Deep Belief Networks are Compact Universal Approximators

Franck Olivier Ndjakou Njeunje

Applied Mathematics and Scientific Computation

May 16, 2016

Outline

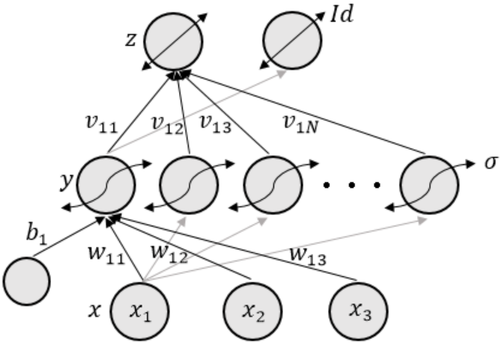
1. Introduction
2. Preliminaries
    - Universal approximation theorem
    - Definition
    - Deep Belief Network
3. Sutekever and Hinton's Method
    - Transfer of Probability
4. Le Roux and Bengio's Method
    - Gray Codes
    - Simultaneous Transfer of Probability
5. Conclusion

Introduction

- Machine learning (by Arthur Samuel - 1959)
- Neural Networks: Set of transformations

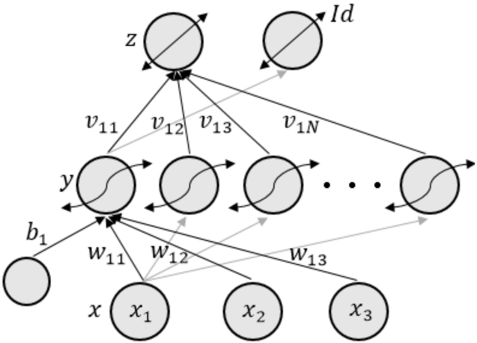
Introduction

- Machine learning (by Arthur Samuel - 1959)
- Neural Networks: Set of transformations

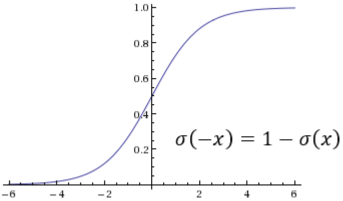


Introduction

- Machine learning (by Arthur Samuel - 1959)
- Neural Networks: Set of transformations



$$\sigma(x) = \frac{1}{1 + \exp(-x)} \qquad (1)$$



Universal approximation theorem [Cybe89]

Let $\varphi(\cdot)$ be a non constant, bounded, and monotonically-increasing continuous function. Let I_m denote the m-dimensional unit hypercube $[0,1]^m$. The space of continuous functions on I_m is denoted by $C(I_m)$. Then, given any function $f \in C(I_m)$ and $\varepsilon > 0$, there exists an integer N, real constants $v_i, b_i \in \mathbb{R}$ and real vectors $w_i \in \mathbb{R}^m$, where $i = 1, \cdots, N$ such that we may define:

$$F(x) = \sum_{i=1}^{N} v_i \varphi\left(w_i^T x + b_i\right) \tag{2}$$

as an approximate realization of the function f where f is independent of φ; that is,

$$|F(x) - f(x)| < \varepsilon \tag{3}$$

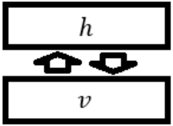
for all $x \in I_m$. In other words, functions of the form $F(x)$ are dense in $C(I_m)$.

Definition

- Generative probabilistic model (GPM)
  - Application: recognition, classification, and generation

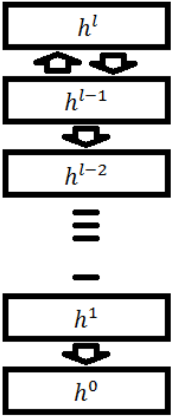
Definition

- Generative probabilistic model (GPM)
    - Application: recognition, classification, and generation
- Restricted Boltzmann machine (RBM) [LeBe08]
    - 2-layer GPM



Definition

- Generative probabilistic model (GPM)
  - Application: recognition, classification, and generation
- Restricted Boltzmann machine (RBM) [LeBe08]
  - 2-layer GPM
- Deep belief network (DBN) [HiOT06]
  - Multilayer GPM
  - First two layer form an RBM



Deep Belief Network

Let $\mathbf{h}^i$ represent the vector of hidden variable at layer i. The model is parametrized as follows:

$$P(\mathbf{h}^0, \mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^l) = P(\mathbf{h}^0|\mathbf{h}^1)P(\mathbf{h}^1|\mathbf{h}^2)\ldots P(\mathbf{h}^{l-2}|\mathbf{h}^{l-1})P(\mathbf{h}^l, \mathbf{h}^{l-1}). \tag{4}$$

The hidden layer $\mathbf{h}^i$ is a binary random vector with elements $\mathbf{h}^i_j$ and

$$P(\mathbf{h}^i|\mathbf{h}^{i+1}) = \prod_{j=1}^{n_i} P(\mathbf{h}^i_j|\mathbf{h}^{i+1}). \tag{5}$$

The element $\mathbf{h}^i_j$ is a *stochatic neuron* or whose binary activation is 1:

$$P(\mathbf{h}^i_j = 1|\mathbf{h}^{i+1}) = \sigma\left(b^i_j + \sum_{k=1}^{n_{i+1}} W^i_{jk}\mathbf{h}^{i+1}_k\right). \tag{6}$$

Universal approximator (DBN)

Let p^* be an arbitrary distribution over binary vectors of n bits. A deep belief network that has p^* as its marginal distribution over $\mathbf{h}^0$ is said to be a **universal approximator**.

This means that for any binary vector $\mathbf{x}$ of n bits, there exist weights and biases such that given $\varepsilon > 0$:

$$|P(\mathbf{h}^0 = \mathbf{x}) - p^*(\mathbf{x})| < \varepsilon \tag{7}$$

Sutekever and Hinton's Method 2008 [SuHi08]

Define an arbitrary sequence $(a_i)_{1 \leq i \leq 2^n}$ of binary vectors in $\{0,1\}^n$.

The goal is to find appropriate weights and biases such that the marginal distribution over the set of outputs to our DBN is the same as the probability distribution over the vectors $(a_i)_{1 \leq i \leq 2^n}$.

In the next few slides I will consider the following example for $n = 4$:

$$\begin{aligned} a_1 &= 1011, \quad p^*(a_1) = 0.1 && (8) \\ a_2 &= 1000, \quad p^*(a_2) = 0.05 && (9) \\ a_3 &= 1001, \quad p^*(a_3) = 0.01 && (10) \\ a_4 &= 1111, \quad p^*(a_4) = 0.02 && (11) \\ &\vdots && (12) \end{aligned}$$

Sutekever and Hinton's Method 2008 [SuHi08]

Consider two consecutive layers $\mathbf{h}$ and $\mathbf{v}$ of size n, with W_{ij} the weight linking unit v_i to unit h_j, b_i the bias of unit v_i and w a positive scalar.

For every positive scalar ε $(0 < \varepsilon < 1)$, there is a weight vector $W_{i,:}$ and a real b_i such that $P(v_i = h_i|\mathbf{h}) = 1 - \varepsilon$.

Indeed, setting:

- $W_{ii} = 2w$
- $W_{ij} = 0$ for $i \neq j$
- $b_i = -w$

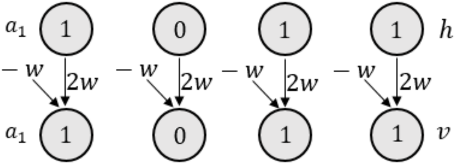
yields a total input to unit v_i of:

$$I(v_i, \mathbf{h}) = 2wh_i - w \qquad (13)$$

Therefore, if $w = \sigma^{-1}(1 - \varepsilon)$, we have $P(v_i = h_i|\mathbf{h}) = 1 - \varepsilon$.

Sutekever and Hinton's Method 2008 [SuHi08]

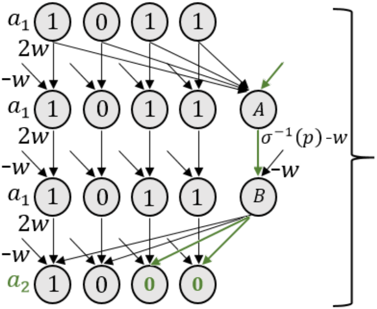
- $W_{ii} = 2w$
- $W_{ij} = 0$ for $i \neq j$
- $b_i = -w$



With $I(v_i, \mathbf{h}) = 2wh_i - w$ and $w = \sigma^{-1}(1-\varepsilon)$:

$$\begin{aligned}
P(v_i = 1 | h_i = 1) &= \sigma(w) = 1 - \varepsilon. && (14)\\
P(v_i = 0 | h_i = 0) &= 1 - P(v_i = 1 | h_i = 0) && (15)\\
&= 1 - \sigma(-w) && (16)\\
&= \sigma(w) = 1 - \varepsilon. && (17)
\end{aligned}$$

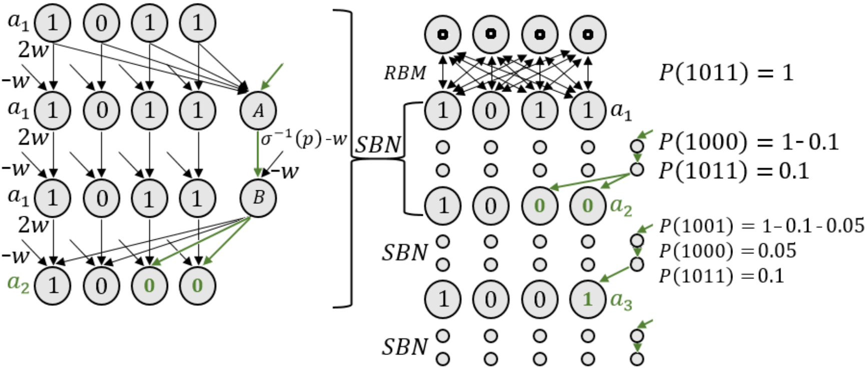
Transfer of Probability



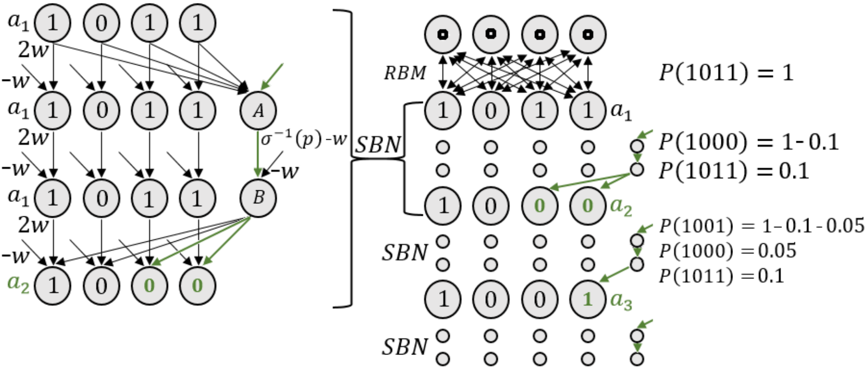
SBN : *3-Layer Sigmoid Belief Network*

- This is what Sutskever and Hinton call transfer of probability.

Transfer of Probability



Transfer of Probability



Number of parameters:

- We need $3(n+1)^2 2^n$ parameters or 3×2^n layers.

Le Roux and Bengio's Method 2010 [LeBe10]: Gray Codes

Gray codes [Gray53] are sequences $(a_i)_{1 \leq i \leq 2^n}$ such that:

- $\cup_k \{a_k\} = \{0,1\}^n$
- $\forall k$ s.t. $2 \leq k \leq 2^n$, $\|a_k - a_{k-1}\|_H = 1$ where $\|\cdot\|_H$ is the Hamming distance

Example for $n = 4$:

$$\begin{array}{rlrllll}
a_1 = & 0000 & a_5 = & 0110 & 1100 & 1010 \\
a_2 = & 0001 & a_6 = & 0111 & 1101 & 1011 \\
a_3 = & 0011 & a_7 = & 0101 & 1111 & 1001 \\
a_4 = & 0010 & \vdots & 0100 & 1110 & 1000
\end{array}$$

Theorem 1

Let $\mathbf{a}_t$ be an arbitrary binary vector in $\{0,1\}^n$ with its last bit equal to 0 and p a scalar. For every positive scalar ε $(0 < \varepsilon < 1)$, there is a weight vector $W_{n,:}$ and a real b_n such that:

- if the binary vector $\mathbf{h}$ is not equal to $\mathbf{a}_t$, the last bit remains unchanged with probability greater than or equal to $1 - \varepsilon$, that is $P(v_n = h_n | \mathbf{h} \neq \mathbf{a}_t) > (1 - \varepsilon)$.
- if the binary vector $\mathbf{h}$ is equal to $\mathbf{a}_t$, its last bit is switched from 0 to 1 with probability $\sigma(p)$.

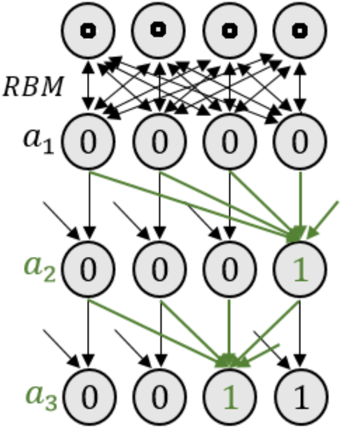
Parameters for Theorem 1

With the following the weights and biases the result in Theorem 1 is achievable:

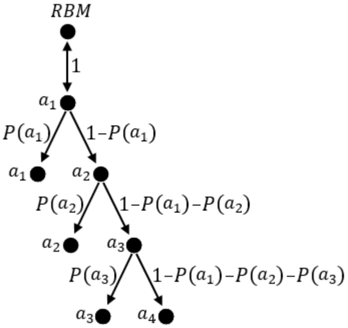
- $W_{nj} = w,\ 1 \leq j \leq k$
- $W_{nj} = -w,\ k+1 \leq j \leq n-1$
- $W_{nn} = nw$
- $b_n = -kw + p$

Number of parameters:

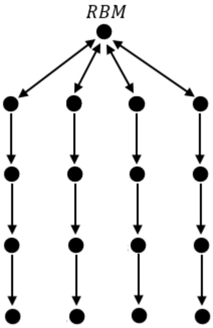
- We need $n^2 2^n$ parameters or 2^n layers.



Simultaneous Transfer of Probability



1 vector at a time.



N vectors at a time.

- We need $n2^n$ parameters or $\frac{2^n}{n}$ layers.

Arrangement of Gray Codes

Let $n = 2^t$. There exist n sequences of vectors of n bits S_i, $0 \leq i \leq n-1$ composed of vectors $S_{i,k}$, $1 \leq k \leq \frac{2^n}{n}$ satisfying the following conditions:

1. $\{S_0, \ldots, S_{n-1}\}$ is a partition of the set of all vectors of n bits.

Example for $n = 4$:

S_0	S_1	S_2	S_3
0000	0100	1000	1100
0001	0110	1001	1110
0011	0111	1011	1111
0010	0101	1010	1101

Arrangement of Gray Codes

Let $n = 2^t$. There exist n sequences of vectors of n bits S_i, $0 \leq i \leq n-1$ composed of vectors $S_{i,k}$, $1 \leq k \leq \frac{2^n}{n}$ satisfying the following conditions:

1. $\{S_0, \ldots, S_{n-1}\}$ is a partition of the set of all vectors of n bits.
2. Every sub-sequence S_i satisfies the second property of Gray codes: The Hamming distance between $S_{i,k}$ and $S_{i,k+1}$ is 1.

Example for $n = 4$:

S_0	S_1	S_2	S_3
0000	0100	1000	1100
0001	0110	1001	1110
0011	0111	1011	1111
0010	0101	1010	1101

Arrangement of Gray Codes

Let $n = 2^t$. There exist n sequences of vectors of n bits S_i, $0 \leq i \leq n-1$ composed of vectors $S_{i,k}$, $1 \leq k \leq \frac{2^n}{n}$ satisfying the following conditions:

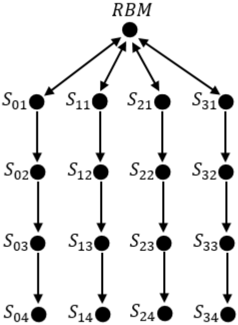
1. $\{S_0, \ldots, S_{n-1}\}$ is a partition of the set of all vectors of n bits.
2. Every sub-sequence S_i satisfies the second property of Gray codes: The Hamming distance between $S_{i,k}$ and $S_{i,k+1}$ is 1.
3. For any two sub-sequences S_i and S_j the bit switched between consecutive vectors ($S_{i,k}$ and $S_{i,k+1}$ or $S_{j,k}$ and $S_{j,k+1}$) is different unless the Hamming distance between $S_{i,k}$ and $S_{j,k}$ is 1.

Example for $n = 4$:

S_0	S_1	S_2	S_3
0000	0100	1000	1100
0001	0110	1001	1110
0011	0111	1011	1111
0010	0101	1010	1101

End Game

- Can we retain the universal approximation property of DBN by transferring probability to n vectors at a time?
- For any binary vector $\mathbf{x}$ of length n, can we still find weights and biases such that $P(\mathbf{h}^0 = \mathbf{x}) = p^*(\mathbf{x})$?



N vectors at a time.

Lemma

Let p^* be an arbitrary distribution over vectors of n bits, where n is again a power of two. A DBN with $\frac{2^n}{n} + 1$ layers such that:

1. for each i, $0 \leq i \leq n-1$, the top RBM between layers $\mathbf{h}^{\frac{2^n}{n}}$ and $\mathbf{h}^{\frac{2^n}{n}-1}$ assigns probability $\sum_k p^*(S_{i,k})$ to $S_{i,1}$.

Lemma

Let p^* be an arbitrary distribution over vectors of n bits, where n is again a power of two. A DBN with $\frac{2^n}{n} + 1$ layers such that:

1. for each i, $0 \leq i \leq n-1$, the top RBM between layers $\mathbf{h}^{\frac{2^n}{n}}$ and $\mathbf{h}^{\frac{2^n}{n}-1}$ assigns probability $\sum_k p^*(S_{i,k})$ to $S_{i,1}$.
2. for each i, $0 \leq i \leq n-1$ and each k, $1 \leq k \leq \frac{2^n}{n} - 1$, we have

$$P(\mathbf{h}^{\frac{2^n}{n}-(k+1)} = S_{i,k+1} | \mathbf{h}^{\frac{2^n}{n}-(k)} = S_{i,k}) = \frac{\sum_{t=k+1}^{\frac{2^n}{n}} p^*(S_{i,t})}{\sum_{t=k}^{\frac{2^n}{n}} p^*(S_{i,t})} \quad (18)$$

$$P(\mathbf{h}^{\frac{2^n}{n}-(k+1)} = S_{i,k} | \mathbf{h}^{\frac{2^n}{n}-(k)} = S_{i,k}) = \frac{p^*(S_{i,k})}{\sum_{t=k}^{\frac{2^n}{n}} p^*(S_{i,t})} \quad (19)$$

Lemma

Let p^* be an arbitrary distribution over vectors of n bits, where n is again a power of two. A DBN with $\frac{2^n}{n}+1$ layers such that:

1. for each i, $0 \leq i \leq n-1$, the top RBM between layers $\mathbf{h}^{\frac{2^n}{n}}$ and $\mathbf{h}^{\frac{2^n}{n}-1}$ assigns probability $\sum_k p^*(S_{i,k})$ to $S_{i,1}$.
2. for each i, $0 \leq i \leq n-1$ and each k, $1 \leq k \leq \frac{2^n}{n} - 1$, we have

$$P(\mathbf{h}^{\frac{2^n}{n}-(k+1)} = S_{i,k+1} | \mathbf{h}^{\frac{2^n}{n}-(k)} = S_{i,k}) = \frac{\sum_{t=k+1}^{\frac{2^n}{n}} p^*(S_{i,t})}{\sum_{t=k}^{\frac{2^n}{n}} p^*(S_{i,t})} \qquad (18)$$

$$P(\mathbf{h}^{\frac{2^n}{n}-(k+1)} = S_{i,k} | \mathbf{h}^{\frac{2^n}{n}-(k)} = S_{i,k}) = \frac{p^*(S_{i,k})}{\sum_{t=k}^{\frac{2^n}{n}} p^*(S_{i,t})} \qquad (19)$$

3. for each k, $1 \leq k \leq \frac{2^n}{n} - 1$, we have

$$P(\mathbf{h}^{\frac{2^n}{n}-(k+1)} = \mathbf{a} | \mathbf{h}^{\frac{2^n}{n}-(k)} = \mathbf{a}) = 1 \quad \mathit{if} \quad \mathbf{a} \notin \cup_i S_{i,k} \qquad (20)$$

has p^* as its marginal distribution over $\mathbf{h}^0$.

Proof of the Lemma

Let $\mathbf{x}$ be an arbitrary binary vector of n bits; there is a pair (i, k) such that $\mathbf{x} = S_{i,k}$. We need to show that:

$$P(\mathbf{h}^0 = S_{i,k}) = p^*(S_{i,k}). \tag{21}$$

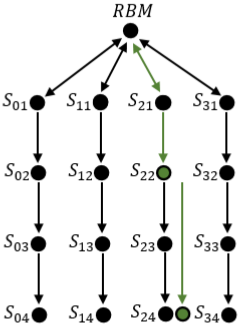
Proof of the Lemma

Let $\mathbf{x}$ be an arbitrary binary vector of n bits; there is a pair (i, k) such that $\mathbf{x} = S_{i,k}$. We need to show that:

$$P(\mathbf{h}^0 = S_{i,k}) = p^*(S_{i,k}). \qquad (21)$$

Example for $n = 4$: if $\mathbf{x} = S_{22}$

$$P(\mathbf{h}^0 = S_{22}) = p^*(S_{22}). \qquad (22)$$



Proof of the Lemma

The marginal probability of $\mathbf{h}^0 = S_{i,k}$ is therefore equal to:

$$
\begin{aligned}
P(\mathbf{h}^0 = S_{i,k}) &= P(\mathbf{h}^{\frac{2^n}{n}-1} = S_{i,1}) & (23)\\
&\cdot \prod_{t=1}^{k-1} P\left(\mathbf{h}^{\frac{2^n}{n}-(t+1)} = S_{i,t+1} | \mathbf{h}^{\frac{2^n}{n}-t} = S_{i,t}\right) & (24)\\
&\cdot P\left(\mathbf{h}^{\frac{2^n}{n}-(k+1)} = S_{i,k} | \mathbf{h}^{\frac{2^n}{n}-k} = S_{i,k}\right) & (25)\\
&\cdot \prod_{t=k+1}^{\frac{2^n}{n}-1} P\left(\mathbf{h}^{\frac{2^n}{n}-(t+1)} = S_{i,k} | \mathbf{h}^{\frac{2^n}{n}-t} = S_{i,k}\right) & (26)
\end{aligned}
$$

Proof of the Lemma

By replacing each of those probabilities by the ones given in the Lemma we get:

$$
\begin{aligned}
P(\mathbf{h}^0 = S_{i,k}) &= \sum_{u=1}^{\frac{2^n}{n}} p^*(S_{i,u}) && (27)\\
&\cdot \prod_{t=1}^{k-1} \frac{\sum_{u=t+1}^{\frac{2^n}{n}} p^*(S_{i,u})}{\sum_{u=t}^{\frac{2^n}{n}} p^*(S_{i,u})} && (28)\\
&\cdot \frac{p^*(S_{i,k})}{\sum_{u=k}^{\frac{2^n}{n}} p^*(S_{i,u})} && (29)\\
&\cdot \mathbf{1}^{\frac{2^n}{n}-1-k} && (30)\\
&= p^*(S_{i,k}) && (31)
\end{aligned}
$$

The last result comes from the cancellation of consecutive terms in the product. This concludes the proof.

Theorem 4

If $n = 2^t$, a DBN composed of $\frac{2^n}{n} + 1$ layers of size n is a universal approximator of distributions over vectors of size n.

Theorem 4

If $n = 2^t$, a DBN composed of $\frac{2^n}{n} + 1$ layers of size n is a universal approximator of distributions over vectors of size n.

Proof of Theorem 4: Using *Lemma*, we now show that it is possible to construct such a DBN.

First, Le Roux and Bengio (2008) showed that an RBM with n hidden units can model any distribution which assigns a non-zero probability to at most n vectors. *Property 1* of the *Lemma* can therefore be achieved.

Proof of Theorem 4

All the subsequent layers are as follows.

- At each layer, the first t bits of $\mathbf{h}^{k+1}$ are copied to the first t bits of $\mathbf{h}^{k}$ with probability arbitrarily close to 1. This is possible as proven earlier.

Proof of Theorem 4

All the subsequent layers are as follows.

- At each layer, the first t bits of $\mathbf{h}^{k+1}$ are copied to the first t bits of $\mathbf{h}^{k}$ with probability arbitrarily close to 1. This is possible as proven earlier.
- At each layer, $n/2$ of the remaining $n - t$ bits are potentially changed to move from one vector in a Gray code sequence to the next with the correct probability (as defined in the *Lemma*).

Proof of Theorem 4

All the subsequent layers are as follows.

- At each layer, the first t bits of $\mathbf{h}^{k+1}$ are copied to the first t bits of $\mathbf{h}^{k}$ with probability arbitrarily close to 1. This is possible as proven earlier.
- At each layer, $n/2$ of the remaining $n-t$ bits are potentially changed to move from one vector in a Gray code sequence to the next with the correct probability (as defined in the *Lemma*).
- The remaining $n/2-t$ bits are copied from $\mathbf{h}^{k+1}$ to $\mathbf{h}^{k}$ with probability arbitrarily close to 1.

Such layers are arbitrarily close to fulfilling the requirements of the second property of the *Lemma*. This concludes the proof.

Conclusion

Deep belief networks are compact universal approximators:

- Sutskever and Hinton method (2008)
  - Transfer of probability
  - We need $3(n+1)^2 2^n$ parameters or 3×2^n layers.

- LeRoux and Bengio improvements (2009)
  - Gray codes
  - Simultaneous transfer of probability
  - We need $n2^n$ parameters or $\frac{2^n}{n}$ layers (given n is a power of 2).

References:

- Le Roux, N., Bengio, Y. (2010). Deep belief networks are compact universal approximators. Neural computation, 22(8), 2192-2207.
- Le Roux, N. and Bengio, Y. (2008). Representational power of restricted boltzmann machines and deep belief networks. Neural Computation, 20(6), 16311649.
- Sutskever, I. and Hinton, G. E. (2008). Deep, narrow sigmoid belief networks are universal approximators. Neural Computation, 20(11), 26292636.
- Hinton, G. E., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18, 15271554.
- Gray, F. (1953). Pulse code communication. U.S. Patent 2,632,058.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 2(4), 303-314.