action is given by

$$\begin{aligned} S[\phi] = \int d^dx\{ & \frac{1+\Delta Z}{2}(\partial_\mu\phi)^2 + \frac{m^2+\Delta m^2}{2}\phi^2 \\ & + \frac{\lambda+\Delta\lambda}{4!}\phi^4 + \Delta C(\partial_\mu\phi\partial_\mu\phi)^2 + \Delta D\phi^6 + ...\} \end{aligned} \tag{5.18}$$

Note that this action is the result of integrating out the high energy modes only and that it does not take into account the rescaling upon which the action changes to

$$S[\phi] = \int b^{-d} d^d x' \{ \frac{1+\Delta Z}{2} b^2 (\partial'_\mu \phi)^2 + \frac{m^2 + \Delta m^2}{2} \phi^2 \\ + \frac{\lambda + \Delta\lambda}{4!} \phi^4 + \Delta C b^4 (\partial'_\mu \phi \partial'_\mu \phi)^2 + \Delta D \phi^6 + ... \} \tag{5.19}$$

The requirement that the free-field Lagrangian $\frac{1}{2}(\partial_\mu \phi)^2$ is written in canonical form i.e. should remain unchanged (thereby creating the so-called Gaussian fixed point of the RG flow in the space of all possible Lagrangians) determines the full RG transformation of the above action.

$$\phi' = [b^{2-d}(1+\Delta Z]^{1/2} \phi \tag{5.20}$$

$$m'^2 = (m^2 + \Delta m^2)[1+\Delta Z]^{-1} b^{-2} \tag{5.21}$$

$$\lambda' = (\lambda + \Delta\lambda)[1+\Delta Z]^{-2} b^{d-4} \tag{5.22}$$

$$C' = \Delta C[(1+\Delta Z]^{-2} b^d \tag{5.23}$$

$$D' = \Delta D[(1+\Delta Z]^{-3} b^{2d-6} \tag{5.24}$$

The new action is given by

$$S[\phi'] = \int d^d x' \{ \frac{1}{2} (\partial'_\mu \phi')^2 + \frac{1}{2} m'^2 \phi'^2 + \frac{1}{4!} \lambda' \phi'^4 + C' (\partial'_\mu \phi' \partial'_\mu \phi')^2 + D' \phi'^6 + ... \} \tag{5.25}$$

5.2 What is Renormalizability ?

Above we mentioned that after an RG transformation, the corrections ξ to the ϕ^4 term $\frac{-\xi}{4!} \int d^d x \phi_l^4(x)$, given by

$$\xi = -4!(\frac{\lambda}{4})^2 \int_{b\Lambda}^{\Lambda} \frac{d^d k}{(2\pi)^d} \frac{1}{(k^2)^2} = -\frac{1}{(2\pi)^d} \frac{1}{\Gamma(d/2)} \frac{(1-b^{d-4})\Lambda^{d-4}}{d-4} \tag{5.26}$$

is not renormalizable in 4 dimensions. This dependence on dimensionality is intimately connected to changes in scale. We have seen that the evolution of operators and couplings in RG. The real question to be asked is how how the change of operators affect the theory across energy scales. If we consider the effect of rescaling (ignoring the quantum corrections we get from integrating out the high momentum modes, the most generalized action

$$S[\phi] = \int d^d x \sum_j c_j O_j(\phi) \tag{5.27}$$

will transform into

$$S[\phi'] = \int d^d x' b^{-d} \sum_j b^{Nd_\phi + M} c_j O_j(\phi') = \int d^d x' \sum_j c'_j O_j(\phi') \tag{5.28}$$

Here, M, N and d_ϕ are the number of derivatives, the number of fields in the operator, and canonical dimension of ϕ, respectively. Comparing the transformed action to the previous action, we have

$$c'_j = b^{Nd_\phi + M - d} c_j = b^{d_{O_j} - d} c_j \tag{5.29}$$

Here, we are interested in the behavior of the operators under continuous RG transformations. It is easily proven that

$$x \frac{dc_j}{dx} = -(d_{O_j} - d) c_j \tag{5.30}$$

which helps classify operators. For instance, if $d_{O_j} = d$, we can conclude that resclaing does not change the coupling. This is called a marginal operator. The other two cases are classified similarly : If $d_{O_j} < d$, the derivative will be positive and c_j will grow with x i.e. going towards the IR regime. These are the so-called relevant operators. The only remaining option is $d_{O_j} > d$ i.e. when the derivative shrinks with x. These so-called irrelevant operators are non-renormalizable. The problem with these operators is that they become suppressed in IR, the higher the cut-off is. In other words, they will not be visible in the large scale theory. In principle, the higher the cut-off, the fast the irrelevant operators become invisible for larger distances.

5.3 Callan-Symanzik Equation

Understanding the RG flow in the language of **correlation functions** is a practical approach to making contact with actual measurements such as amplitudes in QFT. The idea is to investigate how physically measurable quantities behave under RG transformations. The demensionless parameter b introduced above to set the line between low and high momenta can be understood as a parameter controlling the cut-off. Here, we want to understand how amplitudes/**correlation functions** behave under a continuous change of the cut-off.

In calculating the **correlation functions**, the (imposed renormalization condition determines what is called a renormalization scale μ. Variation of this scale defines a renormalization flow of the parameters of the theory. This way, the same problem is cast as the renormalization scale flow. Above we saw that the relationship between the unrenormalized and renormalized fields (ϕ_0 and ϕ) is given by

$$\phi(x) = Z^{-1/2}\phi_0(x) \tag{5.31}$$

As a result, the **correlation functions** transform as

$$\langle T\phi(x_1)...\phi(x_n)\rangle = Z^{-n/2}\langle\phi_0(x_1)...\phi_0(x_n)\rangle \tag{5.32}$$

of which the relevant i.e. connected n-point **correlation functions** transform as

$$G^{(n)}(x_1,...,x_n) = Z^{-n/2}G_0^{(n)}(x_1,...,x_n) \tag{5.33}$$

The non-renormalized/renormalized **correlation functions** depend on ϕ_0/ϕ, m_0/m, λ_0/λ, respectively. Furthermore, while the non-renormalized **correlation functions** depend on the cut-off Λ (but not the renormalization scale), the renormalized **correlation**

functions depend on the renormalization scale μ. This means that

$$\frac{dG_0^{(n)}}{d\mu} = 0 \tag{5.34}$$

a consequence of which is the Callan-Symanzik equation

$$\begin{aligned}&[\mu\frac{\partial}{\partial\mu} + \mu\frac{\partial\lambda}{\partial\mu}\frac{\partial}{\partial\lambda} - n\mu\frac{\partial\eta}{\partial\mu}]G^{(n)}(x_1,...,x_n,\mu,\lambda) = 0 \\ &[\mu\frac{\partial}{\partial\mu} + \beta\frac{\partial}{\partial\lambda} - n\gamma]G^{(n)} = 0\end{aligned} \tag{5.35}$$

In its most general form, the Callan-Symanzik equation can be derived based on the observation that the β function's dependence on the scale comes from the counterterms. A schematic description of an n-point function with a fictitious coupling g is given by

$$\begin{aligned}G^{(n)} = [&-ig + 1PI \text{ loops} \\ &+ \text{vertex counterterms} \\ &+ \text{external leg loops} + \text{external leg counterterms}]\prod_i^n \frac{i}{p_i^2}\end{aligned} \tag{5.36}$$

Or equivalently

$$G^{(n)} = [-ig + 1PI \text{ loops} - i\delta g - ig\sum_{j=1}^{n}(\text{external leg loops} - \delta Z_i)]\prod_i^n \frac{i}{p_i^2} \tag{5.37}$$

where i is the fields/legs that may or may not be different, hence a counter terms δZ_i is inserted for each of these fields. Applying the Callan-Symanzik equation on this generic n-point function (skipping some steps) results in an expression for the β function

$$\beta(g) = \mu\frac{\partial}{\partial\mu}(\frac{1}{2}g\sum_i \delta Z_i - \delta g) \tag{5.38}$$

which can be calculated based on the renormalization scale dependence of the counterterms. A quick glance at the ϕ^4 theory shows that these are giveing by the coefficients of the divergences.

5.4 The beta-function

The CS equation is useful in the context of perturbation theory employed in the renormalization procedure which also helps calculate the β-function. The latter provides an understanding of the flow of QFTs across energy scales. Suppose we calculate a β-function $\beta(g)$ with g is some coupling.

$\beta(g) > 0$: This means that the coupling g increases with the energy E. As the coupling grows beyond a certain limit, the theory becomes non-perturbative but as long as it remains perturbative, the growth of the coupling can be traced.

$\beta(g) < 0$: The coupling decreases with increasing energy. In the limit of infinitely high energies the coupling goes to zero, a so-called asymptotic freedom where the coupling reaches a UV fixed point. On the contrary, the coupling grows towards the IR which also means that the theory no longer is perturbative. For example, if

$$\beta(g) = \frac{dg}{d\ln\mu} = -\frac{1}{2}Cg^3, \;\; C > 0 \tag{5.39}$$

we get

$$g^2(q) = \frac{g^2(\mu_R}{1 + Cg^2(\mu_R)\ln(q/\mu_R)} \tag{5.40}$$

where μ_R is a reference scale. We see a logarithmically decreasing coupling towards the UV with a fixed point at infinity.

$\beta(g) = 0$: The coupling g is energy- or length-scale independent. Scale invariance is an interesting limit of RG flow. This property (also called conformal invariance) is often imposed on theories to learn about their properties otherwise. This vanishing point is a fixed point of the RG flow. If a theory has a positive beta function for small values of the coupling, and the function is well-behaved, it will have to change sign. The coupling grows with the scale as long as the beta function is positive. As the beta-function crosses zero, the behavior is reversed and the coupling shrinks with growing energy scales. This makes the fixed point a UV stable fixed point in the sense that we reach this point no matter how we approach it.

An IR fixed point is possible in the case where the beta function is negative for small values of the coupling constant hence grows with decreasing energy scale. As the function crosses the fixed point and becomes positive, the flow is reversed and the coupling decreases with decreasing scale. An interesting approach is to linearize the beta function in the vicinity of the fixed point $g*$. Assuming that $B > 0$, this means that in the case when β is positive for $g < g*$ and negative when $g > g*$:

$$\begin{aligned}
\beta(g) &\simeq -B(g-g*) \\
\frac{dg}{d\ln\mu} &= -B(g-g*) \\
\frac{dg}{g-g*} &= -Bd\ln\mu \\
\int_{\mu_R}^{q}\frac{dg}{g-g*} &= \int_{\mu_R}^{q} -Bd\ln\mu = \ln(\frac{q}{\mu_R})^B \\
g(q) &\simeq g*+(g(\mu_R)-g*)(\frac{\mu_R}{q})^B
\end{aligned} \tag{5.41}$$

when clearly shows the UV fixed point i.e. the coupling approaches the fixed point in the high energy limit.

CHAPITRE 6

ENERGY-BASED MODELS

It seems difficult to apply RG to models that are not equipped with a **Hamiltonian** or an energy function. There is an interesting interplay between energy-based models in neural networks and the physics of interacting particles. Historically, these models include Hopfield Networks, Boltzmann Machines and RBMs some of which have evolved over time from discrete to more widely applicable or realistic scenarios with continuous variables. The most fundamental characteristic of these models is that they are probabilistic models where a **Lyapunov function** (Here understood as a **Hamiltonian**) is used to describe the probability of any particular state of the system.

It is a fact of life that the presumably complex interactions between microscopic entities of any system is beyond access to us. Instead, we observe the macroscopic properties or what is known as coarse-grained information about the system. One such property is the total energy of a system, defined as a **Hamiltonian** $H(x)$ which determines the energy values of each possible state of a system of N particles where x is a vector whose elements are the degrees of freedom of the particles.

Energy-based learning is an alternative approach to probabilistic estimation for many machine learning tasks where the usual necessity of estimating normalized probability densities that more often than not are intractable, can be avoided. This, simply because the objective is to lower the energy of the system (this is easily seen in Hopfield networks). Still, the only consistent approach is to turn the collection of all possible energies and possible outcomes into a normalized (Gibbs) distribution [42].

$$P(Y|X) = \frac{e^{-\beta E(Y,X)}}{\int_{y\in\mathscr{Y}} e^{-\beta E(Y,X)}} \tag{6.1}$$

This is of course meaningful only if the denominator, the partition function, is tractable.

This is not necessarily a problem since, as we will see in the last chapter of this work, the partition function can be constructed based on the symmetries of the system. In the context of energy-based models, training means finding the best energy function in a family of all possible functions. This is qualitatively measured by a **loss function**al that is minimized during the learning procedure.

Energy-based models assign a probability to each energy level. The probability distribution in question is **Boltzmann distribution**

$$P(x) = \frac{e^{\frac{-H(x)}{T}}}{Z} \tag{6.2}$$

which is derived from Jayne's **maximum entropy principle** [32] stating that the most probable state of a system on the basis of partial information must be one determined by a distribution with largest possible **entropy**.

$$\max_{P(x)} \sum_{x} -P(x) log P(x) \quad s.t. \quad \sum_{x} P(x)H(x) = \langle H(x) \rangle \tag{6.3}$$

The Boltzmann distribution also establishes a relationship between the likelihood of energy levels (the most probably energy levels are those with highest entropy), and the parameter T which is interpreted as temperature. Since a system is a collection of states of different energy, the temperature T can also be viewed as a measure of average energy of the system. In the limit of low temperature (See (6.2) the minima of the energy function i.e. the ground states are more likely. In contrast, in the limit of high or infinite temperature, all states are equally likely.

Inference in EBMs consists of clamping down the values of the observed variables and finding the values of other variables that minimize the energy. When the energy function is not known, learning is the process of finding an energy function in a family of functions such that the correct values of variables are associated with lower energies than the incorrect variables. A **loss function** is one that measures the quality of energy functions in the family to be considered.

Interestingly, while the inference algorithm selects the Y with lowest energy, learning shapes the energy landscape in such a way that the correct examples are associated with lower energy (lower loss) and incorrect examples with higher energies (higher loss).

Note also that this reshaping of the **loss landscape** is crucial in EBMs. For this reason, the per sample energy $E(W,Y^i,X^i)$ cannot be used as a **loss function** simply because it will not push up the **loss landscape** for incorrect answers [42]. On the other hand the negative log-likelihood loss (also known as **maximum mutual information** or **cross entropy** loss) which stems from maximizing the likelihood of the data $P(Y|X)$ satisfies the reshaping criterion for EBMs

$$\mathscr{L}_{NLL} = \frac{1}{P}\sum[E(W,Y^i,X^i) + \frac{1}{\beta}log\int_{y\in\mathscr{Y}} e^{-\beta E(W,y,X^i)}] \tag{6.4}$$

This **loss function** reshapes the energy landscape for every example by pushing up the energy of every example with an amount proportional to its likelihood (second so-called contrastive term)

$$\frac{\partial\mathscr{L}_{NLL}}{\partial W} = \frac{\partial E(W,Y^i,X^i)}{\partial W} - \int_{Y\in\mathscr{Y}} \frac{\partial E(W,Y,X^i)}{\partial W} P(Y|W,X^i) \tag{6.5}$$

Again, this integral is not always tractable [42] due to the similar difficulty of calculating the likelihood $P(Y|W,X^i)$ which in turn is related to the intractable partition function as in many other cases. We will in the last chapter of this work that Statistical Field Theory can circumvent this issue.

One of our ideas was to explore the interplay between the **loss function**, lowering the energy of the system, the principle of **maximum entropy** and RG theory.

CHAPITRE 7

TWO SPECIFIC ENERGY-BASED MODELS

The most fundamental step in designing energy-based systems is the choice of a **Hamiltonian** that reflects the most relevant interactions in the system. For instance, a particle system where energy is assigned to each particle plus pairwise interactions between particles that have two degrees of freedom is known as the Ising model. In this chapter we review two such models both of which are variations of the Ising model. And even if the variations do not seem significantly drastic, these models are much more difficult to handle. In this work, we mainly focused on analyzing Hopfield Networks from various angles.

7.1 Hopfield Networks

Hopfield networks [28], were introduced in 1982. These networks consist of a group of connected neurons each of which are given a certain value (+1 or -1 in discrete networks). The neurons interact through simple rules which can be shown to follow a form of **Hebbian rule** as a result of which Hopfield networks are suitable for repairing corrupted data or retrieving a certain pattern when given a partial pattern.

These networks are associated with an energy function. Training the network amounts to identifying the parameters of the network such that the minima or ground states of the energy function are the states of the input data. In its simplest form (binary valued nodes, one stored pattern or a few different patterns) it is easily proven that each update of the Hopfield network lowers its energy until the network settles down in a configuration of minimum energy (local minima corresponding to stored patterns) after which the updates do not have any effect on the network or the energy.

A binary Hopfield network that stores N d-dimensional patterns $\{\mathbf{x}_i\}_1^N$ in neurons whose activation values are $\{-1,1\}^d$, can be trained by constructing a matrix of outer products of the patterns

$$\mathbf{W} = \sum_1^N \mathbf{x_i}\mathbf{x_i}^T \tag{7.1}$$

which is used to retrieve the pattern corresponding to a particular state/corrupt pattern $\boldsymbol{\xi}$ according to the update rule

$$\boldsymbol{\xi}^{t+1} = sgn(\mathbf{W}\boldsymbol{\xi}^t - \boldsymbol{b}) \tag{7.2}$$

with convergence when $\boldsymbol{\xi}^{t+1} = \boldsymbol{\xi}^t$. Here $\boldsymbol{b}$ is a bias term that can be seen as a direct input to the neurons. The update rule can also be used asynchronously i.e. each component of $\boldsymbol{\xi}^t$ is updated separately until the minimum corresponding to the corrupt pattern in question is reached. Note that removing the bias term is equivalent to the energy of a pattern being equal to the energy of its inverse.

However, this type of (**Hebbian**) learning does not work well if the data vectors are not mutually orthogonal, in which case so-called spurious minima can appear. These minima may be combinations of other data vectors hence do not correspond to the actual data vectors. This way, the minima of the energy function may lead to wrong memory retrieval, a problem that is usually helped (see below) by designing **Lyapunov functions** that lead to a larger memory which in turn contributes to separating the basins of attraction of the local minima.

Another strategy involves so-called unlearning of spurious minima. Here, by modifying the matrix in 7.1 as in

$$w_{ij} = \langle x_i x_j \rangle_{data} - \varepsilon \langle x_i x_j \rangle_{model} \tag{7.3}$$

or equivalently injecting energy into the system, all the energy states , there-among the local and spurious minima, are lifted/moved a process that with a suitable choice of ε can lead to the removal/unlearning of spurious minima.

A more reasonable approach to lowering the energy of the system is in direct correspondence to the Boltzmann distribution being a connection between the energy of the system

$$E = -\frac{1}{2}\sum_{ij} W_{ij}\xi_i\xi_j + \sum_i b_i\xi_i \tag{7.4}$$

and its probability at a particular temperature. In order to mimic a realistic situation as dictated by the Boltzmann distribution, the nodes are flipped at zero temperature only if it lowers the energy, at higher temperature the nodes are flipped both if the energy is decrease and also when it is increased but with the probability $p = e^{\frac{-\Delta E}{T}}$. The nodes are left unchanged with a probability of $1 - p$ otherwise. This is the **Metropolis-Hastings algorithm**.

Useful as the above model has proven to be, perhaps a more realistic approach [27] is to consider a network of biological neurons with graded response. Note that this is also the energy function we used in this work. The graded response is usually taken to be a sigmoid input-output $V_i = g(x_i)$. Provided certain simple conditions on the matrix $\boldsymbol{T}$ below (symmetric and with zero diagonal elements) a Lyapounov function that is guaranteed to converge to stable states is given by

$$E = -\frac{1}{2}\sum_{i \neq j} T_{ij}V_iV_j - \sum_i I_iV_i + \frac{1}{\tau}\sum_i \int^{V_i} g^{-1}(z)dz \tag{7.5}$$

Assuming that the synaptic current has a lag behind the firing rate of the form $e^{\frac{-t}{\tau}}$, the evolution of the state of the network can be described by the ordinary differential equation

$$\frac{dx_i}{dt} = -\frac{x_i}{\tau} + \sum_{ij} T_{ij}V_j + I_j \tag{7.6}$$

It turns out the classic Hopfield networks are not able to retrieve a specific pattern if many similar such patterns are stored in the network. This problem was initially attributed to the memory capacity of the network. Hopfield networks have since been generalized from binary patterns to modern or dense associative memory networks with very large memory capacity. The energy of these is given by

$$E = -\sum_i F(\boldsymbol{x}_i^T \boldsymbol{\xi}) \tag{7.7}$$

where F is a polynomial [40] or an exponential [15] function. It is also shown [15] that the network will converge to a minimum with high probability and with the component-wise update rule

$$\boldsymbol{\xi}^{new}[l] = sgn[-E(\boldsymbol{\xi}[l^+]) + E(\boldsymbol{\xi}[l^-])] \tag{7.8}$$

after only one update of the entire vector $\boldsymbol{\xi}$. Here $\boldsymbol{\xi}[l^+] = 1$ and $\boldsymbol{\xi}[l^-] = -1$ and all the other components remain unchanged.

Note that in the exponential case [15] the authors prove that their network has an exponential memory capacity in terms of the number of neurons. This alone, however, is not the solution since even with an exponentially large memory capacity, the basins of attraction of each stored pattern can be as large as the original Hopfield networks, which are already known to have pattern retrieval problems. The issue is that if the basins of attraction are large and/or overlapping, the network will then converge towards a solution that may be close to a stored pattern or an average of several patterns instead of a particular pattern. In other words, Hopfield networks with large memory capacity can demonstrate a large number of metastable **fixed points**.

A recent generalization of the above modern Hopfield networks to the case of continuous valued inputs/patterns [55] proposes a new energy function and update mechanism that is equivalent to the attention mechanism in transformers, originally introduced in [62]. In this approach, the Lyapounov function i.e. energy is given by

$$E = -\frac{1}{\beta} log \sum_{i=1}^{N} e^{\beta \boldsymbol{x}_i^T \boldsymbol{\xi}} + \frac{1}{2}\boldsymbol{\xi}^T \boldsymbol{\xi} + c \tag{7.9}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ... \boldsymbol{x}_N)$, $c = \frac{1}{\beta} log N + \frac{1}{2} M^2$ and M is the largest norm of all stored patterns. The proposed update rule for this network is given by

$$\boldsymbol{\xi}^{new} = \boldsymbol{X}\mathrm{softmax}(\beta \boldsymbol{X}^{T} \xi) \tag{7.10}$$

This update rule is deduced from the concave-convex procedure (CCCP)[66] which guarantees that the energy function will decrease to a minimum or a saddle point.

7.2 Restricted Boltzmann Machines

Including other interactions in the **Hamiltonian** of Hopfield networks will inevitably lead to less manageable models in terms of the number of parameters. This is circumvented by the introduction of new "particles" to the model who help increase the complexity of the system without encoding the data. These so-called hidden units act as intermediaries between visible units that actually encode the data. These networks are known as Boltzmann Machine and are constructed exactly for the purpose of increasing the complexity of the model. When the interactions between hidden and visible units are set to zero, Boltzmann machines reduce to Hopfield networks.

Both Hopfield networks and Boltzmann machines are difficult to train as each local update of nodes also depends on other nodes. A simplification of Boltzmann machines is to cancel all the intralayer interactions i.e. interactions between the units within each hidden and visible layer. This gives rise to RBMs [41, 50] which is described by the **Hamiltonian**

$$E = -\sum_{i} b_i h_i - \sum_{i} c_i v_i - \sum_{ij} w_{ij} h_i v_j \tag{7.11}$$

where $\boldsymbol{h}$ and $\boldsymbol{v}$ are hidden and visible units, respectively. The direct benefit of training these networks is related to the fact that the energy change in updating visible/hidden nodes is independent from other visible/hidden nodes. As a consequence the conditional probabilities factorize

$$P(\boldsymbol{v}|\boldsymbol{h}) = \prod_i P(v_i|\boldsymbol{h}), \quad P(\boldsymbol{h}|\boldsymbol{v}) = \prod_i P(h_i|\boldsymbol{v}) \tag{7.12}$$

where the individual factors can be calculated separately. The training of RBM is done through the **Gibbs sampling algorithm** by iterating the following procedure until convergence : After fixing the hidden nodes, the individual conditional probability of the state of each node is sampled from $P(v_i|\boldsymbol{h})$. Then the visible nodes are fixed and the conditional probability of the hidden nodes are individually sampled from $P(h_i|\boldsymbol{v})$.

Note that spurious minima can occur in RBMs as well where a process similar to "unlearning" in Hopfield networks is employed to shape the energy landscape such that the minima of the energy function are associated with the data.

CHAPITRE 8

EXPLORATIONS

8.1 Project 1 : Studying phase transitions in Hopfield Networks from the perspective of Mean field theory

In this section we apply MFT to Hopfield networks. The reader is referred to Appendix I for all the calculations and technical details. These networks can be described by a **Hamiltonian** that is somewhat reminiscent of various forms of the Ising model.

$$H(v) = -\frac{1}{2}\sum_{ij} T_{ij} v_i v_j - \sum_i f_i v_i + \sum_i \Phi_i(v_i) \tag{8.1}$$

The problem is significantly simplified if we assume that we are only dealing with nodes that can take on the two values ±1 or equivalently black/white pixels in Hopfield terminology. The main assumption of the MFT is that the systems evolves towards a state where most nodes have values close to the average value of all the nodes. In other words, we assume that $v_i = \langle v_i \rangle + \delta v_i = m + \delta v_i$. This leads to $v_i v_j = \langle v_i \rangle v_j + \langle v_j \rangle v_i - \langle v_i \rangle \langle v_j \rangle$ and the mean field **Hamiltonian** can be written as (See I.1)

$$H_{MF}(v) = ... = -\sum_j [m \sum_i T_{ij} + f_j] v_j + \frac{m^2}{2} \sum_{ij} T_{ij} + \sum_i \Phi_i(v_i) \tag{8.2}$$

If T_{ij} are the elements of a matrix T, we can further simply this expression by defining the sum of the elements in column (or equivalently row) j as $C_j = \sum_i T_{ij}$. The mean field **Hamiltonian** can then be written as (see I.2)

$$H_{MF}(v) = ... = -\sum_j h_j v_j + \frac{m^2}{2} \sum_j C_j + \sum_j \Phi_j(v_j) \tag{8.3}$$

where $h_j = mC_j + f_j$. The end result of this approximation is that the nodes have been decoupled and so each node/spin experiences an effective field h_j. We can use this result to calculate the partition function for the system. As usual Tr refers to sum over all the degrees of freedom of the system, which for simplicity has been chosen to be $+1$ or -1 (See I.3)

$$Z_{MF} = \mathrm{Tr}e^{-\beta H_{MF}} = ... = e^{\frac{-\beta m^2}{2}\Sigma_j C_j} \prod_j \{e^{\beta[h_j - \Phi_j(1)]} + e^{\beta[-h_j - \Phi_j(-1)]}\} \tag{8.4}$$

Following the manipulations after I.3, this expression can be expressed as

$$Z_{MF} = ... = e^{\frac{-\beta m^2}{2}\Sigma_j C_j} \prod_j \{(a_j + b_j)\cosh\beta h_j + (a_j - b_j)\sinh\beta h_j\} \tag{8.5}$$

The partition function of this system can insights into the behavior of the system, including its possible **critical behavior** and phase shifts. First, we note that

$$\frac{\partial Z_{MF}}{\partial h_i} = ... = -\beta \mathrm{Tr}(v_i e^{-\beta H_{MF}}) \tag{8.6}$$

and(see I.6)

$$m = \frac{1}{N}\sum_{k=1}^{N} \langle v_k \rangle == ... = \frac{-1}{\beta N}\sum_{k=1}^{N} \frac{\partial}{\partial h_k} \ln Z_{MF} \tag{8.7}$$

Calculating this entity leads to a transcendental equation whose solutions can be illustrated by graphing. Solving for m results in (see I.7 and I.8)

$$m = ... = -1 + \frac{2}{N}\sum_{k=1}^{N} \frac{b_k}{a_k e^{2\beta(mC_k + f_k)} + b_k} \tag{8.8}$$

Unlike the Ising model or variations of it, the analysis of the behavior of the mean field Hopfield Networks is quite involved. There are a few contributing factors to this : Firstly, the fact that the original model includes T_{ij}, hence not an index independent entity as in the Ising model makes it difficult to analyze Hopfield networks. The ultimate consequence of this is the dependency of the final expression on indices and the difficulty posed by this to calculate a compact expression for the "magnetization" as in the Ising model. The second issue is the inclusion of the gain function term $\sum_j \Phi_j(v_j)$ in the **Hamiltonian**. In other words, despite the simplification provided by the approximation, the magnetization has a contribution from each node.

Suppose we did not include the gain function in the Hopfield network, as it's done in many cases. This is equivalent to setting $\Phi_j(v_j) = 0$ and consequently $a_k = b_k = 1$. The above expression would then simplify to

$$m = \frac{-1}{N}\sum_{k=1}^{N}[1 - \frac{1}{e^{2\beta h_k}+1}] = \frac{-1}{N}\sum_{k=1}^{N}\tanh\beta h_k \tag{8.9}$$

Again, the existence of this sum can be traced back to the indexed term T_{ij} in the Hopfield energy function. However, both in the simpler cases (Ising model) and here, it is easily understood by comparing the two sides of the transcendental equation that there will be three solutions if

$$\frac{d}{dm}\frac{-1}{N}\sum_{k=1}^{N}\tanh\beta h_k|_{m=0} > 1 \tag{8.10}$$

which gives (see I.9)

$$\frac{-1}{N}\sum_{k=1}^{N}\frac{\beta C_k}{\cosh^2(f_k)} > 1 \tag{8.11}$$

In the very special case $f_k = 0$, this is simplified to

$$\beta\frac{-1}{N}\sum_{k=1}^{N}C_k = \beta\frac{-1}{N}\sum_{kl}T_{lk} = \beta M > 1 \tag{8.12}$$

where M denotes the sum of all the matrix elements T_{ij}. In conclusion $\beta M = 1$ defines a **critical temperature** through $k_B T_c = -M/N$ below which there are the three solutions $m = 0$ and $m = \pm m_0$ and above which there is only one solution $m = 0$. Note that the inclusion of f_k in the above calculations forces each $\tanh(mC_k)$ to shift right or left with the amount f_k. Clearly, this poses a major obstacle to drawing conclusions about the existence or number of solutions or even the existence of a **critical temperature**.

8.1.1 Critical Behavior

Here, with the same assumptions as above, we attempt to calculate the critical exponents of the system. As usual, the **critical temperature** T_c was found at $m = 0$. So assuming that we are at the vicinity of this point, after a series expansion and a few other manipulations (See the calculations after I.11) :

$$m = ... = \frac{-1}{N}[\frac{m}{k_B T}\sum_{k=1}^{N} C_k - \frac{m^3}{3k_B^3 T_c^3}\sum_{k=1}^{N} C_k^3] = ... = m(\frac{T_c}{T}) - \frac{N^2 m^3}{3}(\frac{T_c}{T})^3 + ... \tag{8.13}$$

The solution to this equation is either $m = 0$ when $T \rightarrow T_c^+$ or $m = \pm\frac{1}{N}(3t)^{1/2}$ when $T \rightarrow T_c^-$. Here, $t = \frac{T_c - T}{T_c}$.

Let's return to the original case

$$m = -1 + \frac{2}{N}\sum_{k=1}^{N} \frac{b_k}{a_k e^{2\beta(mC_k + f_k)} + b_k}$$

and suppose that the contribution from f_k is small enough to keep the graph of the RHS function centered around $m = 0$. In order to find the **critical temperature**, it makes sense to do as we did previously and calculate the derivative of the above expression w.r.t m. Doing this, we get

$$1 = \frac{2}{N}\sum_{k=1}^{N} \frac{2\beta a_k b_k C_k e^{2\beta(mC_k + f_k)}}{(a_k e^{2\beta(mC_k + f_k)} + b_k)^2}|_{m=0} = \frac{2}{N}\sum_{k=1}^{N} \frac{2\beta a_k b_k C_k e^{2\beta f_k}}{(a_k e^{2\beta f_k} + b_k)^2} \tag{8.14}$$

which in the absence of f_k but presence of the gain function $\Phi_i(v_i)$, hence a_k and b_k simplifies to

$$1 = \frac{4}{N}\sum_{k=1}^{N} \frac{\beta a_k b_k C_k}{(a_k + b_k)^2} \tag{8.15}$$

Defining the **critical temperature** as

$$k_b T_c = \frac{4}{N}\sum_{k=1}^{N} \frac{a_k b_k C_k}{(a_k + b_k)^2} \tag{8.16}$$

we can understand the **critical behavior** of the system as follows. A simple Taylor expansion around $m = 0$ shows that we are not able to find an analytical solution beyond first order (see the calculations before I.18) :

$$m = ... = \frac{-1}{N}\sum_{k=1}^{N}\frac{a_k - b_k}{a_k + b_k}(1 - \frac{T_c}{T}) = \frac{-1}{N}\sum_{k=1}^{N}\frac{a_k - b_k}{a_k + b_k}t \text{ as } T \to T_c^- \tag{8.17}$$

Note that MFT does not necessarily reflect the actual behavior of the system. It is well-known that even for the simplest Ising model, MFT does not describe the system's behavior correctly and in all dimensions since fluctuations may practically be strong enough to bring into question the original idea of the spins/nodes organizing themselves around the mean field.

8.2 Project 2 : Using the Greedy Variational Principle to derive the Feynman propagators of Hopefield Networks

In this project, we explore how/if quantum field theory can be applied directly to Hopfield Networks. This line of work follows the ideas put forward in [48] where it is argued that replacing the usual canonical variables of classical mechanics, x and $\dot{x}$ by v and $\dot{v}$, where v is the output of a neural network, opens up the possibility of applying a path integral formulation of neural network that is then used to derive a wave function for neural network. This wave function is then modified to satisfy a *Schrödinger*-like wave equation, thereby opening up the field for application of quantum mechanical tools.

Based on the fundamental assumption that neural nets are dissipative systems, [48] argues that the **Lagrangian formalism** of least action must be derived not from Euler-Lagrange equations but from the so-called **greedy variational principle**. Then, by interpreting v and $\dot{v}$ as the position and velocity in analytical mechanics, they prove that the equations of motion in dissipative systems can be derived from

$$\frac{\partial_G S}{\partial_G v} = \frac{\partial L}{\partial \dot{v}} \tag{8.18}$$

The application of this principle on Hopfiled Networks with the energy function

$$E(v) = -\frac{1}{2}\sum_{ij} T_{ij} v_i v_j - \sum_i f_i v_i + \sum_i \Phi_i(v_i) \tag{8.19}$$

has consequences that are later used to derive the wave equation of neural nets. Note that here u_i is the output of the last layer of the network, g is a typical threshold (linear or non-linear), $v_i = g(u_i)$ and $\Phi' = g^{-1}$.

The derivation of the path integral formulation of NNs and as a result the wave function and the **Schrödinger's equation** for NNs, are quite involved[1]. It relies heavily on (**a**)

[1] The reader is encouraged to refer to [48] for further details

the interpretation of Feed-forward NNs as a series of discrete weighted sums

$$g\sum_{x_0}...g\sum_{x_{N-1}}\prod_{k=1}^{N-1}w_{x_kx_{k-1}}v^e_{x_1x_0} \tag{8.20}$$

where $w_{x_kx_{k-1}}$ are the synaptic weights connecting the neurons in layers x_k and x_{k-1}, g as described above and $v^e_{x_1x_0}$ is the output of the neuron transmitted from layer 0 to layer 1, and (**b**) the idea that "the neuronal activity models the frequency of the frequency of the actual output, a spiky waveform"[12], allowing us to formally replace e.g. the *k*th neuron's output in the model with

$$e^{iv_kt} \tag{8.21}$$

where v_k is the output frequency and t is the time it takes for the signal to travel along the axon. Equipped with these preliminary observations/assumptions and with additional introduction of two time scales $\Delta x/A$ and $(t_k-t_{k-1})/h$ where A and h are appropriate scaling factors, the transformed equation (8.20) can be interpreted as a collection of **Riemann sums** i.e. discrete versions of path-integrals of the form

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}e^{\frac{i}{h}S}g\frac{dv_1}{A}\cdots g\frac{dv_{N-1}}{A} \tag{8.22}$$

and S, defined by

$$S=\int_0^t L(\dot{v},v)dt \tag{8.23}$$

is interpreted as the action. The above path-integral is then used in a straight-forward application of the quantum mechanical framework to define the wave function $\psi(v,t)$ of the neural network which, after a great deal of manipulations, leads to the wave equation

$$\frac{\partial\psi}{\partial t}=\frac{ih}{2m}\frac{\partial^2\psi}{\partial v^2}+\frac{1}{m}\frac{\partial E}{\partial v}\frac{\partial\psi}{\partial v}-\frac{i}{h}V\psi \tag{8.24}$$

with E as in equation (8.19), $m=1/g'(g^{-1}(v)$ and an appropriate definition of V involving E, m etc [2].

[2] Again, for details and (quite unclear) derivations the reader is referred to the appendix in [48]

It is quite easily argued that the propagator from one point to another is QM is given by

$$< x_f, t_f; x_i, t_i >= \int D[x] e^{i \int_{t_i}^{t_f} dt L} \tag{8.25}$$

A common technical approach is to use an auxiliary field j to calculate the correlation functions in quantum mechanics. Writing

$$< x_f, t_f; x_i, t_i >_j= \int D[x] e^{i \int_{t_i}^{t_f} dt (L + j(t) x(t))} \tag{8.26}$$

the time ordered correlation functions can be be obtained through functional derivatives as follows

$$< x_f, t_f | T(x(t_1), ... x(t_n)) | x_i, t_i >_j= (-i)^n \frac{\delta^n}{\delta j(t_1) ... \delta j(t_n)} < x_f, t_f; x_i, t_i >_j |_{j(t)=0} \tag{8.27}$$

And defining the **generating functional**

$$Z[j] = \lim_{t_f, t_i \to \pm\infty} < x_f, t_f | x_i, t_i >_j \tag{8.28}$$

the master formula for correlation functions is given by

$$< 0 | T(x(t_1), ... x(t_n)) | 0 >_j= (-i)^n \frac{1}{Z[0]} \frac{\delta^n}{\delta j(t_1) ... \delta j(t_n)} Z[j] |_{j(t)=0} \tag{8.29}$$

So far so good. Assuming that Hopfield networks are correct description models of the brain function, we will now examine them in the current context. It is important to note that the construction of the **Lagrangian** from the **Hamiltonian** through Legendre transformation is not possible in this case simply because the **Hamiltonian** here is not written in terms of the canonical variables (generally called p_i and q_i). Thus it has to be constructed indirectly. Again, inspired by the results obtained in (quantum brain) where the authors identify the mass term in the Hopfield network wave function as $m = 1/g'(g^{-1}(v_i)$ (along with certain assumptions), where $v_i = g(u_i)$, $g^{-1}(v_i) = \Phi_i'(v_i)$

and $1/g'(u_i) = \partial g^{-1}(v_i)/\partial v_i$, we write the kinetic term in a more familiar form

$$\frac{1}{2}\sum_i \dot{u}_i^2 g'(u_i) = \frac{1}{2}\sum_i \frac{\dot{v}_i}{g'(u_i)}\frac{\dot{v}_i}{g'(u_i)} g'(u_i) = \frac{1}{2}\sum_i m\dot{v_i}^2 \tag{8.30}$$

This choice has been made based on the assumption that the **Lagrangian** is of the form

$$L = K + \frac{dE(v)}{dt} = K + \frac{\partial E(v)}{\partial v}\frac{dv}{dt} \tag{8.31}$$

where E refers to equation (8.19). As such the greedy extremization of the action "Greedy Least action principle"[48] leads to the equations of motion of the Hopfield network given by

$$\dot{u}_i = \sum_j T_{ij} v_j + f_i - \Phi'(v_i) \tag{8.32}$$

With the addition of the auxiliary fields j_i, as described above, the action is now given by (See II.1)

$$S^j = ... = \int dt \sum_{ij} \left[\delta_{ij}\frac{m\dot{v}_i\dot{v}_j}{2} - T_{ij}v_j\dot{v}_i - \delta_{ij}f_i\dot{v}_j + \delta_{ij}g^{-1}(v_i)\dot{v}_j + \delta_{ij}v_i j_j \right] \tag{8.33}$$

After a series of manipulations (See II.2-II.19) the inverse Fourier transform of the action reads

$$\begin{aligned} S^j = \int dt \Bigg[\sum_i \frac{m\dot{v'}_i^2}{2} - \sum_{ij} T_{ij} v'_i \dot{v'}_j + [G_0 \dot{v}'_j + G_0{}' v'_i \dot{v}'_j + ...] - \sum_i f_i \dot{v'}_i \Bigg] - \\ - \sum_{ij} \frac{\delta_{ij}(G_0 - f_j)}{m} \int dt' j_i(t') D_1(t - t') \\ - \sum_{ij} \frac{\delta_{ij}}{2m} \int dt \int dt' j_i(t) D_1(t - t') j_j(t') \\ - \sum_{ij} \frac{iT_{ij} - i\delta_{ij}G_0{}'}{m^2} \int dt \int dt' j_i(t) D_2(t - t') j_j(t') \end{aligned} \tag{8.34}$$

We observe that the action has been separated into a source independent part that is reminiscent of the action we started with and a part that contains all the dependence on the external sources. In its simplest form, as it is in the case of harmonic oscillator, where $D(t-t')$ are called the **Green's functions**, here defined by

$$D_1(t-t') = \int_{-\infty}^{\infty} \frac{dE}{2\pi} \frac{e^{-iE(t'-t)}}{E^2}, \; D_2(t-t') = \int_{-\infty}^{\infty} \frac{dE}{2\pi} \frac{e^{-iE(t'-t)}}{E^3} \tag{8.35}$$

Changing the variable $z = -E(t-t')$ we can rewrite the above integral and compute it in the complex plane. It turns out that one of the contour integrals diverges (See II.20-II.30).

The roots of these non-glamorous results can be traced back to the **Lagrangian** of the Hopfiled network. In order to generate interesting results akin to those in the case of, for instance, harmonic oscillator, the **Lagrangian** would need to include a kinetic type term $\sim \dot{v}^2$ as well as a term $\sim v^2$. It is in fact the interplay between these two terms that leads to interesting/manageable **Green's functions** can then be used for a path-integral formulation of the Hopfield networks, opening the possibility of investigating all kinds of phenomena such as phase transitions, which we set out to do originally. It should be noted that the truncated Taylor expansion in the above treatment is not to be blamed for this as it would not have lead to the desired result.

8.3 Project 3 : What went wrong ?

It is perhaps time to re-examine the formulation of the **Lagrangian** derived from (8.31) in [48] [3]. The problem is related to the fact that while it is relatively easy to determine and understand what **canonical variables** (often physically meaningful) are in **analytical mechanics**, it is much harder to handle what abstract variables and their derivatives mean and how the Lagrangian or Hamiltonian description of a system should be formulated.

Here, we consider the possibility the Lagrangian not being in a correct form and so we rewrite it (See Appendix III for details). With this new form of the Lagrangian, the action becomes

$$\begin{aligned} S^j &= \int dt\{\frac{1}{2}\sum_i \dot{u}_i^2 g'(u_i) - \frac{1}{2}\sum_{ij} T_{ij} v_i v_j - \sum_i f_i v_i + \sum_i \Phi_i'(v_i) v_i + \sum_i v_i j_i\} \\ &= \int dt \sum_{ij}\left[\frac{m\delta_{ij}}{2}\dot{v}_i\dot{v}_j - \frac{T_{ij}}{2} v_i v_j - \delta_{ij} f_i v_j + \delta_{ij} g^{-1}(v_i) v_j + \delta_{ij} v_i j_j\right] \end{aligned} \tag{8.36}$$

Following similar steps as before, including the truncated Taylor expansion of the term involving g^{-1}, we Fourier transform the terms in the action along with redefining some other quantities (See III.5-III.10 for details)

$$S^j = \sum_{ij}\int \frac{dE}{2\pi}\left[K\tilde{v}_i'(E)\tilde{v}_j'(-E) - \frac{\delta_{ij}}{4K}\tilde{j}_i(E)\tilde{j}_j(-E) + M\delta_{ij}\int dt[\tilde{v}_j'(E) - \frac{\delta_{ij}}{2K}\tilde{j}_i(E)]e^{-iEt}\right] \tag{8.37}$$

Before inverse Fourier transforming this action, we note that the last term is nothing but the shifted variable v_i. We touch upon this subject later when we discuss the ultimate goal of this analysis i.e. using the path integral formalism to calculate the correlation functions. There, we will see that the partition function will be written in a certain form with the requirement that the measure of the path integral is invariant under the transformation from v_i to v_i', as described above. For now, lets just note that the last term in the above action is exactly this transformation or shift of variables.

[3] Unfortunately, this paper has numerous mistakes and at times, also quite ambiguous arguments

The new action is given by

$$S^j = \int dt \left[\sum_i \frac{m\dot{v}'^2_i}{2} - \sum_{ij} \frac{T_{ij}}{2} v'_i v'{}_j + \sum_i [G_0 v'_i + G_0{}' v'_i v'_i + ...] - \sum_i f_i v'{}_i \right] -$$

$$- \sum_{ij} M\delta_{ij} \int dt \int dt' j_i(t) D(t'-t) - \frac{1}{2} \sum_{ij} \delta_{ij} \int dt \int dt' j_i(t) D(t'-t) j_j(t') \quad (8.38)$$

where

$$D(t-t') = ... = \frac{1}{2m\delta_{ij}} \int \frac{dE}{2\pi} \frac{1}{E^2-\omega^2} e^{-iE(t-t')} \quad (8.39)$$

and

$$\omega^2 = \frac{T_{ij} - 2G_0{}'\delta_{ij}}{m\delta_{ij}} \quad (8.40)$$

This is a much more interesting result, reminiscent of the harmonic oscillator but with the special characteristics of the Hopfield networks. Let's note that this **Green's function** satisfies

$$(\frac{\partial^2}{\partial t^2} + \omega^2) D(t-t') = -\delta(t-t') \quad (8.41)$$

which is also similar to what should be expected of **Green's functions**. Computing this integral, the new action reads (For the details see III.10-III.13)

$$S^j = \int dt \left[\sum_i \frac{m\dot{v}'^2_i}{2} - \sum_{ij} \frac{T_{ij}}{2} v'_i v'{}_j + \sum_i [G_0 v'_i + G_0{}' v'_i v'_i + ...] - \sum_i f_i v'{}_i \right] -$$

$$- \sum_{i=j} M \int dt \int dt' j_i(t) D(t'-t) - \frac{1}{2} \sum_{i=j} \int dt \int dt' j_i(t) D(t'-t) j_j(t') \quad (8.42)$$

with the **Green's function**

$$D(t-t') = ... = \frac{-i}{4m\omega'} e^{-i\omega'|t-t'|}, \quad \omega' = \sqrt{\frac{-2G'_0}{m}} \quad (8.43)$$

8.3.1 The path-integral formulation

Here we return to the the main goal of this analysis i.e. to build a path integral formulation for Hopfield networks. The ultimate goal is to use this formalism to investigate the fixed points of the RG transformation and all the knowledge that it entails in terms of universality classes and power-laws etc. Formally, this requires a **generating functional** which is then used to compute the correlation functions of the network. For instance, the two-point function is given by

$$\langle 0|T(v_i(t_1), v_i(t_2)|0\rangle = \frac{(-i)^2}{Z[0]} \frac{\delta^2}{\delta j_i(t_1)\delta j_i(t_2)} Z[j]\bigg|_{j=0} \tag{8.44}$$

The partition function can be written as

$$Z[j] = Z[0]e^{-\frac{i}{2}\sum_{i=j}\iint j_i(t)D(t'-t)j_j(t')}e^{-i\sum_i M\iint j_i(t)D(t'-t)} \tag{8.45}$$

where

$$Z[0] = \int D[v']e^{iS^j[v']} \tag{8.46}$$

Generally speaking, the implicit assumption made here is that the measure of the path integral is invariant under the transformation

$$v_i \to v_i' = v_i + \frac{\delta_{ij}}{2K} j_i \tag{8.47}$$

Failure of the measure of the path integral to satisfy the invariance under this shift is usually considered as a sign of anomaly i.e. a broken symmetry in the classical theory. In our case, the last part of the partition function above

$$e^{-i\sum_i M\iint j_i(t)D(t'-t)} \tag{8.48}$$

seems problematic. One could reason that as far as the computation of the **Green's functions** of any order is concerned, this term does not contribute. Even then, the measure of the path-integral does not seem invariant under this transformation and so there is a problem. We end this part of the investigation with the remark that this term is a remnant from the shift (8.47) of all the linear terms in v_i in the original action.

8.4 Project 4 : Quantum field theoretical approach

The QFT approach is fundamentally different in that it is built on the concept of fields and not traditional canonical variables. As we saw previously, the canonical variables x_i in quantum mechanics or statistical physics (in the current context replaced by the inputs v_i to Hopfield network) are treated as functions of time. In a quantum field theoretical context, the fields $\phi(x)$ are functions of every point in space-time. A reasonable approach seems to be to consider the inputs $v_i(t)$ to the network as field $V_i(x)$ where $x = (t, x)$ refers to space-time.

As such the **Lagrangian** and the action are defined as

$$
\begin{aligned}
L &= \int d^3x \mathscr{L}(\phi(x), \partial_\mu \phi(x)) \\
S &= \int dt d^3x \mathscr{L} = \int d^4x \mathscr{L}(\phi(x), \partial_\mu \phi(x))
\end{aligned} \tag{8.49}
$$

where $\mathscr{L}$ is the **Lagrangian** density. We will proceed to treat the fields as fundamental objects and construct a **Hamiltonian** density through a Legendre transformation of the **Lagrangian** density we have already used but with $v_i(t)$ replaced by the fields $V_i(x)$.

$$
\mathscr{L} = \sum_{ij} \left[\frac{m\delta_{ij}}{2} \dot{V}_i \dot{V}_j - \frac{T_{ij}}{2} V_i V_j - \delta_{ij} f_i V_j + \delta_{ij} g^{-1}(V_i) V_j \right] \tag{8.50}
$$

Defining

$$
\pi_i(x) = \frac{\partial \mathscr{L}}{\partial \dot{V}_i} = \sum_{ij} m\delta_{ij} \dot{V}_j \tag{8.51}
$$

it is straightforward to find the **Hamiltonian**, that is the central object in both quantum mechanics and quantum field theory. This not-withstanding, the main difference with the classical field theory is that both π_i and V_i are now considered as operators, hence the change to $\hat{\pi}_i$ and $\hat{V}_i$, acting on some eigenstates in a Hilbert space according to

$$
\hat{\pi}_i(\mathbf{x}, t)|\pi_i\rangle = \pi_i(\mathbf{x})|\pi_i\rangle, \quad \hat{V}_i(\mathbf{x}, t)|V_i\rangle = V_i(\mathbf{x})|V_i\rangle \tag{8.52}
$$

The **Hamiltonian**

$$\begin{aligned} H &= \int d^3x\Big[\pi_i(x)\partial_t V_i(x) - \mathscr{L}(V_i(x),\partial_\mu V_i(x)\Big] \\ &= \int d^3x \sum_{ij}\left[\frac{\delta_{ij}}{2m}\pi_i\pi_j + \frac{T_{ij}}{2}V_iV_j + \delta_{ij}f_iV_j - \delta_{ij}g^{-1}(V_i)V_j\right] \\ &= \int d^3x \sum_{ij}\left[\frac{\delta_{ij}}{2m}\hat{\pi}_i\hat{\pi}_j + V(\hat{V})\right] \end{aligned} \tag{8.53}$$

can then be used to derive the path integral formulation partition function by dividing up the time interval between the initial and final states $t_f - t_i = N\Delta t$ i.e. somewhat Riemannian approach, and finally arriving at the following in the limit $\Delta t \to 0$.

$$\langle 0|0\rangle = N\int D[V]e^{iS[V]}, \quad S[V] = \int d^4x\mathscr{L} \tag{8.54}$$

Note that apart from using the **Hamiltonian** and the operator formalism, the definition of the correlation functions remains intact. Expressed somewhat differently and with the normalization constant included, the 4-point function, for instance, is defined

$$G^{(4)}(x_1,...,x_4) = \frac{\int D[V]V(x_1)...V(x_4)D[V]e^{iS}}{\int D[V]e^{iS}} \tag{8.55}$$

where the indices have been omitted for clarity. And this is exactly the same as it was defined previously in terms of the functional derivatives of the source augmented partition function. We will derive the 2- and 4-point functions with the new formalism. We start with the action where we expand g^{-1} around zero and reorganize the terms(See IV.1-IV.2).

$$S = ... = \int d^4x \sum_{ij}\left[-\frac{1}{2}V_iO_tV_j + \delta_{ij}MV_j + \delta_{ij}\mathscr{O}(V^3)\right] \tag{8.56}$$

where

$$O_t = m\delta_{ij}\partial_t\partial^t + T_{ij} - 2\delta_{ij}G_0'$$

is an operator. Running through the machinery of QFT, we finally arrive at the **partition**

function (IV.3-IV.6).

$$Z[j] = ... = Z[0]e^{\frac{1}{2}\sum_i \int d^4xd^4y j_i(x)D(x-y)j_i(y)}e^{-\sum_i \int d^4xd^4yMD(x-y)j_i(y)} \tag{8.57}$$

8.4.1 Feynman Rules for Hopfield Networks

The question is whether the extra term above has any effect on the **Green's functions**. The relevance of this consist in the fact that **Green's functions** are considered the building blocks of measurable quantities so a change might result in a measurable quantity. We rewrite the relevant terms of the partition function and compute the **2-point function** (See Appendix IV for all the details)

$$\begin{aligned} G^{(2)}(x_1,x_2) &= \frac{(-i)^2}{Z[0]}\frac{\delta^2}{\delta j_i(x_1)\delta j_i(x_2)}Z[j]\bigg|_{j_i=0} = ... = \\ &= -\left[D(x_1-x_2)+M^2\int d^4xD(x-x_2)\int d^4yD(y-x_1)\right] \end{aligned} \tag{8.58}$$

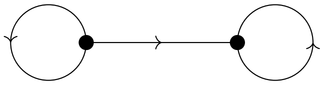
We are in a position to formulate Feynman-like rules :

(1) A propagator is represented by a solid straight line

(2) Solid dots represent a factor $-M$

(3) Loops are the integrals of propagators at the nodes

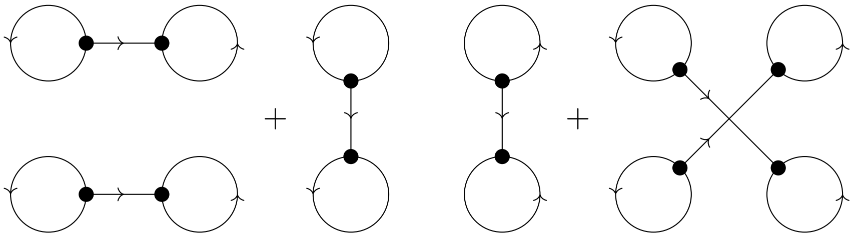
Using the above conventions, the two-point function can be diagrammatically shown as



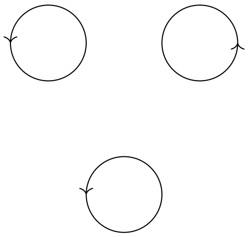
Note that $-M$ at each node is multiplied with the loop terms that originate from the second exponentional in the above action. Computing the 4-point function in a similar manner, we get

$$G^{(4)}(x_1,x_2,x_3,x_4) = \frac{(-i)^4}{Z[0]}\frac{\delta^4}{\delta j_i(x_1)\delta j_i(x_2)\delta j_i(x_3)\delta j_i(x_4)}Z[j]\bigg|_{j_i=0} = \ldots$$
$$= \bigg[D(x_1-x_2)D(x_3-x_4)+D(x_1-x_3)D(x_2-x_4)+D(x_1-x_4)D(x_2-x_3)$$
$$+M^4\int d^4xD(x-x_4)\int d^4yD(y-x_3)\int d^4zD(z-x_2)\int d^4wD(w-x_1)\bigg] \quad (8.59)$$

Following the Feynman-like rules above, the diagrammatic representation of the 4-point function can be seen below. We note that apart from **free**propagation of particles from one point to another, each node is accompanied by an $-M$ representing the second term in the above calculation.



We note that the 4-point function satisfies Wick's theorem in that in consists of products of **2-point functions** i.e. product of propagators with permuted positions. It should also be noted that oddly enough, the odd **Green's functions** have a somewhat unexpected contribution. In these cases, while the first part of the **Green's functions** i.e. the part satisfying Wick's theorem vanishes, the second part does not. As such the contribution from the **3-point Green's function** would be (with our conventions) represented by



8.4.2 Perturbation theory

It seems that what we saw above is just part of the story. A more powerful technique to deal with the correlation functions, particularly to have control over the essential parts of the calculations is to separate the **free theory** and the interaction terms. Then, by making the assumption that the interaction term is equipped with a parameter that is small, making the interactions weak enough to allow for a series expansion of the interaction exponential. Approaching the interactions in the manner, it is clear that this assumption results in suppressing the majority of the interactions to lower order terms.

Below we will investigate the consequences of this approach in the case of Hopfield networks. Recall that after reorganizing the terms in the action, we arrived at

$$S = i\int d^4x \sum_{ij}\left[\frac{m\delta_{ij}}{2}\partial_t V_i \partial V_j - \frac{T_{ij}}{2}V_i V_j + \delta_{ij}G_0' V_i V_j + \delta_{ij}(G_0 - f_i)V_j + \delta_{ij}\mathscr{O}(V^3)\right]$$

where the first three terms are reminiscent of the "**free Lagrangian**". Earlier, we also derived the Feynman propagator based on the **free Lagrangian** and used Wick's theorem to show that the correlation functions lead to combinations of propagators. While it is clear that the remaining $\mathscr{O}(V^3)$ terms can be thought of additional interaction terms, it is not clear how the linear terms in V should be treated. We note that when a source term was added to the action, the shift of the variable that otherwise leads to the separation of the shifted variable from the source, also created an extra term that $\int JD$ that is directly lined to the linear term in the above action. The implications of the appearance of this term are not clear.

However, if we are to treat all other terms outside the **free Lagrangian** as interaction terms, we can consider the linear term as well as all the other terms included in $\mathscr{O}(V^3)$ as perturbations of the **free Lagrangian**. This way, the correlation functions can be calculated up to any desired precision. Separating the interaction term in a general action, schematically, the partition function

$$Z[J] = N\int D[V]e^{i\int\{\mathscr{L}_0+\mathscr{L}_{int}+jV\}} \tag{8.60}$$

A quite straight-forward way of including interactions as perturbation is to implement it in the **generating functional**. Expanding the external source part of the above action, it is quite easy to realize that the **Green's functions** can be written as

$$\begin{aligned} G^{(n)}(x_1,...,x_n) &= \frac{1}{Z[0]}\int D[V]e^{i\int\{\mathscr{L}_0+\mathscr{L}_{int}\}}V(x_1)...V(x_n) \qquad (8.61)\\ &= \frac{1}{Z[0]}\int D[V]e^{i\int\mathscr{L}_0}V(x_1)...V(x_n)\left[1+i\int\mathscr{L}_{int}+\frac{i^2}{2!}\int\mathscr{L}_{int}\int\mathscr{L}_{int}+...\right] \end{aligned}$$

Note that in our case, the **free** and **interaction Lagrangians** are

$$\mathscr{L}_0 = -\sum_{ij}\frac{1}{2}V_iO_tV_j \tag{8.62}$$

$$\mathscr{L}_{int} = \sum_{ij}\delta_{ij}MV_j \tag{8.63}$$

Wick's theorem can then be used to represent the correlations functions and their perturbative corrections to any desired order. Let's assume that M is the parameter in the interaction **Lagrangian** that we referred to. The **2-point function** to the zeroth order in M is then given by the propagator

$$\frac{1}{Z[0]}\int D[V]e^{i\int d^4x\mathscr{L}_0}V(x_1)V(x_2) = D(x_1-x_2) \tag{8.64}$$

Again, higher order terms seem somewhat odd. We know very little **Measure Theory** to conclude that the measure of the path-integral does not survive the kind of variable shift that we saw/did multiple times. This does not in fact have to be the only culprit. The constructed Hopfield Lagrangigan, or the greedy variational principle or one of the many other assumptions made in [48] may have been incorrect.

Or perhaps we should have used a different **Lagrangian/Hamiltonian** in the path-integral ? We dealt with this issue by making a short U-turn to study/review **Statistical Field Theory**. As we will see, the problem lies in the perspective and an extra step in the analysis : The formalism is of course correct but it is in the formulation of the **Free Energy** as a **Hamiltonian** that the problem lies. It turns out that there is a second layer of difficulty at play.

We will see in the discussion of the models beyond the Ising model in section 9.1 that it is quite easy to understand the **free energy** in 9.4 in the case of the Ising model, where magnetization is a natural measure of the macroscopic behavior of the model. It is, however, very difficult to do this in the context of, for instance, Hopfield networks, as the level of abstraction here or in other cases, makes it difficult to replace the **free energy** with a measurable entity, particularly since this measurable entity must satisfy all or parts of the symmetries of the microscopic system as well as other system dependent requirements.

CHAPITRE 9

FROM QUANTUM FIELD THEORY TO STATISTICAL FIELD THEORY

To begin with, and without going further into details, there is a clear analogy between, the partition functions of Statistical Mechanics

$$Z = \sum_{\text{all states}} e^{-\beta E},$$

where E is the energy of the system, statistical field theory

$$Z = \int \mathscr{D} m \, e^{-\beta \int d^d x f[m]},$$

where m is the magnetization of e.g. a ferromagnet, and QFT

$$Z = \int \mathscr{D}[\phi] e^{-\frac{1}{\hbar} \int d^d x \mathscr{L}(\phi)}.$$

where ϕ is a scalar field.

Beyond this resemblance, it is well-known that studying **critical phenomena** requires an understanding of a physical system's long distance behavior. This is equivalent to understanding the role of fluctuations in a physical system. A typical example is the **Landau-Ginzburg model** of ferromagnetism where the scalar field ϕ plays the role of magnetization. A well-studied and simple action for this theory is the scalar ϕ^4 theory

$$S[\phi] = \int \{ \frac{1}{2}(\partial_\mu \phi)^2 + \frac{1}{2} m^2 \phi^2 + \frac{\lambda}{4!} \phi^4 \} + ... \tag{9.1}$$

Using MFT approximation (ignoring fluctuations by definition) is equivalent to considering a spatially uniform field configuration $\partial_\mu \phi = 0$ whose ground state is given by minimizing the action. This leads to the solution

$$\phi = 0 \quad \text{or} \quad \phi = \sqrt{\frac{-6m^2}{\lambda}}$$

In other words, while the first solution describes the paramgnetic **phase** the second solution is accessible for negative $m^2 \sim T - T_c$ i.e. below the critical temperature T_c. This type of transition from one **phase** to another is a so-called spontaneous **symmetry breaking** and does not correspond to the complexity of the actual **phase transition** process.

On the contrary, if the field in question is not restricted to its mean value, the RG evolution describing how the system behaves at different energy scales and how chaning of energy scale is manifested in the parameters (or coupling constants) of the theory is dedcued in the following manner : Let us remember that **coarse-graining** in statistical mechanics can be seen as a change of energy scale from higher to lower at which we want to examine the behavior of the system. At every step of **coarse-graining** the system moves from a microscopic (i.e. high energy) theory towards a macroscopic scale (i.e. lower energies). The functional formulation of RG does this by summing over large momenta of the Fourier transformed theory, hence resulting in a theory describing the system at lower energy scales.

It is about time to explain where we went wrong. It seems that approaching neural networks by treating its input as some kind of canonical variables on which the whole formalism of analytical mechanics can be applied may work. It is, however, at least not straight-forward to extend it beyond this point and to QFT. To begin with, we can question the validity of constructing a **Lagrangian** from static lattice models such as the Ising model (or even Hopfield network). The second issue could be related to the greedy variational approach which may or may not apply in the current situation as it simply ignores the **Euler-Lagrange equations** in the case of Hopfield networks[48].

But even if this approach was correct, what justifies a quantum field theoretical approach ? What is the motivation behind the treatment of input/out of neural networks as fields ? Is it fruitful to apply these methods to Ising model-like systems where no dynamical variables exist and where it is problematic to think of spins or input/output of networks as canonical variables ?

It turns out that many of these questions can be answered by Statistical Field theory which apart from postdating QFT has inherited its methods along with providing a philosophic motivation for what and how the notion of fields can be used in the treatment the Ising model and other systems alike, including Hopfield networks, (R)Boltzmann Machines etc. The two fundamental ingredients of this approach include parallels to the path integral formulation of QFT which in turn provides a reasonable explanation to what should be considered a field in these systems.

It turns out that the answer to the above questions is related to two fundamental facts about nature which consequently also govern much of the model physics' understanding of the world we live in. The world, as we know it, is organized around **Scale** and **Symmetry**. These two facts explain all from **phase transitions** and universality to how (or if) the microscopic world is manifested in what we observe.

It turns out that the partition function of a statistical system cannot be calculated unless in very restricted cases in one and two dimensions beyond which no exact solutions exist. And even if this were possible, we would be only be able to describe the equilibrium states. One way to circumvent this obstacle is **coarse-graining** which also establishes a link between the scales of observation. A lattice can be partitioned into smaller parts, all with its own magnetization (roughly average spin). And so the partition function for a statistical system spin configuration can be calculated (here we also give its counterpart in continuum)

$$Z = \sum_m e^{-\beta F(m)} \to Z = \int dm e^{-\beta N f(m)} \text{ with } F(m) = N f(m) \tag{9.2}$$

The problem is still not solved as we now face the calculation of the **free energy** density. Based on Mean Field approximation, Landau explained a great deal of what is observed in terms of **phase transitions** and universality based on what is known as an **order**

parameter (average magnetization is the simplest case of an order parameter). A reasonable generalization of this approach is the **Ginzburg-Landau Theory** where they moved beyond average magnetization, allowing the magnetization to vary across space, hence promoting it to a **field** $m(x)$ which is then dubbed local order parameter.

This is closely related to the process of coarse graining where lattices are partitioned into cells each of which can be attributed a magnetization. The partition function is now written as

$$Z = \sum_{m(x)} e^{-\beta F[m(x)]} \tag{9.3}$$

which can be written as a path integral if we assume that the magnetization various continuously in space

$$Z = \int D[m] e^{-\beta F[m(x)]} \tag{9.4}$$

But what does the **Landau-Ginzburg free energy** look like and how is it calculated ? It turns out that the **free energy** density can be constructed based on certain requirements that in turn originate from the microscopic lattice we started with. For instance, since the lattice is invariant under translation and rotations, albeit discrete, we could require the same type of invariance from the **free energy**. And since switching the direction of the spins and/or the direction of the external field (we don't go into details), is inherited by the magnetization during the process of **coarse-graining**, the **free energy** should be invariant under Z_2 **symmetry**. Furthermore, since the spins in the underlying Ising model interact locally, we can require the **free energy** to satisfy locality. A final, simplifying assumption is to require the **free energy** density to be an analytic function of magnetization and that it varies slowly across space i.e is mainly dependent on the fields and their gradient.

However, even with these considerations, the space of all possible theories is infinite. The crucial step beyond this point is based on the idea of **universality**. Simply put, since different systems regardless of their perceived complexity on microscopic level behave similarly near the critical point, it is reasonable to search for the simplest of models that give correct predictions instead of dealing with the complex interactions and specific details of every system.

The **Landau-Ginzburg free energy** can be constructed following these observations and requirements. For example, under specific conditions (such as $B = 0$ in the original Ising model) the **symmetry** requirements inherited from the coarse-grained Ising model lattice imply that the **free energy** must include even powers of the magnetization field and its gradients

$$F[\phi(x)] = \int d^d x \Big[\frac{1}{2}\alpha_2 \phi^2 + \frac{1}{4}\alpha_4 \phi^4 + \frac{1}{2}\gamma(\nabla\phi)^2 + ...\Big] \tag{9.5}$$

where we have used ϕ instead of m to signify the field theoretical approach. We should note that the couplings in front of the field configurations are temperature dependent.

We should note that the entire apparatus of field theory relies on perturbation methods and the assumption that higher order terms do not contribute as much as lower order terms. It is very possible that this is not true in some real life cases at which point numerical methods would have to be employed. Assuming that we are still in the perturbative mode, it is common to work with quadratic terms, treating higher order terms as perturbations.

An interesting outcome of this, which is related to what we did previously when we tried to treat the entire Hopfield **Lagrangian** in the context of field theory, is that there will be no linear terms in the fields if we expand the field around one of the saddle point minima of the **free energy** in low temperature i.e. $\pm m$. The absence of linear terms is related to the fact that the field will obey the **Euler-Lagrange equations** of motion. In the work that we did previously, there linear terms were persistent, which might be related to the **greedy variational principle** instead of **Euler-Lagrange equations**. This needs further investigation.

The path integral in the partition function (9.4) is not necessarily easy to calculate even with this generic but simple **free energy**. It can still be done and many interesting quantities, there-among correlation functions, can be extracted from it using the machinery of field theory. In conclusion, comparing the partition functions of statistical field theory and QFT

$$Z = \int D[\phi] e^{-\beta \int d^d x \mathscr{F}(\phi)}, \qquad Z = \int D[\phi] e^{\frac{i}{\hbar} \int d^d x \mathscr{L}(\phi)} \tag{9.6}$$

reveals obvious similarities but also differences. For instance, while the partition function of statistical field theory does not involve a time coordinate, the integral in its counterpart is over space-time. This is dealt with through a Wick rotation of the quantum field theoretical action into its Euclidean version by setting $\tau = it$. This way the Euclidean action will be analogous to the **free energy** in statistical field theory.

As noted above, in its simplest form the **Landau-Ginzburg theory** does not include higher order interaction terms and so adding terms like ϕ^4 to the above action serves the purpose of including fluctuations particularly as they become more important near the critical point. This is of course done only if perturbation theory is meaningful.

9.1 Statistical Field Theory beyond the Ising Model

We have concluded that it is philosophically and practically problematic to directly translate the input/output of a neural network into fields. If this analysis is correct, the Ising model or Hopfield network **Hamiltonians** are used only indirectly in the investigation of statistical physical phenomena. A couple of remarks are in place here : Seeking a similar approach to Hopfield networks or other energy-based models requires a deeper understanding of the meaningfulness of transferring abstract canonical variables, whatever they may be, to the realm of space-time (QFT) or space (Statistical Field Theory). More importantly, it is not entirely clear what constitutes **coarse-graining** in abstract variables since it is at the heart of the leap from the order parameter as the variable m (as in Landau theory and the Mean Field approach) to order parameter as a functional in Statistical Field Theory. This being said, it is also unclear what should be considered an order parameter in abstract settings. Yes, we can study the equivalent of magnetization in every system but even if it can be viewed as a scale-related representation of the collective behavior of a microscopic system, it is not necessarily clear what it means in other networks.

We saw that the form and content of the quantum field theoretical **Landau-Ginzburg Lagrangian** is deduced from the symmetries of the original model. Clearly, without satisfying the analyticity condition, at least in the vicinity of critical points, it is very difficult to compute anything. This means that the **free energy** can only include positive powers of the fields and its gradients (no $1/\phi$ or $1/\phi^2$ etc terms). Secondly, the symmetries of the Hopefiled networks must be inherited by the order parameter and so the **free energy** has to be constrained to include or exclude certain terms. If we are working with the simplest Hopfield networks where the nodes take on ± 1 values, we can enforce a Z_2 **symmetry** on the **free energy** (inverting the pixels of black and white pictures should not change anything). This restricts the **Lagrangian** to include only even powers of the fields or their gradients.

Finally, considering the space of all possible **Lagrangians** in statistical field theory, universality implies that it suffices to choose the simplest possible model. The remaining work is to run through the machinery of QFT to compute the correlation coefficients, scattering amplitudes etc.

9.1.1 Group Theoretical considerations

Besides stating the obvious fact that different **phases** of matter are associated with different symmetries, using **symmetry** to characterize the **phases of matter** leads to conclusions that seem to be conflicting with how they are perceived. **symmetry** places seemingly wildly different types of phenomena in the same so called universality class. The picture is, however, somewhat complicated in that it is not one but two **symmetry groups** that are at play here : The **free energy** as a measure of the behavior of the order parameter, and the ground state of a system can be invariant under same, different or no **symmetry transformations**.

The most general **Landau-Ginzburg free energy** is written based on the order parameter, the choice of which may or many not be obvious, and the **symmetry group** G, under which it transforms. We then require this generic **free energy** to be invariant under the very same **symmetry group** G. As the system undergoes changes, the ground state of the system may be invariant under a different **symmetry group** H. In general, we say that there is a **phase transition** when H changes.

Using the Ising model as a prototype, we know that when $B = 0$, the **free energy symmetry group** is $G = Z_2$. It turns out this **symmetry** persists also above the critical temperature T_c. However, at temperatures below T_c, the **symmetry** of the ground state breaks spontaneously $H = \emptyset$ and so the two **phases** are characterized by the two choices in H, equivalent to m or $-m$. On the other hand, when $B \neq 0$, the **free energy** is no longer invariant under Z_2 i.e. $G = \emptyset$, which is understood as the existence of only one **phase**. In this case it is possible to move across the **phase space** without undergoing a **phase transition**. A (first order) **phase transition** does occur in this case when B is changed at low temperature although according to this classification, there are no two **phases**.

9.1.2 Conformal field theory

We know from the analysis of RG transformations that some operators simply disappear as we move across scales and that many operators of high energy regime might have disappeared in the process of zooming out. It is crucial to understand that the **fixed points** of RG are identified with the critical points of statistical physics. The scale invariance at the **fixed points** (by definition) of RG means that the actual form of the **Lagrangian** i.e. what operators it includes or excludes do not matter.

It turns out that invariance under re-scaling is not the only **symmetry** at the **fixed points** or RG and that the physics is invariant also under larger class of symmetries known as the **conformal symmetry group**. Following the philosophy of **symmetry** and scale as the two major governing principles of modern physics, it is the clear that all the fields and correlation function are restricted by **conformal group**.

CHAPITRE 10

CONCLUSION AND OUTLOOK

The goal of this work was to move beyond empirical findings of recent years and provide a theoretical understanding of phase transitions in various neural networks regardless of their description as energy-based models. This is at least somewhat reasonable because of the parallels we can draw between real world processes in space-time and the empirically but not theoretically explained processes in machine learning that demonstrate remarkable similarities with physical processes.

Although late, too late to say the least, we have gained a reasonable understanding of what mistakes we made and what steps we need to take in order to remedy this situation or perhaps even achieve what we set out to do. This will only be possible if we can find a neural network equivalent of what is called an **order parameter** in statistical field theory. As we saw above, a description of ferromagnetism was made possible by choosing the **free energy** as the order parameter. What can possibly be a good candidate in our case ?

Another crucial factor at play is the study of the **symmetries** of the system and how they should be incorporated in the order parameter, whatever it may be. What are the symmetries of a given energy model ? What obvious or hidden symmetries might they have ? As we explained above, different phases are defined by the **symmetries of the Hamiltonian at different scales**. If the current approach is correct, it should be possible to explain the **empirically observed phase transitions in NNs** based on how the symmetries change (continuously or discontinuously) in the process of **RG transformations**.

Both quantum and statistical field theories play out against the background of space-time. If we are to treat neural networks in the context of these theories, it seems unreasonable to look for abstractions of RG transformations. Even in the context of space-time, it remains to be seen to what extent it makes sense to involve quantum processes in neural network. Related to this, we need to understand whether it is reasonable to demand Lorentz invariance from a theory whose goal answer questions on the inner workings or performance of neural networks.

It seems plausible that the current approach, which hinges upon the construction of an order parameter specifically suited to the symmetries of the Hopfield Networks and/or other energy-based models, will answer many questions about the behavior of these theories, their performance and scaling laws etc. We are curious how we can move beyond Hopfield Networks as primitve RNNs to actual RNNs.

Finally, it seems that conformal field theory may be of help in answering questions about universality and universal constants and so it deserves to be studied particularly thoroughly, if quantum field theory proves to be the correct forum for the treatment of neural networks.

BIBLIOGRAPHIE

[1] Stuart Hameroff 1998. Quantum computation in brain microtubules ? the penrose–hameroff 'orch or' model of consciousness. *Phil. Trans. R. Soc. A.3561869–1896*, 1998. URL http://doi.org/10.1098/rsta.1998.0254.

[2] A. Montanari A. Javanmard. Information and inference. *A Journal of IMA*, 2 :115, 2013.

[3] Madhu Advani, Subhaneil Lahiri, and Surya Ganguli. Statistical mechanics of complex neural systems and high dimensional data. *Journal of Statistical Mechanics : Theory and Experiment*, 2013(03) :P03014, mar 2013. doi : 10.1088/1742-5468/2013/03/p03014. URL https://doi.org/10.1088%2F1742-5468%2F2013%2F03%2Fp03014.

[4] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S. Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11(1) :501–528, 2020. doi : 10.1146/annurev-conmatphys-031119-050745. URL https://doi.org/10.1146/annurev-conmatphys-031119-050745.

[5] E. Bolthausen. An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Commun. Math. Phys.*, 325, 2014. doi : https://doi.org/10.1007/s00220-013-1862-3.

[6] Philip G Breen, Christopher N Foley, Tjarda Boekholt, and Simon Portegies Zwart. Newton versus the machine : solving the chaotic three-body problem using deep neural networks. *Monthly Notices of the Royal Astronomical Society*, 494(2) : 2465–2470, apr 2020. doi : 10.1093/mnras/staa713. URL https://doi.org/10.1093/mnras/staa713.

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Pra-

fulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[8] Cédric Bény. Deep learning and the renormalization group, 2013.

[9] Cédric Bény. Deep learning and the renormalization group, 2013. URL `https://arxiv.org/abs/1301.3124`.

[10] Jonathan Carifio, James Halverson, Dmitri Krioukov, and Brent D. Nelson. Machine learning in the string landscape. *Journal of High Energy Physics*, 2017 (9), sep 2017. doi : 10.1007/jhep09(2017)157. URL `https://doi.org/10.1007/jhep09/282017/29157`.

[11] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová . Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4), dec 2019. doi : 10.1103/revmodphys.91.045002. URL `https://doi.org/10.1103%2Frevmodphys.91.045002`.

[12] Barnes N. S. Curtis H. *Biology*. Worth, 1989.

[13] Ellen De Mello Koch, Robert De Mello Koch, and Ling Cheng. Is deep learning a renormalization group flow ? *IEEE Access*, 8 :106487–106505, 2020. ISSN 2169-3536. doi : 10.1109/access.2020.3000901. URL `http://dx.doi.org/10.1109/ACCESS.2020.3000901`.

[14] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová . Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6), dec 2011. doi : 10.1103/

physreve.84.066106. URL https://doi.org/10.1103%2Fphysreve.84.066106.

[15] Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2) :288–299, May 2017. ISSN 1572-9613. doi : 10.1007/s10955-017-1806-y. URL http://dx.doi.org/10.1007/s10955-017-1806-y.

[16] R. Dengler. Critical phenomena of single and double polymer strands in a solution, 2020. URL https://arxiv.org/abs/2002.03942.

[17] B. Derrida E. Gardner. Optimal storage properties of neural network models. *J. Phys.*, A21 :271–284, 1988.

[18] Andreas Engel and Chris van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001.

[19] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5) :75–174, feb 2010. doi : 10.1016/j.physrep.2009.11.002. URL https://doi.org/10.1016%2Fj.physrep.2009.11.002.

[20] Marylou Gabrie, Eric W Tramel, and Florent Krzakala. Training restricted boltzmann machine via the thouless-anderson-palmer free energy. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 640–648. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5788-training-restricted-boltzmann-machine-via-the-thouless-ander pdf.

[21] Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience, 2015. URL https://arxiv.org/abs/1503.08779.

[22] N. GOLDENFELD. *Lectures on Phase Transitions and the Renormalization Group*. CRC Press, 1992.

[23] David Gross, Yi-Kai Liu, Steven T. Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical Review Letters*, 105 (15), oct 2010. doi : 10.1103/physrevlett.105.150401. URL `https://doi.org/10.1103/physrevlett.105.150401`.

[24] Stuart Hameroff and Roger Penrose. Consciousness in the universe : a review of the 'orch or' theory. *Physics of life reviews*, 11, 08 2013. doi : 10.1016/j.plrev.2013.08.002.

[25] Koji Hashimoto, Sotaro Sugishita, Akinori Tanaka, and Akio Tomiya. Deep learning and the mml :math xmlns :mml="http ://www.w3.org/1998/math/MathML" display="inline"mml :mrowmml :miAdS/mml :mimml :mo//mml :momml :miCFT/mml :mi/mml :
correspondence. *Physical Review D*, 98(4), aug 2018. doi : 10.1103/physrevd.98.046019. URL `https://doi.org/10.1103/physrevd.98.046019`.

[26] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8) :1771–1800, 2002. doi : 10.1162/089976602760128018.

[27] J. J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Nat. Acad. Sci. U.S.A.*, 81(10) :3088, 1984.

[28] JJ Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79 :2554–8, 05 1982. doi : 10.1073/pnas.79.8.2554.

[29] Satoshi Iso, Shotaro Shiba, and Sumito Yokoo. Scale-invariant feature extraction of neural network and renormalization group flow. *Physical Review E*, 97 (5), may 2018. doi : 10.1103/physreve.97.053304. URL `https://doi.org/10.1103%2Fphysreve.97.053304`.

[30] Raban Iten, Tony Metger, Henrik Wilming, Lí dia del Rio, and Renato Renner. Discovering physical concepts with neural networks. *Physical Review Letters*, 124 (1), jan 2020. doi : 10.1103/physrevlett.124.010508. URL `https://doi.org/10.1103%2Fphysrevlett.124.010508`.

[31] Gregg Jaeger. The ehrenfest classification of phase transitions : Introduction and evolution. *Archive for History of Exact Sciences*, 53 :51–81, 05 1998. doi : 10.1007/s004070050021.

[32] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106 :620–630, May 1957. doi : 10.1103/PhysRev.106.620. URL `https://link.aps.org/doi/10.1103/PhysRev.106.620`.

[33] Jason K. Johnson, Dmitry M. Malioutov, and Alan S. Willsky. Lagrangian relaxation for map estimation in graphical models, 2007. URL `https://arxiv.org/abs/0710.0013`.

[34] Yoshiyuki Kabashima, Florent Krzakala, Marc Mezard, Ayaka Sakata, and Lenka Zdeborova. Phase transitions and sample complexity in bayes-optimal matrix factorization. *IEEE Transactions on Information Theory*, 62(7) :4228–4265, jul 2016. doi : 10.1109/tit.2016.2556702. URL `https://doi.org/10.1109%2Ftit.2016.2556702`.

[35] L. P. Kadanoff. *Statics, Dynamics and Renormalization*. World Scientific, 2000.

[36] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[37] Matias Carrasco Kind and Robert J. Brunner. TPZ : photometric redshift PDFs and ancillary information by using prediction trees and random forests. *Monthly Notices of the Royal Astronomical Society*, 432(2) :1483–1501, may 2013. doi : 10.1093/mnras/stt574. URL `https://doi.org/10.1093/mnras/stt574`.

[38] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4) :307–392, 2019. doi : 10.1561/2200000056. URL `https://doi.org/10.1561%2F2200000056`.

[39] Maciej Koch-Janusz and Zohar Ringel. Mutual information, neural networks and the renormalization group. *Nature Physics*, 14(6) :578–582, Mar 2018. ISSN 1745-2481. doi : 10.1038/s41567-018-0081-4. URL `http://dx.doi.org/10.1038/s41567-018-0081-4`.

[40] Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition, 2016.

[41] Nicolas Le Roux and Yoshua Bengio. Representational power of restricted boltzmann machines and deep belief networks. *Neural Comput.*, 20(6) :1631–1649, jun 2008. ISSN 0899-7667. doi : 10.1162/neco.2008.04-07-510. URL `https://doi.org/10.1162/neco.2008.04-07-510`.

[42] Y. LeCun. A tutorial on energy-based learning, 2006. URL `http://www.cs.toronto.edu/~vnair/ciar/lecun1.pdf`.

[43] Patrick M. Lenggenhager, Doruk Efe Gökmen, Zohar Ringel, Sebastian D. Huber, and Maciej Koch-Janusz. Optimal renormalization group transformation from information theory. *Phys. Rev. X*, 10 :011037, Feb 2020. doi : 10.1103/PhysRevX.10.011037. URL `https://link.aps.org/doi/10.1103/PhysRevX.10.011037`.

[44] Henry Lin and Max Tegmark. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7) :299, jun 2017. doi : 10.3390/e19070299. URL `https://doi.org/10.3390%2Fe19070299`.

[45] Henry W. Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well ? *Journal of Statistical Physics*, 168(6) :1223–1247, jul 2017. doi : 10.1007/s10955-017-1836-5. URL `https://doi.org/10.1007%2Fs10955-017-1836-5`.

[46] Pankaj Mehta and David J. Schwab. An exact mapping between the variational renormalization group and deep learning, 2014.

[47] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, 810 :1–124, may 2019. doi : 10.1016/j.physrep.2019.03.001. URL `https://doi.org/10.1016%2Fj.physrep.2019.03.001`.

[48] W. L. Mirnaker. A neural network wave formalism. *Advances in Applied Mathematics*, 37 :19–31, 07 2005.

[49] Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford U. Press, New York, 2009, 2009. ISBN ISBN 978-0-19-857083-7.

[50] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning, 2017.

[51] Long Gang Pang, Kai Zhou, Nan Su, Hannah Petersen, Horst Stöcker, and Xin Nian Wang. Classify qcd phase transition with deep learning. *Nuclear Physics. A*, 982 (C), 1 2019. doi : 10.1016/j.nuclphysa.2018.10.077.

[52] R. Penrose. *The Emperor's New Mind*. Oxford University Press, 1989.

[53] R. Penrose. *Shadows of Mind*. Oxford University Press, 1994.

[54] R. Penrose. On understanding understanding. *Gen. Relativ. Grav. 581-600*, 28, 1996.

[55] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need, 2021.

[56] T. Sejnowski S. Saremi. Hierarchical model of natrual images and the origin of scale invariance, 2013. URL https://www.pnas.org/doi/pdf/10.1073/pnas.1222618110.

[57] Gordon Slade. Self-avoiding walk, spin systems and renormalization. *Proceedings of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 475 (2221) :20180549, jan 2019. doi : 10.1098/rspa.2018.0549. URL https://doi.org/10.1098%2Frspa.2018.0549.

[58] H. S. Seung Sompolinsky H., N. Tishby. Learning from examples in large neural networks. *Phys. Rev. Lettt.*, 65 :1683, 1990.

[59] E. Miles Stoudenmire and David J. Schwab. Supervised learning with quantum-inspired tensor networks, 2016. URL https://arxiv.org/abs/1605.05775.

[60] Richard Sutton. The bitter lesson, 2019. URL http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

[61] R.G. Palmer Thouless D. J., P.W.Anderson. Information and inference. *Philosophical Magazine*, 35 :593, 1977.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[63] Wikipedia. Kepler's laws of planetary motion, 2021.

[64] Wikipedia. Double-slit experiment, 2022.

[65] Wikipedia contributors. The ising model, 2022. URL https://en.wikipedia.org/wiki/Ising_model. [Online ; accessed 9-December-2022].

[66] A. L. Yuille and Anand Rangarajan. The concave-convex procedure. *Neural Comput.*, 15(4) :915–936, apr 2003. ISSN 0899-7667. doi : 10.1162/08997660360581958. URL `https://doi.org/10.1162/08997660360581958`.

Annexe I

Technical details of project 1 : Studying phase transitions in Hopfield Networks from the perspective of Mean field theory

The mean field Hamiltonian

$$\begin{aligned}
H_{MF}(v) &= -\frac{1}{2}\sum_{ij} T_{ij} v_i v_j - \sum_i f_i v_i + \sum_i \Phi_i(v_i) \qquad \text{(I.1)}\\
&= -\frac{1}{2}\sum_{ij} T_{ij}[\langle v_i\rangle v_j + \langle v_j\rangle v_i - \langle v_i\rangle\langle v_j\rangle] - \sum_i f_i v_i + \sum_i \Phi_i(v_i)\\
&= -m\sum_{ij} T_{ij} v_j + \frac{m^2}{2}\sum_{ij} T_{ij} - \sum_i f_i v_i + \sum_i \Phi_i(v_i)\\
&= -\sum_j [m\sum_i T_{ij} + f_j] v_j + \frac{m^2}{2}\sum_{ij} T_{ij} + \sum_i \Phi_i(v_i)
\end{aligned}$$

The mean field Hamiltonian can then be written as

$$\begin{aligned}
H_{MF}(v) &= -\sum_j [mC_j + f_j] v_j + \sum_j [\frac{m^2}{2} C_j + \Phi_j(v_j)]\\
&= -\sum_j h_j v_j + \frac{m^2}{2}\sum_j C_j + \sum_j \Phi_j(v_j) \qquad \text{(I.2)}
\end{aligned}$$

where $h_j = mC_j + f_j$. The mean field partition function

$$\begin{aligned}
Z_{MF} &= \text{Tr}e^{-\beta H_{MF}} \qquad \text{(I.3)}\\
&= \text{Tr}e^{\beta\sum_j h_j v_j - \beta\sum_j [\frac{m^2}{2} C_j + \Phi_j(v_j)]}\\
&= e^{\frac{-\beta m^2}{2}\sum_j C_j}\text{Tr}e^{\beta\sum_j [h_j v_j - \Phi_j(v_j)]}\\
&= e^{\frac{-\beta m^2}{2}\sum_j C_j}\text{Tr}\{e^{\beta[h_1 v_1 - \Phi_1(v_1)]} e^{\beta[h_2 v_2 - \Phi_2(v_2)]} ... e^{\beta[h_N v_N - \Phi_N(v_N)]}\}\\
&= e^{\frac{-\beta m^2}{2}\sum_j C_j}\prod_j\{e^{\beta[h_j - \Phi_j(1)]} + e^{\beta[-h_j - \Phi_j(-1)]}\}
\end{aligned}$$

This expression can be rewritten using the following fact :

$$\begin{aligned} a_1e^x+b_1e^{-x} &= a_1e^x+a_1e^{-x}-a_1e^{-x}+b_1e^{-x} \\ &= 2a_1\cosh x+(b_1-a_1)e^{-x} \\ &= 2a_1\cosh x+(b_1-a_1)(\cosh x-\sinh x) \\ &= (a_1+b_1)\cosh x+(a_1-b_1)\sinh x \end{aligned}$$

We define $a_j = e^{-\beta\Phi_j(1)}$ and $b_j = e^{-\beta\Phi_j(-1)}$, we get

$$\begin{aligned} Z_{MF} &= e^{\frac{-\beta m^2}{2}\sum_j C_j}\prod_j\{a_je^{\beta h_j}+b_je^{-\beta h_j}\} \\ &= e^{\frac{-\beta m^2}{2}\sum_j C_j}\prod_j\{(a_j+b_j)\cosh\beta h_j+(a_j-b_j)\sinh\beta h_j\} \end{aligned} \tag{I.4}$$

We note that

$$\frac{\partial Z_{MF}}{\partial h_i} = ... = -\beta\mathrm{Tr}(v_ie^{-\beta H_{MF}}) \tag{I.5}$$

and

$$m = \frac{1}{N}\sum_{k=1}^{N}\langle v_k\rangle = \frac{1}{N}\sum_{k=1}^{N}\frac{\mathrm{Tr}(v_ke^{-\beta H_{MF}})}{Z_{MF}} = \frac{-1}{\beta N}\sum_{k=1}^{N}\frac{\partial}{\partial h_k}\ln Z_{MF} \tag{I.6}$$

Furthermore

$$\begin{aligned} \ln Z_{MF} &= \frac{-\beta m^2}{2}\sum_j C_j+\sum_{i=1}^{N}\ln[(a_i+b_i)\cosh\beta h_i+(a_i-b_i)\sinh\beta h_i] \\ \frac{\partial}{\partial h_k}\ln Z_{MF} &= \beta\frac{(a_k+b_k)\sinh\beta h_k+(a_k-b_k)\cosh\beta h_k}{(a_k+b_k)\cosh\beta h_k+(a_k-b_k)\sinh\beta h_k} \end{aligned} \tag{I.7}$$

And we get

$$\begin{aligned} m &= \frac{-1}{N}\sum_{k=1}^{N}\frac{(a_k+b_k)\sinh\beta h_k+(a_k-b_k)\cosh\beta h_k}{(a_k+b_k)\cosh\beta h_k+(a_k-b_k)\sinh\beta h_k} \\ &= \frac{-1}{N}\sum_{k=1}^{N}[1-\frac{2b_k}{a_ke^{2\beta h_k}+b_k}] = -1+\frac{2}{N}\sum_{k=1}^{N}\frac{b_k}{a_ke^{2\beta(mC_k+f_k)}+b_k} \end{aligned} \tag{I.8}$$

$$\begin{aligned} \frac{-1}{N}\frac{d}{dm}\sum_{k=1}^{N}\tanh\beta(mC_k+f_k)|_{m=0} &= \frac{-1}{N}\sum_{k=1}^{N}\frac{\beta C_k}{\cosh^2(mC_k+f_k)}|_{m=0} \\ &= \frac{-1}{N}\sum_{k=1}^{N}\frac{\beta C_k}{\cosh^2(f_k)} > 1 \end{aligned} \tag{I.9}$$

In the very special case $f_k = 0$, this is simplified to

$$\beta\frac{-1}{N}\sum_{k=1}^{N}C_k = \beta\frac{-1}{N}\sum_{kl}T_{lk} = \beta M > 1 \tag{I.10}$$

I.0.1 Critical Behavior

Series expansion gives :

$$\begin{aligned} m = \frac{-1}{N}\sum_{k=1}^{N}\tanh\beta h_k &= \frac{-1}{N}\sum_{k=1}^{N}[\beta mC_k - \frac{\beta^3 m^3}{3}C_k^3] \\ &= \frac{-1}{N}[\frac{m}{k_BT}\sum_{k=1}^{N}C_k - \frac{m^3}{3k_B^3T_c^3}\sum_{k=1}^{N}C_k^3] \end{aligned} \tag{I.11}$$

$$\sum_{k=1}^{N}C_k^3 = (\sum_{k=1}^{N}C_k)^3 - 3\sum_{k\neq k'}^{N}C_k^2C_{k'} - 6\sum_{k\neq k'\neq k''}^{N}C_kC_{k'}C_{k''} \tag{I.12}$$

We can then write

$$\begin{aligned} m &= \frac{-1}{N}\{\frac{m}{k_BT}\sum_{k=1}^{N}C_k - \frac{m^3}{3k_B^3T_c^3}[(\sum_{k=1}^{N}C_k)^3 - 3\sum_{k\neq k'}^{N}C_k^2C_{k'} - 6\sum_{k\neq k'\neq k''}^{N}C_kC_{k'}C_{k''}]\} \\ &= \frac{-1}{N}[\frac{mM}{k_BT} - \frac{m^3}{3k_B^3T_c^3}M^3 + \frac{m^3}{k_B^3T_c^3}\sum_{k\neq k'}^{N}C_k^2C_{k'} + 2\frac{m^3}{k_B^3T_c^3}\sum_{k\neq k'\neq k''}^{N}C_kC_{k'}C_{k''}] \end{aligned} \tag{I.13}$$

Consequently

$$m = m(\frac{T_c}{T}) - \frac{N^2m^3}{3}(\frac{T_c}{T})^3 + ... \tag{I.14}$$

The original case

$$m = -1 + \frac{2}{N}\sum_{k=1}^{N} \frac{b_k}{a_k e^{2\beta(mC_k+f_k)} + b_k}$$

The derivative of the above expression w.r.t m.

$$1 = \frac{2}{N}\sum_{k=1}^{N} \frac{2\beta a_k b_k C_k e^{2\beta(mC_k+f_k)}}{(a_k e^{2\beta(mC_k+f_k)} + b_k)^2}|_{m=0} = \frac{2}{N}\sum_{k=1}^{N} \frac{2\beta a_k b_k C_k e^{2\beta f_k}}{(a_k e^{2\beta f_k} + b_k)^2} \tag{I.15}$$

In the absence of f_k but presence of the gain function $\Phi_i(v_i)$

$$1 = \frac{4}{N}\sum_{k=1}^{N} \frac{\beta a_k b_k C_k}{(a_k + b_k)^2} \tag{I.16}$$

Defining the critical temperature as

$$k_b T_c = \frac{4}{N}\sum_{k=1}^{N} \frac{a_k b_k C_k}{(a_k + b_k)^2} \tag{I.17}$$

Taylor expansion around $m = 0$

$$\begin{aligned} m &= \frac{-1}{N}\sum_{k=1}^{N} \frac{a_k - b_k}{a_k + b_k} - \frac{1}{N}\sum_{k=1}^{N} \frac{4a_k b_k C_k}{(a_k + b_k)^2}\beta m + \frac{1}{N}\sum_{k=1}^{N} \frac{4a_k b_k (a_k - b_k) C_k^2}{(a_k + b_k)^3}\beta^2 m^2 + \ldots \\ &= \frac{-1}{N}\sum_{k=1}^{N} \frac{a_k - b_k}{a_k + b_k} - \frac{T_c}{T} m + O(m^2) \end{aligned} \tag{I.18}$$

$$\Rightarrow$$

$$m = \frac{\frac{-1}{N}\sum_{k=1}^{N}\frac{a_k - b_k}{a_k + b_k}}{1 + \frac{T_c}{T}} = \frac{-1}{N}\sum_{k=1}^{N} \frac{a_k - b_k}{a_k + b_k}(1 - \frac{T_c}{T}) = \frac{-1}{N}\sum_{k=1}^{N} \frac{a_k - b_k}{a_k + b_k} t \ \text{ as } \ T \to T_c^- \tag{I.19}$$

Annexe II

Using the Greedy Variational Principle to derive the Feynman propagators of Hopefield Networks

$$\begin{aligned} S^j &= \int dt \{\frac{1}{2}\sum_i \dot{u}_i^2 g'(u_i) + \sum_i [-\sum_j T_{ij} v_j - f_i + \Phi_i'(v_i)]\dot{v}_i + \sum_i v_i j_i\} \\ &= \int dt \sum_i \left[\frac{m\dot{v}_i^2}{2} - \sum_j T_{ij} v_j \dot{v}_i - f_i \dot{v}_i + g^{-1}(v_i)\dot{v}_i + v_i j_i \right] \\ &= \int dt \sum_{ij} \left[\delta_{ij} \frac{m\dot{v}_i \dot{v}_j}{2} - T_{ij} v_j \dot{v}_i - \delta_{ij} f_i \dot{v}_j + \delta_{ij} g^{-1}(v_i)\dot{v}_j + \delta_{ij} v_i j_j \right] \end{aligned} \qquad \text{(II.1)}$$

where we have added the Kronecker delta for simplicity. We then Fourier transform $v_i(t)$ and $j_i(t)$

$$v_i(t) = \int_{-\infty}^{\infty} \frac{dE}{2\pi} e^{-iEt} \tilde{v}_i(E), \quad j_i(t) = \int_{-\infty}^{\infty} \frac{dE}{2\pi} e^{-iEt} \tilde{j}_i(E) \qquad \text{(II.2)}$$

along with using the following relationships

$$\dot{v}_i(t) = \int_{-\infty}^{\infty} \frac{dE}{2\pi} (-iE) e^{-iEt} \tilde{v}_i(E), \quad \int dt e^{-i(E+E')t} = 2\pi\delta(E+E') \qquad \text{(II.3)}$$

to rewrite the above action. We do this term by term

$$\int dt \delta_{ij} \frac{m\dot{v}_i \dot{v}_j}{2} = \frac{m}{2}\delta_{ij} \int \frac{dE}{2\pi} E^2 \tilde{v}_i(E)\tilde{v}_j(-E) \qquad \text{(II.4)}$$

$$\int dt T_{ij} v_i \dot{v}_j = T_{ij} \int \frac{dE}{2\pi} iE \tilde{v}_i(E)\tilde{v}_j(-E) \qquad \text{(II.5)}$$

$$\int dt \delta_{ij} f_i \dot{v}_j = \delta_{ij} f_i \int \frac{dE}{2\pi} \tilde{v}_j(E) e^{-iEt} \qquad \text{(II.6)}$$

$$\int \delta_{ij} dt v_i j_j = \delta_{ij} \int \frac{dE}{2\pi} \tilde{v}_i(E)\tilde{j}_j(-E) \qquad \text{(II.7)}$$

Since there is no known method of calculating $g^{-1}(\tilde{v}_i)$ and since it is not possible to interchange the time integral and the $g^{-1}(\tilde{v}_i)$, which we would have to do in order to proceed, we assume that $g^{-1}(v_i)$ is an analytic function of $v_i(t)$.As such it can be Taylor expanded

$$g^{-1}(v_i) = \sum_{n=0} \frac{g^{-1(n)}(0)}{n!} v_i^n \tag{II.8}$$

For simplicity we will write $g^{-1}(0) = G_0$.The derivatives of this function at zero will be denoted G_0' etc.

$$\begin{aligned}
&\int \delta_{ij} dt g^{-1}(v_i)\dot{v}_j = \int dt \delta_{ij}\{G_0\dot{v}_j + G_0{}' v_i\dot{v}_j + \frac{1}{2!}G_0{}'' v_i^2\dot{v}_j + \frac{1}{3!}G_0{}''' v_i^3\dot{v}_j + ...\} \\
&= G_0\delta_{ij}\int \frac{dE}{2\pi}\tilde{v}_j(E)e^{-iEt} + G_0{}'\delta_{ij}\int \frac{dE}{2\pi}(iE)\tilde{v}_i(E)\tilde{v}_j(-E) \\
&+ \frac{1}{2!}G_0{}''\delta_{ij}\int \frac{dE}{2\pi}\frac{dE_1}{2\pi}\frac{dE_2}{2\pi}(-iE)\tilde{v}_i(E)\tilde{v}_i(E_1)\tilde{v}_i(E_2)\int dte^{-i(E+E_1+E_2)t} \\
&+ ...
\end{aligned} \tag{II.9}$$

It is quite clear that the Taylor expansion has to be truncated. The truncated Fourier transformed action is

$$\begin{aligned}
S^j = \sum_{ij}\int \frac{dE}{2\pi}\Bigg[&\frac{m}{2}\delta_{ij}E^2\tilde{v}_i(E)\tilde{v}_j(-E) - T_{ij}iE\tilde{v}_i(E)\tilde{v}_j(-E) - \delta_{ij}f_i\tilde{v}_j(E)e^{-iEt} \\
&+ [G_0\delta_{ij}\tilde{v}_j(E)e^{-iEt} + G_0{}'\delta_{ij}(iE)\tilde{v}_i(E)\tilde{v}_j(-E) + ...] + \delta_{ij}\tilde{v}_i(E)\tilde{j}_j(-E)\Bigg]
\end{aligned} \tag{II.10}$$

Reorganizing the terms gives

$$S^j = \sum_{ij}\int \frac{dE}{2\pi}\left[K\tilde{v}_i(E)\tilde{v}_j(-E) + M\tilde{v}_j(E)e^{-iEt} + \delta_{ij}\tilde{v}_i(E)\tilde{j}_j(-E)\right] \tag{II.11}$$

where

$$K = \frac{mE^2\delta_{ij}}{2} - iET_{ij} + iEG_0{}'\delta_{ij}, \quad M = G_0\delta_{ij} - \delta_{ij}f_j \tag{II.12}$$

The most common approach here is to decouple $\tilde{v}_i$ from the source $\tilde{j}_i$ through a variable change. Let's assume that we make the variable change

$$\tilde{v}_i(E) = \tilde{v}'_i(E) + \alpha(E)\tilde{j}_i(E) \tag{II.13}$$

and determine $\alpha(E)$ in order to achieve the desired effect. The new action reads

$$\begin{aligned} S^j = \sum_{ij} \int \frac{dE}{2\pi} \Big[& K\tilde{v}'_i(E)\tilde{v}'_j(-E) + \\ & + K\{\alpha(-E)\tilde{v}'_i(E)\tilde{j}_j(-E) + \alpha(E)\tilde{v}'_j(-E)\tilde{j}_i(E)\} + \delta_{ij}\tilde{v}'_i(E)\tilde{j}_j(-E) \\ & + [K\alpha(E)\alpha(-E) + \alpha(E)\delta_{ij}]\tilde{j}_i(E)\tilde{j}_j(-E) \\ & + M\tilde{v}'_j(E)e^{-iEt} + M\alpha(E)\tilde{j}_i(E)e^{-iEt} \Big] \end{aligned} \tag{II.14}$$

Taking into account the symmetries of T_{ij} and δ_{ij}, we note that the integrals in the second line involve terms that either vanish or double. In particular we get two contributions from

$$\int_{-\infty}^{+\infty} \frac{dE}{2\pi} E^2 \alpha(E)\tilde{v}'_j(-E)\tilde{j}_i(E) = -\int_{+\infty}^{-\infty} \frac{dE}{2\pi} E^2 \alpha(-E)\tilde{v}'_j(E)\tilde{j}_i(-E) \tag{II.15}$$

while the following integrals cancel due to sign difference

$$\int_{-\infty}^{+\infty} \frac{dE}{2\pi} E\alpha(E)\tilde{v}'_j(-E)\tilde{j}_i(E) = \int_{+\infty}^{-\infty} \frac{dE}{2\pi} E\alpha(-E)\tilde{v}'_j(E)\tilde{j}_i(-E) \tag{II.16}$$

This means that the mixed terms will vanish if

$$mE^2\delta_{ij}\alpha(-E) + \delta_{ij} = 0 \Rightarrow \alpha(E) = \alpha(-E) = \frac{-1}{mE^2} \tag{II.17}$$

Finally we arrive at

$$\begin{aligned} S^j = \sum_{ij} \int \frac{dE}{2\pi} \Big[& K\tilde{v}'_i(E)\tilde{v}'_j(-E) + [\frac{K}{m^2E^4} - \frac{\delta_{ij}}{mE^2}]\tilde{j}_i(E)\tilde{j}_j(-E) \\ & + M\tilde{v}'_j(E)e^{-iEt} - \frac{M}{mE^2}\tilde{j}_i(E)e^{-iEt} \Big] \end{aligned} \tag{II.18}$$

The inverse Fourier transform of this action is

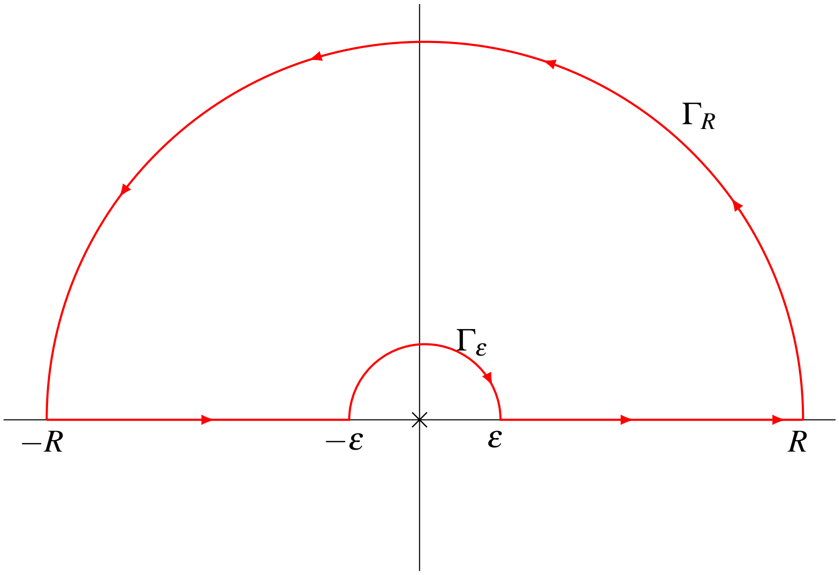
$$
\begin{aligned}
S^j = \int dt \Bigg[\sum_i \frac{m\dot{v}'^2_i}{2} - \sum_{ij} T_{ij} v'_i \dot{v}'_j + [G_0 \dot{v}'_j + G_0{}' v'_i \dot{v}'_j + \ldots] - \sum_i f_i \dot{v}'_i \Bigg] - \\
- \sum_{ij} \frac{\delta_{ij}(G_0 - f_j)}{m} \int dt' j_i(t') D_1(t-t') \\
- \sum_{ij} \frac{\delta_{ij}}{2m} \int dt \int dt' j_i(t) D_1(t-t') j_j(t') \\
- \sum_{ij} \frac{iT_{ij} - i\delta_{ij} G_0{}'}{m^2} \int dt \int dt' j_i(t) D_2(t-t') j_j(t')
\end{aligned}
\tag{II.19}
$$

$$
D_1(t-t') = \int_{-\infty}^{\infty} \frac{dE}{2\pi} \frac{e^{-iE(t'-t)}}{E^2}, \;\; D_2(t-t') = \int_{-\infty}^{\infty} \frac{dE}{2\pi} \frac{e^{-iE(t'-t)}}{E^3} \tag{II.20}
$$

Changing the variable $z = -E(t-t')$ we can rewrite and integral of the form

$$
\int dE \frac{e^{-iE(t-t')}}{E^k} = \int \frac{e^{iz}}{z^k} dz, \; k > 0 \tag{II.21}
$$

Assume that z is a complex variable and use this assumption to calculate the integral with the help of residue calculus, Cauchy's theorem etc. If we Choose the following contour



it is clear that

$$\int_{-R}^{-\varepsilon}\frac{e^{iz}}{z^k}dz-\int_{\Gamma_\varepsilon}\frac{e^{iz}}{z^k}dz+\int_{\varepsilon}^{R}\frac{e^{iz^k}}{z}dz+\int_{\Gamma_R}\frac{e^{iz}}{z^k}dz=0 \tag{II.22}$$

In the integrals over the real axis, we simply set the imaginary part of z equal to zero. This is the integral we are seeking to calculate.

$$\underset{\substack{\lim R\to\infty\\ \lim\varepsilon\to 0}}{\int_{-R}^{-\varepsilon}\frac{e^{ix}}{x^k}dx}+\underset{\substack{\lim R\to\infty\\ \lim\varepsilon\to 0}}{\int_{\varepsilon}^{R}\frac{e^{ix}}{x^k}dx}=\underset{\lim R\to\infty}{\int_{-\infty}^{\infty}\frac{e^{ix}}{x^k}dx}=\int_{-R}^{R}\frac{e^{ix}}{x^k}dx \tag{II.23}$$

Thus

$$\int_{-\infty}^{\infty}\frac{e^{ix}}{x^k}dx=\int_{\Gamma_\varepsilon}\frac{e^{iz}}{z^k}dz-\int_{\Gamma_R}\frac{e^{iz}}{z^k}dz \tag{II.24}$$

We can show that the integral over Γ_R has a vanishing contribution in the limit $R\to\infty$

$$\begin{aligned}|\int_{\Gamma_R}\frac{e^{iz}}{z^k}|&\leq\int_0^{\pi}|\frac{e^{i(R\cos\theta+iR\sin\theta}}{R^ke^{ik\theta}}|Rd\theta=\int_0^{\pi}\frac{e^{-R\sin\theta}}{R^k}d\theta\\ =2\int_0^{\pi/2}\frac{e^{-R\sin\theta}}{R^k}d\theta&\leq 2\int_0^{\pi/2}e^{-2R\theta/\pi}d\theta=\pi R^{-k}(1-e^{-R})\end{aligned} \tag{II.25}$$

which clearly vanishes as R approaches infinity. Note that we have used Jordan's inequality $\frac{2\theta}{\pi}\leq\sin\theta\leq\theta$, valid in $[0,\pi/2]$, to obtain the above result. The integral over the indentation at the simple pole in the origin is evaluated by expanding e^{iz} around the origin and rewriting $z=\varepsilon e^{i\theta}$

$$\int_{\Gamma_\varepsilon}\frac{e^{iz}}{z^k}dz=\sum_0^{\infty}\int_{\Gamma_\varepsilon}\frac{i^nz^{n-k}}{n!}dz=\int_{\Gamma_\varepsilon}\frac{1}{z^k}[1+iz-\frac{z^2}{2}-\frac{iz^3}{3!}+...]dz \tag{II.26}$$

If $k=2$, this integral can be written as

$$\int_{\Gamma_\varepsilon}[\frac{1}{z^2}+\frac{i}{z}+E(z)]dz \tag{II.27}$$

where $E(z)$ is a finite contribution and

$$|\int_{\Gamma_\varepsilon} E(z)dz| \leq \int_{\Gamma_\varepsilon} |E(z)|dz = C\pi\varepsilon \tag{II.28}$$

that vanishes at $\varepsilon \to 0$. This is, however, not the case for

$$\int_{\Gamma_\varepsilon} \frac{1}{z^2}dz = \int_0^\pi \frac{i}{\varepsilon}e^{-i\theta}d\theta \tag{II.29}$$

which diverges in the limit $\varepsilon \to 0$. Using a similar argument we can prove that even

$$\int_{\Gamma_\varepsilon} \frac{1}{z^3}dz \tag{II.30}$$

diverges in the limit $\varepsilon \to 0$.

The roots of these non-glamorous results can be traced back to the Lagrangian of the Hopfiled network. In order to generate interesting results akin to those in the case of, for instance, harmonic oscillator, the Lagrangian would need to include a kinetic type term $\sim \dot{v}^2$ as well as a term $\sim v^2$. It is in fact the interplay between these two terms that leads to interesting/manageable Green's functions can then be used for a path-integral formulation of the Hopfield networks, opening the possibility of investigating all kinds of phenomena such as phase transitions, which we set out to do originally. It should be noted that the truncated Taylor expansion in the above treatment is not to be blamed for this as it would not have lead to the desired result.

Annexe III

What went wrong ?

The Lagrangian in [48]

$$L = K(\dot{v}) - P(v) = K + \frac{\partial E}{\partial v}\frac{dv}{dt} \tag{III.1}$$

where $K = \frac{1}{2}\sum_i \dot{u}_i^2 g'(u_i)$ and E as in (8.19), can be slightly rewritten through the insertion of $\dot{u}_i = \dot{v}_i/g'(u_i)$ and the equations of motion resulting from the greedy action principle

$$\dot{u}_i = -\sum_{ij} T_{ij} v_j - f_i + \Phi_i'(v_i). \tag{III.2}$$

With this new Lagrangian

$$L = \frac{1}{2}\sum_i \dot{u}_i^2 g'(u_i) - \frac{1}{2}\sum_{ij} T_{ij} v_i v_j - \sum_i f_i v_i + \sum_i \Phi_i'(v_i) v_i \tag{III.3}$$

the modified action with the source term is given by

$$\begin{aligned} S^j &= \int dt \{ \frac{1}{2}\sum_i \dot{u}_i^2 g'(u_i) - \frac{1}{2}\sum_{ij} T_{ij} v_i v_j - \sum_i f_i v_i + \sum_i \Phi_i'(v_i) v_i + \sum_i v_i j_i \} \\ &= \int dt \sum_{ij} \left[\frac{m\delta_{ij}}{2} \dot{v}_i \dot{v}_j - \frac{T_{ij}}{2} v_i v_j - \delta_{ij} f_i v_j + \delta_{ij} g^{-1}(v_i) v_j + \delta_{ij} v_i j_j \right] \end{aligned} \tag{III.4}$$

The Fourier transformed action is

$$\begin{aligned} S^j = \sum_{ij} \int \frac{dE}{2\pi} \Bigg[& \frac{m\delta_{ij}}{2} E^2 \tilde{v}_i(E) \tilde{v}_j(-E) - \frac{T_{ij}}{2} \tilde{v}_i(E) \tilde{v}_j(-E) - \delta_{ij} f_i \int dt \tilde{v}_j(E) e^{-iEt} \\ & + \{ \int dt G_0 \delta_{ij} \tilde{v}_j(E) e^{-iEt} + {G_0}' \delta_{ij} \tilde{v}_i(E) \tilde{v}_j(-E) + \ldots \} + \delta_{ij} \tilde{v}_i(E) \tilde{j}_j(-E) \Bigg] \end{aligned} \tag{III.5}$$

We use the abbreviations (note the differences compared to previous cases as well as K below no longer referring to (III.1))

$$K = \frac{mE^2\delta_{ij}}{2} - \frac{T_{ij}}{2} + {G_0}'\delta_{ij}, \quad M = G_0 - f_j \tag{III.6}$$

to simplify this action

$$S^j = \sum_{ij} \int \frac{dE}{2\pi} \left[K\tilde{v}_i(E)\tilde{v}_j(-E) + M\delta_{ij} \int dt \tilde{v}_j(E) e^{-iEt} + \delta_{ij}\tilde{v}_i(E)\tilde{j}_j(-E) \right] \tag{III.7}$$

A change of variable as in the previous analysis leads to

$$\begin{aligned} S^j = \sum_{ij} \int \frac{dE}{2\pi} \Big[& K\tilde{v}'_i(E)\tilde{v}'_j(-E) + \\ & + K\{\alpha(-E)\tilde{v}'_i(E)\tilde{j}_j(-E) + \alpha(E)\tilde{v}'_j(-E)\tilde{j}_i(E)\} + \delta_{ij}\tilde{v}'_i(E)\tilde{j}_j(-E) \\ & + [K\alpha(E)\alpha(-E) + \alpha(E)\delta_{ij}]\tilde{j}_i(E)\tilde{j}_j(-E) \\ & + M\delta_{ij} \int dt [\tilde{v}'_j(E) + \alpha(E)\tilde{j}_j(E)] e^{-iEt} \Big] \end{aligned} \tag{III.8}$$

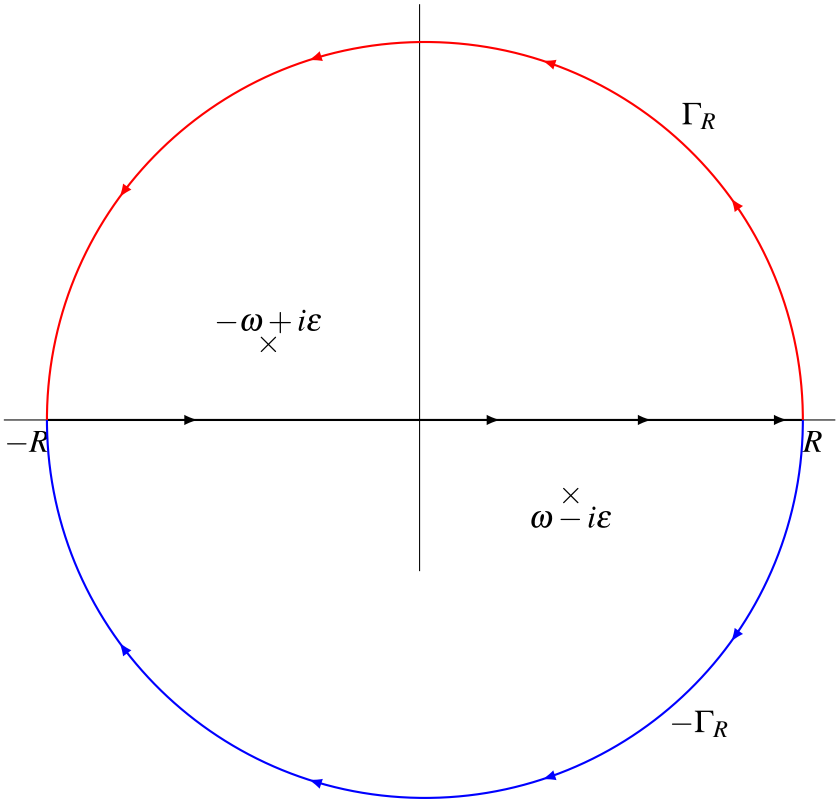
A novelty is that α is now given by

$$\alpha(E) = \frac{-\delta_{ij}}{2K} = \frac{-\delta_{ij}}{mE^2\delta_{ij} - T_{ij} + 2{G_0}'\delta_{ij}} \tag{III.9}$$

and

$$S^j = \sum_{ij} \int \frac{dE}{2\pi} \left[K\tilde{v}'_i(E)\tilde{v}'_j(-E) - \frac{\delta_{ij}}{4K}\tilde{j}_i(E)\tilde{j}_j(-E) + M\delta_{ij} \int dt [\tilde{v}'_j(E) - \frac{\delta_{ij}}{2K}\tilde{j}_i(E)] e^{-iEt} \right] \tag{III.10}$$

The integrand above has poles at $E = \pm\omega$ and so the integral can be done as in the previous case. We can also analytically continue the integrand to the complex plane thereby shifting the two poles to $\omega - i\varepsilon$ and $-\omega + i\varepsilon$. Again the integrals can be evaluated with the help of Cauchy's theorem. The final result is different depending on the time ordering : If $t > t'$, we use the lower contour (blue) as $-\Gamma_R$ its contribution will vanish in the limit $R \to \infty$. We do the integration over $-\Gamma_R$ using a parallel argument.



The final answer, combining the two time ordering scenarios, is given by

$$D(t-t') = \frac{-i}{4m\delta_{ij}\omega} e^{-i\omega|t-t'|} \tag{III.11}$$

Note that this is a formal solution which due to the presence of δ_{ij} in the denominators of ω and $D(t-t')$ is problematic. Observe, however, that terms with different indices are excluded from the onset. Taking this into account, the action reads

$$\begin{aligned} S^j = \int dt \left[\sum_i \frac{m\dot{v'}_i^2}{2} - \sum_{ij} \frac{T_{ij}}{2} v'_i v'_j + \sum_i [G_0 v'_i + G_0{}' v'_i v'_i + \ldots] - \sum_i f_i v'_i \right] - \\ - \sum_{i=j} M \int dt \int dt' j_i(t) D(t'-t) - \frac{1}{2} \sum_{i=j} \int dt \int dt' j_i(t) D(t'-t) j_j(t') \end{aligned} \tag{III.12}$$

we see that the Green's function is

$$D(t-t') = \int \frac{dE}{2\pi} \frac{1}{mE^2 + 2G_0{}'} e^{-iE(t-t')} = \frac{-i}{4m\omega'} e^{-i\omega'|t-t'|} \tag{III.13}$$

with $\omega' = \sqrt{\frac{-2G_0'}{m}}$.

Annexe IV

A Quantum field theoretical approach

The action

$$\begin{aligned} S &= \int d^4x \sum_{ij} \left[\frac{m\delta_{ij}}{2} \partial_t V_i \partial V_j - \frac{T_{ij}}{2} V_i V_j - \delta_{ij} f_i V_j + \delta_{ij} \{ G_0 V_j + G_0' V_i V_j + \frac{G_0''}{2} V_i^2 V_j + ... \} \right] \\ &= \int d^4x \sum_{ij} \left[\frac{m\delta_{ij}}{2} \partial_t V_i \partial V_j - \frac{T_{ij}}{2} V_i V_j + \delta_{ij} G_0' V_i V_j + \delta_{ij} (G_0 - f_i) V_j + \delta_{ij} \mathscr{O}(V^3) \right] \end{aligned} \tag{IV.1}$$

We can use integration by parts in the first term, neglecting a total time derivative to get

$$\begin{aligned} S &= \int d^4x \sum_{ij} \left[V_i \{ -\frac{m\delta_{ij}}{2} \partial_t \partial^t - \frac{T_{ij}}{2} + \delta_{ij} G_0' \} V_j + \delta_{ij} (G_0 - f_i) V_j + \delta_{ij} \mathscr{O}(V^3) \right] \\ &= \int d^4x \sum_{ij} \left[-\frac{1}{2} V_i O_t V_j + \delta_{ij} M V_j + \delta_{ij} \mathscr{O}(V^3) \right] \end{aligned} \tag{IV.2}$$

where $O_t = m\delta_{ij}\partial_t\partial^t + T_{ij} - 2\delta_{ij}G_0'$ is an operator. Adding a source term to the above action (and dropping the terms $\mathscr{O}(V^3)$, the generating functional reads

$$Z[j] = N \int D[V] e^{-i \int d^4x \sum_{ij} \left[\frac{1}{2} V_i O_t V_j - \delta_{ij} M V_j - \delta_{ij} j_i V_j \right]} \tag{IV.3}$$

Once again, but differently this time and without Fourier transforming the variables, we shift the fields $V_i(x)$ with the goal of separating it from the source term. We find that

$$V_i(x) = V_i'(x) + i \int d^4y D(x-y) j_i(y) \tag{IV.4}$$

will achieve the desired result. The shift results in

$$
\begin{aligned}
&-i\int d^4x\sum_{ij}\left[\frac{1}{2}V_iO_tV_j-\delta_{ij}MV_j-\delta_{ij}j_iV_j\right]=\\
&-i\int d^4x\sum_{ij}\left[\frac{1}{2}V_i'O_tV_j'-\delta_{ij}MV_j'-\frac{i}{2}\delta_{ij}\int d^4y\,j_i(x)D(x-y)j_j(x)\right]\\
&-i\delta_{ij}M\int d^4y\,j_j(y)D(x-y)\}
\end{aligned}
\tag{IV.5}
$$

Once again we get the entire action back plus two terms that we will discuss below. The partition function is given by

$$
\begin{aligned}
Z[j]&=N\int D[V']e^{i\int d^4x\mathscr{L}[V']}e^{\frac{1}{2}\sum_i\int d^4xd^4y\,j_i(x)D(x-y)j_i(y)}e^{-\sum_i\int d^4xd^4yMD(x-y)j_i(y)}\\
&=Z[0]e^{\frac{1}{2}\sum_i\int d^4xd^4y\,j_i(x)D(x-y)j_i(y)}e^{-\sum_i\int d^4xd^4yMD(x-y)j_i(y)}
\end{aligned}
\tag{IV.6}
$$

Annexe V

Renormalization Group Theory, Deep Learning and Information Theory

The similarities between RG and DL have been pointed out in multiple works [8, 46, 13]. In most cases, the comparison is done between lattice spin models on which the RG formalism has been applied in the context of statistical physics, and certain types of basic NNs. Since both RG and NNs are procedures that distill large scale structure from complex microscopic interactions, it is not far-fetched to try to draw this parallel and investigate the similarities and differences between these two procedures.

It should be noted that this chapter is part of an exploration attempt to understand the link between Bayesian learning and RG theory. In the end, it turned out this line of research focuses on using ML, in particular the machinery of RBMs, to find ideal coarse-graining schemes in systems that defy description due to our limited knowledge of their microscopic interactions. Although we did not pursue this direction, one of our temporary goals was to explore the relationship between learning and RG i.e. the probable links between how NNs find the minima of the loss landscapes, and the effect of an RG transformation on the loss landscapes. Other than that, the reader can skip this chapter or read it as an "extra".

As data pass from the input through hidden layers to the output layer of the network, it is believed that a NN or a Deep Neural Network (DNN) with its multi-layer architecture successively extracts relevant (hopefully disregarding irrelevant) information. This is akin to the central theme of RG where an iterative coarse-graining scheme is employed to tackle problems involving multiple length scales, a process during which short distance degrees of freedom, sometimes called high momentum or ultraviolet (UV) degrees of freedom, are integrated out. The resulting large scale theory describes the manifestation of the microscopic world in the infrared/macroscopic world.

The **Deep Learning** equivalent of coarse-graining is seen in the context of RBMs as sequential marginalizations over some variables that drive the network from UV/high energy to IR/low energy variables. Similar to the general case of RG, after each sequential marginalization of UV degrees of freedom, new couplings between IR degrees of freedom are induced.

Note that this "integrating-out-equivalency" differs from coarse-graining in RG in that while it is possible to do coarse-graining infinite many times in RG, a process that gives rise to the phase diagram of RG flows, it is not clear how the "RG flow" of the DL type of coarse-graining can be studied. This would clearly be necessary for the analysis of different algorithms, phase transitions, scaling laws and universality. Apart from their structure, NNs also involve various types of learning which adds an extra layer of complexity to the analysis of possible parallels between RG and DL.

Applications of RG procedures in statistical physics are usually neither unique nor exact. While it is fairly straight-forward in simple cases such as 1d and 2d Ising models, the procedure can become increasingly complex and so there are approximate methods such as variational RG [35], a method used in [46] to bridge the gap between RG and DNNs. The latter argues for a one-to-one relationship between variational RG and the mechanism by which a simple NN such as an RBM works. An RG flow i.e. the result of infinite many RG transformations is then equivalent to a DNN of stacked RBMs.

Starting from the Hamiltonian of a typical Ising model 4.1, variational RG finds the coarse-grained Hamiltonian by constructing a parametrized function that encodes the interactions between the coarse-grained and physical spins. The free energy of the coarse-grained system is then calculated from the free energy of the coupled system after integrating out the physical degrees of freedom. The parameters of the coarse-grained system are those that minimize the difference between the free energy of the physical and coarse-grained system.

Note that in the case of RBM, the analogy with the above is that the physical spins are those in the visible layers denoted v_i, and the coarse-grained spins are those in the hidden layers, denoted h_i. With a somewhat simplified notation (using h instead of h_i) this amounts to

$$e^{-H_\lambda^{RG}(h)} = Tr_{\mathbf{v}} e^{T_\lambda(v,h)-H(v)} \tag{V.1}$$

where $H_\lambda^{RG}(h)$ is the coarse-grained Hamiltonian and $H(v)$ is the "exact" Hamiltonian of the Ising model. Variantional RG has to find the parameters λ such that the difference between the free energies of the coarse-grained and original system $\Delta F = F_\lambda(h) - F(v)$ is minimized. When the difference is zero, we get

$$Tr_{\mathbf{v}} e^{T_\lambda(v,h)} = 1 \tag{V.2}$$

Without delving into the general method of finding the function T_λ, [46] argues that both $T_\lambda(v,h)$ and $H_\lambda^{RBM}(v,h)$ encode the connection between the coarse-grained and physical degrees of freedom and that

$$T_\lambda(v,h) = -H_\lambda^{RBM}(v,h) + H(v) \tag{V.3}$$

where λ are the coupling constants (b_i, c_i, w_{ij}) of the RBM Hamiltonian 7.11. Note that this statement is equivalent to saying that the hidden layers of RBM are coarse-grained versions of the physical/visible layer. Not surprisingly, this equation defines a one-to-one map between variational RG in the Ising model and the RBMs where the visible degrees of freedom have been marginalized over. Inserting V.3 in V.1, dividing by the partition function and using the fact that RBM is an energy model

$$p_\lambda(v,h) = \frac{e^{-H_\lambda^{RBM}}(v,h)}{Z}, \quad p_\lambda(v) = \frac{e^{-H_\lambda^{RBM}}(v)}{Z} \tag{V.4}$$

we get

$$H_\lambda^{RG}(v) = H_\lambda^{RBM}(v) \tag{V.5}$$

This result opens up the possibility of formulating the problem in Bayesian terms. If V.2

is satisfied i.e. if the RG transformation is exact, it is easily proven that

$$1 = Tr_h e^{T_\lambda(v,h)} = ... = Tr_h p_\lambda(h|v) e^{-H_\lambda^{RBM}(v)+H(v)} \tag{V.6}$$

which implies that the variational RBM Hamiltonian is identical to the exact Hamiltonian of the Ising model

$$H_\lambda^{RBM}(v) = H(v) \tag{V.7}$$

We know that RBMs find the relevant coupling constants by e.g. mimizing the Kullback-Leibler divergence between the true distribution of data $P([v])$ and the variational distribution $p_\lambda([v])$. When an RG transformation can be performed exactly, since both RBM and the Ising model are engery models, V.7 implies that the

$$D_{KL}(P(v)||p_\lambda(v)) = 0 \tag{V.8}$$

Another implication of the above result is that the function T_λ is identical to the conditional distribution through

$$e^{T_\lambda(h,v)} = p_\lambda(h|v) \tag{V.9}$$

This conditional probability is the start of a more generalized Bayesian approach to RG. The following is based on the information theoretical approach to RG employed in [43]. Suppose we start with an energy model where all the degrees of freedom are denoted X. The joint distribution of all the variables is given by

$$p(X) = \frac{e^{-\beta H(X)}}{Z} \tag{V.10}$$

As usual, we are looking for the coarse-grained Hamiltonian $H^{RG}(X')$. A common approach is to split the Hamiltonian $H(X)$ in two parts, one containing the intra-block interactions with the blocks being the coarse-grained variables, H_0, and one containing the inter-block interactions V

$$H(X) = H_0(X) + V(X) \tag{V.11}$$

Denoting the coarse-grained blocks by h_j, it is clear that if $dim(X) = m$, $X = \cup_{j=1}^{n} h_j$ and $dim(h_j) = m/n,\ \ m > n$. Each such block of units to be coarse-grained will be associated with a unit v_j described by the coarse-grained degrees of freedom X'_j. If H_b is the Hamiltonian of a single block, the total contribution of blocks to the block Hamiltonian is

$$H_0(X) = \sum_{j=1}^{n} H_b(h_j) \tag{V.12}$$

Let's assume that the RG procedure coarse-grains the variables X into X' and that the new distribution in terms of X' is given by

$$p(X') = Tr_X p(X'|X) p(X) \tag{V.13}$$

Since the coarse-graining of the variables in each block is independent from other blocks (translation invariance), we can write

$$p(X'|X) = \prod_{j=1}^{n} p(v_j|h_j) \tag{V.14}$$

Furthermore, the distribution within each block $p(h_j)$ as well as the distribution of total block distributions p_0 can be written as

$$p(h_j) = \frac{e^{-\beta H_b(h_j)}}{Z_b}\ \text{(in blocks)},\quad Z_b = \sum_{X_i \in h_j} e^{-\beta H_b(h_j)} \tag{V.15}$$

$$p_0 = \frac{e^{-\beta H_0(X)}}{Z_0}\ \text{(all blocks)},\quad Z_0 = \sum_{X_i \in X} e^{-\beta H_0(X)} \tag{V.16}$$

Assuming that Z_0 can be factorized (due to factorisation in Hilbert space ?) we have

$$Z_0 = \sum_{X_i \in X} e^{K_0(X)} = \sum_{X_1} \cdots \sum_{X_m} e^{K_0(X)} = \prod_{j=1}^{n} \sum_{X_i \in h_j} e^{K_b(h_j)} = \prod_{j=1}^{n} Z_b \tag{V.17}$$

And we get

$$\frac{e^{-\beta H^{RG}(X')}}{Z'} = Tr_X[p(X'|X)\frac{e^{-\beta H}}{Z}], \quad \text{(Z invariant under coarse-graining)} \tag{V.18}$$

$$\Rightarrow$$

$$\begin{aligned}
e^{-\beta H^{RG}(X')} &= Tr_X[p(X'|X)e^{-\beta H}] && \text{(V.19)}\\
&= Tr_X[p(X'|X)e^{-\beta(H_0+V}] && \text{(V.20)}\\
&= Tr_X[\prod_{j=1}^{n} p(v_j|h_j)e^{-\beta \Sigma H_b(h_j)}e^{-\beta V}] && \text{(V.21)}\\
&= Tr_X[\prod_{j=1}^{n} p(v_j|h_j)e^{-\beta H_b(h_j)}e^{-\beta V}] && \text{(V.22)}\\
&= Tr_X[\prod_{j=1}^{n} p(v_j|h_j)(Z_b p(h_j))e^{-\beta V}] && \text{(V.23)}\\
&= Z_b^n Tr_X[\prod_{j=1}^{n} p(v_j|h_j)p(h_j)e^{-\beta V}] && \text{(V.24)}\\
&= Z_b^n Tr_X[e^{-\beta V}\prod_{j=1}^{n} p(h_j|v_j)p(v_j)] && \text{(V.25)}\\
&= Z_b^n Tr_X[e^{-\beta V}p(X|X')p(X')] && \text{(V.26)}\\
&= Z_b^n p(X')Tr_X[e^{-\beta V}p(X|X')] && \text{(V.27)}\\
&= Z_b^n p(X')\langle e^{-\beta V}\rangle && \text{(V.28)}
\end{aligned}$$

where the average is w.r.t. $p(X|X')$. Furthermore, assuming that V is small (in some sense), the standard procedure is to expand this expression in terms of the cumulants :

The coarse-grained Hamiltonian of is given by

$$-\beta H^{RG}(X') = n\ln Z_b + \ln p(X') + \ln\langle e^{-\beta V}\rangle \tag{V.29}$$

$$= n\ln Z_b + \ln p(X') - \beta\langle V\rangle + \frac{\beta^2}{2}[\langle V^2\rangle - \langle V\rangle^2] + O(V^3) \tag{V.30}$$

V.1 Real Space Mutual Information

The problem with extracting higher level concepts through coarse-graining is that the latter is not unique and that depending on the particular scheme chosen for the problem at hand, RG transformations may lead to very complex calculations. It would, therefore, be important to find a way to perform RG transformation when our knowledge of the microscopic make of a system is limited or even non-existent which will hinder a meaningful coarse-graining scheme. [39] propose an algorithm (RSMI) based on maximizing Real Space (formerly known as Block Spin) Mutual Information, that is capable of identifying the relevant d.o.f , where relevant refers to the terminology of RG. Furthermore, they show that RSMI is capable of performing RG transformations without prior knowledge of the system and that maximizing mutual information naturally leads to certain coarse-graining schemes.

The idea of maximizing mutual information and coarse-graining in RG do not seem too far apart if we take note of the fact that MIM encourages the hidden (coarse-grained) units to couple to combinations of V that are strongly correlated with the environment i.e. carry maximum amount of information.

All has been done in the context of RBMs. Also the idea of MIM seems more like a convenient tool than anything else. Even then, we seem to need to calculate certain microscopic entities and/or make certain assumptions about the nature of the microscopic system. For instance... What about other networks ?

Consider a system described by a set of variables X, partitioned into the variables $X = (\mathscr{O}, \mathscr{E}, \mathscr{B}, V, \mathscr{B}, \mathscr{E}, \mathscr{O})$ where $V, \mathscr{E}, \mathscr{B}, \mathscr{O}$ are the visible/physical , environment, buffer and remaining variables, respectively. For simplicity, it is assumed here that $\mathscr{B} = \emptyset$ and that the variables $\mathscr{O}$ are included in $\mathscr{E}$. The d.o.f of a small visible unit V are to be coarse-grained into new variables H such that the new relevant d.o.f described by H in V depend on both V and $\mathscr{E}$ through the conditional distribution $P_\Lambda(H|V)$. In order to maximize the mutual information between the visible units and the environment, the idea is to define the coarse-grained units H as a composite function of the d.o.f of V. Thus, given Monte-Carlo (MC) samples $(V,\mathscr{E})_i$, the objective is to maximize the mutual information

$$I_\Lambda(H:\mathscr{E}) = \sum_{H,\mathscr{E}} P_\Lambda(H,\mathscr{E}) log \frac{P_\Lambda(H,\mathscr{E})}{P_\Lambda(H)P(\mathscr{E})} \tag{V.31}$$

We make the assumption that the collective distribution of the variables X is given by a Boltzmann distribution

$$P(X) = P(V,\mathscr{B},\mathscr{E}) = \frac{e^{-H(x_i)}}{Z} \tag{V.32}$$

In practice, both $P(X)$ and its marginalizations

$$P(V) = \sum_{\mathscr{B}\mathscr{E}} P(X) \tag{V.33}$$

and

$$P(V,\mathscr{E}) = \sum_{\mathscr{B}} P(X) \tag{V.34}$$

are given by MC samples $(V,\mathscr{B},\mathscr{E})_i$ and restrictions thereof. To this end, two RBMs are used to approximate the distributions $P(V.\mathscr{E})$ and $P(V)$ through contrastive divergence (CD).

$$\Theta\text{-}RBM \xrightarrow{CD} \Theta_{V,\mathscr{E}} \longrightarrow P_\Theta(V,\mathscr{E})$$
$$\Theta\text{-}RBM \xrightarrow{CD} \Theta_V \longrightarrow P_\Theta(V)$$

As seen above, the conditional distribution $P(H|V)$ has a central role is coarse-graining. The core of the RSMI algorithm is to find the parameters of this distribution in such a why that the mutual information between a third RBM is used to find the visible units (through H) and the environment is maximized

$$\Lambda\text{-}RBM \overset{MIM(SGD)}{\longrightarrow} \Lambda = (a_i, b_i, \lambda_{ij}) \longrightarrow P_\Lambda(H|V) = \frac{e^{-E(H,V)}}{\sum_H e^{-E(H,V)}}$$

where $E(H,V)$ is given by the RBM Hamiltonian with parameters (a_i, b_i, λ_{ij}) as previously defined. This is done through SGD. The distribution $P(\mathscr{E})$ can be removed from the expression for mutual information above as it does not depend on any parameters. The function to be maximized after a series of manipulations is

$$\begin{aligned}
A_\Lambda(H:\mathscr{E}) &= \sum_{H,\mathscr{E}} P_\Lambda(H,\mathscr{E}) log \frac{P_\Lambda(H,\mathscr{E})}{P_\Lambda(H)} \\
&= \sum_{H,\mathscr{E}} P_\Lambda(H,\mathscr{E}) log \frac{\sum_V P(V,\mathscr{E}) P_\Lambda(H|V)}{\sum_{V',\mathscr{E}'} P_\Lambda(H,\mathscr{E}',V')} \\
&= \sum_{H,\mathscr{E}} P_\Lambda(H,\mathscr{E}) log \frac{\sum_V P(V,\mathscr{E}) P_\Lambda(H|V)}{\sum_{V',\mathscr{E}'} P_\Lambda(H|V') P(\mathscr{E}',V')} \\
&= \sum_{H,\mathscr{E}} P_\Lambda(H,\mathscr{E}) log \frac{\sum_V P_\Theta(V,\mathscr{E}) P_\Lambda(H,V)/P_\Lambda(V)}{\sum_{V'} P_\Lambda(H,V') P_\Theta(V')/P_\Lambda(V')} \\
&= \sum_{H,\mathscr{E}} P_\Lambda(H,\mathscr{E}) log \frac{\sum_V e^{-E_\Theta(V,\mathscr{E}) - E_\Lambda(H,V) + E_\Lambda(V)}}{\sum_{V'} e^{-E_\Lambda(H,V') - E_\Theta(V') + E_\Lambda(V')}} \\
&= \sum_{H,\mathscr{E}} P_\Lambda(H,\mathscr{E}) log \frac{\sum_V e^{-E_{\Lambda,\Theta}(V,\mathscr{E},H)}}{\sum_{V'} e^{-E_{\Lambda,\Theta}(H,V')}} \\
&= \sum_{H,\mathscr{E}} P_\Lambda(H,\mathscr{E}) log \frac{\sum_V e^{-E_{\Lambda,\Theta}(V,H) - \Delta E}}{\sum_{V'} e^{-E_{\Lambda,\Theta}(H,V')}} \\
&= \sum_{H,\mathscr{E}} P_\Lambda(H,\mathscr{E}) log \langle -\Delta E \rangle_H \approx \sum_{H,\mathscr{E}} P_\Lambda(H,\mathscr{E}) \langle -\Delta E \rangle_H
\end{aligned}$$

where $\Delta E = E_{\Lambda,\Theta}(V,\mathscr{E},H) - E_{\Lambda,\Theta}(V,H)$ and the average is taken over a system with energy $E_{\Lambda,\Theta}(V,H)$ with fixed H. The above surrogate function can be rewritten further

$$A_\Lambda(H:\mathscr{E}) \approx \sum_{H,\mathscr{E}} \sum_V P(V,\mathscr{E}) P_\Lambda(H|V) \langle -\Delta E \rangle_H \tag{V.35}$$

To simplify this function further [39] replace the sums $\sum_{\mathscr{E}} \sum_V$ with the average of $N_{(\mathscr{E},V)}$

MC samples $(\mathscr{E},V)_i$, and use the Λ-RBM and a sample $(V)_i$ to draw a sample $(H(V))_i$ according to the probability distribution $P(H|V)$

$$A_\Lambda(H:\mathscr{E}) \approx \frac{1}{N_{(\mathscr{E},V,H(V))_i)}} \sum_{(V,\mathscr{E},H(V))_i} \langle -\Delta E\rangle_H \tag{V.36}$$

It should be pointed out that there are two distinct issues here : The purpose of casting RG and coarse-graining in information theoretical terms is to use the machinery of DL to solve physics problems where there is a lack of insight in either the microscopic interactions or the inner workings of the system as a whole. Here, it is argued that when we do not know of meaningful coarse-graining schemes, MIM will do the job of finding the most perfect coarse-graining scheme, provide us with the GR flow thereby giving us an insight in the whole universality class of systems that may or may not appear macroscopically similar.

The fact that this has been shown successful in proving that the principle of MIM reproduces what we already know about the RG flow and critical exponents of 1d and 2d Ising model, the passage from RG to information theory seems a bit artificial (at least in the context of RSMI). Yes, the two processes are reminiscent of each other but there is no natural reason to find the parameters of an RBM that maximize MI between the hidden units and the environment. This being said, if the goal of this investigation is to gain insights into the recent years' scaling laws, the question to answer is whether RG is the proper framework for DL, i.e. whether DL algorithms do some kind of sophisticated RG transformations which can then be used to explain the scaling laws.