# Random Walk Formulation of Learning in Restricted Boltzmann Machines

Max Vargas and Kimberly Villalobos
Mentor: Mason Biamonte
Project proposed by: Mason Biamonte
Department of Mathematics
Massachusetts Institute of Technology
August 2, 2017

Abstract. Mehta and Schwab (2014) conjectured that there exists a mapping between deep learning in restricted Boltzmann machines and the renormalization group, which arises in describing the phase transition properties of generalized Ising models and in exorcising pathological infinities that abound in quantum field theory. Rigorous approaches to the renormalization group for generalized Ising models have been developed using random walks. By taking advantage of the Ising form of the energy function defined on restricted Boltzmann machines, we employ such a random walk approach to elucidate the formal connection between deep learning and the renormalization group. This approach not only provides tools for proving fundamental properties of learning algorithms, it also casts deep learning in the framework of rigorous approaches to both critical phenomena and quantum field theory.

## 1. Introduction

A Restricted Boltzmann machine (RBM) provides useful algorithms in machine learning to describe an unknown probability distribution. Given a set of observations, also called training data, RBMs provide approximations to the joint probability distribution of the underlying data, which allows us to sample and make predictions about unseen observations [2]. For example, when the training data is a set of pixels from an image, knowing the joint probability distribution of the pixels grants us the ability to solve tasks related to pattern recognition and machine vision.

RBMs can be regarded as undirected graphical models that represent the probability distribution of what are called visible and hidden variables. The visible units represent the observations from the data, and the hidden variables are introduced to capture dependencies of the visible variables. Suppose that these dependencies between hidden and visible variables are encoded in values $w_{ij}$, where $i$ represents some visible vertex and $j$ some hidden vertex. Learning an RBM means adjusting these parameters $w_{ij}$ such that the marginal probability distribution of the visible units matches the data as accurately as possible [2].

Mehta and Schwab (2014) suggest that the learning techniques for RBMs are closely related to the renormalization group (RG), a coarse-graining procedure from quantum field theory often used to extract information from a lattice spin system [5]. One popular technique regarding renormalization on an Ising-type lattice is known as decimation. Given some Ising lattice, decimation allows you to create a new, smaller lattice where each vertex encodes an average value of a block of vertices in the original lattice. In the machine learning settings, each vertex represents a unit of data, such as a pixel or a bit. Mehta and Schwab work through several examples

displaying a parallel between repeated decimation and learning an RBM through some deep neural net. However, their claim remains at the level of conjecture [5].

It turns out that there is a remarkable connection between quantum fields and what are known as Markov random fields (MRFs). By passing into the imaginary time domain, Brydges demonstrates that a quantum field actually becomes an MRF. Making use of the Osterwalder-Schrader axioms, Brydges even provides conditions in which an MRF may be translated back into a quantum field [4]. Since an RBM is a degenerate case of general MRFs, it may not be surprising that there exists a relation between renormalization in quantum field theory and deep learning in the context of machine learning.

Brydges, Frölich, and Spencer (1982) develop a rigorous random walk formulation of lattice spin systems following the polymer representation by Symanzik (1969) in order to construct bounds on correlation functions in Ising-type models [1,6]. Later, Aizenmann (1985) exhibits a critical connection between random walks and the renormalization group by using the scaling property of Brownian motion to show that intersection properties of random walks generated by independent Brownian motions can be described by the renormalization equation for the beta function of a quantum field theory [3]. The invariance described by Brownian motion's self similarity relates to the RG operations near criticality in that the correlations among vertices remain the same after decimation. This connection between random walks and RG could lead to interpretations of critical points for RBMs in the machine learning setting.

Here we use methods similar to those of Brydges et. al. in [1] to take a step towards completing the formal connection between the renormalization group and learning on a RBM by constructing a random walk representation of correlation functions arising from minimization of the Kullback-Leibler divergence (KL-divergence), a measure of the distance between two probability distributions that allows for the quantification of the error in the distribution predicted by the RBM model [2]. A common technique for learning an RBM is to perform gradient descent on the KL-divergence. We show in section 3 the known fact that the gradient of the KL-divergence can be written as the difference of the expectations of the energy function under the conditional distribution of the hidden layer given the training data and under the joint distribution of the hidden and visible variables described by the model.

In section 5 we adapt the polymer representation by Symanzik in [6] to our RBM model in order to express the partition function in terms or random loops between the hidden and visible layers. In sections 6 and 7, we develop a random walk representation for the expectations described by the training data and by the model separetely. Not only does this random walk representation provide a rigorous toolbox for proving fundamental properties and limitations of learning algorithms on RBMs, it also casts deep learning in the framework of quantum field theory, which leads to an analysis and generalization of learning algorithms that arise naturally in

a field theory setting.

## 2. Background

In the following definitions, we let $G = (V, E)$ be any undirected graph and let $x, y \in V$ denote any two vertices in $G$.

**Definition 1.** *A path from $x$ to $y$ is an ordered subset of edges $\omega^{x,y} \subset E$*

$$\omega^{x,y} = ((x, v_1), (v_1, v_2), ..., (v_{n-1}, y))$$

**Definition 2.** *Let $x, y \in V$ be any two vertices. We call $x$ and $y$ separated by a subset $C \subset V$ if every path from $x$ to $y$ passes through $C$. Specifically, if $x$ and $y$ are separated by $C$ and $\omega^{x,y}$ is a path from $x$ to $y$, then*

$$\Big( \bigcup_{e \in \omega^{x,y}} e \Big) \bigcap C \neq \emptyset.$$

**Definition 3.** *Let $\Sigma$ be the set of all possible paths on a lattice $L$, and define $\mu : \Sigma \to [0, 1]$ as the uniform probability measure. A walk random walk $\omega^N \in \{\omega^{x,y} | x, y \in L \text{ and } |\omega^{x,y}| = N\}$ is a path of size $|\omega^N| = N$ on $L$ chosen under the probablity measure $\mu$, and a random loop $\underline{\omega}^N$ is a special case of a random walk in which $x = y$.*

**Definition 4.** *$n(k|\omega)$ is the number of times the walk $\omega$ hits the vertex $k \in L$*

**Definition 5.** *Given a path $\omega$ on a weighted graph $G = (V, E)$ whose weights are given by an adjacency matrix $W$, we define $W_\omega \equiv \prod_{r \in \omega} W_r$.*

**Definition 6.** *Let $G = (L, W)$ be a graph, where $L = L_V \cup L_H$ is a union of finite sub-lattices of $\mathbb{Z}$ and $W$ is a set of weighted undirected edges. Let $s_i$ be a random variable associated with vertex $i$ and let $p$ be the joint probability distribution of $\boldsymbol{s} = (s_0, s_1, ..., s_{|L|-1})$. Two nodes $x$ and $y \in L$ are separated by a set $C \subset L$ if every path from $x$ to $y$ passes through $C$. Formally, then $C$ separates $x$ and $y$ if for all paths $\omega^{x,y}$ from $x$ to $y$, $\omega^{x,y} \cap C \neq \emptyset$. We say that $p$ fulfills the Markov property with respect to $G$ if for all disjoint subsets $A, B, C \in L$ with all nodes in $A$ and $B$ being separated by $C$, it holds that $p((s_a)_{a \in A} | (s_r)_{r \in B \cup C}) = p((s_a)_{a \in A} | (s_r)_{r \in C})$, where $p(a|b)$ is the conditional probability of $b$ given $a$ under the distribution $p$. In this case $\boldsymbol{s}$ is called a Markov random field (MRF).*

## 3. restricted boltzmann machines

A *Restricted Boltzman Machine* (RBM) is a bipartite undirected graph whose associated random variables $\boldsymbol{s}$ are a MRF. It consists of two independent sets $L_H \subset L$ and $L_V \subset L$ which we call the hidden layer and visible layer, respectively. Let $\Omega_{si}$ be the set of possible outcomes of the random variable $s_i$. Define

$$\Omega_V \equiv \prod_{\boldsymbol{s}_i | i \in L_V} \Omega_{\boldsymbol{s}_i} \tag{3.1}$$

as the state space of the visible layer. For notational reasons, we denote by $\boldsymbol{v} = \{v_1, v_2, ..., v_{N_V}\}$ the vector of random variables on the visible layer, where $N_V = |L_V|$. The same construction gives the objects $\Omega_H$, $\boldsymbol{h}$, and $N_H$.

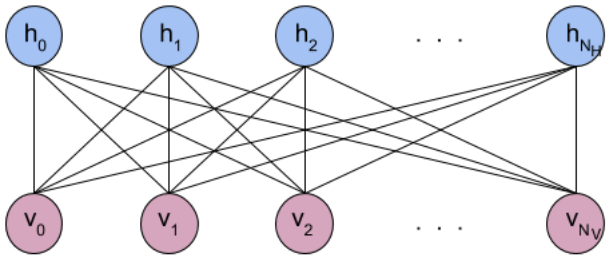


FIGURE 1. Example RBM with $N_H$ hidden variables and $N_V$ visible variables whose edge weights are given by elements of $W$

By the Universal Approximation Theorem proved by Le Roux and Bengio and later improved by Montafur and Ay, we find that for any distribution over $\{1, -1\}^n$ the interactions between the visible and hidden layers can be expressed using the energy function

$$\mathcal{H}(\boldsymbol{v}, \boldsymbol{h}) \equiv -\sum_{i=0}^{N_H-1}\sum_{j=0}^{N_V-1} w_{ij} h_i v_j - \sum_{j=0}^{N_V-1} b_i v_i - \sum_{i=0}^{N_H-1} c_i h_i \tag{3.2}$$

in which $w_{ij}$ is the weight of the edge $v_i h_j$, and $b_j$, $c_i$ are the weights associated with the variables $v_j$ and $h_i$ respectively [8,9]. In the case of the Ising model, the weights $c_i$ and $b_j$ are called a bias or applied magnetic field, where each spin has a magnetic dipole moment that makes the orientation of the spin likely to align along the direction of the field. The joint probability distribution of a specific configuration $\boldsymbol{s} = (\boldsymbol{v}, \boldsymbol{h}) = (s_0, ..., s_{N_S})$ can then be written as

$$p(\boldsymbol{s}) \equiv p(\boldsymbol{v}, \boldsymbol{h}) \equiv \frac{e^{-\beta \mathcal{H}(\boldsymbol{v}, \boldsymbol{h})}}{\mathcal{Z}} \tag{3.3}$$

where

$$\mathcal{Z} \equiv \mathcal{Z}_G(\beta) \equiv \sum_{\boldsymbol{v} \in \Omega_V} \sum_{\boldsymbol{h} \in \Omega_H} e^{-\beta \mathcal{H}(\boldsymbol{v}, \boldsymbol{h})} \tag{3.4}$$

is the normalization constant, better known as the partition function of the RBM. The ability to calculate this partition function by means of summing over all possible configurations of our RBM gives us a powerful tool in calculating other statistical variables. For example, in the context of statistical physics, one often wishes to

calculate what are known as correlation functions. It turns out that the correlation function given by the random variable $v_i h_j$ with $i \in L_V$ and $j \in L_H$ can be calculated via a generating function approach on $\mathcal{Z}$. In particular, we allow ourselves to vary $b_i$ and $c_j$ through some variables $b_i^*$ and $c_j^*$ and the expectation of $v_i h_j$ is

$$\mathbb{E}[v_i h_j] = \left( \frac{\partial}{\partial b_i^*} \frac{\partial}{\partial c_j^*} \mathcal{Z} \right) \Big|_{b_i^* = b_i, c_j^* = c_j}$$

[7]. From now on we consider the case $\beta = 1$, which is the most common scenario in machine learning. However, in order to consider possible interpretations of critical phenomena in RBMs, all the following equations could be easily generalized for an arbitrary $\beta$.

Techniques such as Gibbs sampling are often performed on RBMs to conduct unsupervised learning where the goal is to find the parameters $(w_{ij}, b_k, c_l)$ that best approximate the probability distribution of the data [2], given by

$$p(\boldsymbol{v}) = \sum_{\boldsymbol{h} \in \Omega_H} p(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{\mathcal{Z}} \sum_{\boldsymbol{h} \in \Omega_H} e^{-\mathcal{H}(\boldsymbol{v}, \boldsymbol{h})} \tag{3.5}$$

Once we have a suitable approximation to $p(\boldsymbol{v})$, we can perform the same calculation to find an approximation for the probability distribution over the hidden variables, $p(\boldsymbol{h})$.

$$p(\boldsymbol{h}) = \sum_{\boldsymbol{v} \in \Omega_V} p(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{\mathcal{Z}} \sum_{\boldsymbol{v} \in \Omega_V} e^{-\mathcal{H}(\boldsymbol{v}, \boldsymbol{h})} \tag{3.6}$$

Simply put, we observe the behavior of the visible units and we wish to recover the parameters which give rise to the probability distribution that our RBM follows. Let $q_{N_T}(\boldsymbol{v})$ be the distribution of the observed data, where $N_T$ is the number of training samples. Then as the number of training samples goes to infinity, we have that $q_{N_T}(\boldsymbol{v})$ converges to $q_\infty(\boldsymbol{v}) = p(\boldsymbol{v})$ by the law of large numbers. Despite this convergence, in a practical setting there will almost surely be some degree of approximation error given a finite number of training samples. The Kullback-Leibler divergence (KL-divergence) allows us to measure (from a theoretical perspective) the difference between the observed distribution and the actual distribution [2]. Formally, the KL-divergence from $q_{N_T}$ to $p$ is defined by

$$\begin{aligned} KL(q_{N_T} || p) &\equiv \sum_{\boldsymbol{v} \in \Omega_V} q_{N_T}(\boldsymbol{v}) \ln \frac{q_{N_T}(\boldsymbol{v})}{p(\boldsymbol{v})} \\ &= \sum_{\boldsymbol{v} \in \Omega_V} q_{N_T}(\boldsymbol{v}) \ln q_{N_T}(\boldsymbol{v}) - \sum_{\boldsymbol{v} \in \Omega_V} q_{N_T}(\boldsymbol{v}) \ln p(\boldsymbol{v}). \end{aligned} \tag{3.7}$$

Suppose that we have sampled the visible layer to get $N_T$ samples of training data. Denote the $\mu^{th}$ sample by $\boldsymbol{v}^\mu$. We can achieve a very simple approximation to $q_\infty(\boldsymbol{v})$

by letting

$$q_{N_T}(\boldsymbol{v}) = \frac{1}{N_T}\sum_{\mu=1}^{N_T}\delta(\boldsymbol{v}-\boldsymbol{v}^{\mu}) \tag{3.8}$$

That is, our approximation for the probability distribution of visible variables follows a discrete marginal distribution given by the empirical distribution on the training data.

The learning algorithms for RBMs rely essentially on gradient descent on the KL-divergence, which implies the condition that the marginal distribution of the model becomes a better approximation of the distribution of the data. This gradient is given by

$$\begin{aligned}
\frac{\partial KL(q||p)}{\partial w_{ij}} &= \frac{\partial}{\partial w_{ij}}\sum_{\boldsymbol{v}\in\Omega_V} q(\boldsymbol{v})\log q(\boldsymbol{v}) - \frac{\partial}{\partial w_{ij}}\sum_{\boldsymbol{v}\in\Omega_V} q(\boldsymbol{v})\log p(\boldsymbol{v}) \\
&= -\frac{\partial}{\partial w_{ij}}\sum_{\boldsymbol{v}\in\Omega_V}\frac{1}{N_T}\sum_{\mu=1}^{N_T}\delta(\boldsymbol{v}-\boldsymbol{v}^{\mu})\log p(\boldsymbol{v}) \\
&= -\frac{1}{N_T}\sum_{\mu=1}^{N_T}\frac{\partial}{\partial w_{ij}}\log p(\boldsymbol{v}^{\mu}) \\
&= \frac{1}{N_T}\sum_{\mu=1}^{N_T}\left[\sum_{\boldsymbol{h}\in\Omega_H} p(\boldsymbol{h}|\boldsymbol{v}^{\mu})\frac{\partial\mathcal{H}(\boldsymbol{v}^{\mu},\boldsymbol{h})}{\partial w_{ij}} - \sum_{\substack{\boldsymbol{h},\in\Omega_H \\ \boldsymbol{v}\in\Omega_V}} p(\boldsymbol{v},\boldsymbol{h})\frac{\partial\mathcal{H}(\boldsymbol{v},\boldsymbol{h})}{\partial w_{ij}}\right] \\
&= \frac{1}{N_T}\sum_{\mu=1}^{N_T}\left[\mathbb{E}_{p(\boldsymbol{h}|\boldsymbol{v}^{\boldsymbol{\mu}})}\left[\frac{\partial\mathcal{H}(\boldsymbol{v}^{\mu},\boldsymbol{h})}{\partial w_{ij}}\right] - \mathbb{E}_{p(\boldsymbol{v},\boldsymbol{h})}\left[\frac{\partial\mathcal{H}(\boldsymbol{v},\boldsymbol{h})}{\partial w_{ij}}\right]\right]
\end{aligned} \tag{3.9}$$

Now, recall from the definition of $\mathcal{H}(\boldsymbol{v},\boldsymbol{h})$ that the only term dependent on $w_{ij}$ is $v_i w_{ij} h_j$. Therefore, the derivatives in the argument of the expectation values above simply evaluate to $v_i^{\mu}h_j$ and $v_i h_j$, respecively. Defining $\mathbb{E}^{\mu} = \mathbb{E}_{p(\boldsymbol{h}|\boldsymbol{v}^{\mu})}$ and $\mathbb{E} = \mathbb{E}_{p(\boldsymbol{h}|\boldsymbol{v})}$, we conclude

$$\frac{\partial KL(q||p)}{\partial w_{ij}} = \frac{1}{N_T}\sum_{\mu=1}^{N_T}[\mathbb{E}^{\mu}[v_i^{\mu}h_j] - \mathbb{E}[v_i h_j]] \tag{3.10}$$

Most of the current optimization techniques to find the RBM parameters rely on the value of the gradient in Eq. (3.9), whose calculation is computationally expensive. Gibbs sampling is an algorithm that produces samples from the joint probability distribution of a group of random variables, which is used to generate approximations for the gradient that have a much lower cost than its exact computation. The idea is to update in each time step both, the hidden variables $\boldsymbol{h}$ given $p(\boldsymbol{h}|\boldsymbol{v})$, and the visible variables $\boldsymbol{v}$ given $p(\boldsymbol{v}|\boldsymbol{h})$. This gives rise to a Markov Chain $\boldsymbol{s} = \boldsymbol{s}^{(t)}|t\in\mathbb{N}$ where $\boldsymbol{s}^{(t)} = (s_1^{(t)},...,s_N^{(t)})$ determines the state of the random variables at time $t$. This

chain eventually converges to the distribution of $\boldsymbol{s}$, and then taking a sample from the chain for a sufficiently large $t$ gives a close approximation [2]. It is important to clarify that the random walks arising from this Markov chain are fundamentally different from the random walks that we are going to introduce in this paper, which are time-independent walks on the visible and hidden layers.

## 4. MAIN RESULTS

We take a step towards completing the formal connection between the renormalization group and learning on a RBM by constructing a random walk representation of the expectations arising from the minimization of the KL-divergence in Eq. (3.7). Specifically, if we let $W$ be the weight matrix of our model, and $W^{\mu}$ be a new weight matrix that takes into account the data sample $\mu$ which will be defined more precisely in *proposition 6.1*, we obtain

**Theorem 4.1.**

$$\Delta w_{ij} = \frac{1}{N_T}\sum_{\mu=1}^{N_T}\left[\frac{(2\pi)^{-N_H}\, v_i^{\mu}}{\mathcal{Z}\; P(\boldsymbol{v}^{\mu})}\sum_{\omega^1\in L_s}\mathcal{Y}_j^{\mu}(\omega^1) - \frac{1}{\mathcal{Z}}\sum_{\omega^{\boldsymbol{v}_i,\boldsymbol{h}_j}\subset L_s}\mathcal{X}(\omega^{\boldsymbol{v}_i\boldsymbol{h}_j})\right] \tag{4.1}$$

*where*

$$\begin{aligned}
\mathcal{Y}_j(\omega) &\equiv \sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{1}{2}\right)^n \sum_{\underline{\omega}_1,\ldots,\underline{\omega}_n\subset L_H} W_{\omega}^{\mu}\prod_{k=1}^{n}(-2\mathbf{i}\boldsymbol{a}\boldsymbol{O})_{\underline{\omega}_k}\exp\left[-U_j(\underline{\omega}_1,\underline{\omega}_2,\ldots,\underline{\omega}_n)\right] \\
\exp\left[-U_j(\underline{\omega}_1,\underline{\omega}_2,\ldots,\underline{\omega}_n)\right] &= \prod_{k\in L_H}\int_{\mathbb{R}^{\mathrm{N_H}}} da_k e^{-\mathbf{i}a_i}(2\mathbf{i}a_k d_k)^{\frac{-1}{2}-n(k\,|\,\underline{\omega}_1)-\ldots-n(k\,|\,\underline{\omega}_n)}(2\mathbf{i}a_j)^{-1}
\end{aligned} \tag{4.2}$$

*and*

$$\begin{aligned}
\mathcal{X}(\omega) &\equiv \sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{1}{2}\right)^n \sum_{\underline{\omega}_1,\underline{\omega}_2,\ldots,\underline{\omega}_n}\left(W_{\omega}\prod_{k=1}^{n}W_{\underline{\omega}_k}\right)\exp[-F(\underline{\omega}_1,\ldots,\underline{\omega}_n|\omega)] \\
\exp\left[-F(\underline{\omega}_1,\underline{\omega}_2,\ldots,\underline{\omega}_n)\right] &= \prod_{k\in L_s}\int_{\mathbb{R}^{\mathrm{N_s}}} da_k e^{-\mathbf{i}a_i}(2\mathbf{i}a_k)^{\frac{-1}{2}-n(k\,|\,\underline{\omega}_1)-\ldots-n(k\,|\,\underline{\omega}_n)}
\end{aligned} \tag{4.3}$$

Additionally, we adapt the polymer representation of lattice spin systems developed by Symanzik and elaborated by Brydges et. al. in [1,6] in order to express the partition function of a RBM in terms of random walks that alternate between the visible and hidden layer. Specifically, we show that

$$\mathcal{Z} = (2\pi)^{-N_s}\sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{1}{2}\right)^n \sum_{\underline{\omega}_1,\underline{\omega}_2,\ldots,\underline{\omega}_n\subset L_s}\prod_{k=1}^{n}(W)_{\underline{\omega}_k}\exp\left[-F(\underline{\omega}_1,\underline{\omega}_2,\ldots,\underline{\omega}_n)\right] \tag{4.4}$$

where the edges in the random walks have one end in $L_H$ and the other one in $L_V$ due to the connectivity structure of the RBM.

## 5. Random walk representation of the partition function for RBMs

From now on, we assume that the sample space of the random variable at each vertex of an RBM is $\{\pm 1\}$. In particular, the sample space of the entire RBM can be written as the product of the sample spaces on the hidden and visible layers $\Omega = \Omega_H \times \Omega_V$ where $\Omega_H = \{\pm 1\}^{N_H}$ and $\Omega_V = \{\pm 1\}^{N_V}$. Furthermore, we assume that the bias on each vertex is set to zero ($b_i = c_i = 0$). With these assumptions we lose out on features found in more complex examples such as the Potts model. However, our assumptions do not restrict so much that the model becomes trivial. In fact, Mehta and Schwab provide examples using the binary model on images of binary data to search for underlying structure [5]. Even with these theoretically limiting restrictions on our RBM, these results are still applicable to machine vision, pattern recognition, image classification, and other fields.

**Proposition 5.1.**

$$\mathcal{Z} = (2\pi)^{-N_s} \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{1}{2}\right)^n \sum_{\underline{\omega}_1, \underline{\omega}_2, \ldots, \underline{\omega}_n \subset L_s} \prod_{k=1}^{n} (W)_{\underline{\omega}_k} \exp\left[-F(\underline{\omega}_1, \underline{\omega}_2, ..., \underline{\omega}_n)\right] \quad (5.1)$$

*where*

$$\exp\left[-F(\underline{\omega}_1, \underline{\omega}_2, ..., \underline{\omega}_n)\right] = \prod_{k \in L_s} \int_{\mathbb{R}^{N_s}} da_k e^{-\boldsymbol{i}\boldsymbol{a}_i} (2\boldsymbol{i}a_k)^{\frac{-1}{2} + n(k|\underline{\omega}_1) + \ldots + n(k|\underline{\omega}_n)} \quad (5.2)$$

The above equation presents the partition function $\mathcal{Z}$ in terms of random walks on our RBM. Recalling the form of $W$, we see that any loop $\underline{\omega}$ with an edge $(\boldsymbol{v}_i, \boldsymbol{v}_j)$ or $(\boldsymbol{h}_i, \boldsymbol{h}_j)$ containing two vertices in the same layer will vanish since the weight of

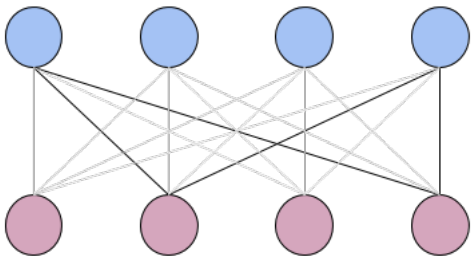

Figure 2. Example a loop outlined in pink which contributes to the partition function of a RBM with four hidden and visible vertices. The other bonds have been grayed out for visual clarity.

any edge between vertices of the same layer are zero. Thus, all walks that contribute to the sum for the partition function must alternate between the visible and hidden layers at each step. Additionally, the expansion allows us to consider several of these terms at once by taking products of loops by multiplying the weightings of all the edges found in each loop. The exponential term acts as a decay factor which forces smaller weights on terms whose loops come back to the same vertex many times.

*Proof.* In order to calculate the partition function by means of random walks on the RBM, we start by introducing the Dirac delta function. This passes us into the continuous regime and our sample space becomes $\Omega' = \mathbb{R}^{N_s}$. This allows us to write the partition function nicely in integral form as

$$\mathcal{Z} = \int_{\Omega'} \prod_{n=0}^{N_S} ds_n \delta(s_n^2 - 1) \exp\left[-\mathcal{H}(\boldsymbol{v}, \boldsymbol{h})\right] \tag{5.3}$$

Notice that $\delta(x)$ is continuous on the positive real line, $\mathbb{R}_{x>0}$. Additionally, $\delta(x)$ trivially decays faster than exponentially and has integrable derivative on $\mathbb{R}_{x>0}$. By these properties of the delta function, we may follow in the fashion of Brydges et. al. in [1]. In particular, we take the following integral representation of the delta function:

$$\delta(s_n^2 - 1) = \frac{1}{2\pi} \int_{\Gamma} da_n \exp\left[\mathbf{i} a_n \left(s_n^2 - 1\right)\right], \tag{5.4}$$

where $\Gamma$ is the contour $\mathrm{Im}(a) = -\lambda$, with $\lambda$ large and positive so that *lemmas 5.2* and *5.3* may be used. Substitution into Eq. (5.3) yields

$$\begin{aligned}
\mathcal{Z} &= \int_{\Omega'} \int_{\Gamma^{N_s}} \prod_{n=0}^{N_S-1} ds_n da_n \frac{e^{-\mathbf{i}a_n}}{2\pi} \exp\left[-\sum_{k=0}^{N_S-1} \mathbf{i} a_k s_k^2\right] \exp\left[-\mathcal{H}(\boldsymbol{v}, \boldsymbol{h})\right] \\
&= \int_{\Gamma^{N_s}} \prod_{n=0}^{N_S-1} da_n \frac{e^{-\mathbf{i}a_n}}{2\pi} \int_{\Omega'} \prod_{n=0}^{N_S-1} ds_n \exp\left[-\sum_{\substack{s_k \in \boldsymbol{v} \\ s_l \in \boldsymbol{h}}} s_k w_{kl} s_l - \sum_{k=0}^{N_S-1} \mathbf{i} a_k s_k^2\right] \\
&= \int_{\Gamma^{N_s}} \prod_{n=0}^{N_S-1} da_n \frac{e^{-\mathbf{i}a_n}}{2\pi} \int_{\Omega'} \prod_{n=0}^{N_S-1} ds_n \exp\left[\sum_{k,l \in L_S} s_k (w_{kl} - \mathbf{i} a_k \delta_{kl}) s_l\right] \\
&= \int_{\Gamma^{N_s}} \prod_{n=0}^{N_S-1} da_n \frac{e^{-\mathbf{i}a_n}}{2\pi} \int_{\Omega'} d\boldsymbol{s} \exp\left[-\frac{1}{2} \boldsymbol{s}^T (2\mathbf{i}\boldsymbol{a} - \boldsymbol{W}) \boldsymbol{s}\right]
\end{aligned} \tag{5.5}$$

where $L_s = \{0, 1, ..., N_s\}$, matrix $\boldsymbol{a}$ is defined such that $\boldsymbol{a}_{kl} = a_k \delta_{kl}$ and $\mathbf{W}$ is the symmetric adjacency matrix for the graph of our RBM. Specifically, $\mathbf{W}$ has the

following properties:

$$\mathbf{W}_{ij} = \begin{cases} w_{ij} & \text{if } \boldsymbol{s}_i \text{ and } \boldsymbol{s}_j \text{ are in different layers} \\ 0 & \text{if } \boldsymbol{s}_i \text{ and } \boldsymbol{s}_j \text{ are in the same layer} \end{cases} \tag{5.6}$$

We now observe that the above integrals are Gaussian over $\mathbf{s}$ and can be evaluated to give us

$$\mathcal{Z} = \int_{\Gamma^{N_S}} \prod_{n=0}^{N_S-1} da_n \frac{e^{-\mathbf{i}a_n}}{2\pi} \det{}^{\frac{-1}{2}}[2\boldsymbol{ia} - \boldsymbol{W}]. \tag{5.7}$$

We now make a bit of a detour to introduce the following lemmas of Brydges, Frölich, and Spencer found in [1] to rewrite $\det^{-1}(2\mathbf{i}\boldsymbol{a} - \boldsymbol{W})$ in terms of random walks on $L_S$.

**Lemma 5.2.** *If $\boldsymbol{M}$ is a real, symmetric finite-dimensional matrix and $\boldsymbol{D}$ is a diagonal matrix of the same dimension, then*

$$\left[(\boldsymbol{D} - \boldsymbol{M})^{-1}\right]_{i,j} = \sum_{N=0}^{\infty} \sum_{\substack{\omega^N : i \to j \\ \omega^N \subset L}} \left( \prod_{r \in \omega^N} M_r \right) \prod_{k \in L} (d_k)^{-n(k\,|\,\omega^N)} \tag{5.8}$$

*where $r$ refers to an ordered pair (also called "step") in the random walk $\omega$.*

*Proof.* By making use of the Neumann series of $(D - M)^{-1}$, we have the expansion

$$[\boldsymbol{D} - \boldsymbol{M}]^{-1} = \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{M}\boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{M}\boldsymbol{D}^{-1}\boldsymbol{M}\boldsymbol{D}^{-1} + \cdots \tag{5.9}$$

We can now look at any single term in the right hand of the above expansion. For example, consider the fourth term, $\boldsymbol{D}^{-1}\boldsymbol{M}\boldsymbol{D}^{-1}\boldsymbol{M}\boldsymbol{D}^{-1}\boldsymbol{M}\boldsymbol{D}^{-1}$. The explicit expression for any entry in this matrix is

$$\begin{aligned} \left[\boldsymbol{D}^{-1}\boldsymbol{M}\boldsymbol{D}^{-1}\boldsymbol{M}\boldsymbol{D}^{-1}\boldsymbol{M}\boldsymbol{D}^{-1}\right]_{i_1,i_4} &= \sum_{i_2,i_3 \in L} D^{-1}_{i_1} M_{i_1,i_2} D^{-1}_{i_2} M_{i_2,i_3} D^{-1}_{i_3} M_{i_3,i_4} D^{-1}_{i_4} \\ &= \sum_{\substack{\omega^4 : i_1 \to i_4 \\ \omega^N \subset L}} \left( \prod_{r \in \omega^N} M_r \right) \prod_{k \in L} (d_k)^{-n(k\,|\,\omega^N)} \end{aligned} \tag{5.10}$$

Similarly, a term in which $D^{-1}$ appears $N$ times will fulfill the same equation, with the difference that the sum will be over all random walks of length $N$ that start and end at the given subindexes. Therefore, the lemma follows by applying equation (5.10) to each term in the right hand side of equation (5.9) and adding them all. □

**Lemma 5.3.** *If $\boldsymbol{M}$ is a real, symmetric finite-dimensional matrix and $\boldsymbol{D}$ is a diagonal matrix of the same dimension, then*

$$\det\left[(\boldsymbol{D}-\boldsymbol{M})^{-1}\right]=\left[\prod_{i\in L}d_i\right]^{-1}\exp\left[\sum_{N=1}^{\infty}\sum_{\underline{\omega}^N\subset L}M_{\underline{\omega}^N}\prod_{k\in L}\left(d_k\right)^{-n(k\,|\,\underline{\omega}^N)}\right] \tag{5.11}$$

*Proof.* The following computation gives the desired result.

$$\begin{aligned}
\det\left[(\boldsymbol{D}-\boldsymbol{M})^{-1}\right] &= \det[\boldsymbol{D}^{-1}]\det{}^{-1}[(\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{M})] \\
&= \det[\boldsymbol{D}^{-1}]\exp\left[-\mathrm{tr}\log(\boldsymbol{I}-\mathbf{D^{-1}M})\right] \\
&= \det[\boldsymbol{D}^{-1}]\exp\left[\sum_{N=1}^{\infty}\frac{1}{N}\mathrm{tr}(\mathbf{D^{-1}M})^N\right] \\
&= \det[\boldsymbol{D}^{-1}]\exp\left[\sum_{N=1}^{\infty}\frac{1}{N}\sum_{i\in L}\sum_{\substack{\omega^N:i\to i\\ \omega^N\subset L}}\left(\prod_{r\in\omega^N}M_r\right)\prod_{k\in L}d_k^{-n(k\,|\,\omega^N)}\right] \\
&= \det[\boldsymbol{D}^{-1}]\exp\left[\sum_{N=1}^{\infty}\sum_{\underline{\omega}^N\subset L}\left(\prod_{r\in\underline{\omega}^N}M_r\right)\prod_{k\in L}\left(d_k\right)^{-n(k\,|\,\underline{\omega}^N)}\right]
\end{aligned} \tag{5.12}$$

□

We can now apply *Lemma 5.3* for $L=L_s$, $A=2\mathbf{i}\boldsymbol{a}$ and $\boldsymbol{M}=W$, obtaining

$$\begin{aligned}
\det{}^{\frac{-1}{2}}\left[2\mathbf{i}\boldsymbol{a}-\boldsymbol{W}\right] &= \left(\det\left[(2\mathbf{i}\boldsymbol{a}-\boldsymbol{W})\right]^{-1}\right)^{\frac{-1}{2}} \\
&= \prod_{k\in L_s}(2\mathbf{i}a_k)^{\frac{-1}{2}}\exp\left[\sum_{N=1}^{\infty}\sum_{\underline{\omega}^N\subset L_s}\left(\prod_{r\in\underline{\omega}^N}(W)_r\right)\prod_{k\in L_s}(2\mathbf{i}a_k)^{-n(k\,|\,\underline{\omega}^N)}\right]^{\frac{1}{2}} \\
&= \prod_{k\in L_s}(2\mathbf{i}a_k)^{\frac{-1}{2}}\exp\left[\frac{1}{2}\sum_{N=1}^{\infty}\sum_{\underline{\omega}^N\subset L_s}(W)_{\underline{\omega}^N}\prod_{k\in L_s}(2\mathbf{i}a_k)^{-n(k\,|\,\underline{\omega}^N)}\right].
\end{aligned} \tag{5.13}$$

It is important to notice that, because the entries of $W$ representing the weights between 2 hidden or 2 visible variables are zero, it is enough to sum over all walks whose vertices alternate between the hidden and visible layer.
After substituting this last expression, Eq (5.7) becomes

$$\mathcal{Z}=\int_{\Gamma^{N_s}}\prod_{i\in L_s}da_i(2\mathbf{i}a_k)^{\frac{-1}{2}}\frac{e^{-\mathbf{i}a_i}}{2\pi}\exp\left[\frac{1}{2}\sum_{N=1}^{\infty}\sum_{\underline{\omega}^N\subset L_s}(W)_{\underline{\omega}^N}\prod_{k\in L_s}(2\mathbf{i}a_k)^{-n(k\,|\,\underline{\omega}^N)}\right] \tag{5.14}$$

Finally, using the Taylor expansion of the exponential the theorem follows.

□

## 6. RANDOM WALK REPRESENTATION OF $\mathbb{E}^\mu$

Recall from section 3 that we have an explicit expression for updating the weight $w_{ij}$ at any step of the gradient descent algorithm in terms of expectation values. We now provide a formulation of each term in Eq. (3.9) in terms of random walks.

**Proposition 6.1.**

$$\mathrm{E}^\mu[v_i^\mu h_j] = \frac{(2\pi)^{-N_H}\, v_i^\mu}{\mathcal{Z}\; P(\boldsymbol{v}^\mu)} \sum_{\omega^1 \in L_s} \mathcal{Y}_j^\mu(\omega^1) \tag{6.1}$$

*where $\omega$ is summed over all random walks that begin at $h_j$,*

$$\mathcal{Y}_j(\omega) \equiv \sum_{n=0}^{\infty} \frac{1}{n!}\left(\frac{1}{2}\right)^n \sum_{\underline{\omega}_1,\ldots,\underline{\omega}_n \subset L_H} W_\omega^\mu \prod_{k=1}^{n} (-2\boldsymbol{i}\boldsymbol{a}\boldsymbol{O})_{\underline{\omega}_k} \exp\left[-U_j(\underline{\omega}_1,\underline{\omega}_2,\ldots,\underline{\omega}_n)\right]$$

$$\exp\left[-U_j(\underline{\omega}_1,\underline{\omega}_2,\ldots,\underline{\omega}_n)\right] = \prod_{k\in L_H} \int_{\Gamma^{N_H}} da_k e^{-\boldsymbol{i}\boldsymbol{a}_i} (2\boldsymbol{i}\boldsymbol{a}_k d_k)^{-n(k|\underline{\omega}_1,\ldots,\underline{\omega}_n)} (2\boldsymbol{i}\boldsymbol{a}_j)^{-1} \tag{6.2}$$

$$n(k|\omega_1,\ldots,\omega_n) = \frac{1}{2} + n(k|\omega_1) + \ldots + n(k|\omega_n),$$

*$W^\mu$ is the symmetric matrix such that*

$$W_{pq}^\mu = \begin{cases} w_{pq} v_q^\mu & \text{if } \; p > q \\ w_{pq} v_p^\mu & \text{otherwise,} \end{cases} \tag{6.3}$$

*and $\boldsymbol{O}$ is a symmetric matrix with zeroes along the diagonal.*

There are several key features in the above proposition. The first is that in eq. (6.1) we are summing over all loops of length one beginning at $h_j$. Then as in our definition of $\mathcal{Y}(\omega)$, we take a path $\omega$ and append loops to it. These new loops are found in a new graph defined by $\boldsymbol{O}$ containing all necessary information about our RBM through some base change matrix.

*Proof.* We know

$$p(\boldsymbol{h}|\boldsymbol{v}) = \frac{p(\boldsymbol{h},\boldsymbol{v})}{p(\boldsymbol{v})} = \frac{\frac{1}{\mathcal{Z}} e^{-\mathcal{H}(\boldsymbol{v},\boldsymbol{h})}}{p(\boldsymbol{v})} \tag{6.4}$$

When we consider the $\mu^{th}$ data sample, the entire visible layer is fixed. In particular, $p(\boldsymbol{h}|\boldsymbol{v}^\mu)$ is a probability distribution for $\boldsymbol{h}$ and we can consider $v_i^\mu$ as a constant when taking the expectation value $\mathbb{E}^\mu[v_i^\mu h_j]$. Once again, we introduce delta functions for the purpose of performing integration. Our sample space for the hidden variables $\boldsymbol{h}$ becomes $\Omega'_H = \mathbb{R}^{\mathrm{N_H}}$ and we get that

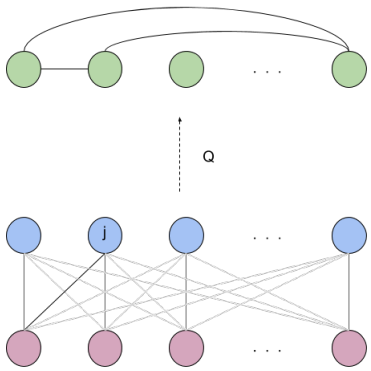


FIGURE 3. Example of a term that contributes to the sum for the correlation function $\mathbb{E}^{\mu}[v_i^{\mu} h_j]$. The connected red and blue nodes represent our original RBM. Notice that we have a path from the visible layer going to the $j^{th}$ hidden vertex. The green nodes denote the hidden layer after we perform the change of variables through some base change matrix, $Q$. In this new hidden layer, we find loops.

$$\begin{aligned}\mathbb{E}^{\mu}[v_i^{\mu} h_j] &= \sum_{\boldsymbol{h}\in\Omega_H} \frac{e^{-\mathcal{H}(\boldsymbol{v^{\mu}},\boldsymbol{h})}}{\mathcal{Z}\; p(\boldsymbol{v^{\mu}})} v_i^{\mu} h_j \\ &= \frac{v_i^{\mu}}{\mathcal{Z}\; p(\boldsymbol{v}^{\mu})} \int_{\Omega'_H} \prod_{n=0}^{N_H-1} dh_n \delta(h_n^2 - 1) \exp\left[-\mathcal{H}(\boldsymbol{h}, \boldsymbol{v}^{\mu})\right] h_j.\end{aligned} \tag{6.5}$$

Continuing as was done for the partition function, we see that the integral over $\Omega'_H$ becomes

$$\begin{aligned}&\int_{\Omega'_H} \prod_{n=0}^{N_H-1} dh_n \delta(h_n^2 - 1) \exp\left[-\mathcal{H}(\boldsymbol{h}, \boldsymbol{v}^{\mu})\right] h_j \\ &= \int_{\Gamma^{N_H}} \prod_{n=0}^{N_H-1} da_n \frac{e^{-\mathbf{i}a_n}}{2\pi} \int_{\Omega'_H} \prod_{k=0}^{N_H-1} dh_k \exp\left[-\mathcal{H}(\boldsymbol{v}, \boldsymbol{h}) - \mathbf{i}\sum_{r=0}^{N_H-1} a_r \left(h_r\right)^2\right] h_j\end{aligned} \tag{6.6}$$

By a process of completing the square, we transform the integral over the hidden variables into a Gaussian form,

$$
\begin{aligned}
&\int_{\Omega'_H} \prod_{k=0}^{N_H-1} dh_k \exp\left[-\mathcal{H}(\boldsymbol{v},\boldsymbol{h}) - \mathbf{i}\sum_{r=0}^{N_H-1} a_r \left(h_r\right)^2\right] h_j \\
&= \int_{\Omega'_H} d\boldsymbol{h} \exp\left[\sum_{k=0}^{N_H-1} h_k \left(\sum_{l=0}^{N_V-1} W_{k,l} v_l^\mu\right) - \sum_{r=0}^{N_H-1} \mathbf{i} a_r h_r^2\right] h_j \\
&= \int_{\Omega'_H} d\boldsymbol{h} \exp\left[\sum_{k=0}^{N_H-1} \left(h_k \sum_{l=0}^{N_V-1} W_{k,l} v_l^\mu - \mathbf{i} a_k h_k^2\right)\right] h_j \\
&= \int_{\Omega'_H} d\boldsymbol{h} \exp\left[\sum_{k=0}^{N_H-1} -\mathbf{i} a_k \left(-2h_k \frac{\sum_{l=0}^{N_V-1} W_{k,l} v_l^\mu}{2\mathbf{i} a_k} + h_k^2\right)\right] h_j \\
&= \int_{\Omega'_H} d\boldsymbol{h} \exp\left[\sum_{k=0}^{N_H-1} -\mathbf{i} a_k \left(h_k - \frac{\sum_{l=0}^{N_V-1} W_{k,l} v_l^\mu}{2\mathbf{i} a_k}\right)^2 + \mathbf{i} a_k \left(\frac{\sum_{l=0}^{N_V-1} W_{k,l} v_l^\mu}{2\mathbf{i} a_k}\right)^2\right] h_j \\
&= \int_{\Omega'_H} d\boldsymbol{h} \exp\left[\sum_{k=0}^{N_H-1} -\mathbf{i} a_k \left(h_k - b_k^\mu\right)^2\right] \prod_{k=0}^{N_H-1} \exp\left[\mathbf{i} a_k \left(b_k^\mu\right)^2\right] h_j
\end{aligned}
\tag{6.7}
$$

where $b_k^\mu = \frac{\sum_{l=0}^{N_V-1} W_{k,l} v_l^\mu}{2\mathbf{i} a_k}$. Define $\mathbf{B}$ as the diagonal matrix such that $\mathbf{B}_{q,q} = \exp\left[\frac{-1}{2}\mathbf{i} a_p \left(b_p^\mu\right)^2\right]$. Then,

$$
\prod_{k=0}^{N_H-1} \exp\left[\mathbf{i} a_k \left(b_k^\mu\right)^2\right] = \det{}^{\frac{-1}{2}}\left[\mathbf{B}\right] \tag{6.8}
$$

and we may move this factor of the determinant out in front of the integral. After substituting $\mathbf{h}$ with $\boldsymbol{\sigma} = \mathbf{h} - \mathbf{b}^\mu$ we get

$$
\begin{aligned}
&\det{}^{\frac{-1}{2}}\left[\mathbf{B}\right] \int_{\Omega'_H} d\boldsymbol{h} \exp\left[\sum_{k=0}^{N_H-1} -\mathbf{i} a_k \left(h_k - b_k^\mu\right)^2\right] h_j \\
&= \det{}^{\frac{-1}{2}}\left[\mathbf{B}\right] \int_{\mathbb{R}^{\mathrm{N_H}}} d\boldsymbol{\sigma} \exp\left[\sum_{k=0}^{N_H-1} -\mathbf{i} a_k \left(\sigma_k\right)^2\right] \left(\sigma_j + b_j^\mu\right) \\
&= \det{}^{\frac{-1}{2}}[\mathbf{B}] \int_{\mathbb{R}^{\mathrm{N_H}}} d\boldsymbol{\sigma} \exp\left[\sum_{k=0}^{N_H-1} -\mathbf{i} a_k \left(\sigma_k\right)^2\right] \sigma_j + \det{}^{\frac{-1}{2}} \int_{\mathbb{R}^{\mathrm{N_H}}} d\boldsymbol{\sigma} \exp\left[\sum_{k=0}^{N_H-1} -\mathbf{i} a_k \left(\sigma_k\right)^2\right] b_j^\mu
\end{aligned}
\tag{6.9}
$$

In the first term, recall that the probability distribution may be factorized among the hidden variables. We get $N_H$ independent Gaussian integrals. In particular, isolating the $j^{th}$ integral we see that the entire first term vanishes. This is because $\exp[-\mathbf{i} a_j (\sigma_j)^2]\sigma_j$ is an odd function in $\sigma_j$. Thus, we only have to consider the second

term

$$
\begin{aligned}
&= \det{}^{\frac{-1}{2}}\left[\mathbf{B}\right] \int_{\mathbb{R}^{N_H}} d\boldsymbol{\sigma} \exp\left[\sum_{k,l\in L_e^H} -\sigma_k \mathbf{i} a_k \delta(l,k)\sigma_l\right] \\
&= \det{}^{\frac{-1}{2}}\left[\mathbf{B}\right] b_j^{\mu} \int_{\mathbb{R}^{N_H}} d\boldsymbol{\sigma} \exp\left[-\frac{1}{2}\boldsymbol{\sigma}^T 2\mathbf{i}\mathbf{a}\boldsymbol{\sigma}\right]
\end{aligned}
\tag{6.10}
$$

where $\mathbf{ia}_{m,n} = \mathbf{i}a_m\delta_{m,n}$. Then, evaluating the Gaussian integral, we get

$$
\begin{aligned}
&\det{}^{\frac{-1}{2}}\left[\mathbf{B}\right] b_j^{\mu} \int_{\mathbb{R}^{N_H}} d\boldsymbol{\sigma} \exp\left[-\frac{1}{2}\boldsymbol{\sigma}^T 2\mathbf{i}\mathbf{a}\boldsymbol{\sigma}\right] \\
&= \det{}^{\frac{-1}{2}}\left[\mathbf{B}\right] b_j^{\mu} \det{}^{\frac{-1}{2}}\left[2\mathbf{i}\mathbf{a}\right]
\end{aligned}
\tag{6.11}
$$

Let $\mathbf{Q}$ be a $N_H$ x $N_H$ symmetric, orthogonal and non diagonal matrix. And let $\mathbf{QBQ}^T = \mathbf{D} + \mathbf{O}$ where $\mathbf{D}$ is a diagonal matrix and $\mathbf{O}$ has zeroes in the diagonal. Then, Equation (6.11) can then be rewritten as

$$
\begin{aligned}
&b_j^{\mu} \det{}^{\frac{-1}{2}}\left[\mathbf{B}\right] \det{}^{\frac{-1}{2}}\left[2\mathbf{i}\mathbf{a}\right] \\
&= b_j^{\mu} \det{}^{\frac{-1}{2}}\left[\mathbf{Q}\mathbf{B}\mathbf{Q}^T\right] \det{}^{\frac{-1}{2}}\left[2\mathbf{i}\mathbf{a}\right] \\
&= b_j^{\mu} \det{}^{\frac{-1}{2}}\left[\mathbf{D}+\mathbf{O}\right] \det{}^{\frac{-1}{2}}\left[2\mathbf{i}\mathbf{a}\right] \\
&= b_j^{\mu} \det{}^{\frac{-1}{2}}\left[2\mathbf{i}\mathbf{a}\mathbf{D} + 2\mathbf{i}\mathbf{a}\mathbf{O}\right] \\
&= b_j^{\mu} \det{}^{\frac{-1}{2}}\left[(2\mathbf{i}\mathbf{a}\mathbf{D}) - (-2\mathbf{i}\mathbf{a}\mathbf{O})\right]
\end{aligned}
\tag{6.12}
$$

Since $(2\mathbf{iaD})$ is a diagonal matrix and $(-2\mathbf{iaO})$ is symmetric with zeroes in the diagonal entries, we can apply *Lemma* 5.3, obtaining

$$
\begin{aligned}
&b_j^{\mu} \det{}^{\frac{-1}{2}}\left[(2\mathbf{i}\mathbf{a}\mathbf{D}) - (-2\mathbf{i}\mathbf{a}\mathbf{O})\right] \\
&= b_j^{\mu}\left[\prod_{p\in L}(2\mathbf{i}a_p d_p)\right]^{-\frac{1}{2}} \exp\left[\frac{1}{2}\sum_{q=1}^{\infty}\sum_{\underline{\omega}^q\subset L_H}(-2\mathbf{i}\mathbf{a}\mathbf{O})_{\underline{\omega}^q}\prod_{k\in L}(2\mathbf{i}a_k d_k)^{-n(k\,|\,\underline{\omega}^q)}\right]
\end{aligned}
\tag{6.13}
$$

And so we can conclude

$$
\begin{aligned}
\mathrm{E}[v_i^{\mu} h_j]^{\mu} = \frac{v_i^{\mu}}{\mathcal{Z}\ P(\boldsymbol{v}^{\mu})} &\int_{\Gamma^{N_H}} \prod_{k=0}^{N_H-1} da_k \frac{(2\mathbf{i}a_k d_k)^{-\frac{1}{2}} e^{-\mathbf{i}a_k} b_j^{\mu}}{2\pi} \\
&\cdot \exp\left[\frac{1}{2}\sum_{q=1}^{\infty}\sum_{\underline{\omega}^q\subset L_H}(-2\mathbf{i}\mathbf{a}\mathbf{O})_{\underline{\omega}^q}\prod_{k=0}^{L_H-1}(2\mathbf{i}a_k d_k)^{-n(k\,|\,\underline{\omega}^q)}\right]
\end{aligned}
\tag{6.14}
$$

Plugging in the definition of $b_j^\mu$ and bringing out the numerator in (6.14), we have

$$\mathrm{E}[v_i^\mu h_j]^\mu = \frac{v_i^\mu \sum_k W_{kj} v_j^\mu}{\mathcal{Z}\ P(\boldsymbol{v}^\mu)} \int_{\Gamma^{N_H}} \prod_{k=0}^{N_H-1} da_k (2\mathbf{i}a_k d_k)^{-\frac{1}{2}} \frac{e^{-\mathbf{i}a_k}}{2\pi} \frac{1}{2\mathbf{i}a_j} \cdot \exp\left[\frac{1}{2}\sum_{q=1}^{\infty} \sum_{\underline{\omega}^q \subset L_H} (-2\mathbf{i}\mathbf{a}\mathbf{O})_{\underline{\omega}^q} \prod_{k\in L_H} (2\mathbf{i}a_k d_k)^{-n(k\,|\,\underline{\omega}^q)}\right] \tag{6.15}$$

Again, the proposition follows after expanding the exponential.

□

## 7. random walk representation of $\mathbb{E}$

Now that we have a representation for the clamped correlation function, we press on to find a formulation of the true correlation functions in terms of random walks. In particular, we will find that

$$\mathbb{E}[\boldsymbol{v}_i \boldsymbol{h}_j] = \frac{\sum_{\omega^{\boldsymbol{v}_i,\boldsymbol{h}_j}} \mathcal{X}(\omega^{\boldsymbol{v}_i \boldsymbol{h}_j})}{\mathcal{Z}}. \tag{7.1}$$

where

$$\mathcal{X}(\omega) \equiv \sum_{n=0}^{\infty} \frac{1}{n!}\left(\frac{1}{2}\right)^n \sum_{\underline{\omega}_1,\underline{\omega}_2,\ldots,\underline{\omega}_n} \left(W_\omega \prod_{k=1}^{n} W_{\underline{\omega}_k}\right) \exp[-F(\underline{\omega}_1,\ldots,\underline{\omega}_n|\omega)]. \tag{7.2}$$

where $\exp[-F(\underline{\omega}_1,\ldots,\underline{\omega}_n|\omega)] = \exp[-F(\underline{\omega}_1,\ldots,\underline{\omega}_n,\omega)]$. Before delving into the proof of the above statement, we take a moment to describe the walks that appear in the above sum. In the expression for $\mathbb{E}[\boldsymbol{v}_i \boldsymbol{h}_j]$ we are summing over paths whose endpoints are $v_i$ and $h_j$. Notice that the each term $\mathcal{X}(\omega)$ looks nearly identical to the expression we found for $\mathcal{Z}$. The main difference is that when we sum over the products of $n$ loops, there is an additional recurring factor $W_\omega$ corresponding to the path found in the argument for $\mathcal{X}$. More so, each term $\mathcal{X}(\omega)$ is damped by additional factors of $(2ia_j)^{-n(i,\omega)}$. This gives a nice geometric interpretation. In the numerator of the expression for $\mathbb{E}[\boldsymbol{v}_i \boldsymbol{h}_j]$, we are effectively summing over all random paths whose endpoints exactly cover the sites $v_i$ and $h_j$. In the denominator, we have a sum over all random loops in our lattice system.

*Proof.* We consider a more general case where we do not distinguish between hidden and visible vertices. Specifically, instead of having visible and hidden spins encoded by $\boldsymbol{v}$ and $\boldsymbol{h}$, we have a single collection of spins denoted $\boldsymbol{s}$, where $s_k$ denotes the $k^{th}$ spin vertex. Starting with the integral representation of the expectation value

$\mathbb{E}[s_i s_j]$.

$$\mathbb{E}[s_i s_j] = \frac{1}{\mathcal{Z}} \int_{\Omega} \int_{\mathbb{R}^{N_S}} \prod_{n=0}^{N_S-1} ds_n da_n \frac{e^{-\mathbf{i}a_n}}{2\pi} \exp\left[ - \sum_{k=0}^{N_S-1} \mathbf{i} a_k s_k^2 \right] \exp\left[-\mathcal{H}(\boldsymbol{v}, \boldsymbol{h})\right] s_i s_j \tag{7.3}$$

We can proceed as we did with the partition function in section 5, using the fact that

$$\int_{\mathbb{R}^{N_s}} d\boldsymbol{s} \;\; e^{\frac{1}{2}\boldsymbol{s}^T \boldsymbol{M} \boldsymbol{s}} \;\; s_i s_j = \det{}^{-\frac{1}{2}} \left[\boldsymbol{M}\right] \exp\left[ \frac{1}{2} \partial_{\boldsymbol{s}}^T \boldsymbol{M}^{-1} \partial_{\boldsymbol{s}} \right] s_i s_j \Big|_{\boldsymbol{s}=0} \tag{7.4}$$

where $\partial_{\boldsymbol{s}} = (\frac{\partial}{\partial s_1}, ..., \frac{\partial}{\partial s_{N_s-1}})^T$ [1], to get

$$\begin{aligned}
&\int_{\mathbb{R}^{N_s}} d\boldsymbol{s} \;\; \exp\left[ \frac{1}{2} \boldsymbol{s}^T (2\boldsymbol{i}\boldsymbol{a} - \boldsymbol{W}) \boldsymbol{s} \;\; \right] s_i s_j \\
&= \det\left[2\boldsymbol{i}\boldsymbol{a} - \boldsymbol{W}\right]^{-1} (2\boldsymbol{i}\boldsymbol{a} - \boldsymbol{W})_{ij}^{-1} \\
&= \prod_{k \in L_s} (2\mathbf{i}a_k)^{\frac{-1}{2}} \exp\left[ \frac{1}{2} \sum_{N=1}^{\infty} \sum_{\underline{\omega}^N \subset L_s} (W)_{\underline{\omega}^N} \prod_{k \in L_s} (2\mathbf{i}a_k)^{-n(k\,|\,\underline{\omega}^N)} \right] \\
&\qquad\qquad \cdot \sum_{N=1}^{\infty} \sum_{\substack{\omega^N : i \to j \\ \omega^N \subset L_s}} (W)_{\omega^N} \prod_{k \in L_s} (2\mathbf{i}a_k)^{-n(k\,|\,\omega^N)}
\end{aligned} \tag{7.5}$$

where the last expression comes from applying *lemma* 5.2 and *lemma* 5.3. Once again, the result follows from applying the Taylor expansion of the exponential in the last expression. □

## 8. FURTHER WORK

Throughout the above discussion we have just provided a setup to begin exploring vital results. From our equation on updating $\Delta w_{ij}$, we have that

$$\Delta w_{ij} = -\delta\{\frac{1}{N_T} \sum_{\mu=1}^{N_T} \mathbb{E}^{\mu}[v_i h_j] - \mathbb{E}[v_i h_j]\} = \delta\{\mathbb{E}[v_i h_j] - \frac{1}{N_T} \sum_{\mu=1}^{N_T} \mathbb{E}^{\mu}[v_i h_j]\}. \tag{8.1}$$

As we found above, the parity of any term $\mathbb{E}^{\mu}[v_i h_j]$ may be positive or negative depending on the value assigned to $v_i$. Then a simple bound for $\Delta w_{ij}$ is

$$\Delta w_{ij} \leq \delta\Big( |\mathbb{E}[v_i h_j]| + |\frac{1}{N_T} \sum_{\mu=1}^{N_T} \mathbb{E}^{\mu}[v_i h_j]| \Big) \leq \delta\Big( |\mathbb{E}[v_i h_j]| + \frac{1}{N_T} \sum_{\mu=1}^{N_T} |\mathbb{E}^{\mu}[v_i h_j]| \Big) \tag{8.2}$$

In [1], Brydges et. al. compute the following upper bound on the correlation functions $\mathbb{E}[v_i h_j]$ in an Ising-type model.

$$\mathbb{E}[v_i h_j] \leq \left( \sum_{\omega : i \to j} \prod_{k \in L} (2\nu)^{-n(k|\omega)} \right) \tag{8.3}$$

In the above equation, $\nu$ is the dimension of the graph. This inequality may be directly carried over for the expectation value $\mathbb{E}[v_i h_j]$. Finding an analogous bound on the expectation $\mathbb{E}^{\mu}[v_i h_j]$ will complete the bound on $\Delta w_{ij}$. Completing this inequality will provide a bound on how fast learning can occur.

Another direction to go from here is to more formally establish the connection between learning an RBM and the renormalization group in an Ising-type lattice. From our discussion on $\mathbb{E}^{\mu}$, we introduced orthogonal matrices so that we retrieve random walks on an equivalent hidden layer as prescribed by the matrix $\mathbf{O}$. Observe that the following block matrix equation must have a solution.

$$\begin{pmatrix} Q_1 & R \\ R^T & Q_2 \end{pmatrix} \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \begin{pmatrix} Q_1^T & R \\ R^T & Q_2^T \end{pmatrix} = \begin{pmatrix} 0 & R' \\ R'^T & 0 \end{pmatrix} \tag{8.4}$$

The solution exists because we can suppose that the matrix we are conjugating with serves as a base-change matrix. Given a basis $\mathbf{B} = (b_1, b_2, ..., b_n)$, there exists a base change operator $T$ that permutes the entries of $\mathbf{B}$. For example, we could have that $T\mathbf{B} = (b_m, b_{m+1}, ...b_n, b_1, b_2, ..., b_{m-1})$. Through this change of basis we are given new hidden and visible layers, allowing us to analyze random walks all over again. More so, by allowing us to create new vertices $x_k$ which are independent from the rest of the graph, we may assume that the hidden layer has $2^M$ vertices for some $M \in \mathbb{N}$. Then by virtue of Eq. (8.4), we may designate values for the new hidden variables via some sampling procedure, giving us new correlation values $\mathbb{E}[v_i^{new} h_j^{new}]$. This process of fixing some fraction of your current graph to create a new graph with similar properties could admit a connection to the practice of decimation in the renormalization group.

One could also follow Aizenmann's analysis on random paths in the renormalization group in [3] where he studies intersection properties of random walks and their implications on the renormalization group. The task here would be to provide an analogous survey on (possibly $n$-layer) RBMs. Once this has been done, it would be natural to take limits where the number of layers you examine as well as number of nodes per layer go to infinity and state results about infinitely deep and infinitely wide neural networks.

## 9. ACKNOWLEDGEMENTS

We would like to thank Mason Biamonte for the many hours of mentoring and guidance throughout this project. We also want to thank Cris Negron for his advice

regarding general practices in the mathematical community. Finally, we thank Professors David Jerison and Ankur Moitra as well as Dr. Slava Gerovitch for putting together the SPUR/SPUR+ program and giving us this opportunity.

## References

1. Brydges, D., Frölich, J., and Spencer, T., *The Random Walk Representation of Classical Spin Systems and Correlation Inequalities*, Commun. Math. Phys., vol. 83, 1982.
2. Fischer, A., and Igel, C., *An Introduction to Restricted Boltzmann Machines*, 2012
3. Aizenmann, *The Intersection of Brownian Paths as a Case Study of a Renormalization Group Method for Quantum Field Theory*, Commun. Math. Phys., vol. 97, 1985.
4. Brydges, D., *What is a Quantum Field Theory?*, Am. Math. Soc., vol 8, 1983.
5. Mehta, P., and Schwab, D., *An exact mapping between the Variational Renormalization Group and Deep Learning*, arXiv preprint, arXiv:1410.3831, 2014.
6. Symanzik, K. *Euclidean quantum field theory.In:Local quantum theory.* Jost, R. (ed.) New York, London: Academic Press, 1969.
7. Kupiainen, A. *Introduction to the Renormalization Group*, Lecture Notes, 2014.
8. Le Roux, Nicolas and Bengio, Yoshua, *Deep Belief Networks Are Compact Universal Approximators* Neural Computation, 2010.
9. Montufar, G., Ay., N, *Refinements of Universal Approximation Results for Deep Belief Networks and Restricted Boltzmann Machines*, arXiv preprint, arXiv:1005.1593, 2010.