Ενισχυτική Μάθηση στην Κβαντική Φυσική Πολλών Σωμάτων και μια Αντιστοιχία Ανάμεσα στην Ομάδα Επανακανονικοποίησης και τα Βαθιά Νευρωνικά Δίκτυα

Δημήτριος Σ. Μπαχτής

Τομέας Φυσικής
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών
Εθνικό Μετσόβιο Πολυτεχνείο

Περίληψη

Στόχος της διπλωματικής είναι η χρήση των νευρωνικών δικτύων ως εργαλεία έρευνας σε προβλήματα εφαρμοσμένης στατιστικής φυσικής. Επιπλέον γίνεται μια μελέτη αντιστοιχίας της συμπίεσης δεδομένων χρησιμοποιώντας τα νευρωνικά δίκτυα και της ομάδος επανακανονικοποίησης πραγματικού χώρου.

Αρχικά, τα νευρωνικά δίκτυα χρησιμοποιήθηκαν σε μια προσέγγιση μάθησης χωρίς επίβλεψη για την μοντελοποίηση της κατανομής πιθανότητας που αναπαρίσταται απο απεικονίσεις του διδιάστατου Ising. Οι απεικονίσεις αυτές έχουν δειγματοληφθεί με τεχνικές Monte Carlo μαρκοβιανων αλυσίδων χρησιμοποιώντας τον αλγοριθμο Metropolis ή τον cluster αλγόριθμο του Wolff. Το νευρωνικό δίκτυο οδηγείται σε ισορροπία ετσι ώστε να παράγει προσεγγιστικές απεικονίσεις απο τις οποίες θα υπολογιστούν παρατηρήσιμες ποσότητες για θερμοκρασίες κοντά στην μετάβαση φάσης. Οι αναμενόμενες τιμές συγκρίθηκαν με αυτές των Monte Carlo προσομοιώσεων και μελετήθηκε η εξάρτηση της ακρίβειας τους απο το πλήθος των κρυφών νευρώνων.

Ακολούθως, θεμελιώθηκε μια αντιστοιχία ανάμεσα στην Ομάδα Επανακανονικοποίησης και τα Δίκτυα Βαθιάς Πεποιθήσεως. Ενα βαθυ νευρωνικό δίκτυο εκπαιδευθηκε κοντά στην μετάβαση φάσης και σχεδιαστηκαν τα receptive fields. Το νευρωνικό δίκτυο εφαρμοζει έναν γενικευμενο μετασχηματισμό που θυμιζει μετασχηματισμό επανακανονικοποίησης πραγματικού χώρου. Η κρίσιμη θερμοκρασία και οι κρίσιμοι εκθέτες του διδιάστατου Ising υπολογίστηκαν για ένα σύστημα μεγέθους $N = 64 * 64$ χρησιμοποιώντας έναν spin-blocking μετασχηματισμό επανακανονικοποίησης. Οι παραπάνω υπολογισμοί είναι ανταγωνιστικοί σε σύγκριση με την εφαρμογή της μεθόδου βαθμισης πεπερασμένου μεγέθους λόγω της χρήσης μικρότερων πλεγμάτων

Τέλος, υλοποιήθηκε μια προσέγγιση ενισχυτικής μάθησης βασισμένη στην Variational Monte Carlo μέθοδο και στην εισαγωγή κβαντικών καταστάσεων στα Restricted Boltzmann Machines. Το νευρωνικό δίκτυο χρησιμοποιήθηκε τότε για τον προσδιορισμό της ενέργειας της θεμελιώδους κατάστασης του μονοδιάστατου Ising μοντέλου διαμήκους πεδίου και οι τιμές συγκρίθηκαν με τις ακριβείς απο διαγωνοποίηση. Στην μέθοδο δεν εμφανίζεται το πρόβλημα προσήμου κάνοντας δυνατή την αντιμετώπιση προβλημάτων στα οποια εμφανίζονταν σχετικές δυσκολίες, οπως για παράδειγμα στην ευρεση των ιδιοτήτων της θεμελιώδους κατάστασης ισχυρά αλληλεπιδρώντων φερμιονίων. Τα αποτελέσματα είναι ανταγωνιστικά συγκρινόμενα με άλλες τεχνικές απο την σχετική

βιβλιογραφία.

Abstract

This thesis was mainly driven by the need to identify what Machine Learning and Statistical Physics can offer to each other.

Initially, Restricted Boltzmann Machines were utilized in an unsupervised setting to model the probability distribution represented by importance-sampled configurations of the $d = 2$ Ising model. Configurations were then drawn from the equilibrium distribution of the neural network in order to calculate expectation values of observables near the second order phase transition. The expectation values compare well with Monte Carlo calculations and show a dependence on the number of hidden units.

A mapping was then established between the Renormalization Group and Deep Belief Networks . To gain further insights to their connection, the receptive fields of a deep neural network trained near the phase transition were visualized. The critical temperature and the critical exponents of the $d = 2$ Ising model were then estimated for a system of size $N = 64 * 64$ using a spin-blocking renormalization group transformation .

Finally, a reinforcement learning approach was implemented based on the variational Monte Carlo method and the introduction of quantum states in Restricted Boltzmann Machines. The neural network was then used to recognize the ground state of the $d = 1$ transverse-field Ising model . The results prove to be competitive when compared with other techniques from relevant literature.

Ευχαριστίες

Ολοκληρώνοντας την μεταπτυχιακή μου εργασία θα ήθελα να εκφράσω τις ευχαριστίες μου προς ένα πλήθος ατόμων.

Στον Αν. Καθηγητή Κ. Αναγνωστόπουλο για την καθοδήγηση και υποστήριξη κατα την επίβλεψη της προπτυχιακής και μεταπτυχιακής μου εργασίας. Επισης η επιλογή του να διαθέσει το βιβλίο του [16] υπο τους όρους της άδειας CC-BY-SA ήταν πηγή εμπνευσης για εμένα. Περισσότερο απο όλα θα ήθελα να τον ευχαριστήσω για το οτι είπε ¨Ναι¨ οταν τον προσέγγισα με τις ιδεες μου για μια μεταπτυχιακή εργασία και μου επέτρεψε να παρω πρωτοβουλίες σχετικά με αυτή.

Στον Καθηγητή Θ. Αλεξόπουλο και στον Επ. Καθηγητή Κ. Κουσουρή για την συμμετοχή τους στην τριμελή επιτροπή.

Στην οικογένεια μου για την υποστήριξη τους και στους φίλους μου για όλες τις χαρούμενες στιγμές που εχουμε ζήσει μαζί.

Περιεχόμενα

1. *Restricted Boltzmann Machines*

1.1.0 *Μηχανές Boltzmann*

Οι *μηχανές Boltzmann*[2, 3] είναι μια κλάση στοχαστικών νευρωνικών δικτύων με ισχυρές βάσεις στην Στατιστική Φυσική. Αντιστοιχούν σε ένα μοντέλο το οποιο αποτελείται απο στοχαστικά στοιχεία-που ονομάζονται και νευρώνες-και αναπαριστά μια κατανομή πιθανότητας. Οι μηχανές Boltzmann περιγράφονται απο ένα σύνολο παραμετρων μεταβολής και μπορούν να χρησιμοποιηθούν για την αναγνώριση χαρακτηριστικών μιας *άγνωστης* κατανομής πιθανότητας. Το νευρωνικό δίκτυο εκπαιδεύεται σε ένα σύνολο απο δεδομένα, μεταβάλλοντας με κατάλληλο τρόπο μεσω μιας διαδικασίας τις παραμέτρους μεταβολής σε κάθε επανάληψη. Εν τέλει, με την χρήση των νευρωνικών δικτύων μια αγνωστη κατανομή πιθανότητας μπορεί να αναπαρασταθεί σε κλειστή μορφή.

Οι *Restricted Boltzmann Machines* είναι μια ειδική περίπτωση των μηχανών Boltzmann. Αποτελούνται απο ένα σύνολο *ορατών* και *κρυφών* νευρώνων και αναπαρίστανται γραφικά σε δύο επίπεδα. Νευρώνες μεταξύ των δύο επιπέδων συνδέονται με μη κατευθυνόμενο τρόπο ενώ δεν επιτρέπεται σύνδεση δυο νευρώνων που βρίσκονται στο ίδιο επίπεδο. Οι ορατοι νευρώνες χρησιμοποιούνται ως είσοδος για τα δεδομένα στο νευρωνικό δίκτυο κατα την εκμάθηση μιας κατανομής. Ταυτόχρονα χρησιμοποιούνται και ως έξοδος για την παραγωγή δειγμάτων απο την κατανομή ισορροπίας του νευρωνικού δικτύου. Οι κρυφοί νευρώνες αναγνωρίζουν με μη γραμμικό τρόπο χαρακτηριστικά και εξαρτησεις που υπάρχουν στην κατανομή κατα την εκπαιδευση.

Οι *Restricted Boltzmann Machines* μπορούν επίσης να προσφέρουν μια λύση σε ημιτελείς παρατηρήσεις. Για παράδειγμα, υπάρχει η δυνατότητα να θέσει κάποιος τμήμα των ορατων νευρώνων ίσο με τις παρατηρήσεις και να δειγματοληπτήσει σε περιθώριες κατανομές για να συμπληρώσει τα υπολειπόμενα στοιχεία. Είναι πολύ σημαντικό, οτι κάποιος μπορεί να χρησιμοποιήσει μια συστοιχία απο *Restricted Boltzmann Machines* για να σχηματίσει ένα βαθύ νευρωνικό δίκτυο το οποιο ονομάζεται Deep Belief Network.

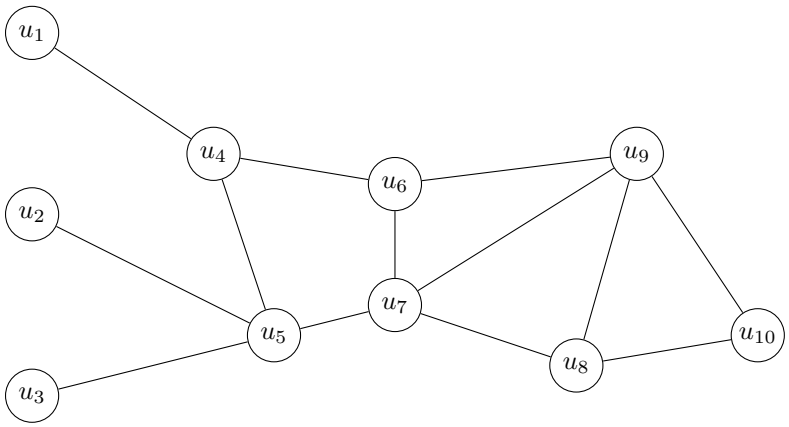
Μια μελέτη ακολουθεί για τις Restricted Boltzmann μηχανές [1] βασισμένη στα πιθανοτικα γραφικά μοντέλα και συγκεκριμένα στα τυχαία μαρκοβιανά πεδία. Αυτή η αυστηρή μαθηματική προσέγγιση επιτρέπει την χρήση θεωρημάτων και αλγορίθμων για μια καλά ορισμένη περιγραφή των *Restricted Boltzmann Machines*.

1.2.0 *Γραφικά Μοντέλα και Τυχαία Μαρκοβιανά Πεδία*

Το πλαίσιο των πιθανοτικών γραφικών μοντέλων απλοποιεί την μελέτη τυχαίων μεταβλητων οι οποίες περιγράφονται απο υπο συνθήκη ιδιότητες ανεξαρτησίας και εξάρτησης με την χρήση γραφημάτων.

Η υπο συνθήκη ανεξαρτησία ανάμεσα σε δύο σύνολα απο τυχαίες μεταβλητές $\boldsymbol{X}$ και $\boldsymbol{Y}$ ορίζεται ως ανεξαρτησία σε όρους ενός επιπλέον συνόλου $\boldsymbol{Z}$ για όλες τις τιμές των $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$. Ισοδύναμα, σε μαθηματικό φορμαλισμό:

$$p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z}) \Rightarrow p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z}) \; and \; p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{y}|\boldsymbol{z}) \tag{1.1}$$



Σχήμα 1.1: Ενα μη κατευθυνόμενο γράφημα $G = (V, E)$. Η γειτονιά του κόμβου u_7 είναι η $\{u_5, u_6, u_8, u_9\}$. Τα υποσύνολα $\{u_9, u_{10}\}$ και $\{u_8, u_9, u_{10}\}$ ορίζουν κλίκες με την $\{u_8, u_9, u_{10}\}$ να είναι μεγιστοτική. Οι κόμβοι u_1 και u_{10} διαχωρίζονται απο τους $\{u_6, u_7\}$.

Ενα μη κατευθυνόμενο γράφημα $G = (V, E)$ είναι μια δομή δεδομένων που αποτελείται απο ένα σύνολο κόμβων V και ένα σύνολο απο μη κατευθυνόμενες ακμές E. Οι μη κατευθυνόμενες ακμές συνδέουν ζευγη απο κόμβους και αναπαριστώνται ως $X_i - X_j$. Είναι δυνατό να οριστεί μια γειτονιά $\mathcal{N}_u$ που αποτελείται απο κόμβους συνδεόμενους με έναν δεδομένο κομβο u:

$$\mathcal{N}_u = \{w \in V : \{w, u\} \in E\} \tag{1.2}$$

Μια κλίκα σε ένα μη κατευθυνόμενο γράφημα $G = (V, E)$ είναι ένα υποσύνολο του V για το οποίο όλοι οι συμπεριλαμβανόμενοι κόμβοι συνδέονται σε ζευγη. Μια δεδομένη κλίκα αποκαλείται μεγιστοτική αν ικανοποιεί την παραπάνω συνθήκη και ταυτόχρονα δεν μπορεί να επεκταθεί με την προσθήκη ενός κόμβου. Ενα μονοπάτι ανάμεσα σε δύο κομβους u_1 και u_m ορίζεται ως μια πεπερασμενη ακολουθία απο κόμβους $\{u_1, u_2, \ldots, u_m \in V\}, \{u_i, u_{i+1}\} \in E, i = 1, \ldots, m-1$. Αν υποθέσουμε ένα σύνολο $\mathcal{V} \subset V$, δύο κόμβοι $u \notin \mathcal{V}$ και $w \notin \mathcal{V}$ διαχωρίζονται αν όλα τα μονοπάτια απο τον u στον w περιλαμβάνουν ένα κόμβο απο το $\mathcal{V}$.

Ας υποθέσουμε ένα μη κατευθυνόμενο γράφημα $G = (V, E)$, για το οποίο σε κάθε κόμβο αντιστοιχεί μια τυχαία μεταβλητή X_u η οποία παίρνει τιμές σε ένα χώρο καταστάσεων $\Lambda_u = \Lambda$. Αν η απο κοινού συνάρτηση κατανομής

ικανοποιεί μια τοπική Markov ιδιότητα, το σύνολο απο τις τυχαίες μεταβλητές $\boldsymbol{X} = (X_u)_{u \in V}$ ορίζουν ένα τυχαίο μαρκοβιανό πεδίο. Η τοπική Markov ιδιότητα υποδηλώνει οτι η υπο συνθήκη κατανομή μιας τυχαίας μεταβλητής είναι ανεξάρτητη για όλες τις άλλες μεταβλητές σε μια αντίστοιχη γειτονιά. Ισοδύναμα:

$$p(x_u|(x_w)_{w \in V \setminus \{u\}}) = p(x_u|(x_w)_{w \in \mathcal{N}_u}), \forall u \in V \; and \; \forall \boldsymbol{x} \in \Lambda^{|V|} \tag{1.3}$$

Θεωρώντας μια αυστηρά θετική κατανομή πιθανότητας, είναι δυνατό να θεμελιωθεί μια ισοδυναμία ανάμεσα στην τοπική Markov ιδιότητα και άλλες Markov ιδιότητες.

Η ολική Markov ιδιότητα ικανοποιείται αν για τρία ξένα υποσύνολα $\mathcal{A}, \mathcal{B}, \mathcal{S} \subset V$, με το $\mathcal{S}$ να διαχωρίζει τους κόμβους στα $\mathcal{A}, \mathcal{B}$, οι τυχαίες μεταβλητές $(X_a)_{a \in \mathcal{A}}$ και $(X_b)_{b \in \mathcal{B}}$ είναι υπο συνθήκη ανεξάρτητες σε όρους της $(X_s)_{s \in \mathcal{S}}$. Ισοδύναμα:

$$p((x_a)_{a \in \mathcal{A}}|(x_t)_{t \in S \cup \mathcal{B}}) = p((x_a)_{a \in \mathcal{A}}|(x_t)_{t \in S}) \tag{1.4}$$

Η ανα δύο Markov ιδιότητα ικανοποιείται αν δύο κόμβοι οι οποίοι δεν είναι γειτονικοί είναι υπο συνθήκοι ανεξάρτητοι σε όρους όλων των άλλων μεταβλητών. Ισοδύναμα αν $\{u, w\} \notin E$ για $\forall \boldsymbol{x} \in \Lambda^{|V|}$:

$$p(x_u, x_w|(x_t)_{t \in V \setminus \{u,w\}}) = p(x_u|(x_t)_{t \in V \setminus \{u,w\}}) p(x_w|(x_t)_{t \in V \setminus \{u,w\}}) \tag{1.5}$$

Ειναι λογικό να αναρωτηθεί κάποιος αν είναι δυνατό να παραγοντοποιηθούν οι κατανομές των τυχαίων μαρκοβιανών πεδίων λόγω της στενής συνδεσης της παραγοντοποίησης των απο κοινού κατανομών πιθανοτήτων και της υπο συνθήκης ανεξαρτησίας για ένα σύνολο τυχαίων μεταβλητών.

Theorem 1 (Hammersley-Clifford). *Μια αυστηρά θετική κατανομή p ικανοποιεί την Markov ιδιότητα για ένα μη κατευθυνόμενο γράφημα $G = (V, E)$ αν και μόνο αν η p παραγοντοποιείται σύμφωνα με το G. [4, 5]*

Για την παραγοντοποίηση μιας κατανομής πιθανότητας ενος μη κατευθυνόμενο γραφήματος με $\mathcal{C}$ μεγιστοτικες κλίκες, ένα σύνολο απο μη αρνητικές συναρτήσεις $\{\psi_C\}_{C \subset \mathcal{C}}$ πρέπει να ικανοποιεί:

$$\forall \boldsymbol{x}, \hat{\boldsymbol{x}} \in \Lambda^{|V|} : (x_c)_{c \in C} = (\hat{x}_c)_{c \in C} \Rightarrow \psi_C(\boldsymbol{x}) = \psi_C(\hat{bmx}) \tag{1.6}$$

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}) \tag{1.7}$$

Ορίζουμε την σταθερά κανονικοποίησης Z ως την συνάρτηση επιμερισμού:

$$Z = \sum_x \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C) \tag{1.8}$$

Για μια αυστηρά θετική κατανομή p, οι συναρτήσεις $\{\psi_C\}_{C \subset \mathcal{C}}$ είναι επίσης θετικές και:

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C) = \frac{1}{Z} e^{\sum_{C \in \mathcal{C}} \ln \psi_C(\boldsymbol{x}_C)} = \frac{1}{Z} e^{-E(\boldsymbol{x})} \tag{1.9}$$

Η συνάρτηση E είναι η συνάρτηση ενέργειας:

$$E = \sum_{C \in \mathcal{C}} \ln \psi_C(\boldsymbol{x}_C) \tag{1.10}$$

Η αυστηρά θετική κατανομή πιθανότητας p κάθε τυχαίου μαρκοβιανού πεδίου είναι τοτε η συνάρτηση Gibbs.

1.2.1 Μαθηση Χωρίς Επίβλεψη Τυχαίων Μαρκοβιανών Πεδίων

Στόχος της μάθησης χωρίς επίβλεψη είναι η μοντελοποίηση μιας άγνωστης κατανομής q που αναπαρίσταται απο ένα σύνολο μη επισημασμένων δεδομένων εκπαίδευσης. Ας υποθέσουμε ένα δεδομένο γραφικό μοντέλο με μια συνάρτηση ενέργειας που εξαρτάται σε ένα σύνολο απο παραμέτρους μεταβολής θ. Η μάθηση χωρίς επίβλεψη αντιστοιχεί στην μεταβολή αυτών των παραμέτρων με στόχο την αναπαράσταση της κατανομής q. Ισοδύναμα επιθυμούμε η κατανομή πιθανότητας του μοντέλου p να είναι επακριβώς ιδια με την κατανομή q. Θα χρησιμοποιηθεί ο συμβολισμός $p(\boldsymbol{x}|\boldsymbol{\theta})$ για να εκφράσει την εξάρτηση της αντίστοιχης κατανομής στις παραμέτρους μεταβολής $\boldsymbol{\theta}$.

Είναι δυνατό να εκτιμηθούν οι παράμετροι ενός μοντέλου με την εκτίμηση μέγιστης πιθανοφάνειας. Για ένα δεδομένο σύνολο απο ανεξάρτητα δεδομένα $S = \{x_i\}$ απο μια άγνωστη κατανομή q οι παράμετροι $\boldsymbol{\theta}$ μεταβάλλονται με στόχο την μεγιστοποίηση της πιθανότητας των S για την κατανομή του τυχαίου μαρκοβιανού πεδίου.

Η πιθανοφάνεια $\mathcal{L} : \Theta \to \mathcal{R}$ προσφέρει μία αντιστοιχία για τις παραμέτρους μεταβολής $\boldsymbol{\theta}$ απο έναν χώρο παραμέτρων Θ σε:

$$\mathcal{L}(\boldsymbol{\theta}|S) = \prod_{i=1}^{l} p(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{1.11}$$

Η ιδέα είναι να βρεθούν παράμετροι $\boldsymbol{\theta}$ που θα μεγιστοποιούν την πιθανοφάνεια του δεδομένου συνόλου εκπαίδευσης. Ισοδύναμα, μπορει να μεγιστοποιηθεί η λογαριθμική πιθανοφάνεια:

$$\ln \mathcal{L}(\boldsymbol{\theta}|S) = \ln \prod_{i=1}^{l} p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{i}^{l} \ln p(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{1.12}$$

Υποθέτοντας οτι η κατανομή του τυχαίου μαρκοβιανού πεδίου είναι η κατανομή Boltzmann δεν είναι δυνατό να βρεθεί μια αναλυτική λύση για ενδιαφέροντα προβλήματα, και πρέπει να χρησιμοποιηθεί μια προσεγγιστική τεχνική όπως η gradient ascent.

Οπως αναφέρθηκε προηγουμένως, η μάθηση χωρίς επίβλεψη αντιστοιχεί σε μια μεταβολη των παραμέτρων μεταβολής θ με στόχο την ακριβή αντιστοίχιση της κατανομής του μοντέλου p στην κατανομή πιθανότητας q που αναπαρίσταται απο ένα σύνολο απο δεδομένα εκπαίδευσης S. Ισοδύναμα, επιθυμούμε την ελαχιστοποίηση της απόστασης ανάμεσα σε δύο κατανομές. Αυτό είναι δυνατό με την ελαχιστοποιηση της Kullback-Leibler απόκλισης, δηλαδη της σχετικής εντροπίας, η οποία για ένα πεπερασμένο χώρο είναι:

$$KL(q||p) = \sum_{\boldsymbol{x}\in\Omega} q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} = \sum_{\boldsymbol{x}\in\Omega} q(\boldsymbol{x}) \ln q(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\Omega} q(\boldsymbol{x}) \ln p(\boldsymbol{x}) \quad (1.13)$$

Εφόσον η $q(x)$ μπορεί να εκτιμηθεί απο το σύνολο δεδομένων εκπαίδευσης S η Kullback-Leibler απόκλιση μπορεί να εκφραστεί σε όρους της λογαριθμικής πιθανοφάνειας μέσα απο τον ορο $-\sum_{x\in\Omega} q(x) \ln p(x)$ με μια εξάρτηση στις παραμέτρους μεταβολής θ. Η Kullback-Leibler απόκλιση είναι μια θετική ποσότητα και είναι μηδενική για την περίπτωση που δύο κατανομές είναι ακριβώς ίδιες. Αρα, η μεγιστοποίηση της λογαριθμικής πιθανοφάνειας μπορεί να επιτευχθεί με την ελαχιστοποίηση της Kullback-Leibler απόκλισης.

Για να βρεθούν παράμετροι μεταβολής που μεγιστοποιούν την λογαριθμική πιθανοφάνεια, πρεπει να χρησιμοποιηθεί ένας αλγόριθμος πρώτης τάξης που αποκαλείται gradient descent. Η gradient descent μεταβάλλει τις παραμέτρους $\boldsymbol{\theta}^{(t)}$ το $\boldsymbol{\theta}^{(t+1)}$ σε κάθε βήμα σύμφωνα με την σχέση:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \frac{\partial}{\partial \boldsymbol{\theta}^{(t)}} \Big(\ln \mathcal{L}(\boldsymbol{\theta}^{(t)}|S) \Big) - \lambda \boldsymbol{\theta}^{(t)} + \nu \Delta \boldsymbol{\theta}^{(t-1)} \quad (1.14)$$

Η ποσότητα $\eta \in \mathbb{R}_0^+$ είναι ο ρυθμος μάθησης. Η προαιρετική ποσότητα λ είναι ο όρος weight decay, που ωθεί τα βάρη σε μικρότερες τιμές. Επιπλέον, η προαιρετική ποσότητα $\nu\Delta\theta^{(t-1)}$ αποκαλείται όρος ορμής και οδηγεί σε γρηγορότερη εκπαίδευση.

Υποθέτουμε την μοντελοποίηση μιας m-διάστασης κατανομής με ένα τυχαίο μαρκοβιανό πεδίο το οποίο αποτελείται απο εναν αριθμό κομβων μεγαλύτερο απο m. Είναι δυνατό να χωριστούν οι μεταβλητές $\boldsymbol{X} = (X_u)_{u\in V}$ σε ορατούς $\boldsymbol{V} = (V_1, V_2, \ldots, V_m)$ και κρυφούς $\boldsymbol{H} = (H_1, H_2, \ldots, H_n)$ νευρώνες, οπου $n = |V| - m$.

Η προσθήκη κρυφών νευρώνων επιτρέπει μια καλύτερη περιγραφή της άγνωστης κατανομής αφού συσχετίσεις ανάμεσα στα δεδομένα μπορούν να εκφραστούν μέσα απο υπο συνθήκη κατανομές. Η κατανομή πιθανότητας Boltzmann του τυχαίου μαρκοβιανού πεδίου είναι τότε μια απο κοινού κατανομή πιθανοτητας πάνω στους ορατούς και κρυφούς νευρώνες $(\boldsymbol{V}, \boldsymbol{H})$. Η περιθώρια κατανομή των $\boldsymbol{V}$ είναι μια άθροιση πάνω στους κρυφούς νευρώνες της απο κοινού κατανομής πιθανότητας:

$$p(u) = \sum_h p(\boldsymbol{u}, \boldsymbol{h}) = \frac{1}{Z} \sum_h e^{-E(\boldsymbol{u},\boldsymbol{h})} \quad (1.15)$$

$$Z = \sum_{\boldsymbol{u},\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})}, \tag{1.16}$$

οπου Z η συνάρτηση επιμερισμού. Αν υποθέσουμε ενα δεδομένο παράδειγμα εκπαίδευσης u απο ένα σύνολο S, η λογαριθμική πιθανοφάνεια είναι:

$$\begin{aligned} \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u}) = \ln p(\boldsymbol{u}|\boldsymbol{\theta}) &= \ln \left(\frac{1}{Z} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})} \right) \\ &= \ln \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})} - \ln \sum_{\boldsymbol{u},\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})} \end{aligned} \tag{1.17}$$

Η παράγωγος της λογαριθμικής πιθανοφάνειας είναι:

$$\begin{aligned} \frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}} \left(\ln \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})} \right) - \frac{\partial}{\partial \boldsymbol{\theta}} \left(\ln \sum_{\boldsymbol{u},\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})} \right) \\ &= -\frac{1}{\sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})}} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})} \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \theta} \\ &\quad + \frac{1}{\sum_{\boldsymbol{u},\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})}} \sum_{\boldsymbol{u},\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})} \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \boldsymbol{\theta}} \end{aligned} \tag{1.18}$$

Δεδομένου οτι η απο συνθήκη πιθανότητα είναι:

$$p(\boldsymbol{h}|\boldsymbol{u}) = \frac{p(\boldsymbol{u},\boldsymbol{h})}{p(\boldsymbol{u})} = \frac{\frac{1}{Z} e^{-E(\boldsymbol{u},\boldsymbol{h})}}{\frac{1}{Z} \sum\limits_{\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})}} = \frac{e^{-E(\boldsymbol{u},\boldsymbol{h})}}{\sum\limits_{\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})}} \tag{1.19}$$

Η παράγωγος της λογαριθμικής πιθανοφάνειας ισούται με:

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial \boldsymbol{\theta}} = -\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \theta} + \sum_{\boldsymbol{u},\boldsymbol{h}} p(\boldsymbol{u},\boldsymbol{h}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \theta} \tag{1.20}$$

Ειναι σημαντικό να παρατηρηθεί οτι οι δύο όροι αντιστοιχούν στις αναμενόμενες τιμές της συνάρτησης ενέργειας για την απο κοινού κατανομή πιθανότητας του μοντέλου και της συνάρτησης ενέργειας για υπο συνθήκη κατανομή των κρυφών νευρώνων με δεδομένο παράδειγμα εκπαίδευσης. Άρα, ειναι πιθανό να υπολογιστούν οι αναμενόμενες τιμές δειγματοληπτώντας ένα αντιπροσωπευτικό υποσύνολο απο τις αντίστοχες κατανομές χρησιμοποιώντας Monte Carlo μαρκοβιανών αλυσίδων.

1.2.2 *Μαρκοβιανές Αλυσίδες Διακριτού Χρόνου*

Ας υποθέσουμε μια ακολουθία απο διακριτές τυχαίες μεταβλητές $\{X^k \big| \ k \in N_0\}$ που παίρνουν τιμές σε έναν χώρο καταστάσεων Ω και για τις οποίες $\forall\, k \geq$

0 και $\forall j, i, i_0, \ldots, i_{k-1} \in \Omega$ ικανοποιούν την Markov ιδιότητα:

$$\begin{aligned} p_{ij}^k &= Pr\Big(X^{(k+1)} = j \big| X^{(k)} = i, X^{(k-1)} = i_{k-1}, \ldots X^{(0)} = 0\Big) \\ &= Pr\Big(X^{(k+1)} = j \big| X^{(k)} = i\Big) \end{aligned} \tag{1.21}$$

Η ακολουθία $\{X^k \big|\ k \in N_0\}$ τότε ορίζει μια αλυσίδα Markov. Η Markov ιδιότητα δηλώνει οτι η διακριτή τυχαία μεταβλητή X_{k+1} εξαρτάται μόνο απο την X_k και όχι απο τις $X_{k-1}, \ldots, X_1, X_0$ αρα η μαρκοβιανή αλυσίδα δεν έχει μνήμη.

Μια μαρκοβιανή αλυσίδα είναι ομογενής στον χρόνο αν οι πιθανότητες μετάβασης είναι χρονικά ανεξάρτητες . Ισοδύναμα αν για $k \geq 0$, $p_{ij}^{(k)} = p_{ij}$. Μια ομογενής μαρκοβιανή αλυσίδα περιγράφεται απο έναν πίνακα μετάβασης $\boldsymbol{P} = (p_{ij})_{i,j \in \Omega}$.

Αν υποθέσουμε οτι η κατανομή πιθανότητας της κατάστασης $X_{(0)}$ δίνεται απο ένα διάνυσμα πιθανότητας $\boldsymbol{\mu^{(0)}} = (\mu^{(0)}(i))_{i \in \Omega}$ με $\mu^{(0)}(i) = Pr\Big(X^{(0)} = i\Big)$, η κατανομή πιθανότητας $\boldsymbol{\mu^{(k)}}$ της τυχαίας μεταβλητής $X^{(k)}$ δίνεται απο την:

$$\boldsymbol{\mu^{(k)T}} = \boldsymbol{\mu^{(0)T}} \boldsymbol{P^k} \tag{1.22}$$

Για να μετακινηθεί η αλυσίδα κατα k βήματα αρκεί ο πολλαπλασιασμός με $\boldsymbol{P}^k$ σύμφωνα με την παραπάνω εξίσωση.

Ορίζεται μια κατανομή ισορροπίας π για την μαρκοβιανή αλυσίδα αν:

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \boldsymbol{P} \tag{1.23}$$

Αν η μαρκοβιανή αλυσίδα βρεθεί σε χρόνο k στην κατανομή ισορροπίας $\boldsymbol{\mu}^k = \boldsymbol{\pi}$ τότε μένει εκεί για πάντα, δηλαδή για όλες τις υπόλοιπες καταστάσεις ισχύει $\boldsymbol{\mu}^{(k+n)} = \boldsymbol{\pi}$, $\forall n \in N$. Μια κατανομή π είναι κατανομή ισορροπίας αν οι πιθανότητες μετάβασης $p_{ij}, i, j \in \Omega$ ικανοποιούν την συνθήκη λεπτομερούς ισορροπίας.

$$\pi(i) p_{ij} = \pi(j) p_{ji},\ \forall\, i, j \in \Omega \tag{1.24}$$

Μια μαρκοβιανή αλυσίδα είναι μη υποβιβάσιμη αν κάθε κατάσταση είναι προσβάσιμη μέτα απο ένα σύνολο πεπερασμένων μεταβάσεων:

$$\forall i, j \in \Omega\ \exists\, k > 0\ with\ Pr\Big(X^{(k)} = j \big| X^{(0)} = i\Big) > 0 \tag{1.25}$$

Ορίζεται ως περίοδος $d(i)$ μιας κατάστασης i ο μέγιστος κοινός διαιρέτης gcd:

$$d(i) = gcd\{k \in N_0 \big| Pr(X^{(k)} = i | X^{(0)} = i > 0)\} \tag{1.26}$$

Αν $d(i) = 1$ για όλες τις καταστάσεις $i \in \Omega$ τότε η μαρκοβιανή αλυσίδα είναι απεριοδική που σημαίνει οτι μπορεί να επιστρέφει σε μια δεδομένη κατάσταση σε ακανόνιστα χρονικά βήματα.

Η ολική απόσταση μεταβολής ανάμεσα σε δύο κατανομές α και β σε έναν πεπερασμένο χώρο πιθανοτήτων Ω είναι:

$$d_V(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}|\boldsymbol{\alpha} - \boldsymbol{\beta}| = \frac{1}{2}\sum_{x\in\Omega}|\alpha(x) - \beta x| \tag{1.27}$$

Μια μαρκοβιανή αλυσίδα που είναι μη υποβιβάσιμη και απεριοδική και για την οποία υπάρχει μια κατανομή ισορροπίας $\boldsymbol{\pi}^T$ θα συγκλίνει στην $\boldsymbol{\pi}^T$ καθώς $k \to \infty$. Πιο συγκεριμένα, θεωρώντας μια αυθαίρετη αρχική κατανομή μ:

$$\lim_{k\to\infty} d_V(\boldsymbol{\mu}^T \boldsymbol{P}^k, \boldsymbol{\pi}^T) = 0. \tag{1.28}$$

Στόχος είναι η κατασκευή μια μαρκοβιανής αλυσίδας που συγκλίνει ασυμπτωτικά στην επιθυμητή κατανομή ισορροπίας έτσι ώστε να δειγματοληφθεί ένα υποσύνολο καταστάσεων απο αυτή. Αυτά τα δειγματα χρησιμοποιούνται για τον υπολογισμό των παρατηρήσιμων ποσοτήτων.

1.2.3 Δειγματοληψία Gibbs

Η δειγματοληψία Gibbs είναι μια τεχνική Monte Carlo μαρκοβιανών αλυσίδων και μπορεί να θεωρηθεί ως μια ειδική περίπτωση του αλγόριθμου Metropolis-Hastings. Η ιδέα είναι να επιλεγεί τυχαία μια κατάσταση απο μια προτεινόμενη κατανομή και να γίνει αποδεκτή σύμφωνα με μια πιθανότητα αποδοχής ενώ ικανοποιείται η συνθήκη λεπτομερούς ισορροπίας.

Ορίζουμε ένα τυχαίο μαρκοβιανό πεδίο που περιγράφεται απο ένα μη κατευθυνόμενο γράφημα $G = (V, E)$, $V = \{1, \ldots, N\}$ με ένα σύνολο απο τυχαίες μεταβλητές $\boldsymbol{X} = (X_1^{(k)}, \ldots, X_N^{(k)})$, $X_i, i \in V$ που παίρνουν τιμές σε ένα πεπερασμένο σύνολο Λ και μια απο κοινού κατανομή για τις $\boldsymbol{X}$ ίση με $\pi(\boldsymbol{x}) = \frac{1}{Z}e^{-\mathcal{E}(\boldsymbol{x})}$.

Υποθέτουμε οτι το τυχαίο μαρκοβιανό πεδίο εξελίσσεται στον χρόνο με διακριτά βήματα και η κατάσταση του δίνεται απο την $\boldsymbol{X}^{(k)} = (X_1^{(k)}, \ldots, X_N^{(k)})$ για χρόνο $k \geq 0$. Αυτή η διακριτή στον χρόνο εξέλιξη μπορεί να θεωρηθεί ως μια μαρκοβιανή αλυσίδα $X = \{\boldsymbol{X}^{(k)} | k \in \mathbb{N}_0\}$ με χώρο καταστάσεων $\Omega = \Lambda^N$.

Μια νέα κατάσταση προτείνεται επιλέγοντας μια τυχαία μεταβλητή X_i, $i \in V$ με μια πιθανότητα q_i. Γίνεται αποδεκτή τότε σύμφωνα με την υπο συνθήκη κατανομή πιθανότητας για δεδομένη κατάσταση $(x_u)_{u\in V\setminus i}$ των υπολοιπόμενων τυχαίων μεταβλητών $(X_u)_{u\in V\setminus i}$. Η τοπική ιδιότητα Markov του τυχαίου μαρκοβιανού πεδίου συνεπάγεται οτι $\pi(\boldsymbol{x_i}|(x_u)_{u\in V\setminus i}) = \pi(\boldsymbol{x_i}|(x_w)_{w\in\mathcal{N}_i})$. Οι πιθανότητες μετάβασης ανάμεσα σε δύο διαφορετικές καταστάσεις $\boldsymbol{x}, \boldsymbol{y}$ με $\boldsymbol{x} \neq \boldsymbol{y}$ ορίζεται ως $p_{\boldsymbol{xy}}$:

$$p_{\boldsymbol{xy}} = \begin{cases} q(i)\pi(y_i|(x_u)_{u\in V\setminus i}), & if\ \exists i \in V\ \forall u \in V\ u \neq i : x_u = y_u \\ 0, & else \end{cases}$$

Η πιθανότητα παραμονής στην ίδια κατάσταση είναι $p_{\boldsymbol{xx}} = q(i)\pi(x_i|(x_u)_{u\in V\setminus i})$. Αν η μαρκοβιανή αλυσίδα είναι μη υποβιβάσιμη και απεριοδική θα συγκλίνει στην κατανομή ισορροπίας και αν η συνθήκη λεπτομερούς ισορροπίας ικανοποιείται τότε η π είναι η επιθυμητή κατανομη ισορροπίας.

Αρχικά πρέπει να δειχθεί οτι ικανοποιείται η συνθήκη λεπτομερούς ισορροπίας. Για την περίπτωση $\boldsymbol{x} = \boldsymbol{y}$ είναι εμφανές. Οταν οι $\boldsymbol{x}$ και $\boldsymbol{y}$ διαφέρουν σε παραπάνω απο μια τυχαία μεταβλητή $p_{\boldsymbol{xy}} = p_{\boldsymbol{yx}} = 0$. Όταν διαφέρουν ακριβώς για μια τυχαία μεταβλητή X_i έχουμε:

$$\begin{aligned} \pi(\boldsymbol{x})p_{\boldsymbol{xy}} &= \pi(\boldsymbol{x})q(i)\pi(y_i|(x_u)_{u\in V\setminus i}) \\ &= \pi(x_i,(x_u)_{u\in V\setminus i})q(i)\frac{\pi(y_i,(x_u)_{u\in V\setminus i})}{\pi((x_u)_{u\in V\setminus i})} \\ &= \pi(y_i,(x_u)_{u\in V\setminus i})q(i)\frac{\pi(x_i,(x_u)_{u\in V\setminus i})}{\pi((x_u)_{u\in V\setminus i})} \\ &= \pi(\boldsymbol{y})q(i)\pi(x_i|(x_u)_{u\in V\setminus i}) \\ &= \pi(\boldsymbol{y})p_{\boldsymbol{xy}} \end{aligned} \tag{1.29}$$

Άρα η συνθήκη λεπτομερούς ισορροπίας ικανοποιείται και η π είναι η κατανομή ισορροπίας. Για να αποδειχθεί οτι η μαρκοβιανή αλυσίδα είναι μη υποβιβάσιμη παρατηρείται ότι εφόσον η π ειναι αυστηρά θετική, τότε οι υπο συνθήκη κατανομές ειναι επίσης θετικές και κάθε πιθανή κατάσταση του τυχαιου μαρκοβιανού πεδίου είναι προσβάσιμη. Επιπλέον, η μαρκοβιανή αλυσίδα ειναι απεριοδική εφόσον $p_{\boldsymbol{xx}} > 0\ \forall \boldsymbol{x} \in \Lambda^n$.

Η ντετερμινιστική επιλογή μια κατάστασης αντιστοιχεί στον periodic Gibbs sampler αλγόριθμο οπου ένα φράγμα μπορεί να οριστεί για τον ρυθμό σύγκλισης:

$$|\boldsymbol{\mu}\boldsymbol{P}^k - \boldsymbol{\pi}| \leq \frac{1}{2}|\boldsymbol{\mu} - \boldsymbol{\pi}|(1 - e^{-N\Delta})^k \tag{1.30}$$

οπου $\boldsymbol{P}$ είναι ο πίνακας μεταβάσεων, $\Delta = \sup_{l\in V}\delta_l$, $\delta_l = \sup\{|\mathcal{E}(\boldsymbol{x}) - \mathcal{E}(\boldsymbol{y})|; x_i = y_i \forall i \in V\ with\ i \neq l\}$, μ είναι μια αυθαίρετη κατανομή και $\frac{1}{2}|\boldsymbol{\mu} - \boldsymbol{\pi}|$ η ολική απόσταση μεταβολής.

1.3.0 Restricted Boltzmann Machines

Έχοντας θεμελιώσει τα τυχαία μαρκοβιανά πεδία είναι δυνατό να ορίσουμε τα Restricted Boltzmann Machines ως τυχαία μαρκοβιανά πεδία που αντιστοιχούν σε διμερή γραμφήματα με μη κατευθυνόμενες ακμές. Αποτελούνται απο m ορατούς νευρώνες $\boldsymbol{V} = \big(V_1, \ldots, V_m\big)$ και n κρυφούς νευρώνες $\boldsymbol{H} =$

$(H_1, \ldots, H_m)$. Δεδομένου οτι στην διπλωματική μελετάται το διδιάστατο Ising και το μονοδιάστατο Ising διαμήκους πεδίου, δηλαδή μοντέλα δυαδικών τιμών, τα Restricted Boltzmann Machines παίρνουν δυαδικές τιμές. Οι τυχαίες μεταβλητές $(\boldsymbol{V}, \boldsymbol{H})$ τότε παίρνουν τιμές $(\boldsymbol{u}, \boldsymbol{h}) \in \{0,1\}^{m+n}$. Η απο κοινού κατανομή πιθανότητας του μοντέλου είναι η κατανομή Boltzmann:

$$p(\boldsymbol{u}, \boldsymbol{h}) = \frac{1}{Z} e^{-E(\boldsymbol{u}, \boldsymbol{h})} \tag{1.31}$$

Η συνάρτηση ενέργειας του μοντέλου $E(\boldsymbol{u}, \boldsymbol{h})$ ορίζεται ως:

$$E(\boldsymbol{u}, \boldsymbol{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} h_i u_j - \sum_{j=1}^{m} b_j u_j - \sum_{i=1}^{n} c_i h_i \tag{1.32}$$

Οι παράμετροι μεταβολής του μοντέλου για $i \in \{1, \ldots, n\}$ και $j \in \{1, \ldots, m\}$ είναι τα βάρη και τα biases b_j και c_i για j ορατό και i κρυφό νευρώνα. Ολες οι παράμετροι μεταβολής παίρνουν πραγματικές τιμές. Το σύνολο των ορατων και κρυφών νευρώνων ορίζει ενα ορατό και κρυφό επίπεδο.

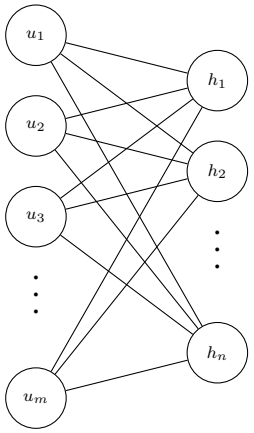
Στα Restricted Boltzmann Machines δεν επιτρέπονται συνδέσεις ανάμεσα σε νευρώνες που ανήκουν στο ίδιο επίπεδο. Αυτό συνεπάγεται μια υπο συνθήκη ανεξαρτησία για στοιχεία σε δεδομένο επίπεδο, δεδομένων των στοιχείων στο άλλο επίπεδο. Σε μαθηματικό συμβολισμό:

$$p(\boldsymbol{h}|\boldsymbol{u}) = \prod_{i=1}^{n} p(h_i|u) \tag{1.33}$$

$$p(\boldsymbol{u}|\boldsymbol{h}) = \prod_{j=1}^{m} p(u_j|\boldsymbol{h}) \tag{1.34}$$

Άρα είναι δυνατό να δειγματοληφθούν οι τιμές των νευρώνων ενός δεδομένου επιπέδου σε ένα βήμα. Συνολικά χρειάζονται δύο βήματα για την δειγματοληψία ορατων και κρυφών νευρώνων. Η περιθώρια κατανομή των νευρώνων στο ορατό επίπεδο είναι:

$$\begin{aligned}
p(\boldsymbol{u}) &= \sum_{h} p(\boldsymbol{u}, \boldsymbol{h}) = \frac{1}{Z}\sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u}, \boldsymbol{h})} \\
&= \frac{1}{Z}\sum_{h_1}\sum_{h_2}\cdots\sum_{h_n} e^{\sum_{j=1}^{m} b_j u_j} \prod_{i=1}^{n} e^{h_i\left(c_i + \sum_{j=1}^{m} w_{ij} u_j\right)} \\
&= \frac{1}{Z} e^{\sum_{j=1}^{m} b_j u_j} \sum_{h_1} e^{h_1\left(c_1 + \sum_{j=1}^{m} w_{1j} u_j\right)} \cdots \sum_{h_n} e^{h_n\left(c_n + \sum_{j=1}^{m} w_{nj} u_j\right)} \\
&= \frac{1}{Z} e^{\sum_{j=1}^{m} b_j u_j} \prod_{i=1}^{n}\sum_{h_i} e^{h_i\left(c_i + \sum_{j=1}^{m} w_{ij} u_j\right)} \\
&= \frac{1}{Z}\prod_{j=1}^{m} e^{b_j u_j} \prod_{i=1}^{n}\left(1 + e^{c_i + \sum_{j=1}^{m} w_{ij} u_j}\right)
\end{aligned} \tag{1.35}$$



Σχήμα 1.2: Ενα Restricted Boltzmann Machine το οποίο είναι ενα διμερές γράφημα με μη κατευθυνόμενες ακμές. Οι νευρώνες u_m και h_n αντιστοιχούν σε ορατα και κρυφά στοιχεία αντίστοιχα. Ο ορος "Restricted" υποδηλώνει οτι δεν επιτρέπονται συνδέσεις για στοιχεία του ίδιου επιπέδου. Οι παράμετροι biases δεν δείχνονται, αλλα σε κάθε νευρώνας αντιστοιχεί και ένα bias

Η παραπάνω έκφραση για την περιθώρια κατανομή υποδεικνύει γιατι τα Restricted Boltzmann Machines είναι product of experts μοντελα [6, 7].

Ένα Restricted Boltzmann Machine αποτελείται απο m ορατά και $k+1$ κρυφά στοιχεία και μπορεί να μοντελοποιήσει μια αγνωστη κατανομή με $\{0,1\}^m$. Η ποσότητα k εκφράζει τον αριθμο των στοιχείων απο $\{0,1\}^m$ που είναι δυνατό να εμφανιστούν με μια μη μηδενική πιθανότητα. Υπάρχει μια εξάρτηση ανάμεσα στα ορατα και κρυφά στοιχεία που απαιτούνται για την μοντελοποίηση μιας κατανομής και ακομα και ένας μικρότερος αριθμος κρυφών στοιχείων θα μπορούσε να είναι αρκετος [8].

Για τον υπολογισμό των υπο συνθήκη πιθανοτήτων ενος δεδομένου κρυφού η ορατου νευρώνα, μπορει να οριστεί ως $\boldsymbol{u}_{-l}$ το σύνολο των ορατών μεταβλητών χωρίς την μεταβλητή l:

$$a_l(\boldsymbol{h}) = -\sum_{i=1}^{n} w_{il} h_i - b_l \tag{1.36}$$

$$\beta(\boldsymbol{u}_{-l}, \boldsymbol{h}) = -\sum_{i=1}^{n} \sum_{j=1, j\neq l}^{m} w_{ij} h_i u_j - \sum_{j=1, j\neq l}^{m} b_j u_j - \sum_{i=1}^{n} c_i h_i \tag{1.37}$$

Η συνάρτηση ενέργειας $E(\boldsymbol{u}, \boldsymbol{h})$ δίνεται τότε απο την σχέση:

$$E(\boldsymbol{u}, \boldsymbol{h}) = \beta(\boldsymbol{u}_{-l}, \boldsymbol{h}) + u_l a_l(\boldsymbol{h}) \tag{1.38}$$

οπου η ποσότητα $u_l a_l(\boldsymbol{h})$ υποδηλώνει μια συλλογή όρων των u_l. Η υπο συνθήκη πιθανότητα του V_l ορατού στοιχεία για δεδομένο κρυφό επίπεδο $\boldsymbol{h}$

ισούται τότε με [9] :

$$
\begin{aligned}
p(V_l = 1|\boldsymbol{h}) = p(V_l = 1|\boldsymbol{u}_{-l}, \boldsymbol{h}) &= \frac{p(V_l = 1, \boldsymbol{u}_{-l}, \boldsymbol{h})}{p(\boldsymbol{u}_{-l}, \boldsymbol{h})} \\
&= \frac{e^{-E(u_{l=1}, \boldsymbol{u}_{-l}, \boldsymbol{h})}}{e^{-E(u_{l=1}, \boldsymbol{u}_{-l}, \boldsymbol{h})} + e^{-E(u_{l=0}, \boldsymbol{u}_{-l}, \boldsymbol{h})}} \\
&= \frac{e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h}) - 1 \cdot a_l(\boldsymbol{h})}}{e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h}) - 1 a_l(\boldsymbol{h})} + e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h}) - 0 \cdot a_l(\boldsymbol{h})}} \\
&= \frac{e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h})} \cdot e^{-a_l(\boldsymbol{h})}}{e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h})} \cdot e^{-a_l(\boldsymbol{h})} + e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h})}} \\
&= \frac{e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h})} e^{-a_l(\boldsymbol{h})}}{e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h})} \cdot (e^{-a_l(\boldsymbol{h})} + 1)} \\
&= \frac{e^{-a_l(\boldsymbol{h})}}{e^{-a_l(\boldsymbol{h})} + 1} \\
&= \frac{\frac{1}{e^{a_l(\boldsymbol{h})}}}{\frac{1}{e^{a_l(\boldsymbol{h})}} + 1} \\
&= \frac{1}{1 + e^{a_l(\boldsymbol{h})}} \\
&= \sigma(-a_l(\boldsymbol{h})) \\
&= \sigma\left(\sum_{i=1}^{n} w_{il} h_i + b_l \right)
\end{aligned}
\tag{1.39}
$$

Η συνάρτηση σ είναι η σιγμοειδής συνάρτηση:

$$
\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1.40}
$$

Μπορεί να προκύψει μια παρόμοια ισότητα για έναν κρυφό νευρώνα με δεδομένο ορατό επίπεδο. Για να μείνουμε πιστοί στον αρχικό συμβολισμο για τα i και j οι δυο εξισώσεις δίνονται παρακάτω:

$$
p(H_i = 1|\boldsymbol{u}) = \sigma\left(\sum_{j=1}^{m} w_{ij} u_j + c_i \right) \tag{1.41}
$$

$$
p(V_j = 1|\boldsymbol{h}) = \sigma\left(\sum_{i=1}^{n} w_{ij} h_i + b_j \right) \tag{1.42}
$$

Αν θέσουμε στην εξίσωση (1.20) του τυχαίου μαρκοβιανού πεδίου την πα-

ράμετρο θ ίση με τα βάρη w_{ij} ο πρώτος όρος δίνει:

$$\begin{aligned}
\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial w_{ij}} &= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) h_i u_j \\
&= \sum_{\boldsymbol{h}} \prod_{k=1}^{n} p(h_k|\boldsymbol{u}) h_i u_j \\
&= \sum_{h_i} \sum_{h_{-i}} p(h_i|\boldsymbol{u}) p(\boldsymbol{h_{-i}}|\boldsymbol{u}) h_i u_j \\
&= \sum_{h_i} p(h_i|\boldsymbol{u}) h_i u_j \underbrace{\sum_{\boldsymbol{h_{-i}}} p(\boldsymbol{h_{-i}}|\boldsymbol{u})}_{=1} \\
&= p(H_i = 1|\boldsymbol{u}) u_j \\
&= \sigma\left(\sum_{j=1}^{m} w_{ij} u_j + c_i\right) u_j
\end{aligned} \tag{1.43}$$

Ο δεύτερος όρος μπορεί να γραφεί ως:

$$\begin{aligned}
\sum_{\boldsymbol{u},\boldsymbol{h}} \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \theta} &= \sum_{\boldsymbol{u}} p(\boldsymbol{u}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \theta} \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}) \sum_{\boldsymbol{u}} p(\boldsymbol{u}|\boldsymbol{h}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \theta}
\end{aligned} \tag{1.44}$$

Παρατηρείται οτι το εξωτερικό άθροισμα και στις δύο περιπτώσεις έχει μια εκθετική πολυπλοκότητα εφόσον είναι μια άθροιση πάνω σε 2^N καταστάσεις. Άρα η ποσότητα που υπολογίζεται δεν μπορεί να επιλυθεί ακόμα και αν το εσωτερικό άθροισμα παραγοντοποιηθεί με ανάλογο τρόπο.

Η παράγωγος της λογαριθμικής πιθανοφάνειας για την περίπτωση οπου η παράμετρος μεταβολής θ ισούται με τα βάρη w_{ij} είναι:

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial w_{ij}} &= -\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial w_{ij}} + \sum_{\boldsymbol{u},\boldsymbol{h}} p(\boldsymbol{u},\boldsymbol{h}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial w_{ij}} \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) h_i u_j - \sum_{\boldsymbol{u}} p(\boldsymbol{u}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) h_i u_j \\
&= p(H_i = 1|\boldsymbol{u}) u_j - \sum_{\boldsymbol{u}} p(\boldsymbol{u}) p(H_i = 1|\boldsymbol{u}) u_j
\end{aligned} \tag{1.45}$$

Χρησιμοποιώντας τον κοινό συμβολισμό της βιβλιογραφίας και υποθέτοντας ένα σύνολο εκπαίδευσης $S = \{u_1, \ldots, u_l\}$,, η μέση τιμή της παραγώγου

της λογαριθμικής πιθανοφάνειας είναι:

$$\begin{aligned}\frac{1}{l}\sum_{\boldsymbol{u}\in S}\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial w_{ij}} &= \frac{1}{l}\sum_{\boldsymbol{u}\in S}\left[-\mathcal{E}_{p(\boldsymbol{h}|\boldsymbol{u})}\left[\frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial w_{ij}}\right]+\mathcal{E}_{p(\boldsymbol{h},\boldsymbol{u})}\left[\frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial w_{ij}}\right]\right]\\ &= \frac{1}{l}\sum_{\boldsymbol{u}\in S}\Big[\mathcal{E}_{p(\boldsymbol{h}|\boldsymbol{u})}[u_i h_j]-\mathcal{E}_{p(\boldsymbol{h},\boldsymbol{u})}[u_i h_j]\Big]\\ &= \langle u_i h_j\rangle_{p(\boldsymbol{h}|\boldsymbol{u})q(\boldsymbol{u})}-\langle u_i h_j\rangle_{p(\boldsymbol{h},\boldsymbol{u})}\end{aligned} \tag{1.46}$$

Η κατανομή q είναι η κατανομή που αναπαρίσταται απο το σύνολο δεδομένων και το παραπάνω αποτέλεσμα μπορεί να γραφτεί ως:

$$\sum_{\boldsymbol{u}\in S}\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial w_{ij}} \propto \langle u_i h_j\rangle_{data}-\langle u_i h_j\rangle_{model} \tag{1.47}$$

Μπορούμε να θέσουμε την παράμετρο θ ίση με το σύνολο των υπολειπόμενων παραμέτρων μεταβολής, δηλαδή τα βάρη b_j και c_i, και να λάβουμε τις ακόλουθες εκφράσεις για τις παραγώγους:

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial b_j} = u_j-\sum_{\boldsymbol{u}}p(\boldsymbol{u})u_j \tag{1.48}$$

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial c_i} = p(H_{i=1}|\boldsymbol{u})-\sum_{\boldsymbol{u}}p(\boldsymbol{u})p(H_i=1|\boldsymbol{u}) \tag{1.49}$$

Είναι δυνατό να εκτιμηθούν οι παραπάνω όροι με Monte Carlo μαρκοβιανών αλυσίδων. Συνεχίζει να υπάρχει ένα πρόβλημα ωστόσο. Ενα αντιπροσωπευτικό υποσύνολο απο δείγματα της κατανομής του μοντέλου θα απαιτούσε την συνεχόμενη δειγματοληψία της μαρκοβιανής αλυσίδας για μεγάλο διάστημα μέχρι να φτάσει σε ισορροπία. Αυτό δεν είναι δυνατό λόγω του υπολογιστικού κόστους και χρειάζεται μια επιπλέον προσέγγιση.

1.3.1 Contrastive Divergence

Η Contrastive Divergence είναι η πιο κοινή προσεγγιστική τεχνική για τον υπολογισμό των αναμενόμενων τιμών στην παράγωγο της λογαριθμικης πιθανοφάνειας υπο την κατανομή του μοντέλου [6, 10, 11, 12, 13]. Συμβολίζεται ως $CD\text{-}k$ οπου k είναι ο αριθμός των βημάτων που εκτελείται.

Αντι για την εφαρμογή διαδοχικών βημάτων δειγματοληψίας Gibbs με στόχο να οδηγηθεί το νευρωνικό δίκτυο σε ισορροπία, είναι δυνατό να εισάγει κάποιος στους ορατούς νευρώνες ένα παράδειγμα εκπαίδευσης $u^{(0)}$ και να εκτελέσει μια αλυσίδα Gibbs για k βήματα, αποκτώντας μια ανακατασκευή $u^{(k)}$. Για ένα μεγάλο αριθμό προβλημάτων, ακόμα και ενα βήμα $k=1$ είναι αρκετό. Η προσέγγιση που προκύπτει για την παράγωγο της λογαριθμικής πιθανοφάνειας

προς μια παράμετρο μεταβολής θ, η οποια είναι μεροληπτική, είναι:

$$CD_k(\boldsymbol{\theta}, \boldsymbol{u^{(0)}}) = -\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}^{(0)}) \frac{\partial E(\boldsymbol{u}^{(0)}, \boldsymbol{h})}{\partial \boldsymbol{\theta}} + \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}^{(k)}) \frac{\partial E(\boldsymbol{u}^{(k)}, \boldsymbol{h})}{\partial \boldsymbol{\theta}} \tag{1.50}$$

Κάποιος μπορεί να επιλέξει να εκτελέσει την Contrastive Divergence σε ολόκληρο το σύνολο δεδομένων S για κάθε βήμα. Ο βέλτιστος τρόπος είναι η εφαρμογή σε ένα υποσύνολο δεδομένων $S' \subset S$, δηλαδή σε ενα mini-batch, ειδικά για μεγάλο αριθμό δεδομένων.

Σε κάθε περίπτωση η Contrastive Divergence είναι μια προσεγγιστική τεχνική και το προκύπτον δείγμα μπορεί να μην είναι απο την κατανομή ισορροπίας του μοντέλου. Η προσέγγιση είναι δηλαδη μεροληπτική.

Το παρακάτω θεώρημα [12] όπως εμφανίζεται στο [1] είναι σημαντικό για μια καλύτερη κατανόηση της Contrastive Divergence:

Theorem 2 (Bengio and Delalleau)**.** *Για μια αλυσίδα Gibbs η οποία συγκλίνει:*

$$\boldsymbol{u}^{(0)} \Rightarrow \boldsymbol{h}^{(0)} \Rightarrow \boldsymbol{u}^{(1)} \Rightarrow \boldsymbol{h}^{(1)} \dots \tag{1.51}$$

Η παράγωγος της λογαριθμικής πιθανοφάνειας ισούται με:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\boldsymbol{u}^{(0)}) = &-\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}^{(0)}) \frac{\partial E(\boldsymbol{u}^{(0)}, \boldsymbol{h})}{\partial \boldsymbol{\theta}} \\ &+ E_{p(\boldsymbol{u}^{(k)})|\boldsymbol{u}^{(0)}} \left[\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}^{(k)}) \frac{\partial E(\boldsymbol{u}^{(k)}, \boldsymbol{h})}{\partial \theta} \right] \\ &+ E_{p(\boldsymbol{u}^{(k)})|\boldsymbol{u}^{(0)}} \left[\frac{\partial \ln p(\boldsymbol{u}^{(k)})}{\partial \theta} \right] \end{aligned} \tag{1.52}$$

και ο τελικός όρος συγκλίνει στο μηδεν καθώς $k \to \infty$.

Ο ρυθμός ανάμιξης της μαρκοβιανής αλυσίδας είναι μια μέτρηση του πόσο γρήγορα οδηγείται στην κατανομή ισορροπίας. Περιγράφεται απο τις πιθανότητες μετάβασης και είναι ένας απο τους παράγοντες-μαζί με τα βήματα εκτέλεσης- που επιδρά στο σφάλμα προσεγγίσεων. Η τάξη μεγέθους των παραμέτρων μεταβολής επιδρά στον ρυθμό ανάμιξης. Αυτό ειναι εμφανές απο τις εκφράσεις για τις υπο συνθήκη πιθανότητες σε όρους της σιγμοειδούς συνάρτησης οπου υψηλές τιμές για τις παραμέτρους μεταβολής αντιστοιχούν σε τιμές κοντά στο μηδεν για τις υπο συνθήκη πιθανότητες. Η μαρκοβιανή αλυσίδα τοτε εξελίσσεται πιο αργα στον χρόνο.

Το παρακάτω θεώρημα [14, 1] δίνει ένα άνω φράγμα για την αναμενόμενη τιμή του σφάλματος προσέγγισης υπο την εμπειρική κατανομή:

Theorem 3 (Fischer and Igel)**.** *Εστω p η περιθώρια κατανομή των ορατων στοιχείων και q η εμπειρική κατανομή που ορίζεται απο ένα σύνολο δειγμάτων $\boldsymbol{u}, \dots, \boldsymbol{u}_l$.*

Ένα άνω φράγμα για την αναμενόμενη τιμή του σφάλματος της CD-k προσέγγισης της παραγώγου προς την παράμετρο θ_a είναι:

$$\left| E_{(q)(\boldsymbol{u}^{(0)})} \left[E_{p(\boldsymbol{u}^{(k)}|\boldsymbol{u}^{(0)})} \left[\frac{\partial \ln p(\boldsymbol{u}^{(k)})}{\partial \theta_a} \right] \right] \right| \leq \frac{1}{2} |q - p| \left(1 - e^{-(m+n)\Delta}\right)^k \tag{1.53}$$

με:

$$\Delta = max \left\{ \underset{l \in \{1,\ldots,m\}}{max} \theta_l, \underset{l \in \{1,\ldots,n\}}{max} \xi_l \right\} \tag{1.54}$$

οπου:

$$\theta_l = max \left\{ \left| \sum_{i=1}^{m} I_{\{w_{il}>0\}} w_{il} + b_l \right|, \left| \sum_{i=1}^{m} I_{\{w_{il}<0\}} w_{il} + b_l \right| \right\} \tag{1.55}$$

και:

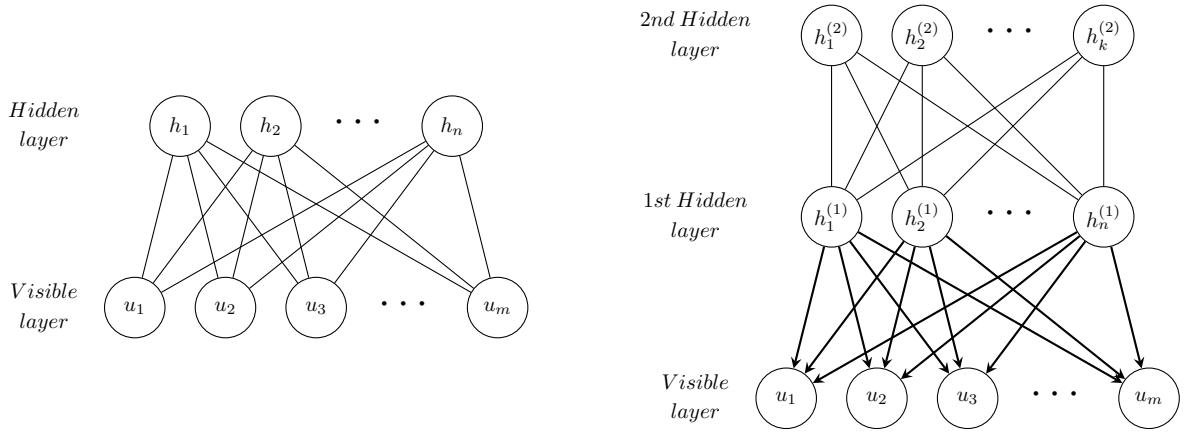
$$\xi_l = max \left\{ \left| \sum_{j=1}^{m} I_{\{w_{lj}>0\}} w_{lj} + c_l \right|, \left| \sum_{j=1}^{m} I_{\{w_{lj}<0\}} w_{lj} + c_l \right| \right\} \tag{1.56}$$

Το φράγμα έχει μια εξάρτηση στο σύνολο των ορατών και αόρατων στοιχείων του Restricted Boltzmann Machine, στην απόλυτη τιμή των παραμέτρων μεταβολής και στην απόσταση μεταβολής ανάμεσα στην κατανομή του μοντέλου και στην αρχική κατανομή της αλυσίδας Gibbs. Δεν είναι απαραίτητο οτι η μάθηση με χρήση της Contrastive Divergence θα δώσει εκπαίδευση μέγιστης πιθανοφάνειας για τις τιμές λόγω του σφάλματος προσέγγισης. Η μεροληψία μπορεί να οδηγήσει και τις παραμέτρους σε σύγκλιση για τιμές που δεν αντιστοιχούν στην μέγιστη πιθανοφάνεια. Επίσης, η πιθανοφάνεια μπορεί να αρχίσει να αποκλίνει μετά απο κάποιες επαναλήψεις αν ο αριθμός των βημάτων k δεν είναι αρκετά μεγάλος. Μια σωστή ρύθμιση του ορου weight decay μπορεί να προσφέρει μια λύση στο τελευταίο πρόβλημα.

1.4.0 Δίκτυα Βαθιάς Πεποιθήσεως

Τα Restricted Boltzmann Machines μπορούν να χρησιμοποιηθούν για την δημιουργία βαθεων νευρωνικών δικτύων. Είναι δυνατη προσθήκη πολλαπλών Restricted Boltzmann Machines σε σειρά έτσι ώστε να σχηματίσουν ενα νευρωνικό δίκτυο βαθιάς πεποιθήσεως [15, 13]. Παρομοίως με τα Restricted Boltzmann Machines δεν επιτρέπονται συνδέσεις ανάμεσα σε νευρώνες του ίδιου επιπέδου. Οι ακμές ανάμεσα στους νευρώνες των δύο τελευταίων κρυφών επιπέδων είναι μη κατευθυνόμενες ενώ όλες οι υπολοιπες είναι κατευθυνόμενες. Κάθε δύο επιπεδα περιγράφονται απο ένα σύνολο παραμέτρων μεταβολής $\{\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{a}\}$.

Ένα δίκτυο βαθιας πεποιθήσεως σχηματίζεται αρχικά εκπαιδεύοντας ένα Restricted Boltzmann Machine χρησιμοποιώντας την Contrastive Divergence.



Σχήμα 1.3: Σύγκριση ενός Restricted Boltzmann Machine και ενός Δικτύου Βαθιάς Πεποιθήσεως το οποιο αποτελείται απο κατευθυνόμενες και μη κατευθυνόμενες ακμές. Το Δίκτυο Βαθιάς Πεποιθήσεως αποτελέιται απο πολλαπλά κρυφά επίπεδα, και παρόμοια με το Restricted Boltzmann Machine δεν επιτρέπονται συνδέσεις αναμεσα σε στοιχεία του ίδιου επιπέδου.

Οι τιμές του κρυφού επιπέδου όταν τοποθετούνται τα δεδομένα εκπαίδευσης ως είσοδος στο ορατο επίπεδο χρησιμοποιούνται για την εκπάιδευση ενος δεύτερου Restricted Boltzmann Machine. Αυτη η διαδικασία μπορεί να επαναληφθεί όσες φορες ειναι επιθυμητο για τον σχηματισμό νέων επιπέδων.

Η δημιουργία δειγμάτων απο ένα δίκτυο βαθιάς πεποιθήσεως είναι δυνατή με διαδοχικές δειγματοληψίες Gibbs στα δύο τελευταία κρυφά επίπεδα μεχρι να φτάσουν σε ισορροπία. Ενα πέρασμα των τιμών στα υπόλοιπα επίπεδα του μοντέλου θα δώσει μια ανακατασκευή στο ορατό επίπεδο.

2. *Το Ising Μοντέλο*

2.1.0 *Το Διδιάστατο Ising Μοντέλο και η Μετάβαση Φάσης Δευτέρας Τάξεως*

Ακολουθεί μια σύντομη αλλα πλήρης εισαγωγή για το μοντέλο Ising στις δύο διαστάσεις και την μετάβαση φάσης δευτέρας τάξεως που εμφανίζει για δεδομένη κρίσιμη θερμοκρασία. Επίσης συμπεριλαμβάνονται και οι προσομοιώσεις Monte Carlo μαρκοβιανών αλυσίδων σύμφωνα με την σχετική βιβλιογραφία [16, 17, 18] και σχετικές δημοσιεύσεις.

Μια προσομοίωση Monte Carlo εκτελείται σε ένα σύστημα που περιγράφεται απο το κανονικό σύνολο για τον υπολογισμό των αναμενόμενων τιμών μιας παρατηρήσιμης ποσότητας $\mathcal{O}$:

$$\langle \mathcal{O} \rangle = \frac{1}{Z} \sum_{k=1}^{K} p_k \mathcal{O}^{(k)} = \frac{1}{Z} \sum_{k=1}^{K} \mathcal{O}^{(k)} e^{-\beta E^{(k)}}. \tag{2.1}$$

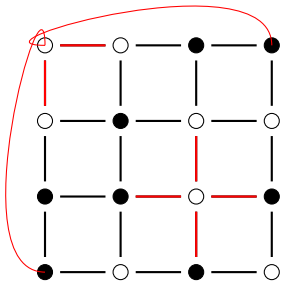
Η αντίστροφη θερμοκρασία $\beta = 1/k_\beta T$ εχει τον ρόλο ενός βάρους στην τιμή της ενέργειας και ορίζει μια χαρακτηριστική κλίμακα ενέργειας. Η σταθερά Boltzmann επιλέγεται ακολούθως να είναι ίση με $k_\beta = 1$. Ο εκθέτης k ορίζει μια κατάσταση του συστήματος στην οποια αντιστοιχεί μια απεικόνιση $\{s_i\}$ με βαθμούς ελευθερίας s_i, $i = 1, \ldots, N$. Μια δεδομένη απεικόνιση έχει μια παρατηρήσιμη ποσότητα $\mathcal{O}^{(k)}$ και μια εσωτερική ενέργεια $E^{(k)}$. Το άθροισμα είναι πάνω σε όλες τις πιθανές απεικονίσεις $\{s_i\}$. Η σταθερά κανονικοποίησης Z η οποία κωδικοποιεί όλη την στατιστική πληροφορία του συστήματος μετρώντας όλες τις καταστάσεις με το σωστό βάρος τους ονομάζεται συνάρτηση επιμερισμού και δίνεται απο την σχέση:

$$Z = Z(\beta) = \sum_{k=1}^{K} e^{-\beta E^{(k)}}. \tag{2.2}$$

Κάθε θερμοδυναμική παρατηρήσιμη ποσότητα μπορεί να υπολογιστεί μέσω της συνάρτησης επιμερισμού Z. Για παράδειγμα η εσωτερική ενέργεια $U \equiv \langle E \rangle$, η ειδική θερμότητα C και η ελεύθερη ενέργεια F δίνονται απο τις σχέσεις:

$$U \equiv \langle E \rangle = -\frac{\partial \ln Z}{\partial \beta} \tag{2.3}$$

$$C = k\beta^2 \frac{\partial^2 \ln Z}{\partial \beta^2} \tag{2.4}$$



Σχήμα 2.1: Ενα $N = 4 * 4$ Ising μοντέλο σε ένα τετραγωνικό πλέγμα με περιοδικές συνοριακές συνθήκες. Οι κόκκινες γραμμές αντιστοιχούν σε αλληλεπιδράσεις πλησιέστερων γειτόνων/ Η περίπτωση επιλογής ενός οριακού σπιν συμπεριλαμβάνεται. Οι μάυροι κύκλοι αντιστοιχούν σε σπιν με τιμες $s_i = -1$ και οι λευκοι σε σπιν με τιμές $s_i = +1$.

$$F = -\frac{1}{\beta} \ln Z \tag{2.5}$$

Ας υποθέσουμε ένα πεδίο με δεδομένη τιμή Y και μια συζυγή μεταβλητή X που έχει συζευχθεί σε αυτό και συμπεριλαμβάνεται ως ένας όρος $-XY$ στη Χαμιλτονιανή. Ειναι δυνατό να υπολογιστεί η αναμενόμενη τιμή της μεταβλητής X ως:

$$\langle X \rangle = -\frac{\partial F}{\partial Y}, \tag{2.6}$$

και η επιδεκτικότητα, δηλαδή η απόκριση της X σε αλλαγές του Y ως:

$$\chi = \frac{\partial \langle X \rangle}{\partial Y} \tag{2.7}$$

Το Ising [19] μοντέλο έχει μια Χαμιλτονιανή:

$$H = -\sum_{\langle ij \rangle} J_{ij} s_i s_j + B \sum_i s_i, \tag{2.8}$$

με πιθανές τιμές σπιν $s_i = \pm 1$ οι οποίες τοποθετούνται σε ενα d-διάστατο πλέγμα. Οι σταθερές J_{ij} μετρούν την ισχύ της αλληλεπίδρασης ανάμεσα σε σπίν και θέτονται ίσες με $J_{ij} = J = 1$, ορίζοντας ένα σιδηρομαγνητικό μοντέλο . Το εξωτερικό μαγνητικό πεδίο B ισούται με μηδέν για το υπόλοιπο του κεφαλαίου. Ο ολικός αριθμός των σπιν ισούται με $N = \prod_{i=1}^{d} n_i$ οπου n_i ειναι ο αριθμός των συνδέσεων και d η διαστατικότητα του μοντέλου. Θα υποθέσουμε αλληλεπίδραση πλησιέστερων γειτόνων $\langle ij \rangle$ και διαστατικότητα ίση με $d = 2$, μελετώντας το τετραγωνικό πλέγμα. Το σύστημα έχει ένα σύνολο απο 2^N απεικονίσεις και είναι αδύνατο να επιλυθεί αναλυτικά κάνοντας όλες τις αθροίσεις, οπότε είναι αναγκαία μια στατιστική προσέγγιση.

Η αναλυτική επίλυση του Onsager[20] για το διδιάστατο Ising, η απλότητα του και το γεγονός οτι εμφανίζει μια μετάβαση φάσης δευτέρας τάξεως για δεδομένη θερμοκρασία:

$$\beta_c = \frac{1}{T_c} = \frac{1}{2} \ln(1 + \sqrt{2}) \approx 0.44068679\ldots, \tag{2.9}$$

το κάνουν ένα ιδανικό μοντέλο για τον έλεγχο τεχνικών προσομοιώσεων. Επίσης είχε μεγάλη επίδραση στην μελέτη της στατιστικής φυσικής και στην κβαντική θεωρία πεδίου. Επιπλέον η αναλυτική λύση του Onsager προσφέρει ακριβείς τιμές για τις σχέσεις βάθμισης, γνωστές και ως κρίσιμοι εκθέτες που έχουν ίδιες τιμές ασχέτως της τοπολογίας του συστήματος ή της μορφής της αλληλεπίδρασης.

Το μηκος συσχετισμού ξ ορίζεται ώς ένα μέτρο της πλεγματικής απόστασης οπου δύο βαθμοί ελευθερίας είναι μετρήσιμα συσχετισμένοι. Η ανηγμένη θερμοκρασια t ορίζεται ως η απόσταση απο το κρίσιμο σημείο:

$$t = \frac{T - T_c}{T_c} = \frac{\beta_c}{\beta} - 1, \tag{2.10}$$

Οι κρίσιμοι εκθετες του Onsager δίνονται απο τις σχέσεις:

$$\begin{aligned} correlation\, length\, \xi &\sim |t|^{-\nu} \\ specific\, heat\, C &\sim |t|^{-a} \\ magnetization\, M &\sim |t|^{\beta} \\ magnetic\, susceptibility\, \chi &\sim |t|^{\gamma} \\ magnetization(field) M &\sim B^{-1/\delta}, (t=0) \\ correlations\ \langle s_i s_j \rangle &\sim |x_i - x_j|^{-d+2-\eta}, \\ & \quad for\, |x_i - x_j| \to \infty, (t=0) \end{aligned} \tag{2.11}$$

$$\nu = 1,\ a = 0,\ \beta = \frac{1}{8},\ \gamma = \frac{7}{4},\ \delta = 15, \eta = \frac{1}{4} \tag{2.12}$$

και θα υπολογιστούν με χρήση της ομάδας επανακανονικοποίησης πραγματικού χώρου.

Ακολουθούν κάποιοι απαραίτητοι ορισμοί. Οι παράμετροι τάξης είναι ενα σημαντικό εργαλείο για την αναγνώριση μεταβάσεων φάσης δευτέρας τάξεως μεσα απο τον χαρακτηρισμό μιας συμμετρίας. Στο Ising μοντέλο η μαγνήτιση M είναι μια παράμετρος τάξης η οποία μηδενίζεται στην άτακτη φάση λόγω της Z_2 συμμετρίας $s_i \to -s_i$ ενω έχει σταθερή τιμή στην φάση πλήρης τάξεως. Η μαγνήτιση είναι μια μη αναλυτική συνάρτηση της θερμοκρασίας.

Οι κρίσιμοι εκθέτες ορίζουν μια κλάση παγκοσμιότητας. Όλα τα μοντέλα στην ίδια κλάση παγκοσμιότητας πρέπει να μοιράζονται τις ίδιες συμμετρίες και διαστατικότητα του χώρου. Εχουν την ίδια συμπεριφορά σε μεγάλου μήκους κλίμακες αφού δεν έχει πια σημασία η μικροσκοπική τους περιγραφή . Η παγκοσμιότητα και το αναλλοίωτο της κλίμακας (scale invariance) εμφανίζονται καθώς το μήκος συσχετισμού του συστήματος $\xi \to \infty$. Το αποκλινων μήκος συσχετισμού ειναι μοναδικώς ορισμένο, αφού όλες οι ποσότητες αποκλίνουν σε όρους μιας παραμέτρου, της ανηγμένης θερμοκρασίας.

Το αναλλοίωτο της κλίμακας συνεπάγεται οτι αλληλεπιδράσεις του μοντέλου σε μεγάλες κλίμακες εξαρτώνται μόνο απο τον λόγο του αντιστοιχου μήκους προς το αποκλίνων μήκος συσχετισμού. Το μήκος συσχετισμού ξεπερνά καθε χαρακτηριστικό μήκος του συστήματος, όπως την πλεγματική απόσταση καθώς αυξάνει η ανηγμένη θερμοκρασία. Σε συνδυασμό με την κλάση παγκοσμιότητας είναι αρκετό να βρεθεί ένα μοντέλο με αναλλοίωτο κλίμακας, στις κατάλληλες διαστάσεις και με την απαραίτητη συμμετρία για να απλοποιηθεί η μελέτη μιας μετάβασης φάσης δευτέρας τάξεως κοντά στο κρίσιμο σημείο.

2.1.1 *Δειγματοληψία με Κριτήριο Σημαντικότητας και η Τεχνική Re-Weighting*

Ο εκτιμητής μιας παρατηρήσιμης ποσότητας $\mathcal{O}$ απο M Monte Carlo μετρήσεις δίνεται απο τη σχέση:

$$\mathcal{O}_M = \frac{\sum_i \mathcal{O}_i p_i^{-1} e^{-\beta E_i}}{\sum_j p_j^{-1} e^{-\beta E_j}} \tag{2.13}$$

Οι πιθανότητες Boltzmann p αντιστοιχούν σε καταστάσεις που δειγματοληπτούνται για δεδομένη θερμοκρασία, και μετατρέπουν την παραπάνω εξίσωση:

$$\mathcal{O}_M = \frac{\sum_i \mathcal{O}_i (e^{-\beta E_i})^{-1} e^{-\beta E_i}}{\sum_j e^{(-\beta E_j)^{-1}} e^{-\beta E_j}} = \frac{1}{M} \sum_i \mathcal{O}_i \tag{2.14}$$

Αν υποθέσουμε στην παραπάνω εξίσωση πιθανότητες Boltzmann μιας θερμοκρασίας β_0 που βρίσκεται ικανοποιητικά κοντά έχουμε:

$$\mathcal{O}_M = \frac{\sum_i \mathcal{O}_i e^{-(\beta-\beta_0)E_i}}{\sum_j e^{-(\beta-\beta_0)E_j}} \tag{2.15}$$

με την πιθανότητα να επεκτείνουμε σε ένα εύρος τιμών $\beta_0 \pm \Delta\beta$ οπου $\Delta\beta \to 0$ στο θερμοδυναμικό όριο [21].

Η συνάρτηση επιμερισμού μπορεί να γραφεί σε όρους της πυκνότητας καταστάσεων $n(E)$, δηλαδή στον αριθμό των απεικονίσεων με εσωτερική ενέργεια E:

$$Z = Z(\beta) = \sum_E n(E) e^{-\beta E} \tag{2.16}$$

Για μία δεδομένη τιμή της β σε ένα εύρος θερμοκρασιών το οποίο βρίσκεται ικανοποιητικά μακριά απο το κρίσιμο σημείο η πυκνότητα πιθανότητας της ενέργειας είναι ίση με:

$$P_\beta(E) = c_\beta n(E) e^{-\beta E} \tag{2.17}$$

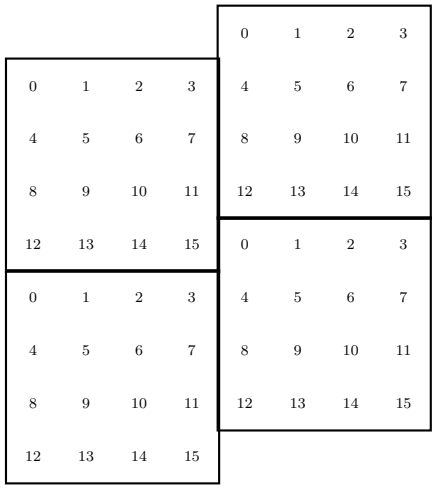
οπου c_β η κατάλληλη σταθερά κανονικοποίησης. Η πυκνότητα πιθανότητας της ενέργειας μεγιστοποιείται κοντά στη μέση τιμή της ενέργειας $E(\beta)$ με ενα πλάτος ανάλογο του τετραγώνου $\sqrt{V}$ οπου V είναι ο όγκος του συστήματος. Ισχύει $N \sim V$ για τις διακυμάνσεις λόγω των τοπικών συσχετίσεων των σπίν μακριά απο την κρίσιμη περιοχή και μια τυπική διακύμανση είναι της τάξης $\sim \sqrt{N}$. Η δυνατότητα επέκτασης σε μια μεταβλητή είναι $\Delta\beta \sim 1/\sqrt{V}$ με $\Delta\beta E \sim \sqrt{V}$ στις διακυμάνσεις του συστήματος. Οι μεγαλύτερες διακυμάνσεις στην κρίσιμη περιοχή επιτρέπουν μεγαλύτερο εύρος επέκτασης

Για μια δεδομένη θερμοκρασία β οι σημαντικές απεικονίσεις είναι αυτές για τις οποίες η πυκνότητα πιθανότητας $P_\beta(E)$ εχει μεγάλες τιμές. Ο στόχος είναι να δειγματοληφθούν απεικονίσεις με τα κατάλληλα βάρη Boltzmann μέσα απο την υλοποίηση μιας Markov διαδικασίας:

$$w_B^{(k)} = e^{-\beta E^{(k)}} \tag{2.18}$$

2.1.2 *Ο Αλγόριθμος Metropolis*

Χρησιμοποιώντας τις μαθηματικές έννοιες για τις μαρκοβιανές αλυσίδες και την δειγματοληψία Gibbs υποθέτουμε μια δεδομένη απεικόνιση k για την οποία



Σχήμα 2.2: Ενα παράδειγμα ελικοειδών συνοριακών συνθηκών που χρησιμοποιείται για την υλοποίηση του αλγόριθμου Metropolis.

η πιθανότητα μετάβασης σε μια απεικόνιση l δίνεται απο την σχέση $W^{(l)(k)} = W[k \to l]$. Ο πίνακας μετάβασης W ορίζεται ως:

$$W = \left(W^{(l)(k)}\right) \tag{2.19}$$

Η συνθήκη της λεπτομερούς ισορροπίας δεν ορίζει ένα σύνολο μοναδικών τιμών για τις πιθανότητες μετάβασης $W^{(l)(k)}$. Ο αλγόριθμος Metropolis [22] ο οποίος χρησιμοποιείται ευρέως και είναι υπολογιστικά απλός, προτείνει για μια δεδομένη απεικόνιση k, νέες απεικονίσεις l με πιθανότητες μετάβασης $f(l,k)$ οι οποίες είναι κανονικοποιημένες:

$$\sum_l f(l,k) = 1 \tag{2.20}$$

Η πιθανότητα να γίνει δεκτή μια νέα απεικόνιση l είναι:

$$w^{(l)(k)} = min\left[1, \frac{P_B^l}{P_B^k}\right] = \begin{cases} 1 & for\ E^l < E^k \\ e^{-\beta(E^l - E^k)} & for\ E^l > E^k \end{cases} \tag{2.21}$$

Ο λόγος αποδοχής ορίζεται ως ο λόγος των αποδεκτων απεικονίσεων προς τις προτεινόμενες κινήσεις. Αυτός ο ορισμός δεν συμπεριλαμβάνει την αποδοχή της τρέχουσας απεικόνισης

Ο αλγόριθμος Metropolis ορίζει τις πιθανότητες μετάβασης:

$$W^{(l)(k)} = f(l,k) w^{(l)(k)}\ for\ l \neq k \tag{2.22}$$

$$W^{(k)(k)} = f(k,k) + \sum_{l \neq k} f(l,k)(1 - w^{(l)(k)}) \tag{2.23}$$

Η συνθήκη συμμετρίας $f(l,k) = f(k,l)$ πρέπει να χρησιμοποιηθεί έτσι ώστε η $W^{(l)(k)}/W^{(k)(k)}$ να ικανοποιεί την συνθηκη λεπτομερούς ισορροπίας

Γενικά υπάρχει η δυνατότητα να χρησιμοποιηθεί ενα μεγάλο πλήθος απο πιθανότητες μετάβασης. Επιπλέον μπορούν να χρησιμοποιηθούν διαφορετικές πιθανότητες αποδοχής που ορίζουν μη συμμετρικές πιθανότητες προτεινόμενων απεικονίσεων.[23]

Οι παρατηρήσιμες ποσότητες που μετρώνται σε κάθε βήμα προσομοίωσης είναι η εσωτερική ενέργεια:

$$E = -\sum_{\langle ij \rangle} s_i s_j, \tag{2.24}$$

και η μαγνήτιση:

$$M = \left| \sum_i s_i \right| \tag{2.25}$$

Ειναι μεγάλης σημασίας να μετρείται η απόλυτη τιμή της μαγνήτισης. Η Z_2 συμμετρία συνεπάγεται οτι απεικονίσεις με κάθε σπιν διαφορετικό έχουν την

ίδια πιθανότητα να εμφανιστούν. Ειναι δυνατό να κανονικοποιηθεί η ενέργεια ανα σπιν ή ανα δεσμό:

$$\langle e \rangle = \frac{1}{N_l} \langle E \rangle = \frac{1}{2N} \langle E \rangle \tag{2.26}$$

Παρομοίως η μαγνήτιση μπορεί να κανονικοποιηθεί ανα σπίν:

$$\langle m \rangle = \frac{1}{N} \langle M \rangle \tag{2.27}$$

Οι υπόλοιπες παρατηρήσιμες ποσότητες που θα υπολογιστουν είναι η ειδική θερμότητα

$$c = \beta^2 N \langle (e - \langle e \rangle^2) \rangle = \beta^2 N (\langle e^2 \rangle - \langle e \rangle^2) \tag{2.28}$$

και η μαγνητική επιδεκτικότητα:

$$\chi = \beta N \langle (m - \langle m \rangle^2) \rangle = \beta N (\langle m^2 \rangle - \langle m \rangle^2) \tag{2.29}$$

2.1.3 *Ο Cluster Αλγόριθμος του Wolff*

Το πλεονέκτημα χρήσης ενός cluster αλγόριθμου είναι η επιλογή μεταβολής της τιμής ολόκληρων συστοιχιών απο σπιν σε κάθε βήμα. Με αυτόν τον τρόπο αντιμετωπίζεται και το φαινόμενο της κρίσιμης επιβράδυνσης κατα την μελέτη μεταβάσεων φάσης δευτέρας τάξεως.

Ενα Swendsen-Wang cluster απο σπιν υλοποιείται με τον σχηματισμό δεσμών ανάμεσα σε γειτονικές θέσεις i και j με πιθανότητα:

$$p_{\langle ij \rangle} = max[0, 1 - exp(-2\beta\delta_{q_i,q_j})] \tag{2.30}$$

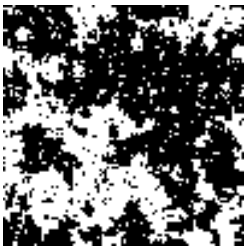
Ο cluster αλγόριθμος του Wolff επιτρέπει την μεταβολή των τιμών μιας μόνο συστοιχίας απο σπιν σε κάθε βήμα. Ο αλγόριθμος περιγράφεται απο τα εξής βήματα:

1. Επιλέγεται τυχαία μια θέση i και αντιστοιχείται σε μια συστοιχία c .
2. Οι πλησιέστεροι γείτονες της πλεγματικής θέσης i προστίθενται στη συστοιχία c με πιθανότητα $p_{\langle ij \rangle}$.
3. Εκτελούνται ακόλουθες επαναλήψεις του παραπάνω βήματος για όλα τα συμπεριλαμβανόμενα σπιν μέχρι να τερματιστεί η διαδικασία.
4. Σε όλες τις πλεγματικές θέσεις που απαρτίζουν την συστοιχία c αντιστοιχεί μια νέα τιμή σπιν $q_i^{'} \neq q_i$ η οποία επιλέγεται ομοιόμορφα απο το $q-1$ σύνολο των υπολοιπόμενων τιμών, οπου $q = 2$ για το μοντέλο που μελετάται σε αυτή την διπλωματική.

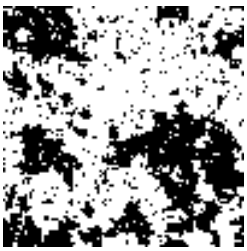
Υπάρχει μια πεπερασμένη πιθανότητα μία συγκεκριμένη θέση q_i να σχηματίσει μια συστοιχία και κάθε σπιν $q_i' \neq q_i$ είναι προσβάσιμο. Επαναλαμβάνοντας το παραπάνω για ένα αριθμό τιμών μεγαλύτερο η ίσο με το μέγεθος του

Σχήμα 2.3: Απεικονίσεις του διδιάστατου Ising μοντέλου που δειγματοληπτούνται με τον αλγόριθμο Wolff για $b = 0.43$ και $L = 100$.

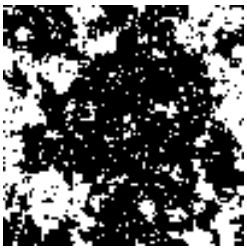
Αρχική απεικόνιση:



1 βημα:



2 βήματα:



συστήματος, υπάρχει μια πεπερασμένη πιθανότητα να συμπεριληφθούν όλες οι πλεγματικές θέσεις. Τελικά η Μαρκοβιανή αλυσίδα είναι μη υποβιβάσιμη.

Για να δειχθεί η ισχύς της συνθήκης λεπτομερούς ισορροπίας, θεωρουμε δυο απεικονίσεις $\{q_i\}$ και $\{q_{i'}\}$ που διαφέρουν ακριβώς κατα την μεταβολή μιας συστοιχίας c. Η c έχει τότε πιθανότητα:

$$P_c = \frac{|c|}{N} \prod_{\langle ij\rangle \in c} p_{\langle ij\rangle} \prod_{\langle ij\rangle \in c, j\neq c} exp(-2\beta\delta_{q_i,q_j}) \tag{2.31}$$

Η συστοιχία c έχει έναν αριθμό απο θέσεις $|c|$. Η ποσότητα $|c|/N$ ορίζει την πιθανότητα να επιλεχθεί ένα σπιν της συστοιχίας κατα το πρώτο βήμα του αλγόριθμου Wolff. Η συστοιχία που διαφέρει ακριβώς κατα μια μεταβολή στην απεικόνιση $\{q_i^{'}\}$ έχει πιθανότητα:

$$P_c^{'} = \frac{|c|}{N} \prod_{\langle ij\rangle \in c} p_{\langle ij\rangle}^{'} \prod_{\langle ij\rangle \in c, j\neq c} exp(-2\beta\delta_{q_i^{'},q_j^{'}}) \tag{2.32}$$

Εφόσον η συστοιχία και στις δύο περιπτώσεις αποτελείται απο πανομοιότυπα σπιν $p_{\langle ij\rangle}^{'} = p_{\langle ij\rangle}$. Τα υπόλοιπα σπιν εκτός της συστοιχίας και στα δύο συστήματα είναι ίδια αφού υποθέσαμε οτι οι δύο απεικονίσεις απλά διαφέρουν κατα την μεταβολή μιας συστοιχίας σπιν c. Η συνθήκη της λεπτομερούς ισορροπίας είναι:

$$\frac{W^{(\{q_i^{'}\},\{q_i\})}}{W^{(\{q_i\},\{q_i^{'}\})}} = \frac{W(\{q_i\}\to\{q_i^{'}\})}{W(\{q_i^{'}\}\to\{q_i\})} = \frac{\exp\Big(-2\beta\sum_{\langle ij\rangle}\delta_{q_i,q_j}\Big)}{\exp\Big(-2\beta\sum_{\langle ij\rangle}\delta_{q_i^{'},q_j^{'}}\Big)} = \frac{\exp(2\beta E)}{\exp(2\beta E^{'})} \tag{2.33}$$

Μια ολόκληρη μελέτη μπορεί να πραγματοποιηθεί για την σύγκριση των δύο αλγορίθμων στην κρίσιμη περιοχή, ωστόσο παραλείπεται. Ο κύριος στοχος είναι η χρήση του νευρωνικού δικτύου για τον υπολογισμό των παρατηρήσιμων ποσοτήτων με υψηλή ακρίβεια.

2.1.4 Ισορροπία και Αυτοσυσχέτιση

Μια τυπική προσομοίωση Monte Carlo μπορεί να χωριστεί σε δύο τμήματα. Στο αρχικό τμήμα προσομοιώσεων εκτος ισορροπίας και στο τμήμα παραγωγής απεικονίσεων εντός ισορροπίας. Το αρχικό κομμάτι αποτελείται απο απεικονίσεις που πρέπει να απορριφθούν αφου δεν έχουν χρησιμότητα στον υπολογισμό των παρατηρήσιμων ποσοτήτων. Οι απεικονίσεις που παράγονται εντός ισορροπίας χρησιμοποιούνται για τις μετρήσεις και των υπολογισμό των αναμενόμενων τιμών.

Ειναι λογικό να αναρωτηθεί κάποιος πόσα βήματα προσομοιώσεων χρειάζονται μέχρι την ισορροπία. Αν και υπάρχουν περιπτώσεις οπου μπορεί να δοθεί μια απάντηση ο γενικός κανόνας ειναι οτι οι απεικονίσεις που θα απορριφθούν πρέπει να διαλεχθούν προσεγγιστικά σωστά. Ο υπολογισμός του ολοκληρωμένου χρόνου αυτοσυσχετισμού, που εισάγεται παρακάτω, δεν είναι

πάντοντε δυνατός. Η συμπερίληψη απεικονίσεων εκτός ισορροπίας στα τελικά δεδομένα συνεπάγεται οτι έχουν γινει υπολογισμοί χρησιμοποιώντας απεικονίσεις που έχουν μηδενική πιθανότητα να εμφανιστούν στην ισορροπία. Η επιδραση της εισαγωγής λανθασμένων απεικονίσεων φθίνει κατα ενα παράγοντα $1/N$ καθώς ο αριθμός των βημάτων προσομοίωσης N αυξάνει. Επίσης μπορεί να ξεπεραστεί απο τα στατιστικά σφαλματα που φθινουν κατα $1/\sqrt{N}$. Είναι πάντα σημαντικό να μην συμπεριληφθύν απεικονίσεις εκτος ισορροπίας για να γινουν σωστες μετρήσεις. Γενικά μπορεί να εμφανιστούν και άλλα προβλήματα κατα την ευρεση ισορροπίας, οπως το σύστημα να μεταβεί σε μια μετασταθή κατάσταση. Σημειώνεται οτι η εύρεση ισορροπίας γίνεται δυσκολότερη καθως το μέγεθος του συστήματος και οι χρονοι αυτοσυσχετισμού αυξάνουν.

Οταν η εύρεση ισορροπίας έχει επιτευχθεί, η αναμενόμενη τιμή $\hat{f}$ μιας παρατηρήσιμης ποσότητας μπορεί να υπολογιστεί απο τον αριθμό των μετρήσεων που έχουν προκύψει απο τα αντίστοιχα βηματα προσομόιωσης. Είναι σημαντικό να γίνουν κατανοητοί οι αυτοσυσχετισμοί που εμφανίζονται σε μια αλυσίδα Markov πριν την διεξαγωγή μιας σωστής μέτρησης.

Ορίζοντας ως x_i τις απεικονίσεις που παράγονται εντός ισορροπίας και υποθέτοντας N μετρήσεις απο μια αλυσίδα Markov έχουμε:

$$f_i = f_i(x_i), i = 1, \dots, N \tag{2.34}$$

Η αλυσίδα Markov είναι διακριτή στον χρόνο και κάθε χρονικό βήμα ανάμεσα σε δύο μετρήσεις f_i, f_{i+1}, που είναι ίσο με ένα βήμα προσομοίωσης, αντιστοιχεί στο ίδιο χρονικό διάστημα.

Ο εκτιμητής της αναμενόμενης τιμής $\hat{f}$:

$$\bar{f} = \frac{1}{N} \sum f_i \tag{2.35}$$

Η συνάρτηση αυτοσυσχέτισης της παρατηρήσιμης ποσότητας f ορίζεται ως:

$$\begin{aligned} \hat{C}(t) &= \hat{C}_{ij} \\ &= \langle (f_i - \langle f_i \rangle)(f_j - \langle f_j \rangle) \rangle \\ &= \langle f_i f_j \rangle - \langle f_i \rangle \langle f_j \rangle \\ &= \langle f_0 f_t \rangle - \hat{f}^2 \end{aligned} \tag{2.36}$$

οπου $t = |i - j|$ και το σύστημα είναι αναλλοίωτο στον χρόνο. Οταν $t \to \infty$:

$$\hat{C}(t) \sim exp\left(- \frac{t}{\tau_{exp}} \right) \tag{2.37}$$

Η ποσότητα τ_{exp} ορίζεται ως ο εκθετικός χρόνος αυτοσυσχέτισης. Η ιδιοτιμή $\lambda_0 = 1$ του πίνακα μεταβάσεων έχει ενα ιδιοδιάνυσμα την πιθανότητα κατανομής. Ο εκθετικός χρόνος αυτοσυσχέτισης μπορεί να εκφραστεί σε όρους

της ιδιοτιμής λ_1 αν υποθέσουμε οτι η f έχει μια μη μηδενική προβολή στην ιδιοκατάσταση. Ο εκθετικός χρόνος αυτοσυσχέτισης ισούται τοτε με:

$$\tau_{exp} = -\ln \lambda_1 \tag{2.38}$$

Ειναι αρκετά σημαντικό να αναφερθεί οτι η διασπορά f σχετίζεται με τις αυτοσυσχετίσεις μέσα απο την έκφραση:

$$\hat{C}(0) = \sigma^2(f) \tag{2.39}$$

Η διασπορά του $\bar{f}$ για τις συναρτήσεις αυτοσυσχέτισης και η μέση τιμή:

$$\begin{aligned}
\sigma^2(\overline{f}) &= \langle (\overline{f} - \hat{f})^2 \rangle \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \langle (f_i - \hat{f})(f_j - \hat{f}) \rangle \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \langle f_i f_j - f_i \hat{f} - f_j \hat{f} + \hat{f}^2 \rangle \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \Big[\langle f_i f_j \rangle - \hat{f}^2 \Big] \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij}
\end{aligned} \tag{2.40}$$

Στο τελευταίο αθροισμα παρατηρείται οτι οροι $|i-j| = 0$ εμφανίζονται για ένα σύνολο απο N φορές και $|i-j| = t$ με $1 \leq t \leq (N-1)$ εμφανίζονται $2(N-t)$ φορές:

$$\sigma^2(\overline{f}) = \frac{1}{N^2} \left[N\hat{C}(0) + 2 \sum_{t=1}^{N-1} (N-t)\hat{C}(t) \right] \tag{2.41}$$

$$\sigma^2(\overline{f}) = \frac{\sigma^2(f)}{N} \left[1 + 2 \sum_{t=1}^{N-1} \Big(1 - \frac{t}{N}\Big) \hat{c}(t) \right] \tag{2.42}$$

$$\hat{c}(t) = \frac{\hat{C}(t)}{\hat{C}(0)} \tag{2.43}$$

Ειναι δυνατή η σύγκριση ανάμεσα στην διασπορά του εκτιμητή f που έχει υπολογιστεί παραπάνω και την έκφραση για την περίπτωση χωρίς συσχέτιση:

$$\sigma^2_{uncorrelated}(\overline{f}) = \frac{\sigma^2(f)}{N} \tag{2.44}$$

Ο διαφορετικός όρος ορίζεται ώς ο ολοκληρωμένος χρόνος αυτοσυσχετισμού:

$$\tau_{int} = \left[1 + 2 \sum_{t=1}^{N-1} \left(1 - \frac{t}{N} \right) \hat{c}(t) \right] \tag{2.45}$$

Η διασπορά για την μέση τιμή των συσχετισμένων δεδομένων είναι μεγαλύτερη κατα ένα παράγοντα t_{int} σε σχέση με την περίπτωση μη συσχέτισης:

$$\tau_{int} = \frac{\sigma^2(\overline{f})}{\sigma^2_{uncorrelated}(\overline{f})}. \tag{2.46}$$

Στο θερμοδυναμικό όριο $N \to \infty$ η σχέση 2.45 είναι:

$$\tau_{int} = 1 + 2\sum_{t=1}^{\infty} \hat{c}(t) \tag{2.47}$$

Η εκτίμηση του ολοκληρωμένου χρόνου αυτοσυσχέτισης δεν είναι ευκολη. Ο εκτιμητής $\overline{\tau}_{int}$ στο θερμοδυναμικό όριο είναι:

$$\overline{\tau}_{int} = 1 + 2\sum_{t=1}^{\infty} \overline{c}(t) \tag{2.48}$$

Η διασπορά του παραπάνω εκτιμητή αποκλίνει:

$$\sigma^2(\bar{t}_{int}) \to \infty \tag{2.49}$$

Αν υποθέσουμε έναν εκτιμητή με εξάρτηση στον χρόνο t:

$$\overline{\tau}_{int}(t) = 1 + 2\sum_{t=1}^{\infty} \overline{c}(t') \tag{2.50}$$

μπορούμε να κάνουμε μια εκτίμηση του ολοκληρωμένου χρόνου αυτοσυσχετισμού αναζητώντας την βέλτιστη τιμή για την οποία ο $\overline{\tau}_{int}(t)$ είναι ανεξάρτητος του t. Μια εναλλακτική μέθοδος βρίσκεται παρακάτω.

2.1.5 *Ανάλυση Binning και Ολοκληρωμένος Χρόνος Αυτοσυσχετισμού*

Υποθέτουμε οτι οι N μετρήσεις της χρονοσειράς έχουν χωριστεί σε N_{bs} bins οπου $N_{bs} \leq N$ και κάθε N_{bs} αποτελείται απο N_b μετρήσεις:

$$N_b = \frac{N}{N_{bs}} \tag{2.51}$$

Τα δεδομένα τα οποια εχουν χωριστεί σε bins είναι οι μέσοι:

$$f_j^{N_b} = \frac{1}{N_b} \sum_{i=1+(j-1)N_b}^{jN_b} f_i, j = 1, \ldots, N_{bs} \tag{2.52}$$

Μια αύξηση στον αριθμό των μετρήσεων μέσα σε κάθε bin θα αντιστοιχούσα σε μια μείωση των αυτοσυσχετίσεων. Τελικά ο αριθμός των μετρήσεων μέσα σε καθε bin θα ήταν μεγαλύτερος απο τον εκθετικό χρόνο αυτοσυσχέτισης τ_{exp} και μόνο bins που βρίσκονται το ένα δίπλα στο άλλο θα ήταν συσχετισμένα. Μια ακομα μεγαλύτερη αυξηση στο μέγεθος κάθε bin θα οδηγούσε σε ακόμα μεγαλύτερες μειώσεις της αυτοσυσχέτισης.

Θεωρούμε όλα τα N_{bs} bins και υπολογίζουμε την μέση τιμή χρησιμοποιώντας τις N_b μετρήσεις:

$$\overline{f_j}^{N_b} = \frac{1}{N_{bs}} \sum_{j=1}^{N_{bs}} f_j^{N_b} \tag{2.53}$$

Επιπλέον θεωρούμε οτι έχουμε ασυσχέτιστα δεδομένα και το σφάλμα ισούται με την τυπική απόκλιση:

$$\sigma = \sqrt{\frac{1}{N_{bs}-1}(\overline{f_j^2} - \overline{f_j}^2)} \tag{2.54}$$

Ο ολοκληρωμένος χρόνος αυτοσυσχέτισης για την περίπτωση $N_b \to \infty$ είναι:

$$\tau_{int} = \lim_{N_b \to \infty} \tau_{int}^{N_b} \tag{2.55}$$

οπου:

$$\tau_{int}^{N_b} = \left(\frac{s^2_{\overline{f}_{N_b}}}{s^2_{\overline{f}}} \right) \tag{2.56}$$

και ο εκτιμητής της διασποράς είναι

$$(s_x^r)^2 = \frac{N}{N-1}(s_x^{'r})^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i^r - \overline{x}^r)^2 \tag{2.57}$$

Απο μια πρακτική άποψη, μια μεγάλη τιμή N_b θα ήταν αρκετή για να υποθέσουμε οτι $N_b \to \infty$. Ο όρος $s^2_{\overline{f}_{N_b}}$ στον $\tau_{int}^{N_b}$ έχει τον κύριο ρόλο για την εκτίμηση του σφάλματος τ_{int} αφού για την περίπτωση $N_b \to \infty$ οι τιμές $s^2_{\overline{f}}$ θα είναι πολύ μικρότερες. Ενα πεπερασμένο μέγεθος N_b μετρήσεων που αντιστοιχεί σε πρακτικά ασυσχέτιστα δεδομένα είναι μια καλή εκτίμηση για τον τ_{int}. Συμφωνα με το κεντρικό οριακό θεώρημα τα δεδομένα που προκύπτουν απο την ανάλυση binning μπορούν να θεωρηθουν ως γκαουσιανά, και η ποσότητα $s^2_{\overline{f}_{N_b}}$ είναι γνωστή αναλυτικά. Τα N_{bs} bins τα οποια έχουν επιλεχθεί ωστε να είναι ανεξάρτητα αντιστοιχούν τότε στο σφάλμα. Για την ποσότητα $s^2_{\overline{f}}$,μπορει να χρησιμοποιηθεί ο αριθμός των δεδομένων που δεν έχουν συσχέτιση :

$$N_{effective} = \frac{N}{\tau_{int}} \tag{2.58}$$

Ο ολοκληρωμένος χρόνος αυτοσυσχετισμού είναι σημαντικός κατα τις προσομοιώσεις Monte Carlo μαρκοβιανών αλυσίδων. Αρχικά, πρεπει να επιλεχθούν βήματα μεγαλύτερα απο τ_{int} για να φτάσει το σύστημα στην ισορροπία. Επιπλέον, γνώση του ολοκληρωμένου χρόνου αυτοσυσχετισμού επιτρέπει τον υπολογισμό της διασποράς για συσχετισμένα δεδομένα:

$$\sigma^2(\overline{f}) = \tau_{int}\frac{\sigma^2(f)}{N} \tag{2.59}$$

Αρα είναι δυνατό να χρησιμοποιηθεί αυτή η γνώση για τον υπολογισμό των σφαλμάτων:

$$\Delta\overline{f} = \sqrt{\sigma^2(\overline{f})} \tag{2.60}$$

Δεν είναι πάντα ευκολο να υπολογιστεί ο ολοκληρωμένος χρόνος αυτοσυσχετισμού, κυρίως σε προσομοιώσεις μεγάλης κλίμακας. Αρα, κάποιος πρέπει να βασιστεί στην ανάλυση binning για τον υπολογισμό σφαλμάτων θεωρρώντας ενα σταθερό αριθμό απο N_{bs} bins. Τελικά, καθώς ο αριθμός των μετρήσεων αυξάνει τα δεδομένα που ανήκουν σε διαφορετικά bin θα γίνουν στατιστικά ανεξάρτητα και η ανάλυση σφαλμάτων θα είναι σωστή.

Μερικές φορές ίσως είναι δυνατό να γίνει μια εκτίμηση της τάξης μεγέθους του τ_{int}. Για την περίπτωση ενός πλέγματος $V = L^d$ κοντά στη κρίσιμη περιοχή, ο ολοκληρωμένος χρόνος αυτοσυσχέτισης αυξάνει ως:

$$\tau_{cpu} = L^{d+z} \tag{2.61}$$

Η ποσότητα z είναι ο δυναμικός κρίσιμος εκθέτης και μπορεί να υπολογιστεί με βάθμιση πεπερασμένου μεγέθους. Αλλες τεχνικές ανάλυσης σφάλματος μπορούν να υλοποιηθούν, όπως η jackknife και η bootstrap αλλα η ανάλυση binning επαρκεί για τα προβλήματα που μελετώνται σε αυτή την διπλωματική.

2.2.0 *Μαθηση Χωρίς Επίβλεψη του $d = 2$ Ising Μοντέλου*

Η κατανομή πιθανότητας που αναπαρίσταται απο απεικονίσεις πραγματικού χώρου του $d = 2$ Ising μοντέλου μπορεί να μοντελοποιηθεί χρησιμοποιώντας την προηγούμενη υλοποίηση των Boltzmann Machines.

Ας υποθέσουμε ένα σύνολο δεδομένων $\{u_i\}$ απο απεικονίσεις σπιν που παράγονται απο μια μαρκοβιανή αλυσίδα Monte Carlo του $d = 2$ Ising μοντέλου για μια πεπερασμένη θερμοκρασία. Αυτό το σύνολο δεδομένων περιγράφεται απο μια κατανομή πιθανότητας q και ο στόχος είναι η ελαχιστοποιηση της Kullback-Leibler απόκλισης ανάμεσα στην κατανομή πιθανότητας q και την κατανομή πιθανότητας p του μοντέλου.

Το νευρωνικό δίκτυο αρχικά εκπαιδεύεται και μετέπειτα χρησιμοποιείται για την παραγωγή προσεγγιστικών απεικονίσεων του Ising μοντέλου μέσω δειγματοληψίας απο την κατανομή ισορροπίας του. Αυτές οι απεικονίσεις χρησιμοποιούνται ακριβώς όπως Monte Carlo δεδομένα και παρατηρήσιμες ποσότητες μπορούν να υπολογιστούν απο αυτές. Το ενδιαφέρον είναι στην μελέτη της κρίσιμης περιοχής και στην παρατήρηση της εξάρτησης της ακρίβειας των αναμενόμενων τιμών σε σχέση με τον αριθμό των κρυφών νευρώνων.

Το νευρωνικό δίκτυο περιγράφεται απο n_v ορατούς νευρώνες και n_h κρυφούς νευρώνες και τα βάρη $\boldsymbol{w}$ είναι οι παράμετροι μεταβολής. Έχουμε υποθέσει ένα επιπλέον νευρώνα για να συμπεριληφθούν τα biases εντός των βαρών και να υπάρχει ταυτόχρονη εκπαίδευση. Για κάθε θερμοκρασία εκπαιδεύεται ένα νευρωνικό δίκτυο σε 100000 απεικονίσεις για τις περιπτώσεις $n_h =$

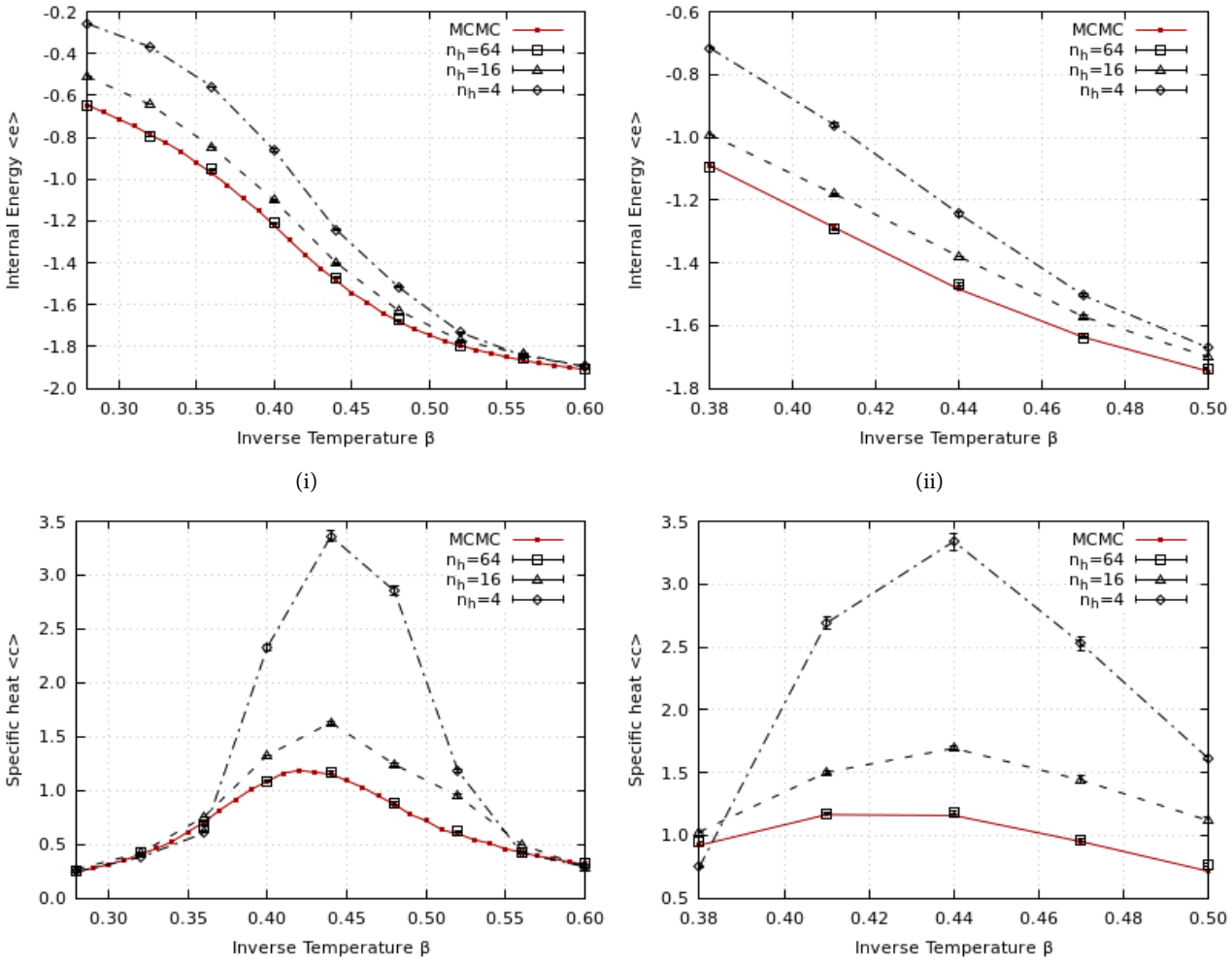
$64, 16, 4$ και για 500 εποχές. Το σύστημα που μελετείται είναι το Ising μοντέλο για $N = L*L = 8*8 = 64$ σε τετραγωνικό πλέγμα. Τα βάρη αρχικοποιούνται ως:

$$\boldsymbol{w} \propto \sqrt{\frac{1}{n_h + n_v}} \tag{2.62}$$

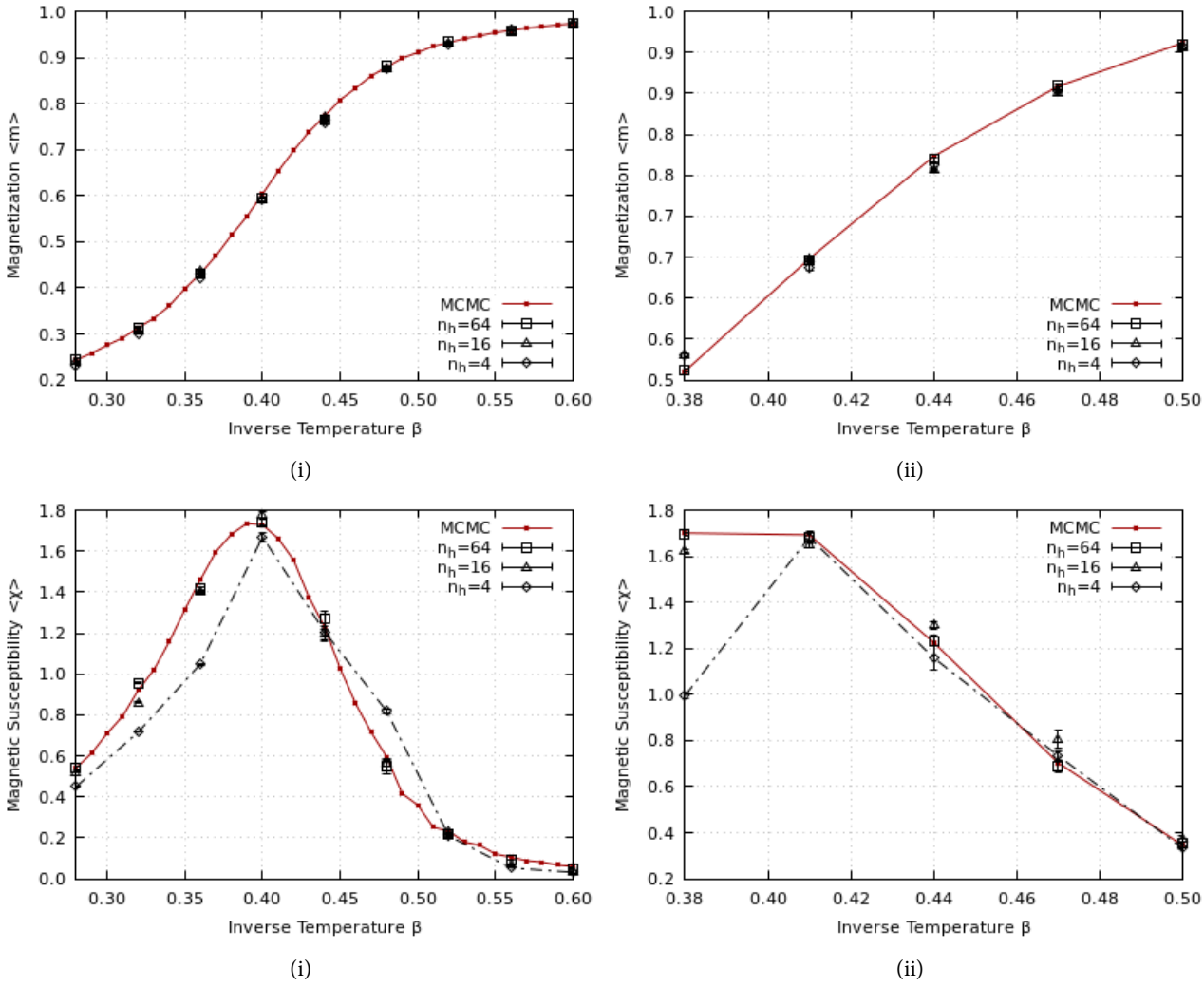
Η μέθοδος contrastive divergence εκτελείται για 20 βήματα και χρησιμοποιείται ενα mini batch απο 50 δείγματα. Ο ρυθμός μάθησης ισούται με $l = 0.01$.

Παρατηρείται απο τον υπολογισμό των παρατηρήσιμων ποσοτήτων οτι το Restricted Boltzmann Machine αναπαράγει απεικονίσεις που δίνουν ακριβέστερα αποτελέσματα μακριά απο την μετάβαση φάσης. ΟΙ παρατηρήσιμες ποσότητες κοντά στην μετάβαση φάσης ειναι ακριβείς για αριθμό κρυφών νευρώνων $n_h = 64$ αλλα όχι για τις άλλες δύο περιπτώσεις $n_h = 16, 4$. Η μαγνήτιση ειναι η μόνη εξαίρεση επειδή το νευρωνικό δίκτυο εκπαιδεύεται σε απεικονίσεις με την μαγνήτιση να ειναι πλήρως κωδικοποιημένη στα δεδομένα. Η ακρίβεια των αποτελεσμάτων κοντά στη μετάβαση φάσης δευτέρας τάξεως έχει μια εξάρτηση στον αριθμό των κρυφών νευρώνων.

Το νευρωνικό δίκτυο έχει πια την δυνατότητα να χρησιμοποιηθεί σε συνδυασμο με δεδομένα Monte Carlo ως ένα βασικό εργαλείο έρευνας αφου είναι δυνατό να ληφθούν σταθερά αποτελέσματα στις περιοχές πλήρης τάξης, αταξιας και στην κρίσιμη περιοχή. Επίσης δίνει την δυνατότητα συμπίεσης του συστήματος μέσα απο την μείωση του αριθμού των κρυφών νευρώνων.



Φιγυρε 2.5: Διαγράμματα της εσωτερικής ενέργειας και της ειδικής θερμότητας ανα σπιν για ενα πλέγμα μεγέθους $N = L*L = 8*8 = 64$. Το Restricted Boltzmann Machine έχει εκπαιδευθει σε δεδομένα απο (i) τον αλγόριθμο Metropolis και (ii) τον αλγόριθμο Wolff. Ειναι εμφανές οτι γενικά το νευρωνικό δίκτυο δίνει καλύτερα αποτελέσματα για αποτελέσματα μακριά απο την κρίσιμη θερμοκρασια $\beta_c \approx 0.4407$ για διαφορετικούς αριθμούς n_h κρυφών στοιχείων. Όταν τα κρυφά στοιχεία είναι ίσα με τα ορατά οι αναμενόμενες τιμές αναπαράγουν σωστα αυτες των Monte Carlo δεδομένων αφού το νευρωνικό δίκτυο έχει την δυνατότητα να μοντελοποιήσει την κατανομή καλύτερα.



Φιγυρε 2.7: Διαγράμματα της μαγνήτισης $< m >$ και της μαγνητικής επιδεκτικότητας $< \chi >$ ανα θέση. Το Restricted Boltzmann Machine έχει εκπαιδευθει σε δεδομένα απο (i) τον αλγόριθμο Metropolis και (ii) τον αλγόριθμο Wolff. Παρατηρούμε οτι για την μαγνήτιση οι αναμενόμενες τιμές που υπολογίζονται απο τις απεικονίσεις του νευρωνικού δικτύου είναι εντός στατιστικού σφάλματος για όλες τις περιπτώσεις κρυφών νευρώνων.

3. Η Ομάδα Επανακανονικοποίησης και τα Δίκτυα Βαθιάς Πεποιθήσεως

3.1.0 Ομάδα Επανακανονικοποίησης Πραγματικού Χώρου

Η Ομάδα Επανακανονικοποίησης είναι μια σημαντική τεχνική στην θεωρητική φυσική για την αντιμετώπιση προβλημάτων πολλαπλής κλίμακας. Ενας μετασχηματισμός επανακανονικοποίησης σε ένα σύστημα ορίζει ένα νέο σύστημα που περιγράφεται απο εναν μικρότερο αριθμό νέων βαθμών ελευθερίας. Είναι δυνατό να εξάγει κάποιος χαρακτηριστικά για αλληλεπιδράσεις μεγάλης κλίμακας, ουσιαστικά εξαλείφοντας βαθμούς αλληλεπίδρασης σε μικρές αποστάσεις.

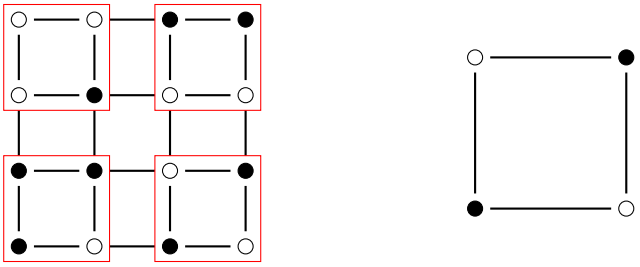
Η επανακανονικοποίηση πραγματικού χώρου είναι μια προσεγγιστική τεχνική που αντικαθιστά σπιν του κανονικού συστήματος με βοηθητικά σπιν, μέσω μιας επαναληπτικής διαδικασίας. Στόχος είναι το μετασχηματισμένο σύστημα να διατηρεί την πληροφορία σε μεγάλη κλίμακα του κανονικού συστήματος. Αυτό ειναι επιτευξιμο απο μια κατάλληλη επιλογή των παραμέτρων που συζευγουν τα βοηθητικά σπιν με αυτα του κανονικού συστήματος. Η επιλογή αυτή βασίζεται στην ελαχιστοποίηση της διαφοράς των ελευθερων ενεργειών των δυο συστημάτων. Η τεχνική μπορεί να εφαρμοστεί ξανά στα βοηθητικά σπιν, δημιουργώντας ενα επιπλέον μετασχηματισμένο σύστημα.

Ας θεωρήσουμε ένα σύνολο απο N δυαδικά σπιν σε ένα πλέγμα, οπου κάθε σπιν έχει μια πιθανή τιμή ± 1. Για ένα δεδομένο σύνολο απο σπίν $\{u_i\}$ με χαμιλτονιανή $H(\{u_i\})$, η πιθανότητα μιας δεδομένης απεικόνισης στην θερμοδυναμική ισορροπία δίνεται απο την κατανομή πιθανότητας Boltzmann:

$$P(\{u_i\}) = \frac{e^{-H(\{u_i\})}}{Z} \tag{3.1}$$

οπου η αντίστροφη θερμοκρασία β είναι ίση με ένα. Η σύνάρτηση επιμερισμού Z δίνεται τότε απο τη σχεση:

$$Z = Tr_{\{u_i\}} e^{-H(\{u_i\})} = \sum_{\{u_i\}} e^{-H(\{u_i\})} \tag{3.2}$$



Σχήμα 3.1: Ενα $N = 4 * 4 = 16$ σύστημα απο σπιν που μετασχηματίζεται σε ένα $N' = 2 * 2$ χρησιμοποιώντας την διαδικασία blocking. Η τιμή του νεου h_j σπιν επιλέγεται σύμφωνα με την πλειοψηφία των τιμών των συμπεριλαμβανομένων αρχικών σπιν. Οταν οι τιμές είναι ίσες, η επιλογή γίνεται τυχαία.

και η ελεύθερη ενέργεια του συστήματος ισούται με:

$$F^u = -\log Z = -\log\left(Tr_{\{u_i\}} e^{-H(\{u_i\})}\right) \tag{3.3}$$

Μια γενικευμένη χαμιλτονιανή του μοντέλου θα είχε εξάρτηση σε ένα σύνολο απο σταθερές σύζευξης $\boldsymbol{K} = \{K_s\}$ που περιγράφουν αλληλεπιδράσεις ανάμεσα σε βαθμούς ελευθερίας πολλαπλών τάξεων:

$$H[\{u_i\}] = -\sum_i K_i u_i - \sum_{i,j} K_{ij} u_i u_j - \sum_{i,j,k} K_{ijk} u_i u_j u_k + \dots \tag{3.4}$$

Είναι δυνατό να εισάγει κάποιος ένα νεο σύνολο $\{h_j\}$ απο M βοηθητικά ή αλλιώς κρυφά σπιν, με $M < N$ και θα υποθέσουμε οτι μια blocking διαδικασία έχει επιλεγεί για τον μετασχηματισμό του αρχικού συστήματος. Τα σπιν μπορούν να χωριστούν σε τετράγωνα, και ορίζεται ένας παραγοντας ανακλιμάκωσης b ο οποιος μειώνει το μέγεθος του αρχικού πλέγματος κατα b σε κάθε διάσταση. Κάθε τετράγωνο θα αντιστοιχείται σε ένα νεο σπιν $\{h_j\}$, η τιμή του οποίου θα είναι ± 1 και θα αποφασιστεί σύμφωνα με την πλειοψηφία των τιμών των σπιν που συμπεριλαμβάνονται στο τετράγωνο. Ενας παράγοντας ανακλιμάκωσης 2 θα μείωνε το σύνολο των σπιν κατα 2^d οπου d ειναι η διαστατικότητα του συστήματος.

Οι αλληλεπιδράσεις μεταξύ των νεων μεταβλητών $\{h_j\}$ έχουν μια εξάρτηση απο τις αλληλεπιδράσεις στο κανονικό σύστημα των $\{u_i\}$ σπιν και περιγράφονται απο μια νέα χαμιλτονιανή με ένα σύνολο απο $\{K^{'}\}$ σταθερές σύζευξης οι οποιες έχουν αντιστοιχιθεί ως $\{K\} \to \{K^{'}\}$:

$$H^{RG}(\{h_j\}) = -\sum_i K_i^{'} h_i - \sum_{i,j} K_{ij}^{'} h_i h_j - \sum_{i,j,k} K_{ijk}^{'} h_i h_j h_k + \dots \tag{3.5}$$

Με ένα μετασχηματισμό επανακανονικοποίησης, εξαλείφονται τα αρχικά σπιν $\{u_i\}$ και δημιουργείται μια νέα περιγραφή του συστήματος μέσα απο τα νεα κρυφά σπιν $\{h_j\}$. Μια συνάρτηση $T_\lambda(\{u_i\}, \{h_j\})$, που εξαρτάται απο ένα σύνολο παραμέτρων $\{\lambda\}$ μπορεί να οριστεί και εκφράζει αλληλεπιδράσεις μεταξύ των κανονικών σπιν και αυτών του μετασχηματισμένου συστήματος. Επίσης ορίζει μια χαμιλτονιανή για τα $\{h_j\}$ μέσω της σχέσης:

$$e^{-H_\lambda^{RG}(\{h_j\})} = Tr_{\{u_i\}} e^{T_\lambda(\{u_i\},\{h_j\}) - H(\{u_i\})} \tag{3.6}$$

Παρομοίως η ελεύθερη ενέργεια ορίζεται ως:

$$F_{\lambda}^{h} = -\log\left(Tr_{\{h_j\}} e^{-H_{\lambda}^{RG}(\{h_j\})}\right) \tag{3.7}$$

Όπως αναφέρθηκε προηγουμένως, το σύνολο των παραμέτρων μεταβολής $\{\lambda\}$ πρέπει να επιλεγεί έτσι ώστε να ελαχιστοποιεί την διαφορά της ελεύθερης ενέργειας των δύο συστημάτων $\Delta F = F_h^{\lambda} - F^u$. Με αυτό τον τρόπο επιβεβαιώνουμε οτι το μετασχηματισμένο σύστημα διατηρεί την πληροφορία σε μεγάλη κλίμακα του κανονικό συστήματος. Παρατηρούμε ότι:

$$\Delta F = 0 \Longleftrightarrow Tr_{\{h_j\}} e^{T_{\lambda}(\{u_i\},\{h_j\}} = 1 \tag{3.8}$$

Ενα μετασχηματισμός επανακανονικοποίησης αποκαλείται ακριβής όταν:

$$Tr_{\{h_j\}} e^{T_{\lambda}(\{u_i\},\{h_j\})} = 1 \tag{3.9}$$

Στην επόμενη ενότητα θα δούμε πως είναι δυνατό να αντιστοιχιθεί η επανακανονικοποίηση πραγματικού χώρου με την συμπίεση δεδομένων των Restricted Boltzmann Machine.

3.2.0 *Μια Αντιστοιχία Ανάμεσα στην Ομάδα Επανακανονικοποίησης και τα Βαθιά Νευρωνικά Δίκτυα*

Για να αντιστοιχιθεί η ομάδα επανακανονικοποίσης με τα βαθιά νευρωνικά δίκτυα πρέπει να γίνει μια κατάλληλη επιλογή για τον τελεστή $T_{\lambda}(\{u_i\},\{h_j\})$.

Ο τελεστής $T_{\lambda}(\{u_i\},\{h_j\})$ περιγράφει αλληλεπιδράσεις ανάμεσα σε σπιν του κανονικού και του μετασχηματισμένου συστήματος. Η συνάρτηση ενέργειας $E(\{u_i\},\{h_j\})$ που ορίζεται στην σχέση 1.32 έχει τον ίδιο ρόλο σε ένα Restricted Boltzmann Machine. Ο τελεστής $T_{\lambda}(\{u_i\},\{h_j\})$ πρεπει να επιλεγεί ως:

$$T(\{u_i\},\{h_j\}) = -E(\{u_i\},\{h_j\}) + H[\{u_i\}] \tag{3.10}$$

Χρησιμοποιώντας την απο κοινού συνάρτηση κατανομής $p_{\lambda}(\{u_i\},\{h_j\})$ του Restricted Boltzmann Machine που ορίζεται στην 1.31, μπορούμε να λάβουμε εκφράσεις της χαμιλτονιανής για τους ορατούς και κρυφούς νευρώνες του Restricted Boltzmann Machine μέσα απο τις περιθώριες κατανομές:

$$p_{\lambda}(\{u_i\}) = \sum_{\{h_j\}} p_{\lambda}(\{u_i\},\{h_j\}) = Tr_{h_j} p_{\lambda}(\{u_i\},\{h_j\}) = \frac{e^{-H_{\lambda}^{RBM}[\{u_i\}]}}{\mathcal{Z}} \tag{3.11}$$

$$p_{\lambda}(\{h_j\}) = \sum_{\{u_j\}} p_{\lambda}(\{u_i\},\{h_j\}) = Tr_{u_i} p_{\lambda}(\{u_i\},\{h_j\}) = \frac{e^{-H_{\lambda}^{RBM}[\{h_j\}]}}{\mathcal{Z}}, \tag{3.12}$$

Οπως αναφέρθηκε προηγουμένως, ο τελεστής $T_\lambda(\{u_i\},\{h_j\})$ ορίζει μια χαμιλτονιανή για τα βοηθητικά σπιν μέσα απο την έκφραση 3.6. Διαιρώντας και τα δύο μέλη της 3.6 με την συνάρτηση επιμερισμού $\mathcal{Z}$ του Restricted Boltzmann Machine:

$$\frac{e^{-H_\lambda^{RG}(\{h_j\})}}{\mathcal{Z}} = \frac{Tr_{\{u_i\}} e^{T_\lambda(\{u_i\},\{h_j\})-H(\{u_i\})}}{\mathcal{Z}} \tag{3.13}$$

Αντικαθιστώντας στην παραπάνω συνάρτηση την έκφραση για τον τελεστή 3.10 έχουμε:

$$\frac{e^{-H_\lambda^{RG}(\{h_j\})}}{\mathcal{Z}} = Tr_{\{u_i\}} \frac{e^{-E(\{u_i\},\{h_j\})}}{\mathcal{Z}} = p_\lambda(\{h_j\}) \tag{3.14}$$

Χρησιμοποιώντας την Χαμιλτονιανή του Restricted Boltzmann Machine για τους κρυφούς νευρώνες:

$$\frac{e^{-H_\lambda^{RG}(\{h_j\})}}{\mathcal{Z}} = \frac{e^{-H_\lambda^{RBM}[\{h_j\}]}}{\mathcal{Z}} \Rightarrow H_\lambda^{RG}[\{h_j\}] = H_\lambda^{RBM}[\{h_j\}] \tag{3.15}$$

Το παραπάνω αποτέλεσμα ορίζει μια ισότητα για την Χαμιλτονιανή του μετασχηματισμένου συστήματος και την χαμιλτονιανή των κρυφών νευρώνων των Restricted Boltzmann Machines. Ισοδύναμα, η περιθώρια κατανομή $p_\lambda(\{h_j\})$ των κρυφών σπιν του Restricted Boltzmann Machine είναι μια Boltzmann κατανομή πιθανότητας με μια Χαμιλτονιανή $H_\lambda^{RG}[\{h_j\}]$. Ο τελεστής $T_\lambda(\{u_i\},\{h_j\})$ είναι μια προσέγγιση της υπο συνθήκη πιθανότητας των κρυφών σπιν για δεδομένα ορατά σπιν:

$$\begin{aligned} e^{T(\{u_i\},\{h_j\})} &= e^{-E(\{u_i\},\{h_j\})+H[\{u_i\}]} \\ &= e^{-E(\{u_i\},\{h_j\})} e^{H[\{u_i\}]} \frac{e^{-H_\lambda^{RBM}[\{u_i\}]}}{e^{-H_\lambda^{RBM}[\{u_i\}]}} \\ &= \frac{p_\lambda(\{u_i\},\{h_j\})}{p_\lambda(\{u_i\})} e^{H[\{u_i\}]-H_\lambda^{RBM}[\{u_i\}]} \\ &= p_\lambda(\{h_j\}|\{u_i\}) e^{H[\{u_i\}]-H_\lambda^{RBM}[\{u_i\}]} \end{aligned} \tag{3.16}$$

Οταν ικανοποιείται η συνθήκη 3.9 για έναν ακριβή μετασχηματισμό επανακανονικοποίησης , η χαμιλτονιανή του κανονικού συστήματος είναι ίση με την χαμιλτονιανή του Restricted Boltzmann Machine $H[\{u_i\}] = H_\lambda^{RBM}[\{u_i\}]$.

$$\begin{aligned} Tr_{h_j} e^{T(\{u_i\},\{h_j\})} &= Tr_{h_j} \frac{p_\lambda(\{u_i\},\{h_j\})}{p_\lambda(\{u_i\})} e^{H[\{u_i\}]-H_\lambda^{RBM}[\{u_i\}]} \\ &= \frac{p_\lambda(\{u_i\})}{p_\lambda(\{u_i\})} e^{H[\{u_i\}]-H_\lambda^{RBM}[\{u_i\}]} \\ &= e^{H[\{u_i\}]-H_\lambda^{RBM}[\{u_i\}]} \\ &= 1 \\ &\Rightarrow H[\{u_i\}] = H_\lambda^{RBM}[\{u_i\}] \end{aligned} \tag{3.17}$$

Επιπλέον μπορούμε να λαβουμε μια έκφραση για τον τελεστή $T(\{u_i\},\{h_j\})$ και την ακριβή υπο συνθήκη πιθανότητα αφού $H[\{u_i\}] - H_\lambda^{RBM}[\{u_i\}] = 0$. Η κατανομή μεταβολών $p_\lambda(u_i)$ μπορεί τοτε να αναπαράγει πλήρως την κατανομή που αντιστοιχεί στα δεδομένα $P(\{u_i\})$ και η Kullback-Leibler απόκλιση είναι ίση με μηδέν $D_{KL}(P(\{u_i\})|p_\lambda(\{u_i\})) = 0$.

Η παραπάνω προσέγγιση έχει θεμελιωθεί στον μηδενισμό της διαφοράς των ελεύθερων ενεργειών μέσα απο ακριβείς μετασχηματισμούς επανακανονικοποίησης και στην περιγραφή συστημάτων μεσα απο χαμιλτονιανές. Ενα πρόβλημα λύνεται με μηχανική μάθηση μέσω καποιων προσεγγίσεων για την ελαχιστοποίηση της Kullback-Leibler απόκλισης. Αυτές οι προσεγγίσεις δημιουργούν μια διαφορετική διαδικασία για τον μετασχηματισμό του συστήματος. Τελικά, είναι σημαντικό να αναφερθει οτι η ακριβής μορφής της ενέργειας $E(\{u_i\},\{h_j\})$ δεν μεταβάλλει τα αποτελέσματα και κάθε είδος Boltzmann Machine μπορεί να χρησιμοποιηθεί.

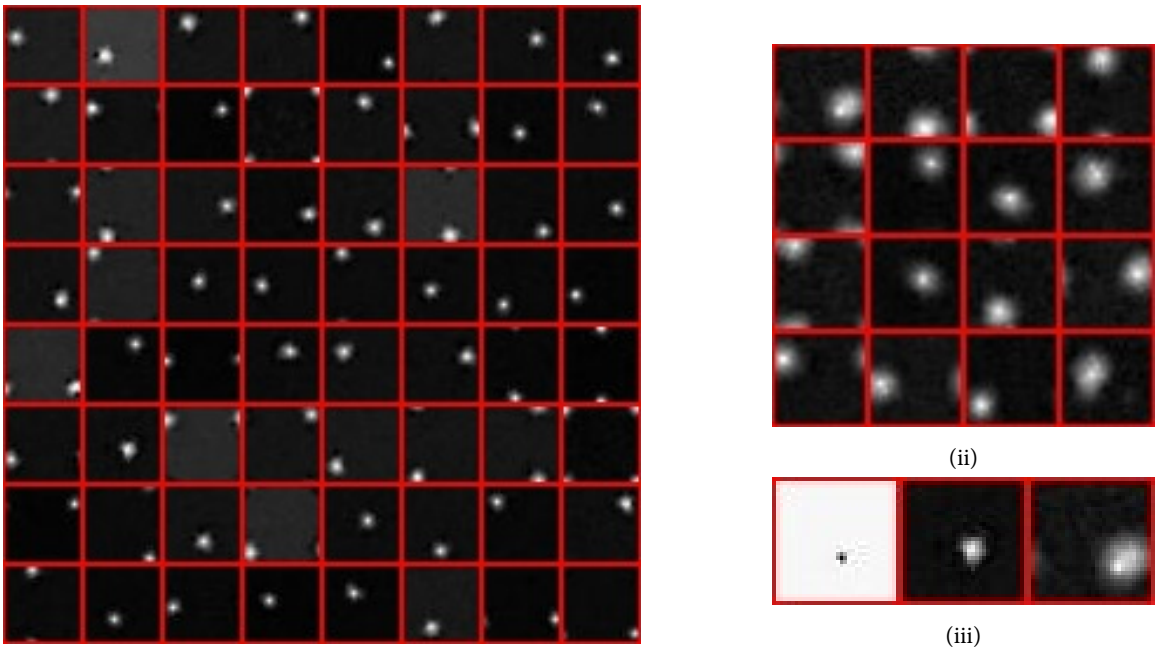
Υλοποιούμε ενα Deep Belief δίκτυο μέ ένα ορατό επίπεδο μεγέθους $n_v = 1024$ και τρία κρυφά επίπεδα μεγέθους $n_h = 256, 64, 16$ που εκπαιδεύεται σε απεικονίσεις του διδιάστατου Ising που αποτελούνται απο 40000 βήματα προσομοίωσης για θερμοκρασία $\beta = 0.43$. Το νευρωνικό δίκτυο εκπαιδεύεται για 400 εποχές με ρυθμό μάθησης $l = 0.1$, L1 weight decay 0.002, μέγεθος mini batch 100 και ορμή 0.5.

Για μια καλύτερη παρατήρηση των αποτελεσμάτων σχεδιάζονται τα *ρεσεπτιε φιελδς* του νευρωνικού δικτύου μέσα απο την αναδρομική σχέση:

$$r^l = r^{(l-1)}W^l, l > 1 \tag{3.18}$$

οπου $r^1 = W^1$. Άρα, αποκτάται μια μέτρηση του τρόπου με τον οποιο ένας κρυφός νευρώνας επηρεάζει νευρώνες στο ορατό στρώμα.

Μια παρόμοια διαδικασία με την blocking υλοποιείται απο το νευρωνικό δίκτυο. Κάθε στοιχείο σε ένα δεδομένο κρυφό στρώμα συζεύγεται με μια συστοιχία απο σπιν του ορατού στρώματος παρόμοιου μεγέθους. Το μέγεθος των συστοιχιών αυξάνεται με τον ρυθμό συμπίεσης, κατα αντιστοιχία με τον παράγοντα ανακλιμάκωσης της ομάδας επανακανονικοποίησης. Η σημαντική διαφορά είναι οτι το νευρωνικό δίκτυο οργανώνεται αυτόνομα κατα την εκπαίδευση.



(i) (ii) (iii)

Φιγυρε 3.3: Αναπαράσταση των receptive fields για το (i) δευτερο και (ii) τρίτο κρυφό επιπεδο. Συμπεριλαμβάνεται και μια αναπαράσταση για αντιπροσωπευτικές περιπτώσεις και των τριών (iii) κρυφών επιπέδων. Το μέγεθος των συζευγμένων νευρώνων αυξάνει με το μέγεθος συμπίεσης. Αυτή είναι η ιδια ιδέα με διαδοχικές επαναλήψεις ενος spin blocking μετασχηματισμού επανακανονικοποίησης.

3.3.0 Σπίν Blocking στο Διδιάστατο Ising

Το κύριο πρόβλημα με την ομάδα επανακανονικοποίησης πραγματικού χώρου είναι η υπόθεση οτι το μετασχηματισμένο σύστημα εμφανίζεται με τις σωστές πιθανότητες Boltzmann,

Υποθέτουμε ένα πλέγμα με μέγεθος σε μια διάσταση L και Monte Carlo απεικονισεις για μια δεδομένη θερμοκρασία T. Η ανακλιμάκωση του συστήματος κατα ενα παράγοντα b θα δημιουργήσει ένα νέο πλέγμα μεγέθους L' με:

$$L' = \frac{L}{b} \tag{3.19}$$

Οι απεικονίσεις του κανονικού συστήματος εμφανίζονται με τις σωστες πιθανότητες Boltzmann. Δεν μπορούμε όμως να ισχυριστούμε οτι οι καταστάσεις που αντιστοιχούν στο νέο πλέγμα L' εμφανίζονται με τις σωστές πιθανότητες Boltzmann που αντιστοιχούν στην ίδια θερμοκρασία T του αρχικού συστήματος. Υποθέτουμε όμως οτι κάτι τέτοιο ισχύει, εισάγοντας σφάλματα στη μέθοδο που δεν μπορούν να ελεγχθούν.

Εφόσον ένας μετασχηματισμός επανακανονικοποίησης πρέπει να διατηρεί τα χαρακτηριστικά του συστήματος σε μεγάλη κλίμακα, το μήκος συσχετισμού ξ θα έπρεπε να είναι περίπου ίδιο. Η μείωση του αριθμού των σπιν κατα b^2 τότε συνεπάγεται οτι το μήκος συσχετισμού του μετασχηματισμένου συστήματος σε όρους πλεγματικής απόστασης πρέπει να είναι:

$$\xi' = \frac{\xi}{b} \tag{3.20}$$

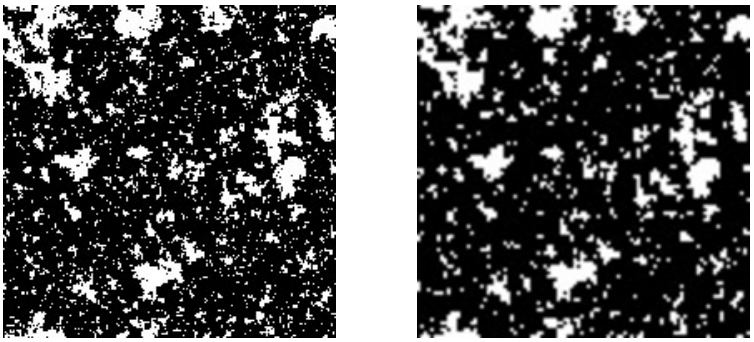
Για την περίπτωση ενός $b = 2$ παράγοντα ανακλιμάκωσης το μήκος συσχετισμού πρέπει να είναι $\xi' = \xi/2$. Άρα οι απεικονίσεις του μετασχηματισμένου συστήματος πρέπει να αντιστοιχούν σε καταστάσεις μιας διαφορετικής θερμοκρασιας T' αφού το μήκος συσχετισμού μεταβάλλεται για διαφορετικές τιμές του T.

Παρομοίως, παρατηρήσιμες ποσότητες που έχουν εξαρτώνται απο την θερμοκρασία οπως η εσωτερική ενέργεια ανα σπιν u θα πρέπει να δίνουν διαφορετικές τιμές όταν υπολογίζονται για το αρχικο και το μετασχηματισμένο σύστημα. Εφόσον οι απεικονίσεις του ανακλιμακωμένου συστήματος αντιστοιχούν σε καταστάσεις που έχουν δειγματοληφθεί για θερμοκρασία T' η εσωτερική ενέργεια ανα σπιν για αυτό το σύστημα θα είναι u'.

Στη κρίσιμη θερμοκρασία ωστόσο τα μήκη συσχετισμού του κανονικού και του μετασχηματισμένου συστήματος έιναι ίσα:

$$\xi = \xi' \; and \; T = T' = T_c \tag{3.21}$$

Όλες οι άλλες εντατικές ιδιότητες όπως η μαγνήτιση, η ειδική θερμότητα, η μαγνητική επιδεκτικότητα και η εσωτερική ενέργεια ανα σπιν είναι επίσης ίσες. Χρησιμοποιώντας re-weighting τεχνικές είναι δυνατό να υπολογιστούν οι παραπάνω ποσότητες επεκτείνοντας σε ένα ευρος θερμοκρασιών για το αρχικό και το μετασχηματισμένο σύστημα. Είναι σημαντικό να χρησιμοποιηθούν οι



(i) (ii)

Σχήμα 3.4: Ένας spin blocking μετασχηματισμός επανακανονικοποίησης με παράγοντα ανακλιμάκωσης $b = 2$ σε ένα πλέγμα μεγέθους $N = 200 * 200$ χρησιμοποιώντας τον κανόνα πλειοψηφίας. Το (ii) μετασχηματισμένο σύστημα το οποιο έχει μεγεθυνθεί για ευκολότερη σύγκριση διατηρεί τα ποιοτικά χαρακτηριστικά μεγάλης κλίμακος του (i) αρχικού συστήματος

τιμές του μετασχηματισμένου συστήματος ως παρατηρήσιμες ποσότητες του αρχικού συστήματος κατα το re-weighting. Αν δεν γνωρίζαμε ήδη το κρίσιμο σημείο, η μέθοδος μπορεί να χρησιμοποιηθεί επαναληπτικα σε θερμοκρασίες υπολογισμένες σε κάθε βήμα ως κρισιμες μέχρι να συγκλινει.

Τα φαινόμενα πεπερασμένου μεγέθους γενικά εισάγουν σφάλματα και μεγαλύτερα συστήματα επιτρέπουν καλύτερους υπολογισμούς. Επιπλέον εισάγουν σφάλματα λόγω της διαφοράς στο μέγεθος του αρχικού και μετασχηματισμένου συστήματος και ένα σύστημα που έχει το ίδιο μέγεθος σε κάθε διάσταση με το μετασχηματισμένο ίσως χρειαστεί να προσομοιωθεί ξεχωριστά για τους υπολογισμούς. Ανάλογα το μοντέλο που μελετάται η υπόθεση οτι οι απεικονίσεις του μετασχηματισμένου συστήματος αντιστοιχούν σε καταστάσεις κάποιας θερμοκρασίας T' μπορεί να είναι μια πολύ σημαντική πηγή σφάλματος.

Ο κύριος στόχος εφαρμογής της ομάδας επανακανονικοποίησης στο Ising μοντέλο είναι ο υπολογισμός των κρίσιμων εκθετών. Μπορει να γίνει μια αντιστοιχία ανάμεσα στην θερμοκρασία T' του μετασχηματισμένου συστήματος και στην θερμορασία T του αρχικού συστήματος. Ορίζοντας ως u και u' τις εσωτερικές ενέργειας αν σπίν των δύο συστημάτων έχουμε:

$$u'(T) = u(T') \tag{3.22}$$

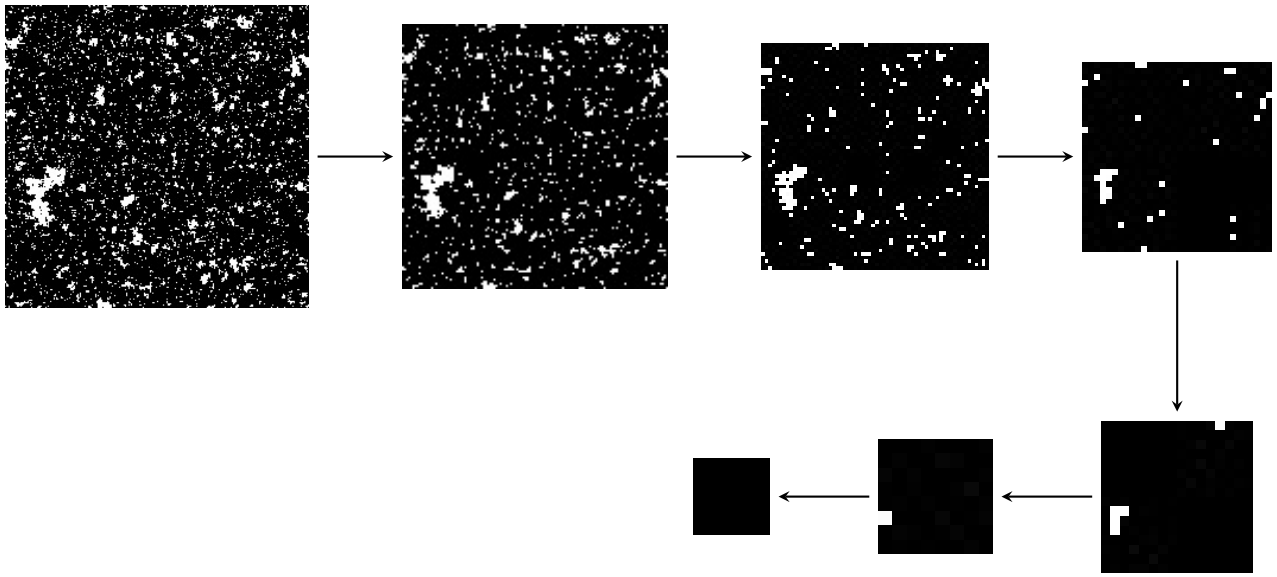
Μπορεί να επιτευχθεί μια αντιστοιχία ανάμεσα στις δύο θερμοκρασίες T' και T ως:

$$T' = u^{-1}(u'(T)) \tag{3.23}$$

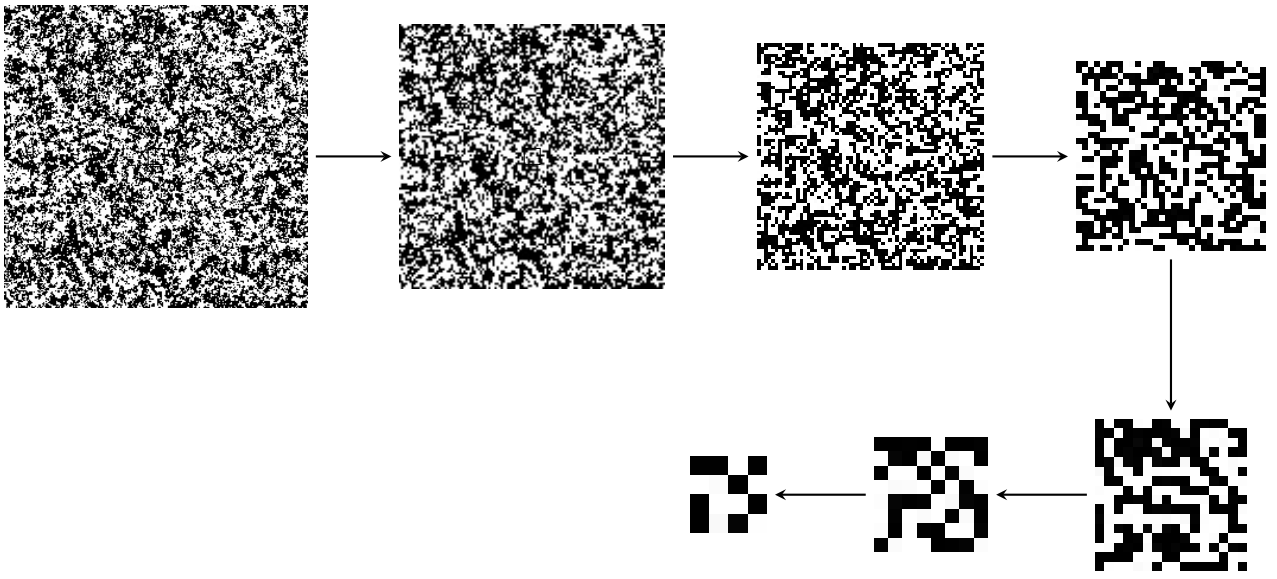
Ειναι ακριβώς αυτή η αντιστοιχία που επιτρέπει τον υπολογισμό των κρίσιμων εκθετών. Η κρίσιμη θερμοκρασία είναι ένα σταθερό σημείο του μετασχηματισμού και αποκαλείται κρίσιμο σταθερο σημείο. Για θερμοκρασίες $T > T_c$ η ανακλιμακωμένη θερμοκρασία είναι μεγαλύτερη απο την θερμοκρασία του αρχικού συστήματος, $T' > T$. Αντίστοιχα $T' < T$ για $T < T_c$. Ενας μετασχηματισμός της ομάδας επανακανονικοποίησης χαρακτηρίζεται τότε απο μια ροή στον χώρο παραμέτρων που για αυτή την περίπτωση είναι μονοδιάστατος και οδηγεί την θερμοκρασία μακριά απο το κρίσιμο σημείο.

Αυτή η συμπεριφορά είναι αναμενόμενη. Για, παράδειγμα η παρουσία συ-

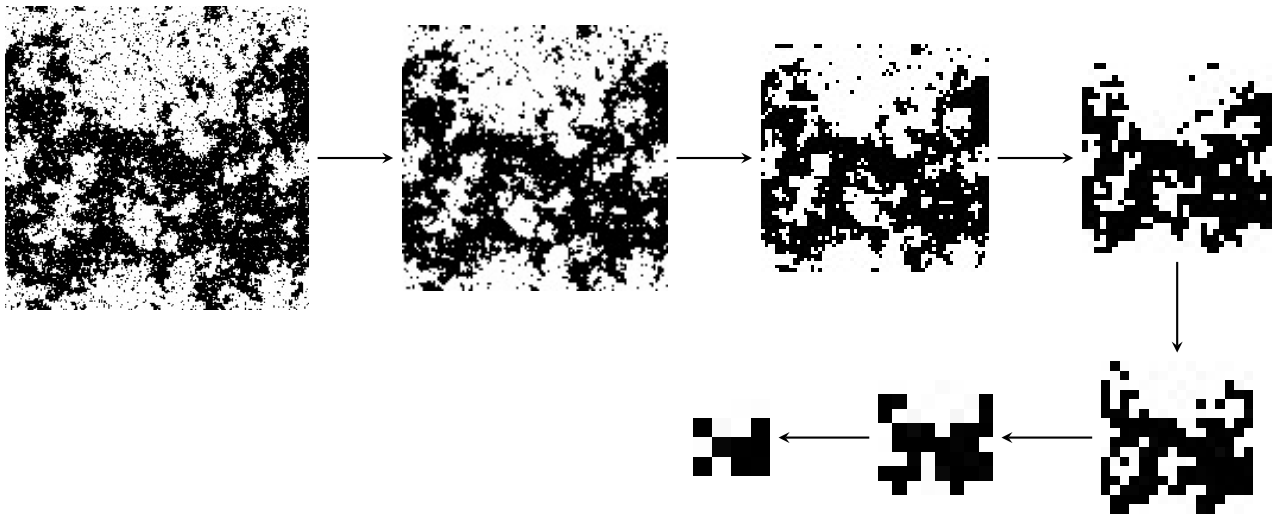
στοιχιών μεγέθους ξ σε θερμοκρασίες μεγαλύτερες της T_c συνεπάγεται οτι ο μετασχηματισμός επανακανονικοποίησης θα σχηματίσει συστοιχίες μικρότερου μεγέθους με $\xi' = \xi/b$. Οι ανακλιμακωμένες απεικονίσεις θα είναι τυπικές μιας υψηλότερης θερμοκρασίας $T' > T$. Για την περίπτωση του μετασχηματισμού επανακανονικοποίησης με $T < T_c$ μια απεικόνιση του συστήματος περιγράφεται απο σπιν τα οποία δείχνουν κυρίως προς μια κατεύθυνση. Κάθε επιπλέον μετασχηματισμός επανακανονικοποίησης θα εξαφάνιζε σπιν που δείχνουν προς την άλλη κατεύθυνση και αποτελούν μια μειονότητα. Οι τελικές απεικονίσεις θα έχουν τότε περισσότερα σπιν που δείχνουν προς την ίδια κατεύθυνση με αυτα του αρχικού συστήματος. Αρα το σύστημα οδηγείται σε πλήρη τάξη και σε χαμηλότερες θερμοκρασίες $T' < T$.



Φιγυρε 3.5: Διαδοχικοί spin blocking μετασχηματισμοί επανακανονικοποίησης χρησιμοποιώντας έναν παράγοντα ανακλιμάκωσης $b = 2$ για ένα Ising μοντέλο μεγέθους $N = 256 * 256$ σε θερμοκρασία $\beta = 0.45$. Κάθε μετασχηματισμός επανακανονικοποίησης ωθεί το σύστημα σε υψηλότερες θερμοκρασίες και αρα σε απεικονίσεις πλήρης τάξεως. Ενα σπιν στο τελικό σύστημα αντιπροσωπεύει 4096 σπιν του αρχικού συστήματος.



Φιγυρε 3.6: Το ίδιο σύστημα οπως παραπάνω για θερμοκρασία $\beta = 0.36$. Κάθε μετασχηματισμός επανακανονικοποίησης οδηγεί το σύστημα σε χαμηλότερες θερμοκρασίες και άρα σε πλήρη αταξία.



Φιγυρε 3.7: Το ίδιο σύστημα οπως παραπάνω για την κρίσιμη θερμοκρασία $\beta_c = 0.4407$. Το σύστημα παραμένει στην ίδια θερμοκρασία άκόμα και μετά απο τους μετασχηματισμούς επανακανονικοποίησης.

3.3.1 *Υπολογισμός των Κρίσιμων Εκθετών*

Ας υποθεσουμε ξανά τον εκθέτη ν

$$\xi \sim |t|^{-\nu} \tag{3.24}$$

Η τιμή t είναι η ανηγμένη θερμοκρασία:

$$t = \frac{T - T_c}{T_c} \tag{3.25}$$

Το μήκος συσχετισμού ξ' του ανακλιμακωμένου συστήματος περιγράφεται επίσης απο την ίδια σχέση, εκτός του οτι τώρα θεωρούμε την θερμοκρασία T':

$$\xi' \sim |t'|^{-\nu} \tag{3.26}$$

Διαιρώντας την σχεση 3.24 με την 3.26 και χρησιμοποιώντας οτι $\xi' = \xi/b$, έχουμε:

$$\left(\frac{t}{t'}\right) = b \tag{3.27}$$

Οι εκφράσεις για τους κρίσιμους εκθέτες έχουν νόημα σε μια περιοχή θερμοκρασιών κοντά στο κρίσιμο σημείο. Άρα χρειαζόμαστε μια σχέση ανάμεσα στα T' και T κοντά στην T_c και αυτο είναι δυνατό με ένα ανάπτυγμα Taylor γύρω απο την T_c:

$$T' - T_c = (T - T_c)\frac{dT'}{dT}\bigg|_{T_c} \tag{3.28}$$

Χρησιμοποιώντας την 3.25 και αντικαθιστώντας τα παραπάνω στην 3.27 έχουμε μια έκφραση για τον κρίσιμο εκθέτη ν:

$$\nu = \frac{\log b}{\log \frac{dT'}{dT}\big|_{T_c}} \tag{3.29}$$

Αν και υπάρχουν σχέσεις βάθμισης ανάμεσα στους κρίσιμους εκθέτες, υπάρχουν περιπτώσεις για τις οποίες πρεπει να γίνουν Monte Carlo προσομοιώσεις για να ελεγχθούν οι σχέσεις βάθμισης. Αρα είναι σημαντικό να υπάρχουν εκφράσεις για την μέτρηση αυτών των εκθετών.

Για την περίπτωση της μαγνήτισης ανα σπιν ο κρίσιμος εκθέτης β είναι:

$$m \sim |t|^{\beta} \tag{3.30}$$

Χρησιμοποιώντας την σχέση 3.24:

$$m \sim \xi^{-\beta/\nu} \tag{3.31}$$

Θεωρώντας ένα μετασχηματισμό επανακανονικοποίσης, το ανακλιμακώμενο σύστημα θα έχει μια μαγνήτιση m' για την οποία:

$$m' \sim \xi'^{-\beta/\nu} \tag{3.32}$$

Ανάλογα με την ίδια διαδικασία όπως παραπάνω, διαιρώντας τις μαγνητίσεις του αρχικού και του ανακλιμακώμενου συστήματος και χρησιμοποιώντας την έκφραση για τα μήκη συσχετισμού, έχουμε:

$$\frac{m'}{m} = b^{\beta/\nu} \tag{3.33}$$

$$\frac{\beta}{\nu} = \frac{\log \frac{m'}{m}}{\log b} \tag{3.34}$$

Η σχέση 3.30 ισχύει για ένα άπειρο σύστημα. Χρησιμοποιώντας τον κανόνα L' Hôpital παίρνουμε την σχέση:

$$\frac{m'}{m} = \frac{dm'/dT}{dm/dT} = \frac{dm'}{dm} \tag{3.35}$$

Η τελική έκφραση για τον κρίσιμο εκθέτη β είναι:

$$\frac{\beta}{\nu} = \frac{\log \frac{dm'}{dm}\big|_{T_c}}{\log b} \tag{3.36}$$

η οποία είναι ανώτερη απο την 3.34 αφού ο όρος dm'/dm δεν διακυμαίνεται πολυ για διαφορετικού μεγέθους συστήματα.

Ανάλογες εκφράσεις μπορούν να υπολογιστούν για τους κρίσιμους εκθέτες α και γ:

$$\frac{\alpha}{\nu} = -\frac{\log \frac{dc'}{dc}\big|_{T_c}}{\log b} \tag{3.37}$$

$$\frac{\gamma}{\nu} = -\frac{\log \frac{d\chi'}{d\chi}\big|_{T_c}}{\log b} \tag{3.38}$$

Η μαγνήτιση του μοντέλου σχετίζεται επίσης με τον κρίσιμο εκθέτη δ για εξωτερικό πεδίο B:

$$m \sim B^{1/\delta} \tag{3.39}$$

Ορίζουμε τώρα έναν κρίσιμο εκθέτη θ που εκφράζει τον τρόπο με τον οποίο το μήκος συσχετισμού διακυμαίνεται καθώς $B \to 0$ στην T_c. Ακολουθώντας την ίδια διαδικασία για τις 3.29 και 3.36 έχουμε:

$$\xi \sim |B|^{-\theta} \tag{3.40}$$

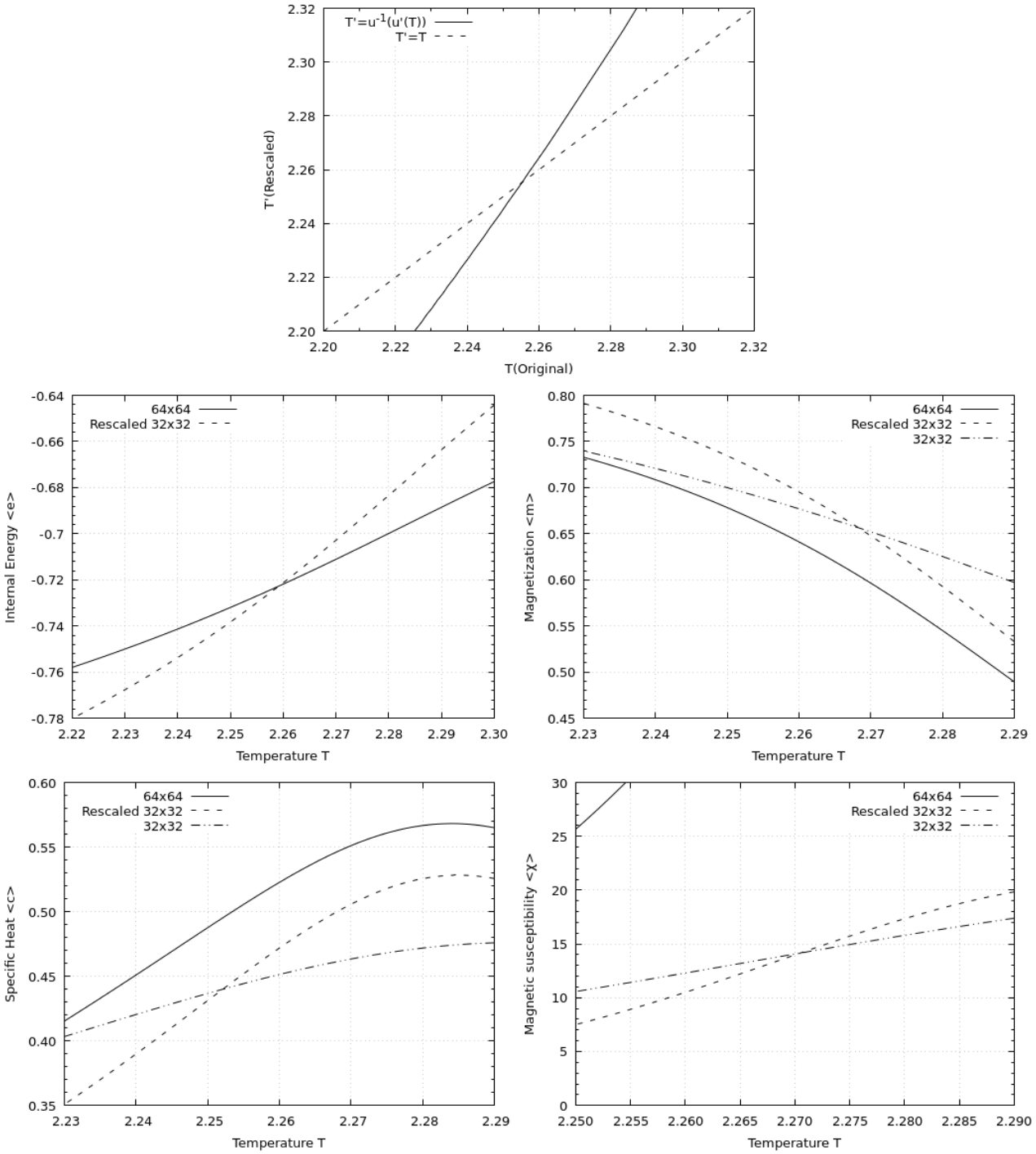
$$\theta = \frac{\log b}{\log \frac{dB'}{dB}\Big|_{B=0}} \tag{3.41}$$

$$\frac{1}{\theta\delta} = \frac{\log \frac{dm'}{dm}\Big|_{B=0}}{\log b} \tag{3.42}$$

Οι παραπάνω σχέσεις δίνουν:

$$\delta = \frac{\log \frac{dB'}{dB}\Big|_{B=0}}{\log \frac{dm'}{dm}\Big|_{B=0}} \tag{3.43}$$

Οι εκθέτες ν και θ που περιγράφουν την απόκλιση του μήκους συσχετισμού ξ σε όρους της κρίσιμης θερμοκρασίας και του εφαρμοζόμενου εξωτερικού πεδίου αντίστοιχα σχετίζονται με τους relevant operators των μετασχηματισμών ομάδας επανακανονικοποίσης και οι υπόλοιποι κρίσιμοι εκθέτες μπορούν να υπολογιστούν απο αυτούς με σχέσεις βάθμισης.



Φιγυρε 3.9: Διαγράμματα για τις παρατηρήσιμες ποσότητες του αρχικού συστήματος $N = 64 * 64$, του μετασχηματισμένου συστήματος $N = 32 * 32$ και ενός ξεχωριστού συστήματος $N = 32 * 32$. Συμπεριλαμβάνεται ένα διάγραμμα για την μετασχηματισμένη θερμοκρασία T' ως συνάρτηση της T για το κρίσιμο σταθερό σημείο του μετασχηματισμού επανακανονικοποίησης. Εχει χρησιμοποιηθεί τεχνική re-weighting για τον προσδιορισμό των παρατηρήσιμων ποσοτήτων σε μεγάλο εύρος θερμοκρασιών. Κάποιες ευθείες δεν τέμνονται λόγω των φαινομένων πεπερασμένου μεγέθους. Χρησιμοποιώντας την μαγνήτιση η εκτίμηση της κρίσιμης θερμοκρασίας είναι $T_c = 2.26821 \Rightarrow \beta_c = 0.4409$. Οι κρίσιμοι εκθετες υπολογίζονται ίσοι με $\alpha = -0.19$, $\beta = 0.101$, $\gamma = 1.744$, $\nu = 1.01$.

4. Ενισχυτική Μάθηση στην Φυσική Πολλών Σωμάτων

4.1.0 Η Αρχή Μεταβολής

Η Variational Monte Carlo μέθοδος καθιστά δυνατή την περιγραφή του προβλήματος εύρεσης της θεμελιώδους κατάστασης ενός συστήματος ως ένα πρόβλημα βελτιστοποίησης. Είναι σημαντικό να γίνει μια αναφορά στην αρχή μεταβολής.

Υποθέτουμε ένα σύστημα που περιγράφεται απο μια χαμιλτονιανή H και για το οποίο δεν μπορούμε να επιλύσουμε την χρονοανεξάρτητη εξίσωση Schrödinger. Το σύστημα αποτελείται απο διακριτές ενέργειες και στόχος είναι να γίνει μια καλή προσέγγιση της ενέργειας της θεμελιώδους κατάστασης E_{gs}:

$$E_{gs} = E_1 \leq E_2 \leq \dots \tag{4.1}$$

Υποθέτουμε κανονικοποιημένες κυματοσυναρτήσεις μεταβολής ψ_{var} που δεν είναι αναγκαίο να είναι ιδιοκαταστάσεις ενέργειας και υποθέτοντας οτι οι ιδιοσυναρτήσεις της H αποτελούν μια πλήρη βάση μπορούμε να επεκτείνουμε σε όρους:

$$\psi_{var} = \sum_n c_n \psi_n, \; with \; H\psi_n = E_n \psi_n \tag{4.2}$$

$$\sum_n |c_n|^2 = 1 \tag{4.3}$$

Η αναμενόμενη τιμή της $\langle H \rangle_{var}$ δίνεται απο την σχέση:

$$\langle H \rangle_{var} = \sum_n E_n |c_n|^2 \tag{4.4}$$

Το σύστημα αποτελείται απο αντίστοιχες διακριτές ενέργειες E_n και η ενέργεια της θεμελιώδους κατάστασης ορίζει ένα κάτω φράγμα:

$$\langle H \rangle_{var} = \sum_n E_n |c_n|^2 \geq \sum_n E_1 |c_n|^2 = E_1 \sum_n |c_n|^2 = E_1 = E_{gs} \tag{4.5}$$

Για την περίπτωση όπου η κυματοσυνάρτηση μεταβολής ψ_{var} δεν ήταν κανονικοποιημένη, μπορούμε να την κανονικοποιήσουμε και να έχουμε μια πιο

γενικευμένη εκφραση:

$$E_{gs} \leq \langle H \rangle_{var} = \frac{\langle \psi_{var}|H|\psi_{var}\rangle}{\langle \psi_{var}|\psi_{var}\rangle} \equiv \mathcal{F}[\psi_{var}] \tag{4.6}$$

Υποθέτοντας μια κυματοσυνάρτηση μεταβολής η αναμενόμενη τιμή της ενέργειας $\langle H \rangle_{var}$ θα είναι μια υπερεκτίμηση της ενέργειας θεμελιώδους κατάστασης E_{gs}. Επιπλέον κάθε $\langle H \rangle_{var}$ θέτει ένα άνω φράγμα για την ακριβή τιμή της E_{gs}. ΟΙ κυματοσυναρτήσεις μεταβολής $\psi_{var}(x; p_1, p_2 \dots p_n)$ εξαρτώνται απο ένα πλήθος παραμέτρων μεταβολής $\{p\}$. Χρησιμοποιώντας μια προσέγγιση βελτιστοποίησης οι παράμετροι $\{p\}$ μπορούν να μεταβληθούν με στόχο την λήψη καλύτερων προσεγγίσεων για την ενέργεια θεμελιώδους κατάστασης. Σε μια ισοδύναμη έκφραση, η $\mathcal{F}[\psi_{var}]$ είναι μια μηχανή που εφαρμόζει ένα σύνολο διαδικασιών και παράγει ένα αριθμητικό αποτέλεσμα.

4.1.1 Η Variational Monte Carlo Μέθοδος και η Ιδιότητα της Μηδενικής Διασποράς

Είναι δυνατό να χρησιμοποιηθεί η Variational Monte Carlo μέθοδος για την έκφραση κβαντικών αναμενόμενων τιμών ως στατιστικές μέσες τιμές με στόχο την καλύτερη κατανόηση της συμπεριφοράς ενός συστήματος σε χαμηλές ενέργειες. Το κύριο πρόβλημα είναι η εισαγωγή ενός ansatz το οποίο ίσως εισάγει κάποια μεροληψία.

Ας υποθέσουμε ενα χώρο Hilbert που αποτελείται απο ένα πλήρες σύνολο βάσης $\{|x\rangle\}$ και ισχύει η συνθήκη πληρότητας:

$$\sum_x |x\rangle \langle x| = \mathcal{I} \tag{4.7}$$

Για μια κυματοσυνάρτηση μεταβολής ψ_{var} ένας τελεστής $\mathcal{O}$ έχει μια κβαντική αναμενόμενη τιμή:

$$\langle \mathcal{O} \rangle = \frac{\langle \psi_{var}|\mathcal{O}|\psi_{var}\rangle}{\langle \psi_{var}|\psi_{var}\rangle} = \frac{\sum_x \langle \psi_{var}|x\rangle \langle x|\mathcal{O}|\psi_{var}\rangle}{\sum_x \langle \psi_{var}|x\rangle \langle x|\psi_{var}\rangle} \tag{4.8}$$

Η παραπάνω σχέση μπορεί να εκφραστει σε μια μορφή δειγματοληψίας με κριτήριο σημαντικότητας:

$$\begin{aligned}
\langle \mathcal{O} \rangle &= \frac{\sum_x \langle \psi_{var}|x\rangle \langle x|\mathcal{O}|\psi_{var}\rangle}{\sum_x \langle \psi_{var}|x\rangle \langle x|\psi_{var}\rangle} \\
&= \frac{\sum_x \langle \psi_{var}|x\rangle \langle x|\psi_{var}\rangle \frac{\langle x|\mathcal{O}|\psi_{var}\rangle}{\langle x|\psi_{var}\rangle}}{\sum_x \langle \psi_{var}|x\rangle \langle x|\psi_{var}\rangle} \\
&= \frac{\sum_x |\psi_{var}(x)|^2 \frac{\langle x|\mathcal{O}|\psi_{var}\rangle}{\langle x|\psi_{var}\rangle}}{\sum_x |\psi_{var}(x)|^2} \\
&= \frac{\sum_x |\psi_{var}(x)|^2 \mathcal{O}_{loc}}{\sum_x |\psi_{var}(x)|^2}
\end{aligned} \tag{4.9}$$

Η ποσότητα $\mathcal{O}_{loc}$ ονομάζεται τοπικός τελεστής και ορίζεται ως:

$$\mathcal{O}_{loc} = \frac{\langle x|\mathcal{O}|\psi_{var}\rangle}{\langle x|\psi_{var}\rangle} \tag{4.10}$$

Ειναι εμφανές οτι η ποσότητα:

$$p(x) = \frac{|\psi_{var}(x)|^2}{\sum_x |\psi_{var}(x)|^2}, \tag{4.11}$$

ορίζει μια πιθανότητα αφου $\sum_x p(x) = 1$ και έχει μη μηδενική τιμή για όλες τις απεικονίσεις. Είναι δυνατό να χρησιμοποιηθεί μια προσέγγιση Monte Carlo μαρκοβιανών αλυσίδων για να δειγματοληφθεί ένα σύνολο απο πεπερασμένες καταστάσεις που έχουν κατανεμηθεί σύμφωνα με μια αντίστοιχη κατανομη ισορροπίας. Η παραπάνω προεγγιση έχει θεμελιωθεί για ένα γενικό τελεστή $\mathcal{O}$ και άρα κάθε παρατηρήσιμη ποσότητα μπορεί να υπολογιστεί στοχαστικά.

Για την περίπτωση της αναμενόμενης τιμής της χαμιλτονιανής $\mathcal{H}$, μπορεί κάποιος να ορίσει έναν τοπικό τελεστή E_{loc} που αποκαλείται τοπική ενέργεια:

$$E_{loc}(x) = \frac{\langle x|\mathcal{H}|\psi_{var}\rangle}{\langle x|\psi_{var}\rangle}. \tag{4.12}$$

Η κβαντική μέση τιμή της $\mathcal{H}^2$ δίνεται απο την σχέση:

$$\begin{aligned}\frac{\langle\psi_{var}|\mathcal{H}^2|\psi_{var}\rangle}{\langle\psi_{var}|\psi_{var}\rangle} &= \frac{\sum_x \langle\psi_{var}|\mathcal{H}|x\rangle\langle x|\mathcal{H}|\psi_{var}\rangle}{\sum_x \langle\psi_{var}|x\rangle\langle x|\psi_{var}\rangle} \\ &= \frac{\sum_x \langle\psi_{var}|x\rangle\langle x|\psi_{var}\rangle \frac{\langle\psi_{var}|\mathcal{H}|x\rangle\langle x|\mathcal{H}|\psi_{var}\rangle}{\langle\psi_{var}|x\rangle\langle x|\psi_{var}\rangle}}{\sum_x \langle\psi_{var}|x\rangle\langle x|\psi_{var}\rangle} \\ &= \frac{\sum_x |\psi_{var}(x)|^2 |E_{loc}(x)|^2}{\sum_x |\psi_{var}(x)|^2}\end{aligned} \tag{4.13}$$

Η διασπορά της τοπικής ενέργειας E_{loc} είναι ίση με την διασπορά της χαμιλτονιανής:

$$\sigma^2_{E_{loc}} = \frac{\langle\psi_{var}|(\mathcal{H}-E)^2|\psi_{var}\rangle}{\langle\psi_{var}|\psi_{var}\rangle} \tag{4.14}$$

Για την περίπτωση οπου η ψ_{var} ειναι μια ιδιοκατάσταση της $\mathcal{H}$ η τοπική ενέργεια E_{loc} είναι σταθερή:

$$E_{loc} = \frac{\langle x|\mathcal{H}|\psi_{var}\rangle}{\langle x|\psi_{var}\rangle} = E\frac{\langle x|\psi_{var}\rangle}{\langle x|\psi_{var}\rangle} = E. \tag{4.15}$$

Το παραπάνω αποτέλεσμα συνεπάγεται οτι η διασπορά είναι μηδενική. Αυτή η ιδιότητα εμφανίζεται μονο στην εκτίμηση των κβαντικών αναμενόμενων τιμών και δηλώνει οτι όσο πιο πολύ πλησιάζουμε σε μια ιδιοκατάσταση, τόσο πιο μικρές γίνονται οι διακυμάνσεις. Ενα κλασικό σύστημα δεν εμφανίζει την ίδια συμπεριφορά, λόγω των θερμικών διακυμάνσεων.

4.2.0 *Ενισχυτική Μάθηση για τον Προσδιορισμό της Θεμελιώδους Κατάστασης*

Για να θεμελιώσουμε μια προσέγγιση βελτιστοποίησης βασισμένη σε παραγώγους πρέπει να εκφραστεί και η παράγωγος της ενέργειας ως μια αναμενόμενη τιμή. Υποθέτοντας μια παράμετρο μεταβολής p_k της κυματοσυνάρτησης η παράγωγος της ενέργειας προς την p_k ισούται με:

$$
\begin{aligned}
\partial_{p_k} \langle \mathcal{H} \rangle &= \partial_{p_k} \frac{\langle \psi_{var} | \mathcal{H} | \psi_{var} \rangle}{\langle \psi_{var} | \psi_{var} \rangle} \\
&= \frac{\partial_{p_k} \big(\langle \psi_{var} | \mathcal{H} | \psi_{var} \rangle \big) \langle \psi_{var} | \psi_{var} \rangle - \langle \psi_{var} | \mathcal{H} | \psi_{var} \rangle \, \partial_{p_k} \big(\langle \psi_{var} | \psi_{var} \rangle \big)}{\big(\langle \psi_{var} | \psi_{var} \rangle \big)^2} \\
&= \frac{\partial_{p_k} \big(\langle \psi_{var} | \mathcal{H} | \psi_{var} \rangle \big)}{\langle \psi_{var} | \psi_{var} \rangle} - \frac{\langle \psi_{var} | \mathcal{H} | \psi_{var} \rangle}{\langle \psi_{var} | \psi_{var} \rangle} \frac{\partial_{p_k} \big(\langle \psi_{var} | \psi_{var} \rangle \big)}{\langle \psi_{var} | \psi_{var} \rangle} \\
&= \frac{\langle \partial_{p_k} \psi_{var} | \mathcal{H} | \psi_{var} \rangle}{\langle \psi_{var} | \psi_{var} \rangle} + \frac{\langle \psi_{var} | \mathcal{H} | \partial_{p_k} \psi_{var} \rangle}{\langle \psi_{var} | \psi_{var} \rangle} - \langle \mathcal{H} \rangle \frac{\langle \partial_{p_k} \psi_{var} | \psi_{var} \rangle + \langle \psi_{var} | \partial_{p_k} \psi_{var} \rangle}{\langle \psi_{var} | \psi_{var} \rangle} \\
&= \frac{\sum_x \langle \partial_{p_k} \psi_{var} | x \rangle \langle x | \mathcal{H} | \psi_{var} \rangle}{\sum_x \langle \psi_{var} | x \rangle \langle x | \psi_{var} \rangle} + \frac{\sum_x \langle \psi_{var} | \mathcal{H} | x \rangle \langle x | \partial_{p_k} \psi_{var} \rangle}{\sum_x \langle \psi_{var} | x \rangle \langle x | \psi_{var} \rangle} \\
&- \langle \mathcal{H} \rangle \frac{\sum_x \langle \partial_{p_k} \psi_{var} | x \rangle \langle x | \psi_{var} \rangle + \sum_x \langle \psi_{var} | x \rangle \langle x | \partial_{p_k} \psi_{var} \rangle}{\sum_x \langle \psi_{var} | x \rangle \langle x | \psi_{var} \rangle} \\
&\simeq \langle E_{loc} D_k^* \rangle - \langle E_{loc} \rangle \langle D_k^* \rangle + c.c.
\end{aligned}
\tag{4.16}
$$

Η ποσότητα $D_k(x)$ που εμφανίζεται στην παραπάνω εξίσωση ορίζεται ως:

$$
D_k(x) = \frac{1}{\langle x | \psi_{var} \rangle} \partial_{p_k} (\langle x | \psi_{var} \rangle) \tag{4.17}
$$

Η ιδέα ειναι να τεθεί η περιθώρια κατανομή ενος νευρωνικού δικτύου με ένα σύνολο απο παραμέτρους μεταβολής $\{p\}$ ίση με την κυματοσυνάρτηση:

$$
\psi(\boldsymbol{x}) = F(\boldsymbol{x}; p_1, p_2, \dots p_{N_p}) \tag{4.18}
$$

Ως νευρωνικό δίκτυο επιλέγεται το Restricted Boltzmann Machine σύμφωνα με την προηγούμενη υλοποίηση. Οι ορατοί και οι κρυφοί νευρώνες θα εκφραστούν ως σ και h αντίστοιχα και οι παράμετροι μεταβολής μπορούν να ειναι μιγαδικες . Οδηγούμαστε σε μια ανάλογη σχεση με την 1.35:

$$
\begin{aligned}
F_{rbm}(\sigma_1^z, \sigma_2^z, \dots \sigma_N^z) &= \sum_{\{h\}} e^{\sum_i \sum_j w_{ij} \sigma_i^z h_j + \sum_j h_j b_j + \sum_i \sigma_i^z a_i} \\
&= e^{\sum_i \sigma_i^z a_i} \sum_{\{h\}} e^{\sum_i \sum_j w_{ij} \sigma_i^z h_j + \sum_j h_j b_j} \\
&= e^{\sum_i \sigma_i^z a_i} \sum_h \prod_j e^{\sum_i w_{ij} \sigma_i^z h_j + h_j b_j} \\
&= e^{\sum_i \sigma_i^z a_i} \prod_j \left(1 + e^{\sum_i w_{ij} \sigma_i^z + b_j} \right)
\end{aligned}
\tag{4.19}
$$

Για την περίπτωση του μονοδιάστατου Ising μοντέλου διαμήκους πεδίου και χρησιμοποιώντας την γνώση οτι η κυματοσυνάρτηση της θεμελιώδους κατάστασης είναι θετικώς ορισμένη, μια επιλογή για την κβαντική κατάσταση του νευρωνικού δικτύου που οδηγεί σε πραγματικές τιμές μεταβολής είναι η:

$$\Psi(\sigma_1^z, \sigma_2^z, \dots \sigma_N^z) = \sqrt{F_{rbm}(\sigma_1^z, \sigma_2^z, \dots \sigma_N^z)} \tag{4.20}$$

Υποθέτοντας μια απεικόνιση σπιν σ του Ising μοντέλου διαμήκους πεδίου, η τοπική ενέργεια ισούται με:

$$E_{loc}(\sigma) = \frac{\langle \sigma|\mathcal{H}|\psi_{var}\rangle}{\langle \sigma|\psi_{var}\rangle} = \sum_{\sigma'} \langle \sigma|\mathcal{H}|\sigma'\rangle \, \frac{\langle \sigma'|\psi_{var}\rangle}{\langle \sigma|\psi_{var}\rangle} \tag{4.21}$$

Η παραπάνω άθροιση είναι πάνω σε $N+1$ απεικονίσεις οπου $\sigma'(0) = \sigma$ είναι η αρχική απεικόνιση και $\langle \sigma|\mathcal{H}|\sigma\rangle = -J\sum_i \sigma_i^z \sigma_{i+1}^z$. Οι υπολοιπες απεικονίσεις $\sigma'(k) = \sigma_1^z \dots \sigma_k^{z'} \dots \sigma_N^z$ προκύπτουν για την μεταβολη της τιμής του k σπιν με $\langle \sigma|\mathcal{H}|\sigma'\rangle = -h$. Παρατηρούμε οτι:

$$\begin{aligned}
\frac{\psi_{var}(\sigma'(k))}{\psi_{var}(\sigma)} &= \frac{\sqrt{e^{\sum_i \sigma_i^z a_i - a_k(1-2\sigma_k^z)} \prod_j \left(1 + e^{\sum_i w_{ij}\sigma_i^z + b_j - w_{jk}(1-2\sigma_k^z)}\right)}}{\sqrt{e^{\sum_i \sigma_i^z a_i} \prod_j \left(1 + e^{\sum_i w_{ij}\sigma_i^z + b_j}\right)}} \\
&= e^{-\frac{1}{2}a_k(1-2\sigma_k^z)} \sqrt{\frac{\prod_j 1 + e^{\sum_i w_{ij}\sigma_i^z + b_j - w_{jk}(1-2\sigma_k^z)}}{\prod_j 1 + e^{\sum_i w_{ij}\sigma_i^z + b_j}}}
\end{aligned} \tag{4.22}$$

$$\begin{aligned}
\ln\left(\frac{\psi_{var}(\sigma'(k))}{\psi_{var}(\sigma)}\right) =& -\frac{1}{2}a_k(1-2\sigma_k^z) \\
&+ \frac{1}{2}\ln\prod_j \left(1 + e^{\sum_i w_{ij}\sigma_i^z + b_j - w_{jk}(1-2\sigma_k^z)}\right) \\
&- \frac{1}{2}\ln\prod_j \left(1 + e^{\sum_i w_{ij}\sigma_i^z + b_j}\right) \\
=& -\frac{1}{2}a_k(1-2\sigma_k^z) \\
&+ \frac{1}{2}\sum_j \ln\left(1 + e^{\sum_i w_{ij}\sigma_i^z + b_j - w_{jk}(1-2\sigma_k^z)}\right) \\
&- \frac{1}{2}\sum_j \ln\left(1 + e^{\sum_i w_{ij}\sigma_i^z + b_j}\right)
\end{aligned} \tag{4.23}$$

Παρομοίως, οι εκφράσεις για τις παραγώγους προς τις παραμέτρους μετα-

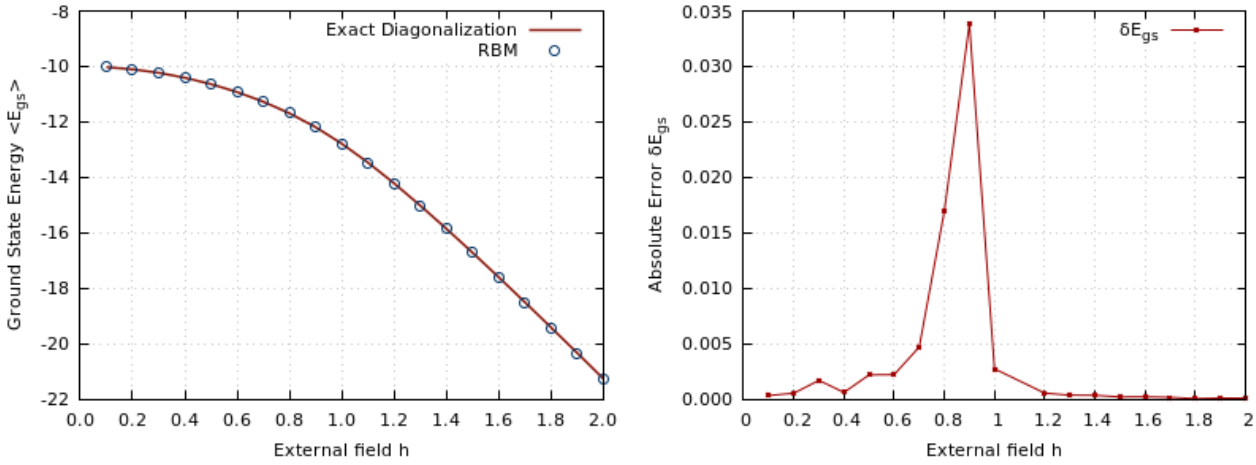
βολής είναι:

$$
\begin{aligned}
D_{a_i}(\sigma) &= \frac{\partial_{a_i}\left(\sqrt{e^{\sum_i \sigma_i^z a_i}\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)}\right)}{\sqrt{e^{\sum_i \sigma_i^z a_i}\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)}} \\
&= \frac{1}{2}\sigma_i^z
\end{aligned}
\tag{4.24}
$$

$$
\begin{aligned}
D_{b_j}(\sigma) &= \frac{\partial_{b_j}\left(\sqrt{e^{\sum_i \sigma_i^z a_i}\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)}\right)}{\sqrt{e^{\sum_i \sigma_i^z a_i}\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)}} \\
&= \frac{1}{2} sig\Big(\sum_i w_{ij}\sigma_i + b_j\Big)
\end{aligned}
\tag{4.25}
$$

$$
\begin{aligned}
D_{w_{ij}}(\sigma) &= \frac{\partial_{w_{ij}}\left(\sqrt{e^{\sum_i \sigma_i^z a_i}\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)}\right)}{\sqrt{e^{\sum_i \sigma_i^z a_i}\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)}} \\
&= \frac{1}{2}\sigma_i^z sig\Big(\sum_i w_{ij}\sigma_i + b_j\Big)
\end{aligned}
\tag{4.26}
$$

Είναι τωρα δυνατό να εκπαιδευθεί ένα νευρωνικό δίκτυο με στόχο την εκτίμηση της ενέργειας της θεμελιώδους κατάστασης. Ενα σύνολο απο απεικονίσεις παράγεται σε κάθε εποχή μεσω δειγματοληψιών Gibbs. Αυτές οι απεικονίσεις χρησιμοποιούνται για τον υπολογισμό των απαραίτητων ποσοτήτων και την μεταβολή των παραμέτρων μεταβολής σύμφωνα με την gradient descent. Οι τεχνικές λεπτομέρειες είναι ίδιες με την υλοποίηση των Restricted Boltzmann Machines στα προηγούμενα κεφάλαια. Η διαφορά έγκειται στο οτι κατα την ενισχυτική μάθηση δειγματοληπτούνται καταστάσεις χρησιμοποιώντας το νευρωνικό δίκτυο και ελαχιστοποιώντας την αναμενόμενη τιμή της ενέργειας. Στα προηγούμενα κεφάλαια στόχος ήταν ο μηδενισμός της Kullback-Leibler απόκλισης ανάμεσα σε δύο κατανομές. Οταν ο αριθμός των εποχών είναι αρκετά μεγάλος το νευρωνικό δίκτυο θα έχει βρεθεί σε ένα ελάχιστο και οι παράμετροι μεταβολής θα αντιστοιχούν σε μια αναπαράσταση της θεμελιώδους κατάστασης.



(ι) (ιι)

Φιγυρε 4.2: Αναμενόμενες τιμές της ενεργειας συναρτήσει του εξωτερικού πεδίου h για ένα μονοδιάστατο Ising μοντέλο διαμήκους πεδίου με $L = 10$ σπιν. Η κόκκινη γραμμή αντιστοιχεί σε υπολογισμούς της ενέργειας θεμελιώδους κατάστασης χρησιμοποιώντας ακριβή διαγωνοποίηση. Για κάθε τιμή του εξωτερικού πεδίου h ένα νευρωνικό δίκτυο εκπαιδεύεται με ρυθμο μάθησης $l = 0.2$, αριθμό κρυφών νευρώνων $n_h = 10$, batch size $b = 100$ και $e = 50000$ εποχές. Η σταθερά σύζευξης ισούται με $J = 1.0$. Επισης σχεδιάζονται τα απολυτα σφάλματα. Το μοντέλο έχει μια κβαντική μετάβαση φάσης για τιμη του εξωτερικού μαγνητικού πεδίου ίση με $h = 1.0$.

5. *Σύνοψη*

Συνοψίζοντας, τα νευρωνικά δίκτυα είναι ενα ενδιαφέρον ερευνητικό εργαλείο για την αντιμετωπιση προβλημάτων τα οποια έχουν προσεγγιστεί μεσω αρχών στατιστικής φυσικής. Η σύνδεση τους με την ομάδα επανακανονικοποίησης είναι επίσης μια ενδιάφερουσα ιδέα η οποία πρέπει να εξεταστεί περαιτέρω. Επιπλέον, υπάρχει η δυνατότητα να χρησιμοποιηθούν συνδυαστικά με τεχνικές Monte Carlo χρησιμοποιώντας μάθηση χωρίς επιβλεψη ή ανεξάρτητα μεσω της προσέγγισης ενισχυτικής μάθησης οπου δίνουν ανταγωνιστικά αποτελέσματα συγκρινόμενα με άλλες τεχνικές απο σχετικές δημοσιεύσεις. Ακολουθεί μια σύνοψη των κύριων θεμάτων που μελετήθηκαν στην διπλωματική.

Αρχικά τα Restricted Boltzmann Machines εκπαιδευθηκαν με στόχο την μοντελοποίηση μιας κατανομής που αποτελείται απο απεικονισεις του διδιάστατου Ising μοντέλου. Οι απεικονίσεις αυτές έχουν δειγματοληφθεί απο μια κατανομή ισορροπίας μεσω Monte Carlo μαρκοβιανών αλυσίδων χρησιμοποιώντας τον αλγόριθμο Metropolis η τον cluster αλγόριθμο του Wolff. Οταν το νευρωνικό δίκτυο έχει εκπαιδευθεί είναι δυνατό να παράγει προσεγγιστικές απεικονίσεις με στοχο την χρήση τους για τον υπολογισμό παρατηρήσιμων ποσοτήτων οπως η εσωτερική ενέργεια, η μαγνήτιση, η ειδική θερμότητα και η μαγνητική επιδεκτικότητα που θα συγκριθούν με αυτες των Monte Carlo μετρήσεων. Τα Restricted Boltzmann Machines μπορούν να αναπαράγουν αποτελέσματα τα οποια ειναι ακριβή οταν ο αριθμός των κρυφών νευρώνων ισούται με τον αριθμό των ορατών για τις περιπτώσεις των φάσεων αταξιας και τάξης αλλα και στην κρίσιμη περίοχη. Επιπλέον, για περιπτωσεις μικρότερου αριθμού κρυφών νευρώνων το σύστημα οδηγείται σε συμπίεση με στόχο τον έλεγχο της ακρίβειας των μετρήσεων και της συσχετισης του αριθμού των κρυφών νευρώνων με τον λόγο συμπίεσης. Τα αποτέλεσματα είναι ακριβή για θερμοκρασίες μακριά απο την μετάβαση φάσης δευτέρας τάξεως αλλα εμφανίζουν ασυμφωνία για θερμοκρασίες κοντα στο κρίσιμο σημείο. Η παραπάνω προσέγγιση υποδεικνύει την χρήση των νευρωνικών δικτύων ως εργαλεία έρευνας εφαρμοσμένων θεματων στατιστικής φυσικής αφού είναι δυνατό να περιγράψουν την πλήρη θερμοδυναμική συμπεριφορά του μοντέλου και να παράγουν καταστάσεις απο τις οποιες είναι δυνατό να υπολογιστούν παρατηρήσιμες ποσότητες σε όλο το ευρος θερμοκρασιών. Επιπλέον τα νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν ως ένα εργαλείο σύνδεσης της πειραματικής με την θεωρητική φυσική. Μια ημιτελής πειραματική παρατήρηση ενος μοντέλου που έχει

επίσης προσομοιωθεί, μπορει να ολοκληρωθεί αν τοποποθετηθεί σε ένα σωστα εκπαιδευμενο νευρωνικό δίκτυο και δειγματοληφθούν οι περιθώριες κατανομές. Επιπλέον σε ένα ήδη εκπαιδευμένο νευρωνικό δίκτυο μπορεί να γίνει εισαγωγή μιας πειραματικης απεικόνισης και να παρατηρηθεί η χρονική της εξέλιξη μέσω διαδοχικών δειγματοληψιών Gibbs. Η εξοικείωση με την παραπάνω προσέγγιση είναι σημαντική διοτι επιτρέπει μια καλύτερη κατανόηση της αντιστοιχίας ανάμεσα στην ομάδα επανακανονικοποίησης και τα βαθια νευρωνικά δίκτυα αφού αυτή η αντιστοιχία θεμελιώνεται για την περίπτωση της μαθησης χωρίς επίβλεψη.

Ακολούθως, εισάγεται η Ομάδα Επανακανονικοποίησης πραγματικού χώρου με στόχο την εκτίμηση της κρίσιμης θερμοκρασίας και των κρίσιμων εκθετών του διδιαστατου Ising μοντέλου. Ενας spin blocking μετασχηματισμός επανακανονικοποίησης υλοποιείται για να παρατηρηθει η ροή στον χώρο παραμέτρων ο οποίος είναι μονοδιαστατος αφου η θερμοκρασία είναι η μονη παράμετρος που οδηγεί το σύστημα σε μετάβαση φάσεως. Η κρίσιμη θερμοκρασία είναι ένα κρίσιμο σταθερό σημείο και προσομοιώσεις για αντίστροφες θερμοκρασίες $\beta > \beta_c$ και $\beta < \beta_c$ οδηγούν το σύστημα προς την πλήρη τάξη και αταξία αντίστοιχα. Ενα συνολο απο απεικονίσεις που έχουν δειγματοληφθεί απο μια κατανομή ισορροπίας προφανώς εμφανίζεται με τις σωστές πιθανότητες Boltzmann. Για την εξαγωγή μετρήσεων είναι σημαντικό να γίνει αναφορά στην υπόθεση οτι οι μετασχηματισμένες καταστάσεις εμφανίζονται με τις σωστές πιθανότητες Boltzmann τους. Αυτή η υπόθεση είναι ακριβώς και ο λόγος για τον οποίο δεν γίνεται να εκτιμηθούν τα σφάλματα στην μέθοδο με χρήση μιας τεχνικής ανάλυσης σφαλμάτων όπως η binning, jackknife και η bootstrap. Οι κρίσιμοι εκθέτες υπολογίστηκαν χρησιμοποιώντας δεδομένα τα οποια έχουν γίνει re-weighted για ένα μεγάλο εύρος θερμοκρασιών απο ένα αρχικό σύστημα μεγέθους $N = 64 * 64$, ενα μετασχηματισμένο σύστημα μεγέθους $N' = 32 * 32$ και ένα επιπλέον σύστημα το οποιο έχει προσομοιωθεί ξεχωριστά με $N = 32 * 32$ έτσι ώστε να αντιμετωπιστούν τα φαινόμενα πεπερασμένου μεγέθους. Για ένα πλήθος των παρατηρήσιμων ποσοτήτων οπως η μαγνήτιση το επιπλέον σύστημα πρέπει να προσομοιωθεί για να γίνει δυνατή η εκτίμηση της κρίσιμης θερμοκρασίας. Τα φαινόμενα πεπερασμένου μεγέθους επηρεάζουν τα αποτελέσματα και τα δεδομένα του αρχικού και του μετασχηματισμένου συστήματος μπορεί να μην τέμνονται στην κρίσιμη θερμοκρασία. Αυτο το πρόβλημα αντιμετωπίζεται με την προσομοίωση ενος επιπλεον συστήματος που έχει το ίδιο μέγεθος με το μετασχηματισμένο. Η μέθοδος δίνει ακριβέστερα αποτελέσματα για μικρότερα πλέγματα σε σύγκριση με την βάθμιση πεπερασμένου μεγέθους η οποια πρέπει να υλοποιηθεί για μεγάλου μεγέθους πλέγματα.

Μια αντιστοιχία θεμελιώθηκε ανάμεσα στην ομάδα επανακανονικοποίησης πραγματικού χώρου και στα νευρωνικά δίκτυα που βασίζονται στην ενεργεια για ακριβείς μετασχηματισμούς επανακανονικοποίησης. Υποθέσαμε ενα γενικευμένο πλέγμα με αλληλεπιδράσεις πολλαπλών τάξεων ανάμεσα στα σπιν. Μια κατάλληλη επιλογή του τελεστή μεταβολής οδηγεί σε μια ισότητα ανάμε-

σα στην Χαμιλτονιανη που περιγράφει τους μετασχηματισμένους βαθμούς ελευθερίας που προκύπτουν απο την εφαρμογή του μετασχηματισμόυ επανακανονικοποίησης και στην Χαμιλτονιανή που περιγράφει τους κρυφούς νευρώνες του Restricted Boltzmann Machine. Το γεγονός οτι και το νευρωνικό δίκτυο είναι ένα μοντέλο στατιστικής φυσικής και περιγραφεται απο μια κατανομή Boltzmann επιτρέπει την εξαγωγή συμπερασμάτων και για τις περιθώριες κατανομές. Οταν ο μετασχηματισμός της ομάδος επανακανονικοποίησης είναι ακριβής η Χαμιλτονιανή του αρχικού συστήματος είναι ίση με την Χαμιλτονιανή των ορατών νευρώνων του Restricted Boltzmann Machine. Αυτό συνεπάγεται οτι η Kullback-Leibler απόκλιση είναι μηδενική και οτι η περιθώρια κατανομή των ορατών νευρώνων του Restricted Boltzmann Machine μπορεί να αναπαράγει πλήρως την εμπειρική κατανομή, δηλαδή τα δεδομένα εκπαίδευσης. Ενα Δικτυο Βαθιάς Πεποιθήσεως υλοποιείται για θερμοκρασία κοντά στην μετάβαση φάσεως δευτέρας τάξεως για το διδιάστατο μοντέλο Ising με στόχο επιπλέον παρατηρήσεις. Τα receptive fields σχεδιάστηκαν για κάθε κρυφό επίπεδο και παρατηρήθηκε οτι οι κρυφοί νευρώνες σχηματίζουν συστοιχίες που συζευγονται με τους ορατούς νευρώνες. Αυτές οι συστοιχίες είναι περίπου ίδιου μεγέθους για ένα δεδομένο κρυφό επίπεδο και το μέγεθος τους αυξάνεται με την αύξηση του λογου συμπίεσης. Η παραπάνω ιδέα είναι ακριβώς ίδια με εναν spin blocking μετασχηματισμό της ομάδας επανακανονικοποίησης. Μια σημαντική διαφορά ειναι οτι το νευρωνικό δίκτυο οργανώνεται αυτόματα για να εφαρμόσει αυτόν τον μετασχηματισμό. Η παραπάνω θεμελίωση έχει γίνει για ακριβείς μετασχηματισμούς επανακανονικοποίησης. Η χρηση προσεγγιστικών τεχνικών για την ελαχιστοποίηση της Kullback-Leibler συνεπάγεται οτι το νευρωνικό δίκτυο εφαρμόζει πρακτικά διαφορετικούς μετασχηματισμούς στα δεδομένα.

Τέλος, ακολουθείται μια προσέγγιση ενισχυτικής μάθησης με στόχο την εκτίμηση της ενέργειας θεμελιώδους κατάστασης του μονοδιάστατου μοντέλου Ising διαμήκους πεδίου. Χρησιμοποιούνται Restricted Boltzmann Machines με ένα σύνολο απο μιγαδικές παραμέτρους μεταβολής με στόχο να περιγράψουν κβαντικές καταστάσεις. Η εκτίμηση της ενέργειας θεμελιώδους κατάστασης μπορει να εκφραστεί ως ένα πρόβλημα βελτιστοποίησης και η Variational Monte Carlo μέθοδος χρησιμοποιείται για να δειγματοληφθούν καταστάσεις απο το νευρωνικό δίκτυο. Απο το πλήθος αυτών των καταστάσεων υπολογίζεται η ενέργεια και η παράγωγος της ενέργειας και εφαρμοζεται η gradient descent. Εκτιμούνται οι ενέργειες θεμελιώδους κατάστασης για ενα πλήθος τιμών εφαρμοζόμενου εξωτερικού μαγνητικού πεδίου κάνοντας προσαρμογή δεδομένων στις τιμές που προκύπτουν οταν το νευρωνικό δίκτυο έχει οδηγηθεί στο ολικό ελάχιστο. Υπολογίζεται τότε το απολυτο σφάλμα με στόχο την σύγκριση των τιμών με αυτές που προκύπτουν απο ακριβή διαγωνοποίηση. Τελικά η ελαχιστοποίηση της ενέργειας δημιουργεί μια αναπαράσταση της θεμελιώδους κατάστασης η οποια έχει κωδικοποιηθεί στο σύνολο των παραμέτρων μεταβολής. Είναι τότε δυνατό να δειγματοληφθούν καταστάσεις αρχικοποιώντας τυχαια τους ορατούς νευρώνες του νευρωνικό δίκτυο και οδηγώντας το σε ι-

σορροπία μέσω δειγματοληψίας Gibbs. Απο αυτο το πλήθος των καταστάσεων είναι δυνατό να υπολογιστουν παρατηρήσιμες ποσότητες χρησιμοποιώντας ως κυματοσυνάρτηση την περιθώρια κατανομή των ορατών νευρώνων. Η μέθοδος έχει υψηλή ακρίβεια συγκρινόμενη με άλλες μεθόδους απο την σχετική βιβλιογραφία και δεν εμφανίζει το πρόβλημα προσήμου, αρα ενδείκνυται να εφαρμοστεί σε αλυτα προβλήματα οπως η εύρεση των ιδιοτήτων της θεμελιώδους κατάστασης ισχυρά αλληλεπιδρωντων φερμιονίων.

Bibliography

[1] Asja Fischer and Christian Igel. An Introduction to Restricted Boltzmann Machines. *Alvarez L., Mejail M., Gomez L., Jacobo J. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2012. Lecture Notes in Computer Science, vol 7441.*, 2012. Springer, Berlin, Heidelberg.

[2] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science 9*, 1985.

[3] Geoffrey E. Hinton. Boltzmann machine. *Scholarpedia, 2(5):1668*, 2007.

[4] Steffen L. Lauritzen. *Graphical models.* Oxford University Press, 1996.

[5] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[6] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation 14, 1771-1800*, 2002.

[7] Max Welling. Product of Experts. *Scholarpedia, 2(10):3879*, 2007.

[8] Guido Montufar and Nihat Ay. Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. *Neural Computation*, 2011.

[9] Yoshua Bengio. *Learning deep architectures for AI.* Foundations and Trends in Machine Learning 21(6), 1601-1621, 2009.

[10] Yoshua Bengion, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing (NIPS 19), pp. 153-160*, 2007. MIT Press.

[11] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation 18(7), 1527-1554*, 2006.

[12] Yoshua Bengio and Olivier Delalleau. Justifying and generalizing contrastive divergence. *Neural Computation 21(6), 1601-1621*, 2009.

[13] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. Learning multiple layers of representation. *Trends in Cognitive Sciences 11(10), 428-434*, 2007.

[14] Asja Fischer and Christian Igel. Bounding the bias of contrastive divergence learning. *Neural Computation 23, 664-673*, 2011.

[15] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neu- ral networks. *Science 313(5786), 504-507*, 2006.

[16] Konstantinos N. Anagnostopoulos. *Computational Physics: A Practical Introduction to Computational Physics and Scientific Computing (Using C++)*. Konstantinos N. Anagnostopoulos and the National Technical University of Athens, 2016.

[17] Bernd A. Berg. *Markov Chain Monte Carlo Simulations and their Statistical Analysis: With Web-Based Fortran Code*. World Scientific Publishing Co. Pte. Ltd, 2004.

[18] M.E.J Newman and G.T Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, 1999.

[19] E. Ising. Beitrag zur Theorie des Ferromagnetizmus. *Z. Phys. 31*, 1925.

[20] Lars Onsager. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physical Review, Series II*, 1944.

[21] Alan M. Ferrenberg and Robert H. Swendsen. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett. 61*, 1988.

[22] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, and Augusta H. Teller. Equation of State Calculations by Fast Computing Machines . *The Journal of Chemical Physics 21, 1087*, 1953.

[23] Wilfred K. Hastings. Monte Carlo sampling methods using Markov chains and their applications . *Biometrika 57: 97-109*, 1970.

[24] Robert H. Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett. 58, 86*, 1987.

[25] C. M. Fortuin and P. W. Kasteleyn. On the random-cluster model: I. Introduction and relation to other models. *Physica, Volume 57, Issue 4*, 1972.

[26] Dietrich Stauffer and Ammon Aharony. *Introduction To Percolation Theory*. CRC Press, 1994.

[27] Henk W. J. Blöte and Youjin Deng. Cluster Monte Carlo simulation of the transverse Ising model. *Phys. Rev. E 66, 066110*, 2002.

[28] Ulli Wolff. Collective Monte Carlo Updating for Spin Systems. *Phys. Rev. Lett. 62, 361*, 1989.

[29] James Propp and David Wilson. Coupling from the Past: a User's Guide, 1997.

[30] H. Flyvbjerg and H. G. Petersen. Error estimates on averages of correlated data . *The Journal of Chemical Physics 91, 461*, 1989.

[31] Giacomo Torlai and Roger G. Melko. Learning thermodynamics with Boltzmann machines. *Phys. Rev. B 94, 165134*, 2016.

[32] Kenneth G.Wilson and J.Kogut. The renormalization group and the ε expansion. *Physics Reports, Volume 12, Issue 2*, 1974.

[33] Kenneth G.Wilson. The renormalization group and critical phenomena. *Rev. Mod. Phys. 55, 583*, 1983.

[34] John Cardy. *Scaling and Renormalization in Statistical Physics*. Cambridge University Press, 1996.

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

[36] Nigel Goldenfeld. *Lectures On Phase Transitions And The Renormalization Group (Frontiers in Physics)* . Addison-Wesley, 1992.

[37] Leo P. Kadanoff, Anthony Houghton, and Mehmet C. Yalabik. Variational approximations for renormalization group transformations. *J. Stat. Phys. 14: 171*, 1976.

[38] Efi Efrati, Zhe Wang, Amy Kolan, and Leo P. Kadanoff. Real-space renormalization in statistical mechanics. *Rev. Mod. Phys. 86, 647*, 2014.

[39] Pankaj Mehta and David J. Schwab. An exact mapping between the Variational Renormalization Group and Deep Learning. *arXiv:1410.3831*, 2014.

[40] David J. Griffiths. *Introduction to Quantum Mechanics*. Prentice Hall, 1995.

[41] P.G.de Gennes. Collective motions of hydrogen bonds. *Solid State Communications, Volume 1, Issue 6*, 1963.

[42] Sei Suzuki, Jun ichi Inoue, and Bikas K. Chakrabarti. *Quantum Ising Phases and Transitions in Transverse Ising Models*. Cambridge University Press, 2013.

[43] W. L. McMillan. Ground State of Liquid He^4. *Phys. Rev. 138, A442*, 1965.

[44] Federico Becca and Sandro Sorella. *Quantum Monte Carlo Approaches for Correlated Systems*. Cambridge University Press, 2017.

[45] Giuseppe Carleo. Machine learning methods for many body physics, 2017. Lectures for the Advanced School on Quantum Science and Quantum technology, ICTP, Trieste, Italy.

[46] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science Vol. 355, Issue 6325, pp. 602-606*, 2017.

[47] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nature Physics*, 2018.

Εκτενής Περίληψη στην Αγγλική Γλώσσα

Ακολουθεί εκτενής περίληψη της διπλωματικής στην Αγγλική Γλώσσα:

Reinforcement Learning in Quantum Many-Body Physics and a Correspondence Between the Renormalization Group and Deep Neural Networks

by

Dimitrios S. Bachtis

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in

Microsystems and Nanodevices

Department of Physics
School of Applied Mathematics and Physical Sciences
National Technical University of Athens

Reinforcement Learning in Quantum Many-Body Physics and a Correspondence Between the Renormalization Group and Deep Neural Networks

by

Dimitrios S. Bachtis

Signed by the examination committee

Name	**Signature**	**Date**
(**Advisor**)		
(**Examiner**)		
(**Examiner**)		

COLOPHON

This thesis was typeset using tufte-latex.

First printing, February 2018

Abstract

This thesis was mainly driven by the need to identify what Machine Learning and Statistical Physics can offer to each other.

Initially, Restricted Boltzmann Machines were utilized in an unsupervised setting to model the probability distribution represented by importance-sampled configurations of the $d = 2$ Ising model. Configurations were then drawn from the equilibrium distribution of the neural network in order to calculate expectation values of observables near the second order phase transition. The expectation values compare well with Monte Carlo calculations and their accuracy shows a dependence on the number of hidden units.

A correspondence was then established between the Renormalization Group and Deep Belief Networks . To gain further insights to their connection, the receptive fields of a deep neural network trained near the phase transition were visualized. The critical temperature and the critical exponents of the $d = 2$ Ising model were then estimated for a system of size $N = 64*64$ using a spin-blocking renormalization group transformation .

Finally, a reinforcement learning approach was implemented based on the variational Monte Carlo method and the introduction of quantum states in Restricted Boltzmann Machines. The neural network was then used to estimate the ground state energy of the $d = 1$ transverse-field Ising model . The results prove to be competitive when compared with other techniques from relevant publications.

Acknowledgements

Completing my studies I would like to express my gratitude to a variety of people.

To Associate Professor Konstantinos N. Anagnostopoulos for his guidance and encouragement while supervising my undergraduate and postgraduate theses. He has also inspired me by releasing his book[1] under a CC-BY-SA license and the software therein under a GNU public license. Above all else I would like to thank him for saying "Yes" when I approached him with my ideas for a master thesis and for allowing me to take initiative on it.

To Professor Theo Alexopoulos and to Assistant Professor Konstantinos Kousouris for their participation in the examination committee.

To my family for their support and to my friends for all the happy moments we have experienced together.

[1] Konstantinos N. Anagnostopoulos. *Computational Physics: A Practical Introduction to Computational Physics and Scientific Computing (Using C++)*. Konstantinos N. Anagnostopoulos and the National Technical University of Athens, 2016

Contents

1. Restricted Boltzmann Machines

1.1.0 Boltzmann Machines

Boltzmann Machines[1] are a class of stochastic neural networks that share strong ties with statistical physics. They correspond to a model that consists of stochastic units and is representing a probability distribution. Boltzmann Machines are described by a set of variational parameters and can be used to extract important features of a *target* probability distribution. A set of data is then used to train the model by properly adjusting it's variational parameters at each iteration. Eventually, one acquires a closed-form approximate representation of the target probability distribution.

The *Restricted Boltzmann Machine* is a special case of Boltzmann Machines. It consists of *visible* and *hidden* units represented visually in two layers and connected through undirected edges. The term "Restricted" implies that no intralayer connections are allowed. Visible units are used as input of the training data when the model *learns* a target distribution. They are also used as output for acquiring samples from the neural network's equilibrium distribution. The hidden units are non-linear detector of features capturing dependencies on training data when they are used as input to the visible layer.

Restricted Boltzmann machines are also able to provide a solution for incomplete observations. One can fix the values of partially completed observations in the visible layer and sample the corresponding marginal distribution eventually acquiring the remaining visible units. They can also serve as classifiers in a supervised setting. It is of high importance that one can stack Restricted Boltzmann Machines to form a deep neural network, called Deep Belief Network, that will be introduced later.

A study follows for Restricted Boltzmann Machines [2] based on probabilistic graphical models and more specifically Markov Random Fields. This rigorous mathematical approach allows the exploitation of theorems and algorithms for a well-established description of RBMs.

[1] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science 9*, 1985; and Geoffrey E. Hinton. Boltzmann machine. *Scholarpedia, 2(5):1668*, 2007

[2] Asja Fischer and Christian Igel. An Introduction to Restricted Boltzmann Machines. *Alvarez L., Mejail M., Gomez L., Jacobo J. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2012. Lecture Notes in Computer Science, vol 7441.*, 2012. Springer, Berlin, Heidelberg

1.2.0 Graphical Models and Markov Random Fields

The framework of probabilistic graphical models simplifies the study of random variables sharing conditional dependence and independence properties by

representing them in graph structure.

Conditional independence between two sets of random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as independence in terms of an additional set $\boldsymbol{Z}$ for all values of $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$. This is equal to saying that:

$$p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z}) \Rightarrow p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z}) \; and \; p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{y}|\boldsymbol{z}) \tag{1.1}$$

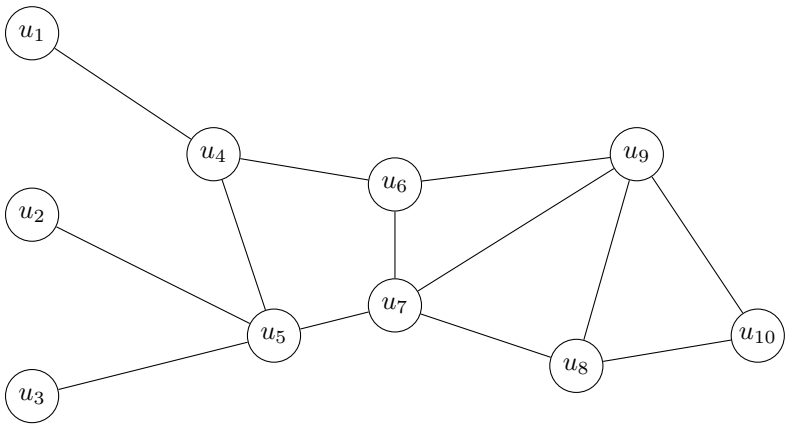


Figure 1.1: An undirected graph $G = (V, E)$. The neighborhood of node u_7 is $\{u_5, u_6, u_8, u_9\}$. The subsets $\{u_9, u_{10}\}$ and $\{u_8, u_9, u_{10}\}$ define cliques with $\{u_8, u_9, u_{10}\}$ being maximal. The nodes u_1 and u_{10} are separated by $\{u_6, u_7\}$.

An *undirected graph* $G = (V, E)$ is a data structure that consists of a set of nodes V and a set of undirected edges E. The undirected edges connect pair of nodes and are depicted as $X_i - X_j$. One can define a *neighborhood* $\mathcal{N}_u$ comprised of nodes that share a connection to a given node u:

$$\mathcal{N}_u = \{w \in V : \{w, u\} \in E\} \tag{1.2}$$

A *clique* in an undirected graph $G = (V, E)$ is a subset of V for which all included nodes are connected by pairs. A given clique is defined as *maximal* if while satisfying the above definition it is impossible to be extended by the addition of a node. A *path* between two nodes u_1 and u_m is defined as a finite sequence of nodes $\{u_1, u_2, \ldots, u_m \in V\}, \{u_i, u_{i+1}\} \in E, i = 1, \ldots, m-1$. If we assume a set $\mathcal{V} \subset V$, two nodes $u \notin \mathcal{V}$ and $w \notin \mathcal{V}$ are *separated* if all paths from u to w include a node from $\mathcal{V}$.

Now, let us assume an undirected graph $G = (V, E)$, for which to every node corresponds a random variable X_u that takes values in a state space $\Lambda_u = \Lambda$. If the joint probability distribution satisfies the *local Markov property*, the set of the random variables $\boldsymbol{X} = (X_u)_{u \in V}$ define a *Markov Random Field*. The local Markov property is satisfied if a random variable that corresponds to a given node has a conditional independence property in terms of all other variables given the corresponding neighborhood. Equally:

$$p(x_u|(x_w)_{w \in V \setminus \{u\}}) = p(x_u|(x_w)_{w \in \mathcal{N}_u}), \forall u \in V \; and \; \forall \boldsymbol{x} \in \Lambda^{|V|} \tag{1.3}$$

Considering a strictly positive probability distribution, it is possible to define other Markov properties that are equivalent to the local Markov property .

The *global Markov property* is satisfied if for three disjoint subsets $\mathcal{A}, \mathcal{B}, \mathcal{S} \subset V$, with $\mathcal{S}$ separating the nodes in $\mathcal{A}, \mathcal{B}$, the random variables $(X_a)_{a\in\mathcal{A}}$ and $(X_b)_{b\in\mathcal{B}}$ are conditionally independent in terms of $(X_s)_{s\in\mathcal{S}}$. Equally:

$$p((x_a)_{a\in\mathcal{A}}|(x_t)_{t\in S\cup\mathcal{B}}) = p((x_a)_{a\in\mathcal{A}}|(x_t)_{t\in S}) \tag{1.4}$$

The *pairwise Markov Property* is satisfied if two nodes that are not adjacent are conditionally independent in terms of all other variables. Equally if $\{u, w\} \notin E$ and $\forall \boldsymbol{x} \in \Lambda^{|V|}$:

$$p(x_u, x_w|(x_t)_{t\in V\setminus\{u,w\}}) = p(x_u|(x_t)_{t\in V\setminus\{u,w\}})p(x_w|(x_t)_{t\in V\setminus\{u,w\}}) \tag{1.5}$$

One can now naturally ask, based on the close connection of factorization of joint probability probability distributions and conditional independence for a set of random variables. if it is possible to factorize Markov Random Field distributions. The theorem by Hammersley-Clifford states that a strictly positive distribution p satisfies the Markov property with respect to an undirected graph $G = (V, E)$ if and only if p factorizes according to G. [3]

[3] Steffen L. Lauritzen. *Graphical models.* Oxford University Press, 1996; and Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009

To factorize a probability distribution over an undirected graph with $\mathcal{C}$ maximal cliques, a set of non-negative functions $\{\psi_C\}_{C\subset\mathcal{C}}$ must satisfy:

$$\forall \boldsymbol{x}, \hat{\boldsymbol{x}} \in \Lambda^{|V|} : (x_c)_{c\in C} = (\hat{x}_c)_{c\in C} \Rightarrow \psi_C(\boldsymbol{x}) = \psi_C(\hat{bmx}) \tag{1.6}$$

$$p(\boldsymbol{x}) = \frac{1}{Z}\prod_{C\in\mathcal{C}} \psi_C(\boldsymbol{x}) \tag{1.7}$$

We define the normalization constant Z as the partition function:

$$Z = \sum_x \prod_{C\in\mathcal{C}} \psi_C(\boldsymbol{x}_C) \tag{1.8}$$

For a strictly positive distribution p, the functions $\{\psi_C\}_{C\subset\mathcal{C}}$ are also positive and:

$$p(\boldsymbol{x}) = \frac{1}{Z}\prod_{C\in\mathcal{C}} \psi_C(\boldsymbol{x}_C) = \frac{1}{Z}e^{\sum_{C\in\mathcal{C}} \ln\psi_C(\boldsymbol{x}_C)} = \frac{1}{Z}e^{-E(\boldsymbol{x})} \tag{1.9}$$

The function E is called the *energy function*:

$$E = \sum_{C\in\mathcal{C}} \ln\psi_C(\boldsymbol{x}_C) \tag{1.10}$$

The strictly positive probability distribution p of any Markov Random Field is then the *Boltzmann (Gibbs) distribution*.

1.2.1 *Unsupervised Markov Random Field Learning*

The goal of *Unsupervised Learning* is the modeling of an unknown target distribution represented by a set of unlabeled training data. Now, let us assume a given graphical model with an energy function that depends on a set of variational parameters θ. Unsupervised Learning corresponds to the adjustment of these parameters in order to reach a representation of the target distribution q. This is equal to saying that we want the model distribution p to match perfectly the unknown target distribution q. The notation $p(\boldsymbol{x}|\boldsymbol{\theta})$ will be used to denote the dependency of the corresponding distribution to the variational parameters $\boldsymbol{\theta}$.

One can estimate the parameters of a statistical model with *maximum-likelihood estimation*. For a given set of independently drawn data $S = \{x_i\}$ from an unknown distribution q the parameters $\boldsymbol{\theta}$ are adjusted in order to maximize the probability of S under the Markov Random Field distribution.

The *likelihood* $\mathcal{L} : \Theta \to \mathcal{R}$ provides a mapping for the variational parameters $\boldsymbol{\theta}$ from a parameter space Θ to:

$$\mathcal{L}(\boldsymbol{\theta}|S) = \prod_{i=1}^{l} p(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{1.11}$$

The idea is to find the parameters $\boldsymbol{\theta}$ that will maximize the likelihood for a given training set. Equally, one can maximize the *log-likelihood*:

$$\ln \mathcal{L}(\boldsymbol{\theta}|S) = \ln \prod_{i=1}^{l} p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{i}^{l} \ln p(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{1.12}$$

Assuming the Markov Random Field distribution is the Boltzmann distribution deriving an analytical solution for problems of interest could be impossible, thus an approximate technique like *gradient ascent* has to be implemented.

As indicated before, unsupervised learning corresponds to an adjustment of the variational parameters θ in order to match the model distribution p to the unknown distribution q represented by a finite set of training data S. Equally, we want to minimize the distance between the two distributions. This is achievable through a minimization of the *Kullback-Leibler Divergence*, i.e the relative entropy, which for a finite space Ω is:

$$KL(q||p) = \sum_{\boldsymbol{x}\in\Omega} q(\boldsymbol{x}) \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} = \sum_{\boldsymbol{x}\in\Omega} q(\boldsymbol{x}) \ln q(\boldsymbol{x}) - \sum_{\boldsymbol{x}\in\Omega} q(\boldsymbol{x}) \ln p(\boldsymbol{x}) \tag{1.13}$$

It is straightforward to notice that since $q(x)$ can be estimated from the training data set S, Kullback-Leibler divergence can be expressed in terms of the log-likelihood through $-\sum_{x\in\Omega} q(x) \ln p(x)$ with a dependence on the variational parameters θ. The Kullback-Leibler divergence is a positive quantity and is zero for the case of the two distributions exactly matching one another. Therefore, maximizing the log-likelihood is equal to minimizing the Kullback-Leibler divergence.

To acquire the variational parameters that maximize the log-likelihood one can use a first-order optimization algorithm called gradient ascent. Gradient ascent updates iteratively the parameters $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$ according to the update rule:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \frac{\partial}{\partial \boldsymbol{\theta}^{(t)}} \left(\ln \mathcal{L}(\boldsymbol{\theta}^{(t)}|S) \right) - \lambda \boldsymbol{\theta}^{(t)} + \nu \Delta \boldsymbol{\theta}^{(t-1)} \tag{1.14}$$

The quantity $\eta \in \mathbb{R}_0^+$ is called the *learning rate*. The optional quantity λ corresponds to *weight decay* which is a penalty term that forces weights to acquire smaller values. Additionally, the optional quantity $\nu \Delta \theta^{(t-1)}$ is the *momentum* term and forces faster updates towards the direction of the gradient based on the previous update.

Now, let us assume the modeling of an m-dimensional target distribution with a Markov Random Field that is comprised with an amount of nodes greater than m. It is possible to separate the variables $\boldsymbol{X} = (X_u)_{u \in V}$ into *visible* $\boldsymbol{V} = (V_1, V_2, \ldots, V_m)$ and latent, also called *hidden*, $\boldsymbol{H} = (H_1, H_2, \ldots, H_n)$, where $n = |V| - m$.

The addition of hidden variables allows a better description of the unknown target probability distribution since correlations between visible units can be expressed through conditional distributions. The Boltzmann probability distribution of the Markov Random Field is then a joint probability distribution over visible and hidden units $(\boldsymbol{V}, \boldsymbol{H})$. The marginal distribution of $\boldsymbol{V}$ is a summation over the hidden units of the joint probability distribution:

$$p(u) = \sum_h p(\boldsymbol{u}, \boldsymbol{h}) = \frac{1}{Z} \sum_h e^{-E(\boldsymbol{u}, \boldsymbol{h})} \tag{1.15}$$

$$Z = \sum_{\boldsymbol{u}, \boldsymbol{h}} e^{-E(\boldsymbol{u}, \boldsymbol{h})}, \tag{1.16}$$

where Z is the partition function. Now, if we consider a single training example u from set S, the log-likelihood is:

$$\begin{aligned} \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u}) = \ln p(\boldsymbol{u}|\boldsymbol{\theta}) &= \ln \left(\frac{1}{Z} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u}, \boldsymbol{h})} \right) \\ &= \ln \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u}, \boldsymbol{h})} - \ln \sum_{\boldsymbol{u}, \boldsymbol{h}} e^{-E(\boldsymbol{u}, \boldsymbol{h})} \end{aligned} \tag{1.17}$$

The gradient of the log-likelihood is:

$$\begin{aligned}\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}}\left(\ln \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})}\right) - \frac{\partial}{\partial \boldsymbol{\theta}}\left(\ln \sum_{\boldsymbol{u},\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})}\right) \\ &= -\frac{1}{\sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})}} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})} \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \theta} \\ &+ \frac{1}{\sum_{\boldsymbol{u},\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})}} \sum_{\boldsymbol{u},\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})} \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \boldsymbol{\theta}}\end{aligned} \tag{1.18}$$

Given that the conditional probability is:

$$p(\boldsymbol{h}|\boldsymbol{u}) = \frac{p(\boldsymbol{u},\boldsymbol{h})}{p(\boldsymbol{u})} = \frac{\frac{1}{Z} e^{-E(\boldsymbol{u},\boldsymbol{h})}}{\frac{1}{Z}\sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})}} = \frac{e^{-E(\boldsymbol{u},\boldsymbol{h})}}{\sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u},\boldsymbol{h})}} \tag{1.19}$$

The gradient of the log-likelihood equals:

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial \boldsymbol{\theta}} = -\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \theta} + \sum_{\boldsymbol{u},\boldsymbol{h}} p(\boldsymbol{u},\boldsymbol{h}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \theta} \tag{1.20}$$

It is important to notice that the two terms correspond to expectation values under $p(\boldsymbol{h}|\boldsymbol{u})$ and $p(\boldsymbol{u},\boldsymbol{h})$. Therefore, it is possible to calculate the expectations by sampling a representative subset from the corresponding distributions through Markov Chain Monte Carlo.

1.2.2 *Discrete-Time Markov Chains*

Let us assume a sequence of discrete random variables $\{X^k \big|\ k \in N_0\}$ taking values in a state space Ω and for which $\forall\, k \geq 0$ and $\forall j, i, i_0, \ldots, i_{k-1} \in \Omega$ they satisfy the *Markov Property*:

$$\begin{aligned}p_{ij}^k &= Pr\Big(X^{(k+1)} = j \Big| X^{(k)} = i, X^{(k-1)} = i_{k-1}, \ldots X^{(0)} = 0\Big) \\ &= Pr\Big(X^{(k+1)} = j \Big| X^{(k)} = i\Big)\end{aligned} \tag{1.21}$$

The sequence $\{X^k \big|\ k \in N_0\}$ then defines a *Markov Chain*. The Markov property indicates that the discrete random variable X_{k+1} depends only upon X_k and not upon $X_{k-1}, \ldots, X_1, X_0$ so the Markov Chain is called memoryless.

A Markov Chain is *time-homogeneous* if the transition probabilities are independent of time. Equally, if for $k \geq 0,\ p_{ij}^{(k)} = p_{ij}$. A time-homogeneous Markov Chain is described by a *transition matrix* $\boldsymbol{P} = (p_{ij})_{i,j\in\Omega}$[4].

[4] The rows of the Transition Matrix add to 1 and it lists all possible states in Ω.

If we assume that the probability distribution of the state $X_{(0)}$ is given by a probability vector $\boldsymbol{\mu^{(0)}} = (\mu^{(0)}(i))_{i\in\Omega}$ with $\mu^{(0)}(i) = Pr\Big(X^{(0)} = i\Big)$, the

probability distribution $\boldsymbol{\mu^{(k)}}$ of random variable $X^{(k)}$ is given by:

$$\boldsymbol{\mu^{(k)T}} = \boldsymbol{\mu^{(0)T}} \boldsymbol{P^k} \tag{1.22}$$

Consequently, taking k further steps in the Markov Chain corresponds to multiplying by $\boldsymbol{P}^k$ according to the above equation.

Now, we define a *stationary* or *equilibrium* distribution π for the Markov Chain if:

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \boldsymbol{P} \tag{1.23}$$

If the Markov chain reaches at time k the stationary distribution $\boldsymbol{\mu}^k = \boldsymbol{\pi}$ then it stays there forever, i.e for all the subsequent states we have $\boldsymbol{\mu}^{(k+n)} = \boldsymbol{\pi}$, $\forall n \in N$. A distribution π is stationary to a Markov Chain if the transition probabilities $p_{ij}, i, j \in \Omega$ satisfy the *detailed balance* condition [5]

$$\pi(i)p_{ij} = \pi(j)p_{ji}, \ \forall \, i, j \in \Omega \tag{1.24}$$

[5] The Detailed Balance is a sufficient but not necessary condition for demonstrating that a given distribution is the *equilibrium* distribution.

A Markov Chain is *irreducible* if every state is reachable through a finite number of transitions[6]:

$$\forall i, j \in \Omega \, \exists \, k > 0 \, with \, Pr\Big(X^{(k)} = j \Big| X^{(0)} = i\Big) > 0 \tag{1.25}$$

[6] Irreducibility guarantees that the Markov Chain cannot be trapped at a subset of the states k. Therefore the initial state can also be chosen freely.

One can define the period $d(i)$ of a state i as the greatest common divisor gcd:

$$d(i) = gcd\{k \in N_0 \big| Pr(X^{(k)} = i | X^{(0)} = i > 0)\} \tag{1.26}$$

If $d(i) = 1$ for all states $i \in \Omega$ then the Markov Chain is *aperiodic* which implies that returning to a given state i is possible at irregular times[7].

[7] If $d(i) > 1$ the Markov chain is periodic. Periodicity implies that at regular intervals the Markov Chain can *only* return to a specific state i. This clearly interferes with our plan of reaching an equilibrium distribution.

The *total variation distance* between two distributions α and β on a finite probability space Ω is:

$$d_V(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}|\boldsymbol{\alpha} - \boldsymbol{\beta}| = \frac{1}{2}\sum_{x \in \Omega} |\alpha(x) - \beta x| \tag{1.27}$$

It is guaranteed for a Markov chain that is irreducible and aperiodic and for which an equilibrium distribution $\boldsymbol{\pi}^T$ exists that it will converge to $\boldsymbol{\pi}^T$ as $k \to \infty$. More specifically, considering an arbitrary initial distribution μ:

$$\lim_{k \to \infty} d_V(\boldsymbol{\mu}^T \boldsymbol{P}^k, \boldsymbol{\pi}^T) = 0. \tag{1.28}$$

The goal is then to construct a Markov Chain that converges asymptotically to the desired equilibrium distribution in order to acquire a subset of samples from it. These samples are then used to calculate expectation values of interest.

1.2.3 *Gibbs Sampling*

Gibbs sampling is a Markov Chain Monte Carlo technique and can be considered a special case of Metropolis-Hastings algorithms. The idea is to randomly choose a state from a *proposal* distribution and accept it based on an *acceptance* probability while satisfying the condition of detailed balance.

Let us define a Markov Random Field described by an undirected graph $G = (V, E)$, $V = \{1, \dots, N\}$ with a set of random variables $\boldsymbol{X} = (X_1^{(k)}, \dots, X_N^{(k)})$, $X_i, i \in V$ that take values in a finite set Λ and a joint probability for $\boldsymbol{X}$ equal to $\pi(\boldsymbol{x}) = \frac{1}{Z} e^{-\mathcal{E}(\boldsymbol{x})}$.

We will assume that the Markov Random Field evolves in time with discrete steps and it's state is given by $\boldsymbol{X}^{(k)} = (X_1^{(k)}, \dots, X_N^{(k)})$ for time $k \geq 0$. This time-discrete evolution can be viewed as a Markov chain $X = \{\boldsymbol{X}^{(k)} | k \in \mathbb{N}_0\}$ with state space $\Omega = \Lambda^N$.

A new state is proposed by choosing a random variable $X_i,\ i \in V$ with a probability q_i . It is then accepted based on the conditional probability distribution given the state $(x_u)_{u \in V \setminus i}$ of the remaining random variables $(X_u)_{u \in V \setminus i}$. The local Markov property of the Markov Random Field implies that $\pi(\boldsymbol{x_i} | (x_u)_{u \in V \setminus i}) = \pi(\boldsymbol{x_i} | (x_w)_{w \in \mathcal{N}_i})$. The transition probability between two different states $\boldsymbol{x}, \boldsymbol{y}$ with $\boldsymbol{x} \neq \boldsymbol{y}$ is defined as $p_{\boldsymbol{xy}}$:

$$p_{\boldsymbol{xy}} = \begin{cases} q(i)\pi(y_i | (x_u)_{u \in V \setminus i}), & if\ \exists i \in V\ \forall u \in V\ u \neq i : x_u = y_u \\ 0, & else \end{cases}$$

The probability to remain in the same state is $p_{\boldsymbol{xx}} = q(i)\pi(x_i | (x_u)_{u \in V \setminus i})$. If the Markov chain is irreducible and aperiodic then it will converge to its equilibrium distribution and if detailed balance is satisfied then π is the desired equilibrium distribution.

First we have to demonstrate that the condition of detailed balance is true. For the case of $\boldsymbol{x} = \boldsymbol{y}$ it is straightforward to prove. When $\boldsymbol{x}$ and $\boldsymbol{y}$ differ by more than one random variables $p_{\boldsymbol{xy}} = p_{\boldsymbol{yx}} = 0$. When they differ *exactly* by one variable X_i we have:

$$\begin{aligned} \pi(\boldsymbol{x}) p_{\boldsymbol{xy}} &= \pi(\boldsymbol{x}) q(i) \pi(y_i | (x_u)_{u \in V \setminus i}) \\ &= \pi(x_i, (x_u)_{u \in V \setminus i}) q(i) \frac{\pi(y_i, (x_u)_{u \in V \setminus i})}{\pi((x_u)_{u \in V \setminus i})} \\ &= \pi(y_i, (x_u)_{u \in V \setminus i}) q(i) \frac{\pi(x_i, (x_u)_{u \in V \setminus i})}{\pi((x_u)_{u \in V \setminus i})} \\ &= \pi(\boldsymbol{y}) q(i) \pi(x_i | (x_u)_{u \in V \setminus i}) \\ &= \pi(\boldsymbol{y}) p_{\boldsymbol{yx}} \end{aligned} \tag{1.29}$$

Therefore the condition of detailed balance holds and π is the equilibrium distribution. To prove the irreducibility of the Markov Chain one has to notice that since π is strictly positive, the conditional distributions will also be positive and

every possible state in the Markov Random Field can be reached. Additionally, the Markov Chain is aperiodic since $p_{\boldsymbol{xx}} > 0 \, \forall \boldsymbol{x} \in \Lambda^n$.

The deterministic choice of a state corresponds to the *periodic Gibbs sampler* algorithm where a bound can be defined for the convergence rate:

$$|\boldsymbol{\mu}\boldsymbol{P}^k - \boldsymbol{\pi}| \leq \frac{1}{2}|\boldsymbol{\mu} - \boldsymbol{\pi}|(1 - e^{-N\Delta})^k \tag{1.30}$$

where $\boldsymbol{P}$ is the transition matrix, $\Delta = \sup_{l \in V} \delta_l$, $\delta_l = \sup\{|\mathcal{E}(\boldsymbol{x}) - \mathcal{E}(\boldsymbol{y})|; x_i = y_i \forall i \in V \; with \; i \neq l\}$, μ an arbitrary distribution and $\frac{1}{2}|\boldsymbol{\mu} - \boldsymbol{\pi}|$ the total variation distance.

1.3.0 *Restricted Boltzmann Machines*

Having established Markov Random Fields in the previous sections, it is now possible to define Restricted Boltzmann Machines as Markov Random Fields corresponding to bipartite graphs with undirected edges. They are comprised of m visible units $\boldsymbol{V} = \big(V_1, \ldots, V_m\big)$ and n hidden units $\boldsymbol{H} = \big(H_1, \ldots, H_m\big)$. Considering that in the subsequent chapters we will study the $d = 2$ Ising model and the $d = 1$ transverse-field Ising model, i.e models consisting of spins with binary values, the focus is on binary Restricted Boltzmann Machines. The random variables $\big(\boldsymbol{V}, \boldsymbol{H}\big)$ then take values $(\boldsymbol{u}, \boldsymbol{h}) \in \{0, 1\}^{m+n}$. The joint probability distribution of the model is the Boltzmann probability distribution:

$$p(\boldsymbol{u}, \boldsymbol{h}) = \frac{1}{Z} e^{-E(\boldsymbol{u}, \boldsymbol{h})} \tag{1.31}$$

The energy function of the model $E(\boldsymbol{u}, \boldsymbol{h})$ is defined as:

$$E(\boldsymbol{u}, \boldsymbol{h}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} h_i u_j - \sum_{j=1}^{m} b_j u_j - \sum_{i=1}^{n} c_i h_i \tag{1.32}$$

The variational parameters of the model for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$ are the weights w_{ij} associating visible with hidden units, and the biases b_j and c_i of the jth visible and ith hidden unit. All the variational parameters are real-valued. The set of visible units defines the *visible layer* and the set of hidden units the *hidden layer*.

In Restricted Boltzmann Machines no intralayer connections are allowed. A conditional independence then holds for units in one layer, given the units in the other layer. In a formal notation:

$$p(\boldsymbol{h}|\boldsymbol{u}) = \prod_{i=1}^{n} p(h_i|u) \tag{1.33}$$

$$p(\boldsymbol{u}|\boldsymbol{h}) = \prod_{j=1}^{m} p(u_j|\boldsymbol{h}) \tag{1.34}$$

This instantly implies that one can sample all variables of a given layer in one step. Thus, only two steps are needed to sample hidden and then visible

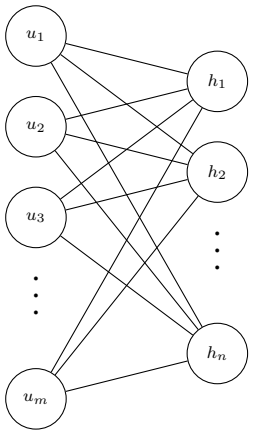


Figure 1.2: A Restricted Boltzmann Machine which is a bipartite graph with undirected edges. The neurons u_m and h_n correspond to the visible and hidden units respectively. The term "Restricted" implies that no intralayer connections are allowed. Biases are not shown, but to each unit corresponds a bias.

units, corresponding to what is known as *block Gibbs sampling*. The marginal distribution of the visible layer is given by:

$$
\begin{aligned}
p(\boldsymbol{u}) &= \sum_{h} p(\boldsymbol{u}, \boldsymbol{h}) = \frac{1}{Z} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{u}, \boldsymbol{h})} \\
&= \frac{1}{Z} \sum_{h_1} \sum_{h_2} \cdots \sum_{h_n} e^{\sum_{j=1}^{m} b_j u_j} \prod_{i=1}^{n} e^{h_i \left(c_i + \sum_{j=1}^{m} w_{ij} u_j\right)} \\
&= \frac{1}{Z} e^{\sum_{j=1}^{m} b_j u_j} \sum_{h_1} e^{h_1 \left(c_1 + \sum_{j=1}^{m} w_{1j} u_j\right)} \cdots \sum_{h_n} e^{h_n \left(c_n + \sum_{j=1}^{m} w_{nj} u_j\right)} \\
&= \frac{1}{Z} e^{\sum_{j=1}^{m} b_j u_j} \prod_{i=1}^{n} \sum_{h_i} e^{h_i \left(c_i + \sum_{j=1}^{m} w_{ij} u_j\right)} \\
&= \frac{1}{Z} \prod_{j=1}^{m} e^{b_j u_j} \prod_{i=1}^{n} \left(1 + e^{c_i + \sum_{j=1}^{m} w_{ij} u_j}\right)
\end{aligned}
\tag{1.35}
$$

The expression of this marginal distribution demonstrates why Restricted Boltzmann Machines can be considered *product of experts* models [8].

A Restricted Boltzmann Machine consisting of m visible and $k+1$ hidden units has the capacity to model a target distribution on $\{0,1\}^m$. The quantity k expresses the number of elements from $\{0,1\}^m$ that are able to appear as an observation with a probability that does not equal zero. There is an association between dependencies among visible units and the number of hidden units required to model a target distribution and even a smaller number of hidden units could prove to be adequate [9].

To calculate conditional probabilities of a single hidden or visible unit, one can define as $\boldsymbol{u}_{-l}$ the set of visible variables excluding the lth one and:

$$
a_l(\boldsymbol{h}) = -\sum_{i=1}^{n} w_{il} h_i - b_l \tag{1.36}
$$

$$
\beta(\boldsymbol{u}_{-l}, \boldsymbol{h}) = -\sum_{i=1}^{n} \sum_{j=1, j \neq l}^{m} w_{ij} h_i u_j - \sum_{j=1, j \neq l}^{m} b_j u_j - \sum_{i=1}^{n} c_i h_i \tag{1.37}
$$

The energy function $E(\boldsymbol{u}, \boldsymbol{h})$ is then given by:

$$
E(\boldsymbol{u}, \boldsymbol{h}) = \beta(\boldsymbol{u}_{-l}, \boldsymbol{h}) + u_l a_l(\boldsymbol{h}) \tag{1.38}
$$

where the quantity $u_l a_l(\boldsymbol{h})$ implies a collection of terms of u_l. The conditional probability of the V_l visible unit given the hidden layer $\boldsymbol{h}$ is then equal to [10] :

[8] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation 14, 1771-1800*, 2002; and Max Welling. Product of Experts. *Scholarpedia, 2(10):3879*, 2007

[9] Guido Montufar and Nihat Ay. Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. *Neural Computation*, 2011

[10] Yoshua Bengio. *Learning deep architectures for AI.* Foundations and Trends in Machine Learning 21(6), 1601-1621, 2009

$$
\begin{aligned}
p(V_l = 1|\boldsymbol{h}) = p(V_l = 1|\boldsymbol{u}_{-l}, \boldsymbol{h}) &= \frac{p(V_l = 1, \boldsymbol{u}_{-l}, \boldsymbol{h})}{p(\boldsymbol{u}_{-l}, \boldsymbol{h})} \\
&= \frac{e^{-E(u_{l=1}, \boldsymbol{u}_{-l}, \boldsymbol{h})}}{e^{-E(u_{l=1}, \boldsymbol{u}_{-l}, \boldsymbol{h})} + e^{-E(u_{l=0}, \boldsymbol{u}_{-l}, \boldsymbol{h})}} \\
&= \frac{e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h}) - 1 \cdot a_l(\boldsymbol{h})}}{e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h}) - 1 a_l(\boldsymbol{h})} + e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h}) - 0 \cdot a_l(\boldsymbol{h})}} \\
&= \frac{e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h})} \cdot e^{-a_l(\boldsymbol{h})}}{e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h})} \cdot e^{-a_l(\boldsymbol{h})} + e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h})}} \\
&= \frac{e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h})} e^{-a_l(\boldsymbol{h})}}{e^{-\beta(\boldsymbol{u}_{-l}, \boldsymbol{h})} \cdot (e^{-a_l(\boldsymbol{h})} + 1)} \\
&= \frac{e^{-a_l(\boldsymbol{h})}}{e^{-a_l(\boldsymbol{h})} + 1} \\
&= \frac{\frac{1}{e^{a_l(\boldsymbol{h})}}}{\frac{1}{e^{a_l(\boldsymbol{h})}} + 1} \\
&= \frac{1}{1 + e^{a_l(\boldsymbol{h})}} \\
&= \sigma(-a_l(\boldsymbol{h})) \\
&= \sigma\left(\sum_{i=1}^{n} w_{il} h_i + b_l \right)
\end{aligned}
\tag{1.39}
$$

The function σ is the sigmoid function:

$$
\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1.40}
$$

One can derive a similar equation for a hidden unit given the visible layer. Following the initial notation of i and j both equations are given below:

$$
p(H_i = 1|\boldsymbol{u}) = \sigma\left(\sum_{j=1}^{m} w_{ij} u_j + c_i \right) \tag{1.41}
$$

$$
p(V_j = 1|\boldsymbol{h}) = \sigma\left(\sum_{i=1}^{n} w_{ij} h_i + b_j \right) \tag{1.42}
$$

Recall the log-likelihood gradient equation (1.20) of the Markov Random

Field. If one sets the parameter θ equal to the weights w_{ij} the first term gives:

$$\begin{aligned}
\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial w_{ij}} &= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) h_i u_j \\
&= \sum_{\boldsymbol{h}} \prod_{k=1}^{n} p(h_k|\boldsymbol{u}) h_i u_j \\
&= \sum_{h_i} \sum_{h_{-i}} p(h_i|\boldsymbol{u}) p(\boldsymbol{h_{-i}}|\boldsymbol{u}) h_i u_j \\
&= \sum_{h_i} p(h_i|\boldsymbol{u}) h_i u_j \underbrace{\sum_{\boldsymbol{h_{-i}}} p(\boldsymbol{h_{-i}}|\boldsymbol{u})}_{=1} \\
&= p(H_i = 1|\boldsymbol{u}) u_j \\
&= \sigma\left(\sum_{j=1}^{m} w_{ij} u_j + c_i \right) u_j
\end{aligned} \tag{1.43}$$

The second term can be written as:

$$\begin{aligned}
\sum_{\boldsymbol{u},\boldsymbol{h}} \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \theta} &= \sum_{\boldsymbol{u}} p(\boldsymbol{u}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \theta} \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}) \sum_{\boldsymbol{u}} p(\boldsymbol{u}|\boldsymbol{h}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial \theta}
\end{aligned} \tag{1.44}$$

Notice that the outer sum in both cases has an exponential complexity since it is a summation over 2^N states. Thus the quantity to be calculated remains intractable even if the inner sum gets factorized in an analogous way.

The derivative of the log-likelihood for the case of setting the variational parameter θ equal to the weights w_{ij} is then:

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial w_{ij}} &= -\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial w_{ij}} + \sum_{\boldsymbol{u},\boldsymbol{h}} p(\boldsymbol{u},\boldsymbol{h}) \frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial w_{ij}} \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) h_i u_j - \sum_{\boldsymbol{u}} p(\boldsymbol{u}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}) h_i u_j \\
&= p(H_i = 1|\boldsymbol{u}) u_j - \sum_{\boldsymbol{u}} p(\boldsymbol{u}) p(H_i = 1|\boldsymbol{u}) u_j
\end{aligned} \tag{1.45}$$

To adopt the common notation in literature, and assuming a training set

$S = \{u_1, \ldots, u_l\}$, the mean value of the derivative of the log-likelihood is:

$$\begin{aligned}\frac{1}{l}\sum_{\boldsymbol{u}\in S}\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial w_{ij}} &= \frac{1}{l}\sum_{\boldsymbol{u}\in S}\left[-\mathcal{E}_{p(\boldsymbol{h}|\boldsymbol{u})}\left[\frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial w_{ij}}\right] + \mathcal{E}_{p(\boldsymbol{h},\boldsymbol{u})}\left[\frac{\partial E(\boldsymbol{u},\boldsymbol{h})}{\partial w_{ij}}\right]\right] \\ &= \frac{1}{l}\sum_{\boldsymbol{u}\in S}\left[\mathcal{E}_{p(\boldsymbol{h}|\boldsymbol{u})}[u_i h_j] - \mathcal{E}_{p(\boldsymbol{h},\boldsymbol{u})}[u_i h_j]\right] \\ &= \langle u_i h_j\rangle_{p(\boldsymbol{h}|\boldsymbol{u})q(\boldsymbol{u})} - \langle u_i h_j\rangle_{p(\boldsymbol{h},\boldsymbol{u})}\end{aligned} \tag{1.46}$$

The distribution q is the distribution represented by the training set and the above result can be rephrased as:

$$\sum_{\boldsymbol{u}\in S}\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial w_{ij}} \propto \langle u_i h_j\rangle_{data} - \langle u_i h_j\rangle_{model} \tag{1.47}$$

Now we can set the parameter θ equal to the rest of the variational parameters, i.e the biases b_j and c_i, and acquire the following expressions for the derivatives:

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial b_j} = u_j - \sum_{\boldsymbol{u}} p(\boldsymbol{u})u_j \tag{1.48}$$

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{u})}{\partial c_i} = p(H_{i=1}|\boldsymbol{u}) - \sum_{\boldsymbol{u}} p(\boldsymbol{u})p(H_i = 1|\boldsymbol{u}) \tag{1.49}$$

One can now sample through Markov Chain Monte Carlo the corresponding terms of the above equations. There is still a problem though. Acquiring a representative subset of samples from the model distribution would require the Markov Chain to run sufficiently enough until it reaches equilibrium. This is clearly not feasible due to computational cost and a further approximation has to be introduced.

1.3.1 *Contrastive Divergence*

Contrastive Divergence is the most common approximation technique for calculating expectations in the log-likelihood gradient under the model distribution [11]. It is often denoted as CD-k where k is the number of steps that it is performed.

Instead of carrying out iterative Gibbs Sampling steps to reach equilibrium in the model distribution, one can set the visible units equal to a training example $u^{(0)}$ and run a Gibbs chain for k steps, acquiring a *reconstruction* $u^{(k)}$. For a large number of problems, even $k = 1$ proves to be enough. The biased approximation of the gradient log-likelihood with respect to a variational parameter θ is then given by:

$$CD_k(\boldsymbol{\theta}, \boldsymbol{u^{(0)}}) = -\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}^{(0)})\frac{\partial E(\boldsymbol{u}^{(0)},\boldsymbol{h})}{\partial \boldsymbol{\theta}} + \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}^{(k)})\frac{\partial E(\boldsymbol{u}^{(k)},\boldsymbol{h})}{\partial \boldsymbol{\theta}} \tag{1.50}$$

[11] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation 14, 1771-1800*, 2002; Yoshua Bengion, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing (NIPS 19), pp. 153-160*, 2007. MIT Press; Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation 18(7), 1527-1554*, 2006; Yoshua Bengio and Olivier Delalleau. Justifying and generalizing contrastive divergence. *Neural Computation 21(6), 1601-1621*, 2009; and Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. Learning multiple layers of representation. *Trends in Cognitive Sciences 11(10), 428-434*, 2007

One can choose to carry out contrastive divergence using the complete data set S at each step resulting in what is known *batch learning*. The optimal way is instead to use a *mini-batch*, i.e a subset $S' \subset S$, especially when dealing with large data sets.

In any case, contrastive divergence is an approximation technique and the resulting sample might not be from the equilibrium distribution of the model. The approximation is thus biased.

The theorem by Bengio and Delalleau [12] as appearing in [13] is important in acquiring a better understanding of Constrastive Divergence. It states that for a Gibbs that is led into convergence:

$$\boldsymbol{u}^{(0)} \Rightarrow \boldsymbol{h}^{(0)} \Rightarrow \boldsymbol{u}^{(1)} \Rightarrow \boldsymbol{h}^{(1)} \dots \tag{1.51}$$

The log-likelihood gradient equals

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\boldsymbol{u}^{(0)}) = &- \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}^{(0)}) \frac{\partial E(\boldsymbol{u}^{(0)}, \boldsymbol{h})}{\partial \boldsymbol{\theta}} \\ &+ E_{p(\boldsymbol{u}^{(k)})|\boldsymbol{u}^{(0)}} \left[\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{u}^{(k)}) \frac{\partial E(\boldsymbol{u}^{(k)}, \boldsymbol{h})}{\partial \theta} \right] \\ &+ E_{p(\boldsymbol{u}^{(k)})|\boldsymbol{u}^{(0)}} \left[\frac{\partial \ln p(\boldsymbol{u}^{(k)})}{\partial \theta} \right] \end{aligned} \tag{1.52}$$

and the final term converges to zero as $k \to \infty$.

The mixing rate of the Markov chain is a measure of how fast it reaches the equilibrium distribution. It is described by the transition probabilities and is one of the factors-along with the number of steps- that influences the approximation error. The magnitude of the variational parameters affects the mixing rate. This is evident from the the expressions of the conditional probabilities in terms of the sigmoid function where high values of the variational parameters correspond to values close to zero or one for the conditional probabilities. The Markov chain then has a slower time evolution.

The following theorem [14] gives an upper bound on the expectation of the CD approximation error under the empirical distribution:

Theorem 1 (Fischer and Igel)**.** *Let p be the marginal distribution of the RBM visible units and q the empirical distribution defined by a set of samples $\boldsymbol{u}, \dots, \boldsymbol{u}_l$. An upper bound on the expectation of the error of the CD-k approximation of the log-likelihood derivative with respect to an RBM parameter θ_a is*

$$\left| E_{(q)(\boldsymbol{u}^{(0)})} \left[E_{p(\boldsymbol{u}^{(k)}|\boldsymbol{u}^{(0)})} \left[\frac{\partial \ln p(\boldsymbol{u}^{(k)})}{\partial \theta_a} \right] \right] \right| \leq \frac{1}{2} |q - p| \left(1 - e^{-(m+n)\Delta}\right)^k \tag{1.53}$$

[12] Yoshua Bengio and Olivier Delalleau. Justifying and generalizing contrastive divergence. *Neural Computation 21(6), 1601-1621*, 2009

[13] Asja Fischer and Christian Igel. An Introduction to Restricted Boltzmann Machines. *Alvarez L., Mejail M., Gomez L., Jacobo J. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2012. Lecture Notes in Computer Science, vol 7441.*, 2012. Springer, Berlin, Heidelberg

[14] Asja Fischer and Christian Igel. Bounding the bias of contrastive divergence learning. *Neural Computation 23, 664-673*, 2011; and Asja Fischer and Christian Igel. An Introduction to Restricted Boltzmann Machines. *Alvarez L., Mejail M., Gomez L., Jacobo J. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2012. Lecture Notes in Computer Science, vol 7441.*, 2012. Springer, Berlin, Heidelberg

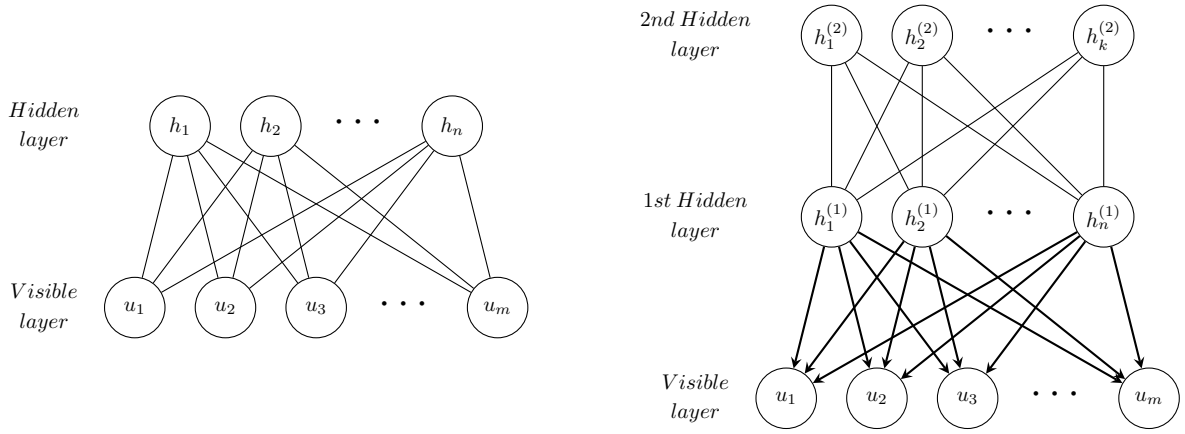


Figure 1.3: A comparison of the (i) Restricted Boltzmann presented in top-down approach and (ii) A Deep Belief Network that features both directed and undirected edges. It consists of multiple hidden layers, and similarly with the Restricted Boltzmann Machine no intralayer connections are allowed.

with

$$\Delta = max\left\{ \max_{l\in\{1,\dots,m\}} \theta_l, \max_{l\in\{1,\dots,n\}} \xi_l \right\} \tag{1.54}$$

where

$$\theta_l = max\left\{ \left| \sum_{i=1}^{m} I_{\{w_{il}>0\}} w_{il} + b_l \right|, \left| \sum_{i=1}^{m} I_{\{w_{il}<0\}} w_{il} + b_l \right| \right\} \tag{1.55}$$

and

$$\xi_l = max\left\{ \left| \sum_{j=1}^{m} I_{\{w_{lj}>0\}} w_{lj} + c_l \right|, \left| \sum_{j=1}^{m} I_{\{w_{lj}<0\}} w_{lj} + c_l \right| \right\} \tag{1.56}$$

The bound shows a dependence on the amount of visible and hidden units of the Restricted Boltzmann Machine, the absolute variational parameters and the variation distance between the model distribution and the initial distribution of the Gibbs chain. It is not necessary that completing a contrastive divergence learning will result in a maximum likelihood training because of the inherent approximation error. The bias might also lead the parameters to converge in values that don't correspond to maximum likelihood. Additionally, the likelihood might start to diverge after some iterations if the contrastive divergence steps k are not high enough. Fine tuning the weight decay can help deal with the last problem.

1.4.0 *Deep Belief Networks*

Restricted Boltzmann Machines are very common in deep learning architectures. One can stack RBMS and form a generative model that consists of multiple

hidden layers called *Deep Belief Network* [15]. Similarly with Restricted Boltzmann Machines no intralayer connections are allowed in a Deep Belief Network. The connections between the last two hidden layers are undirected while all other are directed. Every two layers in a DBN are described by a set of variational parameters $\{\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{a}\}$.

A Deep Belief Network is formed by initially training a Restricted Boltzmann Machine using contrastive divergence. The values of the hidden layer when the data are clamped to the visible layer serve as the training set of a second Restricted Boltzmann Machine . This procedure can then be repeated many times to form new layers.

Generating samples from a DBN is possible by alternate Gibbs sampling of the last two hidden layers until they reach equilibrium. A single pass through the rest of the model will then result in a reconstruction from the visible layer.

[15] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neu- ral networks. *Science 313(5786), 504-507*, 2006; and Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. Learning multiple layers of representation. *Trends in Cognitive Sciences 11(10), 428-434*, 2007

2. The Ising Model

2.1.0 The Ising Model in $d = 2$ and it's Second Order Phase Transition

A short but concise introduction follows for the $2D$ Ising Model and it's second order phase transition. Markov Chain Monte Carlo simulations are also included based on previous chapters, relevant literature [1] and publications.

A Markov Chain Monte Carlo(MCMC) simulation is performed on a system described by the canonical ensemble in order to calculate the expectation value of an observable quantity $\mathcal{O}$:

$$\langle \mathcal{O} \rangle = \frac{1}{Z} \sum_{k=1}^{K} p_k \mathcal{O}^{(k)} = \frac{1}{Z} \sum_{k=1}^{K} \mathcal{O}^{(k)} e^{-\beta E^{(k)}}. \tag{2.1}$$

The inverse temperature $\beta = 1/k_\beta T$ weighs the energy in the exponential and determines a characteristic energy scale. The Boltzmann constant is subsequently chosen to be $k_\beta = 1$. The superscript k denotes a state of the system to which corresponds a configuration $\{s_i\}$ with degrees of freedom s_i, $i = 1, \ldots, N$. A given configuration has an observable quantity $\mathcal{O}^{(k)}$ and an internal energy $E^{(k)}$. The sum is over all possible configurations $\{s_i\}$. The normalizing constant Z which encodes all the statistical information of the system by counting all states with their correct weight is called partition function and is given by:

$$Z = Z(\beta) = \sum_{k=1}^{K} e^{-\beta E^{(k)}}. \tag{2.2}$$

Any thermodynamic observable quantity can be calculated with prior knowledge of the partition function Z. For example the internal energy $U \equiv \langle E \rangle$, the specific heat C and the Helmholtz free energy F are given by:

$$U \equiv \langle E \rangle = -\frac{\partial \ln Z}{\partial \beta} \tag{2.3}$$

$$C = k\beta^2 \frac{\partial^2 \ln Z}{\partial \beta^2} \tag{2.4}$$

$$F = -\frac{1}{\beta} \ln Z \tag{2.5}$$

[1] Konstantinos N. Anagnostopoulos. *Computational Physics: A Practical Introduction to Computational Physics and Scientific Computing (Using C++)*. Konstantinos N. Anagnostopoulos and the National Technical University of Athens, 2016; Bernd A. Berg. *Markov Chain Monte Carlo Simulations and their Statistical Analysis: With Web-Based Fortran Code*. World Scientific Publishing Co. Pte. Ltd, 2004; and M.E.J Newman and G.T Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, 1999

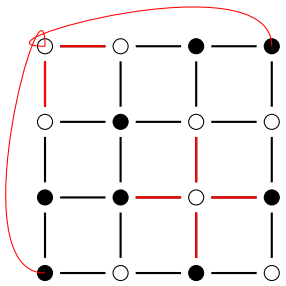


Figure 2.1: A $N = 4 * 4$ Ising Model in a square lattice with periodic boundary conditions. The red lines correspond to the nearest neighbor interaction for the selected spin. The case of selecting a boundary spin is included. Black dots resemble spins facing downward $s_i = -1$ and white dots spins facing upward $s_i = +1$.

Let us assume a field of fixed value Y and a conjugate variable X coupling to it, included in the form $-XY$ in the Hamiltonian. It is possible to calculate the expectation value of X as:

$$\langle X \rangle = -\frac{\partial F}{\partial Y}, \tag{2.6}$$

and the susceptibility, i.e the response of X to changes in Y as:

$$\chi = \frac{\partial \langle X \rangle}{\partial Y} \tag{2.7}$$

The Ising model [2] has a Hamiltonian:

[2] E. Ising. Beitrag zur Theorie des Ferromagnetizmus. *Z. Phys. 31*, 1925

$$H = -\sum_{\langle ij \rangle} J_{ij} s_i s_j + B \sum_i s_i, \tag{2.8}$$

with possible spin values $s_i = \pm 1$ positioned on sites of a hypercubic d-dimensional lattice. The constants J_{ij} measure the strength of interaction between spins and will be set to $J_{ij} = J = 1$, defining a ferromagnetic model[3]. The external magnetic field B is set to zero for the rest of the chapter. The total amount of spins equals $N = \prod_{i=1}^{d} n_i$ where n_i are the edge lengths of the lattice and d the dimensionality of the model. We will assume nearest-neighbor interactions $\langle ij \rangle$ and a dimensionality of $d = 2$, studying the square lattice. The system has a total number of 2^N configurations and is impossible to solve by carrying out the summation except for very small sizes. Thus, the need for a statistical approach arises.

[3] $J > 0$ defines a ferromagnetic model while $J < 0$ defines an anti-ferromagnetic model.

Thanks to the analytical solution of the 2D Ising Model[4], it's relevant simplicity and the fact that it exhibits a second order phase transition for the critical temperature:

[4] Lars Onsager. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physical Review, Series II*, 1944

$$\beta_c = \frac{1}{T_c} = \frac{1}{2} \ln(1 + \sqrt{2}) \approx 0.44068679\ldots, \tag{2.9}$$

it is a perfect candidate for testing novel simulation techniques. It also had a great impact in the study of statistical physics and quantum field theory. Additionally, Onsager's analytical solution provides exact values of the scaling relations known as *critical exponents* that take universal values irrelevant of the system's topology or form of interaction.

Let us define the correlation length ξ as a measure of the lattice spacing where two degrees of freedom are measurably correlated. The reduced temperature t is defined as the distance from the critical point:

$$t = \frac{T - T_c}{T_c} = \frac{\beta_c}{\beta} - 1, \tag{2.10}$$

Onsager's exponent values are then given by:

$$\begin{aligned} correlation\ length\ \xi &\sim |t|^{-\nu} \\ specific\ heat\ C &\sim |t|^{-a} \\ magnetization\ M &\sim |t|^{\beta} \\ magnetic\ susceptibility\ \chi &\sim |t|^{\gamma} \\ magnetization(field) M &\sim B^{-1/\delta}, (t=0) \\ correlations\ \langle s_i s_j \rangle &\sim |x_i - x_j|^{-d+2-\eta}, \\ & \quad for\ |x_i - x_j| \to \infty, (t=0) \end{aligned} \tag{2.11}$$

$$\nu = 1,\ a = 0,\ \beta = \frac{1}{8},\ \gamma = \frac{7}{4},\ \delta = 15, \eta = \frac{1}{4} \tag{2.12}$$

We will revisit the critical exponents later as they will be calculated with Real-Space Renormalization Group.

Before moving on we shall proceed with some necessary definitions. *Order parameters* are an important tool in the identification of a second order phase transition as they characterize a symmetry of the system. In the Ising Model the magnetization M is an order parameter vanishing in the disordered phase due to the Z_2 symmetry $s_i \to -s_i$ [5] whereas it has a constant value for the ordered phase. The magnetization then is, a non analytic function of the temperature.

[5] The Hamiltonian of the Ising Model is the same under the reflection symmetry $s_i \to -s_i$ on a configuration $\{s_i\}$ of spins, while the magnetization M is opposite.

The critical exponents define a *universality class*. All models in the same universality class must share same symmetries and dimensionality of space. They exhibit the same large-scale behavior as their microscopic description becomes irrelevant. Universality and scale invariance appear as the correlation length of the system $\xi \to \infty$. The diverging correlation length is uniquely defined, as all quantities diverge in terms of one parameter, the reduced temperature.

Scale invariance denotes that large scale interactions of the model only depend on the ratio of the corresponding length to the diverging correlation length. The correlation length exceeds any characteristic length scale of the system, e.g the lattice spacing, as the reduced temperature is decreased. In accordance with the universality class it is enough to find a scale-invariant model in the appropriate dimension that has the required symmetry to simplify the study of a second order phase transition near the critical point.

2.1.1 *Importance Sampling and Re-weighting*

The estimator of an observable quantity $\mathcal{O}$ from M Monte Carlo measurements $\mathcal{O}_i$ is given by the equation:

$$\mathcal{O}_M = \frac{\sum_i \mathcal{O}_i p_i^{-1} e^{-\beta E_i}}{\sum_j p_j^{-1} e^{-\beta E_j}} \tag{2.13}$$

The Boltzmann probabilities p correspond to the states sampled at the speci-

fied temperature, recasting the above equation to an importance sampled form:

$$\mathcal{O}_M = \frac{\sum_i \mathcal{O}_i (e^{-\beta E_i})^{-1} e^{-\beta E_i}}{\sum_j e^{(-\beta E_j)^{-1}} e^{-\beta E_j}} = \frac{1}{M} \sum_i \mathcal{O}_i \tag{2.14}$$

If we assume at the above equation Boltzmann probabilities of a sufficiently close temperature β_0 we have:

$$\mathcal{O}_M = \frac{\sum_i \mathcal{O}_i e^{-(\beta-\beta_0)E_i}}{\sum_j e^{-(\beta-\beta_0)E_j}} \tag{2.15}$$

with the possibility of re-weighting in a range $\beta_0 \pm \Delta\beta$ where $\Delta\beta \to 0$ in the thermodynamic limit [6].

[6] Alan M. Ferrenberg and Robert H. Swendsen. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett. 61*, 1988

Rewriting the partition function in terms of the density of states $n(E)$, i.e the number of configurations with internal energy E we have:

$$Z = Z(\beta) = \sum_E n(E) e^{-\beta E} \tag{2.16}$$

For a given value of β in a range of temperatures that are sufficiently distant from the critical point the energy probability density is:

$$P_\beta(E) = c_\beta n(E) e^{-\beta E} \tag{2.17}$$

with c_β the appropriate normalization constant. The energy probability density peaks around the average value of $E(\beta)$ with a width proportional to the square root of $\sqrt{V}$ where V is the volume of the system. It's due to the local correlation of the spins, away from the critical region, that the fluctuations are of $N \sim V$ and a typical fluctuation is $\sim \sqrt{N}$. The re-weighting range mentioned earlier is of $\Delta\beta \sim 1/\sqrt{V}$ keeping $\Delta\beta E \sim \sqrt{V}$ within the fluctuations of the system. We note that the larger fluctuations present in the critical region allow for larger re-weighting ranges.

It is possible to classify for a given temperature β the importance of configurations based on the large values of the probability density $P_\beta(E)$. One must then sample configurations with the appropriate Boltzmann weights through the implementation of a Markov process:

$$w_B^{(k)} = e^{-\beta E^{(k)}} \tag{2.18}$$

2.1.2 *Metropolis algorithm*

Using the mathematical concepts introduced for Markov Chains and Gibbs Sampling in 1.2.2 and 1.2.3 we assume a given configuration k for which the transition probability to obtain a configuration l is given by $W^{(l)(k)} = W[k \to l]$. The transition matrix W is then defined as:

$$W = \left(W^{(l)(k)}\right) \tag{2.19}$$

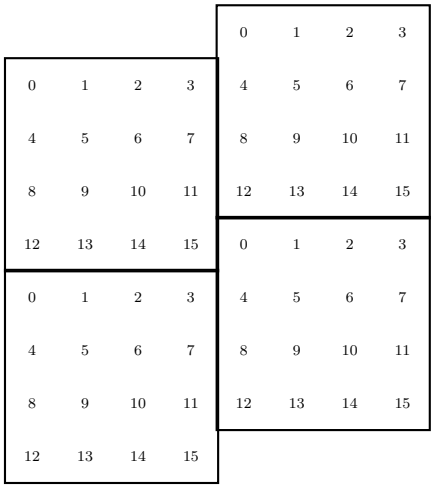


Figure 2.2: An example of helical boundary conditions chosen for the implementation of the Metropolis algorithm.

The condition of detailed balance does not imply a unique set of values for the transition probabilities $W^{(l)(k)}$. The Metropolis algorithm [7] which is widely used and computationally simple, proposes for a given configuration k, new configurations l with transition probabilities $f(l,k)$ which are normalized:

$$\sum_{l} f(l,k) = 1 \tag{2.20}$$

The probability to accept a new configuration l is:

$$w^{(l)(k)} = min\left[1, \frac{P_B^l}{P_B^k}\right] = \begin{cases} 1 & for\ E^l < E^k \\ e^{-\beta(E^l - E^k)} & for\ E^l > E^k \end{cases} \tag{2.21}$$

The *acceptance rate* can be defined as a ratio of accepted new configurations over proposed moves. This definition does not include as accepted the proposal of the current configuration .

The Metropolis algorithm results in the transition probabilities:

$$W^{(l)(k)} = f(l,k) w^{(l)(k)}\ for\ l \neq k \tag{2.22}$$

$$W^{(k)(k)} = f(k,k) + \sum_{l \neq k} f(l,k)(1 - w^{(l)(k)}) \tag{2.23}$$

In order for the condition of detailed balance to be satisfied for $W^{(l)(k)}/W^{(k)(k)}$ the symmetry condition below must be used:

$$f(l,k) = f(k,l) \tag{2.24}$$

In general the probability density $f(l,k)$ could be unconstrained and allow a variety of choices for the transition probabilities. It is also possible to use different acceptance probabilities resulting in non-symmetric proposal probabilities[8].

The observable quantities to be measured at each sweep are the internal energy:

$$E = -\sum_{\langle ij \rangle} s_i s_j, \tag{2.25}$$

and the magnetization

$$M = \left| \sum_i s_i \right| \tag{2.26}$$

It is of high importance to measure the absolute value of the magnetization. The Z_2 symmetry implies that configurations with every spin opposite have equal probability to appear. One can normalize the energy *per link*, acquiring the expression:

$$\langle e \rangle = \frac{1}{N_l} \langle E \rangle = \frac{1}{2N} \langle E \rangle \tag{2.27}$$

[7] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, and Augusta H. Teller. Equation of State Calculations by Fast Computing Machines . *The Journal of Chemical Physics 21, 1087*, 1953

[8] Wilfred K. Hastings. Monte Carlo sampling methods using Markov chains and their applications . *Biometrika 57: 97-109*, 1970

Similarly the magnetization can be normalized *per spin*:

$$\langle m \rangle = \frac{1}{N} \langle M \rangle \tag{2.28}$$

The rest of the quantities that will be of interest in our calculations are the specific heat:

$$c = \beta^2 N \langle (e - \langle e \rangle^2) \rangle = \beta^2 N (\langle e^2 \rangle - \langle e \rangle^2) \tag{2.29}$$

and the magnetic susceptibility:

$$\chi = \beta N \langle (m - \langle m \rangle^2) \rangle = \beta N (\langle m^2 \rangle - \langle m \rangle^2) \tag{2.30}$$

2.1.3 *Wolff's Cluster Algorithm*

The development of a cluster Markov Chain Monte Carlo algorithm came as a result of a mapping from the q state Potts Model to a percolation model of spin clusters.[9]

The benefit of using a cluster algorithm is the option to flip clusters of spins at every sweep, therefore reducing or eliminating the critical slowing effect during the study of second order phase transitions.

The formulation of Swendsen-Wang clusters is achievable by creating bonds among neighboring sites i and j with probability:

$$p_{\langle ij \rangle} = max[0, 1 - exp(-2\beta\delta_{q_i, q_j})] \tag{2.31}$$

The Wolff cluster algorithm [10] that allows the flipping of one cluster per sweep will be implemented. The steps describing the algorithm are:

1. We randomly choose a site i and assign it to a corresponding cluster c .
2. The nearest neighbors of site i are added to the cluster based on the probability $p_{\langle ij \rangle}$.
3. Subsequent iterations of the above step are executed on all added sites until the process is terminated.
4. All sites of the cluster c are assigned a new spin value $q_i^{'} \neq q_i$ which is chosen uniformly from the set of $q - 1$ alternative values, where $q = 2$ for the model studied in this thesis .

There is a finite probability that a single site q_i can formulate a cluster and every spin $q_i' \neq q_i$ is reachable. Repeating the above for a number of times greater or equal to the size of the system, there is a finite probability that all sites can be included. This guarantees the irreducibility of the Markov Chain.

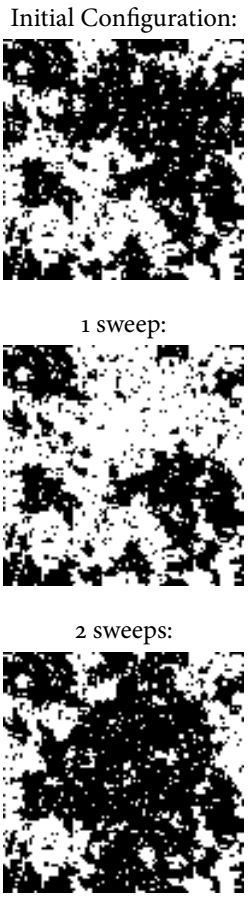


Figure 2.3: Configurations of the 2D Ising Model sampled with the Wolff Algorithm for $b = 0.43$ and $L = 100$.

[9] Robert H. Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett. 58, 86*, 1987; C. M. Fortuin and P. W. Kasteleyn. On the random-cluster model: I. Introduction and relation to other models. *Physica, Volume 57, Issue 4*, 1972; Dietrich Stauffer and Ammon Aharony. *Introduction To Percolation Theory*. CRC Press, 1994; and Henk W. J. BlÃűte and Youjin Deng. Cluster Monte Carlo simulation of the transverse Ising model. *Phys. Rev. E 66, 066110*, 2002

[10] Ulli Wolff. Collective Monte Carlo Updating for Spin Systems. *Phys. Rev. Lett. 62, 361*, 1989

To show that the condition of detailed balance holds one has to consider two configurations $\{q_i\}$ and $\{q_{i'}\}$ that differ *exactly by one* flip of a cluster c. The cluster c has then a probability:

$$P_c = \frac{|c|}{N} \prod_{\langle ij \rangle \in c} p_{\langle ij \rangle} \prod_{\langle ij \rangle \in c, j \neq c} exp(-2\beta\delta_{q_i,q_j}) \tag{2.32}$$

The cluster c has a number of sites $|c|$. The quantity $|c|/N$ defines the probability to pick a spin of the cluster at the first step of the Wolff Algorithm. The cluster that differs by exactly one flip in the configuration $\{q_i^{'}\}$ has a probability:

$$P_c^{'} = \frac{|c|}{N} \prod_{\langle ij \rangle \in c} p_{\langle ij \rangle}^{'} \prod_{\langle ij \rangle \in c, j \neq c} exp(-2\beta\delta_{q_i^{'},q_j^{'}}) \tag{2.33}$$

Since the cluster in both cases consists of identical spins $p_{\langle ij \rangle}^{'} = p_{\langle ij \rangle}$. The rest of the spins outside the clusters in both systems are identical since we have assumed that they differ only by a flip of c. The condition of detailed balance is then:

$$\frac{W^{(\{q_i^{'}\},\{q_i\})}}{W^{(\{q_i\},\{q_i^{'}\})}} = \frac{W(\{q_i\} \to \{q_i^{'}\})}{W(\{q_i^{'}\} \to \{q_i\})} = \frac{\exp\left(-2\beta \sum_{\langle ij \rangle} \delta_{q_i,q_j}\right)}{\exp\left(-2\beta \sum_{\langle ij \rangle} \delta_{q_i^{'},q_j^{'}}\right)} = \frac{\exp(2\beta E)}{\exp(2\beta E')} \tag{2.34}$$

A whole study can be conducted in order to compare the performance of the two algorithms in the critical region. This study though overlaps with the author's undergraduate thesis and is therefore skipped. The focus is on demonstrating that the neural network can calculate observable quantities with high accuracy using both algorithms.

2.1.4 *Equilibration and Autocorrelation*

A typical Markov Chain Monte Carlo simulation can be separated into two essential parts. The *equilibration* part, also known as thermalization, and the *production* part. The equilibration part consists of the initial sweeps performed in order to reach the equilibrium distribution which are of no use to the calculation of expectation values, and are thus to be discarded. The production part consists of the subsequent sweeps after reaching equilibrium which are used for measurements and the calculation of the expectation values.

One can naturally wonder how many sweeps are needed to reach the equilibrium distribution. Even though a rigorous answer can be offered for some cases [11] the general consensus is that the discarded sweeps must be chosen approximately right. The measurement of the integrated autocorrelation time, to be introduced below, cannot always be achieved in practice. The inclusion of non-equilibrium configurations in the final data means that configurations with zero probability in the equilibrium distribution contribute in the expectation values. This contribution declines by $1/N$ as the amount of equilibrium sweeps

[11] James Propp and David Wilson. Coupling from the Past: a User's Guide, 1997

N increases and can also be overcome by the statistical errors which decline like $1/\sqrt{N}$. It is always important to exclude the equilibration part in order to conduct proper equilibrium statistics. Also generally there are other possible problems that can arise during the equilibration part, like the system reaching a metastable configuration. We note that equilibration becomes a more serious problem as the size of the system increases and when autocorrelation times are large.

Once the equilibration part is complete, the expectation value $\hat{f}$ of some observable quantity can be calculated by an amount of measurements from corresponding sweeps. It is important though to understand the autocorrelations inherent in a Markov Chain before conducting a proper measurement.

Defining as x_i configurations produced in equilibrium and assuming N measurements from a Markov Chain we have:

$$f_i = f_i(x_i), i = 1, \ldots, N \tag{2.35}$$

The Markov Chain is time-discrete and each time step between two measurements f_i, f_{i+1}, which corresponds to a sweep, resembles the same amount of time.

The estimator of the expectation value $\hat{f}$ is:

$$\bar{f} = \frac{1}{N} \sum f_i \tag{2.36}$$

The autocorrelation function of the observable quantity f is defined as:

$$\begin{aligned} \hat{C}(t) &= \hat{C}_{ij} \\ &= \langle (f_i - \langle f_i \rangle)(f_j - \langle f_j \rangle) \rangle \\ &= \langle f_i f_j \rangle - \langle f_i \rangle \langle f_j \rangle \\ &= \langle f_0 f_t \rangle - \hat{f}^2 \end{aligned} \tag{2.37}$$

where $t = |i - j|$ and the system is time-translation invariant. When $t \to \infty$:

$$\hat{C}(t) \sim exp\left(- \frac{t}{\tau_{exp}} \right) \tag{2.38}$$

The quantity τ_{exp} is defined as the *exponential autocorrelation time*. The eigenvalue $\lambda_0 = 1$ of the transition matrix has as an eigenvector the equilibrium distribution. The exponential autocorrelation time can be expressed in terms of the eigenvalue λ_1 if we assume that f has a projection which does not equal zero on the eigenstate. The exponential autocorrelation time is then:

$$\tau_{exp} = - \ln \lambda_1 \tag{2.39}$$

It is of importance to mention that the variance of f is related with the autocorrelations through the expression:

$$\hat{C}(0) = \sigma^2(f) \tag{2.40}$$

The variance of $\bar{f}$ for the autocorrelation functions and the mean is expressed by:

$$\begin{aligned}\sigma^2(\overline{f}) &= \langle(\overline{f}-\hat{f})^2\rangle \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\langle(f_i-\hat{f})(f_j-\hat{f})\rangle \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\langle f_i f_j - f_i\hat{f} - f_j\hat{f} + \hat{f}^2\rangle \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\Big[\langle f_i f_j\rangle - \hat{f}^2\Big] \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} C_{ij}\end{aligned} \tag{2.41}$$

The last sum can be rephrased by noticing that $|i-j|=0$ appears a total of N times and $|i-j|=t$ with $1 \leq t \leq (N-1)$ appears $2(N-t)$ times:

$$\sigma^2(\overline{f}) = \frac{1}{N^2}\left[N\hat{C}(0) + 2\sum_{t=1}^{N-1}(N-t)\hat{C}(t)\right] \tag{2.42}$$

$$\sigma^2(\overline{f}) = \frac{\sigma^2(f)}{N}\left[1 + 2\sum_{t=1}^{N-1}\Big(1-\frac{t}{N}\Big)\hat{c}(t)\right] \tag{2.43}$$

$$\hat{c}(t) = \frac{\hat{C}(t)}{\hat{C}(0)} \tag{2.44}$$

One can make a comparison then between the variance of the estimator f calculated above and the expression for the uncorrelated case:

$$\sigma^2_{\text{uncorrelated}}(\overline{f}) = \frac{\sigma^2(f)}{N} \tag{2.45}$$

The different term is defined as the *integrated autocorrelation time*:

$$\tau_{int} = \left[1 + 2\sum_{t=1}^{N-1}\left(1-\frac{t}{N}\right)\hat{c}(t)\right] \tag{2.46}$$

It is then straightforward to notice that the variance of the mean for correlated data is larger by a factor of t_{int} than the uncorrelated case:

$$\tau_{int} = \frac{\sigma^2(\overline{f})}{\sigma^2_{\text{uncorrelated}}(\overline{f})}. \tag{2.47}$$

In the thermodynamic limit $N \to \infty$ the equation 2.46 is:

$$\tau_{int} = 1 + 2\sum_{t=1}^{\infty}\hat{c}(t) \tag{2.48}$$

Estimating the integrated autocorrelation time is not an easy task. The estimator $\overline{\tau}_{int}$ in the thermodynamic limit is:

$$\overline{\tau}_{int} = 1 + 2\sum_{t=1}^{\infty} \overline{c}(t) \tag{2.49}$$

The variance of the above estimator then diverges:

$$\sigma^2(\overline{t}_{int}) \to \infty \tag{2.50}$$

If we assume an estimator that depends on t :

$$\overline{\tau}_{int}(t) = 1 + 2\sum_{t=1}^{\infty} \overline{c}(t') \tag{2.51}$$

we can make an estimate of the integrated autocorrelation time by searching for the best value where $\overline{\tau}_{int}(t)$ is independent of t . An alternative method follows below.

2.1.5 *Binning Analysis and Integrated Autocorrelation Time*

It is possible to acquire an estimation of the integrated autocorrelation time by using the expression 2.47. [12].

[12] H. Flyvbjerg and H. G. Petersen. Error estimates on averages of correlated data . *The Journal of Chemical Physics 91, 461*, 1989

Let us assume that we have separated the N time series measurements into N_{bs} bins where $N_{bs} \leq N$ and each N_{bs} consists of N_b measurements:

$$N_b = \frac{N}{N_{bs}} \tag{2.52}$$

The data which have been binned are the averages:

$$f_j^{N_b} = \frac{1}{N_b} \sum_{i=1+(j-1)N_b}^{jN_b} f_i, j = 1, \ldots, N_{bs} \tag{2.53}$$

An increase in the amount of measurements inside each bin would correspond to a decrease in the autocorrelations. Eventually the amount of measurements inside each bin would be greater than the exponential autocorrelation time τ_{exp} and only bins that are next to each other would be autocorrelated. An even greater increase in the size of each bin would lead to even further reductions for the autocorrelations.

Assuming now all N_{bs} bins we can calculate the mean using the N_b measurements inside each bin:

$$\overline{f_j}^{N_b} = \frac{1}{N_{bs}} \sum_{j=1}^{N_{bs}} f_j^{N_b} \tag{2.54}$$

We are now considering that we have *uncorrelated* data and the error bar is equal to the standard deviation:

$$\sigma = \sqrt{\frac{1}{N_{bs}-1}(\overline{f_j^2} - \overline{f_j}^2)} \tag{2.55}$$

The integrated autocorrelation time for the case $N_b \to \infty$ is:

$$\tau_{int} = \lim_{N_b \to \infty} \tau_{int}^{N_b} \tag{2.56}$$

where:

$$\tau_{int}^{N_b} = \left(\frac{s^2_{\overline{f}_{N_b}}}{s^2_{\overline{f}}} \right) \tag{2.57}$$

and the unbiased estimator of the variance is

$$(s_x^r)^2 = \frac{N}{N-1} (s_x^{'r})^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i^r - \overline{x}^r)^2 \tag{2.58}$$

From a practical point of view, a large value of N_b would be enough to assume $N_b \to \infty$. The $s^2_{\overline{f}_{N_b}}$ present in the numerator of $\tau_{int}^{N_b}$ have the main role in estimating the error of τ_{int} since for the case of $N_b \to \infty$ the values $s^2_{\overline{f}}$ will be much smaller. A finite size N_b of measurements that corresponds to practically uncorrelated data is a decent estimate for τ_{int}. Considering that using the central limit theorem the data resulting from binning can be approximated as Gaussian, the quantity $s^2_{\overline{f}_{N_b}}$ is known analytically. The N_{bs} bins chosen to be independent then correspond to the error. For the $s^2_{\overline{f}}$ in the denominator, the effective number of data that have no correlation can be used:

$$N_{\text{effective}} = \frac{N}{\tau_{int}} \tag{2.59}$$

The integrated autocorrelation time is important during the Markov Chain Monte Carlo simulation. Initially, a choice of sweeps that are far greater than τ_{int} must be executed during the equilibration part in order for the system to arrive at equilibrium. Additionally, knowledge of the integrated autocorrelation time allows the calculation the variance for correlated data:

$$\sigma^2(\overline{f}) = \tau_{int} \frac{\sigma^2(f)}{N} \tag{2.60}$$

Therefore, one can use this knowledge to calculate the errors:

$$\Delta \overline{f} = \sqrt{\sigma^2(\overline{f})} \tag{2.61}$$

It is not always easy to calculate the integrated autocorrelation time though, particularly in large scale simulations. Therefore, one might have to rely on using the binning method for error estimation by considering a constant amount of bins N_{bs}. Eventually, as the number of measurements increases the data that constitute different bins will become statistically independent and the error analysis will be reasonable.

Sometimes we might be able to derive an estimation about the magnitude of the τ_{int}. For the case of $V = L^d$ lattice near the critical region, the integrated

autocorrelation time increases as:

$$\tau_{cpu} = L^{d+z} \tag{2.62}$$

The quantity z is the dynamic critical exponent and it can estimated using finite size scaling extrapolations. Other error analysis techniques could be implemented, like the *jackknife* and *bootstrap* method but the binning analysis should suffice for the problems studied.

2.2.0 *Unsupervised Learning of the d=2 Ising Model*

The probability distribution represented by real-space configurations of the $d = 2$ Ising model can be modeled using our prior implementation of Restricted Boltzmann Machines.

Let us assume a data set $\{u_i\}$ of spin configurations produced by a Markov Chain Monte Carlo simulation of the $d = 2$ Ising model for a finite temperature using either the Metropolis or Wolff's cluster algorithm. This data set is described by an empirical probability distribution q and the neural network by the model distribution p. The main goal is the minimization of the Kullback-Leibler divergence between q and p so as they exactly match one another.

When the neural network training is complete it can be led into equilibrium by randomly initializing the visible units and executing alternate Gibbs sampling. It is then possible to sample approximate states of the $d = 2$ Ising model from the neural network's equilibrium distribution in order to calculate observable quantities. These approximate states, which are also called reconstructions, are treated exactly like the Markov Chain Monte Carlo data . The interest is on studying the critical region and observe the dependence between the number of hidden units and the accuracy of the expectation values calculated from the corresponding configurations.

The neural network is described by n_v visible units and n_h hidden units. The weights $\boldsymbol{w}$ are the variational parameters since we have assumed an extra node fixed to value one in order to use one update rule. For each temperature a neural network is trained on 100000 configurations for the cases of $n_h = 64, 16, 4$ units and 500 epochs. The system studied is the Ising model with a size $N = L * L = 8 * 8 = 64$ in a square lattice and nearest neighbor interactions. The weights are initialized as:

$$\boldsymbol{w} \propto \sqrt{\frac{1}{n_h + n_v}} \tag{2.63}$$

Contrastive divergence is carried out for a number of 20 steps and the training is on a mini batch of 50 samples. The learning rate is set to $l = 0.01$ and no weight decay or momentum term is included.

We notice from the calculation of observable quantities that the Restricted Boltzmann Machine can reproduce configurations that give more accurate results away from the phase transition. The observable quantities calculated near the

critical point are accurate for $n_h = 64$ number of hidden units but not for the other two cases $n_h = 16, 4$. For the internal energy, non accurate results are expected. The neural network is unaware of the constraint we have imposed in the calculation which is the nearest neighbor interaction. This information isn't encoded in the data set and a smaller number of hidden units cannot capture the correct dependencies in order to give accurate results. The magnetization is the opposite case. Since the training is conducted on spin configurations the magnetization is encoded in the data set. It is a summation over all spins without a dependence on some local constraint. The neural network is able to capture the correct behavior for all cases of hidden units. The results for the specific heat and the magnetic susceptibility show some inaccuracy for a smaller number of hidden units. This is due to the fact that critical fluctuations arise near the phase transition and a larger number of hidden units is required in order to model them . The accuracy of the results near the second-order phase transition is then clearly dependent on the number of hidden units.

The Restricted Boltzmann Machine has then the capacity to be used as a basic research tool since consistent results can be achieved across the ordered, disordered and critical regions. The unsupervised setting implies that it has to be used in conjunction with a set of Monte Carlo or Molecular Dynamics data. Additionally, it also offers the option to "compress" the system through reduction of the amount of hidden units. This approach is of high importance since the correspondence between the Renormalization Group and energy based Deep Learning is established in the next chapters for the case of unsupervised learning.

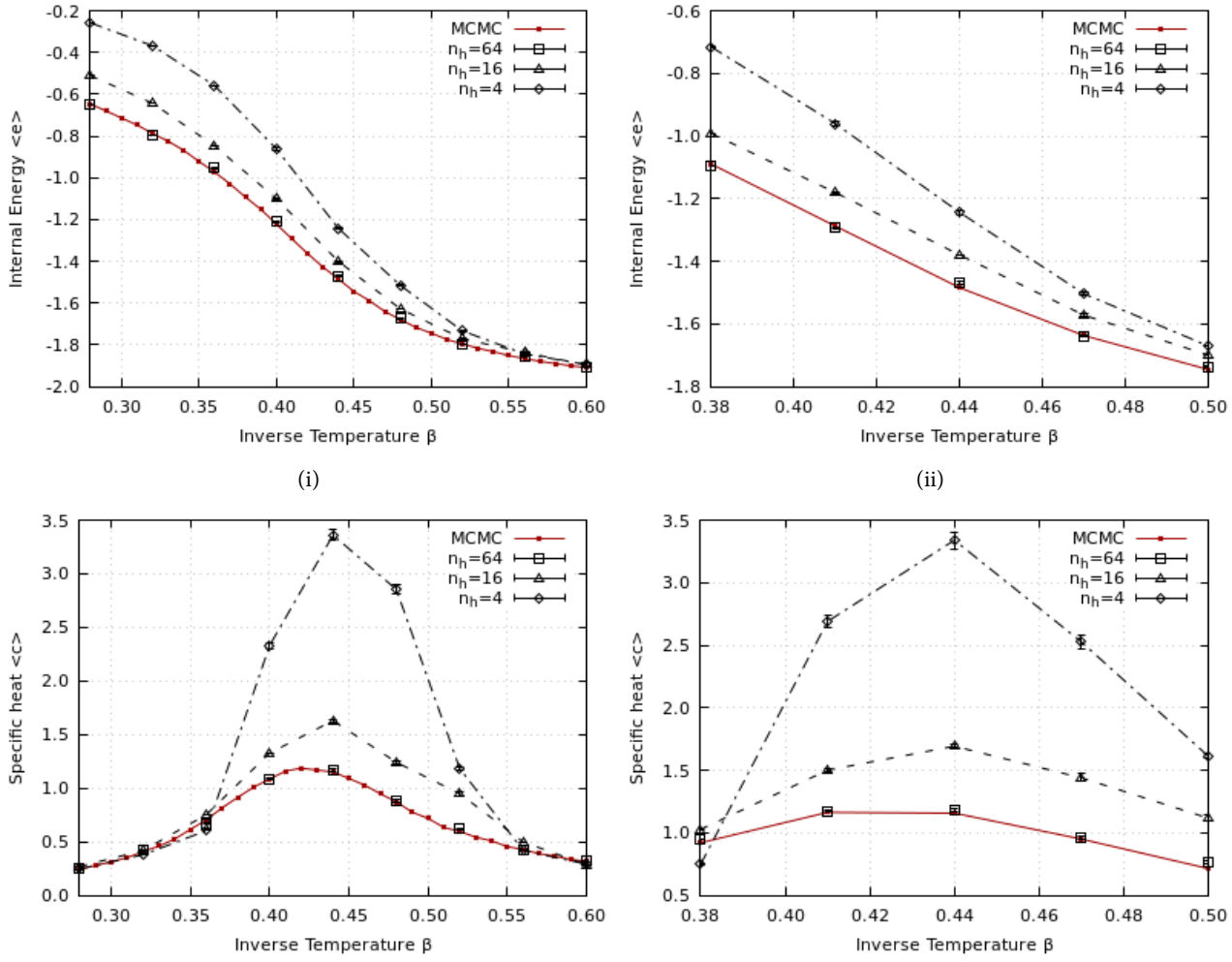


Figure 2.5: Figures of the internal energy and the specific heat per site for a lattice of site $N = L * L = 8 * 8 = 64$. The Restricted Boltzmann Machine has been trained on data from (i) the Metropolis algorithm and (ii) Wolff's cluster algorithm. It is evident that in general the neural network works best away from the critical temperature $\beta_c \approx 0.4407$ for different numbers n_h of hidden units. When hidden units are equal to visible units the expectation values compare well with the ones from Monte Carlo since the RBM is able to model the probability distribution better.

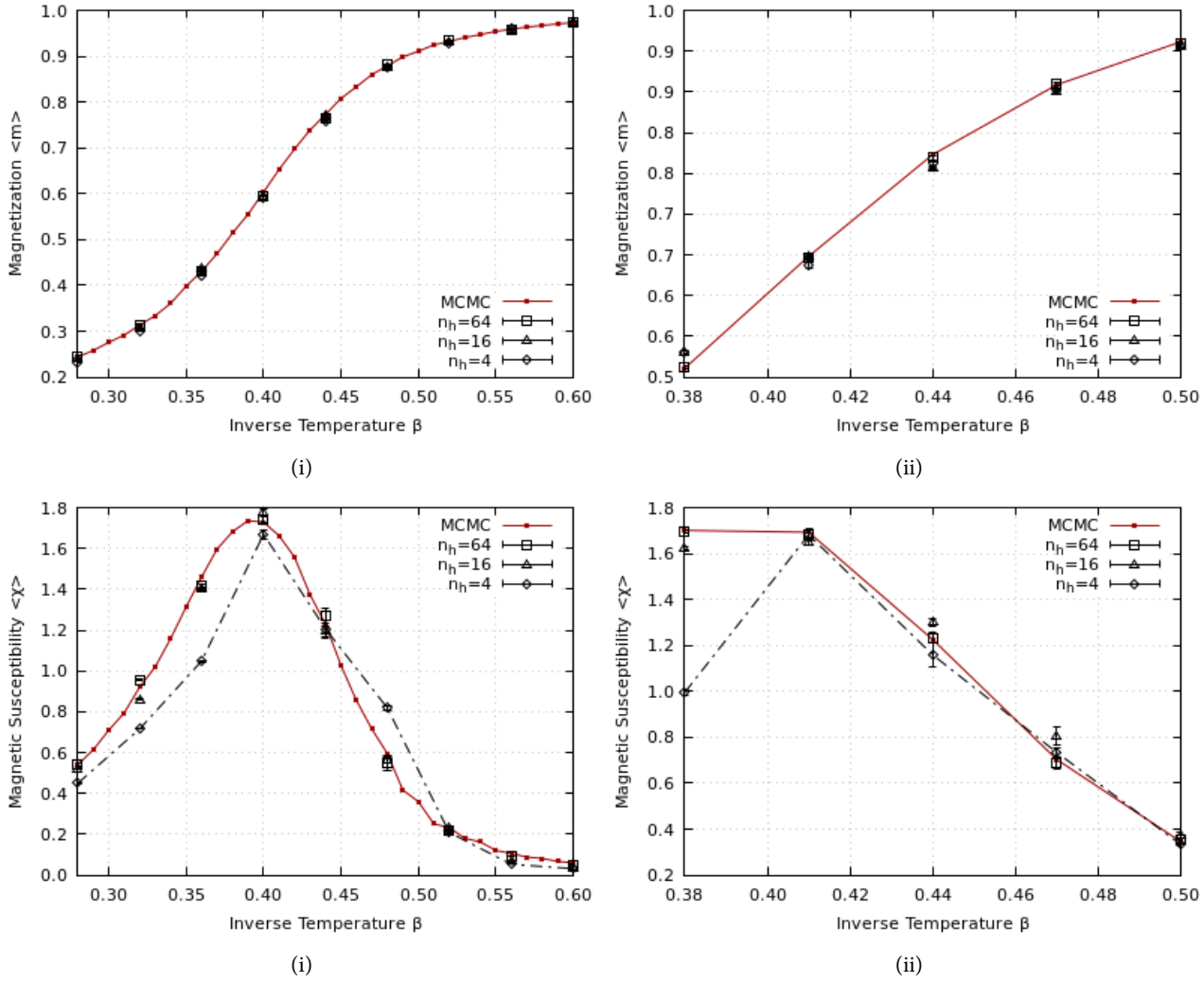


Figure 2.7: Figures of magnetization $< m >$ and magnetic susceptibility $< \chi >$ per site . The Restricted Boltzmann Machine has been trained on data from (i) the Metropolis algorithm and (ii) Wolff's cluster algorithm. We notice that for the case of magnetization the expectation values calculated from configurations produced by the RBM are accurate for all cases of hidden units.

3. *The Renormalization Group and Deep Belief Networks*

3.1.0 *Real-Space Renormalization Group*

The Renormalization Group[1] is an important technique in theoretical physics when confronting problems of many length scales. One can implement a renormalization group transformation and arrive at a coarse-grained description of a system. It is then possible to extract features that describe large scale interactions while integrating out degrees of freedom at short distances.

Among the Renormalization group techniques, the real-space renormalization [2] is an approximate method that replaces the physical spins with auxiliary ones, which we will also call *hidden*, through an iterative procedure. The goal is to ensure that the rescaled system preserves the information at large scale of the original system. This can be achieved by an appropriate choice of the parameters that couple the auxiliary spins with the ones on the original spin system. The choice is based on minimizing the difference of the free energies between the two systems. The technique can be performed again on the auxiliary spins, resulting in a new coarse-grained description of the system.

Let's consider an ensemble of N binary spins on a lattice, where each spin can have a possible value of ± 1 . Given a set of spins $\{u_i\}$ with Hamiltonian $H(\{u_i\})$, the probability of a spin configuration in thermal equilibrium is given by the Boltzmann probability distribution:

$$P(\{u_i\}) = \frac{e^{-H(\{u_i\})}}{Z} \tag{3.1}$$

where the inverse temperature β is set to one. The partition function Z is then given by:

$$Z = Tr_{\{u_i\}} e^{-H(\{u_i\})} = \sum_{\{u_i\}} e^{-H(\{u_i\})} \tag{3.2}$$

and the free energy of the system:

$$F^u = -\log Z = -\log \left(Tr_{\{u_i\}} e^{-H(\{u_i\})} \right) \tag{3.3}$$

A general Hamiltonian of the model would depend on a set of coupling constants $\boldsymbol{K} = \{K_s\}$ that describe interactions between degrees of freedom of

[1] Kenneth G.Wilson and J.Kogut. The renormalization group and the Ïţ expansion. *Physics Reports, Volume 12, Issue 2*, 1974; and Kenneth G.Wilson. The renormalization group and critical phenomena. *Rev. Mod. Phys. 55, 583*, 1983

[2] John Cardy. *Scaling and Renormalization in Statistical Physics*. Cambridge University Press, 1996; Leo P. Kadanoff. *Statics, Dynamics and Renormalization* . World Scientific, 2000; Nigel Goldenfeld. *Lectures On Phase Transitions And The Renormalization Group (Frontiers in Physics)* . Addison-Wesley, 1992; Leo P. Kadanoff, Anthony Houghton, and Mehmet C. Yalabik. Variational approximations for renormalization group transformations. *J. Stat. Phys. 14: 171*, 1976; and Efi Efrati, Zhe Wang, Amy Kolan, and Leo P. Kadanoff. Real-space renormalization in statistical mechanics. *Rev. Mod. Phys. 86, 647*, 2014

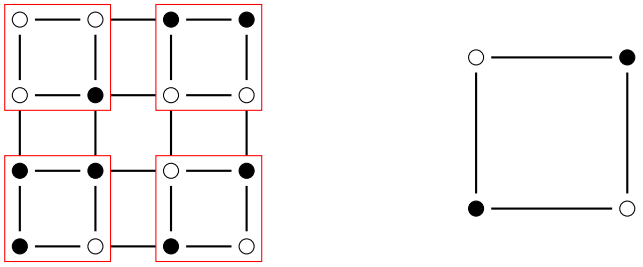


Figure 3.1: A $N = 4 * 4 = 16$ spin system rescaled to a $N' = 2 * 2$ spin system using the blocking procedure with a rescaling factor of $b = 2$. The final value of the h_j coarse-grained spin is decided based on the majority of values of the included spins. When they are equal, the choice is made randomly.

various orders:

$$H[\{u_i\}] = -\sum_i K_i u_i - \sum_{i,j} K_{ij} u_i u_j - \sum_{i,j,k} K_{ijk} u_i u_j u_k + \dots \tag{3.4}$$

One can now introduce a new set $\{h_j\}$ of M hidden spins, with $M < N$, and we will assume that the blocking procedure has been chosen to coarse-grain the original system. The spins can now been grouped into blocks described by a rescaling factor b, reducing the size of the original lattice by b in each direction. Each block corresponds to a hidden spin $\{h_j\}$, whose value can be ± 1 and is chosen depending on the majority of the values of the coarse-grained spins inside each block. A rescaling factor of 2 would reduce the amount of spins by 2^d, where d is the dimensionality of the system.

The interactions between the new coarse-grained variables $\{h_j\}$ have a dependence on the original interactions of the $\{u_i\}$ spins and can be described by a new Hamiltonian with a different set of $\{K^{'}\}$ coupling constants that have been mapped as $\{K\} \to \{K^{'}\}$:

$$H^{RG}(\{h_j\}) = -\sum_i K_i^{'} h_i - \sum_{i,j} K_{ij}^{'} h_i h_j - \sum_{i,j,k} K_{ijk}^{'} h_i h_j h_k + \dots \tag{3.5}$$

With a Renormalization Group transformation, one can integrate out the original spins $\{u_i\}$, resulting in a complete description of the system through the coarse-grained spins $\{h_j\}$. A function $T_\lambda(\{u_i\}, \{h_j\})$, based on a set of parameters $\{\lambda\}$ can then be defined and it expresses interactions between the original spins and the ones of the rescaled system. It also defines a Hamiltonian for the $\{h_j\}$ through:

$$e^{-H_\lambda^{RG}(\{h_j\})} = Tr_{\{u_i\}} e^{T_\lambda(\{u_i\},\{h_j\}) - H(\{u_i\})} \tag{3.6}$$

Similarly a free energy is defined by:

$$F_\lambda^h = -\log\left(Tr_{\{h_j\}} e^{-H_\lambda^{RG}(\{h_j\})}\right) \tag{3.7}$$

As mentioned before, the set of variational parameters $\{\lambda\}$ must be chosen so as to minimize the energy difference of the two systems, $\Delta F = F_h^\lambda - F^u$.

In this way, we can guarantee that the rescaled system preserves the large scale information of the original system. Notice that:

$$\Delta F = 0 \Longleftrightarrow Tr_{\{h_j\}} e^{T_\lambda(\{u_i\},\{h_j\}} = 1 \tag{3.8}$$

A renormalization group transformation is called exact when:

$$Tr_{\{h_j\}} e^{T_\lambda(\{u_i\},\{h_j\})} = 1 \tag{3.9}$$

In the following section we will see how to map real-space renormalization group with Restricted Boltzmann Machine compression.

3.2.0 *A Correspondence between the Renormalization Group and Energy-Based Deep Learning*

In order to map the renormalization group with Boltzmann-Based deep neural networks we must make an appropriate choice for the operator $T_\lambda(\{u_i\},\{h_j\})$.

The operator $T_\lambda(\{u_i\},\{h_j\})$ describes interactions between spins in the original and rescaled system. The energy function $E(\{u_i\},\{h_j\})$ defined in 1.32 has the same role in a Restricted Boltzmann Machine. The variational operator $T_\lambda(\{u_i\},\{h_j\})$ must be chosen as:

$$T(\{u_i\},\{h_j\}) = -E(\{u_i\},\{h_j\}) + H[\{u_i\}] \tag{3.10}$$

Using the joint probability $p_\lambda(\{u_i\},\{h_j\})$ of the Restricted Boltzmann Machine defined in 1.31, we can acquire expressions of Hamiltonians for the visible and hidden units of the RBM through the marginal distributions:

$$p_\lambda(\{u_i\}) = \sum_{\{h_j\}} p_\lambda(\{u_i\},\{h_j\}) = Tr_{h_j} p_\lambda(\{u_i\},\{h_j\}) = \frac{e^{-H_\lambda^{RBM}[\{u_i\}]}}{\mathcal{Z}} \tag{3.11}$$

$$p_\lambda(\{h_j\}) = \sum_{\{u_j\}} p_\lambda(\{u_i\},\{h_j\}) = Tr_{u_i} p_\lambda(\{u_i\},\{h_j\}) = \frac{e^{-H_\lambda^{RBM}[\{h_j\}]}}{\mathcal{Z}}, \tag{3.12}$$

As mentioned before, $T_\lambda(\{u_i\},\{h_j\})$ defines a Hamiltonian for the auxiliary spins through the expression 3.6. Dividing both sides of 3.6 with the partition function $\mathcal{Z}$ of the RBM we get:

$$\frac{e^{-H_\lambda^{RG}(\{h_j\})}}{\mathcal{Z}} = \frac{Tr_{\{u_i\}} e^{T_\lambda(\{u_i\},\{h_j\})-H(\{u_i\})}}{\mathcal{Z}} \tag{3.13}$$

Substituting in the above equation the expression for the variational operator 3.10 we acquire:

$$\frac{e^{-H_\lambda^{RG}(\{h_j\})}}{\mathcal{Z}} = Tr_{\{u_i\}} \frac{e^{-E(\{u_i\},\{h_j\})}}{\mathcal{Z}} = p_\lambda(\{h_j\}) \tag{3.14}$$

Using the RBM Hamiltonian for the hidden units then gives:

$$\frac{e^{-H_\lambda^{RG}(\{h_j\})}}{\mathcal{Z}} = \frac{e^{-H_\lambda^{RBM}[\{h_j\}]}}{\mathcal{Z}} \Rightarrow H_\lambda^{RG}[\{h_j\}] = H_\lambda^{RBM}[\{h_j\}] \tag{3.15}$$

The above result leads to an equality for the Hamiltonian of the rescaled system and the Hamiltonian for the hidden spins of the Restricted Boltzmann Machine . Therefore, the same Hamiltonian describes both. Equally, one can state that the marginal distribution $p_\lambda(\{h_j\})$ of the hidden spins on the RBM is a Boltzmann probability distribution with a Hamiltonian $H_\lambda^{RG}[\{h_j\}]$. The operator $T_\lambda(\{u_i\},\{h_j\})$ approximates the conditional probability of the hidden spins given the visible spins:

$$\begin{aligned} e^{T(\{u_i\},\{h_j\})} &= e^{-E(\{u_i\},\{h_j\})+H[\{u_i\}]} \\ &= e^{-E(\{u_i\},\{h_j\})} e^{H[\{u_i\}]} \frac{e^{-H_\lambda^{RBM}[\{u_i\}]}}{e^{-H_\lambda^{RBM}[\{u_i\}]}} \\ &= \frac{p_\lambda(\{u_i\},\{h_j\})}{p_\lambda(\{u_i\})} e^{H[\{u_i\}]-H_\lambda^{RBM}[\{u_i\}]} \\ &= p_\lambda(\{h_j\}|\{u_i\}) e^{H[\{u_i\}]-H_\lambda^{RBM}[\{u_i\}]} \end{aligned} \tag{3.16}$$

When the condition of an exact Renormalization Group transformation 3.9 is satisfied, the Hamiltonian of the original system is equal to the Hamiltonian of the Restricted Boltzmann Machine $H[\{u_i\}] = H_\lambda^{RBM}[\{u_i\}]$:

$$\begin{aligned} Tr_{h_j} e^{T(\{u_i\},\{h_j\})} &= Tr_{h_j} \frac{p_\lambda(\{u_i\},\{h_j\})}{p_\lambda(\{u_i\})} e^{H[\{u_i\}]-H_\lambda^{RBM}[\{u_i\}]} \\ &= \frac{p_\lambda(\{u_i\})}{p_\lambda(\{u_i\})} e^{H[\{u_i\}]-H_\lambda^{RBM}[\{u_i\}]} \\ &= e^{H[\{u_i\}]-H_\lambda^{RBM}[\{u_i\}]} \\ &= 1 \\ &\Rightarrow H[\{u_i\}] = H_\lambda^{RBM}[\{u_i\}] \end{aligned} \tag{3.17}$$

Additionally we acquire an expression for the operator $T(\{u_i\},\{h_j\})$ and the exact conditional probability since $H[\{u_i\}] - H_\lambda^{RBM}[\{u_i\}] = 0$. The variational distribution $p_\lambda(u_i)$ can then reproduce completely the distribution encoded in the data $P(\{u_i\})$ which implies that the Kullback-Leibler divergence is equal to zero $D_{KL}(P(\{u_i\})|p_\lambda(\{u_i\})) = 0$.

The above approach is established on minimizing the difference of free energies through exact transformations and on describing the systems in terms of Hamiltonians. On the contrary, approaching a Machine Learning problem usually entails approximations in order to minimize the Kullback-Leibler divergence. These approximations result in a different procedure that coarse-grains the system. Finally, it is important to note that the explicit form of the joint energy function $E(\{u_i\},\{h_j\})$ does not alter the results and any variation of Boltzmann Machines can be used.

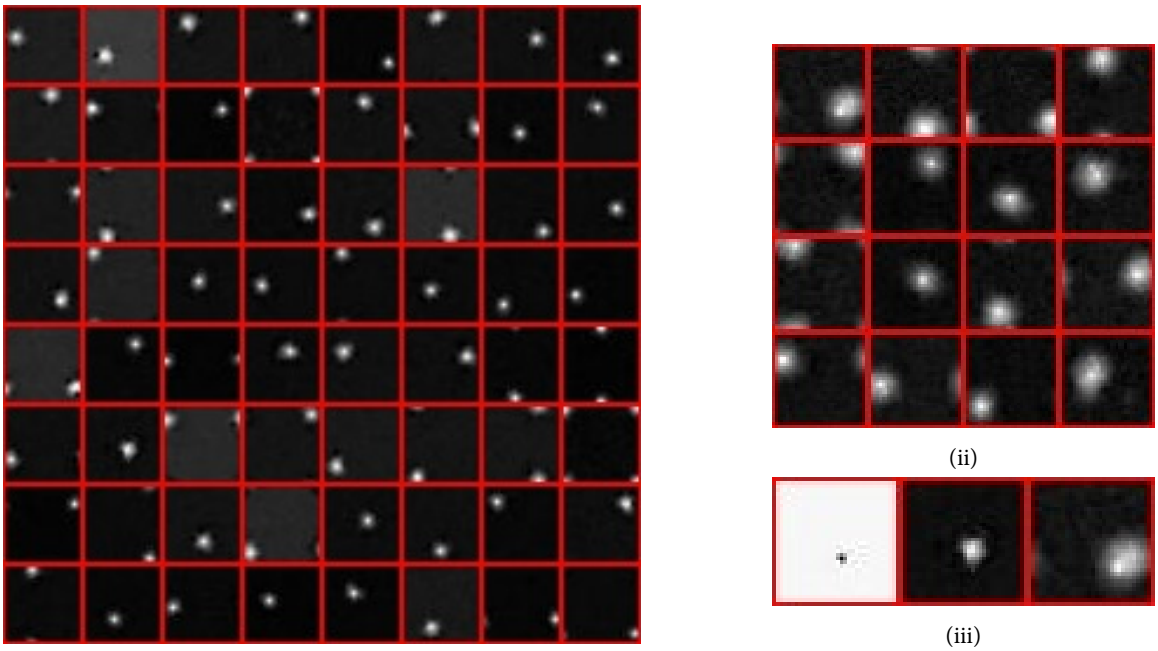
We now implement a Deep Belief Network with one visible layer of size $n_v = 1024$ and three hidden layers of size $n_h = 256, 64, 16$ that is trained on importance sampled configurations consisting of 40000 sweeps for temperature $\beta = 0.44$ of the $d = 2$ Ising model. The neural network is trained for 400 epochs with learning rate $l = 0.1$, mini batch size 100 and momentum 0.5. A L1 weight decay term is used by subtracting the sign values of the weight matrix multiplied by 0.002 .

To gain further insights between the established mapping we can visualize the *receptive fields* of the deep belief network through the recursion relation:

$$r^l = r^{(l-1)} W^l, l > 1 \tag{3.18}$$

where $r^1 = W^1$. Therefore, we acquire a measure of how a unit in a given hidden layer influences units in the visible layer.

It is evident from the figures, that a similar procedure to spin blocking is implemented by the neural network. Every unit in a given hidden layer couples to a spin block of approximately the same size in the visible layer. The size of the blocks increases with the rate of compression which is the same idea for an increasing rescaling factor in the renormalization group. The important difference is that the neural network applies this procedure in an *autonomous* way during training.



(i) (ii) (iii)

Figure 3.3: Visualization of the receptive fields for the second (i) and third (ii) hidden layer. A visualization of a representative receptive field for all three (iii) hidden layers is included. The size of the blocks/clusters increases with the rate of compression. This is the same idea with iterative applications of a spin-blocking renormalization group transformation.

3.3.0 *A Spin Blocking Transformation for the Ising Model in $d = 2$*

The main problem with Real-Space Renormalization Group is the assumption that the configurations of the rescaled system appear with their correct Boltzmann probabilities.

Let us assume a lattice of dimension L with Monte Carlo importance-sampled configurations for a given temperature T. The rescaling of the system by a factor of b will result in a new lattice of size L' with:

$$L' = \frac{L}{b} \tag{3.19}$$

The original configurations obviously appear with their correct Boltzmann probabilities since they have been sampled from an equilibrium distribution. But we cannot claim that the states corresponding to the new lattice L' appear with *their* correct Boltzmann probabilities if the Hamiltonian of the original system is used. It is the assumption that they do which introduces errors in the use of this method.

Since a renormalization group transformation must preserve the large-scale features of the system, the correlation length ξ is presumed to remain approximately same. But the reduction of the amount of spins by b in each dimension implies that the correlation length of the rescaled system must equal:

$$\xi' = \frac{\xi}{b} \tag{3.20}$$

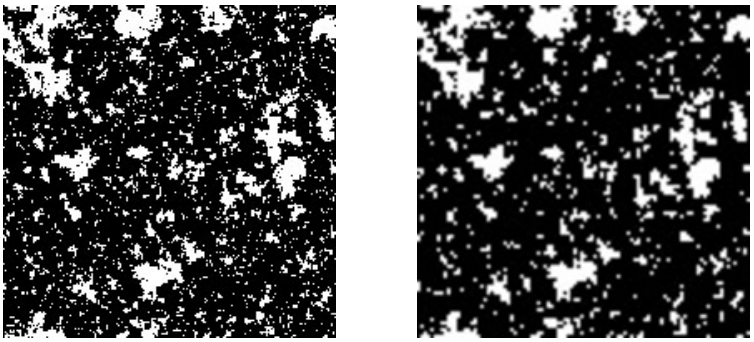
For the case of a $b = 2$ rescaling factor the above equation indicates that the correlation length should be $\xi' = \xi/2$. Thus the configurations of the rescaled system must correspond to states of a different temperature T' since the correlation length changes for different values of T.

Similarly, observable quantities that have a dependence on temperature like the internal energy per spin u should give different values when calculated for the original and rescaled system. Since the configurations of the rescaled system correspond to states sampled at a temperature T' then the internal energy per spin u' should be the correct value for the corresponding system.

At the critical temperature though the correlation lengths of the original and rescaled system are the same:

$$\xi = \xi' \; and \; T = T' = T_c \tag{3.21}$$

All other intensive quantities like the internal energy per spin are also equal. With the use of re-weighting techniques it is possible to calculate the above quantities by extrapolating in a range of temperatures for both the original and the rescaled system. It is of importance to use the values of the rescaled system as observable quantities of the *original* system when using re-weighting. If there is no prior knowledge of the critical point, the method can be used iteratively on temperatures calculated at every step as "critical" until it converges.



(i) (ii)

Figure 3.4: A spin blocking renormalization group transformation with rescaling factor $b = 2$ on a lattice of size $N = 200 * 200$ using the majority rule. The rescaled system (ii) of size $N = 100 * 100$ which has been resized for easier comparison captures the large-scale characteristics of the original system (i).

Finite size effects do generally introduce errors in the calculations and larger system sizes allow better calculations. Additionally they also introduce errors because of the difference in the size of the original and rescaled system and a system sharing the same size with the rescaled one might have to be simulated separately in order to acquire results. Depending on the model studied the assumption that the configurations of the rescaled system correspond to states of some temperature T' can be a major source of problems.

Our primary concern of applying renormalization group transformations in the Ising model is the calculation of critical exponents. We can map the temperatures T' of the rescaled system and T of the original system. Denoting u and u' the internal energies per spin of the two systems we have:

$$u'(T) = u(T') \tag{3.22}$$

A mapping between the two temperatures T' and T is established:

$$T' = u^{-1}(u'(T)) \tag{3.23}$$

It is precisely this mapping that allows the calculation of critical exponents. The critical temperature defines a fixed point of the transformation which is called the *critical fixed point*. For temperatures $T > T_c$ the rescaled temperature is greater than the temperature of the original system, $T' > T$. Similarly $T' < T$ for $T < T_c$. A renormalization group transformation is then characterized by a flow through the parameter space which for this case is one-dimensional since the temperature is the only parameter that induces a phase transition. The flow leads the rescaled temperature away from the fixed point.

This behavior is expected. For example, the presence of clusters with size ξ in temperatures greater than T_c implies that the renormalization group transformation will create clusters of smaller size with $\xi' = \xi/b$. The rescaled configurations then will be typical of a higher temperature $T' > T$. For the case of renormalization group transformations with $T < T_c$ a configuration of the system is described by spins that mostly point to a specific direction. Every subsequent spin blocking transformation with the majority rule will in a sense "omit" the spins that point in the opposite direction as they are a minority. The resultant

configurations will then have even more spins pointing in the same direction as the original system. These configurations then correspond to systems with lower temperatures $T' < T$.

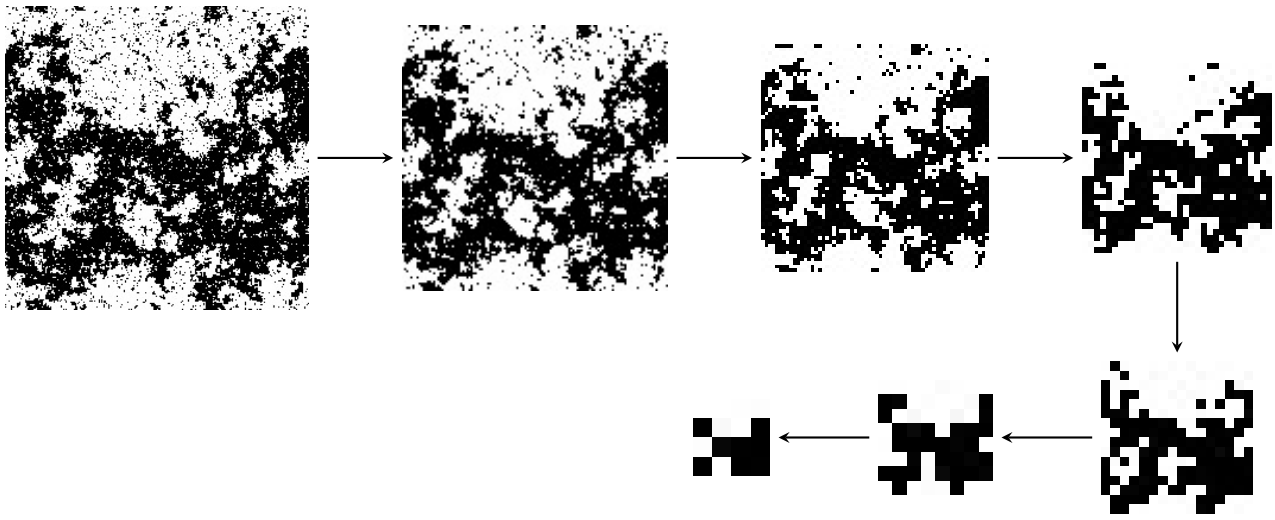


Figure 3.5: Consecutive spin-blocking Renormalization Group Transformations using a rescaling factor of $b = 2$ for an Ising Model of size $N = 256 * 256$ at the critical temperature $\beta_c = 0.4407$. The system remains at the same temperature despite the renormalization group transformations. One spin at the last system represents 4096 spins of the original system.

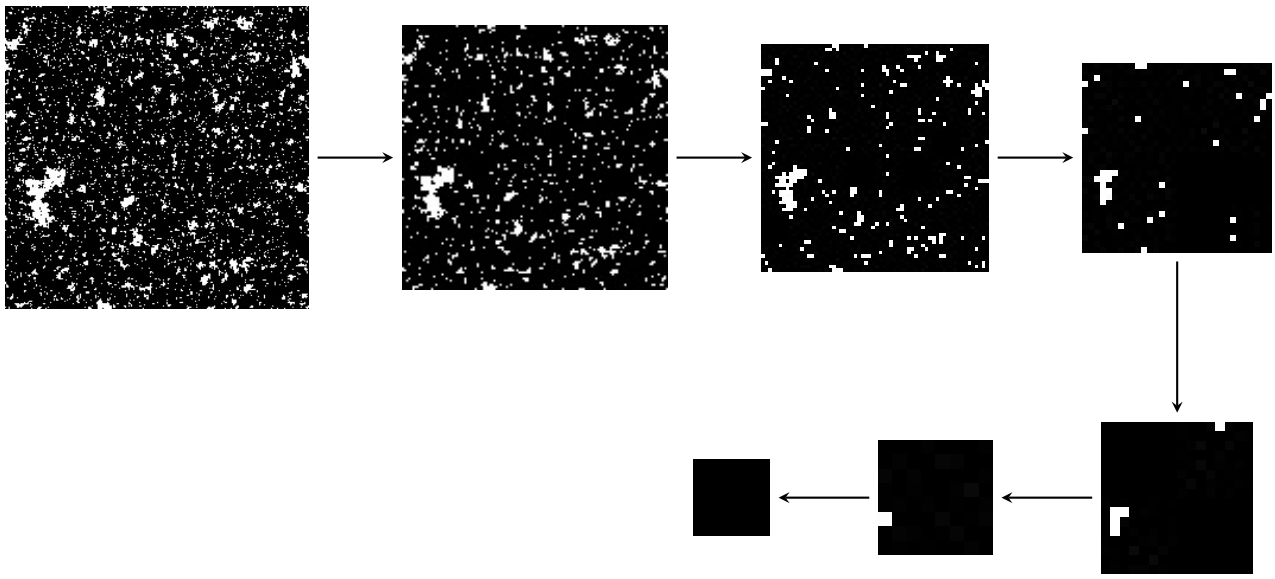


Figure 3.6: The same system as above but for temperature $\beta = 0.45$. Every spin blocking transformation leads the system to a higher inverse temperature and therefore to more ordered configurations.

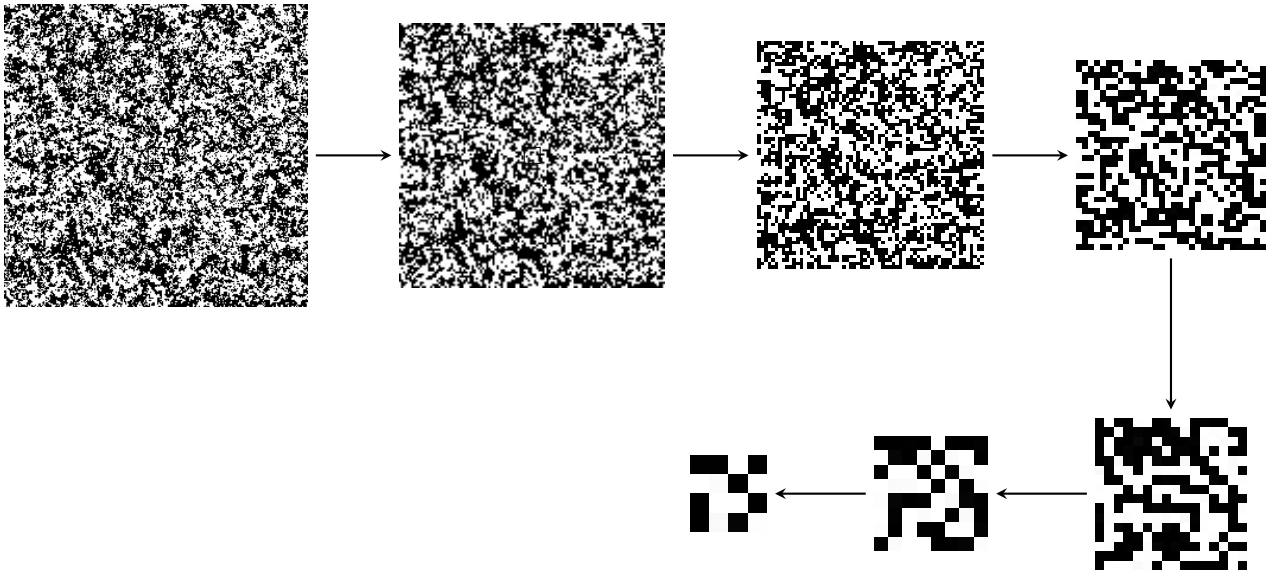


Figure 3.7: The same system as above for temperature $\beta = 0.36$. Every spin blocking transformation leads the system to lower inverse temperatures and therefore towards complete disorder.

3.3.1 *Estimating Onsager's Critical Exponents*

Now, let us consider again the exponent ν:

$$\xi \sim |t|^{-\nu} \tag{3.24}$$

The variable t is the reduced temperature:

$$t = \frac{T - T_c}{T_c} \tag{3.25}$$

The correlation length ξ' of the rescaled system is also expressed by the same equation except we must now consider the temperature T':

$$\xi' \sim |t'|^{-\nu} \tag{3.26}$$

Dividing 3.24 by 3.26 and using $\xi' = \xi/b$, we have:

$$\left(\frac{t}{t'}\right) = b \tag{3.27}$$

The expressions for the critical exponents only have meaning in a region of temperatures close to the critical point. Therefore, we need to acquire a relationship between T' and T near T_c and this is possible by linearizing the renormalization group transformation with a Taylor series expansion to leading order around T_c:

$$T' - T_c = (T - T_c)\frac{dT'}{dT}\bigg|_{T_c} \tag{3.28}$$

Using 3.25 and substituting the above into 3.27 we have an expression for the critical exponent ν:

$$\nu = \frac{\log b}{\log \frac{dT'}{dT}\big|_{T_c}} \tag{3.29}$$

Even though there are *scaling relations* which relate critical exponents with one another, there are cases for which Monte Carlo calculations might be performed in order to actually test these scaling relations. Therefore, it is important to have expressions that allow us to measure these exponents.

For the case of magnetization per spin the critical exponent β is :

$$m \sim |t|^{\beta} \tag{3.30}$$

With the use of equation 3.24:

$$m \sim \xi^{-\beta/\nu} \tag{3.31}$$

Now, considering a renormalization group transformation, the rescaled system will have a magnetization m' for which:

$$m' \sim \xi'^{-\beta/\nu} \tag{3.32}$$

Following the same procedure as above and dividing the magnetizations of the original and the rescaled system while using the expression that relates the correlation lengths, we have:

$$\frac{m'}{m} = b^{\beta/\nu} \tag{3.33}$$

$$\frac{\beta}{\nu} = \frac{\log \frac{m'}{m}}{\log b} \tag{3.34}$$

The validity of the expression 3.30 holds for an infinite system. With the use of L' Hôpital's rule we acquire the limiting value of m'/m :

$$\frac{m'}{m} = \frac{dm'/dT}{dm/dT} = \frac{dm'}{dm} \tag{3.35}$$

The resulting expression for the critical exponent β is:

$$\frac{\beta}{\nu} = \frac{\log \frac{dm'}{dm}\big|_{T_c}}{\log b} \tag{3.36}$$

which is superior to 3.34 since dm'/dm does not fluctuate much with different system sizes.

Equivalent equations can be derived with the same procedure as above for the critical exponents α and γ giving as result:

$$\frac{\alpha}{\nu} = -\frac{\log \frac{dc'}{dc}\big|_{T_c}}{\log b} \tag{3.37}$$

$$\frac{\gamma}{\nu} = -\frac{\log \frac{d\chi'}{d\chi}\big|_{T_c}}{\log b} \tag{3.38}$$

The magnetization of the model also depends on the critical exponent δ for an external field B:

$$m \sim B^{1/\delta} \tag{3.39}$$

We now define a critical exponent θ that expresses the way the correlation length diverges as $B \to 0$ at T_c. We then acquire, following the same ideas for 3.29 and 3.36:

$$\xi \sim |B|^{-\theta} \tag{3.40}$$

$$\theta = \frac{\log b}{\log \frac{dB'}{dB}\Big|_{B=0}} \tag{3.41}$$

$$\frac{1}{\theta\delta} = \frac{\log \frac{dm'}{dm}\Big|_{B=0}}{\log b} \tag{3.42}$$

The equations above give:

$$\delta = \frac{\log \frac{dB'}{dB}\Big|_{B=0}}{\log \frac{dm'}{dm}\Big|_{B=0}} \tag{3.43}$$

The exponents ν and θ which describe the divergence of the correlation length ξ in terms of critical temperature and applied external field respectively are associated with the relevant operators of the renormalization group transformation and the rest of the critical exponents can be calculated from them using scaling relations.

A system of size $N = 64 * 64$ is studied using a spin blocking renormalization group transformation of rescaling factor $b = 2$. Re-weighting has been used to calculate the observable quantities for a large range of temperatures. Some lines do not cross due to finite size effects. Using the magnetization the critical temperature is estimated to be $T_c = 2.26821 \Rightarrow \beta_c = 0.4409$. The critical exponents are calculated as $\alpha = -0.19$, $\beta = 0.101$, $\gamma = 1.744$, $\nu = 1.01$. Errors cannot be included due to the assumptions discussed previously. The technique is superior to finite size scaling extrapolations since it gives accurate results for dramatically smaller sizes of lattices.

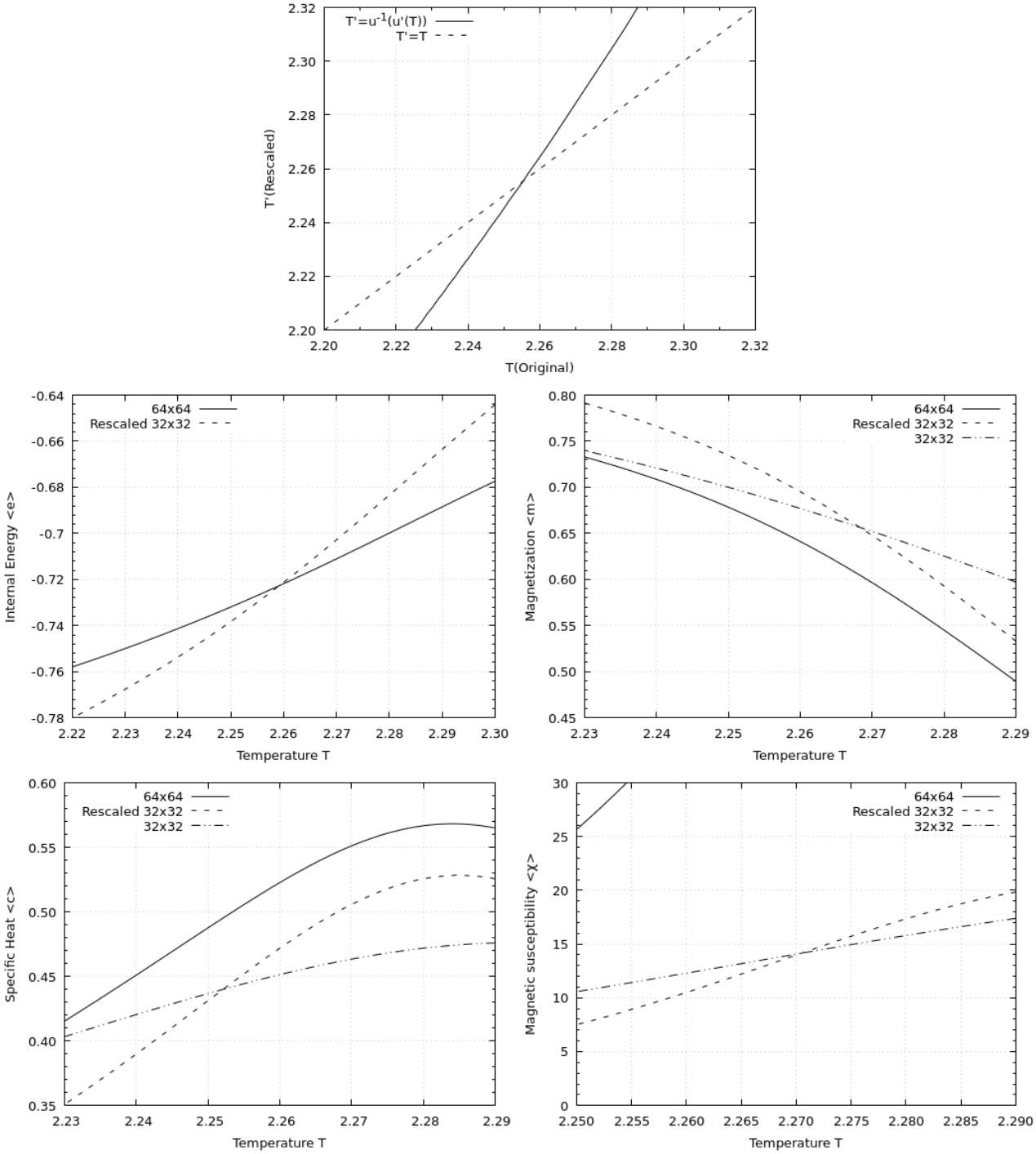


Figure 3.9: Figures of the observable quantities for the original system of $N = 64 * 64$, the rescaled system of $N = 32 * 32$ and a separate system of $N = 32 * 32$. A figure of the rescaled temperature T' as function of T is also included for the critical fixed point of the renormalization group transformation.

4. Reinforcement Learning in Many-Body Physics

4.1.0 The Transverse Field Ising Model in $d = 1$.

The transverse field Ising Model is the simplest system to exhibit a quantum phase transition on zero temperature through a variation of the external field. It is possible to map the quantum phase transition in a given dimension d with the phase transition induced by thermal fluctuations for the $d + z$ Ising model, where z is the dynamical exponent. Therefore both transitions belong in the same universality class.

The thermal fluctuations that gave rise to the second order phase transition differ from the quantum fluctuations present in the transverse field Ising model studied in this chapter. The quantum fluctuations also induce a phase transition that leads the system from order to disorder. The model was introduced by de Gennes [1] in order to study the ferroelectric phase of KDP. The Hamiltonian of the model is given by:

[1] P.G.de Gennes. Collective motions of hydrogen bonds. *Solid State Communications, Volume 1, Issue 6*, 1963

$$H = -J \sum_{\langle i,j \rangle} S_i^z S_j^z - h \sum_i S_i^x \tag{4.1}$$

where S^a are the Pauli operators, h is the transverse field that determines the tunneling term and J is the coupling constant of the cooperative term that is taken greater than zero, defining a ferromagnetic system.

It is possible to carry out an exact diagonalization for the model in $d = 1$ in order to compare the exact values with the results from the neural network reinforcement learning approach. The model has a vast amount of experimental applications and the interested reader is referred to [2].

[2] Sei Suzuki, Jun ichi Inoue, and Bikas K. Chakrabarti. *Quantum Ising Phases and Transitions in Transverse Ising Models*. Cambridge University Press, 2013

4.2.0 The Variational Principle

The Variational Monte Carlo method [3] will allow us to solve the problem of approximating the ground state energy of a system through an optimization approach. It is important then to set the necessary foundations by first establishing the variational principle.

[3] W. L. McMillan. Ground State of Liquid He^4. *Phys. Rev. 138, A442*, 1965; and Federico Becca and Sandro Sorella. *Quantum Monte Carlo Approaches for Correlated Systems*. Cambridge University Press, 2017

Let us consider a system described by a Hamiltonian H for which we cannot solve the time-independent Schrödinger equation. The system consists of discrete energies and the goal is to acquire reliable approximations of the

ground-state energy E_{gs}:

$$E_{gs} = E_1 \leq E_2 \leq \dots \tag{4.2}$$

Assuming *normalized* variational wave functions ψ_{var} which do not necessarily need to be energy eigenstates and considering that the eigenfunctions of H form a complete set we can expand in terms:

$$\psi_{var} = \sum_n c_n \psi_n, \; with \; H\psi_n = E_n \psi_n \tag{4.3}$$

$$\sum_n |c_n|^2 = 1 \tag{4.4}$$

The expectation value $\langle H \rangle_{var}$ is then given by:

$$\langle H \rangle_{var} = \sum_n E_n |c_n|^2 \tag{4.5}$$

The system consists of corresponding discrete energies E_n and therefore the ground state energy defines a lower bound:

$$\langle H \rangle_{var} = \sum_n E_n |c_n|^2 \geq \sum_n E_1 |c_n|^2 = E_1 \sum_n |c_n|^2 = E_1 = E_{gs} \tag{4.6}$$

Considering the case where the variational wave function ψ_{var} was not normalized, we can acquire a general expression by normalizing it :

$$E_{gs} \leq \langle H \rangle_{var} = \frac{\langle \psi_{var} | H | \psi_{var} \rangle}{\langle \psi_{var} | \psi_{var} \rangle} \equiv \mathcal{F}[\psi_{var}] \tag{4.7}$$

Assuming a given *trial* wave function the expectation value of the variational energy $\langle H \rangle_{var}$ will naturally be an overestimation of the ground state energy E_{gs}. Additionally, $\langle H \rangle_{var}$ will always set an upper bound for the exact value of E_{gs}. The trial wave functions $\psi_{var}(x; p_1, p_2 \dots p_n)$ also have a dependence on a set of variational parameters $\{p\}$. By using an optimization approach the parameters $\{p\}$ can be adjusted accordingly in order to acquire better approximations of the ground state energy. In an equal description, $\mathcal{F}[\psi_{var}]$ is a machine applying a set of operations and providing a resulting numerical output.

4.3.0 *The Variational Monte Carlo Method and the Zero-Variance Property*

One can use the Variational Monte Carlo method in order to calculate quantum expectation values as statistical averages and acquire a better understanding of a system's low energy behavior. The main problem is the necessary establishment of an *ansatz* [4] that might introduce a relevant bias.

[4] An ansatz is a set of mathematical tools that will have to be confirmed practically.

Let us assume that the Hilbert space is spanned by a complete basis set $\{|x\rangle\}$, and the completeness relation holds:

$$\sum_x |x\rangle \langle x| = \mathcal{I} \tag{4.8}$$

For a variational wave function ψ_{var} an operator $\mathcal{O}$ has a quantum expectation value:

$$\langle \mathcal{O} \rangle = \frac{\langle \psi_{var} | \mathcal{O} | \psi_{var} \rangle}{\langle \psi_{var} | \psi_{var} \rangle} = \frac{\sum_x \langle \psi_{var} | x \rangle \langle x | \mathcal{O} | \psi_{var} \rangle}{\sum_x \langle \psi_{var} | x \rangle \langle x | \psi_{var} \rangle} \tag{4.9}$$

The above equation can be expressed in an importance sampled form:

$$\begin{aligned} \langle \mathcal{O} \rangle &= \frac{\sum_x \langle \psi_{var} | x \rangle \langle x | \mathcal{O} | \psi_{var} \rangle}{\sum_x \langle \psi_{var} | x \rangle \langle x | \psi_{var} \rangle} \\ &= \frac{\sum_x \langle \psi_{var} | x \rangle \langle x | \psi_{var} \rangle \frac{\langle x | \mathcal{O} | \psi_{var} \rangle}{\langle x | \psi_{var} \rangle}}{\sum_x \langle \psi_{var} | x \rangle \langle x | \psi_{var} \rangle} \\ &= \frac{\sum_x |\psi_{var}(x)|^2 \frac{\langle x | \mathcal{O} | \psi_{var} \rangle}{\langle x | \psi_{var} \rangle}}{\sum_x |\psi_{var}(x)|^2} \\ &= \frac{\sum_x |\psi_{var}(x)|^2 \mathcal{O}_{loc}}{\sum_x |\psi_{var}(x)|^2} \end{aligned} \tag{4.10}$$

The quantity $\mathcal{O}_{loc}$ appearing in the above equation is called the *local estimator* and is defined as:

$$\mathcal{O}_{loc} = \frac{\langle x | \mathcal{O} | \psi_{var} \rangle}{\langle x | \psi_{var} \rangle} \tag{4.11}$$

It is now straightforward to notice that the quantity:

$$p(x) = \frac{|\psi_{var}(x)|^2}{\sum_x |\psi_{var}(x)|^2}, \tag{4.12}$$

defines a probability [5]. It is then possible to use a Markov Chain Monte Carlo approach in order to sample a finite set of states distributed according to the corresponding equilibrium distribution . Also the above approach was established for a general operator $\mathcal{O}$ and therefore any observable quantity of interest can be computed stochastically.

[5] $\sum_x p(x) = 1$ and it has a non-negative value for all configurations.

For the case of the expectation value of the Hamiltonian $\mathcal{H}$, one can define a local estimator E_{loc} called local energy:

$$E_{loc}(x) = \frac{\langle x | \mathcal{H} | \psi_{var} \rangle}{\langle x | \psi_{var} \rangle}. \tag{4.13}$$

The quantum average of $\mathcal{H}^2$ is then given by:

$$\begin{aligned} \frac{\langle \psi_{var} | \mathcal{H}^2 | \psi_{var} \rangle}{\langle \psi_{var} | \psi_{var} \rangle} &= \frac{\sum_x \langle \psi_{var} | \mathcal{H} | x \rangle \langle x | \mathcal{H} | \psi_{var} \rangle}{\sum_x \langle \psi_{var} | x \rangle \langle x | \psi_{var} \rangle} \\ &= \frac{\sum_x \langle \psi_{var} | x \rangle \langle x | \psi_{var} \rangle \frac{\langle \psi_{var} | \mathcal{H} | x \rangle \langle x | \mathcal{H} | \psi_{var} \rangle}{\langle \psi_{var} | x \rangle \langle x | \psi_{var} \rangle}}{\sum_x \langle \psi_{var} | x \rangle \langle x | \psi_{var} \rangle} \\ &= \frac{\sum_x |\psi_{var}(x)|^2 |E_{loc}(x)|^2}{\sum_x |\psi_{var}(x)|^2} \end{aligned} \tag{4.14}$$

The variance of the local energy E_{loc} is equal to the variance of the Hamiltonian:

$$\sigma^2_{E_{loc}} = \frac{\langle\psi_{var}|(\mathcal{H}-E)^2|\psi_{var}\rangle}{\langle\psi_{var}|\psi_{var}\rangle} \tag{4.15}$$

One can consider the case where ψ_{var} is an eigenstate of $\mathcal{H}$ and observe that the local energy E_{loc} is constant:

$$E_{loc} = \frac{\langle x|\mathcal{H}|\psi_{var}\rangle}{\langle x|\psi_{var}\rangle} = E\frac{\langle x|\psi_{var}\rangle}{\langle x|\psi_{var}\rangle} = E. \tag{4.16}$$

The above result implies that the variance is zero. This zero-variance property is a feature exclusive to the estimation of quantum expectation values and indicates that the closer we get to an eigenstate, the smaller the fluctuations become. A classical system does not exhibit the same behavior, due to the thermal fluctuations.

4.4.0 Reinforcement Learning

In order to establish a gradient-based optimization approach the energy derivative must also be expressed as an expectation value. Assuming a variational parameter p_k of the wave function the energy derivative with respect to p_k equals:

$$\begin{aligned}
\partial_{p_k}\langle\mathcal{H}\rangle &= \partial_{p_k}\frac{\langle\psi_{var}|\mathcal{H}|\psi_{var}\rangle}{\langle\psi_{var}|\psi_{var}\rangle} \\
&= \frac{\partial_{p_k}\big(\langle\psi_{var}|\mathcal{H}|\psi_{var}\rangle\big)\langle\psi_{var}|\psi_{var}\rangle - \langle\psi_{var}|\mathcal{H}|\psi_{var}\rangle\,\partial_{p_k}\big(\langle\psi_{var}|\psi_{var}\rangle\big)}{\big(\langle\psi_{var}|\psi_{var}\rangle\big)^2} \\
&= \frac{\partial_{p_k}\big(\langle\psi_{var}|\mathcal{H}|\psi_{var}\rangle\big)}{\langle\psi_{var}|\psi_{var}\rangle} - \frac{\langle\psi_{var}|\mathcal{H}|\psi_{var}\rangle}{\langle\psi_{var}|\psi_{var}\rangle}\frac{\partial_{p_k}\big(\langle\psi_{var}|\psi_{var}\rangle\big)}{\langle\psi_{var}|\psi_{var}\rangle} \\
&= \frac{\langle\partial_{p_k}\psi_{var}|\mathcal{H}|\psi_{var}\rangle}{\langle\psi_{var}|\psi_{var}\rangle} + \frac{\langle\psi_{var}|\mathcal{H}|\partial_{p_k}\psi_{var}\rangle}{\langle\psi_{var}|\psi_{var}\rangle} - \langle\mathcal{H}\rangle\frac{\langle\partial_{p_k}\psi_{var}|\psi_{var}\rangle + \langle\psi_{var}|\partial_{p_k}\psi_{var}\rangle}{\langle\psi_{var}|\psi_{var}\rangle} \\
&= \frac{\sum_x\langle\partial_{p_k}\psi_{var}|x\rangle\langle x|\mathcal{H}|\psi_{var}\rangle}{\sum_x\langle\psi_{var}|x\rangle\langle x|\psi_{var}\rangle} + \frac{\sum_x\langle\psi_{var}|\mathcal{H}|x\rangle\langle x|\partial_{p_k}\psi_{var}\rangle}{\sum_x\langle\psi_{var}|x\rangle\langle x|\psi_{var}\rangle} \\
&- \langle\mathcal{H}\rangle\frac{\sum_x\langle\partial_{p_k}\psi_{var}|x\rangle\langle x|\psi_{var}\rangle + \sum_x\langle\psi_{var}|x\rangle\langle x|\partial_{p_k}\psi_{var}\rangle}{\sum_x\langle\psi_{var}|x\rangle\langle x|\psi_{var}\rangle} \\
&\simeq \langle E_{loc}D^*_k\rangle - \langle E_{loc}\rangle\langle D^*_k\rangle + c.c.
\end{aligned} \tag{4.17}$$

The quantity $D_k(x)$ appearing in the above equation is defined as:

$$D_k(x) = \frac{1}{\langle x|\psi_{var}\rangle}\partial_{p_k}(\langle x|\psi_{var}\rangle) \tag{4.18}$$

The idea is to set the marginal distribution of the visible units equal to the

wave-function:

$$\psi(\boldsymbol{x}) = F(\boldsymbol{x}; p_1, p_2, \dots p_{N_p}) \tag{4.19}$$

The Restricted Boltzmann Machine will be chosen based on our previous implementation. The visible and hidden units will now be expressed as σ and h respectively and the variational parameters can be complex-valued. This leads to an expression analogous to equation 1.35 :

$$\begin{aligned} F_{rbm}(\sigma_1^z, \sigma_2^z, \dots \sigma_N^z) &= \sum_{\{h\}} e^{\sum_i \sum_j w_{ij}\sigma_i^z h_j + \sum_j h_j b_j + \sum_i \sigma_i^z a_i} \\ &= e^{\sum_i \sigma_i^z a_i} \sum_{\{h\}} e^{\sum_i \sum_j w_{ij}\sigma_i^z h_j + \sum_j h_j b_j} \\ &= e^{\sum_i \sigma_i^z a_i} \sum_h \prod_j e^{\sum_i w_{ij}\sigma_i^z h_j + h_j b_j} \\ &= e^{\sum_i \sigma_i^z a_i} \prod_j \left(1 + e^{\sum_i w_{ij}\sigma_i^z + b_j}\right) \end{aligned} \tag{4.20}$$

For the case of the transverse-field Ising Model in $d = 1$ and using the prior knowledge that the wave function describing the ground state is positive definite, an optimal choice for the neural network quantum state resulting in real valued variational parameters is [6] :

[6]

$$\Psi(\sigma_1^z, \sigma_2^z, \dots \sigma_N^z) = \sqrt{F_{rbm}(\sigma_1^z, \sigma_2^z, \dots \sigma_N^z)} \tag{4.21}$$

Assuming a spin configuration σ of the transverse-field Ising Model, and considering the locality of the Hamiltonian, the local energy is equal to:

$$E_{loc}(\sigma) = \frac{\langle\sigma|\mathcal{H}|\psi_{var}\rangle}{\langle\sigma|\psi_{var}\rangle} = \sum_{\sigma'} \langle\sigma|\mathcal{H}|\sigma'\rangle \frac{\langle\sigma'|\psi_{var}\rangle}{\langle\sigma|\psi_{var}\rangle} \tag{4.22}$$

The above summation is over $N + 1$ configurations where $\sigma'(0) = \sigma$ is the configuration with no spins flipped and $\langle\sigma|\mathcal{H}|\sigma\rangle = -J\sum_i \sigma_i^z\sigma_{i+1}^z$. The remaining $\sigma'(k) = \sigma_1^z \dots \sigma_k^{z'} \dots \sigma_N^z$ configurations are generated by flipping the k spin with $\langle\sigma|\mathcal{H}|\sigma'\rangle = -h$. Notice that:

$$\begin{aligned} \frac{\psi_{var}(\sigma'(k))}{\psi_{var}(\sigma)} &= \frac{\sqrt{e^{\sum_i \sigma_i^z a_i - a_k(1-2\sigma_k^z)} \prod_j \left(1 + e^{\sum_i w_{ij}\sigma_i^z + b_j - w_{jk}(1-2\sigma_k^z)}\right)}}{\sqrt{e^{\sum_i \sigma_i^z a_i} \prod_j \left(1 + e^{\sum_i w_{ij}\sigma_i^z + b_j}\right)}} \\ &= e^{-\frac{1}{2}a_k(1-2\sigma_k^z)} \sqrt{\frac{\prod_j 1 + e^{\sum_i w_{ij}\sigma_i^z + b_j - w_{jk}(1-2\sigma_k^z)}}{\prod_j 1 + e^{\sum_i w_{ij}\sigma_i^z + b_j}}} \end{aligned} \tag{4.23}$$

$$\begin{aligned}\ln\left(\frac{\psi_{var}(\sigma'(k))}{\psi_{var}(\sigma)}\right) =& -\frac{1}{2}a_k(1-2\sigma_k^z)\\ &+\frac{1}{2}\ln\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j-w_{jk}(1-2\sigma_k^z)}\right)\\ &-\frac{1}{2}\ln\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)\\ =& -\frac{1}{2}a_k(1-2\sigma_k^z)\\ &+\frac{1}{2}\sum_j\ln\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j-w_{jk}(1-2\sigma_k^z)}\right)\\ &-\frac{1}{2}\sum_j\ln\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)\end{aligned} \tag{4.24}$$

Similarly, one acquires expressions for the derivatives with respect to the variational parameters:

$$\begin{aligned}D_{a_i}(\sigma) &= \frac{\partial_{a_i}\left(\sqrt{e^{\sum_i \sigma_i^z a_i}\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)}\right)}{\sqrt{e^{\sum_i \sigma_i^z a_i}\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)}}\\ &= \frac{1}{2}\sigma_i^z\end{aligned} \tag{4.25}$$

$$\begin{aligned}D_{b_j}(\sigma) &= \frac{\partial_{b_j}\left(\sqrt{e^{\sum_i \sigma_i^z a_i}\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)}\right)}{\sqrt{e^{\sum_i \sigma_i^z a_i}\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)}}\\ &= \frac{1}{2}sig\Big(\sum_i w_{ij}\sigma_i+b_j\Big)\end{aligned} \tag{4.26}$$

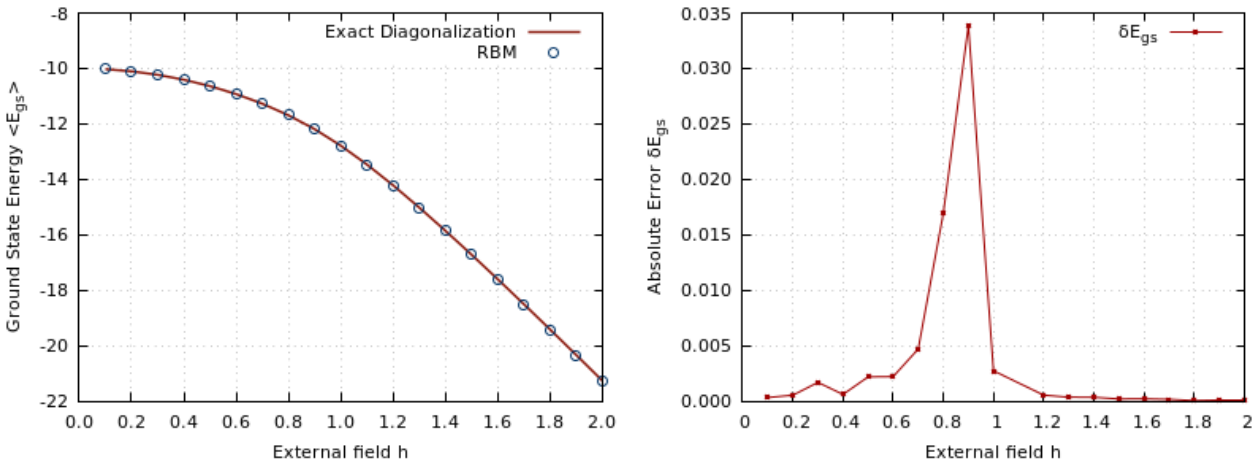
$$\begin{aligned}D_{w_{ij}}(\sigma) &= \frac{\partial_{w_{ij}}\left(\sqrt{e^{\sum_i \sigma_i^z a_i}\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)}\right)}{\sqrt{e^{\sum_i \sigma_i^z a_i}\prod_j\left(1+e^{\sum_i w_{ij}\sigma_i^z+b_j}\right)}}\\ &= \frac{1}{2}\sigma_i^z sig\Big(\sum_i w_{ij}\sigma_i+b_j\Big)\end{aligned} \tag{4.27}$$

It is now possible to train a neural network in order to estimate the energy of the ground state. An amount of configurations is generated at every epoch for the visible units through Gibbs sampling. These configurations, which have been mapped to the Ising Model spins, are then used to calculate the necessary quantities and update the variational parameters accordingly with gradient descent.

The technical details are similar to the implementation of Restricted Boltzmann Machines in the preceding chapters. The difference is the reinforcement learning approach established mainly by sampling states using the neural network and minimizing the expectation value of the energy. Therefore, the quantity of interest is optimized from the beginning. In unsupervised learning the goal was the minimization of the Kullback-Leibler divergence between two distributions.

When the number of epochs is sufficiently enough the neural network will have reached a minimum and it's variational parameters will have be tuned to correspond to a representation of the ground state. It is then possible to randomly initialize the visible units and lead the Restricted Boltzmann Machine into equilibrium in order to sample states. These states can then be used, in conjunction with the wave function chosen before, in order to calculate other quantities of interest besides the local energy.

For every value of the external field h and constant $J = 1.0$ a neural network has been trained with learning rate $l = 0.2$, hidden units $n_h = 10$, batch size $b = 100$, epochs $e = 50000$. Once the neural network has reached the minimum, the expectation value of the local energy will have some small fluctuations. A polynomial fit can be used in order to acquire values of higher accuracy. Additionally absolute errors have also been drawn. This neural network approach is highly accurate and does not suffer from the sign problem. Therefore it can be used on a variety of hard-to-solve problems like determining the exact ground state of strongly interacting fermions.



(i) (ii)

Figure 4.2: Expectation values of (i) the ground state energy in terms of the external field h for a transverse-field Ising model in $d = 1$ with $L = 10$ spins. The red line corresponds to the ground state energy as calculated with exact diagonalization. Absolute errors are also drawn in (ii). The system is led into a quantum phase transition when the external field is $h = 1.0$.

5. *Discussion*

Neural networks prove to be an interesting research tool when confronting problems on a statistical physics based approach. The applications of the Restricted Boltzmann Machines in this thesis have also the capacity to be used in a quite exotic way.

For example consider the unsupervised learning approach for the $d = 2$ Ising model. The Restricted Boltzmann Machine is able to capture the thermodynamic behavior when the hidden neurons are equal to the degrees of freedom. By thermodynamic behavior, the correct transition rate between states is meant. This implies that the problem studied is literally moved on to the neural network. One can discard the Monte Carlo measurements and sample states from the machine. What is even more important, is that if one can observe degrees of freedom experimentally, then these observations instead of Monte Carlo measurements can be used to train the neural network.

Let us assume a real system that has a slow evolution in time, and the same system simulated with a Monte Carlo approach. A Restricted Boltzmann Machine can be trained on configurations of the simulated system. One has then the option to initialize the neural network on an experimentally observed configuration and inspect it's evolution in time.

Additionally, neural networks offer the option to "compress" the system studied. They appear to be implementing a generalized blocking transformation on the data set. It is of great importance, based on the correspondence between Renormalization Group and neural networks, that this coarse-graining is realized in an automated way. Further reduction of hidden neurons is equal to higher compression rates and the transformation is reminiscent of the spin blocking implemented for a variating rescaling factor. The idea is definitely worthy of further research.

The Real-Space Renormalization Group is also demonstrated to be a superior technique when compared with finite size scaling extrapolations for the estimation of the critical temperature and the critical exponents. Even though the method suffers from uncontrolled errors introduced by assuming that the rescaled configurations appear with their correct Boltzmann probabilities, it gives very accurate results for very small sizes of lattices.

Finally, the Reinforcement Learning approach established by introducing quantum states on neural networks proves to be competitive when compared

with other relevant methods. The estimation of the ground state energy for the transverse-field Ising model in $d = 1$ compares well with results from exact diagonalization. Additionally once the neural network has reached a minimum it can be led into equilibrium in order to sample states and calculate other observable quantities of interest. Since this approach is not influenced by the sign problem, it can be used to deal with hard-to-solve problem like gaining insights on the low-energy behavior of strongly interacting fermions.

A. Code: Reproducing Results

The code used in this thesis is written in C, C++ and Python and is available at https://github.com/dbachtis. The interested reader can also find code for the Ising Model at [1]. Every program, even the ones in Python, uses the *getopt* C library function to parse arguments from the command line. Following this approach makes it easy to automate the production of results by using *shell scripts*.

Instead of listing all code a pedagogical example will be shown on how to reproduce results. The idea is the same for all relevant code listed in the above page. Let us examine the unsupervised learning of Restricted Boltzmann Machines. The *tcsh* shell script that calls the programs is:

[1] Konstantinos N. Anagnostopoulos. *Computational Physics: A Practical Introduction to Computational Physics and Scientific Computing (Using C++)*. Konstantinos N. Anagnostopoulos and the National Technical University of Athens, 2016

```
#!/bin/tcsh -f

set visible = 64
set hidden  = (64 16)
set learn   = 0.01
set wdecay  = 0.0
set batch   = 50
set cd      = 20
set epochs  = 100
set start   = 0
set imomentum = 0.0
set fmomentum = 0.0
set betas   = (0.40 0.44 0.48)

foreach beta ($betas)
 foreach hid ($hidden)
    python rbm.py -v $visible -h $hid -l $learn -w $wdecay
↪  -b $batch -c $cd -f 8b${beta}conf.csv -e $epochs -s
↪  $start -m $imomentum -M $fmomentum
    mv errors.dat ./weights/errors.dat
    mv weights wN${visible}b${beta}h${hid}
 end
end
```

For each temperature *beta*, and then for each number of hidden units *hid*, the shell script calls *rbm.py* with a set of mandatory arguments. The python program reads the data from the input file, then trains the Restricted Boltzmann Machine and saves the weights and biases produced at each epoch in a folder *./weights*. It then renames the folder and continues the same procedure. The python code is:

```
from __future__ import print_function
import numpy as np, pandas as pd
import getopt, sys, os


def main():

        #Initialization
        try:
            opts, args = getopt.getopt(sys.argv[1:],
↪  "v:h:l:w:b:c:f:e:s:m:M:", ["help", "output="])
        except getopt.GetoptError as err:
            print(err)
            usage()
            sys.exit(2)
        for o,a in opts:
            if o=="-v":
                visible=int(a)
            elif o =="-h":
                hidden=int(a)
            elif o=="-l":
                learn=float(a)
            elif o =="-w":
                wdecay=float(a)
            elif o=="-b":
                batch_size=int(a)
            elif o =="-c":
                cd=int(a)
            elif o=="-f":
                fname=str(a)
            elif o=="-e":
                epochs=int(a)
            elif o =="-s":
                start=int(a)
            elif o=="-m":
                momentum=float(a)
            elif o =="-M":
                final_momentum=float(a)
```

```
            else:
                assert False, "unhandled option"

        if (start==0):
            #mu, sigma= 0, 0.01
            #self.weights=
↪   np.random.normal(mu,sigma,(self.visible,self.hidden))
            weights = np.sqrt(1./(hidden+visible)) *
↪   np.random.randn(visible, hidden)

↪   #self.hidden_bias=np.random.normal(mu,sigma,(1,self.hidden))
            hidden_bias = np.sqrt(1./(hidden+visible)) *
↪   np.random.randn(1, hidden)

↪   #self.visible_bias=np.random.normal(mu,sigma,(1,self.visible))
            visible_bias = np.sqrt(1./(hidden+visible)) *
↪   np.random.randn(1,visible)
        else:
            weights =np.loadtxt("w.dat")
            hidden_bias=np.loadtxt("hb.dat")
            visible_bias=np.loadtxt("vb.dat")

        data=pd.read_csv(fname,delimiter='
↪   ',index_col=False, header=None).values

        #Training
        winc=np.zeros((visible,hidden))
        hbinc=np.zeros((1,hidden))
        vbinc=np.zeros((1,visible))

        os.mkdir("weights")

        f=open('errors.dat','w')
        f.close()

        for epoch in range(start,epochs):
            training_error=0

            np.random.shuffle(data)
            b=np.vsplit(data,data.shape[0]/batch_size)
            for batch in b:

↪   hidden_probabilities=sigmoid(np.dot(batch,weights)+hidden_bias)
```

```
                hidden_states=hidden_probabilities >
↪ np.random.rand(batch_size,hidden)

                poswinc= np.dot(batch.T,
↪ hidden_probabilities)

↪ poshbinc=np.sum(hidden_probabilities,axis=0,
↪ keepdims=True)
                posvbinc=np.sum(batch,axis=0,keepdims=True)
                for x in xrange(0,cd):

↪ visible_probabilities=sigmoid(np.dot(hidden_states,
↪ weights.T)+visible_bias)
                    visible_states=visible_probabilities >
↪ np.random.rand(batch_size,visible)


↪ hidden_probabilities=sigmoid(np.dot(visible_states,weights)+hidden_bias)
                    hidden_states=hidden_probabilities >
↪ np.random.rand(batch_size,hidden)
                negwinc= np.dot(visible_states.T,
↪ hidden_probabilities)

↪ neghbinc=np.sum(hidden_probabilities,axis=0,
↪ keepdims=True)

↪ negvbinc=np.sum(visible_states,axis=0,keepdims=True)
                winc= momentum*winc+ learn *
↪ ((poswinc-negwinc)/(batch_size) -wdecay*weights)
                #winc= momentum*winc+ learn *
↪ ((poswinc-negwinc)/(batch_size)
↪ -wdecay*np.divide(np.abs(weights),weights,
↪ out=np.zeros_like(np.abs(weights)),where=weights!=0 ))
                hbinc=momentum*hbinc+learn *
↪ (poshbinc-neghbinc)/(batch_size)
                vbinc=momentum*vbinc+learn *
↪ (posvbinc-negvbinc)/(batch_size)
                weights += winc
                hidden_bias+=hbinc
                visible_bias+=vbinc


↪ training_error+=np.sum((batch-visible_states) **2)
```

```
↪  training_error=float(training_error)/(data.shape[0]*visible)

            f=open('errors.dat','a')
            f.write( str(epoch) + ' ' + str(training_error)
↪  + '\n')
            f.close()
            np.savetxt("./weights/w" + str(epoch) + ".dat"
↪  ,weights,fmt='%f',delimiter=' ')
            np.savetxt("./weights/hb" + str(epoch) + ".dat"
↪  ,hidden_bias,fmt='%f',delimiter=' ')
            np.savetxt("./weights/vb" + str(epoch) + ".dat"
↪  ,visible_bias,fmt='%f',delimiter=' ')
            print("Epoch %s: training-error:%s" %
↪  (epoch,training_error))
            if((epoch-start) == 20):
                momentum=final_momentum

def sigmoid(x):
        x = np.clip( x, -500, 500 )
        return 1.0 / (1.0 + np.exp(-x))

if __name__ == "__main__":
    main()
```

Once the weights and biases have been saved, another tcsh shell script named *reconstruct* calls *rec.py* can be called in order to choose a set of weights and biases and start a Gibbs Chain from randomly initialized visible units.

```
#!/bin/tcsh -f

set visible = 64
set hidden  = (64 16)
set epoch    = 49
set betas   = (0.40 0.44 0.48)
set reconstructions = 100000

foreach beta ($betas)
 foreach hid ($hidden)
    python rec.py -v $visible -h $hid -r $reconstructions
↪  -w ./wN${visible}b${beta}h${hid}/w${epoch}.dat -b
↪  ./wN${visible}b${beta}h${hid}/hb${epoch}.dat -a
↪  ./wN${visible}b${beta}h${hid}/vb${epoch}.dat
    mv rec.dat rN${visible}b${beta}h${hid}.dat
 end
```

```
end
```

These visible states can then be saved and be treated as configurations of the Ising Model in order to calculate expectation values. They do not differ-from a practical point of view when doing calculations- from configurations sampled through Monte Carlo. Obviously, one has to be careful that the neural network has reached equilibrium and one might also want to initiate more than one Gibbs chains to reduce autocorrelations. The *rec.py* code is:

```
from __future__ import print_function
import numpy as np
import getopt, sys

def main():

        try:
            opts, args = getopt.getopt(sys.argv[1:],
↪  "v:h:r:w:b:a:", ["help", "output="])
        except getopt.GetoptError as err:
            print(err)
            usage()
            sys.exit(2)
        for o,a in opts:
            if o=="-v":
                visible=int(a)
            elif o =="-h":
                hidden=int(a)
            elif o=="-r":
                reconstructions=int(a)
            elif o=="-w":
                wfile=str(a)
            elif o=="-b":
                hbfile=str(a)
            elif o=="-a":
                vbfile=str(a)
            else:
                assert False, "unhandled option"

        weights=np.loadtxt(wfile)
        hidden_bias=np.loadtxt(hbfile)
        visible_bias=np.loadtxt(vbfile)

        f_handle = file('rec.dat','w')
        f_handle.close()
```

```
        visible_states=np.random.randint(2, size=(1,
↪  visible))
        for x in range(reconstructions):


↪  hidden_probs=sigmoid(np.dot(visible_states,weights)+hidden_bias)
            hidden_states=hidden_probs >
↪  np.random.rand(1,hidden)


↪  visible_probs=sigmoid(np.dot(hidden_states,weights.T)+visible_bias)
            visible_states = visible_probs >
↪  np.random.rand(1,visible)
            f_handle = file('rec.dat','a')

↪  np.savetxt(f_handle,visible_states,fmt='%i',delimiter='
↪  ')
            f_handle.close()

def sigmoid( x):
        x = np.clip( x, -500, 500 )
        return 1.0 / (1 + np.exp(-x))

if __name__=='__main__':
    main()
```

The rest of the code for Reinforcement Learning or for the Renormalization Group works similarly. The Deep Belief network is also trained using an appropriate shell script and the *rbm.py* listed above.

Bibliography

[1] Asja Fischer and Christian Igel. An Introduction to Restricted Boltzmann Machines. *Alvarez L., Mejail M., Gomez L., Jacobo J. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2012. Lecture Notes in Computer Science, vol 7441.*, 2012. Springer, Berlin, Heidelberg.

[2] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science 9*, 1985.

[3] Geoffrey E. Hinton. Boltzmann machine. *Scholarpedia, 2(5):1668*, 2007.

[4] Steffen L. Lauritzen. *Graphical models.* Oxford University Press, 1996.

[5] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[6] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation 14, 1771-1800*, 2002.

[7] Max Welling. Product of Experts. *Scholarpedia, 2(10):3879*, 2007.

[8] Guido Montufar and Nihat Ay. Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. *Neural Computation*, 2011.

[9] Yoshua Bengio. *Learning deep architectures for AI.* Foundations and Trends in Machine Learning 21(6), 1601-1621, 2009.

[10] Yoshua Bengion, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing (NIPS 19), pp. 153-160*, 2007. MIT Press.

[11] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation 18(7), 1527-1554*, 2006.

[12] Yoshua Bengio and Olivier Delalleau. Justifying and generalizing contrastive divergence. *Neural Computation 21(6), 1601-1621*, 2009.

[13] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. Learning multiple layers of representation. *Trends in Cognitive Sciences 11(10), 428-434*, 2007.

[14] Asja Fischer and Christian Igel. Bounding the bias of contrastive divergence learning. *Neural Computation 23, 664-673*, 2011.

[15] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neu- ral networks. *Science 313(5786), 504-507*, 2006.

[16] Konstantinos N. Anagnostopoulos. *Computational Physics: A Practical Introduction to Computational Physics and Scientific Computing (Using C++)*. Konstantinos N. Anagnostopoulos and the National Technical University of Athens, 2016.

[17] Bernd A. Berg. *Markov Chain Monte Carlo Simulations and their Statistical Analysis: With Web-Based Fortran Code*. World Scientific Publishing Co. Pte. Ltd, 2004.

[18] M.E.J Newman and G.T Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, 1999.

[19] E. Ising. Beitrag zur Theorie des Ferromagnetizmus. *Z. Phys. 31*, 1925.

[20] Lars Onsager. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physical Review, Series II*, 1944.

[21] Alan M. Ferrenberg and Robert H. Swendsen. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett. 61*, 1988.

[22] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, and Augusta H. Teller. Equation of State Calculations by Fast Computing Machines . *The Journal of Chemical Physics 21, 1087*, 1953.

[23] Wilfred K. Hastings. Monte Carlo sampling methods using Markov chains and their applications . *Biometrika 57: 97-109*, 1970.

[24] Robert H. Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett. 58, 86*, 1987.

[25] C. M. Fortuin and P. W. Kasteleyn. On the random-cluster model: I. Introduction and relation to other models. *Physica, Volume 57, Issue 4*, 1972.

[26] Dietrich Stauffer and Ammon Aharony. *Introduction To Percolation Theory*. CRC Press, 1994.

[27] Henk W. J. BlÃűte and Youjin Deng. Cluster Monte Carlo simulation of the transverse Ising model. *Phys. Rev. E 66, 066110*, 2002.

[28] Ulli Wolff. Collective Monte Carlo Updating for Spin Systems. *Phys. Rev. Lett. 62, 361*, 1989.

[29] James Propp and David Wilson. Coupling from the Past: a User's Guide, 1997.

[30] H. Flyvbjerg and H. G. Petersen. Error estimates on averages of correlated data . *The Journal of Chemical Physics 91, 461*, 1989.

[31] Giacomo Torlai and Roger G. Melko. Learning thermodynamics with Boltzmann machines. *Phys. Rev. B 94, 165134*, 2016.

[32] Kenneth G.Wilson and J.Kogut. The renormalization group and the Ït expansion. *Physics Reports, Volume 12, Issue 2*, 1974.

[33] Kenneth G.Wilson. The renormalization group and critical phenomena. *Rev. Mod. Phys. 55, 583*, 1983.

[34] John Cardy. *Scaling and Renormalization in Statistical Physics*. Cambridge University Press, 1996.

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

[36] Nigel Goldenfeld. *Lectures On Phase Transitions And The Renormalization Group (Frontiers in Physics)* . Addison-Wesley, 1992.

[37] Leo P. Kadanoff, Anthony Houghton, and Mehmet C. Yalabik. Variational approximations for renormalization group transformations. *J. Stat. Phys. 14: 171*, 1976.

[38] Efi Efrati, Zhe Wang, Amy Kolan, and Leo P. Kadanoff. Real-space renormalization in statistical mechanics. *Rev. Mod. Phys. 86, 647*, 2014.

[39] Pankaj Mehta and David J. Schwab. An exact mapping between the Variational Renormalization Group and Deep Learning. *arXiv:1410.3831*, 2014.

[40] David J. Griffiths. *Introduction to Quantum Mechanics*. Prentice Hall, 1995.

[41] P.G.de Gennes. Collective motions of hydrogen bonds. *Solid State Communications, Volume 1, Issue 6*, 1963.

[42] Sei Suzuki, Jun ichi Inoue, and Bikas K. Chakrabarti. *Quantum Ising Phases and Transitions in Transverse Ising Models*. Cambridge University Press, 2013.

[43] W. L. McMillan. Ground State of Liquid He^4. *Phys. Rev. 138, A442*, 1965.

[44] Federico Becca and Sandro Sorella. *Quantum Monte Carlo Approaches for Correlated Systems*. Cambridge University Press, 2017.

[45] Giuseppe Carleo. Machine learning methods for many body physics, 2017. Lectures for the Advanced School on Quantum Science and Quantum technology, ICTP, Trieste, Italy.

[46] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science Vol. 355, Issue 6325, pp. 602-606*, 2017.

[47] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nature Physics*, 2018.