Inductive Principles for Restricted Boltzmann Machine Learning

Benjamin M. Marlin, Kevin Swersky, Bo Chen and Nando de Freitas
Department of Computer Science
University of British Columbia
{bmarlin, kswersky, bochen, nando}@cs.ubc.ca

Abstract

Recent research has seen the proposal of several new inductive principles designed specifically to avoid the problems associated with maximum likelihood learning in models with intractable partition functions. In this paper, we study learning methods for binary restricted Boltzmann machines (RBMs) based on *ratio matching* and *generalized score matching*. We compare these new RBM learning methods to a range of existing learning methods including stochastic maximum likelihood, contrastive divergence, and pseudo-likelihood. We perform an extensive empirical evaluation across multiple tasks and data sets.

1 Introduction

The prevalence of maximum likelihood (ML) as an inductive principle for estimating the parameters of statistical models is based on two key properties: asymptotic consistency and asymptotic efficiency (Fisher, 1922, p. 316). Asymptotic consistency means that the bias in the estimated parameters goes to zero as the sample size goes to infinity. Asymptotic efficiency means that the variance in the estimated parameters attains the minimum possible value among all consistent estimators as the sample size goes to infinity. Maximum likelihood estimation is simultaneously known to suffer from a variety of defects including a potentially large bias in the small sample setting, the possibility of un-bounded likelihood functions, the possibility that the maximum of the likelihood is in a region of low probability mass, to name only a few (MacKay, 2003, p. 305-306).

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

Our primary interest in this paper is the estimation of models where the partition function involves an exponential number of terms. This poses a fundamental problem for maximum likelihood estimation as the exact computation of the partition function (and thus the likelihood function) is intractable. This is an orthogonal issue to the more commonly encountered defects with maximum likelihood estimation outlined above, but is an extremely important problem in models for high dimensional discrete data including Ising models (MacKay, 2003, p. 400), restricted Boltzmann machines (Smolensky, 1986), and discrete exponential family harmonium models (Welling et al., 2005).

There are essentially two approaches to dealing with the intractability of the partition function in such models. The first approach is to approximately maximize the likelihood. The second approach is to select an alternative inductive principle that explicitly avoids the problems associated with an intractable partition function. Examples of the approximation approach include stochastic approximation-based maximum likelihood learning methods (Younes, 1989), as well as approximation methods based on belief propagation (Wainwright et al., 2003).

Well known examples of alternative inductive principles include the principle of maximum pseudo likelihood (PL) (Besag, 1975), and the principle of minimum contrastive divergence (CD) (Hinton, 2000). PL avoids the problems due to the partition function by defining an alternative criterion function based on products of conditional distributions. While CD is an alternative to maximum likelihood, the exact computation of the CD criterion function is itself intractable and applications of CD also require stochastic approximation.

Recently, two new alternative principles have been proposed for dealing with the problem of intractable partition functions that are applicable to models of discrete data: ratio matching (RM) (Hyvärinen, 2007) and generalized score matching (GSM) (Lyu, 2009). Like the principle of maximum pseudo-likelihood, these

principles avoid the partition function by defining different criteria based on conditional distributions.

This paper makes three main contributions. First is the development of a ratio matching-based learning method for binary restricted Boltzmann machines. The second is a correction to the derivation of generalized score matching and an analysis of the corrected objective function. Last is a comparison of ratio matching to existing learning methods for RBMs including approximate maximum likelihood, contrastive divergence, and pseudo-likelihood. We perform an extensive empirical evaluation across multiple tasks including density estimation, classification, novelty detection, and de-noising. We use a range of data sets including hand written digits, images, and text.

2 The Restricted Boltzmann Machine

A restricted Boltzmann machine is a two-layer undirected graphical model where the first layer consists of observed data variables (or visible units), and the second layer consists of latent variables (or hidden units). The visible layer is fully connected to the hidden layer via pair-wise potentials, while both the visible and hidden layers are restricted to have no within-layer connections.

We define D to be the number of data dimensions and K to be the number of hidden units. The space of visible vectors for a binary RBM is $\mathcal{X} = \{0,1\}^D$, while the space of hidden unit vectors is $\mathcal{H} = \{0,1\}^K$. We define $\boldsymbol{x}$ to be a visible vector of size $D \times 1$, and $\boldsymbol{h}$ to be hidden vector of size $K \times 1$. We use the notation $\boldsymbol{x}_{-d}$ to indicate the sub-vector formed from $\boldsymbol{x}$ by removing the d^{th} dimension. We use capital letters to denote random variables and lower case letters to denote their instantiations. We define N to be the number of data cases in a data set and $\boldsymbol{x}_n$ to be the n^{th} data case.

The RBM model can be defined in terms of the energy function of a joint configuration of the hidden and visible vectors $E(\boldsymbol{x}, \boldsymbol{h})$ as seen in Equation 2.1 where W are the weight parameters, b are the visible unit bias parameters and c are the hidden unit bias parameters. We represent W as a $D \times K$ matrix, b as a $D \times 1$ vector and c as a $K \times 1$ vector. We define $\theta = \{W, b, c\}$ to be the complete set of parameters for the model. Note that we follow the convention that high probability configurations have low energy.

$$E_\theta(\boldsymbol{x}, \boldsymbol{h}) = -(\boldsymbol{x}^T W \boldsymbol{h} + \boldsymbol{x}^T b + \boldsymbol{h}^T c) \tag{2.1}$$

The joint probability $P_\theta(\boldsymbol{x}, \boldsymbol{h})$ can in turn be defined through the energy $E_\theta(\boldsymbol{x}, \boldsymbol{h})$ as seen in Equation 2.2. The partition function $\mathcal{Z}$ of the RBM model is given in Equation 2.3. We see that the partition function involves a sum over all 2^D elements of the set $\mathcal{X}$, as well as the 2^K elements of the set $\mathcal{H}$. The probability of the visible vector $P_\theta(\boldsymbol{x})$ is given in Equation 2.4 and is obtained by marginalizing over the space $\mathcal{H}$ of hidden vectors.

$$P_\theta(\boldsymbol{x}, \boldsymbol{h}) = \frac{1}{\mathcal{Z}} \exp\left(-E_\theta(\boldsymbol{x}, \boldsymbol{h})\right) \tag{2.2}$$

$$\mathcal{Z} = \sum_{\boldsymbol{x}' \in \mathcal{X}} \sum_{\boldsymbol{h}' \in \mathcal{H}} \exp\left(-E_\theta(\boldsymbol{x}', \boldsymbol{h}')\right) \tag{2.3}$$

$$P_\theta(\boldsymbol{x}) = \frac{1}{\mathcal{Z}} \sum_{h \in \mathcal{H}} \exp\left(-E_\theta(\boldsymbol{x}, \boldsymbol{h})\right) \tag{2.4}$$

An important property of binary hidden unit restricted Boltzmann machines is that marginalization over the hidden vectors can be carried out analytically. This allows for an equivalent definition of the RBM model in terms of the *free energy* $F(\boldsymbol{x})$ as seen in Equation 2.5. We Note that the partition function given in Equation 2.6 still involves marginalizing over all 2^D elements of the set $\mathcal{X}$, rendering the computation of the partition function intractable for even moderate values of D.

$$P_\theta(\boldsymbol{x}) = \frac{1}{\mathcal{Z}} \exp\left(-F_\theta(\boldsymbol{x})\right) \tag{2.5}$$

$$\mathcal{Z} = \sum_{\boldsymbol{x}' \in \mathcal{X}} \exp\left(-F_\theta(\boldsymbol{x}')\right) \tag{2.6}$$

$$F_\theta(\boldsymbol{x}) = -\left(\boldsymbol{x}^T b + \sum_{k=1}^{K} \log\left(1 + \exp\left(\boldsymbol{x}^T W_k + c_k\right)\right)\right) \tag{2.7}$$

As we will see in Section 3, these two different views of the RBM model are useful for deriving different learning algorithms. The first view where the hidden units are explicitly represented (Equations 2.1 to 2.4) is useful for deriving stochastic approximation methods based on alternating sampling of the visible and hidden vectors. The second view where the hidden units have been analytically marginalized away (Equations 2.7 to 2.6) is useful for applying inductive principles that can not naturally deal with the presence of latent variables.

3 Inductive Principles

In this section we derive learning algorithms for the RBM model based on a number of inductive principles. We begin by deriving the standard stochastic approximation algorithm for maximum likelihood learning. Next, we turn to contrastive divergence, maximum pseudo-likelihood, ratio matching and generalized score matching.

3.1 Approximate Maximum Likelihood

The maximum likelihood principle states that we should select the parameters θ that assign the highest probability to the observed data. The maximum likelihood criterion function is given in Equation 3.8. The maximum likelihood principle can equivalently be viewed as selecting the parameters to minimize the Kullback-Leibler divergence $KL(P_e||P_\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}} P_e(\boldsymbol{x})\Big(\log P_e(\boldsymbol{x}) - \log P_\theta(\boldsymbol{x})\Big)$. This provides an important connection to other inductive principles based on minimizing differences between model and empirical distributions.

$$f^{ML}(\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}} P_e(\boldsymbol{x}) \log P_\theta(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N} \log P_\theta(\boldsymbol{x}_n) \tag{3.8}$$

The gradient of the maximum likelihood objective function is given in Equation 3.9. The gradient of the free energy $F_\theta(\boldsymbol{x})$ with respect to the weights and biases is given in Equation 3.10.

$$\nabla f^{ML} = \frac{-1}{N}\sum_{n=1}^{N}\left(\nabla F_\theta(\boldsymbol{x}_n) - \sum_{\boldsymbol{x'}\in\mathcal{X}} P_\theta(\boldsymbol{x'})\nabla F_\theta(\boldsymbol{x'})\right) \tag{3.9}$$

$$\nabla F_\theta(\boldsymbol{x}) = \{-\boldsymbol{x}E[\boldsymbol{h}|\boldsymbol{x}]^T,\ -\boldsymbol{x},\ -E[\boldsymbol{h}|\boldsymbol{x}]\} \tag{3.10}$$

Equation 3.11 gives the standard stochastic approximation to the maximum likelihood gradient where we replace the expectation under $P_\theta(\boldsymbol{x})$ by a sum over samples $\tilde{\boldsymbol{x}}_s$ drawn according to $P_\theta(\boldsymbol{x})$.

$$\nabla f^{ML} \approx -\left(\frac{1}{N}\sum_{n=1}^{N}\nabla F_\theta(\boldsymbol{x}_n) - \frac{1}{S}\sum_{s=1}^{S}\nabla F_\theta(\tilde{\boldsymbol{x}}_s)\right) \tag{3.11}$$

Sampling from $P_\theta(\boldsymbol{x})$ can be accomplished using a block Gibbs sampler applied to the joint distribution of visible and hidden vectors $P_\theta(\boldsymbol{x}, \boldsymbol{h})$. We alternate between sampling $\boldsymbol{x}$ according to $P_\theta(\boldsymbol{x}|\boldsymbol{h})$ and sampling $\boldsymbol{h}$ according to $P_\theta(\boldsymbol{h}|\boldsymbol{x})$. Importantly, the hidden units are all conditionally independent given the visible units and the visible units are all conditionally independent given the hidden units.

A naive implementation of the stochastic approximation algorithm would require running this Markov chain to equilibrium after every parameter update to draw a new set of samples. Younes (1989) showed that one can instead alternate between simulating the Markov chain for a single step and performing a parameter update using a small learning rate. This method has also been suggested by Tieleman (2008).

3.2 Contrastive Divergence

The inductive principle underlying contrastive divergence (CD) is to select the parameters θ that minimize the difference between the KL divergences $KL(P_e||P_\theta)$ and $KL(Q_\theta^t||P_\theta)$ (Hinton, 2000). Q_θ^t is defined to be the distribution obtained by applying t steps of the standard Gibbs kernel to the empirical distribution P_e. The criterion function for CD is given below.

$$f^{CD}(\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}} P_e(\boldsymbol{x}) \log\left(\frac{P_e(\boldsymbol{x})}{P_\theta(\boldsymbol{x})}\right) - Q_\theta^t(\boldsymbol{x}) \log\left(\frac{Q_\theta^t(\boldsymbol{x})}{P_\theta(\boldsymbol{x})}\right) \tag{3.12}$$

After an initial approximation of the CD gradient, we arrive at the form given in Equation 3.13, which involves an intractable expectation with respect to the t-step distribution Q_θ^t. We obtain a stochastic approximation algorithm by replacing the expectation with a sum over samples $\tilde{\boldsymbol{x}}_n$ drawn from Q_θ^t.

$$\nabla f^{CD} \approx \frac{-1}{N}\sum_{n=1}^{N}\left(\nabla F_\theta(\boldsymbol{x}_n) - \sum_{\boldsymbol{x'}\in\mathcal{X}} Q_\theta^t(\boldsymbol{x'})\nabla F_\theta(\boldsymbol{x'})\right) \tag{3.13}$$

$$\approx \frac{-1}{N}\sum_{n=1}^{N}\left(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\tilde{\boldsymbol{x}}_n)\right) \tag{3.14}$$

CD as proposed by Hinton (2000) uses a particular alternating Gibbs sampler where one Gibbs chain is run for t steps starting from each data case $\boldsymbol{x}_n$, resulting in N samples $\tilde{\boldsymbol{x}}_n$. Note that unlike the algorithm of Younes (1989), the chain is restarted at the data after every parameter update. Typically t is chosen to be quite small. We use $t = 1$ in the experiments that follow.

3.3 Maximum Pseudo-Likelihood

The inductive principle underlying maximum pseudo-likelihood states that we should select the parameters that maximize the product of all one-dimensional conditional distributions $P_\theta(x_d|\boldsymbol{x}_{-d})$ given the observed data as seen in Equation 3.15.

$$f^{PL}(\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{d=1}^{D} P_e(\boldsymbol{x}) \log P_\theta(x_d|\boldsymbol{x}_{-d}) \tag{3.15}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D} \log P_\theta(x_{dn}|\boldsymbol{x}_{-dn}) \tag{3.16}$$

We can equivalently view pseudo-likelihood as minimizing a sum of differences between the model and empirical distributions via the pseudo KL divergence $PKL(P_e||P_\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{d=1}^{D} P_e(\boldsymbol{x})\Big(\log P_e(x_d|\boldsymbol{x}_{-d}) -$

$\log P_\theta(x_d|\boldsymbol{x}_{-d})\Big)$. The gradient of the pseudo likelihood objective function is given in Equation 3.17 where $\bar{\boldsymbol{x}}_n^i$ is the data vector $\boldsymbol{x}_n$ with the value on the i^{th} dimension "flipped" as seen in Equation 3.18.

$$\nabla f^{PL} = \frac{-1}{N} \sum_{n,d} P_\theta(\bar{\boldsymbol{x}}_{dn}^d|\boldsymbol{x}_{-dn}) \left(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\bar{\boldsymbol{x}}_n^d)\right) \tag{3.17}$$

$$\bar{x}_{dn}^i = \begin{cases} 1 - x_{dn} & \text{if } i = d \\ x_{dn} & \text{if } i \neq d \end{cases} \tag{3.18}$$

3.4 Ratio Matching

The inductive principle underlying ratio matching can be interpreted as selecting the parameters θ to minimize a weighted sum of ℓ_2 distances between pairs of one-dimensional conditional distributions under the model and the empirical distribution. The ratio matching criterion function is given in Equation 3.19 (Hyvärinen, 2007).

$$f^{RM}(\theta) = \sum_{\boldsymbol{x} \in \mathcal{X}} \sum_{d=1}^{D} \sum_{\xi \in \{0,1\}} P_e(\boldsymbol{x}) \Big(P_\theta(X_d = \xi|\boldsymbol{x}_{-d}) - P_e(X_d = \xi|\boldsymbol{x}_{-d}) \Big)^2 \tag{3.19}$$

As shown by Hyvärinen (2007), the ratio matching criterion function can be reduced to the form given in Equation 3.20 where we replace the sum over the empirical distribution by a sum over individual training vectors, and C is a constant that does not depend on the model parameters.

$$f^{RM}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} g^2(u_{dn}) + C \tag{3.20}$$

$$g(u) = \frac{1}{1+u}, \quad u_{dn} = P_\theta(\boldsymbol{x}_n)/P_\theta(\bar{\boldsymbol{x}}_n^d) \tag{3.21}$$

This form of the criterion function depends on a ratio of probabilities where $\bar{\boldsymbol{x}}_n^d$ is defined in Equation 3.18. Importantly, both $P_\theta(\boldsymbol{x}_n)$ and $P_\theta(\bar{\boldsymbol{x}}_n^d)$ have the same partition function, so the partition functions cancel out in the ratio. The gradient of the ratio matching objective function is given in Equation 3.22.

$$\nabla f^{RM} = \frac{2}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} g(u_{dn})^3 u_{dn} \left(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\bar{\boldsymbol{x}}_n^d)\right) \tag{3.22}$$

3.5 Generalized Score Matching

The inductive principle underlying generalized score matching also involves an ℓ_2 distance applied to the difference of the inverses of the conditional probabilities as seen in Equation 3.23 (Lyu, 2009).

$$f^{GSM}(\theta) = \sum_{\boldsymbol{x} \in \mathcal{X}} \sum_{d=1}^{D} P_e(\boldsymbol{x}) \Big(\frac{1}{P_\theta(x_d|\boldsymbol{x}_{-d})} - \frac{1}{P_e(x_d|\boldsymbol{x}_{-d})} \Big)^2 \tag{3.23}$$

The generalized score matching criterion can be reduced to a form that only depends on ratios of probabilities under the RBM model as shown in Equation 3.24 (we use the definition of $\bar{\boldsymbol{x}}_n^d$ introduced in Equation 3.18). Note that the original formula in Lyu's paper contains an error corresponding to $g(u) = u^{-2} + u^2$, which we correct here (Lyu, 2009).

$$f^{GSM}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} g(u_{dn}) + C \tag{3.24}$$

$$g(u) = u^{-2} - 2u, \quad u_{dn} = P_\theta(\boldsymbol{x}_n)/P_\theta(\bar{\boldsymbol{x}}_n^d) \tag{3.25}$$

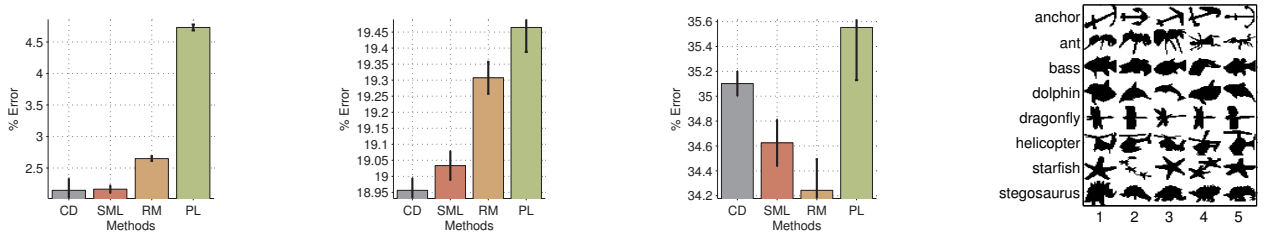
The gradient of the GSM objective function is given in Equation 3.26.

$$\nabla f^{GSM} = \frac{2}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} (u_{dn}^{-2} - u_{dn}) \left(\nabla F_\theta(\boldsymbol{x}_n) - \nabla F_\theta(\bar{\boldsymbol{x}}_n^d)\right) \tag{3.26}$$

Interestingly, the corrected reduced form can be unbounded below in the limit as u goes to positive infinity since it is dominated by the negative linear term $-2u$. The derivation of the reduced form given by Lyu begins by dropping the constant terms $P_e(\boldsymbol{x})/P_e(x_d|\boldsymbol{x}_{-d})^2$. Consider two configurations $\boldsymbol{x}$ and $\boldsymbol{x}'$ that are equal on every dimension except d where $x'_d = 1 - x_d$. In the limit as $P_e(\boldsymbol{x}) = \epsilon$ goes to zero while $P_e(\boldsymbol{x}') = \alpha$ goes to a non-zero limit, the constant contribution will be $(\epsilon + \alpha)^2/\epsilon$, which goes to infinity. As the corresponding model conditional $P_\theta(x_d|\boldsymbol{x}_{-d})$ goes to zero (or equivalently the ratio of model probabilities for $\boldsymbol{x}'$ and $\boldsymbol{x}$ goes to infinity), the objective function (Equation 3.23) with the constant terms dropped goes to negative infinity. Unfortunately, this means that the reduced GSM objective function can not be applied to high-dimensional real data where many $P_e(\boldsymbol{x})$ will be zero. The basic objective function is also problematic since this same analysis shows that the we can not simply ignore the contributions from data cases where $P_e(\boldsymbol{x}) = 0$. Indeed, the GSM objective function puts maximum weight on getting the corresponding model conditional distributions to exactly equal zero. In the subsequent experiments, we apply GSM with synthetic data only to illustrate its behavior as some $P_e(\boldsymbol{x})$ go to zero.

3.6 Discussion

The common presentation of the learning methods derived from each inductive principle makes the relation-



(a) MNIST (b) 20News (c) CalTech (d) CalTech Data Sample

Figure 1: (a)-(c) Test set classification error. Lower classification error indicates better performance. (d) shows a sample of the CalTech101 Silhouettes data.

ships between them much easier to discern. First, we observe that all the methods deal with the intractable partition function by only computing conditional probabilities or ratios. In SML and CD, this is used within the Gibbs sampler while in PL, RM, and GSM it is used directly in the definition of the inductive principle. Next, we can easily see that the gradient for each method is based on the difference between the gradient of the free energy evaluated at a data case and the gradient of the free energy evaluated at a "fantasy" case. In SML the fantasy cases are sampled from the equilibrium distribution of the model, while in CD they are sampled using t steps of the Gibbs kernel applied to each data case. By contrast, PL, RM and GSM all use the same fixed set of fantasy cases consisting of the one-neighbours of each data case. PL, RM, and GSM differ from each other only in terms of the weighting of the gradient contributions. Finally, our analysis of GSM shows that not all objective functions and resulting gradient weighting terms will be well behaved. It is particularly important from an optimization perspective that the objective functions be bounded below.

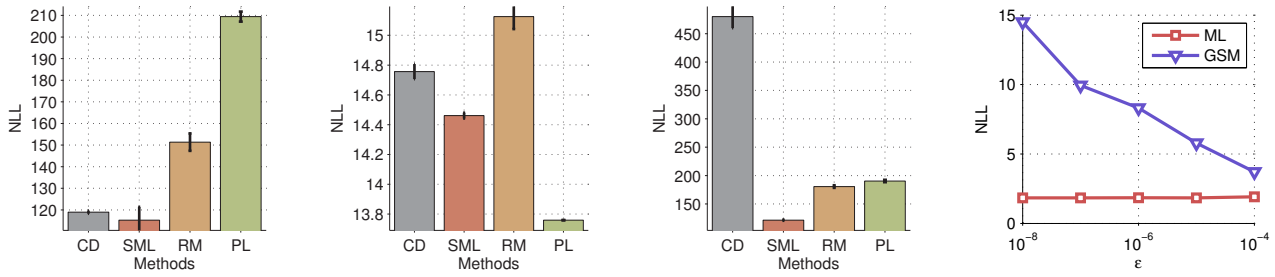
4 Data and Training Protocols

We use three data sets in the experimental comparisons: MNIST[1], the small 20-Newsgroups[2] data set, and a new binary image data set derived from CalTech101.[3] MNIST consists of 28×28-size images of hand-written digits from 0 through 9. We binarize the images and divide the database into a training set of $50,000$ examples and a validation set of $10,000$ examples, and use the standard $10,000$ example testset. The small 20-Newsgroups data set contains newsgroup postings divided into four classes of groups. Each posting is represented by binary vectors over a vocabulary of 100 words. We split this data set into 8500 training, 1245 validation, and 6497 test examples respectively. The final data set we use is derived from the object outlines contained in the CalTech101 annotations data set. The object outlines were centered and scaled on a 28×28 image plane and rendered as filled black regions on a white background creating a silhouette of each object. We call this data set *Caltech101 Silhouettes* [4]. We show examples from several of the 101 classes in Figure 1(d). The training set contains 4100 examples with at least 20, and at most 100 examples from each class. The remaining images were split among a test set and validation set of size 2307 and 2264 respectively.

Our training protocol begins with 100 iterations of Stochastic Gradient Descent (SGD) on mini-batches of 100 data cases using momentum and iterate averaging as acceleration techniques (Robbins and Monro, 1951). To reduce computation time, we select the SGD learning rate, momentum parameter and whether or not to use iterate averaging separately for each method by minimizing a performance measure on a smaller subset of each data set. For CD and SML, we chose one-step reconstruction error as the performance measure, while for RM and PL we used their objective function values. We found that CD benefited most from the use of momentum without averaging, while the rest of the methods did well without momentum, but with averaging.

We fine-tune methods with computable objective functions, using an LBFGS optimizer (Nocedal and Wright, 2000, p. 224). We select a weight decay setting specific to each data set, experiment, and method by training on the full training set with a validation set held-out. Lastly, we do five full training runs for each data set, experiment, and method using the selected learning rate, momentum, iterate averaging and weight decay settings and average the results. The learning rate range was $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The momentum range was $\{0, 0.3, 0.5, 0.8\}$. The weight decay range was $\{10^{-5}, 10^{-4}, 10^{-3}\}$. We use

[1] `http://yann.lecun.com/exdb/mnist`
[2] `http://www.cs.toronto.edu/~roweis/data.html`
[3] `http://www.vision.caltech.edu/Image\_Datasets`
[4] `http://www.cs.ubc.ca/~bmarlin/data/`



(a) MNIST (b) 20News (c) CalTech (d) Synthetic

Figure 2: (a) to (c) show test set negative log likelihood (NLL) estimated using AIS. (d) shows a synthetic data experiment comparing GSM and exact ML.

500 hidden units for all experiments.

5 Experiments and Results

• **Classification Error:** We test the feature extraction performance of each learning algorithm by estimating a support vector machine classifier based on the class labels y_n and mean-field hidden unit activation vectors $\hat{\boldsymbol{h}}_n = E[\boldsymbol{h}|\boldsymbol{x}_n]$. We use a smooth, multi-class, linear SVM with an L2-penalty selected to minimize validation set error (Lee and Mangasarian, 2001). We report the corresponding test-set error in Figure 1. The results show that stochastic maximum likelihood (SML) has the most consistent performance across the data sets.

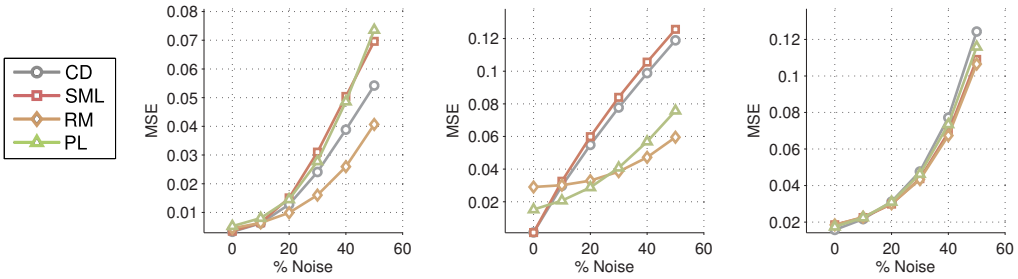
• **Test Set Log Likelihood:** There have been significant recent advances in Monte Carlo methods for estimating the log partition function of an RBM that are computationally feasible for post-training analysis. We compute test set log likelihoods by estimating the log partition function once for each trained model using the annealed importance sampling (AIS) method proposed by Murray and Salakhutdinov (2009). We show the results of this evaluation in Figure 2. We see a win for the stochastic approximate maximum likelihood (SML) method, which achieves the best average test set log likelihood on MNIST and CalTech and the second best on 20 Newsgroups. This is not surprising since SML is the only method specifically trained to optimize the likelihood.

Figure 2(d) shows an experiment comparing generalized score matching and exact maximum likelihood in terms of log likelihood as the probability of some data cases goes to zero. We use 9 data dimensions and 10 hidden units. We randomly select 6 data configurations from the 2^9 configurations and assign each a weight of 1. We assign all other configurations a weight of ϵ. We normalize the weights over all 2^9 data configurations to form the empirical probability distribution. The graph shows the log likelihood under models trained by exact ML and by GSM. As predicted by our analysis, GSM diverges as ϵ approaches zero while ML is unaffected.

• **De-Noising:** Reconstruction and de-noising are commonly used to asses the performance of RBM models and auto-encoders. We consider a de-noising task where we select a certain fraction of bits in each test data case $\boldsymbol{x}_n$ and set them to 0 or 1 with even probability, creating a noisy version of the data case $\widetilde{\boldsymbol{x}}_n$. We then compute the 1-step mean field reconstruction of $\boldsymbol{x}_n$ using $\hat{\boldsymbol{h}}_n = E[\boldsymbol{h}|\widetilde{\boldsymbol{x}}_n]$ and $\hat{\boldsymbol{x}}_n = E[\boldsymbol{x}_n|\hat{\boldsymbol{h}}_n]$. We measure the average per-dimension reconstruction error using mean squared error between $\boldsymbol{x}_n$ and $\hat{\boldsymbol{x}}_n$: $(1/ND)\sum_{n=1}^{N}||\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n||_2^2$. The results of this analysis are presented in Figure 3 (note that in the error bars are inside the line markers and were not plotted). These results show a consistent advantage for ratio matching (RM) across the three data sets.

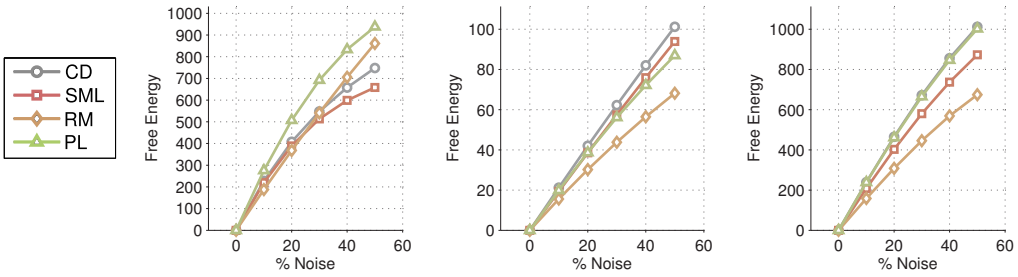
• **Novelty Detection:** We consider a novelty detection task that looks at how the free energy of test cases varies as we add random noise. We select bits at random and set them to 0 or 1 with even probability. We report the relative free energy defined as the difference between the free energy at a given noise level and the free energy at the zero-noise baseline. The results for this task are presented in Figure 4 (a) to (c) (note that in the error bars are inside the line markers and were not plotted). Pseudo-likelihood is consistently the most sensitive to noise with a rate of free energy increase that is similar to or higher than the other methods on all of the data sets.

• **Visualization:** As a final qualitative assessment of the trained RBM models, we display the learned weights W and visible biases b for each training method on the MNIST data set. We select the regularization setting that results in the lowest reconstruction error on the MNIST data set. The weights are shown in Figure 5 where the top left cell in each figure is the



(a) MNIST (b) 20News (c) CalTech

Figure 3: Test set de-noising mean squared error as a function of percent noisy bits



(a) MNIST (b) 20News (c) CalTech

Figure 4: Relative test set free energy as a function of percent noisy bits.

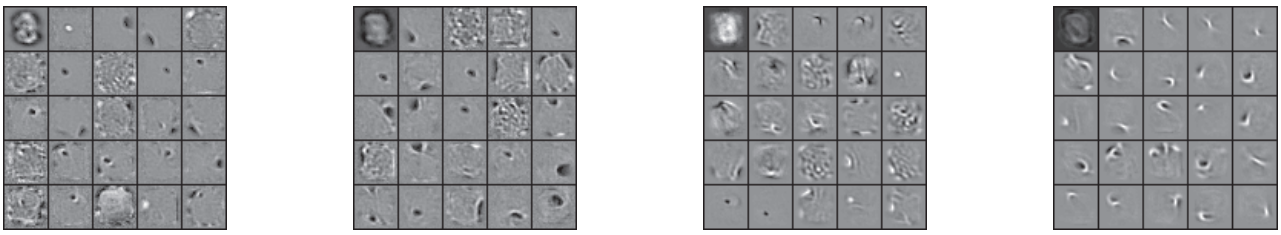
visible bias vector. We sort the hidden units from highest to lowest average activation on test data, and show the weights for every 20^{th} unit in the sorted list. The weight matrix for each hidden unit can be thought of as a linear filter or feature detector where a larger filter response will result a larger value of $P_\theta(h_k = 1|\boldsymbol{x})$.

Figure 5(a) shows that contrastive divergence (CD) learns a mixture of localized spot filters and noisy non-localized filters. This result is in excellent agreement with previous results by both Hinton (2007, Figure 3) and Ranzato et al. (2007, Figure 1(e)), which show an almost identical mixture of spot filters and non-localized filters. The filters learned by stochastic maximum likelihood (SML) are very similar to those learned by CD as seen in Figure 5(b). Pseudo-likelihood learns a mixture of very narrow spot filters, short/narrow stroke filters, and noisy non-localized filters as seen in Figure 5(c). Finally, Figure 5(d) shows that ratio matching (RM) learns highly localized stroke filters. Interestingly, the RM filters are very similar to previous results on MNIST obtained using sparse RBMs (Lee et al., 2008, Figure 2) and other sparse coding models (Ranzato et al., 2007, Figure 1(d)), even though the ratio matching objective does not include an explicit sparsity term.

• **Computation Time:** Basic implementations of pseudo-likelihood (PL) and ratio matching (RM) have a computational complexity that is approximately D times higher than stochastic maximum likelihood (SML) and contrastive divergence (CD) due to the fact that CD and SML consider one fantasy data case per training case, while PL and RM consider D fantasy cases corresponding to the D possible single bit flips. A more careful implementation of PL and RM can reduce this gap by re-using intermediate computations. On the 100 dimensional 20 Newsgroups data we observe that PL and RM are approximately 10 times slower than SML and CD. On the 784 dimensional MNIST and CalTech101 data sets, they are approximately 16 times slower.

6 Conclusions

Our analysis in Section 3 shows that all of the methods we consider differ only in the distribution of fantasy data they use and how they weight the gradient contributions from different training cases. However, these differences are meaningful as the methods exhibit very different theoretical and empirical characteristics. Our analysis of the generalized score matching criterion has revealed that it is ill suited for use on real data since it is not well behaved when some data cases have zero probability. Empirically, we find that the stochastic maximum likelihood (SML) method has consistently better performance in terms of density estimation, which is explained by the fact that it is the only method based on the maximum likelihood



(a) CD (b) SML (c) PL (d) RM

Figure 5: Learned weights and visible biases on the MNIST data set. The top left cell in each figure is the visible bias vector. Black corresponds to a weight of -1 while white corresponds to a weight of $+1$.

principle. The classification results show that all the methods give fairly comparable performance, although SML contrastive divergence (CD) are definitely better on MNIST. RM has a consistent advantage over the other methods in terms of de-noising, which corroborates well with the observation that the filters it produces are localized strokes instead of spots. On the novelty detection task pseudo-likelihood exhibits the most sensitivity. Finally, our most efficient implementations for ratio matching and pseudo likelihood are still an order of magnitude slower than SML and CD. Taking computation time into account, SML is certainly the most attractive method.

Future work on alternative inductive principles for RBMs will need to seriously consider the issue of computational complexity. We are currently investigating whether the complexity of RM and PL can be effectively reduced by considering fewer bit flips for each training case. We are also investigating the use of score matching in Gaussian/Binary RBMs where it has essentially the same computational complexity as SML. Finally, we plan to investigate the application of accelerated RM and PL methods for greedy layer-wise training of deep architectures.

Acknowledgments

The authors would like to thank the CIFAR NCAP program, NSERC, and the Killam Trusts for supporting this work. We would also like to thank Siwei Lyu for his personal correspondence regarding GSM.

References

J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London Series A.*, 222:309–368, 1922.

G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14, 2000.

G. E. Hinton. To recognize shapes, first learn to generate images. *Progress in brain research*, 165:535–547, 2007.

Aapo Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, 2007.

H. Lee, C. Ekanadham, and A. Y. Ng. Sparse deep belief net model for visual area V2. In *Advances in Neural Information Processing Systems 20*. MIT Press, 2008.

Y.J. Lee and O.L. Mangasarian. SSVM: A smooth support vector machine for classification. *Computational optimization and Applications*, 20(1):5–22, 2001.

Siwei Lyu. Interpretation and generalization of score matching. In *Uncertainty in Artificial Intelligence 25*, 2009.

D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

I. Murray and R. Salakhutdinov. Evaluating probabilities under high-dimensional latent variable models. In *Advances in Neural Information Processing Systems 21*, pages 1137–1144, 2009.

J. Nocedal and S. Wright. *Numerical Optimization*. Springer, Second edition, 2000.

M. Ranzato, Y-L. Boureau, and Y. LeCun. Sparse feature learning for deep belief networks. In *Advances in Neural Information Processing Systems 19*. MIT Press, 2007.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407, 1951.

P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*, pages 194–281, 1986.

T. Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *International Conference on Machine Learning 25*, pages 1064–1071, 2008.

M. Wainwright, T. Jaakkola, and A. Willsky. Tree-reweighted belief propagation algorithms and approximate ML estimation via pseudo-moment matching. In *Workshop on Artificial Intelligence and Statistics 9*, 2003.

M. Welling, M. Rosen-Zvi, and G. Hinton. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems 17*, pages 1481–1488, 2005.

L. Younes. Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields*, 82(4):625–645, 1989.