Pattern Recognition & Bioinformatics Laboratory
Delft University of Technology
http://prb.tudelft.nl/

P.O. Box 5013
2600 GA Delft
The Netherlands

February 2, 2011

EWI-PRB TR 2011–001

Discriminative Restricted Boltzmann Machines are Universal Approximators for Discrete Data

Laurens van der Maaten
Pattern Recognition & Bioinformatics Laboratory
Delft University of Technology

Abstract

This report proofs that discriminative Restricted Boltzmann Machines (RBMs) are universal approximators for discrete data by adapting existing universal approximation proofs for generative RBMs.

Discriminative Restricted Boltzmann Machines are Universal Approximators for Discrete Data

Laurens van der Maaten
Pattern Recognition & Bioinformatics Laboratory
Delft University of Technology

1 Introduction

A discriminative Restricted Boltzmann Machine (RBM) models is a conditional variant of the RBM [1, 2, 4] that models the conditional distribution $p(\mathbf{y}|\mathbf{x})$ as

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \sum_{\mathbf{z}} \exp(\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{z} + \mathbf{z}^{\mathrm{T}}\mathbf{V}\mathbf{y} + \mathbf{b}^{\mathrm{T}}\mathbf{z} + \mathbf{c}^{\mathrm{T}}\mathbf{y}), \tag{1}$$

where $Z(\mathbf{x})$ represents the partition function

$$Z(\mathbf{x}) = \sum_{\mathbf{y}'} \sum_{\mathbf{z}'} \exp(\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{z}' + \mathbf{z}'^{\mathrm{T}}\mathbf{V}\mathbf{y}' + \mathbf{b}^{\mathrm{T}}\mathbf{z}' + \mathbf{c}^{\mathrm{T}}\mathbf{y}'). \tag{2}$$

In this note, we proof the following theorem for discriminative RBMs:

Universal Approximation Theorem. *For data $\mathbf{x} \in \mathcal{X} = \{0, 1\}^D$, a discriminative RBM can represent any conditional distribution $p(\mathbf{y}|\mathbf{x})$ arbitrarily well in terms of Kullback-Leibler divergence.*

Since discrete data can be expressed exactly in terms of a binary representation (e.g., using a 1-of-D representation), the theorem applies to discrete data, too. The proof of the above theorem is an adaptation of an earlier proof on the representational power of generative RBMs [3].

2 Proof

Denote potentials by $F(\mathbf{x}, \mathbf{y}, \mathbf{z})$, the conditional distribution modeled by a discriminative RBM can be written as

$$p(\mathbf{y}|\mathbf{x}) = \frac{\sum_{\mathbf{z}} F(\mathbf{x}, \mathbf{y}, \mathbf{z})}{\sum_{\mathbf{y}', \mathbf{z}'} F(\mathbf{x}, \mathbf{y}', \mathbf{z})'} = \frac{\sum_{\mathbf{z}} \exp(\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{z} + \mathbf{z}^{\mathrm{T}}\mathbf{V}\mathbf{y} + \mathbf{b}^{\mathrm{T}}\mathbf{z} + \mathbf{c}^{\mathrm{T}}\mathbf{y})}{\sum_{\mathbf{y}'} \sum_{\mathbf{z}'} \exp(\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{z}' + \mathbf{z}'^{\mathrm{T}}\mathbf{V}\mathbf{y}' + \mathbf{b}^{\mathrm{T}}\mathbf{z}' + \mathbf{c}^{\mathrm{T}}\mathbf{y}')}, \tag{3}$$

where we assume that $\mathbf{x}$ is a binary vector, and $\mathbf{y}$ is a 1-of-K vector. We denote the discriminative RBM that has one additional hidden unit with parameters $\mathbf{w}$, $\mathbf{v}$, and b by $p_{wvb}(\mathbf{y}|\mathbf{x})$:

$$p_{wvb}(\mathbf{y}|\mathbf{x}) = \frac{(1 + \exp(\mathbf{w}^{\mathrm{T}}\mathbf{x} + \mathbf{v}^{\mathrm{T}}\mathbf{y} + b)) \sum_{\mathbf{z}} F(\mathbf{x}, \mathbf{y}, \mathbf{z})}{\sum_{\mathbf{y}'} (1 + \exp(\mathbf{w}^{\mathrm{T}}\mathbf{x} + \mathbf{v}^{\mathrm{T}}\mathbf{y}' + b)) \sum_{\mathbf{z}'} F(\mathbf{x}, \mathbf{y}', \mathbf{z}')}. \tag{4}$$

Let $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ be an arbitrary $(\mathbf{x}, \mathbf{y})$-pair for which we wish to change increase the conditional probability $p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})$. Also, we define the parameters $\hat{\mathbf{w}} = a(\tilde{\mathbf{x}} - \frac{1}{2})$, $\hat{\mathbf{v}} = a(\tilde{\mathbf{y}} - \frac{1}{2})$, and $\hat{b} =$

$-\hat{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}} - \hat{\mathbf{v}}^{\mathrm{T}}\tilde{\mathbf{y}} + \lambda$ with $a, \lambda \in \mathbb{R}$. For this setting of the parameters, we investigate the limit of the term induced by the new hidden unit as the variable a goes to infinity. In particular, we find the limits

$$\lim_{a\to\infty}(1+\exp(\hat{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}} + \hat{\mathbf{v}}^{\mathrm{T}}\tilde{\mathbf{y}} + \hat{b})) = \lim_{a\to\infty}(1+\exp(\lambda)) = 1+\exp(\lambda), \tag{5}$$

$$\forall \mathbf{x} \neq \tilde{\mathbf{x}} : \lim_{a\to\infty}(1+\exp(\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{x} + \hat{\mathbf{v}}^{\mathrm{T}}\tilde{\mathbf{y}} + \hat{b})) = \lim_{a\to\infty}\left(1+\exp\left(a\left(\tilde{\mathbf{x}} - \frac{1}{2}\right)^{\mathrm{T}}(\mathbf{x}-\tilde{\mathbf{x}}) + \lambda\right)\right) = 1, \tag{6}$$

$$\forall \mathbf{y} \neq \tilde{\mathbf{y}} : \lim_{a\to\infty}(1+\exp(\hat{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}} + \hat{\mathbf{v}}^{\mathrm{T}}\mathbf{y} + \hat{b})) = \lim_{a\to\infty}\left(1+\exp\left(a\left(\tilde{\mathbf{y}} - \frac{1}{2}\right)^{\mathrm{T}}(\mathbf{y}-\tilde{\mathbf{y}}) + \lambda\right)\right) = 1, \tag{7}$$

$$\forall \mathbf{x} \neq \tilde{\mathbf{x}}, \mathbf{y} \neq \tilde{\mathbf{y}} : \lim_{a\to\infty}(1+\exp(\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{x} + \hat{\mathbf{v}}^{\mathrm{T}}\mathbf{y} + \hat{b})) = \lim_{a\to\infty}\left(1+\exp\left(a\left(\tilde{\mathbf{x}} - \frac{1}{2}\right)^{\mathrm{T}}(\mathbf{x}-\tilde{\mathbf{x}}) + a\left(\tilde{\mathbf{y}} - \frac{1}{2}\right)^{\mathrm{T}}(\mathbf{y}-\tilde{\mathbf{y}}) + \lambda\right)\right) = 1. \tag{8}$$

In the derivation of these limits, we use the fact that $\mathbf{x}$, $\tilde{\mathbf{x}}$, $\mathbf{y}$, and $\tilde{\mathbf{y}}$ are binary vectors. It is not obvious how to obtain the same limits for the continuous case, i.e., when $\mathbf{x} \in \mathbb{R}^D$. The key observation is that in the limit when a goes to infinity, for this specific choice of parameters for the new hidden unit, the contribution of the new hidden unit to the product over all hidden units is always 1, except for the particular pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ we picked. In other words, the hidden unit has no effect on the unnormalized conditional probabilities, except for the $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ pair[1].

To formalize this, we can consider the behavior of the conditional distribution $p_{\hat{w}\hat{v}\hat{b}}(\mathbf{y}|\tilde{\mathbf{x}})$ in the limit when a goes to infinity. We first work out the limit for the $\forall \mathbf{y} \neq \tilde{\mathbf{y}}$ case, to find

$$\forall \mathbf{y} \neq \tilde{\mathbf{y}} : \lim_{a\to\infty} p_{\hat{w}\hat{v}\hat{b}}(\mathbf{y}|\tilde{\mathbf{x}}) = \lim_{a\to\infty} \frac{(1+\exp(\hat{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}} + \hat{\mathbf{v}}^{\mathrm{T}}\mathbf{y} + \hat{b}))\sum_{\mathbf{z}} F(\tilde{\mathbf{x}}, \mathbf{y}, \mathbf{z})}{\sum_{\mathbf{y}'}(1+\exp(\hat{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}} + \hat{\mathbf{v}}^{\mathrm{T}}\mathbf{y}' + \hat{b}))\sum_{\mathbf{z}'} F(\tilde{\mathbf{x}}, \mathbf{y}', \mathbf{z}')} \tag{9}$$

$$= \frac{\sum_{\mathbf{z}} F(\tilde{\mathbf{x}}, \mathbf{y}, \mathbf{z})}{(1+\exp(\lambda))\sum_{\mathbf{z}'} F(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \mathbf{z}') + \sum_{\mathbf{y}'\neq\tilde{\mathbf{y}}}\sum_{\mathbf{z}'} F(\tilde{\mathbf{x}}, \mathbf{y}', \mathbf{z}')} \tag{10}$$

$$= \frac{1}{1+\exp(\lambda)\frac{\sum_{\mathbf{z}} F(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \mathbf{z})}{\sum_{\mathbf{y}'}\sum_{\mathbf{z}'} F(\tilde{\mathbf{x}}, \mathbf{y}', \mathbf{z}')}} \frac{\sum_{\mathbf{z}} F(\tilde{\mathbf{x}}, \mathbf{y}, \mathbf{z})}{\sum_{\mathbf{y}'}\sum_{\mathbf{z}'} F(\tilde{\mathbf{x}}, \mathbf{y}', \mathbf{z}')}, \tag{11}$$

where we use the fact that $\tilde{\mathbf{y}}$ is an element in the sum over $\mathbf{y}'$. Noting that $p(\mathbf{y}|\mathbf{x}) = \frac{\sum_{\mathbf{z}} F(\mathbf{x}, \mathbf{y}, \mathbf{z})}{\sum_{\mathbf{y}'}\sum_{\mathbf{z}'} F(\mathbf{x}, \mathbf{y}', \mathbf{z}')}$, we obtain the following limit

$$\forall \mathbf{y} \neq \tilde{\mathbf{y}} : \lim_{a\to\infty} p_{\hat{w}\hat{v}\hat{b}}(\mathbf{y}|\tilde{\mathbf{x}}) = \frac{p(\mathbf{y}|\tilde{\mathbf{x}})}{1+\exp(\lambda)p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})}. \tag{12}$$

Similarly, we can derive

$$\lim_{a\to\infty} p_{\hat{w}\hat{v}\hat{b}}(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}) = \frac{(1+\exp(\lambda))p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})}{1+\exp(\lambda)p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})}, \tag{13}$$

$$\forall \mathbf{x} \neq \tilde{\mathbf{x}} : \lim_{a\to\infty} p_{\hat{w}\hat{v}\hat{b}}(\tilde{\mathbf{y}}|\mathbf{x}) = p(\tilde{\mathbf{y}}|\mathbf{x}), \tag{14}$$

$$\forall \mathbf{x} \neq \tilde{\mathbf{x}}, \mathbf{y} \neq \tilde{\mathbf{y}} : \lim_{a\to\infty} p_{\hat{w}\hat{v}\hat{b}}(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x}). \tag{15}$$

Hence, we can arbitrarily increase the conditional probability $p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})$ for a pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ by choosing the appropriate λ, whilst uniformly decreasing the other conditional probabilities $p(\mathbf{y}|\tilde{\mathbf{x}})$ through the normalization factor. Perhaps surprisingly, this does not affect the conditional

[1]We note here that there is an effect on other conditional probabilities through the normalization term. We explore this effect below.

probabilities $p(\mathbf{y}|\mathbf{x})$ for any other data point $\mathbf{x} \neq \tilde{\mathbf{x}}$. We can use this property to construct a discriminative RBM that exactly matches each possible distribution $p(\mathbf{y}|\mathbf{x})$ by repeating the procedure below for each of the exponential number of $\mathbf{x}$'s.

Denote the discriminative RBM with i+1 hidden units as p^i, and define p^0 as a discriminative RBM in which all weights and biases (i.e., $\mathbf{w}_0$, $\mathbf{v}_0$, and b_0) are set to 0. Hence, p^0 defines a uniform distribution over $\mathbf{y}$ for $\forall \mathbf{x}$; for a particular $\tilde{\mathbf{x}}$, we obtain that $p^0(\mathbf{y}|\tilde{\mathbf{x}}) = \frac{1}{K}$. Now, we index all k possible label vectors $\mathbf{y}$ that have non-zero probability (given $\tilde{\mathbf{x}}$) by integers from 1 to $k \leq K$ (which gives us $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k$), and sort them according to the true conditional distribution we wish to obtain such that

$$0 < p(\mathbf{y}_1|\tilde{\mathbf{x}}) \leq p(\mathbf{y}_2|\tilde{\mathbf{x}}) \leq \cdots \leq p(\mathbf{y}_k|\tilde{\mathbf{x}}). \tag{16}$$

Next, we define the second hidden unit with parameters $\mathbf{w}_1 = a_1(\tilde{\mathbf{x}} - \frac{1}{2})$, $\mathbf{v}_1 = a_1(\mathbf{y}_1 - \frac{1}{2})$, and $b_1 = -\mathbf{w}_1^{\mathrm{T}}\tilde{\mathbf{x}} - \mathbf{w}_1^{\mathrm{T}}\mathbf{y}_1 + \lambda_1$. For this specification of the parameters for the second hidden unit, we have shown above that

$$\lim_{a_1 \to \infty} p^1(\mathbf{y}_1|\tilde{\mathbf{x}}) = \frac{(1 + \exp(\lambda_1))\frac{1}{K}}{1 + \exp(\lambda_1)\frac{1}{K}}, \text{ and} \tag{17}$$

$$\forall i > 1 : \lim_{a_1 \to \infty} p^1(\mathbf{y}_i|\tilde{\mathbf{x}}) = \frac{\frac{1}{K}}{1 + \exp(\lambda_1)\frac{1}{K}}. \tag{18}$$

Indeed, we can make $p^1(\mathbf{y}_1|\tilde{\mathbf{x}})$ arbitrarily close to 1 by increasing λ_1, whilst maintaining a uniform distribution over $\mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_k$. In the next step, we can similarly introducte a third hidden unit with parameters $\mathbf{w}_2$, $\mathbf{v}_2$, and b_2, and set the corresponding λ_2 in such a way that $\frac{p^2(\mathbf{y}_2|\tilde{\mathbf{x}})}{p^2(\mathbf{y}_1|\tilde{\mathbf{x}})} = \frac{p(\mathbf{y}_2|\tilde{\mathbf{x}})}{p(\mathbf{y}_1|\tilde{\mathbf{x}})}$. To see that this is possible, note (1) that $\frac{p(\mathbf{y}_2|\tilde{\mathbf{x}})}{p(\mathbf{y}_1|\tilde{\mathbf{x}})} \geq 1$ because of the ordering and $\frac{p^1(\mathbf{y}_2|\tilde{\mathbf{x}})}{p^1(\mathbf{y}_1|\tilde{\mathbf{x}})} \leq 1$ because it was impossible to decrease $p^1(\mathbf{y}_1|\tilde{\mathbf{x}})$, and thus $\frac{p(\mathbf{y}_2|\tilde{\mathbf{x}})}{p(\mathbf{y}_1|\tilde{\mathbf{x}})} \geq \frac{p^1(\mathbf{y}_2|\tilde{\mathbf{x}})}{p^1(\mathbf{y}_1|\tilde{\mathbf{x}})}$, and note (2) that we can arbitrarily increase $p^2(\mathbf{y}_2|\tilde{\mathbf{x}})$ whilst uniformly multiplying the other conditional probabilities by a constant factor. We can repeat this procedure for the consecutive values $\lambda_3, \lambda_4, \ldots, \lambda_k$. Note that once we set a ratio correctly, the ratio is not altered by later hidden unit additions. Hence, the result of this procedure is a discriminative RBM $p^k(\mathbf{y}|\mathbf{x})$ that – for our particular $\tilde{\mathbf{x}}$ – correctly preserves log-ratios for all label vectors with non-zero probability

$$\frac{p^k(\mathbf{y}_2|\tilde{\mathbf{x}})}{p^k(\mathbf{y}_1|\tilde{\mathbf{x}})} = \frac{p(\mathbf{y}_2|\tilde{\mathbf{x}})}{p(\mathbf{y}_1|\tilde{\mathbf{x}})}, \frac{p^k(\mathbf{y}_3|\tilde{\mathbf{x}})}{p^k(\mathbf{y}_2|\tilde{\mathbf{x}})} = \frac{p(\mathbf{y}_3|\tilde{\mathbf{x}})}{p(\mathbf{y}_2|\tilde{\mathbf{x}})}, \ldots, \frac{p^k(\mathbf{y}_k|\tilde{\mathbf{x}})}{p^k(\mathbf{y}_{k-1}|\tilde{\mathbf{x}})} = \frac{p(\mathbf{y}_k|\tilde{\mathbf{x}})}{p(\mathbf{y}_{k-1}|\tilde{\mathbf{x}})}. \tag{19}$$

At the same time, for the label vectors with zero probability $\mathbf{y}_{k+1}, \ldots \mathbf{y}_K$, the discriminative RBM $p^k(\mathbf{y}|\tilde{\mathbf{x}})$ still specifies a uniform distribution, i.e., $p^k(\mathbf{y}_{k+1}|\tilde{\mathbf{x}}) = \cdots = p^k(\mathbf{y}_K|\tilde{\mathbf{x}})$. Since all ratios are correct, all conditional probabilities $p^k(\mathbf{y}_1|\tilde{\mathbf{x}}), \ldots, p^k(\mathbf{y}_k|\tilde{\mathbf{x}})$ only still need to be multiplied by a constant, i.e., the remaining probability mass in the uniform distribution needs to be distributed among all label vectors with non-zero probabilities. Specifically, we can see that $p^k(\mathbf{y}_1|\tilde{\mathbf{x}}) = \nu p(\mathbf{y}_1|\tilde{\mathbf{x}}), p^k(\mathbf{y}_2|\tilde{\mathbf{x}}) = \nu p(\mathbf{y}_2|\tilde{\mathbf{x}}), \ldots, p^k(\mathbf{y}_k|\tilde{\mathbf{x}}) = \nu p(\mathbf{y}_k|\tilde{\mathbf{x}})$, where the value of ν is given by $\nu = 1 - (K - k)p^k(\mathbf{y}_K|\tilde{\mathbf{x}})$. In addition, because we never changed any of the ratios that involve the label vector $\mathbf{y}_1$ after we fixed λ_1, we also have

$$\text{for } k > i \geq K : \qquad \frac{p^k(\mathbf{y}_1|\tilde{\mathbf{x}})}{p^k(\mathbf{y}_i|\tilde{\mathbf{x}})} = \frac{p^1(\mathbf{y}_1|\tilde{\mathbf{x}})}{p^1(\mathbf{y}_i|\tilde{\mathbf{x}})} = 1 + \exp(\lambda_1). \tag{20}$$

As a result, we obtain the equations

$$\text{for } k > i \geq K : \qquad p^k(\mathbf{y}_1|\tilde{\mathbf{x}}) = p(\mathbf{y}_1|\tilde{\mathbf{x}})[1 - (K - k)p^k(\mathbf{y}_i|\tilde{\mathbf{x}})] = (1 + \exp(\lambda_1))p^k(\mathbf{y}_i|\tilde{\mathbf{x}}), \tag{21}$$

$$\text{for } 0 < i \leq k : \qquad p^k(\mathbf{y}_i|\tilde{\mathbf{x}}) = p^k(\mathbf{y}_1|\tilde{\mathbf{x}})\frac{1 + \exp(\lambda_i)}{1 + \exp(\lambda_1)} = (1 + \exp(\lambda_1))p^k(\mathbf{y}_K|\tilde{\mathbf{x}}). \tag{22}$$

Solving the above equations, we obtain

$$\text{for } k > i \geq K: \qquad (1+\exp(\lambda_1))p^k(\mathbf{y}_i|\tilde{\mathbf{x}}) = p(\mathbf{y}_1|\tilde{\mathbf{x}})\left[1-(K-k)p^k(\mathbf{y}_i|\tilde{\mathbf{x}})\right] \tag{23}$$

$$p(\mathbf{y}_1|\tilde{\mathbf{x}}) = \left[(1+\exp(\lambda_1)) + (K-k)p(\mathbf{y}_1|\tilde{\mathbf{x}})\right] p^k(\mathbf{y}_i|\tilde{\mathbf{x}}) \tag{24}$$

$$p^k(\mathbf{y}_i|\tilde{\mathbf{x}}) = \frac{p(\mathbf{y}_1|\tilde{\mathbf{x}})}{1+\exp(\lambda_1)+(K-k)p(\mathbf{y}_1|\tilde{\mathbf{x}})}, \tag{25}$$

and filling in this result into the RHS of Equation 22, we obtain

$$\text{for } 0 < i \leq k: \qquad p^k(\mathbf{y}_i|\tilde{\mathbf{x}}) = p(\mathbf{y}_1|\tilde{\mathbf{x}})\frac{1+\exp(\lambda_i)}{1+\exp(\lambda_1)+(K-k)p(\mathbf{y}_1|\tilde{\mathbf{x}})} \tag{26}$$

$$= p(\mathbf{y}_i|\tilde{\mathbf{x}})\frac{1+\exp(\lambda_1)}{1+\exp(\lambda_1)+(K-k)p(\mathbf{y}_1|\tilde{\mathbf{x}})}, \tag{27}$$

where we made use of the fact that $0 < i \leq k, 0 < j \leq k : \frac{p^k(\mathbf{y}_i|\tilde{\mathbf{x}})}{p^k(\mathbf{y}_j|\tilde{\mathbf{x}})} = \frac{p(\mathbf{y}_i|\tilde{\mathbf{x}})}{p(\mathbf{y}_j|\tilde{\mathbf{x}})}$. From the two solutions above, it is straightforward to see that $p^k(\mathbf{y}_i|\tilde{\mathbf{x}})$ approaches $p(\mathbf{y}_i|\tilde{\mathbf{x}})$ in the limit where λ_1 goes to infinity. Hence, we can take the limit $\lambda_1 \to \infty$ to approach the desired conditional distribution $p(\mathbf{y}|\tilde{\mathbf{x}})$.

We can express this by investigating the Kullback-Leibler divergence $KL(p(\mathbf{y}|\tilde{\mathbf{x}})||p^k(\mathbf{y}|\tilde{\mathbf{x}}))$ between the distribution we aim to obtain and the discriminative RBM distribution

$$KL(p(\mathbf{y}|\tilde{\mathbf{x}})||p^k(\mathbf{y}|\tilde{\mathbf{x}})) = \sum_{i=1}^{K} p(\mathbf{y}_i|\tilde{\mathbf{x}}) \log p(\mathbf{y}_i|\tilde{\mathbf{x}}) - \sum_{i=1}^{K} p(\mathbf{y}_i|\tilde{\mathbf{x}}) \log p^k(\mathbf{y}_i|\tilde{\mathbf{x}}) \tag{28}$$

$$= \sum_{i=1}^{k} p(\mathbf{y}_i|\tilde{\mathbf{x}}) \log p(\mathbf{y}_i|\tilde{\mathbf{x}}) - \sum_{i=1}^{k} p(\mathbf{y}_i|\tilde{\mathbf{x}}) \log p^k(\mathbf{y}_i|\tilde{\mathbf{x}}) \tag{29}$$

$$= -\sum_{i=1}^{k} p(\mathbf{y}_i|\tilde{\mathbf{x}}) \log(1+\exp(\lambda_1)) + \sum_{i=1}^{k} p(\mathbf{y}_i|\tilde{\mathbf{x}}) \log\left(1+\exp(\lambda_1)+(K-k)p(\mathbf{y}_1|\tilde{\mathbf{x}})\right) \tag{30}$$

$$= \sum_{i=1}^{k} p(\mathbf{y}_i|\tilde{\mathbf{x}}) \log\left(1+\frac{(K-k)p(\mathbf{y}_1|\tilde{\mathbf{x}})}{1+\exp(\lambda_1)}\right). \tag{31}$$

We can investigate the Kullback-Leibler divergence in the limit where λ_1 goes to infinity, to find

$$\lim_{\lambda_1\to\infty} KL(p(\mathbf{y}|\tilde{\mathbf{x}})||p^k(\mathbf{y}|\tilde{\mathbf{x}})) = \lim_{\lambda_1\to\infty} \sum_{i=1}^{k} p(\mathbf{y}_i|\tilde{\mathbf{x}}) \log\left(1+\frac{(K-k)p(\mathbf{y}_1|\tilde{\mathbf{x}})}{1+\exp(\lambda_1)}\right) = 0. \tag{32}$$

Hence, for a particular data point $\tilde{\mathbf{x}}$, the discriminative RBM $p^k(\mathbf{y}|\tilde{\mathbf{x}})$ can approach the conditional distribution $p(\mathbf{y}|\tilde{\mathbf{x}})$ arbitrarily well in terms of Kullback-Leibler divergence (by letting λ_1 go to infinity).

Note that we have already shown that the above procedure only alters $p(\mathbf{y}|\tilde{\mathbf{x}})$ for the specific data point $\tilde{\mathbf{x}}$, but that it does not alter $p(\mathbf{y}|\mathbf{x})$ for $\forall \mathbf{x} \neq \tilde{\mathbf{x}}$ (in Equation 14 and 15). After adapting the discriminative RBM for a particular data point $\tilde{\mathbf{x}}$, the distribution over $\mathbf{y}$ for all other data vectors $\mathbf{x} \neq \tilde{\mathbf{x}}$ is thus still uniform. Hence, we can repeat the above procedure for every $\mathbf{x} \in \mathcal{X} = \{0,1\}^D$. □

Acknowledgments

The author thanks Lawrence Saul for helpful discussions. This work was performed while the author was at University of California, San Diego. The work was supported by the Netherlands Organisation for Scientific Research (NWO; grant no. 680.50.0908), the EU-FP7 NoE on Social Signal Processing, and by award number 0812576 from the National Science Foundation.

References

[1] Y. Freund and D. Haussler. Unsupervised learning of distributions on binary vectors using two layer networks. Technical Report UCSC-CRL-94-25, University of California, Santa Cruz, 1994.

[2] H. Larochelle and Y. Bengio. Classification using discriminative Restricted Boltzmann Machines. In *Proceedings of the* 25^{th} *International Conference on Machine Learning*, pages 536–543, 2008.

[3] N. Le Roux and Y. Bengio. Representational power of Restricted Boltzmann Machines and Deep Belief Networks. *Neural Computation*, 20(6):1631–1649, 2008.

[4] T. Schmah, G.E. Hinton, R. Zemel, S.L. Small, and S. Strother. Generative versus discriminative RBM models for classification of fMRI images. In *Advances in Neural Information Processing Systems*, volume 21, pages 1409–1416, 2009.