A NEURAL NETWORK BASED ON FIRST PRINCIPLES

Paul M Baggenstoss

Fraunhofer FKIE, Fraunhoferstr 20,
53343 Wachtberg, Germany

ABSTRACT

In this paper, a Neural network is derived from first principles and assuming a maximum entropy (MaxEnt) prior and that a network layer starts with a linear transformation. A posterior distribution for the input data given the linear transformation output is derived. A new theorem is used to find a closed-form expression for the posterior and for its mean, which is a MaxEnt data reconstruction employing a special activation function. Combining layers results in an auto-encoder with conventional feed-forward analysis network and a type of linear Bayesian belief network in the reconstruction path. The new theorem unifies previous results relevant to some special cases. Methods for sampling the posterior are provided.

Index Terms**—** Neural networks, Maximum Entropy, Activation Functions, Projected Belief Network

1. INTRODUCTION

1.1. Motivation

Despite the brilliant success of deep networks, there has been insufficient attention payed to statistical optimality. Networks and their activation functions are generally selected empirically to learn general functions [1]. In generative networks, the activation functions revolve around approximating the expected value of generating distributions that are selected for tractability [2–4] or are empirically determined [5]. Despite the elegant mathematical formulations, restricted Boltzmann machines (RBMs) [6], and variation autoencoders [7], the models are also selected based on tractability or empirical performance. This paper seeks to derive the network structure and activation function from first principles by deducing the network structure from the *a posteriori* distribution of the visible data given the layer output.

1.2. Problem Statement

Given a high-dimensional input data $\mathbf{x} \in \mathbb{R}^N$, a lower-dimensional feature is computed by linear transformation, $\mathbf{z} = \mathbf{W}'\mathbf{x}$, where $\mathbf{z} \in \mathbb{R}^M$, and $M < N$. The goal is to derive an expression for $p(\mathbf{x}|\mathbf{z})$, from which the optimal reconstruction network and activation function can be inferred. The prior $p_0(\mathbf{x})$, is specified by the principle of maximum entropy (MaxEnt). The main idea is illustrated in Figure 1. The diagram shows two network layers, but we will focus on just the first layer for now. The input data $\mathbf{x}$ is operated on by a linear transformation: $\mathbf{z} = \mathbf{W}'\mathbf{x}$. A bias and activation function are applied prior to the next layer, but this is not relevant to analyzing the first layer. For now, the question is what can be inferred about $p(\mathbf{x})$ from $\mathbf{z}$, bypassing layer 2 (See "bypass" in Fig. 1). The remaining components in layer 1 are described below and layer 2 is explained in Section 4.

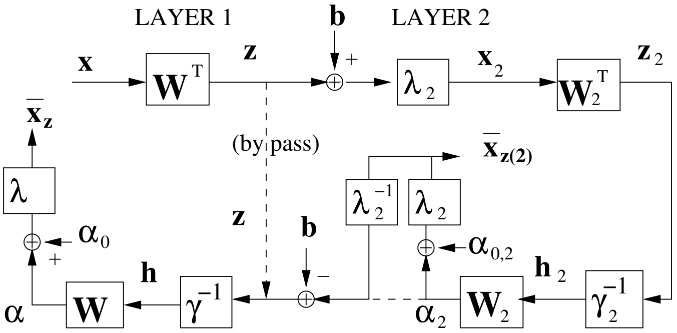

Fig. 1. Block diagram.

2. MATHEMATICAL APPROACH

2.1. Prior Distribution

To proceed, it is necessary to define the *a priori* distribution $p_0(\mathbf{x})$ that quantifies the expectation about $\mathbf{x}$ before feature $\mathbf{z}$ is measured. The principle of maximum entropy (MaxEnt) [8] proposes that the entropy of a distribution, given by $H\{p(\mathbf{x})\} = -\int_{\mathbf{x}} p(\mathbf{x}) \log(p(\mathbf{x}))\, d\mathbf{x}$ should be as high as possible subject to the known constraints. These distributions are generally of the exponential class [9]. Consider the following univariate exponential class of distributions:

$$\log p(x;\alpha) = \alpha x + bx^2 - \log Z(\alpha), \tag{1}$$

where the dependence on b has been removed from the notation because it is fixed by the choice of prior. Parameter α plays a special role because it controls the distribution mean.

This class encompasses Gaussian, exponential, and their truncated variants and includes all the MaxEnt distributions that will be necessary in this discussion. Let the expected value of distribution (1) be written as a function of α as

$$\lambda(\alpha) \triangleq \mathbb{E}\{x;\alpha\} = \int_x x\, p(x;\alpha)\, \mathrm{d}x. \tag{2}$$

In keeping with maximum entropy, $p_0(\mathbf{x})$ should be constructed from N independent univariate distributions (1) as follows

$$\log p_0(\mathbf{x}) = \sum_{i=1}^{N} \log p(x_i;\alpha_0). \tag{3}$$

This class includes independent and identically distributed (*iid*) Gaussian, exponential, and their truncated variants, and they have highest entropy among all multivariate distributions under constraints that will be proposed.

2.2. Manifold Distribution

Conditioned on knowing $\mathbf{z}$, $\mathbf{x}$ can only exist on the set

$$\mathcal{M}(\mathbf{z}) = \{\mathbf{x} : \mathbf{W}'\mathbf{x} = \mathbf{z}\}. \tag{4}$$

This is the set (a manifold) of all possible values of $\mathbf{x}$ that exactly reproduce the measured value $\mathbf{z}$. The posterior is therefore a manifold distribution

$$p(\mathbf{x}|\mathbf{z}) = p_0(\mathbf{x}) / \left(\int_{\mathbf{x}\in\mathcal{M}(\mathbf{z})} p_0(\mathbf{x})\, \mathrm{d}\mathbf{x} \right), \quad \mathbf{x} \in \mathcal{M}(\mathbf{z}), \tag{5}$$

which is $p_0(\mathbf{x})$ projected onto the manifold, then normalized so it integrates to 1. To draw samples from (5), samples are drawn from the manifold $\mathcal{M}(\mathbf{z})$ with probability proportional to the value of the prior distribution $p_0(\mathbf{x})$. It can be shown [10] that the denominator in (5) can be written

$$\int_{\mathbf{x}\in\mathcal{M}(\mathbf{z})} p_0(\mathbf{x})\, \mathrm{d}\mathbf{x} = p_0(\mathbf{z}),$$

which is the prior feature distribution, i.e. distribution of $\mathbf{z}$ under the assumption that $\mathbf{x} \sim p_0(\mathbf{x})$. Rewriting (5),

$$p(\mathbf{x}|\mathbf{z}) = \frac{p_0(\mathbf{x})}{p_0(\mathbf{z})}, \quad \mathbf{x} \in \mathcal{M}(\mathbf{z}). \tag{6}$$

Note that the denominator has a fixed value on the manifold, so the manifold distribution is shaped only by $p_0(\mathbf{x})$. This quantity is known in the method of PDF projection [10, 11].

2.3. Surrogate Density and Main Theoem

Despite the simple form of (6), it is not useful for sampling or determining the mean of $p(\mathbf{x}|\mathbf{z})$, and is not even a proper distribution, having infinite density on an infinitely thin manifold. To find a proper distribution that approximates (5), we use a surrogate density [12], which is a proper distribution that shares the properties of (5), which are (a) item probability mass concentrated on the manifold $\mathcal{M}(\mathbf{z})$, (b) mean $\bar{\mathbf{x}}_z \in \mathcal{M}(\mathbf{z})$ (because $\mathcal{M}(\mathbf{z})$ is convex), and (c) density on the manifold proportional to $p_0(\mathbf{x})$. The following theorem gives form to the surrogate density.

Theorem 1 *Let prior $p_0(\mathbf{x})$ be written as (3) with univariate densities $p(x;\alpha_0)$ of class (1) with mean $\lambda(\alpha_0)$. Then, the surrogate density for $\mu_z(\mathbf{x})$ can be written*

$$\log p(\mathbf{x};\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}) = \sum_{i=1}^{N} \log p(x_i;\alpha_0 + \alpha_i), \tag{7}$$

where $\boldsymbol{\alpha} = \mathbf{W}\mathbf{h}_z$, and $\mathbf{h}_z$ is the solution of

$$\mathbf{W}'\lambda\left(\boldsymbol{\alpha}_0 + \mathbf{W}\mathbf{h}\right) = \mathbf{z}. \tag{8}$$

Furthermore, the mean of the surrogate density is asymptotically (for large N) equal to the mean and centroid of the manifold $\mathcal{M}(\mathbf{z})$ and equals

$$\bar{\mathbf{x}}_z = \lambda(\boldsymbol{\alpha}_0 + \mathbf{W}\mathbf{h}_z). \tag{9}$$

Outline of Proof:. To show that solution $\mathbf{h}_z$ solving (8) exists, it is shown in Section 2.4, that (8) is the same as the saddlepoint (SP) equation for the SP expansion of $p_0(\mathbf{z})$. Since for the exponential family (1), the SP expansion exists over the entire range of $\mathbf{z}$ (see [13] appendix), it appears, and is supported by numerous experiments, that the solution exists whenever $\mathbf{z}$ is valid, i.e. whenever $\mathbf{z} = \mathbf{W}'\mathbf{x}$ for a sample $\mathbf{x}$ in the support of $p_0(\mathbf{x})$. Since $\bar{\mathbf{x}}_z = \lambda\left(\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}\right)$, it is clear that $\mathbf{W}'\bar{\mathbf{x}}_z = \mathbf{z}$, meeting property (b) for a surrogate density. Using (3),(1), the gradient of $\log p(\mathbf{x};\boldsymbol{\alpha}_0 + \boldsymbol{\alpha})$ with respect to $\mathbf{x}$ is $\left[\frac{\partial \log p(\mathbf{x};\boldsymbol{\alpha}_0+\boldsymbol{\alpha})}{\partial \mathbf{x}}\right] = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha} + 2b\mathbf{x}$. In order that (7) is proportional to $p_0(\mathbf{x})$ on the manifold, it is necessary that the component of this gradient in any direction parallel to the manifold (i.e. orthogonal to columns of $\mathbf{W}$) is the same as for the prior $p_0(\mathbf{x})$. This can be mathematically written $\mathbf{B}'\left[\boldsymbol{\alpha}_0 + \boldsymbol{\alpha} + 2b\mathbf{x}\right] = \mathbf{B}'\left[\boldsymbol{\alpha}_0 + 2b\mathbf{x}\right]$, for orthonormal matrix $\mathbf{B}$ spanning the linear subspace orthogonal to the columns of $\mathbf{W}$. It is then clear that $\boldsymbol{\alpha}$ must be fully orthogonal to $\mathbf{B}$, therefore of the form $\boldsymbol{\alpha} = \mathbf{W}\mathbf{h}$. This fulfills property (c) of the surrogate density. To fulfill property (a), it can be shown that the probability mass of the surrogate density indeed converges to the manifold for large N (see [12], Appendix A).

To simplify notation, we define the function $\gamma(\mathbf{h}) = \mathbf{W}'\lambda\left(\mathbf{W}\mathbf{h}\right) = \mathbf{z}$ and its inverse: $\mathbf{h}_z = \gamma^{-1}(\mathbf{z})$. The concept of $\gamma^{-1}(\mathbf{z})$ is illustrated in Figure 1. Feature $\mathbf{z}$, is converted to $\mathbf{h}$ through $\gamma^{-1}(\mathbf{z})$, then multiplied by $\mathbf{W}$ to raise the dimension back to N, and finally passed through activation function $\lambda(\ )$ to produce $\bar{\mathbf{x}}_z$. Optionally, it can be passed to the generating distributions $p(\mathbf{x};\boldsymbol{\alpha}_0+\boldsymbol{\alpha})$ for stochastic generation. According to the definition of $\gamma^{-1}(\ )$, it is clear that $\mathbf{W}'\bar{\mathbf{x}}_z=\mathbf{z}$, or in other words, the feature $\mathbf{z}$ is recovered exactly when $\bar{\mathbf{x}}_z$ is processed by the forward path. In this role, $\gamma^{-1}(\mathbf{z})$ acts as a non-linearity (but is not applied element-wise). Despite the iterative solution of $\gamma^{-1}(\mathbf{z})$, its derivatives are easly calculated from $\gamma(\mathbf{h})$, so are amenable to back-propagation training for optmizing the network parameters.

2.4. Properties of the Surrogate Density.

The surrogate density converges to the posterior $p(\mathbf{x}|\mathbf{z})$, and so the mean of the surrogate density approaches the mean of $p(\mathbf{x}|\mathbf{z})$. This convergence occurs quickly and low dimension as has been demonstrated in certain cases (see fig. 8 in [12]). The surrogate density mean $\bar{\mathbf{x}}_z$ given by (9) enjoys numerous properties. As conditional mean estimator, it has many well-known optimal properties [14]. Another special case of (9) corresponds to autoregressive spectral estimation, which can be generalized for conditioning on any linear function of the spectrum, such as MaxEnt inversion of MEL band features [12]. A special case of (9) is mathematically the same as classical maximum entropy image reconstruction [15, 16]. It is also not surprising, given form (6), that the surrogate density has a close relationship to $p_0(\mathbf{z})$. In fact, it can be shown that $\mathbf{h}_z$ is also the saddlepoint for the SP approximation to $p_0(\mathbf{z})$. The equivalence of the SP equation to (8) can be seen in ([17], equation (25), page 2245), which is the general SP equation for the distribution of the linear sum of independent random variables, and it is easily shown that $c'(b_n)=\lambda(\alpha_n)$. It can also be shown that $\mathbf{h}_z$ is the maximum likelihood estimate of $\mathbf{h}$ under the likelihood function (7) [13].

3. THREE CASES OF $\mathbb{X}$

In the following sections, the MaxEnt prior $p_0(\mathbf{x})$ is defined, and the distribution mean $\lambda(\alpha)$, the MaxEnt activation function, is provided for three cases of the range of $\mathbf{x}$, denoted by $\mathbb{X}$. In addition, it is explained how to sample from the manifold $\mathcal{M}(\mathbf{z})$. Note that manifold sampling is exact sampling of $p(\mathbf{x}|\mathbf{z})$, which differs from sampling from the surrogate density. However, experiments have demonstrated the almost perfect correspondence between the two distributions (e.g. Figures 8,10,11 in [18]).

3.1. Unit hypercube $\mathbb{U}^N$

In *unit hypercube*, denoted by $\mathbb{U}^N$, elements of $\mathbf{x}$ are in the range $[0,\ 1]$, the case for intensity images, or if $\mathbf{x}$ is the output of a sigmoid activation function. The uniform prior is the MaxEnt distribution in $[0,\ 1]$, $p_0(\mathbf{x})=1$, which is the trivial case of (1) with $\alpha_0=0$, $b=0$. Sampling from $\mathcal{M}(\mathbf{z})$ uniformly within $\mathbb{U}^N$ is done using a type of Monte Carlo Markov chain (MCMC) called hit-and-run [19], with modification for $\mathbb{U}^N$ as explained in detail in ([12], Sec. V, p. 2465). For the surrogate density, with $\alpha_i\neq 0$, a truncated exponential distribution (TED) is produced, $p(x;\alpha)=\frac{\alpha}{e^{\alpha}-1}\ e^{\alpha x},\quad 0>x>1$. The activation function is the TED nonlinearity [12, 20]

$$\lambda(\alpha)=\frac{e^{\alpha}}{e^{\alpha}-1}-\frac{1}{\alpha} \tag{10}$$

which resembles the sigmoid, $\lambda(\alpha)\simeq\sigma(\alpha/3)$. This problem has been studied in detail in ([12], Sec. V, p. 2465).

3.2. Positive Quadrant $\mathbb{P}^N$

We assume that elements of $\mathbf{x}$ are positive, so exist in the positive quadrant of $\mathbb{R}^N$, denoted by $\mathbb{P}^N$. This happens if $\mathbf{x}$ is the output of an previous network layer and a rectifying activation function was used, or if $\mathbf{x}$ is some kind of spectral or intensity data that is inherently positive. There is no proper MaxEnt distribution on the open interval $[0,\ \infty]$ without constraining the mean or variance, resulting in two solutions.

3.2.1. Positive Quadrant $\mathbb{P}^N$, constrained mean (exponential prior)

The mean can be constrained by including the statistic $t_1(\mathbf{x})=\sum_{i=1}^N\ x_i$ in the feature, then the mean constraint is implicit in the conditioning on $\mathbf{z}$. This is conveniently achieved by modifying $\mathbf{W}$ so that one of the columns is a constant. With constrained mean, the exponential distribution is MaxEnt on $[0,\ \infty]$, the case of (1) with $\alpha_0=1$, $b=0$. Then

$$p_0(\mathbf{x})=e^{\sum_{i=1}^N\ x_i}.$$

Inclusion of $t_1(\mathbf{x})$ in the feature also insures that $p_0(\mathbf{x})$ is constant on the manifold, because $p_0(\mathbf{x})=e^{t_1(\mathbf{x})}$, meaning that the manifold distribution is uniform. Methods for uniform sampling in a simplex or convex subspace based on Monte Carlo Markov chain (MCMC) have been developed [19] and is treated in detail in ([12], Sec. IV, p. 2460). The activation function for the exponential prior is quite strange, $\lambda(\alpha)=\frac{1}{\alpha}$. This case is mathematically the same as classical maximum entropy image reconstruction [15, 16].

3.2.2. Positive Quadrant $\mathbb{P}^N$, constrained variance (truncated Gaussian prior)

If we are willing to assume a fixed variance, the truncated Gaussian mean parameter 0 and variance parameter 1 (not the same as mean 0 and variance 1) provides the distribution with maximum entropy on $[0,\ \infty]$ [9]. This is the case of (1) with $\alpha_0=0$, $b=1$. This can also be written

$$p_0(\mathbf{x})=\prod_{i=1}^{N}2\mathcal{N}(x_i),\quad x_i>0,\forall i, \tag{11}$$

where $\mathcal{N}(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$. To sample $\mathcal{M}(\mathbf{z})$ with this prior, an MCMC method similar to the exponential case (given in ([12], Sec. IV, p. 2460) can be used. One starts with a sample $\mathbf{x}$ that lies on the manifold. We then choose a direction (a vector orthogonal to the column space of $\mathbf{W}$) to move, and determine the line segment on this line for which $\mathbf{x}$ remains positive. The distance moved along this direction is drawn from a truncated Gaussian distribution with these limits. The process repeats by selecing a new direction. The activation function is the mean of the truncated Gaussian:

$$\lambda(\alpha) = \alpha + \frac{\mathcal{N}(\alpha)}{\Phi(\alpha)} \tag{12}$$

which resembles softplus (see Figure 2).

3.3. Unconstrained $\mathbb{R}^N$

There is no proper MaxEnt distribution on the open interval $[-\infty,\ \infty]$ without constraining the variance. In many cases, data has been normalized, so we are justified in using a standard Normal prior, which is the maximum entropy distribution on $\mathbb{R}^N$ for known variance [9]. Because we assume $p_0(\mathbf{x})$ is standard normal (Gaussian with zero mean and variance 1), sampling from (6) is trivial. All samples on the manifold can be written $\mathbf{x} = \bar{\mathbf{x}}_z + \mathbf{B}\mathbf{u}$, where $\bar{\mathbf{x}}_z = \mathbf{W}\mathbf{h}_z$, where $\mathbf{h}_z = \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{z}$, and $\mathbf{B}$ is the same as in the proof of Theorem 1. To conform to the assumed prior distribution, $\mathbf{u}$ is a set of $(N-M)$ independent Gaussian random variables of zero mean and variance 1. The activation function is linear, $\lambda(x) = x$.

3.4. Summary and Remarks

A network layer structure has been inferred from the posterior $p(\mathbf{x}|\mathbf{z})$ and a MaxEnt prior $p_0(\mathbf{x}) = p(\mathbf{x};\boldsymbol{\alpha}_0)$ (see Fig. 1). To reconstruct $\mathbf{x}$, the linear transformation output $\mathbf{z} = \mathbf{W}'\mathbf{x}$ is passed through the dimension-preserving function γ^{-1}, which exists whenever $\mathbf{x}$ is in the support of $p_0(\mathbf{x})$, then multiplied by $\mathbf{W}$. A bias $\boldsymbol{\alpha}_0$ is added, then an activation function $\lambda(\ )$ that is derived from $p_0(\mathbf{x})$ is applied. The resulting data vector is the conditional mean $\mathbb{E}\{\mathbf{x}|\mathbf{z}\}$ and enjoys many optimality properties. The activation functions are:
Uniform prior: TED nonlinearity (10),
exponential prior: $\lambda(\alpha) = 1/\alpha$,
truncated Gaussian prior: T-G nonlinearity (12),
Gaussian prior: $\lambda(\alpha) = \alpha$.

These activation functions sometimes resemble commonly-used functions (see Fig. 2). Note that the T-G nonlinearity approaches the rectified linear unit (RELU) as $\sigma_0^2 \rightarrow 0$. Alternatively, $\lambda(\ )$ can be replaced by the generating distribution $p(\mathbf{x};\boldsymbol{\alpha}_0 + \boldsymbol{\alpha})$ for stochastic generation. For a single layer, this would produce an RBM with deterministic forward path.

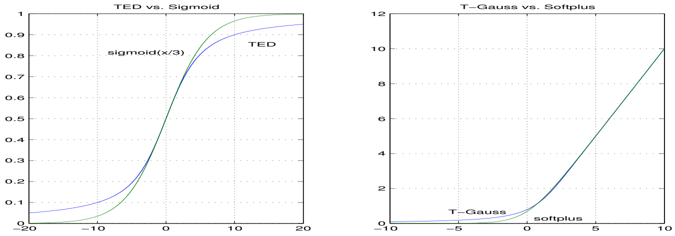

Fig. 2. Left: TED activation compared to Sigmoid. Right: T-Gauss activation compared to Softplus.

4. BUILDING A NETWORK

In Figure 1, a 2-layer network is created by adding another MaxEnt layer. The forward path (top) is a standard feed-forward network employing the MaxEnt activation functions. The data is first passed through a bias and activation function $\lambda_2(\ )$ before being presented to the second layer's linear transform. Note that after layer 2 reconstructs its input in the backward path $(\bar{\mathbf{x}}_{z(2)})$ the activation function $\lambda_2(\ )$ and bias must be inverted before being processed by γ^{-1}. However, because the forward activation function $\lambda_2(\ )$ is the same as the MaxEnt activation function for layer 2, then $\lambda_2(\ )$ cancels $\lambda_2^{-1}(\ )$, resulting in a simplified backward path! It is also worth noting that in the backward (reconstruction) path, stochastic generation using $p(\mathbf{x};\boldsymbol{\alpha}_0 + \boldsymbol{\alpha})$ can be used in place of activation functions to create stochastic networks.

The reverse path (bottom) consists of applying $\gamma^{-1}(\mathbf{z})$ (after removal of bias, if needed), followed by dimension-increasing transformation by the layer weight matrices (same matrix used in the forward path). This eliminates the need for separate reconstruction weights, and decreases network parameter count. This has been called a deterministic projected belief network [18, 21] and has been shown to significantly out-perform a standard auto-encoder of exactly the same specification [21].

5. CONCLUSIONS

In this paper, a new theorem has been presented that provides a closed-form asymptotic (large N) expression for the conditional mean $\bar{\mathbf{x}}_z = \mathbb{E}\{\mathbf{x}|\mathbf{z}\}$ given the output $\mathbf{z}$ of a dimension-reducing linear transformation. The computation of the conditional mean resembles a linear Bayesian belief network layer with special non-linear function preceding the linear transformation and special activation function. The theorem is generalized for a class of exponential family prior distributions and with support on $\mathbb{R}^N$, the positive quadrant $\mathbb{P}^N$, and the unit hypercube $\mathbb{U}^N$. Methods to sample the posterior $p(\mathbf{x}|\mathbf{z})$ are provided. The method can be extended to multiple layers to form genertive networks.

6. REFERENCES

[1] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. Cambridge, MA: MIT press, 2016.

[2] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 2010*, 2010.

[3] M. Zhou, "Softplus regressions and convex polytopes," *arXiv preprint arXiv:1608.06383*, 2016.

[4] S. Ravanbakhsh, B. Póczos, J. Schneider, D. Schuurmans, and R. Greiner, "Stochastic neural networks with monotonic activation functions," *Proceedings of the 19 th International Conference on Artificial Intelligence and Statistics (AISTATS), Cadiz, Spain*, 2016.

[5] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[6] M. Welling, M. Rosen-Zvi, and G. Hinton, "Exponential family harmoniums with an application to information retrieval," *Advances in neural information processing systems*, 2004.

[7] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[8] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proceedings of IEEE*, vol. 70, no. 9, pp. 939–952, 1982.

[9] J. N. Kapur, *Maximum Entropy Models in Science and Engineering*. Wiley (Eastern), 1993.

[10] P. M. Baggenstoss, "Maximum entropy PDF design using feature density constraints: Applications in signal processing," *IEEE Trans. Signal Processing*, vol. 63, June 2015.

[11] P. M. Baggenstoss, "The PDF projection theorem and the class-specific method," *IEEE Trans Signal Processing*, pp. 672–685, March 2003.

[12] P. M. Baggenstoss, "Uniform manifold sampling (UMS): Sampling the maximum entropy pdf," *IEEE Transactions on Signal Processing*, vol. 65, pp. 2455–2470, May 2017.

[13] O. Barndorff-Nielsen and D. R. Cox, "Edgeworth and saddle-point approximations with statistical applications," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 41, no. 3, pp. 279–299, 1979.

[14] S. Kay, *Fundamentals of Statisticsl Signal Processing, Estimation Theory*. Prentice Hall, Upper Saddle River, New Jersey, USA, 1993.

[15] S. J. Wernecke and L. R. D'Addario, "Maximum entropy image reconstruction," *IEEE Trans. Computers*, vol. C-26, no. 4, pp. 351–364, 1977.

[16] G. Wei and H. Zhen-Ya, "A new algorithm for maximum entropy image reconstruction," in *Proceedings of ICASSP-87*, vol. 12, pp. 595–597, April 1987.

[17] S. M. Kay, A. H. Nuttall, and P. M. Baggenstoss, "Multidimensional probability density function approximations for detection, classification, and model order selection," *IEEE Transactions on Signal Processing*, vol. 49, pp. 2240–2252, Oct 2001.

[18] P. M. Baggenstoss, "On the duality between belief networks and feed-forward neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2018.

[19] S. Kiatsupaibul, R. Smith, and Z. Zabinsky, "An analysis of a variation of hit-and-run for uniform sampling from general regions," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 21, no. 3, 2011.

[20] P. M. Baggenstoss, "Evaluating the RBM without integration using pdf projection," in *Proceedings of EUSIPCO 2017, Island of Kos, Greece*, Aug 2017.

[21] P. M. Baggenstoss, "Applications of projected belief networks (pbn)," in *Proceedings of EUSIPCO 2019*, (La Coruña, Spain), Sep 2019.