Journal of Machine Learning Research 17 (2016) 1-21 Submitted 9/15; Revised 6/16; Published 8/16

Universal Approximation Results for the Temporal Restricted Boltzmann Machine and the Recurrent Temporal Restricted Boltzmann Machine

Simon Odense SIODENSE@UVIC.CA
Roderick Edwards EDWARDS@UVIC.CA
Department of Mathematics
University of Victoria
Victoria, BC, 3800 Finnerty Rd, Canada

Editor: Yoshua Bengio

Abstract

The Restricted Boltzmann Machine (RBM) has proved to be a powerful tool in machine learning, both on its own and as the building block for Deep Belief Networks (multi-layer generative graphical models). The RBM and Deep Belief Network have been shown to be universal approximators for probability distributions on binary vectors. In this paper we prove several similar universal approximation results for two variations of the Restricted Boltzmann Machine with time dependence, the Temporal Restricted Boltzmann Machine (TRBM) and the Recurrent Temporal Restricted Boltzmann Machine (RTRBM). We show that the TRBM is a universal approximator for Markov chains and generalize the theorem to sequences with longer time dependence. We then prove that the RTRBM is a universal approximator for stochastic processes with finite time dependence. We conclude with a discussion on efficiency and how the constructions developed could explain some previous experimental results.

Keywords: TRBM, RTRBM, machine learning, universal approximation

1. Introduction

Modeling temporal sequences has been an important problem in machine learning because of the natural time dependence in many data sets. The Restricted Boltzmann Machine (RBM), a type of probabilistic neural network, has become popular as a result of the use of an efficient learning algorithm called contrastive divergence (Hinton, 2002). In particular, its use in the construction of Deep Belief Networks (Hinton and Osindero, 2006) has led to widespread use in a number of machine learning tasks. One major drawback of the basic model, however, is the difficulty in using these models to capture temporal dependence in a data set. Several refinements of the model have attempted to combine the efficient statistical modeling of RBMs with the dynamic properties of Recurrent Neural Networks (Hinton and Osindero, 2006)(Taylor et al., 2006)(Le Roux and Bengio, 2008). These include the Temporal Restricted Boltzmann Machine (TRBM), the Recurrent Temporal Restricted Boltzmann Machine (RTRBM), and the Conditional Restricted Boltzmann Machine (CRBM). Boltzmann machines, and RBMs have been shown to be universal approximators for probability distributions on binary vectors (Freund and Haussler, 1991)(Younes, 1996)(Le Roux and

Bengio, 2008). Furthermore the related Deep Belief Networks have also been shown to be universal approximators even when each hidden layer is restricted to a relatively small number of hidden nodes (Sutskever and Hinton, 2010)(Le Roux and Bengio, 2010)(Montufar and Ay, 2011). The universal approximation of CRBMs follows immediately from that of Boltzmann machines (Montufar et al., 2014). The question we wish to address here is the universal approximation of stochastic processes by TRBMs and RTRBMs.

1.1 The Restricted Boltzmann Machine

An RBM defines a probability distribution over a set of binary vectors $x \in \{0,1\}^n = X$ as follows

$$P(v,h) = \exp(v^\top W h + c^\top v + b^\top h)/Z$$

where the set of binary vectors X is partitioned into visible and hidden units $X = V \times H$ and Z is the normalization factor, in other words $Z = \sum_{v,h} \exp(v^\top W h + c^\top v + b^\top h)$. This distribution is entirely defined by (W, b, c) and is referred to as a Boltzmann Distribution. We are generally concerned with the marginal distribution of the visible units. When we refer to the distribution of an RBM we are referring to the marginal distribution of its visible units. The marginal distribution of a single visible node is given by

$$P(v_i = 1|h) = \sigma\left(\sum_j w_{i,j} h_j + c_i\right)$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$. A similar equation holds for the hidden units. Variations of the RBM which use real-valued visible and hidden units (or mixes of the two) exist but will not be considered here.

1.2 Approximation

In order to measure how well one distribution approximates another we use the Kullback-Leibler divergence, which for discrete probability distributions is given by

$$KL(R||P) = \sum_v R(v) \log\left(\frac{R(v)}{P(v)}\right),$$

where v ranges over the sample space of R and P. It can be shown that for any $\epsilon > 0$, given a probability distribution R on V there is a Boltzmann Distribution given by an RBM P such that $KL(R||P) < \epsilon$ (Le Roux and Bengio, 2008)(Freund and Haussler, 1991). Our goal now is to prove the same result where R satisfies certain temporal dependency conditions and P is a TRBM.

2. Universal Approximation Results for the TRBM

A TRBM defines a probability distribution on a sequence $x^T = (x^{(0)}, ..., x^{(T-1)})$, $x^{(i)} \in \{0,1\}^n$, $x^{(i)} = (v^{(i)}, h^{(i)})$, given by

$$P(v^{(t)}, h^{(t)}|h^{(t-1)}) = \frac{\exp(v^{(t)\top} W h^{(t)} + c^\top v^{(t)} + b^\top h^{(t)} + h^{(t)\top} W' h^{(t-1)})}{Z(h^{(t-1)})},$$

$$P(v^T, h^T) = \left(\prod_{k=1}^{T-1} P(v^{(k)}, h^{(k)}|h^{(k-1)}) \right) P_0(v^{(0)}, h^{(0)}).$$

This distribution is defined by the same parameters as the RBM along with the additional parameters W'. The TRBM can be seen as an RBM with a dynamic hidden bias determined by $W'h^{(t-1)}$. The initial distribution, $P_0(v^{(0)}, h^{(0)})$, is the same as $P(v^{(t)}, h^{(t)}|h^{(t-1)})$ with $h^{(0)\top} b_{init}$ replacing $h^{(t)\top} W' h^{(t-1)}$ for some initial hidden bias b_{init}. Note that W' is not symmetric in general. We call the connections between $h^{(t-1)}$ and $h^{(t)}$ with weights in W' temporal connections.

2.1 Universal Approximation Results for the Basic TRBM

Our approximation results will deal with distributions which are time-homogeneous and have finite time dependence. These distributions can be written in the form

$$R(v^T) = \left(\prod_{k=m}^{T-1} R_1(v^{(k)}|v^{(k-1)}, ..., v^{(k-m)}) \right) R_0(v^{(0)}, ..., v^{(m-1)})$$

where R_1 is the transition probability and R_0 is the initial distribution. We first show that a TRBM can approximate a Markov chain (distributions of the above form with $m = 1$) for a finite number of time steps to arbitrary precision. We begin by proving a lemma. Here P_t is the marginal distribution of P over $(v^{(0)}, \ldots, v^{(t)})$. Similarly $R_{0,t}$ is the marginal distribution of R_0 over $(v^{(0)}, \ldots, v^{(t)})$.

Lemma 1: *Let R be a distribution on a finite sequence of length T of n-dimensional binary vectors that is time homogeneous with finite time dependence. Given a set of distributions $\mathbf{P}$ on the same sequences, if for every $\epsilon > 0$ we can find a distribution $P \in \mathbf{P}$ such that for every v^T,*
$KL(R_1(\cdot|v^{(t-1)}, ..., v^{(t-m)})||P_t(\cdot|v^{(t-1)}, ..., v^{(0)})) < \epsilon$ for $m \leq t < T-1$,
$KL(R_{0,t}(\cdot|v^{(t-1)}, ..., v^{(0)})||P_t(\cdot|v^{(t-1)}, ..., v^{(0)})) < \epsilon$ for $0 < t < m$,
and $KL(R_{0,0}(\cdot)||P_0(\cdot)) < \epsilon$,
then we can find distributions $P \in \mathbf{P}$ to approximate R to arbitrary precision.
Proof: The proof is given in the appendix.
Now we use this lemma to prove our first universal approximation theorem. In this case $\mathbf{P}$ is the set of distributions given by a TRBM. Note that throughout this paper P will often be used to refer to a TRBM. When we say P is a TRBM we are referring to the distribution associated with a set of parameters defining the TRBM.

Theorem 1: *Let R be a distribution over a sequence of length T of binary vectors of length n that is time homogeneous and satisfies the Markov property. For any $\epsilon > 0$ there exists a TRBM defined on sequences of length T of binary vectors of length n with distribution P such that $KL(R||P) < \epsilon$.*

Proof: By the previous lemma we will be looking for a TRBM that can approximate the transition probabilities of R along with its initial distribution. The proof will rely on the universal approximation properties of RBMs. The idea is that given one of the 2^n configurations of the visible units, v, there is an RBM with distribution P_v approximating $R_1(\cdot|v^{(t-1)} = v)$ to a certain precision. The universal approximation results for RBMs tell us that this approximation can be made arbitrarily precise for an RBM with enough hidden units. Furthermore, this approximation can be done with visible biases set to 0 (Le Roux and Bengio, 2008). We thus set all visible biases of our TRBM to 0 and include each of the approximating RBMs without difficulty. We label these RBMs $H_1, ..., H_{2^n}$. Given a specific configuration of the visible nodes v, H_v refers to the RBM chosen to approximate $R_1(\cdot|v^{(t-1)} = v)$.

The challenge then is to signal the TRBM which of the 2^n RBMs should be active at the next time step. To do this we include 2^n additional hidden nodes which we will call *control nodes*, each corresponding to a particular configuration of the visible units. Thus we add 2^n control nodes, $h_{c,1}, ..., h_{c,2^n}$ corresponding to the hidden nodes $H_1, ..., H_{2^n}$. Again, given a particular visible configuration v, we denote the corresponding control node by $h_{c,v}$. The set of all control nodes will be denoted H_c. Note that (c, v) is the label of $h_{c,v}$ and weights involving $h_{c,v}$ will be denoted $w_{(c,v),i}$ or $w'_{j,(c,v)}$. The control nodes will signal which of the H_i's should be active at the next time step. To accomplish this we will choose parameters such that when v is observed at time $t-1$, $h_{c,v}$ will be on at time $t-1$ with a probability close to 1 and every other control node will be off at time $t-1$ with probability close to 1. Each $h_{c,v}$ will have strong negative temporal connections to every $H_{v'}$ with $v' \neq v$, in essence turning off every RBM corresponding to $R_1(\cdot|v^{(t-1)} = v')$, and leaving the RBM corresponding to $R_1(\cdot|v^{(t-1)} = v)$ active (see Fig. 1). We will break the proof down into four parts and we must be able to choose parameters that satisfy all four conditions.

First, we must be able to choose parameters so that given $v^{(t-1)} = v$, the probability that $h_{c,v}^{(t-1)} = 1$ can be made arbitrarily close to 1 and the probability that $h_{c,v'}^{(t-1)} = 1$ can be made arbitrarily close to 0. Second, we must have that the control nodes have no impact on the visible distribution at the same time step so the transition probabilities of the TRBM will still approximate $R_1(\cdot|v^{(t-1)} = v)$. Third, we must be able to choose parameters so that given $h_{c,v}^{(t-1)} = 1$ and $h_{c,v'}^{(t-1)} = 0$, the probability that any nodes in $H_{v'}$ are on at time t can be made arbitrarily close to 0 for $v' \neq v$. Finally, we must be able to approximate the initial distribution R_0.

Note in the TRBM temporal data cannot flow directly from the visible nodes to the hidden nodes at the next time step. In contrast, by using the visible nodes at time t as input and the visible nodes at time $t+1$ as output, in the CRBM there is a direct relationship between the visible nodes at time t and the hidden nodes at time $t+1$. The CRBM is known to be a universal approximator (Montufar et al., 2014). The main challenge for the TRBM is to

encode the visible state in the control unit so that information can be passed to the hidden nodes at the next time step without the encoding changing the visible distribution. This is covered in Steps 1 and 2. Step 3 verifies that the correct distribution can be recovered from the encoding and Step 4 shows we can simulate the initial distribution without changing the rest of the machine.

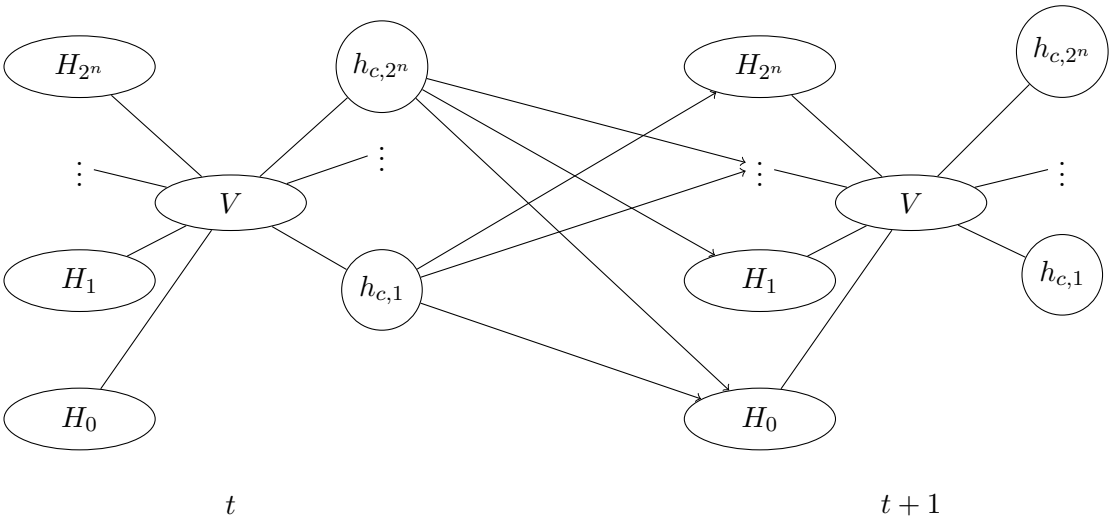

Figure 1: Interactions between sets of nodes within and between time steps: each $h_{c,v}$ turns off every $H_{v'} : v \neq v'$ in the subsequent time step. $h_{c,v}$ will only be on if v is observed at time t and collectively the control nodes H_c have negligible effect on the visible distribution. H_0 is an additional set of hidden units that will be used to model the initial distribution.

To choose temporal connections, define $w'_{j,(c,v)}$ to be $-\alpha$ if $h_j \in H_{v'}$ where $v' \neq v$ and 0 otherwise. Let every other $w'_{i,j} = 0$. In particular, remembering that W' is not necessarily symmetric, we have $w'_{(c,v),j} = 0$ for all h_j. The only parameters left to define are the biases and visible-hidden connections of H_c. Let $b_{(c,v)} = -(k - 0.5)\beta$ where k is the number of visible nodes on in v. Finally, define the connections from the visible units to $h_{c,v}$ by $w_{i,(c,v)} = \beta$ if v_i is on in the configuration v and $-\beta$ otherwise. Here the parameters of the control nodes are completely determined by α and β. We will proceed by showing that all necessary conditions are satisfied when α and β are large enough.

A note on notation. Throughout the proof H will denote the set of hidden nodes and $H^{(t)}$ will denote the set of configurations of hidden nodes at time t. Similarly $H_v^{(t)}$ will denote the set of configurations of the hidden nodes in which $h_i^{(t)} = 0$ if $h_i \notin H_v$. Similar conventions are used for H_c. $(H \backslash H_c)^{(t)}$ then denotes the set of configurations of non-control nodes. This is used in scenarios where we want to sum over a certain subset of hidden nodes and ignore the others, which is equivalent to simply setting all other nodes to 0. $H_{c,v}^{(t)}$ denotes the set of configurations of $h^{(t)}$ with $h_{c,v}^{(t)} = 1$ and $h_{c,v'}^{(t)} = 0$ for $v \neq v'$. $\bar{H}_{c,v}^{(t)}$ denotes the set of configurations $H^{(t)} \backslash H_{c,v}^{(t)}$. See Appendix A for a partial list of relevant notation.

Step 1:
For this step we show that as $\beta \to \infty$ we have $P(H^{(t)}_{c,v^{(t)}}|v^{(t)},\dots,v^{(0)}) \to 1$. Note that given the visible state at time t, the state of the control nodes at time t is conditionally independent of all other previous states. With this in mind, we can write the probability of a control node being on at time t as

$$P(h^{(t)}_{c,v} = 1|v^{(t)},...,v^{(0)}) = \sigma\left(\sum_i v^{(t)}_i w_{i,(c,v)} + b_{(c,v)}\right),$$

where σ is the logistic function. Note that for all $v^{(t)} \in \{0,1\}^n$ and $v \in \{0,1\}^n$, $\sum\limits_i v^{(t)}_i w_{i,(c,v)} = a\beta - b\beta$ where a is the number of nodes on in both v and $v^{(t)}$ and b is the number of nodes on in $v^{(t)}$ but off in v. Since $b_{(c,v)} = -(k-0.5)\beta$ if $v \neq v^{(t)}$, then either $a < k$, in which case $\sum\limits_i v^{(t)}_i w_{i,(c,v)} + b_{(c,v)} \leq -0.5\beta$, or $a = k$ and $b \geq 1$, which again implies $\sum\limits_i v^{(t)}_i w_{i,(c,v)} + b_{(c,v)} \leq -0.5\beta$. If $v = v^{(t)}$ then $a\beta - b\beta + b_{(c,v)} = k\beta - (k-0.5)\beta = 0.5\beta$. Thus if $v = v^{(t)}$,

$$\sigma\left(\sum_i v^{(t)}_i w_{i,(c,v)} + b_{(c,v)})\right) = \sigma(0.5\beta).$$

Otherwise

$$\sigma\left(\sum_i v^{(t)}_i w_{i,(c,v)} + b_{(c,v)}\right) \leq \sigma(-0.5\beta).$$

So as $\beta \to \infty$, $P(H^{(t)}_{c,v^{(t)}}|v^{(t)},...,v^{(0)}) \to 1$. In other words, for all $v^{(t)},...,v^{(0)}$ and all $\epsilon_0 > 0$ there exists some β_0 such that $\beta > \beta_0$ implies $|1 - P(H^{(t)}_{c,v^{(t)}}|v^{(t)},...,v^{(0)})| = P(\bar{H}^{(t)}_{c,v^{(t)}}|v^{(t)},...,v^{(0)}) < \epsilon_0$.

Step 2:
Here we show that by making β large enough we can make the effect of the control nodes on the visible distribution at the same time step negligible. Take any $v^{(t)}$, for all $h^{(t-1)}$, we have

$$P(v^{(t)}|h^{(t-1)}) = \frac{P(v^{(t)}|h^{(t-1)})}{\sum\limits_{v^{(t)}} P(v^{(t)}|h^{(t-1)})}$$

$$= \frac{\sum\limits_{h^{(t)} \in H^{(t)}_{c,v^{(t)}}} P(v^{(t)},h^{(t)}|h^{(t-1)}) + \sum\limits_{h^{(t)} \in \bar{H}^{(t)}_{c,v^{(t)}}} P(v^{(t)},h^{(t)}|h^{(t-1)})}{\sum\limits_{v^{(t)}} \sum\limits_{h^{(t)} \in H^{(t)}_{c,v^{(t)}}} P(v^{(t)},h^{(t)}|h^{(t-1)}) + \sum\limits_{v^{(t)}} \sum\limits_{h^{(t)} \in \bar{H}^{(t)}_{c,v^{(t)}}} P(v^{(t)},h^{(t)}|h^{(t-1)})}. \quad (1)$$

We also have that $P(v^{(t)},h^{(t)}|h^{(t-1)}) = P(h^{(t)}|v^{(t)},h^{(t-1)})P(v^{(t)}|h^{(t-1)})$ for all $h^{(t)}, v^{(t)}$ and by definition

$$\sum_{h^{(t)} \in \bar{H}^{(t)}_{c,v^{(t)}}} P(h^{(t)}|v^{(t)},h^{(t-1)})P(v^{(t)}|h^{(t-1)}) = P(\bar{H}^{(t)}_{c,v^{(t)}}|v^{(t)},h^{(t-1)})P(v^{(t)}|h^{(t-1)}).$$

By Step 1, there exists a β_0 such that for any $\beta > \beta_0$ we have $P(\bar{H}^{(t)}_{c,v^{(t)}}|v^{(t)},...,v^{(0)}) < \epsilon_0$ for all $v^{(t)},...,v^{(t-1)}$. Since the only connections going to a control node are from the visible units, given $v^{(t)}$, the state of the control nodes are conditionally independent of $v^{(t-1)},...,v^{(0)}$ and $h^{(t-1)}$. Since $P(v^{(t)}|h^{(t-1)}) < 1$, we have that $\beta > \beta_0$ implies that $P(\bar{H}^{(t)}_{c,v^{(t)}}|v^{(t)},h^{(t-1)})P(v^{(t)}|h^{(t-1)}) < \epsilon_0$, giving us that for $\beta > \beta_0$ and all $v^{(t)}$,

$$\sum_{h^{(t)} \in \bar{H}^{(t)}_{c,v^{(t)}}} P(v^{(t)},h^{(t)}|h^{(t-1)}) < \epsilon_0. \tag{2}$$

Note that this inequality is independent of α. Increasing α has no effect on $P(\bar{H}^{(t)}_{c,v^{(t)}}|v^{(t)},h^{(t-1)})$ and $P(v^{(t)}|h^{(t-1)})$ is bounded above by 1 so even after increasing α arbitrarily the inequality will hold with the same choice of β. Looking back to equation (1), as $\beta \to \infty$ the right hand terms in both the numerator and denominator go to 0. Consider $\sum_{h^{(t)} \in H^{(t)}_{c,v^{(t)}}} P(v^{(t)},h^{(t)}|h^{(t-1)})$. For all $v^{(t)}$, this is bounded above by 1 and since we are summing over $h^{(t)} \in H^{(t)}_{c,v^{(t)}}$, Step 1 tells us this is strictly increasing in β. This tells us that limit of the numerator and denominator of (1) are both finite and non-zero giving us that

$$\lim_{\beta\to\infty} P(v^{(t)}|h^{(t-1)}) = \lim_{\beta\to\infty} \frac{\sum_{h^{(t)} \in H^{(t)}_{c,v^{(t)}}} P(v^{(t)},h^{(t)}|h^{(t-1)})}{\sum_{v^{(t)}} \sum_{h^{(t)} \in H^{(t)}_{c,v^{(t)}}} P(v^{(t)},h^{(t)}|h^{(t-1)})}.$$

Define

$$\tilde{P}(v^{(t)}|h^{(t-1)}) := \frac{\sum_{h^{(t)} \in H^{(t)}_{c,v^{(t)}}} P(v^{(t)},h^{(t)}|h^{(t-1)})}{\sum_{v^{(t)}} \sum_{h^{(t)} \in H^{(t)}_{c,v^{(t)}}} P(v^{(t)},h^{(t)}|h^{(t-1)})}$$

$$= \frac{\sum_{h^{(t)} \in H^{(t)}_{c,v^{(t)}}} \exp\left(\sum_{i,j:h_j \in (H \setminus H_c)} v_i^{(t)} h_j^{(t)} w_{i,j} + \sum_{j:h_j \in (H\setminus H_c)} b_j h_j^{(t)} + 0.5\beta + \sum_{i,j} h_i^{(t)} h_j^{(t-1)} w'_{i,j}\right)}{\sum_{v^{(t)}} \sum_{h^{(t)} \in H^{(t)}_{c,v^{(t)}}} \exp\left(\sum_{i,j:h_j \in (H\setminus H_c)} v_i^{(t)} h_j^{(t)} w_{i,j} + \sum_{j:h_j \in (H\setminus H_c)} b_j h_j^{(t)} + 0.5\beta + \sum_{i,j} h_i^{(t)} h_j^{(t-1)} w'_{i,j}\right)}$$

$$= \frac{\sum_{h^{(t)} \in H^{(t)}_{c,v^{(t)}}} \exp\left(\sum_{i,j:h_j \in (H\setminus H_c)} v_i^{(t)} h_j^{(t)} w_{i,j} + \sum_{j:h_j \in (H\setminus H_c)} b_j h_j^{(t)} + \sum_{i,j} h_i^{(t)} h_j^{(t-1)} w'_{i,j}\right)}{\sum_{v^{(t)}} \sum_{h^{(t)} \in H^{(t)}_{c,v^{(t)}}} \exp\left(\sum_{i,j:h_j \in (H\setminus H_c)} v_i^{(t)} h_j^{(t)} w_{i,j} + \sum_{j:h_j \in (H\setminus H_c)} b_j h_j^{(t)} + \sum_{i,j} h_i^{(t)} h_j^{(t-1)} w'_{i,j}\right)},$$

but this is just the probability of $v^{(t)}$ when we remove the control nodes. Thus for any $v^{(t)}$ and $\epsilon_1 > 0$ there exists a β_0 such that $\beta > \beta_0$ implies that for all $h^{(t-1)}$, $|P(v^{(t)}|h^{(t-1)}) - \tilde{P}(v^{(t)}|h^{(t-1)})| < \epsilon_1$. Furthermore, this is unchanged by increasing α.

Step 3:
In this step, remembering that P_v is the distribution of the RBM corresponding to $R_1(\cdot|v^{(t-1)} = v)$, we show that as α and β are increased to infinity, if $h^{(t-1)} \in H_{c,v}^{(t-1)}$ then $P(v^{(t)}|h^{(t-1)}) \to P_v(v^{(t)})$ for all $v^{(t)}$. First note that since the states of any two hidden nodes at time t are independent, $P(h_j^{(t)} = 1|v^{(t)}, h^{(t-1)}) = \tilde{P}(h_j^{(t)} = 1|v^{(t)}, h^{(t-1)})$. Here $\tilde{P}$ is the system without control nodes, defined in the previous step. Take any $v^{(t)}$ and consider some configuration $h^{(t-1)} \in H_{c,v}^{(t-1)}$. We have $h_{c,v'}^{(t-1)} = 0$ for all $v' \neq v$ and $w_{j,(c,v)} = 0$ for $h_j \in H_v$, giving us

$$\tilde{P}(h_j^{(t)} = 1|v^{(t)}, h^{(t-1)}) = \sigma\left(\sum_i v_i^{(t)} w_{i,j} + b_j\right).$$

This is $P_v(h_j^{(t)} = 1|v^{(t)})$. Now take a hidden unit $h_j \in H_{v'}$ with $v' \neq v$. Since $v' \neq v$ and $h^{(t-1)} \in H_{c,v}^{(t-1)}$, then $h_{c,v}^{(t-1)} = 1$ and $w_{j,(c,v)} = -\alpha$. This gives us

$$\tilde{P}(h_j^{(t)} = 1|v^{(t)}, h^{(t-1)}) = \sigma\left(\sum_i v_i^{(t)} w_{i,j} + b_j - \alpha\right).$$

Since h_j is not a control node, $w_{i,j}$ is fixed for all v_i. Thus as $\alpha \to \infty$, $\tilde{P}(h_j^{(t)} = 1|v^{(t)}, h^{(t-1)}) \to 0$. So for any $\epsilon_0 > 0$ there exists α_0 such that $\alpha > \alpha_0$ implies that if $h_j \in H_{v'}$ with $v \neq v'$, $|\tilde{P}(h_j^{(t)} = 1|v^{(t)}, h^{(t-1)})| < \epsilon_0$. Now we have

$$\tilde{P}(v^{(t)}|h^{(t-1)}) = \sum_{h^{(t)} \in (H \setminus H_c)^{(t)}} \tilde{P}(v^{(t)}, h^{(t)}|h^{(t-1)})$$

$$= \sum_{h^{(t)} \in H_v^{(t)}} \tilde{P}(v^{(t)}, h^{(t)}|h^{(t-1)}) + \sum_{h^{(t)} \in (H \setminus H_c)^{(t)} \setminus H_v^{(t)}} \tilde{P}(v^{(t)}, h^{(t)}|h^{(t-1)}). \quad (3)$$

Note that $\tilde{P}(v^{(t)}, h^{(t)}|h^{(t-1)}) = \tilde{P}(h^{(t)}|v^{(t)}, h^{(t-1)})\tilde{P}(v^{(t)}|h^{(t-1)})$, $h_j^{(t)}$ and $h_i^{(t)}$ are independent for all i, j, and $\tilde{P}(v^{(t)}|h^{(t-1)}) < 1$. Thus we have that $\alpha > \alpha_0$ implies that if $h^{(t)} \in (H \setminus H_c)^{(t)} \setminus H_v^{(t)}$, then $\tilde{P}(h^{(t)}|v^{(t)}, h^{(t-1)})\tilde{P}(v^{(t)}|h^{(t-1)}) < \epsilon_0$. So as $\alpha \to \infty$, the right hand term of (3) goes to 0. So for any ϵ_1 there exists an α_1 such that $\alpha > \alpha_1$ implies $|\tilde{P}(v^{(t)}|h^{(t-1)}) - \sum_{h^{(t)} \in H_v^{(t)}} \tilde{P}(v^{(t)}, h^{(t)}|h^{(t-1)})| < \epsilon_1$. Note that since there are a finite number of configurations, $v^{(t)}$, we can take α large enough so that this is true for all $v^{(t)}$. So for any ϵ_1 we can choose $\alpha > \alpha_1$ so that

$$\left| \tilde{P}(v^{(t)}|h^{(t-1)}) - \frac{\sum_{h^{(t)} \in H_v^{(t)}} \tilde{P}(v^{(t)}, h^{(t)}|h^{(t-1)})}{\sum_{v^{(t)}, h^{(t)} \in H_v^{(t)}} \tilde{P}(v^{(t)}, h^{(t)}|h^{(t-1)})} \right| < \epsilon_1,$$

but we have

$$\frac{\sum_{h^{(t)}\in H_v^{(t)}} \tilde{P}(v^{(t)},h^{(t)}|h^{(t-1)})}{\sum_{v^{(t)},h^{(t)}\in H_v^{(t)}} \tilde{P}(v^{(t)},h^{(t)}|h^{(t-1)})} = \frac{\sum_{h^{(t)}\in H_v^{(t)}} \exp\left(\sum_{i,j} v_i^{(t)} h_j^{(t)} w_{i,j} + \sum_j b_j h_j^{(t)}\right)}{\sum_{v^{(t)},h^{(t)}\in H_v^{(t)}} \exp\left(\sum_{i,j} v_i^{(t)} h_j^{(t)} w_{i,j} + \sum_j b_j h_j^{(t)}\right)}$$

$$= P_v(v^{(t)}).$$

To summarize, Step 3 tells us that for any given $v^{(t)}$ and all $h^{(t-1)} \in H_{c,v}^{(t-1)}$ and any $\epsilon_1 > 0$, there exists α_1 such that $\alpha > \alpha_1$ implies that $|\tilde{P}(v^{(t)}|h^{(t-1)}) - P_v(v^{(t)})| < \epsilon_1$.

Step 4:
Finally, we must also be able to approximate the initial distribution R_0 to arbitrary precision. We know there is an RBM H_0 with visible biases 0 whose Boltzmann distribution can approximate R_0 to a certain precision. Include this machine in our TRBM. Now we define the initial biases. Let $b_{i,init} = \gamma$ for every $h_i \in H_0$ and $b_{c,v,init} = 0$ for all (c, v). Set $b_{i,init} = -\gamma$ for all other hidden nodes. Add $-\gamma$ to the biases of H_0. Call the distribution of this modified machine $\hat{P}$. By Step 2, for any $v^{(t)}$, and any $\epsilon_0 > 0$, there exists β_0 such that $\beta > \beta_0$ implies $|\hat{P}_0(v^{(0)}, h^{(0)}) - \tilde{\hat{P}}_0(v^{(0)}, h^{(0)})| < \epsilon_0$. If $h_k \in H_v$ for some v, we have

$$\tilde{\hat{P}}_0(h_k^{(0)} = 1|v^{(0)}) = \sigma\left(\sum_i v_i^{(0)} w_{i,k} + b_k - \gamma\right),$$

and for $h_j \in H_0$ we have

$$\tilde{\hat{P}}_0(h_j^{(0)} = 1|v^{(0)}) = \sigma\left(\sum_i v_i^{(0)} w_{i,j} + b_j\right).$$

Note that $\tilde{\hat{P}}_0$ does not depend on α or β. Following the same logic as in Step 3, for any $\epsilon_0 > 0$, there exists γ_0 such that $\gamma > \gamma_0$ implies $\tilde{\hat{P}}_0(v^{(0)}, h^{(0)}) < \epsilon_0$ if $h^{(0)} \in (H \backslash H_c)^{(t)} \backslash H_0^{(t)}$ for some H_v. So for all $v^{(t)}$, $\hat{P}_0(v^{(0)})$ can be made arbitrarily close to the probability of $v^{(0)}$ in the Boltzmann distribution of H_0, which by construction approximates R_0. At subsequent time steps, for each $h_j \in H_0$ we have $P(h_j^{(t)} = 1|h^{(t-1)}, v^{(t)}) = \sigma(\sum_i w_{i,j} v_i^{(t)} + b_j - \gamma)$. This can be made arbitrarily close to 0 by making γ arbitrarily large, so $P(h_j^{(t)} = 1, v^{(t)}|h^{(t-1)})$ can be made arbitrarily close to 0. Thus for any $\epsilon_0 > 0$ there exists γ_1 such that $\gamma > \gamma_1$ implies

$$|\hat{P}(v^{(t)}|h^{(t-1)}) - \sum_{h^{(t)} \notin H_0^{(t)}} \hat{P}(v^{(t)} h^{(t)}|h^{(t-1)})| < \epsilon_0. \tag{4}$$

But since γ does not appear anywhere else for $t > 0$, $\sum_{h^{(t)} \notin H_0^{(t)}} \hat{P}(v^{(t)} h^{(t)}|h^{(t-1)}) = P(v^{(t)}|h^{(t-1)})$.
Note that this construction allows the control nodes to be active in the first time step and to transmit temporal data without disturbing the initial distribution.

Now we put the four steps together. Given an arbitrary $0 < t < T$, we can write each $P(v^{(t)}|v^{(t-1)}, ..., v^{(0)})$ as

$$\sum_{h^{(t-1)}} P(v^{(t)}, h^{(t-1)}|v^{(t-1)}, ..., v^{(0)})$$

$$= \sum_{h^{(t-1)}} P(v^{(t)}|h^{(t-1)}, v^{(t-1)}, ..., v^{(0)})P(h^{(t-1)}|v^{(t-1)}, ..., v^{(0)})$$

$$= \sum_{h^{(t-1)}} P(v^{(t)}|h^{(t-1)})P(h^{(t-1)}|v^{(t-1)}, ..., v^{(0)}).$$

Step 1 tells us that if $h^{(t-1)} \notin H^{(t-1)}_{c,v^{(t-1)}}$, then $\lim_{\beta\to\infty} P(h^{(t-1)}|v^{(t-1)}, ..., v^{(0)}) = 0$. Step 2 tells us that $\lim_{\beta\to\infty} P(v^{(t)}|h^{(t-1)}) = \tilde{P}(v^{(t)}|h^{(t-1)})$. Since P is continuous in terms of β, for any ϵ_1, there exists β_0 such that $\beta > \beta_0$ implies

$$\begin{aligned} &| \sum_{h^{(t-1)}} P(v^{(t)}|h^{(t-1)})P(h^{(t-1)}|v^{(t-1)}, ..., v^{(0)}) - \\ &\sum_{h^{(t-1)} \in H^{(t-1)}_{c,v^{(t-1)}}} \tilde{P}(v^{(t)}|h^{(t-1)})P(h^{(t-1)}|v^{(t-1)}, ..., v^{(0)})| < \epsilon_1. \end{aligned} \tag{5}$$

Step 3 tells us that for any $\epsilon_0 > 0$, there exists an α_0 such that for all $h^{(t-1)} \in H^{(t-1)}_{c,v^{(t-1)}}$, if $\alpha > \alpha_0$ we have $|\tilde{P}(v^{(t)}|h^{(t-1)}) - P_{v^{(t-1)}}(v^{(t)})| < \epsilon_0$. So for any ϵ_1, there exists an α_0 such that $\alpha > \alpha_0$ implies

$$\begin{aligned} &| \sum_{h^{(t-1)} \in H^{(t-1)}_{c,v^{(t-1)}}} \tilde{P}(v^{(t)}|h^{(t-1)})P(h^{(t-1)}|v^{(t-1)}, ..., v^{(0)}) - \\ &P_{v^{(t-1)}}(v^{(t)}) \sum_{h^{(t-1)} \in H^{(t-1)}_{c,v^{(t-1)}}} P(h^{(t-1)}|v^{(t-1)}, ..., v^{(0)})| < \epsilon_1. \end{aligned} \tag{6}$$

Again by Step 1, as β goes to infinity, $\sum_{h^{(t-1)} \in H_{c,v^{(t-1)}}} P(h^{(t-1)}|v^{(t-1)}, ..., v^{(0)}) \to 1$, so for any ϵ_1 there exists β_1 such that $\beta > \beta_1$ implies that

$$|P_v^{(t-1)}(v^{(t)}) \sum_{h^{(t-1)} \in H^{(t-1)}_{c,v^{(t-1)}}} P(h^{(t-1)}|v^{(t-1)}, ..., v^{(0)}) - P_v^{(t-1)}(v^{(t)})| < \epsilon_1. \tag{7}$$

Now take any $\epsilon_2 > 0$ and take $\epsilon_1 < \epsilon_2/4$ with corresponding $\beta_0, \beta_1, \alpha_0$ so that the inequalities in (5), (6) and (7) hold. Then from Step 4 there exist γ_0, β_2 such that $\gamma > \gamma_0$ and $\beta > \beta_2$ implies that (4) holds. Then taking $\beta > \max(\beta_0, \beta_1, \beta_2)$, $\alpha > \alpha_0$, $\gamma > \gamma_0$ and applying the triangle inequality to (4), (5), (6), and (7) we have that $|\hat{P}(v^{(t)}|v^{(t-1)}, ..., v^{(0)}) - P_v^{(t-1)}(v^{(t)})| < \epsilon_2$. Since there are a finite number of configurations $v^{(t)}, v^{(t-1)}, ..., v^{(0)}$, we can choose α, β, γ so that this holds for all $v^{(t)}, v^{(t-1)}, ..., v^{(0)}$ and by construction, $KL(R(\cdot|v^{(t-1)})||P_{v^{(t-1)}}) < \epsilon$

for some arbitrarily chosen ϵ. Since the KL-divergence as a function of α, β, and γ is continuous, for any $\epsilon' > 0$ we can find $\alpha_1, \beta_2, \gamma_1$ such that $\alpha > \alpha_1$, $\beta > \beta_2$, $\gamma > \gamma_1$ implies that $|KL(R(\cdot|v^{(t-1)})||P(\cdot|v^{(t-1)}, ..., v^{(0)})) - KL(R(\cdot|v^{(t-1)})||P_{v^{(t-1)}}))| < \epsilon'$. And $KL(R(\cdot|v^{(t-1)})||P_{v^{(t-1)}})) < \epsilon$ for some arbitrarily chosen ϵ. So we can choose parameters such that $KL(R(\cdot|v^{(t-1)})||P(\cdot|v^{(t-1)}, ..., v^{(0)})) < \epsilon$. By the same argument, Step 4 tells us that we can choose parameters so that $KL(R_0||P_0) < \epsilon$. Thus by Lemma 1 the result holds. ■

Note that following the remark after the proof of Lemma 1, if we have a TRBM which approximates R over T time steps to a certain precision, it also approximates R over $t < T$ time steps to at least the same precision since the construction satisfies the conditions of Lemma 1.

2.2 The Generalized TRBM

The TRBM used in the previous section is a restricted instance of a more generalized model described by Sutskever et al. (2006). In the full model we allow explicit long-term hidden-hidden connections as well as long-term visible-visible connections. In this paper we will not consider models with visible-visible temporal interaction. From a practical standpoint any learning algorithm operating on a class of models with visible-visible interactions would be able to make those connections arbitrarily small if it helped, so in practice the class of models with visible-visible temporal connections is bigger than the one without any. The generalized TRBM is given by

$$P(v^{(t)}, h^{(t)}|h^{(t-1)}, ..., h^{(0)})$$

$$= \frac{\exp(v^{(t)\top} W h^{(t)} + c^\top v^{(t)} + b^\top h^{(t)} + h^{(t)\top} W^{(1)} h^{(t-1)} + ... + h^{(t)\top} W^{(m)} h^{(t-m)})}{Z(h^{(t-1)}, ..., h^{(t-m)})},$$

where we have a finite number of weight matrices $W^{(i)}$ used to determine the bias at time t. We replace $W^{(k)} h^{(t-k)}$ with an initial bias $b^{(k)}_{init}$ if $k > t$. The distribution $P(v^T, h^T)$ is then given by

$$P(v^T, h^T) = \left(\prod_{t=m}^{T-1} P(v^{(t)}, h^{(t)}|h^{(t-1)}, ..., h^{(t-m)}) \right) \left(\prod_{t=1}^{m-1} P(v^{(t)}, h^{(t)}|h^{(t-1)}, ..., h^{(0)}) \right)$$

$$\times P_0(v^{(0)}, h^{(0)}).$$

If we drop the restriction that R be a Markov chain we can generalize the previous theorem so that R is any distribution homogeneous in time with a finite time dependence.

Theorem 2: *Let R be a distribution over a sequence of length T of binary vectors of length n that is time homogeneous and has finite time dependence. For any $\epsilon > 0$ there exists a generalized TRBM, P, such that $KL(R||P) < \epsilon$.*

Proof: The initial part of the proof is identical to the proof of Theorem 1. Let m be the time dependence of R. Then for each visible sequence $v^{(t-1)}, ..., v^{(t-m)}$ we construct a TRBM P by adding sets of hidden units $H_{v^{(t-1)}, ..., v^{(t-m)}}$ with parameters chosen to approximate $R_1(\cdot|v^{(t-1)}, ..., v^{(t-m)})$. Note that although the indicies here are written as $v^{(t-1)}, ..., v^{(t-m)}$,

they do not depend on the time step t. Rather, there is one set of hidden nodes added for each configuration of an m-length sequence of visible nodes. The superscripts are added to distinguish different vectors in the sequence as well as emphasize how the connections should be made.

For each visible configuration v we add a control unit $h_{c,v}$ with the same bias and visible-hidden connections (determined by a parameter β) as in the construction for Theorem 1. If $i \leq m$, define the i-step temporal connections as $w^{(i)}_{(c,v),j} = -\alpha$ if $h_j \in H_{v^{(t-1)},...,v^{(t-i)},...,v^{(t-m)}}$ with $v^{(t-i)} \neq v$ and 0 otherwise. All other temporal connections are set to 0. Then repeating Step 1, Step 2, and Step 3 in Theorem 1, by making α and β sufficiently large we can make $KL(R_1(\cdot|v^{(t-1)},...,v^{(t-m)})||P(\cdot|v^{(t-1)},...,v^{(0)}))$ arbitrarily small for all v^T.

To finish the proof we must modify the machine to approximate the m initial distributions as well. In practice, one could train an RBM with the first m time steps as input in order to simulate the initial distribution. In this case the remainder of the proof is identical to step 4 of Theorem 1. The proof for the general TRBM as defined above is more intricate. In order to simulate the initial distributions with the general TRBM, First set $b^{(k)}_{init}$ to $-\gamma$ for each node $h \in H_{v^{(t-1)},...,v^{(t-m)}}$ and all k, and set $b^{(k)}_{init}$ to 0 for every control node. Now for each sequence $v^{(i-1)},...,v^{(0)}$ with $i < m$ add a set of hidden units $H_{v^{(i-1)},...,v^{(0)}}$ to approximate $R_{0,i}(\cdot|v^{(i-1)},...,v^{(0)})$ to a certain precision. For each i, call the set of all of these hidden units $H_{(i)}$. Connect each of these sets to the control nodes in the same way as done previously. In other words if $h_j \in H_{v^{(i-1)},...,v^{(0)}}$ then $w^{(l)}_{j,(c,v)} = -\alpha$ if $v^{(i-l)} \neq v$ and 0 otherwise. Add $-\gamma$ to the bias of each h_j if $h_j \in H_{(i)}$ for some i. For each $h_j \in H_{(i)}$ let $b^{(l)}_{init} = -\gamma$ for $l \neq i$ and $b^{(i)}_{init} = (m-i+2)\gamma$.

Start by choosing β so that $|P(v^{(i)}|v^{(i-1)},...,v^{(0)}) - \tilde{P}(v^{(i)}|v^{(i-1)},...,v^{(0)})| < \epsilon_0$. This can be done for any $\epsilon_0 > 0$ by the argument in Theorem 1 Step 2. Now for time $l < m$, for any non-control node $h_j \notin H_{(l)}$, and all $h^{(l-1)},...,h^{(0)}$, $\tilde{P}(h^{(l)}_j = 1|v^{(l)},h^{(l-1)},...,h^{(0)}) \leq \sigma(\sum w_{i,j}v^{(l)}_i + b_j - \gamma)$. This tends to 0 as $\gamma \to \infty$. So for any $\epsilon_1 > 0$, the argument in Theorem 1 Step 3 tells us we can choose γ large enough so that

$$|\tilde{P}(v^{(l)}|v^{(l-1)},...,v^{(0)}) - \sum_{h^{(l)} \in H^{(l)}_{(l)}} \tilde{P}(v^{(l)},h^{(l)}|v^{(l-1)},...,v^{(0)})| < \epsilon_1. \tag{8}$$

Furthermore if $h_j \in H_{(l)}$ then

$$\tilde{P}(h^{(l)}_j = 1|v^{(l)},h^{(l-1)},...,h^{(0)})$$

$$= \sigma\left(\sum_i w_{i,j}v^{(l)}_i + b_j + (m-i+2)\gamma - (m-i+2)\gamma + \sum_i w^{(1)}_{i,j}h^{(l-1)}_i + ... + \sum_i w^{(l)}_{i,j}h^{(0)}_j\right)$$

$$= \sigma\left(\sum w_{i,j}v^{(l)}_i + b_j + \sum_i w^{(1)}_{i,j}h^{(l-1)}_i + ... + \sum_i w^{(l)}_{i,j}h^{(0)}_j\right).$$

So $\sum_{h^{(l)} \in H_{(l)}^{(l)}} \tilde{P}(v^{(l)}, h^{(l)} | v^{(l-1)}, ..., v^{(0)})$ does not depend on γ. Using the same argument as in Step 3 of Theorem 1, for all $\epsilon_1 > 0$ there exists α_0 so that $\alpha > \alpha_0$ implies that

$$| \sum_{h^{(l)} \in H_{(l)}^{(l)}} \tilde{P}(v^{(l)}, h^{(l)} | v^{(l-1)}, ..., v^{(0)}) - \sum_{h^{(l)} \in H_{v^{(l-1)}, ..., v^{(0)}}} \tilde{P}(v^{(l)}, h^{(l)} | v^{(l-1)}, ..., v^{(0)}) | < \epsilon_1.$$

But the second term is just the probability of $v^{(l)}$ under the Boltzmann distribution of $H_{v^{(l-1)}, ..., v^{(0)}}$, so using continuity of the KL-divergence along with the triangle inequality gives us the second and third condition for Lemma 1. Finally note that for $t \geq m$, if $h_j \in H_{(l)}$ for any l and all $h^{(t-1)}, ..., h^{(t-m)}$, $P(h_j^{(t)} = 1 | v^{(t)}, h^{(t-1)}, ..., h^{(t-m)}) \leq \sigma(\sum w_{i,j} v_i^{(l)} + b_j - \gamma)$. So for any ϵ_1, we can take γ large enough such that

$$|\tilde{P}(v^{(t)} | v^{(t-1)}, ..., v^{(0)}) - \sum_{h^{(t)} \in (H \setminus H_c \setminus H_{(l)})^{(l)}} \tilde{P}(v^{(t)}, h^{(t)} | v^{(t-1)}, ..., v^{(0)})| < \epsilon_1. \tag{9}$$

Since for $t \geq m$, γ does not appear anywhere in the second term, this leaves us with the machine described in the first part of the proof, thus the first condition for Lemma 1 also holds. ■

3. Universal Approximation Results for the Recurrent Temporal Restricted Boltzmann Machines

The TRBM gives a nice way of using a Boltzmann Machine to define a probability distribution that captures time dependence in data, but it turns out to be difficult to train in practice (Sutskever et al., 2008). To fix this, a slight variation of the model, the RTRBM, was introduced. The key difference between the TRBM and the RTRBM is the use of deterministic real values denoted $h^{(t)}$. We will denote the probabilistic binary hidden units at time t by $h'^{(t)}$. The distribution defined by an RTRBM, Q, is

$$Q(v^T, h'^T) = \left(\prod_{k=1}^{T-1} Q(v^{(k)}, h'^{(k)} | h^{(k-1)}) \right) Q_0(v^{(0)}, h'^{(0)}).$$

Here $Q(v^{(t)}, h'^{(t)} | h^{(t-1)})$ is defined as

$$Q(v^{(t)}, h'^{(t)} | h^{(t-1)}) = \frac{\exp(v^{(t)\top} W h'^{(t)} + c^\top v^{(t)} + b^\top h'^{(t)} + h'^{(t)\top} W' h^{(t-1)})}{Z(h^{(t-1)})},$$

and h^T is a sequence of real-valued vectors defined by

$$\begin{aligned} h^{(t)} &= \sigma(W v^{(t)} + W' h^{(t-1)} + b), \\ h^{(0)} &= \sigma(W v^{(0)} + b_{init} + b), \end{aligned} \tag{10}$$

where σ is the logistic function and b_{init} is an initial bias. Q_0 is once again an initial distribution defined as a Boltzmann Distribution with bias $b + b_{init}$. The difference between

the RTRBM and the TRBM is the use of the sequence of real valued vectors h^T for the temporal connections. At each time step each hidden node h_i takes on two values, a deterministic $h_i^{(t)}$ and a probabilistic $h_i^{'(t)}$. The fact that the temporal parameters are calculated deterministically makes learning more tractable in these machines (Sutskever et al., 2008).

Theorem 3: *Let R be a distribution over a sequence of length T of binary vectors of length n that is time homogeneous and has finite time dependence. For any ϵ there exists an RTRBM, Q, such that $KL(R||Q) < \epsilon$.*
Proof: As in Theorem 2, for each configuration $v^{(t-1)}, ..., v^{(t-m)}$, include hidden units $H_{v^{(t-1)},...,v^{(t-m)}}$ with parameters so that the KL distance between the visible distribution of the Boltzmann machine given by $H_{v^{(t-1)},...,v^{(t-m)}}$ with these parameters and the distribution $R_1(\cdot|v^{(t-1)}, ..., v^{(t-m)})$ is less than ϵ'. Now for each possible visible configuration v add the control node $h_{c,v}$ with the same biases and visible-hidden weights as in Theorems 1 and 2 (determined entirely by parameter β). In Theorem 2, $h_{c,v}$ had i-step temporal connections from $h_{c,v}$ to every $H_{v^{(t-1)},...,v^{(t-m)}}$ with $v^{(t-i)} \neq v$. The proof will proceed by showing that each of these i-step temporal connections can instead be given by a chain of nodes in the RTRBM. We wish to show that we can add i hidden units connecting $h_{c,v}$ to every hidden node in $H_{v^{(t-1)},...,v^{(t-m)}}$ such that if $h_{c,v}$ is on at time $t-i$, it will have the same effect on the distribution of a node in $H_{v^{(t-1)},...,v^{(t-m)}}$ as it does in the general TRBM, and the i additional hidden units do not effect the visible distribution. If we can achieve this then the same proof will hold for the RTRBM. This will be done as follows.

For each $h_{c,v}$ and each $1 \leq k < m$, add an additional hidden unit with 0 bias and no visible-hidden connections. For each $h_{c,v}$, label these $m-1$ hidden units $g_{(v,1)}, ..., g_{(v,m-1)}$. Since these nodes have no visible-hidden connections they have no effect on the visible distribution at the current time step. For $1 < k < m-1$, let $w'_{(v,k),(v,k+1)} = 1$. Let $w'_{(c,v),(v,1)} = 1$ and $w'_{(v,k),j} = -\alpha$ if $h_j \in H_{v^{(t-1)},...,v^{(t-k-1)},...,v^{(t-m)}}$ with $v^{(t-k-1)} \neq v$, and $w'_{(v,k),j} = 0$ otherwise (see Fig. 2). Given a sequence $v^{(t-1)}, ..., v^{(t-m)}$, consider the probability that $h_j^{'(t)} = 1$ for some hidden unit $h_j \in H_{v^{(t-1)'},...,v^{(t-m)'}}$ where $v^{(t-k)'} \neq v^{(t-k)}$ for some k. Then by construction there is some hidden node g_{k-1} with $w'_{(v,k-1),j} = -\alpha$. The value of $g_{k-1}^{(t-1)}$ is calculated recursively by $g_{k-1}^{(t-1)} = \sigma(g_{k-2}^{(t-2)}) = \sigma^{k-1}(h_{c,v^{(t-k)}}^{(t-k)}) = \sigma^k(0.5\beta)$. Since k is bounded, by making β arbitrarily large we make g_{k-1} arbitrarily close to 1 and thus make $h_j^{(t)'} w'_{(v,k-1),j} g_{k-1}^{(t-1)}$ arbitrarily close to $-\alpha$.

Now suppose we have $h_j^{(t)'} = 1$ for some hidden unit in $H_{v^{(t-1)},...,v^{(t-m)}}$. Then every temporal connection is $-\alpha$, and $g_{(v,k)}^{(t-1)} = \sigma^k(-0.5\beta)$ for every $g_{(v,k)}$, so again by making β arbitrarily large we make the temporal terms arbitrarily close to 0. Thus as $\beta \to \infty$, the temporal terms from $h_{c,v}^{(t-i)}$ are either $-\alpha$ or 0 as needed.

We know from Theorem 2 that we can construct a TRBM with distribution P such that for $t \geq m$ and all $v^{(t)}$, $|P(v^{(t)}|v^{(t-1)}, ..., v^{(0)}) - R(v^{(t)}|v^{(t-1)}, ..., v^{(t-m)})| < \epsilon$ for any $\epsilon > 0$. The above argument shows that we can construct an RTRBM by replacing the connections $w_{(c,v),j}^{(i)}$ in the TRBM with the chain described above so that for any $\epsilon' > 0$

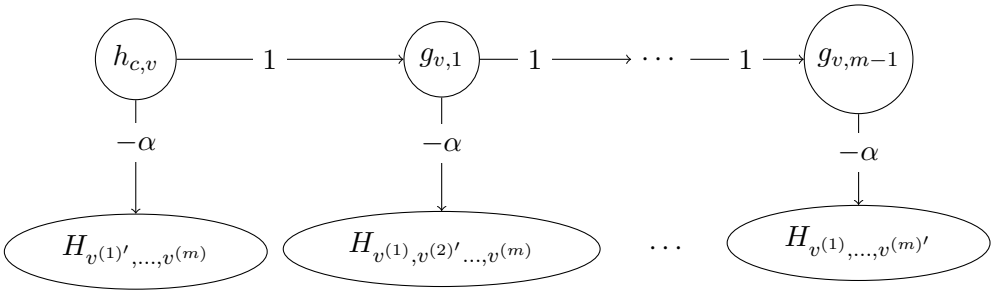

Figure 2: The temporal connections of a control node. Each $g_{v,i}$ connects to every $H_{v^{(1)},\ldots,v^{(i+1)},\ldots,v^{(m)}}$ with $v^{(i+1)\prime} \neq v$ and $h_{c,v}$ connects to every $H_{v^{(1)\prime},v^{(2)}\ldots,v^{(m)}}$ with $v^{(1)\prime} \neq v$.

there exist α_0, β_0 such that $\alpha > \alpha_0$ and $\beta > \beta_0$ imply that for all $v^{(t)}$ with $t \geq m$, $|Q(v^{(t)}|v^{(t-1)}, ..., v^{(0)}) - P(v^{(t)}|v^{(t-1)}, ..., v^{(0)})| < \epsilon'$. Note that since the chains are of length m, Q only depends on the previous m visible configurations. Then, once again applying the triangle inequality and continuity of the KL-divergence we can satisfy the first condition of Lemma 1.

To finish the proof our machine must also be able to approximate the initial m distributions. Again this could be easily done should we choose to use an RBM to approximate the distribution of the first m time step. Instead we provide a construction to simulate the first m distributions in the RTRBM using the definition given above. As before we will use the construction of Theorem 2 and replace the long-term temporal connections with a chain. To begin, add each $H_{(i)}$ described in Theorem 2 with the temporal connections again replaced by the chains described in the above step. Now we just need to replace the m initial biases. First add $-\gamma$ to the bias of each node in $H_{(i)}$. Add hidden units $l_0, ..., l_{m-1}$ with connections between them $w_{(l,i),(l,i+1)} = \delta$ and biases $-\delta$ for l_0, l_1 and -0.5δ for $l_2, ..., l_{m-1}$. For every $i > 0$, define the temporal connections to be 2γ from l_i to every node in $H_{(i)}$ and -2γ to every node in $H_{v^{(t-1)},\ldots,v^{(t-m)}}$ for all $v^{(t-1)}, ..., v^{(t-m)}$. Now set the initial biases for every l_i to be 0 except for l_0. Set this initial bias to be 2δ. Define the initial bias for every other non-control node to be $-\delta$ with the exception of $H_{(0)}$ whose initial bias is 0 (see Fig 3.).

First we calculate the values $l_i^{(t)}$. Since l_0 has bias of $-\delta$ and initial bias of 2δ, we have $l_0^{(0)} = \sigma(\delta)$, and $l_1^{(1)} = \sigma(\delta\sigma(\delta) - \delta)$. Taking the limit as $\delta \to \infty$ we have $l_0^{(0)} = 1$ and

$$\lim_{\delta\to\infty} l_1^{(1)} = \lim_{\delta\to\infty} \sigma(\delta(\sigma(\delta) - 1)) = \lim_{\delta\to\infty} \sigma\left(\frac{-\delta}{1 + \exp(\delta)}\right) = \sigma(0) = 0.5.$$

Next we calculate the limit of $l_2^{(2)}$ as $\delta \to \infty$:

$$\lim_{\delta\to\infty} l_2^{(2)} = \lim_{\delta\to\infty} \sigma(\delta l_1^{(1)} - 0.5\delta) = \lim_{\delta\to\infty} \sigma\left(\frac{\delta}{\frac{1}{(l_1^{(1)} - 0.5)}}\right) = \lim_{\delta\to\infty} \sigma\left(\frac{-(l_1^{(1)} - 0.5)^2}{\frac{d}{d\delta} l_1^{(1)}}\right).$$

Note that $\frac{d}{d\delta} l_1^{(1)}$ is finite and non-zero, so evaluating the limit we get $\lim_{\delta\to\infty} l_2^{(2)} = \sigma(0) = 0.5$. Then by induction $l_i^{(i)} = 0.5$ for $i > 1$. Now we look at the case where $j \neq i$. For $j > 0$

we have $l_0^{(j)} = \sigma(-\delta)$, $l_1^{(j+1)} = \sigma(l_0^{(j)} - \delta)$ and $l_k^{(j+k)} = \sigma(l_{k-1}^{(j+k-1)} - 0.5\delta)$ for $k > 1$. So for $j > i$ we have in the limit that $l_i^{(j)} = 0$. For $j < i$, we know $l_{j-i}^{(0)} \leq (-0.5\delta)$ so $l_{j-i+1}^{(1)} = \sigma(l_{j-i}^{(0)} - 0.5\delta)$, etc., so that in the limit we have $l_j^{(i)} = 0$. We conclude that for any $\epsilon > 0$, there exists δ_0 such that $\delta > \delta_0$ implies that for all $i > 0$ and all $j \neq i$, $|l_0^{(0)} - 1| < \epsilon, |l_i^{(i)} - 0.5| < \epsilon$, and $|l_i^{(j)}| < \epsilon$.

When $l_0^{(0)} = 0$, $l_i^{(i)} = 0.5$, and $l_i^{(j)} = 0$, we have that for $t < m$, if $h_j \in H_{v^{(t-1)},...,v^{(t-m)}}$ or $h_j \in H_{(i)}$ with $i \neq t$ then the bias is at most $b_j - \gamma$ and if $h_j \in H_{(t)}$ then the temporal connections from $l_1, ..., l_{m-1}$ are γ, which cancels the γ subtracted initially so the added bias is 0. For $t \geq m$ the temporal connections from $l_i, ..., l_{m-1}$ to all $H_{v^{(t-1)},...,v^{(t-m)}}$ are 0 and the added bias to each node in $H_{(i)}$ is $-\gamma$. This is exactly the machine described in the second part of Theorem 2.

Putting this together, we first note that since each l_i has no visible connections we can ignore their binary values much in the same way that we can ignore the chains in the first part of the proof. Now for $t \neq 0$ and any $\epsilon > 0$ and any $v^{(t)}$, first we choose $\beta > \beta_0$ so that $|Q(v^{(t)}|v^{(t-1)}, ..., v^{(0)}) - \tilde{Q}(v^{(t)}|v^{(t-1)}, ..., v^{(0)})| < \epsilon'$, then we choose $\delta > \delta_0$ so that $|\tilde{Q}(v^{(t)}|v^{(t-1)}, ..., v^{(0)}) - \tilde{\tilde{Q}}(v^{(t)}|v^{(t-1)}, ..., v^{(0)})| < \epsilon'$ where $\tilde{\tilde{Q}}$ is the distribution obtained by replacing $l_0^{(0)}$ with 1, $l_i^{(i)}$ with 0.5 for $i > 0$, and $l_j^{(i)}$ with 0. As noted above this distribution is exactly the construction from the second part of Theorem 2. Finally, by the reasoning of Theorem 1 Step 4, by making δ large we make the initial distribution arbitrarily close to $H_{(0)}$ allowing us to approximate the distribution for the first time step. So the first, second and third conditions of the lemma are satisfied by the same argument used in Theorem 2. ■

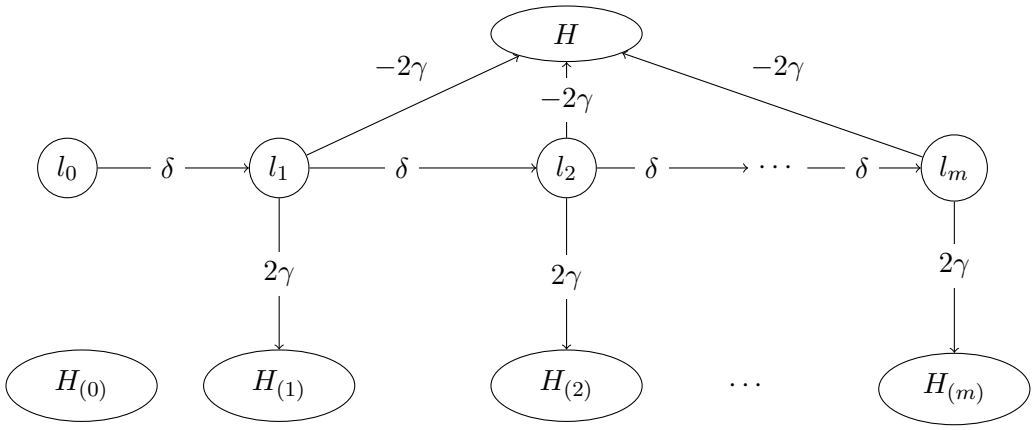

Figure 3: The temporal connections of the initial chain. H is the machine defined in the first part of the proof. Each connection from an l_i to H or $H_{(k)}$ only goes to non-control nodes in the set (control nodes include the chains $g_{(v,1)}, ..., g_{(v,m)}$). The l_i's have no visible connections.

4. Conclusion

The proofs above have shown that generalized TRBMs and RTRBMs both satisfy the same universal approximation condition. In the proof of universal approximation for the RTRBM we take the weights large enough so that the real-valued hidden sequence becomes approximately binary. This suggests that the same proof could be adapted to the basic TRBM. However, the TRBM seems to have difficulty modeling distributions with long term time dependence. After an RTRBM was trained on videos of bouncing balls the machine was able to model the movement correctly and the hidden units contained chains of length two as described in the above proof (Sutskever et al., 2008). On the other hand the TRBM did not have this structure and modeled the motion of balls as a random walk which is what one might expect for a machine that is unable to use velocity data by modeling two-step time dependencies. Given the likely equivalence in representational power of the two models, this discrepancy of results is best explained by the efficiency of the learning algorithm for the RTRBM in comparison to the TRBM.

At first glance the constructions used here seem quite inefficient. For Theorem 1 we require $2^n(G(n)+1)$ hidden nodes where G is the number of hidden nodes required to approximate an arbitrary distribution on n nodes with a Restricted Boltzmann Machine. It is important to note that the number of nodes required here, although large, depends only on the number of visible units and the process we wish to approximate, not the number of time steps for which we wish to approximate R. If the required number of hidden nodes had depended on the number of time steps then the TRBM and RTRBM would be essentially pointless as the RBM can do the same for any finite number of time steps. In contrast, the CRBM has a comparatively small lower bound on the number of hidden units required to approximate a set of conditional distributions (Montufar et al., 2014). Nonetheless, the above proofs are constructive and give only an upper bound on the required number of hidden units. Furthermore, we made no assumptions about $KL(R_1(\cdot|v^{(t-1)})||R_1(\cdot|v^{(t-1)'}))$. Even if $v^{(t-1)}$ and $v^{(t-1)'}$ are similar vectors, the resulting distributions may be quite different, so to guarantee the result in full generality we could need a whole new set of hidden units to approximate R_1 for each pattern $v^{(t-1)}$. With this in mind, we might expect $2^n(G(n))$ to be a reasonable lower bound. In practice, similar vectors in the previous time step should produce similar distributions for the current time step. For example, looking at consecutive frames in video data, we expect that two similar frames at a certain time step will lead to similar frames in the next time step. To formalize this we could impose the restriction $KL(R_1(\cdot|v^{(t-1)})||R_1(\cdot|v^{(t-1)'})) < f(d(v^{(t-1)}, v^{(t-1)'}))$, where f is a bounded function and d is a metric on $\{0,1\}^n$. With this condition we could hope to find a more efficient TRBM to approximate R than the one given in the proof.

Without making this additional assumption, the most obvious way to increase efficiency is to obtain a better upper bound on G. We know that the bounds given by Le Roux et al. (2008) are not the lowest possible upper bounds for G (Montufar and Ay, 2011). In practice, multiple layers of RBMs are often stacked, leading to a Deep Belief Network. Several papers have investigated the universal approximation properties of Deep Belief Networks (Sutskever and Hinton, 2010)(Le Roux and Bengio, 2010)(Montufar and Ay, 2011). Re-

calling the constructions used in the previous proofs, by replacing the RBMs modeling the transition probabilities with Deep Belief Networks we end up with a column structure in which certain control nodes in a column send negative feedback to the other columns. This structure bears an interesting resemblance to the structure of the visual cortex (Goodhill and Carreira-Perpinán, 2006) suggesting that perhaps the two are computationally similar.

Acknowledgments

The authors thank NSERC and the University of Victoria for partial funding of this work, and anonymous reviewers for helpful comments.

Appendix A.

The following table lists notations used for the labels and states for the nodes in the previous proofs

H_0	*a set of hidden nodes whose distribution approximates* R_0
$H_{(i)}$	*a set of hidden nodes used to approximate the distribution at time* i
H_v	*Hidden nodes whose distribution approximates* $R_1(\cdot\|v^{(t-1)}=v)$
$h_{c,v}$	*the control node corresponding to the configuration* v *of the visible units*
H_c	*the set of all control nodes*
$H_{c,v}^{(t)}$	*the set of configurations of the hidden nodes at time* t *with* $h_{c,v}^{(t)}=1$ *and* $h_{c,v'}^{(t)}=0$
$\bar{H}_{c,v}^{(t)}$	*the set of configurations of hidden nodes at time* t *not in* $H_{c,v}^{(t)}$
$g_{(v,i)}$	*the* i^{th} *node in a chain connecting* $h_{(c,v)}$ *to the visible nodes*
l_i	*the* i^{th} *node in the initial chain connecting* l_0 *with the rest of the* H

Appendix B.

In this appendix we provide a proof for Lemma 1
Proof: For an arbitrary $\epsilon > 0$, we need to find a $P \in \mathbf{P}$ such that $KL(R||P) < \epsilon$, where the KL-divergence is

$$KL(R||P) = \sum_{v^T} R(v^T) \log\left(\frac{R(v^T)}{P(v^T)}\right).$$

We can write $P(v^T)$ as

$$\left(\prod_{t=1}^{T-1} P(v^{(t)}|v^{(t-1)},...,v^{(0)})\right) P(v^{(0)}),$$

and by assumption

$$R(v^T) = \left(\prod_{t=m}^{T-1} R_1(v^{(t)}|v^{(t-1)},...,v^{(t-m)})\right) \prod_{i=1}^{m-1} R_0(v^{(i)}|(v^{(i-1)},...,v^{(0)}) R_0(v^{(0)}).$$

Then expanding out the log in the KL-divergence gives us

$$KL(R||P) = \sum_{v^T}\sum_{t=m}^{T-1} R(v^T) \log\left(\frac{R_1(v^{(t)}|v^{(t-1)},...,v^{(t-m)})}{P(v^{(t)}|v^{(t-1)},...,v^{(0)})}\right)$$

$$+ \sum_{v^T}\sum_{t=1}^{m-1} R(v^T) \log\left(\frac{R_0(v^{(t)}|v^{(t-1)},...,v^{(0)})}{P(v^{(t)}|v^{(t-1)},...,v^{(0)})}\right) + \sum_{v^T} R(v^T) \log\left(\frac{R_0(v^{(0)})}{P(v^{(0)})}\right).$$

We can decompose $R(v^T)$ into $R(v^{(T-1)},...,v^{(t)}|v^{(t-1)},...,v^{(0)})R(v^{(t-1)},...,v^{(0)})$ so for a given t we can write

$$\sum_{v^T} R(v^T) \log\left(\frac{R_1(v^{(t)}|v^{(t-1)},...,v^{(t-m)})}{P(v^{(t)}|v^{(t-1)},...,v^{(0)})}\right)$$

$$= \sum_{v^{(t-1)},...,v^{(0)}} R(v^{(t-1)},...,v^{(0)})$$

$$\times \left(\sum_{v^{(t)}} \sum_{v^{(T-1)},...,v^{(t+1)}} R(v^{(T-1)},...,v^{(t)}|v^{(t-1)},...,v^{(0)}) \log \left(\frac{R_1(v^{(t)}|v^{(t-1)},...,v^{(t-m)})}{P(v^{(t)}|v^{(t-1)},...,v^{(0)})} \right) \right)$$

$$= \sum_{v^{(t-1)},...,v^{(0)}} R(v^{(t-1)},...,v^{(0)}) KL(R_1(\cdot|v^{(t-1)},...,v^{(t-m)})||P_t(\cdot|v^{(t-1)},...,v^{(0)})).$$

Since $R(v^{(t-1)},...,v^{(0)}) < 1$ for all $v^{(t-1)},...,v^{(0)}$ we have

$$\sum_{v^T} R(v^T) \log \left(\frac{R_1(v^{(t)}|v^{(t-1)},...,v^{(t-m)})}{P(v^{(t)}|v^{(t-1)},...,v^{(0)})} \right)$$

$$\leq 2^{tn} \sum_{v^t} KL(R_1(\cdot|v^{(t-1)},...,v^{(t-m)})||P_t(\cdot|v^{(t-1)},...,v^{(0)})).$$

The same logic applies for the cases with $t < m$.
By hypothesis there exists $P \in \mathbf{P}$ such that for every v^T and every ϵ',
$KL(R_1(\cdot|v^{(t-1)},...,v^{(t-m)})||P_t(\cdot|v^{(t-1)},...,v^{(0)})) < \epsilon'$ for $t \geq m$,
$KL(R_{0,t}(\cdot|v^{(t-1)},...,v^{(0)})||P_t(\cdot|v^{(t-1)},...,v^{(0)})) < \epsilon'$ for every $0 < t < m$ and
$KL(R_0||P_0) < \epsilon'$.
This gives us

$$\begin{aligned} KL(R||P) &\leq \sum_{t=m}^{T-1} 2^{tn} \sum_{v^t} KL(R_1(\cdot|v^{(t-1)},...,v^{(t-m)})||P_t(\cdot|v^{(t-1)},...,v^{(0)})) \\ &+ \sum_{t=1}^{m-1} 2^{tn} \sum_{v^t} KL(R_0(\cdot|v^{(t-1)},...,v^{(0)})||P(\cdot|v^{(t-1)},...,v^{(0)})) \\ &+ KL(R_0||P_0) \\ &< \sum_{t=m}^{T-1} 4^{tn}\epsilon' + \sum_{t=0}^{m-1} 4^{tn}\epsilon'. \end{aligned}$$

Then we merely choose an ϵ' so that this expression is less than ϵ and choose a corresponding $P \in \mathbf{P}$. ■

Notice in the proof that T was chosen arbitrarily and in the last line of the proof we see that decreasing T provides a tighter bound on the KL-divergence so any distribution which approximates R with a certain upper bound on the KL-divergence for T time steps will approximate R with at most the same upper bound on the KL-divergence for $t < T$ time steps.

References

Y. Freund and D. Haussler. Unsupervised learning of distributions of binary vectors using 2 layer networks. *Advances in Neural Information Processing Systems*, 4:912–919, 1991.

G. Goodhill and M. Carreira-Perpinán. *Cortical Columns*. Macmillan, first edition, 2006.

G. Hinton. Training a product of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

Geoffrey E. Hinton and Simon Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1553, 2006.

N. Le Roux and Y. Bengio. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, 20:1631–1649, 2008.

N. Le Roux and Y. Bengio. Deep belief networks are compact universal approximators. *Neural Computation*, 22:2192 – 2207, 2010.

G. Montufar and N. Ay. Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. *Neural Computation*, 23:1306–1319, 2011.

G. Montufar, N. Ay, and K. Ghazi-Zahedi. Geometry and expressive power of conditional restricted Boltzmann machines. *Journal of Machine Learning(Preprint)*, 2014. URL `http://arxiv.org/abs/1402.3346`.

I. Sutskever and G. Hinton. Learning multilevel distributed representations for high-dimensional sequences. *Proceeding of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 544 – 551, 2006.

I. Sutskever and G. Hinton. Deep, narrow sigmoid belief networks are universal approximators. *Neural Computation*, 20:2192 – 2207, 2010.

I. Sutskever, G. Hinton, and G. Taylor. The recurrent temporal restricted Boltzmann machine. *Advances in Nueral Information Processing Systems*, 21:1601–1608, 2008.

G. Taylor, G. Hinton, and S. Roweis. Modeling human motion using binary latent variables. *Advances in Neural Information Processing Systems*, 19:1345–1352, 2006.

L. Younes. Synchronous Boltzmann machines can be universal approximators. *Applied Mathematics Letters*, 9:109–113, 1996.