Stochastic Neural Networks with Monotonic Activation Functions

Siamak Ravanbakhsh **Barnabás Póczos** **Jeff Schneider**
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213

Dale Schuurmans **Russell Greiner**
University of Alberta, Edmonton, AB T6G 2E8, Canada

Abstract

We propose a Laplace approximation that creates a stochastic unit from any smooth monotonic activation function, using only Gaussian noise. This paper investigates the application of this stochastic approximation in training a family of Restricted Boltzmann Machines (RBM) that are closely linked to Bregman divergences. This family, that we call exponential family RBM (Exp-RBM), is a subset of the exponential family Harmoniums that expresses family members through a choice of smooth monotonic non-linearity for each neuron. Using contrastive divergence along with our Gaussian approximation, we show that Exp-RBM can learn useful representations using novel stochastic units.

1 Introduction

Deep neural networks (LeCun et al., 2015; Bengio, 2009) have produced some of the best results in complex pattern recognition tasks where the training data is abundant. Here, we are interested in deep learning for generative modeling. Recent years has witnessed a surge of interest in directed generative models that are trained using (stochastic) back-propagation (*e.g.*, Kingma and Welling, 2013; Rezende et al., 2014; Goodfellow et al., 2014). These models are distinct from deep energy-based models – including deep Boltzmann machine (Hinton et al., 2006) and (convolutional) deep belief network (Salakhutdinov and Hinton, 2009; Lee et al., 2009) – that rely on a bipartite graphical model called restricted Boltzmann machine (RBM) in each layer. Although, due to their use of additive Gaussian noise, the stochastic units that we introduce in this paper can be used for reparametrization trick in stochastic back-propagation, this paper is limited to applications in RBM.

To this day, the choice of stochastic units in RBM has been constrained to well-known members of the exponential family; in the past RBMs have used units with Bernoulli (Smolensky, 1986), Gaussian (Freund and Haussler, 1994; Marks and Movellan, 2001), categorical (Welling et al., 2004), Gamma (Welling et al., 2002) and Poisson (Gehler et al., 2006) conditional distributions. The exception to this specialization, is the Rectified Linear Unit that was introduced with a (heuristic) sampling procedure (Nair and Hinton, 2010).

This limitation of RBM to well-known exponential family members is despite the fact that Welling et al. (2004) introduced a generalization of RBMs, called Exponential Family Harmoniums (EFH), covering a large subset of exponential family with bipartite structure. The architecture of EFH does not suggest a procedure connecting the EFH to *arbitrary non-linearities* and more importantly a general sampling procedure is missing.[1] We introduce a useful subset of the EFH, which we call exponential family RBMs (Exp-RBMs), with an approximate sampling procedure addressing these shortcomings.

The basic idea in Exp-RBM is simple: restrict the sufficient statistics to identity function. This allows definition of each unit using only its mean stochastic activation, which is the non-linearity of the neuron. With this restriction, not only we gain interpretability, but also trainability; we show that it is possible

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

[1] As the concluding remarks of Welling et al. (2004) suggest, this capability is indeed desirable:"A future challenge is therefore to start the modelling process with the desired non-linearity and to subsequently introduce auxiliary variables to facilitate inference and learning."

to efficiently sample the activation of these stochastic neurons and train the resulting model using contrastive divergence. Interestingly, this restriction also closely relates the generative training of Exp-RBM to discriminative training using the matching loss and its regularization by noise injection.

In the following, Section 2 introduces the Exp-RBM family and Section 3 investigates learning of Exp-RBMs via an efficient approximate sampling procedure. Here, we also establish connections to discriminative training and produce an interpretation of stochastic units in Exp-RBMs as an infinite collection of Bernoulli units with different activation biases. Section 4 demonstrates the effectiveness of the proposed sampling procedure, when combined with contrastive divergence training, in data representation.

2 The Model

The conventional RBM models the joint probability $p(v, h \mid W)$ for visible variables $v = [v_1, \ldots, v_i, \ldots, v_I]$ with $v \in \mathcal{V}_1 \times \ldots \times \mathcal{V}_I$ and hidden variables $h = [h_1, \ldots, h_j, \ldots h_J]$ with $h \in \mathcal{H}_1 \times \ldots \times \mathcal{H}_J$ as

$$p(v, h \mid W) = \exp(-E(v, h) - A(W)).$$

This joint probability is a Boltzmann distribution with a particular energy function $E : \mathcal{V} \times \mathcal{H} \to \mathbb{R}$ and a normalization function A. The distinguishing property of RBM compared to other Boltzmann distributions is the conditional independence due to its bipartite structure.

Welling et al. (2004) construct Exponential Family Harmoniums (EFH), by first constructing independent distribution over individual variables: considering a hidden variable h_j, its sufficient statistics $\{t_b\}_b$ and canonical parameters $\{\tilde{\eta}_{j,b}\}_b$, this independent distribution is

$$p(h_j) = r(h_j) \exp\left(\sum_b \tilde{\eta}_{j,b}\, t_b(h_j) - A(\{\tilde{\eta}_{j,b}\}_b) \right)$$

where $r : \mathcal{H}_j \to \mathbb{R}$ is the *base measure* and $A(\{\eta_{i,a}\}_a)$ is the normalization constant. Here, for notational convenience, we are assuming functions with distinct inputs are distinct – *i.e.*, $t_b(h_j)$ is not necessarily the same function as $t_b(h_{j'})$, for $j' \neq j$.

The authors then combine these independent distributions using quadratic terms that reflect the bipartite structure of the EFH to get its joint form

$$p(v, h) \propto \exp\left(\sum_{i,a} \tilde{\nu}_{i,a}\, t_a(v_i) + \sum_{j,b} \tilde{\eta}_{j,b}\, t_b(h_j) + \sum_{i,a,j,b} W_{i,j}^{a,b} t_a(v_i) t_b(h_j) \right) \quad (1)$$

where the normalization function is ignored and the base measures are represented as additional sufficient statistics with fixed parameters. In this model, the conditional distributions are

$$p(v_i \mid h) = \exp\left(\sum_a \nu_{i,a} t_a(v_j) - A(\{\nu_{i,a}\}_a) \right)$$
$$p(h_j \mid v) = \exp\left(\sum_b \eta_{j,b} t_b(h_j) - A(\{\eta_{j,b}\}_b) \right)$$

where the *shifted* parameters $\eta_{j,b} = \tilde{\eta}_{j,b} + \sum_{i,a} W_{i,j}^{a,b} t_a(v_i)$ and $\nu_{i,a} = \tilde{\nu}_{i,a} + \sum_{j,b} W_{i,j}^{a,b} t_b(h_j)$ incorporate the effect of evidence in network on the random variable of interest.

It is generally not possible to efficiently sample these conditionals (or the joint probability) for arbitrary sufficient statistics. More importantly, the joint form of Equation (1) and its energy function are "obscure". This is in the sense that the base measures $\{r\}$, depend on the choice of sufficient statistics and the normalization function $A(W)$. In fact for a fixed set of sufficient statistics $\{t_a(v_i)\}_i, \{t_b(h_j)\}_j$, different compatible choices of normalization constants and base measures may produce diverse subsets of the exponential family. Exp-RBM is one such family, where sufficient statistics are identity functions.

2.1 Bregman Divergences and Exp-RBM

Exp-RBM restricts the sufficient statistics $t_a(v_i)$ and $t_b(h_j)$ to single identity functions v_i, h_j for all i and j. This means the RBM has a single weight matrix $W \in \mathbb{R}^{I \times J}$. As before, each hidden unit j, receives an input $\eta_j = \sum_i W_{i,j} v_i$ and similarly each visible unit i receives the input $\nu_i = \sum_j W_{i,j} h_j$. [2]

Here, the conditional distributions $p(v_i \mid \nu_i)$ and $p(h_j \mid \eta_j)$ have a single *mean parameter*, $f(\eta) \in \mathcal{M}$, which is equal to the mean of the conditional distribution. We could freely assign any desired continuous and monotonic non-linearity $f : \mathbb{R} \to \mathcal{M} \subseteq \mathbb{R}$ to represent the mapping from canonical parameter η_j to this mean parameter: $f(\eta_j) = \int_{\mathcal{H}_j} h_j p(h_j \mid \eta_j)\, \mathrm{d}h_j$. This choice of f defines the conditionals

$$p(h_j \mid \eta_j) = \exp\left(-D_f(\eta_j \,\|\, h_j) + g(h_j) \right) \quad (2)$$
$$p(v_i \mid \nu_i) = \exp\left(-D_f(\nu_i \,\|\, v_i) + g(v_i) \right)$$

where g is the base measure and D_f is the Bregman divergence for the function f.

[2] Note that we ignore the "bias parameters" $\tilde{\nu}_i$ and $\tilde{\eta}_j$, since they can be encoded using the weights for additional hidden or visible units ($h_j = 1, v_i = 1$) that are clamped to one.

The Bregman divergence (Bregman, 1967; Banerjee et al., 2005) between h_j and η_j for a monotonically increasing transfer function (corresponding to the activation function) f is given by[3]

$$D_f(\eta_j \,\|\, h_j) = -\eta_j h_j + F(\eta_j) + F^*(h_j) \tag{3}$$

where F with $\frac{\mathrm{d}}{\mathrm{d}\eta}F(\eta_j) = f(\eta_j)$ is the anti-derivative of f and F^* is the anti-derivative of f^{-1}. Substituting this expression for Bregmann divergence in Equation (2), we notice both F^* and g are functions of h_j. In fact, these two functions are often not separated (*e.g.*, McCullagh et al., 1989). By separating them we see that some times, g simplifies to a constant, enabling us to approximate Equation (2) in Section 3.1.

Example 2.1. Let $f(\eta_j) = \eta_j$ be a linear neuron. Then $F(\eta_j) = \frac{1}{2}\eta_j^2$ and $F^*(h_j) = \frac{1}{2}h_j^2$, giving a Gaussian conditional distribution $p(h_j \mid \eta_j) = e^{-\frac{1}{2}(h_j-\eta_j)^2 - g(h_j)}$, where $g(h_j) = -\log(\sqrt{2\pi})$ is a constant.

2.2 The Joint Form

So far we have defined the conditional distribution of our Exp-RBM as members of, using a single mean parameter $f(\eta_j)$ (or $f(\nu_i)$ for visible units) that represents the activation function of the neuron. Now we would like to find the corresponding joint form and the energy function.

The problem of relating the local conditionals to the joint form in graphical models goes back to the work of Besag (1974).It is easy to check that, using the more general treatment of Yang et al. (2012), the joint form corresponding to the conditional of Equation (2) is

$$p(v,h \mid W) \;=\; \exp\Bigg(v^T \cdot W \cdot h \tag{4}$$
$$-\sum_i \big(F^*(v_i) + g(v_i)\big) - \sum_j \big(F^*(h_j) + g(h_j)\big) - A(W)\Bigg)$$

where $A(W)$ is the joint normalization constant. It is noteworthy that only the anti-derivative of f^{-1}, F^* appears in the joint form and F is absent. From this, the energy function is

$$E(v,h) = -v^T \cdot W \cdot h \tag{5}$$
$$+ \sum_i \big(F^*(v_i) + g(v_i)\big) + \sum_j \big(F^*(h_j) + g(h_j)\big).$$

Example 2.2. For the sigmoid non-linearity $f(\eta_j) = \frac{1}{1+e^{-\eta_j}}$, we have $F(\eta_j) = \log(1+e^{\eta_j})$ and $F^*(h_j) = (1-h_j)\log(1-h_j) + h_j\log(h_j)$ is the negative entropy. Since $h_j \in \{0,1\}$ only takes extreme values, the negative entropy $F^*(h_j)$ evaluates to zero:

$$p(h_j \mid \eta_j) = \exp\Big(h_j\eta_j - \log(1+\exp(\eta_j)) + g(h_j)\Big) \tag{6}$$

Separately evaluating this expression for $h_j = 0$ and $h_j = 1$, shows that the above conditional is a well-defined distribution for $g(h_j) = 0$, and in fact it turns out to be the sigmoid function itself – *i.e.*, $p(h_j = 1 \mid \eta_j) = \frac{1}{1+e^{-\eta_j}}$. When all conditionals in the RBM are of the form Equation (6) – *i.e.*, for a binary RBM with a sigmoid non-linearity, since $\{F(\eta_j)\}_j$ and $\{F(\nu_i)\}_i$ do not appear in the joint form Equation (4) and $F^*(0) = F^*(1) = 0$, the joint form has the simple and the familiar form $p(v,h) = \exp\big(v^T \cdot W \cdot h - A(W)\big)$.

3 Learning

A consistent estimator for the parameters W, given observations $\mathcal{D} = \{v^{(1)}, \dots, v^{(N)}\}$, is obtained by maximizing the marginal likelihood $\prod_n p(v^{(n)} \mid W)$, where the Equation (4) defines the joint probability $p(v,h)$. The gradient of the log-marginal-likelihood $\nabla_W\big(\sum_n \log(p(v^{(n)} \mid W))\big)$ is

$$\frac{1}{N}\sum_n \mathbb{E}_{p(h|v^{(n)},W)}[h \cdot (v^{(n)})^T] \;-\; \mathbb{E}_{p(h,v|W)}[h \cdot v^T] \tag{7}$$

where the first expectation is w.r.t. the observed data in which $p(h \mid v) = \prod_j p(h_j \mid v)$ and $p(h_j \mid v)$ is given by Equation (2). The second expectation is w.r.t. the model of Equation (4).

When discriminatively training a neuron $f(\sum_i W_{i,j}v_i)$ using input output pairs $\mathcal{D} = \{(v^{(n)}, h_j^{(n)})\}_n$, in order to have a loss that is convex in the model parameters $W_{:j}$, it is common to use a *matching loss* for the given transfer function f (Helmbold et al., 1999). This is simply the Bregman divergence $D_f(f(\eta_j^{(n)}) \| h_j^{(n)})$, where $\eta_j^{(n)} = \sum_i W_{i,j}v_i^{(n)}$. Minimizing this matching loss corresponds to maximizing the log-likelihood

[3] The conventional form of Bregman divergence is $D_f(\eta_j \,\|\, h_j) = F(\eta_j) - F(f^{-1}(h_j)) - h_j(\eta_j - f^{-1}(h_j))$, where F is the anti-derivative of f. Since F is strictly convex and differentiable, it has a Legendre-Fenchel dual $F^*(h_j) = \sup_{\eta_j}\langle h_j, \eta_j\rangle - F(\eta_j)$. Now, set the derivative of the r.h.s. w.r.t. η_j to zero to get $h_j = f(\eta_j)$, or $\eta_j = f^{-1}(h_j)$, where $F^*(h_j)$ is the anti-derivative of $f^{-1}(h_i)$. Using the duality to switch f and f^{-1} in the above we can get $F(f^{-1}(h_j)) = h_j f^{-1}(h_j) - F^*(h_j)$. By replacing this in the original form of Bregman divergence we get the alternative form of Equation (3).

unit name	non-linearity $f(\eta)$	Gaussian approximation	conditional dist $p(h \mid \eta)$
Sigmoid (Bernoulli) Unit	$(1+e^{-\eta})^{-1}$	-	$\exp\{\eta h - \log(1+\exp(\eta))\}$
Noisy Tanh Unit	$(1+e^{-\eta})^{-1} - \frac{1}{2}$	$\mathcal{N}(f(\eta), (f(\eta)-1/2)(f(\eta)+1/2))$	$\exp\{\eta h - \log(1+\exp(\eta)) + \text{ent}(h) + g(h)\}$
ArcSinh Unit	$\log(\eta + \sqrt{1+\eta^2})$	$\mathcal{N}(\sinh^{-1}(\eta), (\sqrt{1+\eta^2})^{-1})$	$\exp\{\eta h - \cosh(h) + \sqrt{1+\eta^2} - \eta \sin^{-1}(\eta) + g(h)\}$
Symmetric Sqrt Unit (SymSqU)	$\text{sign}(\eta)\sqrt{\lvert\eta\rvert}$	$\mathcal{N}(f(\eta), \sqrt{\lvert\eta\rvert}/2)$	$\exp\{\eta h - \lvert h\rvert^3/3 - 2(\eta^2)^{\frac{3}{4}}/3 + g(h)\}$
Linear (Gaussian) Unit	η	$\mathcal{N}(\eta, 1)$	$\exp\{\eta h - \frac{1}{2}(\eta^2) - \frac{1}{2}(h^2) - \log(\sqrt{2\pi})\}$
Softplus Unit	$\log(1+e^{\eta})$	$\mathcal{N}(f(\eta), (1+e^{-\eta})^{-1})$	$\exp\{\eta h - 2\text{Li}_2(-e^{\eta}) - h\log(1-e^h) + y\log(e^{\eta}-1) + g(h)\}$
Rectified Linear Unit (ReLU)	$\max(0, \eta)$	$\mathcal{N}(f(\eta), \mathbb{I}(\eta > 0))$	-
Rectified Quadratic Unit (ReQU)	$\max(0, \eta\lvert\eta\rvert)$	$\mathcal{N}(f(\eta), \mathbb{I}(\eta > 0)\eta)$	-
Symmetric Quadratic Unit (SymQU)	$\eta\lvert\eta\rvert$	$\mathcal{N}(\eta\lvert\eta\rvert, \lvert\eta\rvert)$	$\exp\{\eta h - \lvert\eta\rvert^3/3 - 2(h^2)^{\frac{3}{4}}/3 + g(h)\}$
Exponential Unit	e^{η}	$\mathcal{N}(e^{\eta}, e^{\eta})$	$\exp\{\eta h - e^{\eta} - h(\log(y) - 1) + g(h)\}$
Sinh Unit	$\frac{1}{2}(e^{\eta} - e^{-\eta})$	$\mathcal{N}(\sinh(\eta), \cosh(\eta))$	$\exp\{\eta h - \cosh(\eta) + \sqrt{1+h^2} - h\sin^{-1}(h) + g(h)\}$
Poisson Unit	e^{η}	-	$\exp\{\eta h - e^{\eta} - y!\}$

Table 1: *Stochastic units, their conditional distribution (Equation* (2)*) and the Gaussian approximation to this distribution. Here* $\text{Li}(\cdot)$ *is the polylogarithmic function and* $\mathbb{I}(\text{cond.})$ *is equal to one if the condition is satisfied and zero otherwise.* $\text{ent}(p)$ *is the binary entropy function.*

of Equation (2), and it should not be surprising that the gradient $\nabla_{W_{:j}}\big(\sum_n D_f(f(\eta_j^{(n)}) \| h_j^{(n)})\big)$ of this loss w.r.t. $W_{:j} = [W_{1,j}, \ldots, W_{M,j}]$

$$\sum_n f(\eta_j^{(n)})(v^{(n)})^T \; - \; h_j^{(n)}(v^{(n)})^T$$

resembles that of Equation (7), where $f(\eta_j^{(n)})$ above substitutes h_j in Equation (7).

However, note that in generative training, h_j is not simply equal to $f(\eta_j)$, but it is sampled from the exponential family distribution Equation (2) with the mean $f(\eta_j)$ – that is $h_j = f(\eta_j) + \text{noise}$. This extends the previous observations linking the discriminative and generative (or regularized) training – via Gaussian noise injection – to the noise from other members of the exponential family (*e.g.*, An, 1996; Vincent et al., 2008; Bishop, 1995) which in turn relates to the regularizing role of generative pretraining of neural networks (Erhan et al., 2010).

Our sampling scheme (next section) further suggests that when using output Gaussian noise injection for regularization of arbitrary activation functions, the **variance of this noise should be scaled by the derivative of the activation function.**

3.1 Sampling

To learn the generative model, we need to be able to sample from the distributions that define the expectations in Equation (7). Sampling from the joint model can also be reduced to alternating conditional sampling of visible and hidden variables (*i.e.*, block Gibbs sampling). Many methods, including contrastive divergence (CD; Hinton, 2002), stochastic maximum likelihood (a.k.a. persistent CD Tieleman, 2008) and their variations (*e.g.*, Tieleman and Hinton, 2009; Breuleux et al., 2011) only require this alternating sampling in order to optimize an approximation to the gradient of Equation (7).

Here, we are interested in sampling from $p(h_j \mid \eta_j)$ and $p(v_i \mid \nu_i)$ as defined in Equation (2), which is in general non-trivial. However some members of the exponential family have relatively efficient sampling procedures (Ahrens and Dieter, 1974). One of these members that we use in our experiments is the Poisson distribution.

Example 3.1. For a Poisson unit, a Poisson distribution

$$p(h_j \mid \lambda) = \frac{\lambda^{h_j}}{h_j!} e^{-\lambda} \tag{8}$$

represents the probability of a neuron firing h_j times in a unit of time, given its average rate is λ. We can define Poisson units within Exp-RBM using $f_j(\eta_j) = e^{\eta_j}$, which gives $F(\eta_j) = e^{\eta_j}$ and $F^*(h_j) = h_j(\log(h_j) - 1)$. For $p(h_j \mid \eta_j)$ to be properly normalized, since $h_j \in \mathbb{Z}^+$ is a non-negative integer, $F^*(h_j) + g(h_j) = \log(h_j!) \approx F^*(h_j)$ (using Sterling's approximation). This gives $p(h_j \mid \eta_j) = \exp\big(h_j\eta_j - e^{\eta_j} - \log(h_j!)\big)$ which is identical to distribution of Equation (8), for $\lambda = e^{\eta_j}$. This means, we can use any available sampling routine for Poisson distribution to learn the parameters for an exponential family RBM where some units are Poisson. In Section 4, we use a modified version of Knuth's method (Knuth, 1969) for Poisson sampling.

By making a simplifying assumption, the following Laplace approximation demonstrates how to use Gaussian noise to sample from general conditionals in Exp-RBM, for "any" smooth and monotonic non-linearity.

Proposition 3.1. *Assuming a constant base measure* $g(h_i) = c$*, the distribution of* $p(h_j \| \eta_j)$ *is to the second*

order approximated by a Gaussian

$$\exp\Big(- D_f(\eta_j \,\|\, h_j) + c \Big) \approx \mathcal{N}(\, h_j \mid f(\eta_j),\, f'(\eta_j)\,) \tag{9}$$

where $f'(\eta_j) = \frac{\mathrm{d}}{\mathrm{d}\eta_j} f(\eta_j)$ *is the derivative of the activation function.*

Proof. The mode (and the mean) of the conditional Equation (2) for η_j is $f(\eta_j)$. This is because the Bregman divergence $D_f(\eta_j \| h_j)$ achieves minimum when $h_j = f(\eta_j)$. Now, write the Taylor series approximation to the target log-probability around its mode

$$\begin{aligned}
&\log(\, p(\varepsilon + f(\eta_j) \mid \eta_j\,)) \\
&= -D_f(\eta_j \| \varepsilon + f(\eta_j)) + c \\
&= \eta_j f(\eta_j) - F^*(f(\eta_j)) - F(\eta_j) \\
&+ \varepsilon(\eta_j - f^{-1}(f(\eta_j)) + \frac{1}{2}\varepsilon^2(\frac{-1}{f'(\eta_j)}) + \mathcal{O}(\varepsilon^3) \qquad (10a)\\
&= \eta_j f(\eta_j) - (\eta_j f(\eta_j) - F(\eta_j)) - F(\eta_j) \qquad (10b)\\
&+ \varepsilon(\eta_j - \eta_j) + \frac{1}{2}\varepsilon^2(\frac{-1}{f'(\eta_j)}) + \mathcal{O}(\varepsilon^3) \\
&= -\frac{1}{2}\frac{\varepsilon^2}{f'(\eta_j)} + \mathcal{O}(\varepsilon^3) \qquad (10c)
\end{aligned}$$

In Equation (10a) we used the fact that $\frac{\mathrm{d}}{\mathrm{d}y} f^{-1}(y) = \frac{1}{f'(f^{-1}(y))}$ and in Equation (10b), we used the conjugate duality of F and F^*. Note that the final unnormalized log-probability in Equation (10c) is that of a Gaussian, with mean zero and variance $f'(\eta_j)$. Since our Taylor expansion was around $f(\eta_j)$, this gives us the approximation of Equation (9). □

3.1.1 Sampling Accuracy

To exactly evaluate the accuracy of our sampling scheme, we need to evaluate the conditional distribution of Equation (2). However, we are not aware of any analytical or numeric method to estimate the base measure $g(h_j)$. Here, we replace $g(h_j)$ with $\tilde{g}(\eta_j)$, playing the role of a normalization constant. We then evaluate

$$p(h_j \mid \eta_j) \approx \exp\big(- D_f(\eta_j \,\|\, h_j) + \tilde{g}(\eta_j)\big) \tag{11}$$

where $\tilde{g}(\eta_j)$ is numerically approximated for each η_j value. Figure 1 compares this density against the Gaussian approximation $p(h_j \mid \eta_j) \approx \mathcal{N}(\, f(\eta_j),\, f'(\eta_j)\,)$. As the figure shows, the densities are very similar.

3.2 Bernoulli Ensemble Interpretation

This section gives an interpretation of Exp-RBM in terms of a Bernoulli RBM with an infinite collection of Bernoulli units. Nair and Hinton (2010) introduce the softplus unit, $f(\eta_j) = \log(1+e^{\eta_j})$, as an approximation to the rectified linear unit (ReLU) $f(\eta_j) = \max(0, \eta_j)$. To have a probabilistic interpretation for this nonlinearity, the authors represent it as an infinite series of Bernoulli units with shifted bias:

$$\log(1 + e^{\eta_j}) = \sum_{n=1}^{\infty} \sigma(\eta_j - n + .5) \tag{12}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. This means that the sample y_j from a softplus unit is effectively the number of active Bernoulli units. The authors then suggest using $h_j \sim \max(0, \mathcal{N}(\eta_j, \sigma(\eta_j))$ to sample from this type of unit. In comparison, our Proposition 3.1 suggests using $h_j \sim \mathcal{N}(\log(1 + e^{\eta_j}), \sigma(\eta_j))$ for softplus and $h_j \sim \mathcal{N}(\max(0, \eta_j), \mathit{step}(\eta_j))$ – where $\mathit{step}(\eta_j)$ is the step function – for ReLU. Both of these are very similar to the approximation of (Nair and Hinton, 2010) and we found them to perform similarly in practice as well.

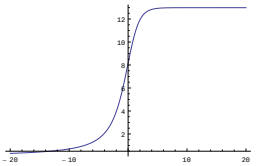

Figure 2: *Numerical approximation to the integral* $\int_{\mathcal{H}_j} \exp\big(- D_f(\eta_j \| h_j)\big) \mathrm{d}h_j$ *for the softplus unit* $f(\eta_j) = \log(1 + e^{\eta_j})$*, at different* η_j*.*

Note that these Gaussian approximations are assuming $g(\eta_j)$ is constant. However, by numerically approximating $\int_{\mathcal{H}_j} \exp\big(- D_f(\eta_j \| h_j)\big) \mathrm{d}h_j$, for $f(\eta_j) = \log(1 + e^{\eta_j})$, Figure 2 shows that the integrals are not the same for different values of η_j, showing that the base measure $g(h_j)$ is not constant for ReLU. In spite of this, experimental results for pretraining ReLU units using Gaussian noise suggests the usefulness of this type of approximation.

We can extend this interpretation as a collection of (weighted) Bernoulli units to any non-linearity f. For simplicity, let us assume $\lim_{\eta \to -\infty} f(\eta) = 0$ and $\lim_{\eta \to +\infty} f(\eta) = \infty$[4], and define the following series of Bernoulli units: $\sum_{n=0}^{\infty} \alpha\sigma(f^{-1}(\alpha n))$, where the given parameter α is the weight of each unit. Here, we are

[4]The following series and the sigmoid function need to be adjusted depending on these limits. For example, for the case where h_j is antisymmetric and unbounded (*e.g.*, $f(\eta_j) \in \{\sinh(\eta_j), \sinh^{-1}(\eta_j), \eta_j\,|\eta_j|\}$), we need to change the domain of Bernoulli units from $\{0, 1\}$ to $\{-.5, +.5\}$. This corresponds to changing the sigmoid to hyperbolic tangent $\frac{1}{2}\tanh(\frac{1}{2}\eta_j)$. In this case, we also need to change the bounds for n in the series of Equation (13) to $\pm\infty$.

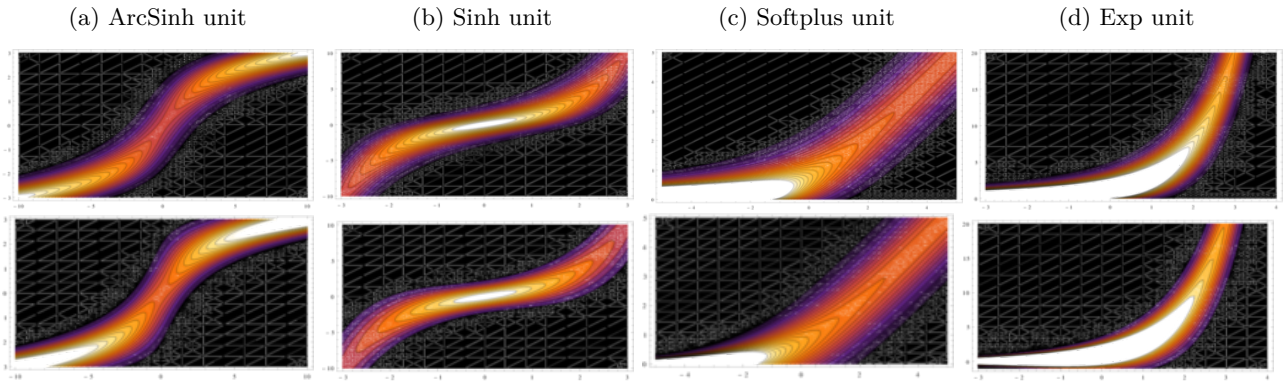

Figure 1: *Conditional probability of Equation* (11) *for different stochastic units (top row) and the Gaussian approximation of Proposition 3.1 (bottom row) for the same unit. Here the horizontal axis is the input* $\eta_j = \sum_i W_{i,j} v_i$ *and the vertical axis is the stochastic activation* h_j *with the intensity* $p(h_j \mid \eta_j)$*. see Table 1 for more details on these stochastic units.*

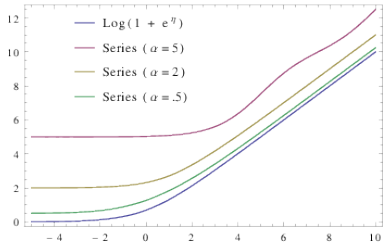

Figure 3: *reconstruction of ReLU by as a series of Bernoulli units with shifted bias.*

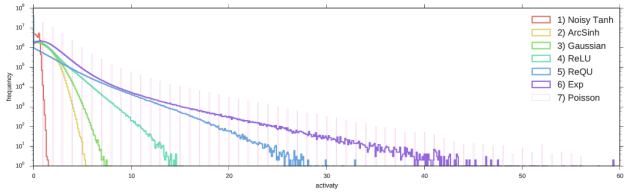

Figure 4: *Histogram of hidden variable activities on the MNIST test data, for different types of units. Units with heavier tails produce longer strokes in Figure 5. Note that the linear decay of activities in the log-domain correspond to exponential decay with different exponential coefficients.*

defining a new Bernoulli unit with a weight α for each α unit of change in the value of f. Note that the underlying idea is similar to that of inverse transform sampling (Devroye, 1986). At the limit of $\alpha \to 0^+$ we have

$$f(\eta_j) \approx \alpha \sum_{n=0}^{\infty} \sigma(\eta_j - f^{-1}(\alpha n)) \tag{13}$$

that is $\hat{h}_j \sim p(h_j \mid \eta_j)$ is the weighted sum of active Bernoulli units. Figure 4(a) shows the approximation of this series for the softplus function for decreasing values of α.

4 Experiments and Discussion

We evaluate the representation capabilities of Exp-RBM for different stochastic units in the following two sections. Our initial attempt was to adapt Annealed Importance Sampling (AIS; Salakhutdinov and Murray, 2008) to Exp-RBMs. However, estimation of the importance sampling ratio in AIS for general Exp-RBM proved challenging. We consider two alternatives: 1) for large datasets, Section 4.1 qualitatively evaluates the filters learned by various units and; 2) Section 4.2 evaluates Exp-RBMs on a smaller dataset where we can use indirect sampling likelihood to quantify the generative quality of the models with different activation functions.

Our objective here is to demonstrate that a combination of our sampling scheme with contrastive divergence (CD) training can indeed produce generative models for a diverse choice of activation function.

4.1 Learning Filters

In this section, we used CD with a single Gibbs sampling step, 1000 hidden units, Gaussian visible units[5], mini-batches and method of momentum, and selected the learning rate from $\{10^{-2}, 10^{-3}, 10^{-4}\}$ using reconstruction error at the final epoch.

The MNIST handwritten digits dataset (LeCun et al., 1998) is a dataset of 70,000 "size-normalized and centered" binary images. Each image is 28×28 pixel, and represents one of $\{0, 1, \ldots, 9\}$ digits. See the first row of Figure 5 for few instances from MNIST dataset. For this dataset we use a momentum of .9 and train each model for 25 epochs. Figure 5 shows the filters of dif-

[5]Using Gaussian visible units also assumes that the input data is normalized to have a standard deviation of 1.

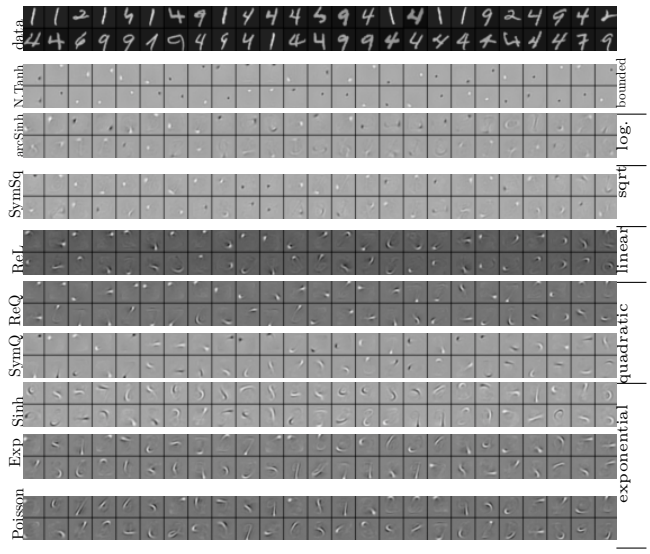

Figure 5: *Samples from the MNIST dataset (first two rows) and the filters with highest variance for different Exp-RBM stochastic units (two rows per unit type). From top to bottom the non-linearities grow more rapidly, also producing features that represent longer strokes.*

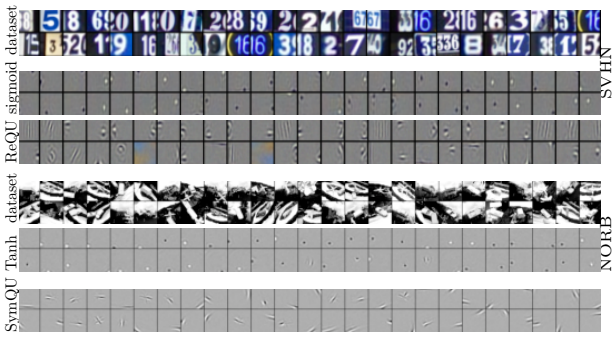

Figure 6: *Samples and the receptive fields of different stochastic units for from the* ***(top three rows)*** *SVHN dataset and* ***(bottom three rows)*** 48×48 *(non-stereo) NORB dataset with jittered objects and cluttered background. Selection of the receptive fields is based on their variance.*

ferent stochastic units; see Table 1 for details on different stochastic units. Here, the units are ordered based on the asymptotic behavior of the activation function f; see the right margin of the figure. This asymptotic change in the activation function is also evident from the hidden unit activation histogram of Figure 4(b), where the activation are produced on the test set using the trained model.

These two figures suggest that transfer functions with faster asymptotic growth, have a more heavy-tailed distributions of activations and longer strokes for the MNIST dataset, also hinting that they may be preferable in learning representation (*e.g.*, see Olshausen and Field, 1997). However, this comes at the cost of trainability. In particular, for all exponential units, due to occasionally large gradients, we have to reduce the learning rate to 10^{-4} while the Sigmoid/Tanh unit remains stable for a learning rate of 10^{-2}. Other factors that affect the instability of training for exponential and quadratic Exp-RBMs are large momentum and small number of hidden units. Initialization of the weights could also play an important role, and sparse initialization (Sutskever et al., 2013; Martens, 2010) and regularization schemes (Goodfellow et al., 2013) could potentially improve the training of these models. In all experiments, we used uniformly random values in $[-.01, .01]$ for all unit types. In terms of training time, different Exp-RBMs that use the Gaussian noise and/or Sigmoid/Tanh units have similar computation time on both CPU and GPU.

Figure 6(top) shows the receptive fields for the street-view house numbers (SVHN) (Netzer et al., 2011) dataset. This dataset contains 600,000 images of digits in natural settings. Each image contains three RGB values for 32×32 pixels. Figure 6(bottom) shows few filters obtained from the jittered-cluttered NORB dataset (LeCun et al., 2004). NORB dataset contains 291,600 stereo $2 \times (108 \times 108)$ images of 50 toys under different lighting, angle and backgrounds. Here, we use a sub-sampled 48×48 variation, and report the features learned by two types of neurons. For learning from these two datasets, we increased the momentum to .95 and trained different models using up to 50 epochs.

4.2 Generating Samples

The USPS dataset (Hull, 1994) is relatively smaller dataset of 9,298, 16×16 digits. We binarized this data and used 90%, 5% and 5% of instances for training, validation and test respectively; see Figure 7 (first two rows) for instances from this dataset. We used Tanh activation function for the $16 \times 16 = 256$ *visible* units of the Exp-RBMs[6] and 500 hidden units of different types: 1) Tanh unit; 2) ReLU; 3) ReQU and 4)Sinh unit.

We then trained these models using CD with 10 Gibbs sampling steps. Our choice of CD rather than alternatives that are known to produce better generative models, such as Persistent CD (PCD; Tieleman, 2008), fast PCD (FPCD; Tieleman and Hinton, 2009) and (rates-FPCD; Breuleux et al., 2011) is due to practical reasons; these alternatives were unstable for some activation functions, while CD was always well-behaved. We ran CD for 10,000 epochs with three different learning rates $\{.05, .01, .001\}$ for each model.

[6] Tanh unit is similar to the sigmoid/Bernoulli unit, with the difference that it is (anti)symmetric $v_i \in \{-.5, +.5\}$.

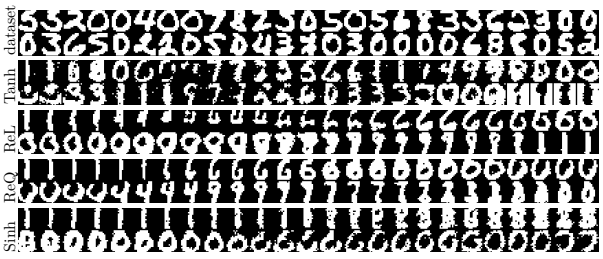

Figure 7: *Samples from the USPS dataset (first two rows) and few of the consecutive samples generated from different Exp-RBMs using rates-FPCD.*

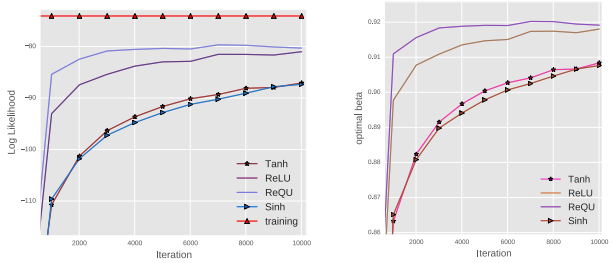

Figure 8: *Indirect Sampling Likelihood of the test data (left) and β^* for the density estimate (right) at different epochs (x-axis) for USPS dataset.*

Note that here, we did not use method of momentum and mini-batches in order to to minimize the number of hyper-parameters for our quantitative comparison. We used rates-FPCD [7] to generate $9298 \times \frac{90}{100}$ samples from each model – *i.e.*, the same number as the samples in the training set. We produce these sampled datasets every 1000 epochs. Figure 7 shows the samples generated by different models at their final epoch, for the "best choices" of sampling parameters and learning rate.

We then used these samples $D_{sample} = \{v^{(1)}, \ldots, v^{(N=9298)}\}$, from each model to estimate the Indirect Sampling Likelihood (ISL; Breuleux et al., 2011) of the validation set. For this, we built a non-parametric density estimate

$$\hat{p}(v;\beta) = \sum_{n=1}^{N} \prod_{j=1}^{256} \beta^{\mathbb{I}(v_j^{(n)}=v_j)}(1-\beta)^{\mathbb{I}(v_j^{(n)}\neq v_j)} \qquad (14)$$

and optimized the parameter $\beta \in (.5, 1)$ to maximize the likelihood of the validation set – that is $\beta^* = \arg_\beta \max \prod_{v \in D_{valid}} \hat{p}(v, \beta)$. Here, $\beta = .5$ defines a uniform distribution over all possible binary images, while for $\beta = 1$, only the training instances have a non-zero probability.

[7] We used 10 Gibbs sampling steps for each sample, zero decay of fast weights – as suggested in (Breuleux et al., 2011) – and three different fast rates $\{.01, .001, .0001\}$.

We then used the density estimate for β^* as well as the best rates-FPCD sampling parameter to evaluate the ISL of the *test set*. At this point, we have an estimate of the likelihood of test data for each hidden unit type, for every 1000 iteration of CD updates. The likelihood of the test data using the density estimate produced *directly from the training data*, gives us an upper-bound on the ISL of these models.

Figure 8 presents all these quantities: for each hidden unit type, we present the results for the learning rate that achieves the highest ISL. The figure shows the estimated log-likelihood (left) as well as β^* (right) as a function of the number of epochs. As the number of iterations increases, all models produce samples that are more representative (and closer to the training-set likelihood). This is also consistent with β^* values getting closer to $\beta^*_{training} = .93$, the optimal parameter for the training set.

In general, we found stochastic units defined using ReLU and Sigmoid/Tanh to be the most numerically stable. However, for this problem, ReQU learns the best model and even by increasing the CD steps to 25 and also increasing the epochs by a factor of two we could not produce similar results using Tanh units. This shows that a non-linearities outside the circle of well-known and commonly used exponential family, can sometimes produce more powerful generative models, even using an "approximate" sampling procedure.

Conclusion

This paper studies a subset of exponential family Harmoniums (EFH) with a single sufficient statistics for the purpose of learning generative models. The resulting family of distributions, Exp-RBM, gives a freedom of choice for the activation function of individual units, paralleling the freedom in discriminative training of neural networks. Moreover, it is possible to efficiently train arbitrary members of this family. For this, we introduced a principled and efficient approximate sampling procedure and demonstrated that various Exp-RBMs can learn useful generative models and filters.

References

Joachim H Ahrens and Ulrich Dieter. Computer methods for sampling from gamma, beta, poisson and bionomial distributions. *Computing*, 12(3):223–246, 1974.

Guozhong An. The effects of adding noise during backpropagation training on a generalization performance. *Neural Computation*, 8(3):643–674, 1996.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *JMLR*, 6:1705–1749, 2005.

Yoshua Bengio. Learning deep architectures for ai. *Foundations and trends in ML*, 2(1), 2009.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR CMMP*, 7(3):200–217, 1967.

Olivier Breuleux, Yoshua Bengio, and Pascal Vincent. Quickly generating representative samples from an rbm-derived process. *Neural Computation*, 23(8):2058–2073, 2011.

L. Devroye. *Non-uniform random variate generation*. Springer-Verlag, 1986. ISBN 9783540963059.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *JMLR*, 11:625–660, 2010.

Yoav Freund and David Haussler. *Unsupervised learning of distributions of binary vectors using two layer networks*. Computer Research Laboratory [University of California, Santa Cruz], 1994.

Peter V Gehler, Alex D Holub, and Max Welling. The rate adapting poisson model for information retrieval and object recognition. In *Proceedings of the 23rd international conference on Machine learning*, pages 337–344. ACM, 2006.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.

David P Helmbold, Jyrki Kivinen, and Manfred K Warmuth. Relative loss bounds for single neurons. *Neural Networks, IEEE Transactions on*, 10(6):1291–1304, 1999.

Geoffrey Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8): 1771–1800, 2002.

Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Jonathan J Hull. A database for handwritten text recognition research. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(5):550–554, 1994.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Donald E Knuth. Seminumerical algorithms. the art of computer programming, 1969.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *CVPR, 2004*, volume 2, pages II–97, 2004.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM, 2009.

Tim K Marks and Javier R Movellan. Diffusion networks, product of experts, and factor analysis. In *Proc. Int. Conf. on Independent Component Analysis*, pages 481–485, 2001.

James Martens. Deep learning via hessian-free optimization. In *ICML-10*, pages 735–742, 2010.

Peter McCullagh, John A Nelder, and P McCullagh. *Generalized linear models*, volume 2. Chapman and Hall London, 1989.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML-10*, pages 807–814, 2010.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop*, volume 2011, page 5. Granada, Spain, 2011.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Ruslan Salakhutdinov and Geoffrey E Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In *ICML-08*, pages 872–879. ACM, 2008.

Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. 1986.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML-13*, pages 1139–1147, 2013.

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. ACM, 2008.

Tijmen Tieleman and Geoffrey Hinton. Using fast weights to improve persistent contrastive divergence. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1033–1040. ACM, 2009.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML-08*, pages 1096–1103, 2008.

Max Welling, Simon Osindero, and Geoffrey E Hinton. Learning sparse topographic representations with products of student-t distributions. In *Advances in neural information processing systems*, pages 1359–1366, 2002.

Max Welling, Michal Rosen-Zvi, and Geoffrey E Hinton. Exponential family harmoniums with an application to information retrieval. In *NIPS*, pages 1481–1488, 2004.

Eunho Yang, Genevera Allen, Zhandong Liu, and Pradeep K Ravikumar. Graphical models via generalized linear models. In *NIPS*, pages 1358–1366, 2012.