

# Learning Feature Hierarchies with Centered Deep Boltzmann Machines

Grégoire Montavon  
Klaus-Robert Müller\*

Machine Learning Group  
Technische Universität Berlin  
Franklinstr. 28/29  
10587 Berlin, Germany

GMONTAVON@CS.TU-BERLIN.DE  
KLAUS-ROBERT.MUELLER@TU-BERLIN.DE

## Abstract

Deep Boltzmann machines are in principle powerful models for extracting the hierarchical structure of data. Unfortunately, attempts to train layers jointly (without greedy layer-wise pretraining) have been largely unsuccessful. We propose a modification of the learning algorithm that initially recenters the output of the activation functions to zero. This modification leads to a better conditioned Hessian and thus makes learning easier. We test the algorithm on real data and demonstrate that our suggestion, the *centered deep Boltzmann machine*, learns a hierarchy of increasingly abstract representations and a better generative model of data.

## 1. Introduction

Deep Boltzmann machines (DBM, Salakhutdinov and Hinton, 2009) are in principle powerful models for extracting the hierarchical structure of data (Montavon et al., 2012). Unfortunately, attempts to train layers jointly (without greedy layer-wise pretraining) have been mostly unsuccessful. As we will argue later in greater detail, a possible reason for this could be that the mapping of net activities onto the sigmoid nonlinearities is not centered to zero by default.

In this paper, we propose to recenter the output of each unit to zero by rewriting the energy as a function of centered states  $\xi = x - \beta$  where  $\beta$  is an offset parameter. The reparameterization of the energy function leads to a better conditioned Hessian of the estimated model log-likelihood. The centered Boltzmann machine is easy to implement as the reparameterization leaves the associated Gibbs distribution invariant.

We train a centered deep Boltzmann machine on the MNIST data set. Empirical results show that the centered DBM is able to learn a top-layer representation that contains useful discriminative features and to produce a good generative model of data. In addition, the centered DBM learns faster and is more stable than its non-centered counterpart.

**Related work** The case for using centered nonlinearities has already been made by LeCun et al. (1998) and Glorot and Bengio (2010) in the context of backpropagation networks, showing that the logistic function generally performs poorly compared to its centered counterpart, the hyperbolic tangent. The idea of centering was also proposed by

---

\*Also at the Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Korea

Tang and Sutskever (2011) in the context of restricted Boltzmann machines but was restricted to data centering.

## 2. Centered Boltzmann Machines

In this section, we introduce the centered Boltzmann machine. In the following, the sigmoid function is defined as  $\text{sigm}(x) = \frac{e^x}{e^x + 1}$ ,  $x \sim \mathcal{B}(p)$  denotes that the variable  $x$  is drawn randomly from a Bernoulli distribution of parameter  $p$  and  $\langle \cdot \rangle_P$  denotes the expectation operator with respect to a probability distribution  $P$ . All these operations apply element-wise to the input vector.

A Boltzmann machine is a network of  $M_x$  interconnected binary units that associates to each state  $x \in \{0, 1\}^{M_x}$  the energy

$$E(x; \theta) = -x^\top W x - x^\top b$$

where  $\theta = \{W, b\}$  groups the model parameters. The matrix  $W$  of size  $M_x \times M_x$  is symmetric and contains the connection strength between units. The vector  $b$  of size  $M_x$  contains the biases associated to each unit. A probability is associated to each state according to the Gibbs distribution

$$p(x; \theta) = \frac{e^{-E(x; \theta)}}{\sum_x e^{-E(x; \theta)}}$$

where the term in the denominator is the partition function that makes probabilities sum to one. For the centered Boltzmann machine, we rewrite the energy as a function of centered states

$$E(x; \theta) = -(x - \beta)^\top W (x - \beta) - (x - \beta)^\top b$$

where  $\theta = \{W, b, \beta\}$  and where the vector  $\beta$  contains the offsets associated to each unit of the network. Setting  $\beta = \text{sigm}(b_0)$  where  $b_0$  is the initial bias enforces the initial centering of the Boltzmann machine. From these equations, we can derive the conditional probability

$$p(x_i = 1 | x_{-i}; \theta) = \text{sigm}(b_i + \sum_{j \neq i} W_{ij}(x - \beta)_j)$$

of each unit and the gradient of the model log-likelihood with respect to  $W$  and  $b$ :

$$\begin{aligned} \frac{\partial}{\partial W} \langle \log p(x; \theta) \rangle_{\text{data}} &= \langle (x - \beta)(x - \beta)^\top \rangle_{\text{data}} - \langle (x - \beta)(x - \beta)^\top \rangle_{\text{model}} \\ \frac{\partial}{\partial b} \langle \log p(x; \theta) \rangle_{\text{data}} &= \langle x - \beta \rangle_{\text{data}} - \langle x - \beta \rangle_{\text{model}} \end{aligned}$$

### 2.1 Stability of the Centered Boltzmann Machine

In this section, we look at the stability of the underlying optimization problem. We argue that when the sigmoid is centered, the Hessian is better conditioned (see Figure 2), and

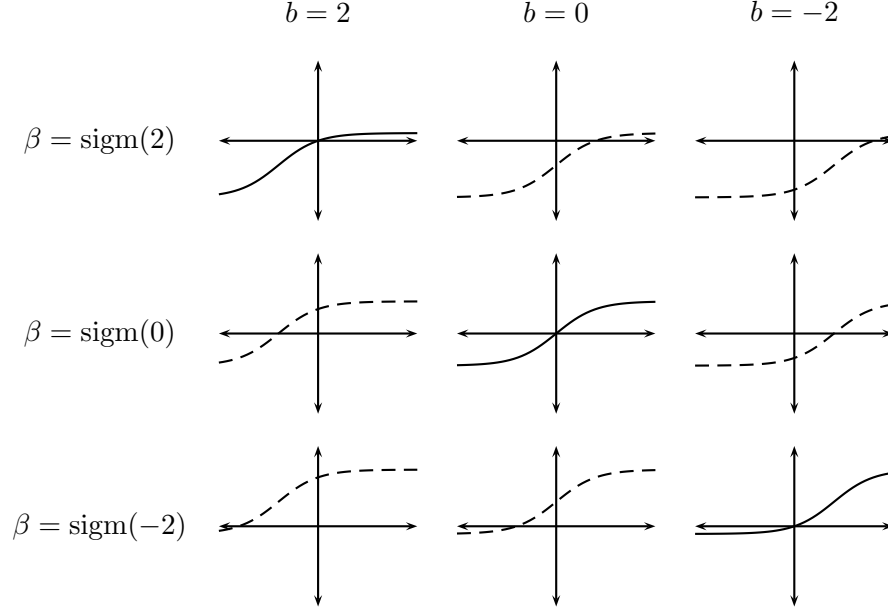


Figure 1: Example of sigmoids with different biases and offsets. The three non-dashed sigmoids are said to be *centered* because they cross the origin. We show that centering sigmoids leads to a better conditioned Hessian.

therefore, the learning algorithm is more stable. We define  $\xi$  as the centered state  $\xi = x - \beta$ . The derivative of the model log-likelihood with respect to the weight vector takes the form

$$\frac{\partial}{\partial W} \langle \log p(x; \theta) \rangle_{\text{data}} = \langle \xi \xi^\top \rangle_{\text{data}} - \langle \xi \xi^\top \rangle_W$$

where  $\langle \cdot \rangle_W$  designates the expectation with respect to the probability distribution associated to a model of weight parameter  $W$ . Using the definition of the directional derivative, the second derivative with respect to a random direction  $V$  (which is equal to the projected Hessian  $\mathbf{H}V$ ) can be expressed as:

$$\begin{aligned} \mathbf{H}V &= \frac{\partial}{\partial V} \left( \frac{\partial}{\partial W} \langle \log p(x; W) \rangle_{\text{data}} \right) \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \left( \frac{\partial}{\partial W} \langle \log p(x; W + hV) \rangle_{\text{data}} - \frac{\partial}{\partial W} \langle \log p(x; W) \rangle_{\text{data}} \right) \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \left( (\langle \xi \xi^\top \rangle_{W+hV, \text{data}} - \langle \xi \xi^\top \rangle_{W+hV}) - (\langle \xi \xi^\top \rangle_{W, \text{data}} - \langle \xi \xi^\top \rangle_W) \right) \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \left( \langle \xi \xi^\top \rangle_{W+hV, \text{data}} - \langle \xi \xi^\top \rangle_{W, \text{data}} \right) - \lim_{h \rightarrow 0} \frac{1}{h} \left( \langle \xi \xi^\top \rangle_{W+hV} - \langle \xi \xi^\top \rangle_W \right) \end{aligned}$$

From the last line, we can see that the Hessian can be decomposed into a data-dependent term and a data-independent term. A remarkable fact is that in absence of hidden units, the data-dependent part of the Hessian is zero, because the model—and therefore, the perturbation of the model—have no influence on the states. The conditioning of the optimization

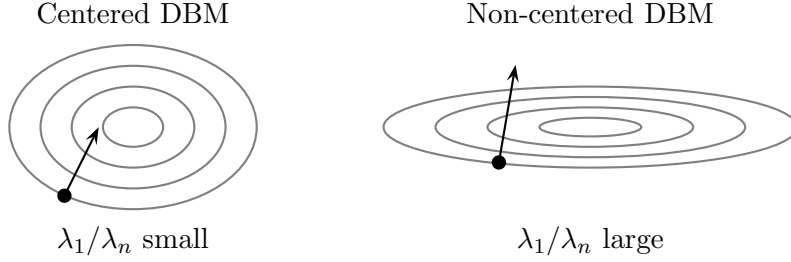


Figure 2: Relation between the conditioning number  $\lambda_1/\lambda_n$  and the shape of the optimization problem. Gradient descent is easier to achieve when the conditioning number is small.

problem can therefore be analyzed exclusively from the perspective of the model without even looking at the data. The data-dependent term is likely to be small even in the presence of hidden variables due to the sharp reduction of entropy caused by the clamping of visible units to data.

We can think of a well-conditioned model as a model for which a perturbation of the model parameter  $W$  in any direction  $V$  causes a well-behaved perturbation of state expectations  $\langle \xi \xi^\top \rangle_W$ . Pearlmutter (1994) showed that in a Boltzmann machine with no hidden units, the projected Hessian can be further reduced to

$$\mathbf{H}V = \langle \xi \xi^\top \rangle_W \cdot \langle D \rangle_W - \langle \xi \xi^\top D \rangle_W \quad \text{where} \quad D = \frac{1}{2} \xi^\top V \xi \quad (1)$$

thus, getting rid of the limit and leading to numerically more accurate estimates. LeCun et al. (1998) showed that the stability of the optimization problem can be quantified by the *conditioning number* defined as the ratio between the largest eigenvalue  $\lambda_1$  and the smallest eigenvalue  $\lambda_n$  of  $\mathbf{H}$ . A geometrical interpretation of the conditioning number is given in Figure 2. A low rank approximation of the Hessian can be obtained as

$$\hat{\mathbf{H}} = \mathbf{H}(V_0 | \dots | V_n) = (\mathbf{H}V_0 | \dots | \mathbf{H}V_n) \quad (2)$$

where the columns of  $(V_0 | \dots | V_n)$  form a basis of independent unit vectors that projects the Hessian on a low-dimensional random subspace. The conditioning number can then be estimated by performing a singular value decomposition of the projected Hessian  $\hat{\mathbf{H}}$  and taking the ratio between the largest and smallest resulting eigenvalues.

We estimate below the conditioning number  $\lambda_1/\lambda_n$  of a fully connected Boltzmann machine of 50 units at initial state ( $W = 0$ ) for different bias and offset parameters  $b$  and  $\beta$  using Equation 1 and 2:

$\lambda_1/\lambda_n$	$b = 2$	$b = 0$	$b = -2$
$\beta = \text{sigm}(2)$	<b>2.26</b>	21.97	839.59
$\beta = \text{sigm}(0)$	83.43	<b>2.75</b>	95.57
$\beta = \text{sigm}(-2)$	866.00	22.95	<b>2.24</b>

These numerical estimates clearly exhibit the better conditioning occurring when the sigmoid is centered. The more than 100-fold factor between the conditioning number of non-centered and centered Boltzmann machines is striking.

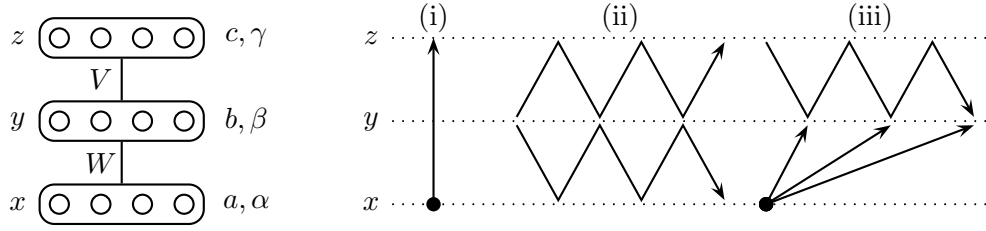


Figure 3: On the left, diagram of a two-layer deep Boltzmann machine along with its parameters. On the right, different sampling methods: (i) a feed-forward pass on the network starting from a data point, (ii) the path followed by the alternate Gibbs sampler and (iii) the path followed by the alternate Gibbs sampler when the input is clamped to data.

## 2.2 Centered Deep Boltzmann Machines

For technical and practical reasons, it is common to introduce a structure to the Boltzmann machine by restricting the connections between its units. A typical structure is the deep Boltzmann machine (DBM, Salakhutdinov and Hinton, 2009) in which units are organized in a deep layered architecture. The layered structure of the DBM has two advantages: first, it gives a specific role to units at each layer so that we can easily build top layer kernels that exploit the hierarchical structure of data. Second, the layered structure of the DBM can be folded into a bipartite graph from which it is easy to derive an efficient alternate Gibbs sampler. In the case of the two-layer deep Boltzmann machine shown in Figure 3, the energy function associated to each state  $(x, y, z) \in \{0, 1\}^{M_x + M_y + M_z}$  takes the form

$$E(x, y, z; \theta) = - (y - \beta)^\top W (x - \alpha) - (z - \gamma)^\top V (y - \beta) \\ - (x - \alpha)^\top a - (y - \beta)^\top b - (z - \gamma)^\top c$$

where  $\theta = \{W, V, a, b, c, \alpha, \beta, \gamma\}$  groups the model parameters. *Data-independent* states can be sampled using the following alternate Gibbs sampler:

$$\{x \sim \mathcal{B}(\text{sigm}(W^\top(y - \beta) + a)) \quad ; \quad z \sim \mathcal{B}(\text{sigm}(V(y - \beta) + c))\} \quad (3)$$

$$y \sim \mathcal{B}(\text{sigm}(W(x - \alpha) + V^\top(z - \gamma) + b)). \quad (4)$$

The same Gibbs sampler can be used for sampling *data-dependent* states at the difference that the input units  $x$  are clamped to the data. We show below a basic algorithm based on persistent contrastive divergence for training a two-layer centered DBM:

**Basic algorithm for training a 2-layer centered DBM:**

```

 $W, V = 0, 0$ 
 $a, b, c = \text{sigm}^{-1}(\langle x \rangle_{\text{data}}), b_0, c_0$ 
 $\alpha, \beta, \gamma = \text{sigm}(a), \text{sigm}(b), \text{sigm}(c)$ 
initialize free particle  $(x_m, y_m, z_m) = (\alpha, \beta, \gamma)$ 
loop
  initialize data particle  $(x_d, y_d, z_d) = (\text{pick}(\text{data}), \beta, \gamma)$ 
  loop
     $y_d \sim \mathcal{B}(\text{sigm}(W(x_d - \alpha) + V^\top(z_d - \gamma) + b))$ 
     $z_d \sim \mathcal{B}(\text{sigm}(V(y_d - \beta) + c))$ 
  end loop
   $y_m \sim \mathcal{B}(\text{sigm}(W(x_m - \alpha) + V^\top(z_m - \gamma) + b))$ 
   $x_m \sim \mathcal{B}(\text{sigm}(W^\top(y_m - \beta) + a))$ 
   $z_m \sim \mathcal{B}(\text{sigm}(V(y_m - \beta) + c))$ 
   $W = W + \eta \cdot [(y_d - \beta)(x_d - \alpha)^\top - (y_m - \beta)(x_m - \alpha)^\top]$ 
   $V = V + \eta \cdot [(z_d - \gamma)(y_d - \beta)^\top - (z_m - \gamma)(y_m - \beta)^\top]$ 
   $a = a + \eta \cdot (x_d - x_m)$ 
   $b = b + \eta \cdot (y_d - y_m)$ 
   $c = c + \eta \cdot (z_d - z_m)$ 
end loop

```

### 3. Discriminative Analysis

We present the method introduced by Montavon et al. (2011) that measures how the representation evolves layer after layer in a deep network. It is based on the theoretical insight that the projection of the input distribution onto the hidden units of each layer provides a function space that can be thought of as a representation or a feature extractor.

The method aims to characterize this function space by constructing a kernel for each layer that approximates the implicit transfer function between the input and the layer and measuring how much these kernels “match” the task of interest. The approach is theoretically motivated by the work of Braun et al. (2008) showing that projections on the leading components of the implicit kernel feature map (Schölkopf et al., 1998) obtained with a finite and typically small number of samples  $n$  are close with essentially multiplicative errors to their asymptotic counterparts. In the following lines, we describe the principal steps of the analysis:

Let  $X$  and  $T$  be two matrices of  $n$  rows representing respectively the inputs and labels of a data set of  $n$  samples. Let

$$f : x \mapsto f_L \circ \dots \circ f_1(x)$$

be a deep network of  $L$  layers. We build a hierarchy of increasingly “deep” kernels

$$\begin{aligned} k_{0,\sigma}(x, x') &= \kappa_\sigma(x, x') \\ k_{1,\sigma}(x, x') &= \kappa_\sigma(f_1(x), f_1(x')) \\ &\vdots \\ k_{L,\sigma}(x, x') &= \kappa_\sigma(f_L \circ \dots \circ f_1(x), f_L \circ \dots \circ f_1(x')) \end{aligned}$$

that subsume the mapping performed by more and more layers of the deep network and where  $\kappa_\sigma$  is an RBF kernel of scale  $\sigma$ . For each kernel  $k_{l,\sigma}$ , we can compute the empirical kernel  $K_{l,\sigma}$  of size  $n \times n$  and its eigenvectors  $u_{l,\sigma}^1, \dots, u_{l,\sigma}^n$  sorted by decreasing magnitude of their respective eigenvalues  $\lambda_{l,\sigma}^1, \dots, \lambda_{l,\sigma}^n$ .

We measure how good a representation is with respect to a certain task by measuring whether the task is contained within the leading principal components of the representation. The matrix

$$U_{l,\sigma}^d = (u_{l,\sigma}^1 \mid \dots \mid u_{l,\sigma}^d)$$

spans the  $d$  leading kernel principal components of empirical kernel. The error is obtained as the residuals of the projection of the labels  $T$  on the  $d$  leading components of the mapped distribution:

$$e_T(l, d, \sigma) = \|T - U_{l,\sigma}^d U_{l,\sigma}^{d\top} T\|_F^2$$

Curves  $(e(l, 0, \sigma), \dots, e(l, d, \sigma))$  represent how well the task can be solved as we add more and more principal components of the data distribution. These curves can be interpreted as learning curves as the regularization imposed by the rank of the kernel feature space determines the number of samples that are necessary in order to train the model effectively. Therefore, the number of observed kernel principal components  $d$  closely relates to the amount of label information given to the learning machine. Small values for  $d$  cover the “one-shot” learning regime where the model is asked to generalize from very few observations. On the other hand, large values for  $d$  cover the other extreme case where label information is abundant, and where the representation has to be rich enough in order to encode any subtle variation of the learning problem. For practical purposes, these curves can be reduced as follows:

$$e_T(l, d) = \min_{\sigma} e_T(l, d, \sigma) \tag{5}$$

$$e_T(l) = \frac{1}{n} \sum_{d=1}^n e_T(l, d) \tag{6}$$

These compact measures of how well layer  $l$  represent  $T$  make it easier to compare the layer-wise evolution of the representation for different architectures.

#### 4. Generative Analysis

Here, we present an analysis that estimates the likelihood of the learned Boltzmann machine (Salakhutdinov and Hinton, 2010) based on annealed importance sampling (AIS, Neal,

2001). We describe here the basic analysis. Salakhutdinov and Hinton (2010) introduced more elaborate procedures for particular types of Boltzmann machines such as restricted, semi-restricted and deep Boltzmann machines.

A deep Boltzmann machine associates to each input  $x$  a probability

$$p(x; \theta) = \frac{\Psi(\theta, x)}{Z(\theta)}$$

$$\begin{aligned} \text{where } \Psi(\theta, x) &= \sum_{y,z} p^*(x, y, z; \theta) \\ Z(\theta) &= \sum_{x,y,z} p^*(x, y, z; \theta) \end{aligned}$$

and where  $p^*(x, y, z; \theta) = e^{-E(x,y,z;\theta)}$  is the unnormalized probability of state  $(x, y, z)$ . Computing  $\Psi(\theta, x)$  and  $Z(\theta)$  analytically is intractable because of the exponential number of elements involved in the sum. Let us rewrite the ratio of partition functions as follows:

$$p(x; \theta) = \frac{\Psi(\theta, x)}{Z(\theta)} = \frac{\frac{\Psi(\theta, x)}{\Psi(0, x)}}{\frac{Z(\theta)}{Z(0)}} \cdot \frac{\Psi(0, x)}{Z(0)} \quad (7)$$

It can be first noticed that the ratio of base-rate partition functions ( $\theta = 0$ ) is easy to compute as  $\theta = 0$  makes units independent. It has the analytical form

$$\frac{\Psi(0, x)}{Z(0)} = \frac{1}{2^{M_x}}. \quad (8)$$

The two other ratios in Equation 7 can be estimated using annealed importance sampling. The annealed importance sampling method proceeds as follows:

**Annealed importance sampling:**

1. Generate a sequence of states  $\xi_1, \dots, \xi_T$  using a sequence of transition operators  $\mathcal{T}(\xi, \xi'; \theta_0), \dots, \mathcal{T}(\xi, \xi'; \theta_K)$  that leave  $p(\xi)$  invariant, that is,
  - Draw  $\xi_0$  from the base model (e.g. a random vector of zero and ones)
  - Draw  $\xi_1$  given  $\xi_0$  using  $\mathcal{T}(\xi, \xi'; \theta_1)$
  - ...
  - Draw  $\xi_K$  given  $\xi_{K-1}$  using  $\mathcal{T}(\xi, \xi'; \theta_K)$
2. Compute the importance weight

$$\omega_{\text{AIS}} = \frac{p^*(\xi_1; \theta_1)}{p^*(\xi_1; \theta_0)} \cdot \frac{p^*(\xi_2; \theta_2)}{p^*(\xi_2; \theta_1)} \cdot \dots \cdot \frac{p^*(\xi_K; \theta_K)}{p^*(\xi_K; \theta_{K-1})}$$



It can be shown that if the sequence of models  $\theta_0, \theta_1, \dots, \theta_K$  where  $\theta_0 = 0$  and  $\theta_K = \theta$  evolves slowly enough, the importance weight obtained with the annealed importance sampling procedure is an estimate for the ratio between the partition function of the model  $\theta$  and the partition function of the base rate model.

In our case,  $\xi$  denotes the state  $(x, y, z)$  of the DBM and the transition operator  $\mathcal{T}(\xi, \xi'; \theta)$  is the alternate Gibbs sampler defined in Equation 3. We can now compute the two ratios of partition functions of Equation 7 as

$$\frac{Z(\theta)}{Z(0)} \approx \mathbb{E}[\omega_{\text{AIS}}] \quad \text{and} \quad \frac{\Psi(\theta, x)}{\Psi(0, x)} \approx \mathbb{E}[\nu_{\text{AIS}}(x)] \quad (9)$$

where  $\omega_{\text{AIS}}$  is the importance weight resulting from the annealing process with the freely running Gibbs sampler and  $\nu_{\text{AIS}}$  is the importance weight resulting from the annealing with input units clamped to the data point. Substituting Equation 8 and 9 into Equation 7, we obtain

$$p(x; \theta) \approx \frac{\mathbb{E}[\nu_{\text{AIS}}(x)]}{\mathbb{E}[\omega_{\text{AIS}}]} \cdot \frac{1}{2^{M_x}}$$

and therefore, the log-likelihood of the model is

$$\mathbb{E}_X[\log(p(x; \theta))] \approx \mathbb{E}_X[\log \mathbb{E}[\nu_{\text{AIS}}(x)]] - \log \mathbb{E}[\omega_{\text{AIS}}] - M_x \log(2). \quad (10)$$

Generally, computing an average of the importance weight  $\nu_{\text{AIS}}$  for each data point  $x$  can take a long time. In practice, we can use an approximation to this computation where the estimate is computed with a single AIS run for each point. In that case, it follows from Jensen's inequality that

$$\mathbb{E}_X[\log \nu_{\text{AIS}}(x)] - \log \mathbb{E}[\omega_{\text{AIS}}] \leq \mathbb{E}_X[\log \mathbb{E}[\nu_{\text{AIS}}(x)]] - \log \mathbb{E}[\omega_{\text{AIS}}]. \quad (11)$$

Consequently, this approximation tends to produce slightly pessimistic estimates of the model log-likelihood, however the variance of  $\nu_{\text{AIS}}$  is low compared to the variance of  $\omega_{\text{AIS}}$  because the clamping of visible units to data points sharply reduces the diversity of AIS runs. We find that this approximation is sufficiently accurate for the purpose of this paper, that is, demonstrating the importance of centering deep Boltzmann machines.

## 5. Experimental Setup

In this section, we describe the different parameters used to train the deep Boltzmann machines and to perform the discriminative and generative analysis. These parameters correspond to reasonable choices, most of which have been validated by previous research work.

**Architecture** We consider two-layer deep Boltzmann machines made of 784 input units, 400 intermediate units and 100 top units. The initial biases and offsets for visible units are set to  $a_0 = \text{sigm}^{-1}(\langle x \rangle_{\text{data}})$  and  $\alpha = \text{sigm}(a)$ . We consider different initial biases ( $b_0, c_0 = -2$ ,  $b_0, c_0 = 0$  and  $b_0, c_0 = 2$ ) and offsets ( $\beta, \gamma = \text{sigm}(-2)$ ,  $\beta, \gamma = \text{sigm}(0)$  and  $\beta, \gamma = \text{sigm}(2)$ ) for the hidden units. These offsets and initial biases correspond to the sigmoids plotted in Figure 1.

**Data** We train the DBMs on a binary version of the MNIST handwritten digits data set where the activation threshold is set to 0.5 (medium gray). The MNIST training set consists of 60,000 samples. Each sample is a binary image of size  $28 \times 28$  representing a handwritten digit and is fed to the DBM as a 784-dimensional binary vector.

**Inference** We use persistent contrastive divergence (Tieleman, 2008) to train the network and keep track of 25 free particles in background of the learning procedure. We use a Gibbs sampling estimation to collect *both* the data-independent and data-dependent statistics. The rationale for this is that the more classical mean field estimation of data statistics (Salakhutdinov and Hinton, 2009) tends to artificially drive the DBM to sparsity due to the convex/concave shape of the sigmoid function. At each step of the learning procedure, we run 5 iterations of the alternate Gibbs sampler for collecting the data-dependent statistics and one iteration for updating the data-independent statistics.

**Learning** We use a stochastic gradient descent on the approximate log-likelihood with minibatches of size 25 and a learning rate  $\eta = 0.0005$  for each layer. For practical purposes, the minibatch size is set equivalent to the number of particles for persistent contrastive divergence (Hinton, 2010). We consider models trained for  $10^0$ ,  $10^{0.5}$ ,  $10^1$ ,  $10^{1.5}$  and  $10^2$  epochs.

**Model averaging** We use a variant of averaged stochastic gradient descent (Polyak and Juditsky, 1992; Tieleman and Hinton, 2009; Xu, 2011) for reducing the parameter noise. We compute at each step  $k$  the new parameter estimate  $\theta_{\text{avg}} \leftarrow \frac{k_c}{k+k_c} \cdot \theta + \frac{k}{k+k_c} \cdot \theta_{\text{avg}}$  with  $k_c = 10$  in order to only remember the last 10% of the training procedure.

**Discriminative analysis** The analysis is performed on a subset of 500 samples drawn randomly from the MNIST test set. Representations at each layer are built by running a Gibbs sampler for 100 iterations with the input clamped to data and taking the mean activation of each unit. Discriminative performance is measured as the projection residuals of the labels (see Equation 5) and the area under the error curve (see Equation 6). Results are produced with candidate scale parameters of the Gaussian kernel  $\sigma^2 = 1, 10, 100, 1000$  and 10000.

**Generative analysis** The generative analysis is performed on a subset of 500 samples drawn randomly from the MNIST test set. Generative performance is measured as the estimated log-likelihood of the model given the test data (see Equation 10). We estimate the partition function  $Z(\theta)/Z(0)$  using 500 AIS runs. We estimate each 500 partition functions  $\Psi(\theta, x)/\Psi(0, x)$  using a single AIS run. Each AIS run has length  $K = 2500$  where model parameter at the  $k^{\text{th}}$  step of the annealing process is defined as  $\theta_k = 1 - (1 - k\theta/K)^2$ . This sequence of parameters implies that annealing starts with large parameter updates and finishes with very small updates.

## 6. Results

Table 1 corroborates the importance of centering for better discriminating in the top layer of a deep Boltzmann machine. As it can be seen in Figure 5 (left), discriminative performance of the top layer can be further improved by training the network for a longer time.

AUC error	$b_0, c_0 = 2$	$b_0, c_0 = 0$	$b_0, c_0 = -2$
$\beta, \gamma = \text{sigm}(2)$	<b>0.119</b>	0.194	0.285
$\beta, \gamma = \text{sigm}(0)$	0.133	<b>0.090</b>	0.127
$\beta, \gamma = \text{sigm}(-2)$	0.368	0.323	<b>0.114</b>

Table 1: Discriminative performance after 10 epochs in the top layer of the deep Boltzmann machine as measured by Equation 6 for different configurations of initial bias and offset. The lower the AUC error the better. In each case, centering sigmoids leads to better discrimination in the top layer.

$\langle \log p(x; \theta) \rangle_{\text{data}}$	$b_0, c_0 = 2$	$b_0, c_0 = 0$	$b_0, c_0 = -2$
$\beta, \gamma = \text{sigm}(2)$	<b>-81.5*</b>	-86.5*	-88.9*
$\beta, \gamma = \text{sigm}(0)$	-83.5*	<b>-81.2*</b>	-85.6*
$\beta, \gamma = \text{sigm}(-2)$	-88.1*	-83.3*	<b>-80.4*</b>

Table 2: Generative performance after 10 epochs in terms of estimated model log-likelihood  $\langle \log p(x; \theta) \rangle_{\text{data}}$  for different configurations of initial bias and offset. The generative performance is less sensitive to the initial conditioning of the DBM than the top layer discriminative performance as the top-level units can simply be discarded, leading essentially to a more robust one-layer generative model.

Table 2 further supports the importance of centering, showing that centered DBMs learn a better generative model of data. However, the advantage is not as strong as for the discriminative case. Indeed, units in the top layer are not critical for generative performance as the learning algorithm can simply discard them and learn a one-layer shallow generative model instead.

Figure 4 and 5 highlight the importance of centering for faster and more stable learning. The models emerging from the centered deep Boltzmann machine have systematically better discriminative properties in the top layer and good generative properties. While a non-centered DBM may ultimately learn a model which is as good as the one produced by a centered DBM, it may also diverge.

Figure 6 and 7 show that each model is able to learn reasonable first-layer filters but that second-layer filters learned by a centered DBM tend to be more varied than those learned by a non-centered DBM. This higher variety of second layer filters suggests that the centered DBM produces a richer top-level representation. The argument is corroborated by Figure 9 showing that, in absence of centering mechanism, the projection of the data on the top layer representation tends to form a simplistic low-dimensional manifold that may still contain useful features (for example, discriminating the digit “1” from other digits) but, on the other hand, that also discards a lot of potentially useful discriminative features. As

---

\* In some other research work, authors are computing a lower bound of the log probability instead of a direct estimate of it, thus making a direct comparison impossible. Also, estimates of log probability become increasingly inaccurate as the model  $\theta$  complexifies.

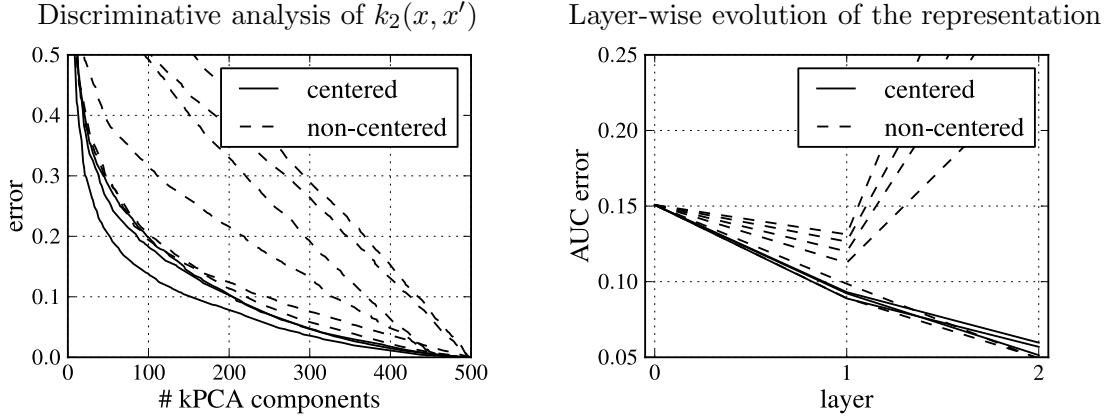


Figure 4: On the left, residuals of the projection of the labels in the leading components of the top layer kernel  $k_2(x, x')$  (see Equation 5) after 10 epochs. On the right, layer-wise evolution of the representation in terms of area under the error curve (see Equation 6) after 100 epochs. Centered DBMs are more stable than non-centered ones. Top layer representations are clearly better than the input.

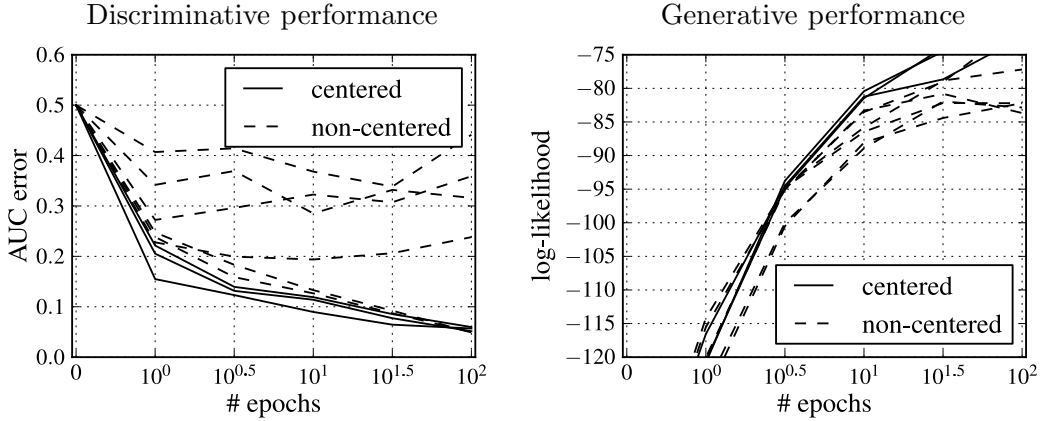


Figure 5: Convergence speed of centered and non-centered DBMs (in terms of top layer AUC error and model log-likelihood). Centered DBMs learn faster and are more stable than non-centered ones. Note that the estimate of the log-likelihood from Equation 10 becomes inaccurate as the model becomes more complex (after 10 epochs).

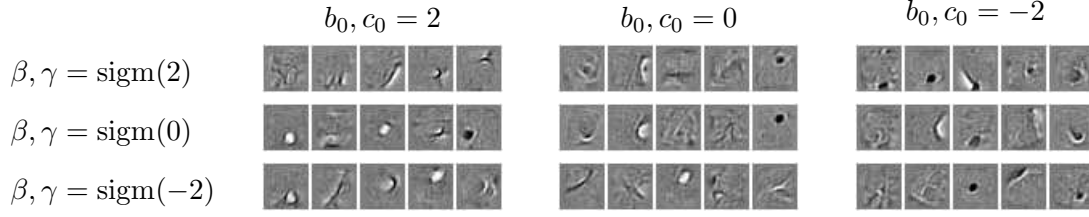


Figure 6: Examples of first-layer filters of the DBM for different bias and offset parameters after 100 epochs. These filters are rendered using a linear backprojection of top layer units onto the input space. Each model is producing reasonable first-layer filters, suggesting that one-layer networks (i.e. restricted Boltzmann machines) are less sensitive to the quality of the conditioning of the parameter space.

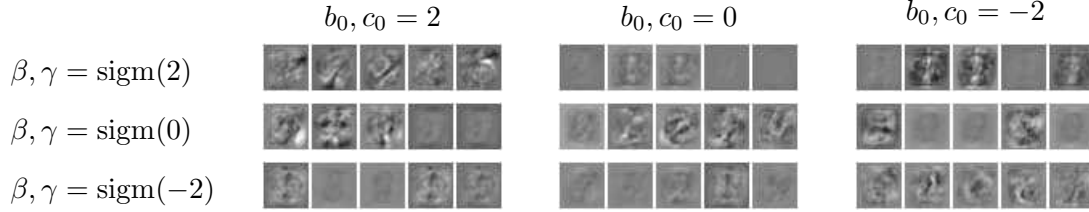


Figure 7: Examples of second-layer filters of the DBM for different bias and offset parameters after 100 epochs. These filters are rendered using a linear backprojection of intermediate layer units onto the input space. Here, we can clearly see that the diversity of filters is higher when the DBM is centered.

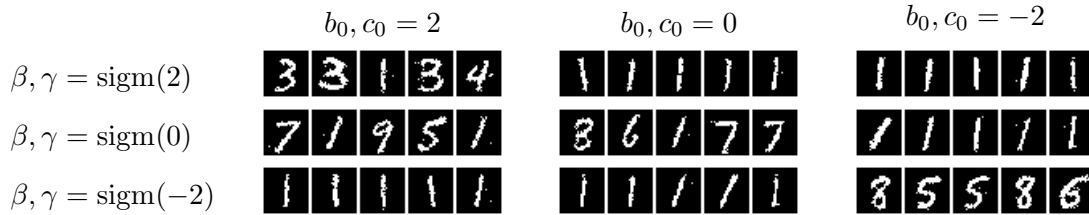


Figure 8: Examples of digits generated by the DBM for different bias and offset parameters after 10 epochs. The degenerated second layer of the non-centered DBM seems to have a negative impact on the balance between different classes.

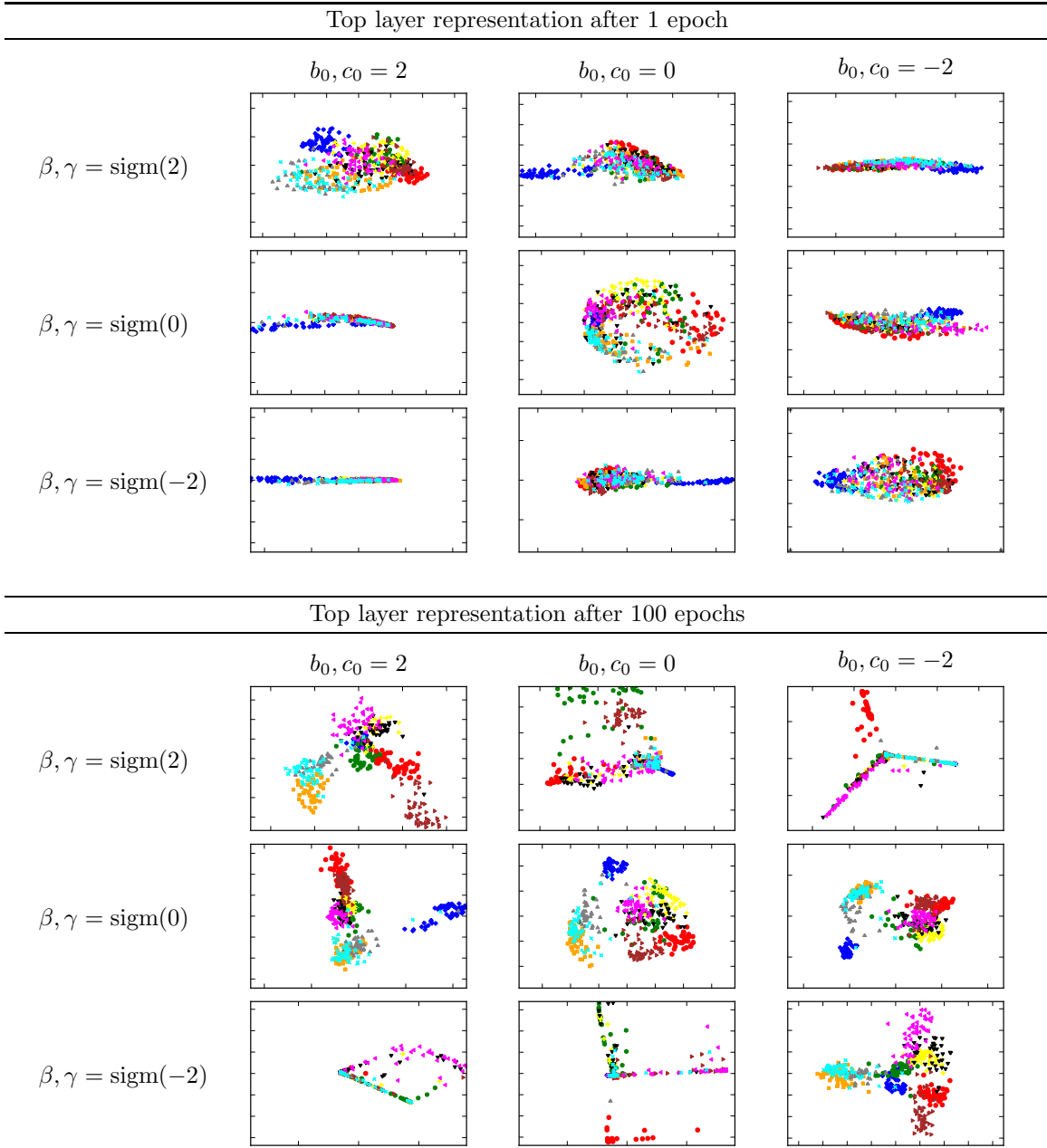


Figure 9: 2-kPCA visualization of the top-level representation in the DBM for different bias and offset parameters at different stages of training. Points are colored according to their label (“0”=red, “1”=blue, “2”=green, “3”=yellow, “4”=orange, “5”=black, “6”=brown, “7”=gray, “8”=magenta, “9”=cyan). Non-centered DBMs tend to collapse the data onto a simplistic low-dimensional manifold in the top layer representation. On the other other hand, in the centered DBM, we can clearly observe in the late stage of training the emergence of clusters corresponding to labels.

suggested by Figure 8, the top-layer simplistic representation may even negatively affect the generative properties of the model by perturbing the balance between different classes.

## 7. Conclusion

We presented a simple modification of the deep Boltzmann machine that centers the output of the sigmoids by rewriting the energy function as a function of centered states. This centered version of the deep Boltzmann machine is easy to implement as it simply involves a reparameterization of the energy function. A theoretical motivation for centering is that it leads to a better conditioning of the Hessian of the optimization criterion.

This simple modification allows to learn efficiently a deep Boltzmann machine without greedy layer-wise pretraining. Experiments on real data corroborate the benefits of centering, showing that the centered deep Boltzmann machine learns faster and is more stable than its non-centered counterpart. In addition, the centered deep Boltzmann machine produces useful discriminative features in the top layer and a good generative model of data.

Training hierarchies of many layers is still tedious and requires many iterations. Understanding whether the difficulty comes from a difficult optimization problem or from the exhaustion of statistical information in the data set remains to be done. Also, despite an initial good conditioning of the Hessian, it can not be excluded that the solution progressively drifts towards degenerate regions of the parameter space throughout the learning procedure. Strategies to dynamically maintain the solution within well-behaved regions of the parameter space or to better descend the objective function also need to be further investigated.

## Acknowledgments

This work was supported by the World Class University Program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology, under Grant R31-10008.

## References

- Mikio L. Braun, Joachim Buhmann, and Klaus-Robert Müller. On relevant dimensions in kernel feature spaces. *Journal of Machine Learning Research*, 9:1875–1908, Aug 2008.
- Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research - Proceedings Track*, 9:249–256, 2010.
- Geoffrey E. Hinton. A practical guide to training restricted Boltzmann machines. Technical report, University of Toronto, 2010.
- Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient backprop. In Genevieve B. Orr and Klaus-Robert Müller, editors, *Neural Networks—Tricks of the trade LNCS 1524*, pages 5–50. Springer, 1998.

- Grégoire Montavon, Mikio L. Braun, and Klaus-Robert Müller. Kernel analysis of deep networks. *Journal of Machine Learning Research*, 12:2563–2581, 2011.
- Grégoire Montavon, Mikio L. Braun, and Klaus-Robert Müller. Deep Boltzmann machines as feed-forward hierarchies. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.
- Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.
- Barak A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994.
- Boris T. Polyak and Anatoly B. Juditsky. Acceleration of stochastic approximation by averaging. *Siam J. Control Optim*, 30(4):838–855, 1992.
- Ruslan Salakhutdinov and Geoffrey E. Hinton. Deep Boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 5, pages 448–455, 2009.
- Ruslan Salakhutdinov and Geoffrey E. Hinton. An efficient learning procedure for deep Boltzmann machines. Technical Report MIT-CSAIL-TR-2010-037, MIT, 2010.
- Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.
- Yichuan Tang and Ilya Sutskever. Data normalization in the learning of restricted Boltzmann machines. Technical Report UTML-TR-11-2, Department of Computer Science, University of Toronto, 2011.
- Tijmen Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1064–1071, 2008.
- Tijmen Tieleman and Geoffrey E. Hinton. Using fast weights to improve persistent contrastive divergence. In *Proceedings of the 26th international conference on Machine learning*, pages 1033–1040, 2009.
- Wei Xu. Towards optimal one pass large scale learning with averaged stochastic gradient descent. *CoRR*, abs/1107.2490, 2011.