On the Representational Efficiency of Restricted Boltzmann Machines

James Martens∗ Arkadev Chattopadhyay+ Toniann Pitassi∗ Richard Zemel∗
∗Department of Computer Science 
University of Toronto 
{jmartens,toni,zemel}@cs.toronto.edu
+School of Technology & Computer Science
Tata Institute of Fundamental Research
arkadev.c@tifr.res.in

Abstract

This paper examines the question: What kinds of distributions can be efficiently represented by Restricted Boltzmann Machines (RBMs)? We characterize the RBM’s unnormalized log-likelihood function as a type of neural network, and through a series of simulation results relate these networks to ones whose representational properties are better understood. We show the surprising result that RBMs can efficiently capture any distribution whose density depends on the number of 1’s in their input. We also provide the first known example of a particular type of distribution that provably cannot be efficiently represented by an RBM, assuming a realistic exponential upper bound on the weights. By formally demonstrating that a relatively simple distribution cannot be represented efficiently by an RBM our results provide a new rigorous justification for the use of potentially more expressive generative models, such as deeper ones.

1 Introduction

Standard Restricted Boltzmann Machines (RBMs) are a type of Markov Random Field (MRF) characterized by a bipartite dependency structure between a group of binary visible units $x \in \{0, 1\}^n$ and binary hidden units $h \in \{0, 1\}^m$. Their energy function is given by:

$$E_\theta(x, h) = -x^\top W h - c^\top x - b^\top h$$

where $W \in \mathbb{R}^{n \times m}$ is the matrix of weights, $c \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ are vectors that store the input and hidden biases (respectively) and together these are referred to as the RBM’s parameters $\theta = \{W, c, b\}$. The energy function specifies the probability distribution over the joint space $(x, h)$ via the Boltzmann distribution $p(x, h) = \frac{1}{Z_\theta} \exp(-E_\theta(x, h))$ with the partition function $Z_\theta$ given by $\sum_{x,h} \exp(-E_\theta(x, h))$. Based on this definition, the probability for any subset of variables can be obtained by conditioning and marginalization, although this can only be done efficiently up to a multiplicative constant due to the intractability of the RBM’s partition function [Long and Servedio, 2010].

RBMs have been widely applied to various modeling tasks, both as generative models (e.g. [Salakhutdinov and Murray, 2008, Hinton, 2000, Courville et al., 2011, Martin et al., 2010, Tang and Sutskever, 2011]), and for pre-training feed-forward neural nets in a layer-wise fashion (Hinton and Salakhutdinov, 2006). This method has led to many new applications in general machine learning problems including object recognition and dimensionality reduction. While promising for practical applications, the scope and basic properties of these statistical models have only begun to be studied.

As with any statistical model, it is important to understand the expressive power of RBMs, both to gain insight into the range of problems where they can be successfully applied, and to provide justification for the use of potentially more expressive generative models. In particular, we are interested in the question of how large the number of hidden units $m$ must be in order to capture a particular distribution to arbitrarily high accuracy. The question of size is of practical interest, since very large models will be computationally more demanding (or totally impractical), and will tend to overfit a lot more during training.
It was shown in [Freund and Haussler, 1994], and later [Le Roux and Bengio, 2008] that for binary-valued \( x \), any distribution over \( x \) can be realized (up to an approximation error which vanishes exponentially quickly in the magnitude of the parameters) by an RBM, as long as \( m \) is allowed to grow exponentially fast in input dimension (\( n \)). Intuitively, this construction works by instantiating, for each of the up to \( 2^n \) possible values of \( x \) that have support, a single hidden unit which turns on only for that particular value of \( x \) (with overwhelming probability), so that the corresponding probability mass can be individually set by manipulating that unit’s bias parameter. An improvement to this result was obtained by [Montufar and Ay, 2011], however this construction still requires that \( m \) grow exponentially fast in \( n \).

Recently, [Montufar et al., 2011] generalized the construction used by Le Roux and Bengio (2008) so that each hidden unit turns on for, and assigns probability mass to, not just a single \( x \). Specifically, they showed that any \( k \)-component mixture of product distributions over the free variables of mutually disjoint cubic sets can be approximated arbitrarily well by an RBM with \( m = k \) hidden units.

Unfortunately, families of distributions that are of this specialized form (for some \( m = k \) bounded by a polynomial function of \( n \)) constitute only a very limited subset of all distributions that have some kind of meaningful/interesting structure. For example, this result would not allow us to efficiently construct simple distributions where the mass is a function of \( \sum_i x_i \) (e.g., for \( p(x) \propto \text{PARITY}(x) \)).

In terms of what kinds of distributions provably cannot be efficiently represented by RBMs, even less is known. [Cueto et al., 2009] characterized the distributions that can be realized by a RBM with \( k \) parameters as residing within a manifold inside the entire space of distributions on \( \{0,1\}^n \) whose dimension depends on \( k \). For sub-exponential \( k \) this implies the existence of distributions which cannot be represented. However, this kind of result gives us no indication of what these hard-to-represent distributions might look like, leaving the possibility that they might all be structureless or otherwise uninteresting.

In this paper we first develop some tools and simulation results which relate RBMs to certain easier-to-analyze approximations, and to neural networks with 1 hidden layer of threshold neurons, for which many results about representational efficiency are already known [Maass, 1992; Maass et al., 1994; Hajnal et al., 1999]. This opens the door to a range of potentially relevant complexity results, some of which we apply in this paper.

Next, we present a construction that shows how RBMs with \( m = n^2 + 1 \) can produce arbitrarily good approximations to any distribution where the mass is a symmetric function of the inputs (that is, it depends on \( \sum_i x_i \)). One example of such a function is the (in)famous PARITY function, which was shown to be hard to compute in the perceptron model by the classic Minsky and Papert book from 1968.

Having ruled out such distributions as candidates for ones that are hard for RBMs to represent, we provide a concrete example of one which is only slightly more complex, and yet whose representation by an RBM requires \( m \) to grow exponentially with \( n \) (assuming an exponential upper bound on the size of the RBM’s weights). Because this distributions is particular simple, it can be viewed as a special case of many other more complex types of distributions, and thus our results speak to the hardness of representing those distributions with RBMs as well.

Our results thus provide a pair of explicit example distributions, the latter of which is only a slightly more complicated version of the former, that provide a fine delineation between what is “easy” for RBMs to represent, and what is “hard”. Perhaps more importantly, they demonstrate that the distributions that cannot be efficiently represented by RBMs can have a relatively basic structure, and are not simply random in appearance as one might hope given the previous results. This provides a new form of rigorous justification for the use of deeper generative models such as Deep Boltzmann Machines [Salakhutdinov and Hinton, 2009], and contrastive backpropagation networks [Hinton et al., 2006] over standard RBMs.

The rest of the paper is organized as follows. Section 2 characterizes the unnormalized log-likelihood as a type of neural network (called an “RBM network”) and shows how this type is related to single hidden layer neural networks of threshold neurons, and to an easier-to-analyze approximation (which we call a “hardplus RBM network”). Section 3 describes a \( m = n^2 + 1 \) construction for distributions whose mass is a function of \( \sum_i x_i \), and in Section 4 we present an exponential lower bound on \( m \) for a slightly more complicated class of explicit distributions. Note that all proofs can be found in the Appendix.
2 RBM networks

2.1 Free energy networks

In an RBM, the (negative) unnormalized log probability of $x$, after $h$ has been marginalized out, is known as the free energy. Denoted by $F_\theta(x)$, the free energy satisfies the property that $p(x) = \exp(-F_\theta(x))/Z_\theta$ where $Z_\theta$ is the usual partition function.

It is well known (see Appendix A.1 for a derivation) that, due to the bipartite structure of RBMs, computing $F$ is tractable and has a particularly nice form:

$$F_\theta(x) = -c^\top x - \sum_j \log(1 + \exp(x^\top [W][j] + b_j))$$

where $[W][j]$ is the $j$-th column of $W$.

Because the free energy completely determines the log probability of $x$, it fully characterizes an RBM’s distribution. So studying what kinds of distributions an RBM can represent amounts to studying the kinds of functions that can be realized by the free energy function for some setting of $\theta$.

2.2 RBM networks

The form of an RBM’s free energy function can be expressed as a standard feed-forward neural network, or equivalently, a real-valued circuit, where instead of using hidden units with the usual sigmoidal activation functions, we have $m$ “neurons” (a term we will use to avoid confusion with the original meaning of a “unit” in the context of RBMs) that use the softplus activation function:

$$\text{soft}(y) = \log(1 + \exp(y))$$

Note that at the cost of increasing $m$ by one (which does not matter asymptotically) and introducing an arbitrarily small approximation error, we can assume that the visible biases ($c$) of an RBM are all zero. To see this, note that up to an additive constant, we can very closely approximate $c^\top x$ by $\text{soft}(K + c^\top x) \approx K + c^\top x$ for a suitably large value of $K$ (i.e., $K \gg \|c\|_1 \geq \max_x(c^\top x)$).

Proposition 11 in the Appendix quantifies the very rapid convergence of this approximation as $K$ increases.

These observations motivate the following definition of an RBM network, which computes functions with the same form as the negative free energy function of an RBM (assumed to have $c = 0$), or equivalently the log probability (negative energy) function of an RBM. RBM networks are illustrated in Figure 1.

Definition 1 A RBM network with parameters $W, b$ is defined as a neural network with one hidden layer containing $m$ softplus neurons and weights and biases given by $W$ and $b$, so that each neuron $j$’s output is $\text{soft}([W][j] + b_j)$. The output layer contains one neuron whose weights and bias are given by $1 \equiv [11\ldots1]^\top$ and the scalar $B$, respectively.

For convenience, we include the bias constant $B$ so that RBM networks shift their output by an additive constant (which does not affect the probability distribution implied by the RBM network since any additive constant is canceled out by $\log Z$ in the full log probability).
2.3 Hardplus RBM networks

A function which is somewhat easier to analyze than the softplus function is the so-called hardplus function (aka 'plus' or 'rectification'), defined by:

\[ \text{hard}(y) = \max(0, y) \]

As their names suggest, the softplus function can be viewed as a smooth approximation of the hardplus, as illustrated in Figure 1. We define a hardplus RBM network in the obvious way: as an RBM network with the softplus activation functions of the hidden neurons replaced with hardplus functions.

The strategy we use to prove many of the results in this paper is to first establish them for hardplus RBM networks, and then show how they can be adapted to the standard softplus case via simulation results given in the following section.

2.4 Hardplus RBM networks versus (Softplus) RBM networks

In this section we present some approximate simulation results which relate hardplus and standard (softplus) RBM networks.

The first result formalizes the simple observation that for large input magnitudes, the softplus and hardplus functions behave very similarly (see Figure 1, and Proposition 11 in the Appendix).

**Lemma 2.** Suppose we have a softplus and hardplus RBM networks with identical sizes and parameters. If, for each possible input \( x \in \{0, 1\}^n \), the magnitude of the input to each neuron is bounded from below by \( C \), then the two networks compute the same real-valued function, up to an error (measured by \( |\cdot| \)) which is bounded by \( m \exp(-C) \).

The next result demonstrates how to approximately simulate a RBM network with a hardplus RBM network while incurring an approximation error which shrinks as the number of neurons increases. The basic idea is to simulate individual softplus neurons with groups of hardplus neurons that compute what amounts to a piece-wise linear approximation of the smooth region of a softplus function.

**Theorem 3.** Suppose we have a (softplus) RBM network with \( m \) hidden neurons with parameters bounded in magnitude by \( C \). Let \( p > 0 \). Then there exists a hardplus RBM network with \( \leq 2m^2p \log(mp) + m \) hidden neurons and with parameters bounded in magnitude by \( C \) which computes the same function, up to an approximation error of \( 1/p \).

Note that if \( p \) and \( m \) are polynomial functions of \( n \), then the simulation produces hardplus RBM networks whose size is also polynomial in \( n \).

2.5 Thresholded Networks and Boolean Functions

Many relevant results and proof techniques concerning the properties of neural networks focus on the case where the output is thresholded to compute a Boolean function (i.e. a binary classification). In this section we define some key concepts regarding output thresholding, and present some basic propositions that demonstrate how hardness results for computing Boolean functions via thresholding yield analogous hardness results for computing certain real-valued functions.

We say that a real-valued function \( g \) represents a Boolean function \( f \) with margin \( \delta \) if for all \( x \) \( g \) satisfies \( |g(x)| \geq \delta \) and \( \text{thresh}(g(x)) = f(x) \), where \( \text{thresh} \) is the 0/1 valued threshold function defined by:

\[ \text{thresh}(a) = \begin{cases} 1 & : a \geq 0 \\
0 & : a < 0 \end{cases} \]

We define a thresholded neural network (a distinct concept from a “threshold network”, which is a neural network with hidden neurons whose activation function is \( \text{thresh} \)) to be a neural network whose output is a single real value, which is followed by an application of the threshold function. Such a network will be said to compute a given Boolean function \( f \) with margin \( \delta \) (similar to the concept of “separation” from [Maass et al. (1994)]) if the real valued input \( g \) to the final threshold represents \( f \) according to the above definition.

While the output of a thresholded RBM network does not correspond to the log probability of an RBM, the following observation spells out how we can use thresholded RBM networks to establish lower bounds on the size of an RBM network required to compute certain simple functions (i.e., real-valued functions that represent certain Boolean functions):
Proposition 4. If no thresholded RBM network of size $m$ can compute some Boolean function $f(x)$ with margin $\delta$, then no RBM network of the same size can compute any real-valued function $g$ which represents $f$ with margin $\delta$.

This statement clearly holds if we replace each instance of “RBM network” with “hardplus RBM network” above.

Using Theorem 3, we can prove a more interesting result which states that any lower bound result for thresholded hardplus RBMs implies a somewhat weaker lower bound result for standard RBM networks:

Proposition 5. If no thresholded hardplus RBM network of size $\leq 4m^2 \log (2m/\delta)/\delta + m$ with parameters bounded in magnitude by $C$ ($C$ can be $\infty$) can compute some Boolean function $f(x)$ with margin $\delta/2$, then no RBM network of size $\leq m$ with parameters bounded in magnitude by $C$ can compute a function which represents $f$ with margin $\delta$.

This proposition implies that any exponential lower bound on the size of a thresholded hardplus RBM network will yield an exponential lower bound for (softplus) RBM networks that compute functions of the given form, provided that the margin $\delta$ is bounded from below by some function of the form $1/poly(n)$.

Intuitively, if $f$ is a Boolean function and no RBM network of size $m$ can compute a real-valued function that represents $f$ (with a margin $\delta$), this means that no RBM of size $m$ can represent any distribution where the log probability of each member of $\{x|f(x) = 1\}$ is at least $2\delta$ higher than each member of $\{x|f(x) = 0\}$. In other words, RBMs of this size cannot generate any distribution where the two “classes” implied by $f$ are separated in log probability by more than $2\delta$.

2.6 RBM networks versus standard neural networks

Viewing the RBM log probability function through the formalism of neural networks (or real-valued circuits) allows us to make use of known results for general neural networks, and helps highlight important differences between what an RBM can effectively “compute” (via its log probability) and what a standard neural network can compute.

There is a rich literature studying the complexity of various forms of neural networks, with diverse classes of activation functions, e.g., [Maass (1992); Maass et al. (1994); Hajnal et al. (1993)]. RBM networks are distinguished from these, primarily because they have a single hidden layer and because the upper level weights are constrained to be 1.

For some activation functions this restriction may not be significant, but for soft/hard-plus neurons, whose output is always positive, it makes particular computations much more awkward (or perhaps impossible) to express efficiently. Intuitively, the $j^{th}$ softplus neuron acts as a “feature detector”, which when “activated” by an input s.t. $x^\top w_j + b_j \gg 0$, can only contribute positively to the log probability of $x$, according to an (asymptotically) affine function of $x$ given by that neuron’s input. For example, it is easy to design an RBM network that can (approximately) output 1 for input $x = 0$ and 0 otherwise (i.e., have a single hidden neuron with weights $-M$ for a large $M$ and bias $b$ such that $\text{soft}(b) = 1$), but it is not immediately obvious how an RBM network could efficiently compute (or approximate) the function which is 1 on all inputs except $x = 0$, and 0 otherwise (it turns out that a non-obvious construction exists for $m = n$). By comparison, standard threshold networks only requires 1 hidden neuron to compute such a function.

In fact, it is easy to show [1] that without the constraint on upper level weights, an RBM network would be, up to a linear factor, at least as efficient at representing real-valued functions as a neural network with 1 hidden layer of threshold neurons. From this, and from Theorem 4.1 of [Maass et al. (1994)], it follows that a thresholded RBM network is, up to a polynomial increase in size, at least as efficient at computing Boolean functions as 1-hidden layer neural networks with any “sigmoid-like” activation function [2] and polynomially bounded weights.

1 To see this, note that we could use 2 softplus neurons to simulate a single neuron with a “sigmoid-like” activation function (i.e., by setting the weights that connect them to the output neuron to have opposite signs). Then, by increasing the size of the weights so the sigmoid saturates in both directions for all inputs, we could simulate a threshold function arbitrarily well, thus allowing the network to compute any function computable by a one hidden layer threshold network while only using only twice as many neurons.

2 This is a broad class and includes the standard logistic sigmoid. See [Maass et al. 1994] for a precise technical definition.
2.7 Simulating hardplus RBM networks by a one-hidden-layer threshold network

Here we provide a natural simulation of hardplus RBM networks by threshold networks with one hidden layer. Because this is an efficient (polynomial) and exact simulation, it implies that a hardplus RBM network can be no more powerful than a threshold network with one hidden layer, for which several lower bound results are already known.

**Theorem 6.** Let \( f \) be a real-valued function computed by a hardplus RBM network of size \( m \). Then \( f \) can be computed by a single hidden layer threshold network, of size \( mn \). Furthermore, if the weights of the RBM network have magnitude at most \( C \), then the weights of the corresponding threshold network have magnitude at most \((n+1)C\).

3 \( n^2 + 1 \)-sized RBM networks can compute any symmetric function

In this section we present perhaps the most surprising results of this paper: a construction of an \( n^2 \)-sized RBM network (or hardplus RBM network) for computing any given symmetric function of \( x \). Here, a symmetric function is defined as any real-valued function whose output depends only on the number of 1-bits in the input \( x \). This quantity is denoted \( X \equiv \sum_i x_i \). A well-known example of a symmetric function is PARITY.

Symmetric functions are already known\(^3\) to be computable by single hidden layer threshold networks (Hajnal et al., 1993) with \( m = n \). Meanwhile (qualified) exponential lower bounds on \( m \) exist for functions which are only slightly more complicated (Hajnal et al., 1993; Forster, 2002).

Given that hardplus RBM networks appear to be strictly less expressive than such threshold networks (as discussed in Section 2.6), it is surprising that they can nonetheless efficiently compute functions that test the limits of what those networks can compute efficiently.

**Theorem 7.** Let \( f : \{0, 1\}^n \to \mathbb{R} \) be a symmetric function defined by \( f(x) = t_k \) for \( \sum_i x_i = k \). Then (i) there exists a hardplus RBM network, of size \( n^2 + 1 \), and with weights polynomial in \( n \) and \( t_1, \ldots, t_k \) that computes \( f \) exactly, and (ii) for every \( \epsilon \) there is a softplus RBM network of size \( n^2 + 1 \), and with weights polynomial in \( n, t_1, \ldots, t_k \) and \( \log(1/\epsilon) \) that computes \( f \) within an additive error \( \epsilon \).

The high level idea of this construction is as follows. Our hardplus RBM network consists of \( n \) “building blocks”, each composed of \( n \) hardplus neurons, plus one additional hardplus neuron, for a total size of \( m = n^2 + 1 \). Each of these building blocks is designed to compute a function of the form:

\[
\max(0, \gamma X(e - X))
\]

for parameters \( \gamma > 0 \) and \( e > 0 \). This function, examples of which are illustrated in Figure 2, is quadratic from \( X = 0 \) to \( X = e \) and is 0 otherwise.

The main technical challenge is then to choose the parameters of these building blocks so that the sum of \( n \) of these “clipped quadratics”, plus the output of the extra hardplus neuron (which handles

---

\(^3\)The construction in [Hajnal et al., 1993] is only given for Boolean-valued symmetric functions but can be generalized easily.
the $X = 0$ case), yields a function that matches $f$, up to a additive constant (which we then fix by setting the bias $B$ of the output neuron). This would be easy if we could compute more general clipped quadratics of the form $\max(0, \gamma(X - q)(e - X))$, since we could just take $q = k - 1/2$ and $e = k + 1/2$ for each possible value $k$ of $X$. But the requirement that $g = 0$ makes this more difficult since significant overlap between non-zero regions of these functions will be unavoidable. Figure 2 depicts an example of the solution to this problem as given in our proof of Theorem 7.

Note that this construction is considerably more complex than the well-known construction used for computing symmetric functions with 1 hidden layer threshold networks [Hajnal et al. (1993)]. While we cannot prove that ours is the most efficient possible construction RBM networks, we can prove that a construction directly analogous to the one used for 1 hidden layer threshold networks—where each individual neuron computes a symmetric function—cannot possibly work for RBM networks.

To see this, first observe that any neuron that computes a symmetric function must compute a function of the form $g(\beta X + b)$, where $g$ is the activation function and $\beta$ is some scalar. Then noting that both $\text{soft}(y)$ and $\text{hard}(y)$ are convex functions of $y$, and that the composition of an affine function and a convex function is convex, we have that each neuron computes a convex function of $X$. Then because the positive sum of convex functions is convex, the output of the RBM network (which is the unweighted sum of the output of its neurons, plus a constant) is itself convex in $X$. Thus the symmetric functions computable by such RBM networks must be convex in $X$, a severe restriction which rules out most examples.

4 Lower bounds on the size of RBM networks for certain functions

4.1 Existential results

In this section we prove a result which establishes the existence of functions which cannot be computed by RBM networks that are not exponentially large.

Instead of identifying non-representable distributions as lying in the complement of some low-dimensional manifold (as was done previously), we will establish the existence of Boolean functions which cannot be represented with a sufficiently large margin by the output of any sub-exponentially large RBM network. However, this result, like previous such existential results, will say nothing about what these Boolean functions actually look like.

To prove this result, we will make use of Proposition 5 and a classical result of Muroga [1971] which allows us to discretize the incoming weights of a threshold neuron (without changing the function it computes), thus allowing us to bound the number of possible Boolean functions computable by 1-layer threshold networks of size $m$.

Theorem 8. Let $F_{m,n}$ represent the set of those Boolean functions on $\{0, 1\}^n$ that can be computed by a thresholded RBM network of size $m$ with margin $\delta$. Then, there exists a fixed number $K$ such that,

$$|F_{m,n}| \leq 2^{\text{poly}(s(m,n,\delta))}, \text{ where } s(m,\delta,n) = \frac{4m^2n}{\delta} \log \left( \frac{2m}{\delta} \right) + m.$$

In particular, when $m^2 \leq \delta 2^{\alpha n}$, for any constant $\alpha < 1/2$, the ratio of the size of the set $F_{m,n}$ to the total number of Boolean functions on $\{0, 1\}^n$ (which is $2^{2^n}$), rapidly converges to zero with $n$.

4.2 Qualified lower bound results for the IP function

While interesting, existential results such as the one above does not give us a clear picture of what a particular hard-to-compute function for RBM networks might look like. Perhaps these functions will resemble purely random maps without any interesting structure. Perhaps they will consist only of functions that require exponential time to compute on a Turing machine, or even worse, ones that are non-computable. In such cases, not being able to compute such functions would not constitute a meaningful limitation on the expressive efficiency of RBM networks.

In this sub-section we present strong evidence that this is not the case by exhibiting a simple Boolean function that provably requires exponentially many neurons to be computed by a thresholded RBM network, provided that the margin is not allowed to be exponentially smaller than the weights. Prior to these results, there was no formal separation between the kinds of unnormalized log-likelihoods realizable by polynomially sized RBMs, and the class of functions computable efficiently by almost any reasonable model of computation, such as arbitrarily deep Boolean circuits.
The Boolean function we will consider is the well-known “inner product mod 2” function, denoted $IP(x)$, which is defined as the parity of the the inner product of the first half of $x$ with the second half (we assume for convenience that $n$ is even). This function can be thought of as a strictly harder to compute version of PARITY (since PARITY is trivially reducible to it), which as we saw in Section 4 can be efficiently computed by thresholded RBM network (indeed, an RBM network can efficiently compute any possible real-valued representation of PARITY). Intuitively, $IP(x)$ should be harder than PARITY, since it involves an extra “stage” or “layer” of sequential computation, and our formal results with RBMs agree with this intuition.

There are many computational problems that $IP$ can be reduced to, so showing that RBM networks cannot compute $IP$ thus proves that RBMs cannot efficiently model a wide range of distributions whose unnormalized log-likelihoods are sufficiently complex in a computational sense. Examples include distributions whose density involves the computation of functions such as binary multiplication, graph connectivity, and more (see Corollary 3.5 of [Hajnal et al., 1993]).

Using the simulation of hardplus RBM networks by 1 hidden layer threshold networks (Theorem 6, and Proposition 5) and an existing result about the hardness of computing $IP$ by 1 hidden layer thresholded networks of bounded weights due to [Hajnal et al., 1993], we can prove the following basic result:

**Theorem 9.** If $m < \min \left\{ \frac{2n^{1/2}}{\epsilon}, \frac{2n^{1/3}}{\epsilon^{1/2} \log(2^{2/\delta})}, \frac{2n^{1/2}}{\epsilon^{1/4} \log(2^{2/\delta})} \right\}$ then no RBM network of size $m$, whose weights are bounded in magnitude by $C$, can compute a function which represents $n$-dimensional $IP$ with margin $\delta$. In particular, for $C$ and $1/\delta$ bounded by polynomials in $n$, for $n$ sufficiently large, this condition is satisfied whenever $m < 2^{(1/3-\epsilon)n}$ for some $\epsilon > 0$.

While the above theorem is easy to prove from known results and the simulation/hardness results given in previous sections, by generalizing the techniques used in [Hajnal et al., 1993], we can (with much more effort) derive a stronger result. This gives an improved bound on $m$ and lets us partially relax the magnitude bound on parameters so that they can be arbitrarily negative:

**Theorem 10.** If $m < \frac{\delta}{2 \max \{\log 2, nC + \log 2\}} \cdot 2^{n/4}$, then no RBM network of size $m$, whose weights are upper bounded in value by $C$, can compute a function which represents $n$-dimensional $IP$ with margin $\delta$. In particular, for $C$ and $1/\delta$ bounded by polynomials in $n$, for $n$ sufficiently large, this condition is satisfied whenever $m < 2^{(1/4-\epsilon)n}$ for some $\epsilon > 0$.

The general theorem we use to prove this second result (Theorem 17 in the Appendix) requires only that the neural network have 1 hidden layer of neurons with activation functions that are monotonic and contribute to the top neuron (after multiplication by the outgoing weight) a quantity which can be bounded by a certain exponentially growing function of $n$ (that also depends on $\delta$). Thus this technique can be applied to produce lower bounds for much more general types of neural networks, and thus may be independently interesting.

It is also interesting to note that Theorem 17 appears to be tight in the sense that none of the hypotheses can be removed. That is, for neurons with general non-montonic activation functions, or for neurons with monotonic activation functions whose output magnitude violates the aforementioned bounds, there are example networks that can efficiently compute any real-valued function. Thus, to improve this result (e.g. removing the weight bounds) it appears one would need to use a stronger property of the particular activation function than monotonicity.

### 5 Conclusions and Future Work

In this paper we significantly advanced the theoretical understanding of the representational efficiency of RBMs. We treated the RBM’s unnormalized log likelihood as a neural network which allowed us to relate an RBM’s representational efficiency to that of threshold networks, which are much better understood. We showed that, quite surprisingly, RBMs can efficiently represent distributions that are given by symmetric functions such as PARITY, but cannot efficiently represent distributions which are slightly more complicated, assuming an exponential bound on the weights. This provides rigorous justification for the use of potentially more expressive/deeper generative models.

Going forward, some promising research directions and open problems include characterizing the expressive power of Deep Boltzmann Machines and more general Boltzmann machines, and proving an exponential lower bound for some specific distribution without any qualifications on the weights.

### Acknowledgments

This research was supported by NSERC. JM is supported by a Google Fellowship; AC by a Ramanujan Fellowship of the DST, India.
References


