Representational Power of Restricted Boltzmann Machines and Deep Belief Networks

Nicolas Le Roux and Yoshua Bengio
Presented by Colin Graber

Introduction

- Representational abilities of functions with some sort of compositional structure is a well-studied problem
  - Neural networks, kernel machines, digital circuits
- 2-level architectures of some of these have been shown to be able to represent any function
- Efficiency of representation has been shown to be improved as depth increases
- What about for Restricted Boltzmann Machines and Deep Belief Networks?

Questions Addressed By Paper

1. What sorts of distributions can be represented by Restricted Boltzmann Machines?
2. What benefits do adding additional layers to RBMs (thus creating Deep Belief Networks) give us?

Summary of Main Results

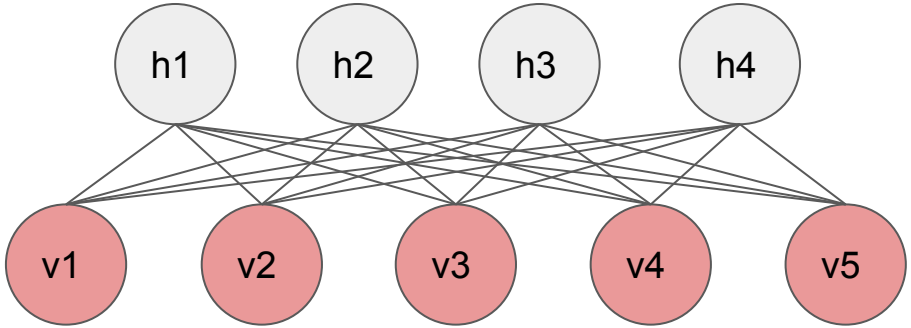
Restricted Boltzmann Machines:

- Increasing the number of hidden units improves representational ability
- With an unbounded number of units, any distribution over $\{0,1\}^n$ can be approximated arbitrarily well

Deep Belief Networks:

- Adding additional layers using greedy contrastive divergence training does not provide additional benefit
- There remain open questions about the benefits additional layers add

Recap: Restricted Boltzmann Machines



- Bipartite graphs consisting of visible units ($\mathbf{v}$) and hidden units ($\mathbf{h}$)
- The joint distribution has the following form (where $E(\mathbf{v}, \mathbf{h})$ is called the energy of the state $(\mathbf{v}, \mathbf{h})$:

$$p(\mathbf{v}, \mathbf{h}) \propto \exp(-E(\mathbf{v}, \mathbf{h})) = e^{\mathbf{h}^T W \mathbf{v} + b^T v + c^T \mathbf{h}}$$

Recap: Restricted Boltzmann Machines (2)

- The structure of the model simplifies the computation of certain values:
  - $P(\mathbf{v}|\mathbf{h}) = \prod_j P(v_j|\mathbf{h})$
  - $P(v_j = 1|\mathbf{h}) = \text{sigm}(b_j + \sum_i W_{ij} h_i)$
  - $P(\mathbf{h}|\mathbf{v}) = \prod_i P(h_i|\mathbf{v})$
  - $P(h_i = 1|\mathbf{v}) = \text{sigm}(c_i + \sum_j W_{ij} v_j)$
- In this paper, all units have values in {0, 1}

Recap: Deep Belief Networks

- Essentially, RBMs with additional layers of hidden units
- The joint distribution is written in the following way:

$$p(\mathbf{v}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \ldots, \mathbf{h}^{(\ell)}) = P(\mathbf{v}|\mathbf{h}^{(1)})P(\mathbf{h}^{(1)}|\mathbf{h}^{(2)}) \ldots P(\mathbf{h}^{(\ell-2)}|\mathbf{h}^{(\ell-1)})p(\mathbf{h}^{(\ell-1)}, \mathbf{h}^{(\ell)})$$

- $p(\mathbf{h}^{(l\text{-}1)}, \mathbf{h}^{(l)})$ (i.e. the marginal distribution over the top two layers) is an RBM
- Note on notation: we can define $\mathbf{h}^{(0)}$ to equal $\mathbf{v}$

Recap: Deep Belief Networks (2)

- As for RBMs, the structure of the model makes certain computations simpler:

$$P(\mathbf{h}^{(k)}|\mathbf{h}^{(k+1)}) = \prod_i P(\mathbf{h}_i^{(k)}|\mathbf{h}^{(k+1)})$$

$$P(\mathbf{h}_i^{(k)} = 1|\mathbf{h}^{(k+1)}) = \text{sigm}\left(b_i^{(k)} + \sum_j W_{ij}^{(k)}\mathbf{h}_j^{(k+1)}\right)$$

Summary of Main Results

Restricted Boltzmann Machines:

- Increasing the number of hidden units improves representational ability
- With an unbounded number of units, any distribution over $\{0,1\}^n$ can be approximated arbitrarily well

Deep Belief Networks:

- Adding additional layers using greedy contrastive divergence training does not provide additional benefit
- There remain open questions about the benefits additional layers add

What do we mean by “Representational Ability”?

- We have some empirical distribution $p_0(\mathbf{v})$ that is defined by the observed data
- A RBM represents a marginal distribution $p(\mathbf{v})$ over the visible units
- “Quality of representation” is measured by the KL divergence between p_0 and p (which we want to be small)
- Decreasing KL divergence is equivalent to increasing log-likelihood of data:

$$KL(p_0||p) = \sum_{\mathbf{v}} p_0(\mathbf{v}) \log \frac{p_0(\mathbf{v})}{p(\mathbf{v})} = -H(p_0) - \frac{1}{N} \sum_{i=1}^{N} \log p(\mathbf{v}^{(i)})$$

Some Notation

R_p : the set of RBMs having marginal distribution $p(\mathbf{v})$

$R_{p_{w,c}}$: the set of RBMs obtained by adding a hidden unit with weight w and bias c to an RBM from R_p

$p_{w,c}$: The marginal distribution over visible units for any RBM in R_p

$$z(\mathbf{v}, \mathbf{h}) = \exp(\mathbf{h}^T W \mathbf{v} + B^T \mathbf{v} + C^T \mathbf{h})$$

Equivalence Classes of RBMs

Lemma 2.1. Let R_p be the equivalence class containing the RBMs whose associated marginal distribution over the visible units is p. The operation of adding a hidden unit to an RBM of R_p preserves the equivalence class. Thus, the set of RBMs composed of an RBM of R_p and an additional hidden unit is also an equivalence class (meaning that all the RBMs of this set have the same marginal distribution over visible units).

Takeaway - the results we are about to prove are independent of the exact form of the RBM

Effect of adding unit with infinite negative bias

Lemma 2.2. Let p be the distribution over binary vectors $\mathbf{v}$ in $\{0, 1\}^d$, obtained with an RBM R_p and let $p_{w,c}$ be the distribution obtained when adding a hidden unit with weights w and bias c to R_p. Then $\forall p, \forall w \in \mathrm{R^d}, p = p_{w,-\infty}$

Effect of adding hidden units

Theorem 2.3. Let p_0 be an arbitrary distribution over $\{0,1\}^n$ and let R_p be an RBM with marginal distribution p over the visible units such that $KL(p_0 \| p) > 0$. Then there exists an RBM $Rp_{w,c}$ composed of R_p and an additional hidden unit with parameters (w,c) whose marginal distribution $p_{w,c}$ over the visible units achieves $KL(p_0 \| p_{w,c}) < KL(p_0 \| p)$

Proof Sketch (Theorem 2.3)

1. Write down definition of $KL(p_0 || p_{w,c})$
2. Rearrange to get expression of form

   $KL(p_0 || p_{w,c})$ - $KL(p_0 || p) = Z$

3. Show that Z is negative

Proof of Theorem 2.3

Step 1: Write $KL(p_0||p_{w,c})$ in terms of $KL(p_0||p)$

We start with the definition of KL divergence:

$$KL(p_0||p_{w,c}) = \sum_{\mathbf{v}} p_0(\mathbf{v}) \log p_0(\mathbf{v}) - \sum_{\mathbf{v}} p_0(\mathbf{v}) \log p_{w,c}(\mathbf{v})$$

Let's expand this term

Proof of Theorem 2.3 (2)

We can “push in” the sum corresponding to the new hidden unit:

$$
\begin{aligned}
p_{w,c}(\mathbf{v}) &= \frac{\sum_{\widetilde{\mathbf{h}}} \exp\left(\mathbf{h}^T W \mathbf{v} + h_{n+1} w^T \mathbf{v} + B^T \mathbf{v} + C^T \mathbf{h} + c h_{n+1}\right)}{\sum_{\widetilde{\mathbf{h}^{(0)}}, \mathbf{v}^0} \exp\left(\mathbf{h}^{(0)^T} W \mathbf{v}^0 + h^{(0)}_{n+1} w^T \mathbf{v}^0 + B^T \mathbf{v}^0 + C^T \mathbf{h}^{(0)} + c h^{(0)}_{n+1}\right)} \\
&= \frac{\sum_{\mathbf{h}} z(\mathbf{v}, \mathbf{h}) \left(1 + \exp\left(w^T \mathbf{v} + c\right)\right)}{\sum_{\mathbf{h}^{(0)}, \mathbf{v}^0} z(\mathbf{v}^0, \mathbf{h}^{(0)}) \left(1 + \exp\left(w^T \mathbf{v}^0 + c\right)\right)} \\
&= \frac{\left(1 + \exp\left(w^T \mathbf{v} + c\right)\right) \sum_{\mathbf{h}} z(\mathbf{v}, \mathbf{h})}{\sum_{\mathbf{v}^0, \mathbf{h}^{(0)}} \left(1 + \exp\left(w^T \mathbf{v}^0 + c\right)\right) z(\mathbf{v}^0, \mathbf{h}^{(0)})}
\end{aligned}
$$

Proof of Theorem 2.3 (3)

Substituting this into our earlier equation:

$$
\begin{aligned}
KL(p_0||p_{w,c}) &= \sum_{\mathbf{v}} p_0(\mathbf{v}) \log p_0(\mathbf{v}) - \sum_{\mathbf{v}} p_0(\mathbf{v}) \log p_{w,c}(\mathbf{v}) \\
&= -H(p_0) - \sum_{\mathbf{v}} p_0(\mathbf{v}) \log \left(\frac{\left(1 + \exp\left(w^T \mathbf{v} + c\right)\right) \sum_{\mathbf{h}} z(\mathbf{v}, \mathbf{h})}{\sum_{\mathbf{v}^0, \mathbf{h}^{(0)}} \left(1 + \exp\left(w^T \mathbf{v}^0 + c\right)\right) z(\mathbf{v}^0, \mathbf{h}^{(0)})} \right) \\
&= -H(p_0) - \sum_{\mathbf{v}} p_0(\mathbf{v}) \log \left(1 + \exp\left(w^T \mathbf{v} + c\right)\right) - \sum_{\mathbf{v}} p_0(\mathbf{v}) \log \left(\sum_{\mathbf{h}} z(\mathbf{v}, \mathbf{h}) \right) \\
&\quad + \sum_{\mathbf{v}} p_0(\mathbf{v}) \log \left(\sum_{\mathbf{v}^0, \mathbf{h}^{(0)}} \left(1 + \exp\left(w^T \mathbf{v}^0 + c\right)\right) z(\mathbf{v}^0, \mathbf{h}^{(0)}) \right)
\end{aligned}
$$

Proof of Theorem 2.3 (4)

We want to simplify $\log(1 + \exp(w^{\mathrm{T}}\mathbf{v} + c))$.

Consider the Taylor series expansion of the natural logarithm:

$$\ln(z) = \frac{(z-1)^1}{1} - \frac{(z-1)^2}{2} + \frac{(z-1)^3}{3} - \frac{(z-1)^4}{4} + \cdots$$

If we assume is $w^{\mathrm{T}}\mathbf{v} + c$ is a large negative value for all v (which we can, since we set the parameter values), then we can use the approximation

$$\log(1 + x) = x + o_{x \to 0}(x)$$

Proof of Theorem 2.3 (5)

Second term:

$$\sum_{\mathbf{v}} p_0(\mathbf{v}) \log\left(1+\exp\left(w^T\mathbf{v}+c\right)\right) = \sum_{\mathbf{v}} p_0(\mathbf{v}) \exp\left(w^T\mathbf{v}+c\right) + o_{c\to-\infty}\left(\exp(c)\right)$$

Proof of Theorem 2.3 (6)

Last term:

$$\left(\sum_{\mathbf{v}} p_0(\mathbf{v})\right) \log \left(\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} \left(1+\exp\left(w^T\mathbf{v}^0+c\right)\right) z(\mathbf{v}^0,\mathbf{h}^{(0)})\right)$$

$$= \log\left(\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} z(\mathbf{v}^0,\mathbf{h}^{(0)})\right) + \log\left(1+\frac{\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} \exp\left(w^T\mathbf{v}^0+c\right) z(\mathbf{v}^0,\mathbf{h}^{(0)})}{\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} z(\mathbf{v}^0,\mathbf{h}^{(0)})}\right)$$

$$= \log\left(\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} z(\mathbf{v}^0,\mathbf{h}^{(0)})\right) + \frac{\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} \exp\left(w^T\mathbf{v}^0+c\right) z(\mathbf{v}^0,\mathbf{h}^{(0)})}{\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} z(\mathbf{v}^0,\mathbf{h}^{(0)})} + o_{c\to-\infty}\left(\exp(c)\right)$$

Proof of Theorem 2.3 (7)

More simplification:

$$\frac{\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} \exp\left(w^T\mathbf{v}^0 + c\right) z(\mathbf{v}^0,\mathbf{h}^{(0)})}{\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} z(\mathbf{v}^0,\mathbf{h}^{(0)})} = \sum_{\mathbf{v}} \exp\left(w^T\mathbf{v} + c\right) \frac{\sum_{\mathbf{h}^{(0)}} z(\mathbf{v},\mathbf{h}^{(0)})}{\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} z(\mathbf{v}^0,\mathbf{h}^{(0)})}$$

$$= \sum_{\mathbf{v}} \exp\left(w^T\mathbf{v} + c\right) p(\mathbf{v})$$

Proof of Theorem 2.3 (8)

Finally, we can substitute in everything we just derived:

$$\begin{aligned} KL(p_0||p_{w,c}) &= -H(p_0) - \sum_{\mathbf{v}} p_0(\mathbf{v}) \exp\left(w^T\mathbf{v} + c\right) + \sum_{\mathbf{v}} p(\mathbf{v}) \exp\left(w^T\mathbf{v} + c\right) + o_{c\to-\infty}\left(\exp(c)\right) \\ &- \sum_{\mathbf{v}} p_0(\mathbf{v}) \log\left(\sum_{\mathbf{h}} z(\mathbf{v},\mathbf{h})\right) + \log\left(\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} z(\mathbf{v}^0,\mathbf{h}^{(0)})\right) \\ &= KL(p_0||p) + \sum_{\mathbf{v}} \exp\left(w^T\mathbf{v} + c\right)\left(p(\mathbf{v}) - p_0(\mathbf{v})\right) + o_{c\to-\infty}\left(\exp(c)\right) \end{aligned}$$

Which gives us what we wanted:

$$KL(p_0||p_{w,c}) - KL(p_0||p) = \exp(c) \sum \exp\left(w^T\mathbf{v}\right)\left(p(\mathbf{v}) - p_0(\mathbf{v})\right) + o_{c\to-\infty}\left(\exp(c)\right)$$

Proof of Theorem 2.3 (9)

Now, we want to show that there exists a w such that

$\sum_{\mathbf{v}} \exp\left(w^T \mathbf{v}\right) \left(p(\mathbf{v}) - p_0(\mathbf{v})\right)$ is negative.

Since $p \neq p_0$, there exists an input $\hat{\mathbf{v}}$ such that $p(\hat{\mathbf{v}}) < p_0(\hat{\mathbf{v}})$

Using this fact, we will now prove that a positive scalar a exists such that defining $\hat{w} = a\left(\hat{\mathbf{v}} - \frac{1}{2}e\right)$ (with $e = [1, \ldots, 1]^{\mathrm{T}}$) gives us the condition above.

Proof of Theorem 2.3 (10)

We can decompose the target sum in the following way:

$$\sum_{\mathbf{v}} \exp\left(\widehat{w}^T\mathbf{v}\right)\left(p(\mathbf{v}) - p_0(\mathbf{v})\right) = \exp(\widehat{w}^T\widehat{\mathbf{v}})\left(\sum_{\mathbf{v}} \frac{\exp\left(\widehat{w}^T\mathbf{v}\right)}{\exp(\widehat{w}^T\widehat{\mathbf{v}})}\left(p(\mathbf{v}) - p_0(\mathbf{v})\right)\right)$$

$$= \exp(\widehat{w}^T\widehat{\mathbf{v}})\left(p(\widehat{\mathbf{v}}) - p_0(\widehat{\mathbf{v}}) + \sum_{\mathbf{v}\neq\widehat{\mathbf{v}}} \frac{\exp\left(\widehat{w}^T\mathbf{v}\right)}{\exp(\widehat{w}^T\widehat{\mathbf{v}})}\left(p(\mathbf{v}) - p_0(\mathbf{v})\right)\right)$$

This is negative

What about this?

Proof of Theorem 2.3 (11)

For $\mathbf{v} \neq \widehat{\mathbf{v}}$, we have:

$$\begin{aligned}
\frac{\exp(\widehat{w}^T \mathbf{v})}{\exp(\widehat{w}^T \widehat{\mathbf{v}})} &= \exp\left(\widehat{w}^T (\mathbf{v} - \widehat{\mathbf{v}})\right) \\
&= \exp\left(a \left(\widehat{\mathbf{v}} - \frac{1}{2} e\right)^T (\mathbf{v} - \widehat{\mathbf{v}})\right) \\
&= \exp\left(a \sum_i \left(\widehat{\mathbf{v}}_i - \frac{1}{2}\right) (\mathbf{v}_i - \widehat{\mathbf{v}}_i)\right)
\end{aligned}$$

Proof of Theorem 2.3 (12)

Let's look at this term:

$$\exp\left(a\sum_i\left(\widehat{\mathbf{v}}_i-\frac{1}{2}\right)(\mathbf{v}_i-\widehat{\mathbf{v}}_i)\right)$$

- No matter which of the possible four assignments you give to $\widehat{\mathbf{v}}_i$ and $\mathbf{v}_i$, the terms of the sum are less than or equal to zero
- Thus, as a approaches infinity, this term approaches zero

Proof of Theorem 2.3 (13)

Thus, going back to the expression we want to prove is negative:

$$\begin{aligned}\sum_{\mathbf{v}} \exp\left(\widehat{w}^T\mathbf{v}\right)\left(p(\mathbf{v})-p_0(\mathbf{v})\right) &= \exp(\widehat{w}^T\widehat{\mathbf{v}})\left(\sum_{\mathbf{v}}\frac{\exp\left(\widehat{w}^T\mathbf{v}\right)}{\exp(\widehat{w}^T\widehat{\mathbf{v}})}\left(p(\mathbf{v})-p_0(\mathbf{v})\right)\right)\\ &= \exp(\widehat{w}^T\widehat{\mathbf{v}})\left(p(\widehat{\mathbf{v}})-p_0(\widehat{\mathbf{v}})+\sum_{\mathbf{v}\neq\widehat{\mathbf{v}}}\frac{\exp\left(\widehat{w}^T\mathbf{v}\right)}{\exp(\widehat{w}^T\widehat{\mathbf{v}})}\left(p(\mathbf{v})-p_0(\mathbf{v})\right)\right)\end{aligned}$$

We know the following:

$$\sum_{\mathbf{v}} \exp\left(\widehat{w}^T\mathbf{v}\right)\left(p(\mathbf{v})-p_0(\mathbf{v})\right) \sim_{a\to+\infty} \exp(\widehat{w}^T\widehat{\mathbf{v}})(p(\widehat{\mathbf{v}})-p_0(\widehat{\mathbf{v}})) < 0$$

Hence, an a exists which makes the difference in divergences negative.

Summary of Main Results

Restricted Boltzmann Machines:

- Increasing the number of hidden units improves representational ability
- With an unbounded number of units, any distribution over $\{0,1\}^n$ can be approximated arbitrarily well

Deep Belief Networks:

- Adding additional layers using greedy contrastive divergence training does not provide additional benefit
- There remain open questions about the benefits additional layers add

RBMs are Universal Approximators

Theorem 2.4. Any distribution over $\{0, 1\}^n$ can be approximated arbitrarily well (in the sense of the KL divergence) with an RBM with $k + 1$ hidden units where k is the number of input vectors whose probability is not 0.

Theorem 2.4: Proof Sketch

We construct a RBM in the following way:

- Each hidden unit is “assigned” one of the possible input vectors $\mathbf{v}_i$ such that, when $\mathbf{v}_i$ is the input:
  - All other hidden units have probability of zero of being “on”
  - The corresponding hidden unit has probability $\text{sigmoid}(\lambda_i)$ of being “on”
- Values for weights and λ parameters are chosen such that:
  - λ_i is tied with $p(\mathbf{v}_i)$
  - When all hidden units except for *i* are off, $p(\mathbf{v}_i \mid \mathbf{h}) = 1$
  - When all of the hidden units are off (which happens w/probability $1\text{-sigmoid}(\lambda_i)$),

Proof of Theorem 2.4

In the previous proof, we had:

$$p_{w,c}(\mathbf{v}) = \frac{\left(1 + \exp\left(w^T \mathbf{v} + c\right)\right) \sum_{\mathbf{h}} z(\mathbf{v}, \mathbf{h})}{\sum_{\mathbf{v}^0, \mathbf{h}^{(0)}} \left(1 + \exp\left(w^T \mathbf{v}^0 + c\right)\right) z(\mathbf{v}^0, \mathbf{h}^{(0)})}$$

Let $\tilde{\mathbf{v}}$ be an arbitrary input vector. Define a weight vector in the same way we did during the last proof; namely,

$$\widehat{w} = a\left(\tilde{\mathbf{v}} - \frac{1}{2}\right)$$

Proof of Theorem 2.4 (2)

Define another parameter $\widehat{c} = -\widehat{w}^T \tilde{\mathbf{v}} + \lambda$, with $\lambda \in \mathbb{R}$. Note the following fact, which we will be using next:

$$\lim_{a \to \infty} 1 + \exp(\widehat{w}^T \mathbf{v} + \widehat{c}) = 1 \quad \text{for } \mathbf{v} \neq \tilde{\mathbf{v}}$$

$$1 + \exp(\widehat{w}^T \tilde{\mathbf{v}} + \widehat{c}) = 1 + \exp(\lambda)$$

Proof of Theorem 2.4 (3)

Using this fact, we get that, for $\mathbf{v} \neq \tilde{\mathbf{v}}$:

$$\begin{aligned}
\lim_{a\to\infty} p_{\widehat{w},\widehat{c}}(\mathbf{v}) &= \frac{\sum_{\mathbf{h}} z(\mathbf{v},\mathbf{h})}{\sum_{\mathbf{v}^0\neq\tilde{\mathbf{v}},\mathbf{h}^{(0)}} z(\mathbf{v}^0,\mathbf{h}^{(0)}) + \sum_{\mathbf{h}^{(0)}} \left(1+\exp\left(\widehat{w}^T\tilde{\mathbf{v}}+\widehat{c}\right)\right) z(\tilde{\mathbf{v}},\mathbf{h}^{(0)})} \\
&= \frac{\sum_{\mathbf{h}} z(\mathbf{v},\mathbf{h})}{\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} z(\mathbf{v}^0,\mathbf{h}^{(0)}) + \sum_{\mathbf{h}^{(0)}} \exp\left(\lambda\right) z(\tilde{\mathbf{v}},\mathbf{h}^{(0)})} \\
&= \frac{\sum_{\mathbf{h}} z(\mathbf{v},\mathbf{h})}{\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} z(\mathbf{v}^0,\mathbf{h}^{(0)})} \frac{1}{1+\exp(\lambda)\frac{\sum_{\mathbf{h}^{(0)}} z(\tilde{\mathbf{v}},\mathbf{h}^{(0)})}{\sum_{\mathbf{v}^0,\mathbf{h}^{(0)}} z(\mathbf{v}^0,\mathbf{h}^{(0)})}} \\
&= \frac{p(\mathbf{v})}{1+\exp(\lambda)p(\tilde{\mathbf{v}})}
\end{aligned}$$

Proof of Theorem 2.4 (4)

Using the same derivation, we get that

$$\lim_{a\to\infty} p_{\hat{w},\hat{c}}(\tilde{\mathbf{v}}) = \frac{[1+\exp(\lambda)]p(\tilde{\mathbf{v}})}{1+\exp(\lambda)p(\tilde{\mathbf{v}})} \qquad \lim_{a\to\infty} p_{\hat{w},\hat{c}}(\mathbf{v}) = \frac{p(\mathbf{v})}{1+\exp(\lambda)p(\tilde{\mathbf{v}})}$$

What did we learn?

We have figured out a way of adding a hidden unit to a RBM that increases the probability of a single input vector and decreases uniformly the probabilities of every single other input vector.

Additionally, if $p(\tilde{\mathbf{v}}) = 0$, then $p_{\hat{w},\hat{c}}(\tilde{\mathbf{v}}) = 0$ for all λ

Proof of Theorem 2.4 (5)

Let's build a RBM. First, let's index the input vectors in the following way:

$$p_0(\mathbf{v}_{k+1}) = \ldots = p_0(\mathbf{v}_{2^n}) = 0 < p_0(\mathbf{v}_1) \leq p_0(\mathbf{v}_2) \leq \ldots \leq p_0(\mathbf{v}_k)$$

We'll use p^i to represent the distribution of an RBM with i [added] hidden units

We start with a RBM with weights and biases all set to zero; this induces a uniform marginal distribution over the visible units:

$$p^0(\mathbf{v}_1) = \ldots = p^0(\mathbf{v}_{2^n}) = 2^{-n}$$

Proof of Theorem 2.4 (6)

Next, we add a hidden unit with the following parameters:

$$w_1 = a_1(\mathbf{v}_1 - \tfrac{1}{2}) \qquad\qquad c_1 = -w_1^T \mathbf{v}_1 + \lambda_1$$

As mentioned previously, this gives us:

$$\lim_{a_1 \to +\infty} p^1(\mathbf{v}_1) = \frac{[1 + \exp(\lambda_1)]2^{-n}}{1 + \exp(\lambda_1)2^{-n}}$$

$$\lim_{a_1 \to +\infty} p^1(\mathbf{v}_i) = \frac{2^{-n}}{1 + \exp(\lambda_1)2^{-n}} \quad \forall i \geq 2$$

Proof of Theorem 2.4 (7)

- Now, we can add another hidden node, this time based on the second input vector, thus giving us p^2
- We set λ_2 such that the following ratio is true:

$$\frac{p^2(\mathbf{v}_2)}{p^2(\mathbf{v}_1)} = \frac{p(\mathbf{v}_2)}{p(\mathbf{v}_1)}$$

- We can do this, since we can increase p(v2) arbitrarily and $\frac{p(\mathbf{v}_2)}{p(\mathbf{v}_1)} \geq \frac{p^1(\mathbf{v}_2)}{p^1(\mathbf{v}_1)}$
- This ratio will continue to be true as additional hidden nodes are added, since the probabilities of all vectors (besides the vector under consideration) are multiplied by the same factor at each step

Proof of Theorem 2.4 (8)

- After adding k hidden nodes, the following equations are true:

$$\frac{p^k(\mathbf{v}_k)}{p^k(\mathbf{v}_{k-1})} = \frac{p(\mathbf{v}_k)}{p(\mathbf{v}_{k-1})}, \quad \dots \quad , \frac{p^k(\mathbf{v}_2)}{p^k(\mathbf{v}_1)} = \frac{p(\mathbf{v}_2)}{p(\mathbf{v}_1)}$$

$$p^k(\mathbf{v}_{k+1}) = \quad \dots \quad = p^k(\mathbf{v}_{2^n})$$

- These imply that $p^k(\mathbf{v}_1) = \nu_k p(\mathbf{v}_1), \dots, p^k(\mathbf{v}_k) = \nu_k p(\mathbf{v}_k)$, where $\nu_k = 1 - (2^n - k)p^k(\mathbf{v}_{2^n})$

- Additionally, $\frac{p^k(\mathbf{v}_1)}{p^k(\mathbf{v}_{2^n})} = \frac{p^1(\mathbf{v}_1)}{p^1(\mathbf{v}_{2^n})} = 1 + \exp(\lambda_1)$ which implies

$$p^k(\mathbf{v}_1) = p(\mathbf{v}_1)[1 - (2^n - k)p^k(\mathbf{v}_{2^n})] = (1 + \exp(\lambda_1))p^k(\mathbf{v}_{2^n})$$

Proof of Theorem 2.4 (9)

A tiny bit of derivation gives us the following results:

$$p^k(\mathbf{v}_i) = \frac{p(\mathbf{v}_1)}{1+\exp(\lambda_1)+p(\mathbf{v}_1)(2^n-k)} \quad \text{for } i > k$$

$$p^k(\mathbf{v}_i) = p(\mathbf{v}_i)\frac{1+\exp(\lambda_1)}{1+\exp(\lambda_1)+p(\mathbf{v}_1)(2^n-k)} \quad \text{for } i \leq k$$

Using the logarithmic series identity around 0 again, we can then show that this RBM has the behavior we wanted:

$$KL(p||p^k) = \sum_i p(\mathbf{v}_i)\frac{(2^n-k)p(\mathbf{v}_i)}{1+\exp(\lambda_1)} + o(\exp(-\lambda_1)) \underset{\lambda_1 \to \infty}{\longrightarrow} 0$$

Summary of Main Results

Restricted Boltzmann Machines:

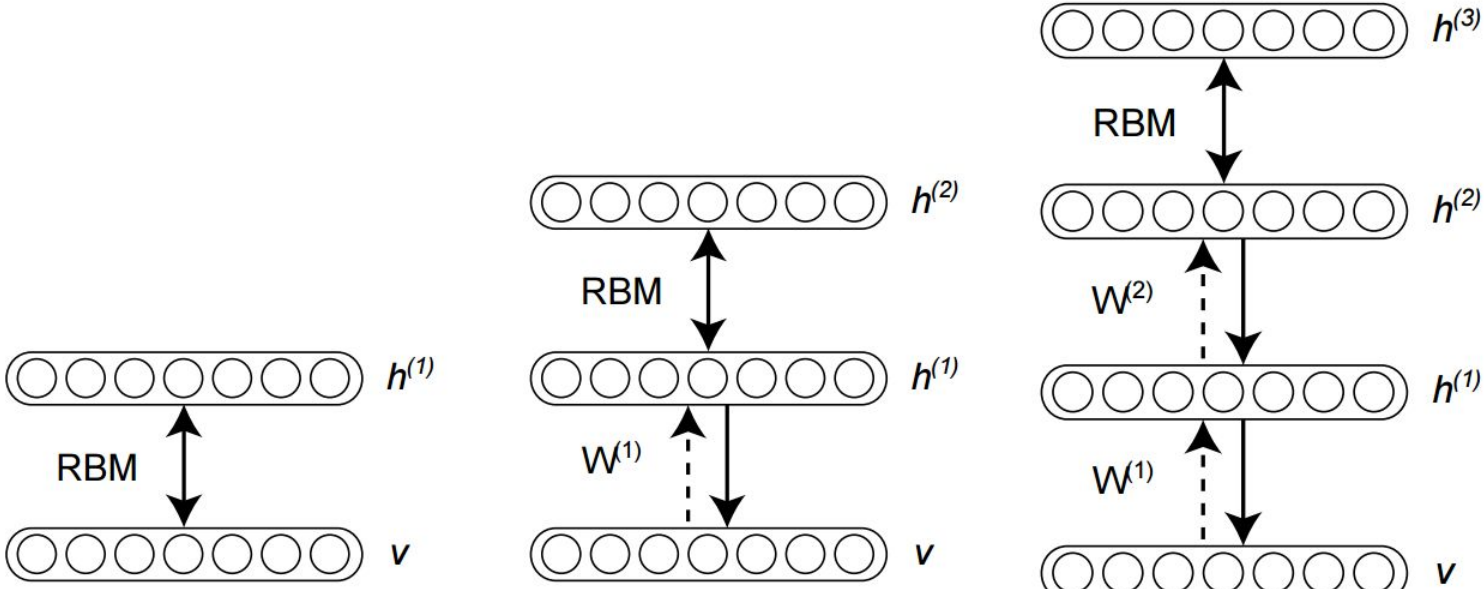
- Increasing the number of hidden units improves representational ability
- With an unbounded number of units, any distribution over $\{0,1\}^n$ can be approximated arbitrarily well

Deep Belief Networks:

- Adding additional layers using greedy contrastive divergence training does not provide additional benefit
- There remain open questions about the benefits additional layers add

Recap: Greedy Training Method for DBNs

- Proposed in (Hinton et al., 2006)
  - Training is completed by adding one layer at a time
  - When training a new layer, the weights of all previous layers are fixed & the top layer is trained as an RBM
- Problems with this method:
  - The training procedure does not take into account the fact that additional layers will be added in the future
  - This unnecessarily restricts the forms of the distributions that intermediate layers may learn



(a) Stage 1 (b) Stage 2 (c) Stage 3

DBN Greedy Training Objective

In the greedy training procedure, a lower bound to the likelihood (called the **variational bound**) is maximized. For example, for the second layer of a two-layer DBN, we have:

$$\log p(\mathbf{v}) \geq \sum_{\mathbf{h}^{(1)}} Q(\mathbf{h}^{(1)}|\mathbf{v}) \left[\log p(\mathbf{h}^{(1)}) + \log P(\mathbf{v}|\mathbf{h}^{(1)})\right] - \sum_{\mathbf{h}^{(1)}} Q(\mathbf{h}^{(1)}|\mathbf{v}) \log Q(\mathbf{h}^{(1)}|\mathbf{v})$$

Since we fix the weights of the first layer, the only term in this expression that we can optimize is $p(\mathbf{h}^{(1)})$

DBN Greedy Training Objective (2)

- It turns out that there's an analytic formulation for the optimal solution to $p(\mathbf{h}^{(1)})$:

$$p^*(\mathbf{h}^{(1)}) = \sum_{\mathbf{v}} p_0(\mathbf{v}) Q(\mathbf{h}^{(1)} | \mathbf{v})$$

- From theorem 2.4, we know this can be approximated arbitrarily well by some RBM
- If we use an RBM that approximates this value, what distribution is being modeled by our DBN?

Marginal Distribution achieved by this DBN

Proposition 3.1: In a 2-layer DBN, using a second layer RBM achieving $p^*(\mathbf{h}^{(1)})$, the model distribution p is equal to p_1, where p_1 is the distribution obtained by starting with p_0 clamped to the visible units $\mathbf{v}$, sampling $\mathbf{h}^{(1)}$ given $\mathbf{v}$, and then sampling $\mathbf{v}$ given $\mathbf{h}^{(1)}$.

Proof of Proposition 3.1

We start with the analytic formulation of the marginal distribution for the “top” RBM:

$$p^*(\mathbf{h}^{(1)}) = \sum_{\tilde{\mathbf{h}}^0} p_0(\tilde{\mathbf{h}}^0) Q(\mathbf{h}^{(1)}|\tilde{\mathbf{h}}^0)$$

Now substitute this into the expression for the marginal distribution over the bottom layer:

$$\begin{aligned} p(\mathbf{h}^{(0)}) &= \sum_{\mathbf{h}^{(1)}} P(\mathbf{h}^{(0)}|\mathbf{h}^{(1)}) p^*(\mathbf{h}^{(1)}) \\ &= \sum_{\tilde{\mathbf{h}}^0} p_0(\tilde{\mathbf{h}}^0) \sum_{\mathbf{h}^{(1)}} Q(\mathbf{h}^{(1)}|\tilde{\mathbf{h}}^0) P(\mathbf{h}^{(0)}|\mathbf{h}^{(1)}) \\ p(\mathbf{h}^{(0)}) &= p_1(\mathbf{h}^{(0)}) \end{aligned}$$

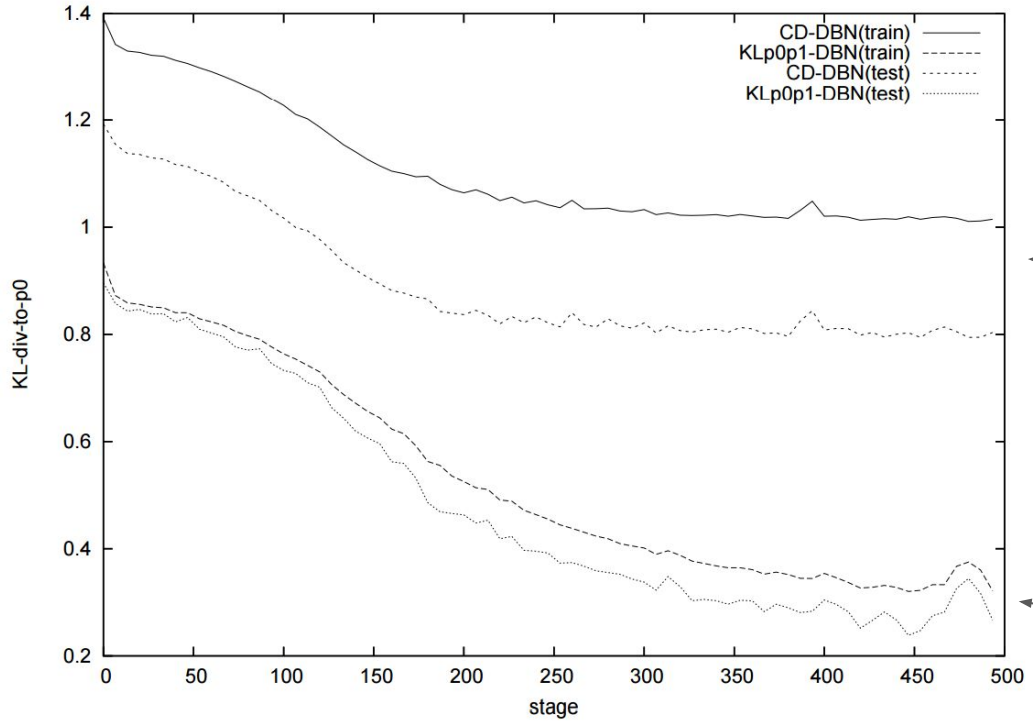
Takeaways from Proposition 3.1

- Using this training procedure, the best KL divergence we can achieve is $KL(p_0||p_1)$
- Achieving $KL(p_0||p_1) = 0$ requires that $p_0 = p_1$
  - In this case, we wouldn't need an extra layer!
- Does this mean there's no benefit to depth for RBMs?
  - Not necessarily - we may be able to do better by optimizing some other bound besides the variational bound
  - Also - the two layer network achieves its given approximation of p_0 using only one "up-down" step, while the one layer network achieves this after an infinite number of steps

Proposed Alternative Training Criterion

- Rather than use the previous training objective for intermediate layers, could it be better to try to optimize over $KL(p_0||p_1)$?
- Experiment:
  - Generated a toy dataset consisting of 60 bit vectors of length 10, with either one, two, or three consecutive bits turned on
  - Trained two two-layer DBNs with the same number of nodes per layer
  - The first used the contrastive divergence objective for the intermediate layer, while the second minimized $KL(p_0||p_1)$ using gradient descent
  - The second layer for both was trained using contrastive divergence

Experimental Results



Contrastive Divergence Results

KL Divergence Results

Summary of Main Results

Restricted Boltzmann Machines:

- Increasing the number of hidden units improves representational ability
- With an unbounded number of units, any distribution over $\{0,1\}^n$ can be approximated arbitrarily well

Deep Belief Networks:

- Adding additional layers using greedy contrastive divergence training does not provide additional benefit
- There remain open questions about the benefits additional layers add

Open Questions

Using an argument from an earlier paper, we know that every distribution that can be represented by an l-layer DBN with n units per layer can also be represented by an $(l+1)$-layer DBN with n units per layer. This begs the following questions:

- Are there distributions that can be represented by the latter but not the former?
- What distributions can be represented using an unbounded number of layers?

Summary of Main Results

Restricted Boltzmann Machines:

- Increasing the number of hidden units improves representational ability
- With an unbounded number of units, any distribution over $\{0,1\}^n$ can be approximated arbitrarily well

Deep Belief Networks:

- Adding additional layers using greedy contrastive divergence training does not provide additional benefit
- There remain open questions about the benefits additional layers add