Evaluating the RBM Without Integration Using PDF Projection

Paul M. Baggenstoss
Fraunhofer FKIE, Fraunhoferstrasse 20
53343 Wachtberg, Germany
Email: p.m.baggenstoss@ieee.org

***Abstract*—In this paper, we apply PDF projection to arrive at an exact closed-form expression for the marginal distribution of the visible data of a restricted Boltzmann machine (RBM) without requiring integrating over the distribution of the hidden variables or needing to know the partition function. We express the visible data marginal as a projected PDF based on a set of sufficient statistics. When a Gaussian mixture model (GMM) is used to estimate the PDF of the sufficient statistics, then we arrive at a combined RBM/GMM model that serves as a general-purpose PDF estimator and Bayesian classifier. The approach extends recusively to compute the input distribution of a multi-layer network. We demonstrate the method using the MNIST handwritten character data set.**

I. Introduction

A. Background and Problem Statement

A restricted Boltzmann machine (RBM) is a type of recurrent neural network made popular by Hinton and collegues [1]. The RBM describes the joint distribution between a so-called *visible data* vector, $\mathbf{x}$ and *hidden data* vector $\mathbf{h}$. We assume throughout the paper that $\mathbf{h}$ is of lower dimension than $\mathbf{x}$. The interaction between $\mathbf{x}$ and $\mathbf{h}$ is defined by the conditional distributions $p(\mathbf{x}|\mathbf{h};\mathbf{\Lambda})$ and $p(\mathbf{h}|\mathbf{x};\mathbf{\Lambda})$, which are specified. One generates samples of the pair $\mathbf{x},\mathbf{h}$ by starting with some sample of either vector, say $\mathbf{h}$, then generate $\mathbf{x}$ according to $p(\mathbf{x}|\mathbf{h};\mathbf{\Lambda})$, then generate $\mathbf{h}$ according to $p(\mathbf{h}|\mathbf{x};\mathbf{\Lambda})$, and so on. After a number of these forward-backward iterations, (Gibbs sampling), the distribution converges to the theoretical distribution

$$p(\mathbf{x},\mathbf{h};\mathbf{\Lambda}) = \frac{e^{-E(\mathbf{x},\mathbf{h};\mathbf{\Lambda})}}{K(\mathbf{\Lambda})}, \tag{1}$$

where $\mathbf{\Lambda}$ is the set of parameters of the RBM, $E(\mathbf{x},\mathbf{h};\mathbf{\Lambda})$ is the *energy function* that describes the interaction between data vectors $\mathbf{x}$ and $\mathbf{h}$, and $K(\mathbf{\Lambda})$ is the normalization constant, called *partition function*

$$K(\mathbf{\Lambda}) = \int_{\mathbf{h}}\int_{\mathbf{x}} e^{-E(\mathbf{x},\mathbf{h};\mathbf{\Lambda})}\mathrm{d}\mathbf{h}\,\mathrm{d}\mathbf{x}, \tag{2}$$

that can in general cannot be found in closed form.

The conditional distributions are shown in diagramatic form in Figure 1. In the "backward" direction, we generate the $\mathbf{x}\in\mathcal{R}^N$ from $\mathbf{h}\in\mathcal{R}^M$ by generating the intermediate variable $\boldsymbol{\alpha}\in\mathcal{R}^N$ using the affine transformation $\boldsymbol{\alpha} = \mathbf{W}\mathbf{h}+\mathbf{b}$, then generating the elements of $\mathbf{x}$ independently according to the node generation distribution $x_i \sim p(x_i;\alpha_i), \quad 1\le i\le N$. We will discuss typical node generation distributions below. To generate $\mathbf{h}$ from $\mathbf{x}$ ("forward" path), we generate the intermediate variable $\boldsymbol{\beta}\in\mathcal{R}^M$ by the affine transformation $\boldsymbol{\beta} = \left[\Sigma^{-1}\mathbf{W}\right]'\mathbf{x}+\mathbf{c}$, where Σ is the diagonal matrix of variances $\sigma_1^2,\sigma_2^2\ldots\sigma_N^2$ (assumed to be the identiy matrix for non-Gaussian RBMs), then generate the individual elements of $\mathbf{h}$ independently according to the node generation distribution $h_i \sim p(h_j;\beta_j), \quad 1\le j\le M$. Note that $\mathbf{W}$ is the same matrix as was used in the backward path. The RBM is normally trained using *contrastive divergence* (CD) so that the marginal distribution

$$p(\mathbf{x};\mathbf{\Lambda}) = \int_{\mathbf{h}} p(\mathbf{x},\mathbf{h};\mathbf{\Lambda})\,\mathrm{d}\mathbf{h} = \int_{\mathbf{h}} \frac{e^{-E(\mathbf{x},\mathbf{h};\mathbf{\Lambda})}}{K(\mathbf{\Lambda})}\,\mathrm{d}\mathbf{h} \tag{3}$$

approximates the distribution of some training data $\mathbf{x}$ [1]. The problem of the RBM that we seek to solve is the intractability of (3), which must be approximated [2]. Even if we recast (3) using Bayes theorem

$$p(\mathbf{x};\mathbf{\Lambda}) = \int_{\mathbf{h}} p(\mathbf{x}|\mathbf{h};\mathbf{\Lambda})\;p(\mathbf{h};\mathbf{\Lambda})\,\mathrm{d}\mathbf{h}, \tag{4}$$

this requires knowing $p(\mathbf{h};\mathbf{\Lambda})$, a kind of chicken and egg problem.

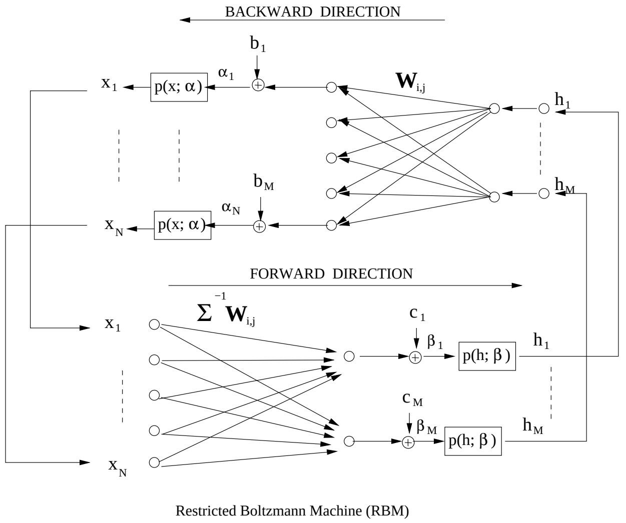

Fig. 1. Restricted Boltzmann Machine (RBM). For non-Gaussian input RBMs, $\Sigma = \mathbf{I}$.

The RBM is very popular because the deterministic approximation of its forward path (replacing the node generation distributions with their expected value) can be used to initialize a feed-forward network layer (stacked RBMs). Theoretically, the RBM can be seen as an implementation of Factor analysis with an infinite number of mixture components and can also be trained using the efficient contrastive divergence (CD) algorithm. But, generating from the model requires Gibbs-sampling, the partition function (2) cannot in general be computed, and determining the marginal distribution (3) requires integration. These factors make the RBM ill-suited as a general-purpose generative model, such as for example the Gaussian mixture model (GMM). The purpose of this paper is to lift these disadvantages.

B. Types of RBMs

The type of RBM is determined by the input and output node generation distributions $p(x;\alpha)$ and $p(h;\beta)$, respectively. Generally, in the literature, we see the discrete Bernoulli RBM created by the sigmoid node generation function $p(x_i = 1;\alpha_i) = \frac{1}{1+e^{-\alpha_i}}$, or, for continuous-valued data, the Gaussian RBM created by the Gaussian node generation function $p(x_i;\alpha_i,\sigma_i^2) = (2\pi\sigma_i^2)^{-1/2} e^{-\frac{(x_i-\alpha_i)^2}{2\sigma_i^2}}$. We find it also useful to use the truncated exponential distribution (TED) node generation function for continuous-valued data limited to the range $[0,\ 1]$, $p(x;\alpha) = C(\alpha)\ e^{\alpha x}$, where $C(\alpha) = \left(\frac{\alpha}{e^{\alpha}-1}\right)$. The mean of this density is $\lambda(\alpha) = \frac{e^{\alpha}}{e^{\alpha}-1} - \frac{1}{\alpha}$. Due to space limitations, we will discuss only the Gauss-Gauss and TED-TED RBMs. Figure 1 can be used for reference.

TED-TED RBM. The TED-TED RBM has the energy function $E(\mathbf{x},\mathbf{h};\mathbf{\Lambda}) = -\mathbf{x}'\mathbf{b} - \mathbf{c}'\mathbf{h} - \mathbf{x}'\mathbf{W}\mathbf{h}$, which can be verified using the procedure of Welling et al [3], and is the same as the energy function for the Bernoulli-Bernoulli RBM. The conditional distribution of $\mathbf{x}$ given $\mathbf{h}$ is $p(\mathbf{x}|\mathbf{h};\mathbf{\Lambda}) = \prod_{i=1}^{N} C(\alpha_i)e^{\alpha_i x_i} = C(\boldsymbol{\alpha})e^{\boldsymbol{\alpha}'\mathbf{x}}$, where $\boldsymbol{\alpha} = \mathbf{b} + \mathbf{W}\mathbf{h}$. The conditional distribution of $\mathbf{h}$ given $\mathbf{x}$ is $p(\mathbf{h}|\mathbf{x};\mathbf{\Lambda}) = \prod_{j=1}^{M} C(\beta_j)e^{\beta_j h_j} = C(\boldsymbol{\beta})e^{\boldsymbol{\beta}'\mathbf{x}}$, where $\boldsymbol{\beta} = \mathbf{c} + \mathbf{W}'\mathbf{x}$. The RBM parameter set is $\mathbf{\Lambda} = \{\mathbf{b},\mathbf{c},\mathbf{W}\}$.

Gauss-Gauss RBM. The formulation for the Gauss-Gauss RBMs is lifted from [4]. The energy function is: $E(\mathbf{x},\mathbf{h};\mathbf{\Lambda}) = \frac{1}{2}(\mathbf{x}-\mathbf{b})'\Sigma^{-1}(\mathbf{x}-\mathbf{b}) - \mathbf{x}'\Sigma^{-1}\mathbf{W}\mathbf{h} + \frac{1}{2}(\mathbf{h}-\mathbf{c})'(\mathbf{h}-\mathbf{c})$. The conditional distribution of $\mathbf{x}$ is $p(\mathbf{x}|\mathbf{h};\mathbf{\Lambda}) = (2\pi)^{-N/2}|\Sigma|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\alpha})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\alpha})}$, where $\boldsymbol{\alpha} = \mathbf{b}+\mathbf{W}\mathbf{h}$. The conditional distribution of $\mathbf{h}$ is $p(\mathbf{h}|\mathbf{x};\mathbf{\Lambda}) = (2\pi)^{-M/2}e^{-\frac{1}{2}(\mathbf{h}-\boldsymbol{\beta})'(\mathbf{h}-\boldsymbol{\beta})}$, where $\boldsymbol{\beta} = \mathbf{c} + \mathbf{W}'\Sigma^{-1}\mathbf{x}$. The RBM parameter set is $\mathbf{\Lambda} = \{\mathbf{b},\mathbf{c},\mathbf{W},\Sigma\}$.

II. Mathematical Results

A. Main Result

Let $\mathbf{z} = T(\mathbf{x})$ be some deterministic mapping from $\mathcal{X} \in \mathcal{R}^N$ to $\mathcal{Z} \in \mathcal{R}^M$, where $M < N$. Let $p(\mathbf{x};H_0)$ be some known reference distribution on $\mathcal{X}$, and let $p(\mathbf{z};H_0)$ be the corresponding distribution imposed on $\mathcal{Z}$ by transformation T. Let $p(\mathbf{x};\Lambda)$ be some unknown distribution that we want to approximate. The projected PDF is given by

$$\hat{p}(\mathbf{x};\mathbf{\Lambda}) = \frac{p(\mathbf{x};H_0)}{p(\mathbf{z};H_0)}\hat{p}(\mathbf{z};\mathbf{\Lambda}), \tag{5}$$

where $\hat{p}(\mathbf{z};\mathbf{\Lambda})$ is an estimate of the distribution of $\mathbf{z}$ when $\mathbf{x}$ is drawn from $p(\mathbf{x};\Lambda)$. It can be shown [5] that (5) is a PDF (integrates to 1), and is consistent with $\hat{p}(\mathbf{z};\mathbf{\Lambda})$, meaning that samples drawn from $\hat{p}(\mathbf{x};\mathbf{\Lambda})$ and passed through transformation T will have exactly distribution $\hat{p}(\mathbf{z};\mathbf{\Lambda})$. It is therefore a reasonable estimate of $p(\mathbf{x};\Lambda)$ and under certain conditions [5], (5) is the maximum entropy (MaxEnt) PDF among all PDFs consistent with $\hat{p}(\mathbf{z};\mathbf{\Lambda})$. But, more importantly, (5) becomes *exact*, meaning that $\hat{p}(\mathbf{x};\mathbf{\Lambda}) \rightarrow p(\mathbf{x};\mathbf{\Lambda})$, under the following two conditions [6]: (a) that $\hat{p}(\mathbf{z};\mathbf{\Lambda}) = p(\mathbf{z};\mathbf{\Lambda})$, and (b), that $\mathbf{z}$ is a sufficient statistic for the likelihood ratio $L(\mathbf{x}) = p(\mathbf{x};\mathbf{\Lambda})/p(\mathbf{x};H_0)$. This happens if we can write $L(\mathbf{x}) = h(T(\mathbf{x}))$, for some function h [7]. Thus, if $\mathbf{z}$ is a sufficient statistic for $L(\mathbf{x})$ and if we know $p(\mathbf{z};\mathbf{\Lambda})$, then

$$\frac{p(\mathbf{x};H_0)}{p(\mathbf{z};H_0)}p(\mathbf{z};\mathbf{\Lambda}) = p(\mathbf{x};\mathbf{\Lambda}), \tag{6}$$

which is an exact expression. We can exploit this to get an exact expression for (4) without integration.

B. Sufficient Statistics for various RBM types

TED-TED RBM. Substituting the energy function $E(\mathbf{x},\mathbf{h};\mathbf{\Lambda})$ for the TED-TED RBM into (3) , we get

$$p(\mathbf{x};\mathbf{\Lambda}) = \frac{1}{K(\mathbf{\Lambda})}e^{\mathbf{x}'\mathbf{b}}\int_{\mathbf{h}} e^{[\mathbf{W}'\mathbf{x}+\mathbf{c}]'\mathbf{h}}\,\mathrm{d}\mathbf{h} \tag{7}$$

For data limited to $0 < x_i < 1$, the uniform reference hypothesis $p(\mathbf{x};H_0) = 1$, which assumes $\{\mathbf{x}_i\}$ are *iid* uniform-distributed, provides the MaxEnt projected PDF [5]. Therefore, the likelihood ratio $L(\mathbf{x}) = p(\mathbf{x};\mathbf{\Lambda})/p(\mathbf{x};H_0)$ is the same as (7). Now, it is obvious that (7) may be written as a function of the sufficient statistic $\mathbf{z} = T(\mathbf{x}) = [\mathbf{W}'\mathbf{x},\ \mathbf{b}'\mathbf{x}] = \mathbf{W}_b'\mathbf{x}$, where $\mathbf{W}_b = [\mathbf{W},\ \mathbf{b}]$. Figure 2 shows the feed-forward network layer for extracting the sufficient statistic for the TED-TED RBM, where $\Sigma = \mathbf{I}$. We have added the bias $\mathbf{c}$ and subsequent non-linearity $\hat{\mathbf{h}} = f(\mathbf{z}+\mathbf{c})$, where $f(\ )$ is the activation function that produces the expected value of the node generation distribution. Note that $\hat{\mathbf{h}}$ is a deterministic approximation to the RBM's random hidden variable $\mathbf{h}$ (conditional mean), in the same way as is done in stacked RBMs [1]. Note that in Figure 2, the auxilliary statistic $\mathbf{b}'\mathbf{x}$ is treated like other nodes by adding a suitable bias c_0, and passing it through a non-linearity.

Gauss-Gauss RBM. For the Gauss-Gauss RBM, we use the reference hypothesis

$$p(\mathbf{x};H_0) = (2\pi)^{-N/2}|\Sigma|^{-1/2}\ e^{-\frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x}}, \tag{8}$$

which assumes $\{\mathbf{x}_i\}$ are *iid* Gaussian zero-mean with variance σ_i^2. Substituting the energy function for the Gauss-Gauss RBM into (3) , and dividing by the reference hypothesis, the resulting likelihood ratio becomes

$$L(\mathbf{x}) = \frac{p(\mathbf{x};\mathbf{\Lambda})}{p(\mathbf{x};H_0)} = K'\int_{\mathbf{h}} e^{\mathbf{x}'\Sigma^{-1}[\mathbf{W}\mathbf{h}+\mathbf{b}]}\ e^{-\frac{1}{2}(\mathbf{h}-\mathbf{c})'(\mathbf{h}-\mathbf{c})}\,\mathrm{d}\mathbf{h}, \tag{9}$$

where K' is a constant independent of $\mathbf{x}$. It is easy to show that $L(\mathbf{x})$ can be written in terms of the sufficient statistic $\mathbf{z} = T(\mathbf{x}) = \mathbf{W}_b'\Sigma^{-1}\mathbf{x}$, where $\mathbf{W}_b$ is defined above. This is illustrated in Figure 2 which shows the feed-forward network layer for extracting the sufficient statistic, and additional processing to produce $\hat{\mathbf{h}}$, which approximates the RBM's hidden variable.

C. Discussion and Implemetation

We have suggested replacing the integral expression (4) by (6), so let's compare the two approaches. Both methods assume we know the distribution of the lower-dimensional variable ($\mathbf{h}$ or $\mathbf{z}$). But, $\mathbf{h}$ is distributed jointly with $\mathbf{x}$ - it requires Gibbs sampling in order to generate samples of $\mathbf{h}$, whereas $\mathbf{z}$ is deterministically dependent upon $\mathbf{x}$, so we only need to convert the visible data $\mathbf{x}$ to $\mathbf{z}$ using a fixed transformation. Even more important is the fact that (4) requires integration, whereas (6) is evaluated at just one value pair ($\mathbf{x}$,$\mathbf{z}$).

The disadvantage of (6) is that evaluating $p(\mathbf{z}; H_0)$ is sometimes non-trivial and requires knowing the exact distribution of the sufficient statistic $\mathbf{z}$ when $\mathbf{x}$ is drawn from $p(\mathbf{x}; H_0)$. This problem has been studied in detail and solutions exist for a wide range of features [8], [9], [10], [11], [5]. For the TED-TED RBM which uses the uniform reference distribution, $p(\mathbf{z}; H_0)$ is derived in Section VII. For the Gauss-Gauss RBM, $p(\mathbf{x}; H_0)$ is given by (8), so $p(\mathbf{z}; H_0)$ is Gaussian with covariance $\mathbf{W}_b' \Sigma^{-1} \mathbf{W}_b$. As illustrated in Figure 2, it is convenient to first define the "whitened" statistic $\mathbf{y} = \Sigma^{-1/2}\mathbf{x}$, whose distribution under H_0 is the canonical Gaussian $p(\mathbf{y}; H_0) = (2\pi)^{-N/2}\ e^{-\frac{1}{2}\mathbf{y}'\mathbf{y}}$. Then, the sufficient statistic is re-written in terms of $\mathbf{y}$: $\mathbf{z} = \mathbf{W}_b' \Sigma^{-1/2}\mathbf{y}$.

D. Energy Statistic (ES)

In Figure 2 (bottom), we have shown optional "energy statistic" (ES). An ES is an optional scalar statistic, but insures the maximum entropy property of the PDF projection [11], [5]. As explained in [5], when the input data is limited to $0 \leq x_i \leq 1$, (i.e. TED-TED RBM) no energy statistic is strictly required. For the Gauss-Gauss RBM, we could use the second-order energy statistic $e = \log(\mathbf{y}'\mathbf{y})$ [5], but it is better to make the energy statistic approximately independent of $\hat{\mathbf{h}}$ by subtracting the conditional mean first: $e = \log\left[(\mathbf{y} - \hat{\mathbf{y}}_h)'(\mathbf{y} - \hat{\mathbf{y}}_h)\right]$, where $\hat{\mathbf{y}}_h = \Sigma^{-1/2}(\mathbf{W}\hat{\mathbf{h}} + \mathbf{b})$. This orthogonalized energy statistic is a good feature because it is an indicator of fit to the PDF (4). The energy statistic is appended to $\mathbf{z}$. Use of the energy statistic changes $p(\mathbf{z}; H_0)$, which must be analyzed in the method given in [5] Section III.B. The energy statistic can also be used for re-synthesis of the input data according to the MaxEnt method [11], [5].

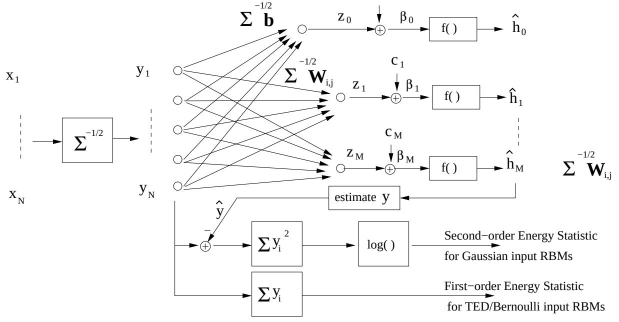

Fig. 2. Feed-forward NN that extracts the sufficient statistic for an RBM. For TED or Bernoulli input RBM's, $\Sigma = \mathbf{I}$ and $\mathbf{y} = \mathbf{x}$.

E. RBM Analysis

We now reduce (2) to a single integral of dimension M, so that we can evaluate (3) numerically.

Analysis of TED-TED RBM. For the TED-TED RBM, $K(\mathbf{\Lambda})$ can be written $K(\mathbf{\Lambda}) = \int_{\mathbf{h}}\int_{\mathbf{x}} e^{\left\{\mathbf{x}'\boldsymbol{\alpha}+\mathbf{c}'\mathbf{h}\right\}} \mathrm{d}\mathbf{h}\,\mathrm{d}\mathbf{x}$, where $\boldsymbol{\alpha} = \mathbf{b} + \mathbf{W}\mathbf{h}$. But, since the integrand with $\mathbf{h}$ fixed is a TED distribution on $\mathbf{x}$, it integrates to the inverse of the TED constant $C(\alpha)$, so we can reduce this to

$$K(\mathbf{\Lambda}) = \int_{\mathbf{h}} \frac{e^{\mathbf{c}'\mathbf{h}}}{C(\boldsymbol{\alpha})}\ \mathrm{d}\mathbf{h}, \tag{10}$$

which can be integrated numerically. Once we have $K(\mathbf{\Lambda})$, we can compute the marginal of $\mathbf{x}$ for any sample $\mathbf{x}$ as follows. We rewrite (3) as $p(\mathbf{x}; \mathbf{\Lambda}) = \frac{e^{\mathbf{x}'\mathbf{b}}}{K(\mathbf{\Lambda})}\int_{\mathbf{h}} e^{\boldsymbol{\beta}'\mathbf{h}}\mathrm{d}\mathbf{h}$, where $\boldsymbol{\beta} = \mathbf{c} + \mathbf{W}'\mathbf{x}$, and where it can be seen that for a fixed $\mathbf{x}$ it has a TED distribution over $\mathbf{h}$, which integrates to the inverse of the TED constant $C(\boldsymbol{\beta})$. Therefore,

$$p(\mathbf{x}; \mathbf{\Lambda}) = \frac{1}{K(\mathbf{\Lambda})}\frac{e^{\mathbf{x}'\mathbf{b}}}{C(\boldsymbol{\beta})}. \tag{11}$$

Analysis of Gauss-Gauss RBM. The Gauss-Gauss RBM is special because the partition function can be determined analytically. Note that the Gauss-Gauss energy function can be written

$$\frac{1}{2}\left[\begin{array}{c}(\mathbf{x}-\tilde{\mathbf{b}})\\(\mathbf{h}-\tilde{\mathbf{c}})\end{array}\right]'\left[\begin{array}{cc}\Sigma^{-1} & -\Sigma^{-1}\mathbf{W}\\ -\mathbf{W}'\Sigma^{-1} & \mathbf{I}\end{array}\right]\left[\begin{array}{c}(\mathbf{x}-\tilde{\mathbf{b}})\\(\mathbf{h}-\tilde{\mathbf{c}})\end{array}\right] + C, \tag{12}$$

where C is independent of $\mathbf{x}$ and $\mathbf{h}$, and the adapted mean vectors $\tilde{\mathbf{b}}$ and $\tilde{\mathbf{c}}$ are related to $\mathbf{b}$ and $\mathbf{c}$ by:

$$\left[\begin{array}{c}\tilde{\mathbf{b}}\\ \tilde{\mathbf{c}}\end{array}\right] = \left[\begin{array}{cc}\mathbf{I} & -\mathbf{W}\\ -\mathbf{W}'\Sigma^{-1} & \mathbf{I}\end{array}\right]^{-1}\left[\begin{array}{c}\mathbf{b}\\ \mathbf{c}\end{array}\right].$$

It is therefore clear that the Gauss-Gauss RBM is a multivariate Gaussian distribution and the marginal of $\mathbf{x}$ is also Gaussian with mean $\tilde{\mathbf{b}}$ and covariance equal to the upper $N \times N$ diagonal block of the inverse of the kernel matrix in quadratic form (12).

III. Numerical Validation

We now compare (6) with (4) evaluated numerically.

A. TED-TED RBM

We trained a TED-TED RBM of output dimension $M = 3$ using contrastive divergence on data from the MNIST handwritten character corpus. We created 1000 independent samples of $p(\mathbf{x}; \mathbf{\Lambda})$, by Gibbs-sampling the RBM. Starting with independent uniformly-distributed random data, we performed 500 forward-backward passes of the RBM. The ending sample of $\mathbf{x}$ was then regarded as a sample of $p(\mathbf{x}; \mathbf{\Lambda})$. We computed $K(\mathbf{\Lambda})$ using (10) using MATLAB's *integral3()* function. For PDF projection, we implemented (6) using uniform reference hypothesis, evaluated $p(\mathbf{z}; H_0)$ using the method in Section VII, and approximated $p(\mathbf{z}; \mathbf{\Lambda})$ from the 1000 samples of RBM "training data" using a Gaussian mixture. We used no energy statistic in this experiment. For validation, we selected 400 samples of the 1000 RBM training samples. For each sample we computed the marginal using (11), plotting it on the X-axis, then using (6), plotting it on the Y-axis in Figure 3 (left).

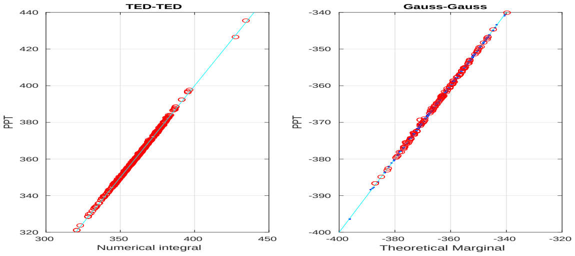

Fig. 3. X-axis: analytic or numerically-determined marginal, Y-axis PDF projection. Left: TED-TED, Right: Gauss-Gauss. Circles: Gibbs-sampled data. Dots: Direct-sampled data.

B. Gauss-Gauss RBM

We repeated the same experiment using a Gauss-Gauss RBM. We expanded the MNIST data to the real line by applying an inverse sigmoid function to the pixel values, so that the pixel data was approximately zero mean and most of the data was in the range [-10,10]. We then trained a Gauss-Gauss RBM using contrastive divergence on some data of character "3". To provide training data of $p(\mathbf{x};\mathbf{\Lambda})$, we used two approaches: (a) Gibbs-sampling the RBM, (b) directly sampling the RBM's theoretical marginal (see Section II-E). For the PDF projection approach, we evaluated (6), using no energy statistic. Under reference hypothesis (8), $p(\mathbf{z};H_0)$ is Gaussian with covariance $\mathbf{W}_b'\Sigma^{-1}\mathbf{W}_b$. We trained a Gaussian mixture to approximate $p(\mathbf{z};\mathbf{\Lambda})$. We compared (6) with the theoretical Gaussian marginal for a 400 sample subset of the 1000 RBM training samples. This is plotted in Figure 3 (right) for the Gibbs-sampled data (dots) and direct-sampled data (circles). No significant difference can be seen.

IV. Applications

A general-purpose PDF estimator for high-dimensional data can be created by training the RBM on $\mathbf{x}$, extracting the sufficient statistic $\mathbf{z}$, and approximating $p(\mathbf{z})$ with a GMM, then applying equation (6). Alternatively, the method can be applied recursively to a multi-stage network by approximating $p(\mathbf{z})$ by another RBM (stacked RBM). This results in the exact PDF of the full network[1].

V. Classification Experiment

A. MNIST Data description

The MNIST OCR data corpus [12] set consists of ten handwritten digits 0-9 divided into two sub-corpora: the training sub-corpus with about 6000 training samples of each digit, and the testing sub-corpus with about 1000 testing samples of each digit. We downsampled the 28×28 images 2:1 to 14×14, giving an input data dimension of $N = 14\times14 = 196$. For classification, we selected the digits "3", "8", and "9", and gathered all the data together into a pool of 7000 samples per digit. Then, in each of ten random trials, we chose an independent training set of 200 samples from the pool, and kept the rest for testing. In each trial, the total number of testing samples was 20324. For GMM and Gauss-input RBMs, we used expanded data (see Section III-B).

B. Results

We trained an RBM separately on the data of each class m using CD to obtain the RBM parameter $\mathbf{\Lambda}_m$. Then, we extracted the corresponding sufficient statistic, denoted by $\mathbf{z}_m$, whos PDF $\hat{p}(\mathbf{z}_m;H_m)$ was modeled by a GMM. We then applied formula (6) as the likelihood function, with $p(\mathbf{z}|\mathbf{\Lambda})$ replaced by $\hat{p}(\mathbf{z}_m;H_m)$. The classification performance was averaged over ten independent trials, each time selecting 200 training samples at random, and using the rest as test data. In Figure 4, we show the performance as a function of M for various RBM types and for GMM only. The GMM-only classifier modeled the input data $\mathbf{x}$ directly. On the plot, we see the error in percent for GMM alone as a function of the number of mixture components (plotted on the X axis as variable M) for diagonal covariance matrices and full covariance matrices. We also see the performance of RBM/GMM for various RBM types with the sufficient statistic modeled only using 2 GMM modes. Plotted are TED-TED, Gauss-TED, and Gauss-Gauss RBMs. For the Gauss-Gauss RBM, we show results with and without the energy statistic. It is clear that the energy statistic helps. In Figure 5, we see the performance as a function of standard deviation floor γ for each Generative model on Figure 4, at the optimum value of M. To implement "standard deviation floor", we added the value $\gamma^2\hat{\sigma}_i^2$ to the i-th entry of main diagonal of the GMM's covariance matrices, where $\hat{\sigma}_i^2$ is the sample variance of data dimension i.

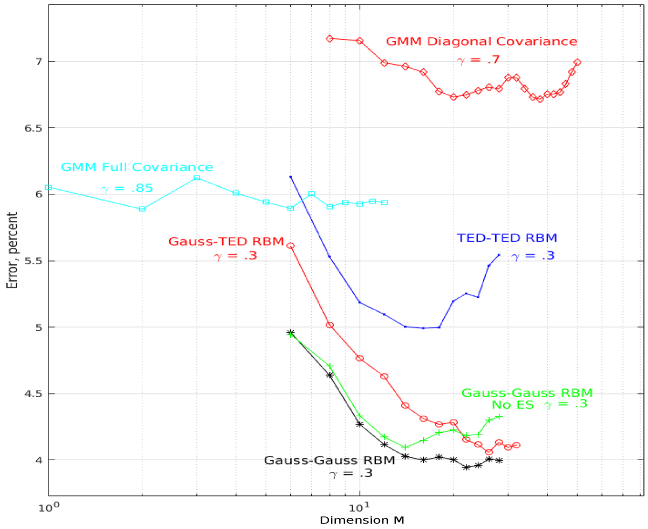

Fig. 4. Performance as a function of M.

VI. Conclusions

In this paper, we have presented a means to compute the visible data marginal of the RBM without integration. We did this by finding the sufficient statistic of the RBM's theoretical density, then using PDF projection to compute the PDF of the visible data. We compared the method with the theoretical

[1] The only approximation would be the PDF of the output of the final layer.

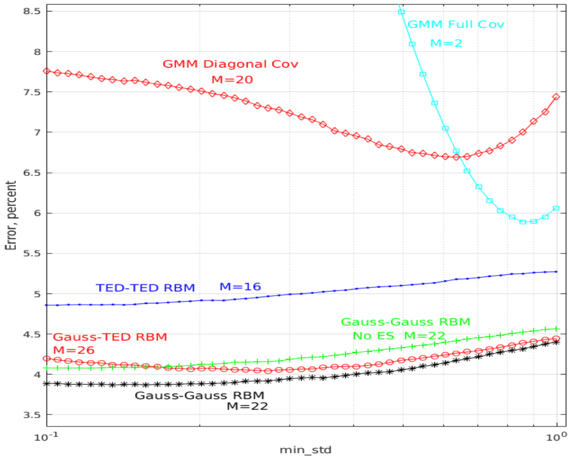

Fig. 5. Performance as a function of standard deviation floor γ.

marginal obtained by analysis or numerical integration, and they matched. Then we applied the method to classifying handwritten characters and it performed significantly better than GMM alone.

VII. Appendix

A. The Saddle point approximation.

Let $\mathbf{z} = T(\mathbf{x})$ be some dimension-reducing transformation and assume we need to compute $p(\mathbf{z}; H_0)$ where $p(\mathbf{x}; H_0)$ is known. The types of transformations T and reference hypotheses H_0 for which analytic expressions are available are limited. For a broader class of statistics, the exact moment generating function (MGF) is often known and can be inverted using the saddle-point approximation (SPA) [8]. It should be kept in mind that although the SPA is an "approximation", the accuracy is not degraded in the PDF tails and the result even converges to the true value as N becomes large and the feature approaches Gaussian by Central Limit Theorem.

Let feature $\mathbf{z} \in \mathcal{R}^M$. The moment-generating function for $p(\mathbf{z})$ is given by $g_z(\boldsymbol{\lambda}) = \mathcal{E}\left\{\exp\left(\boldsymbol{\lambda}'\mathbf{z}\right)\right\}$ for M-dimensional Laplace transform variable $\boldsymbol{\lambda}$. Then, $\mathbf{z}$ is obtained by the inverse Laplace transform: $p_z(\mathbf{z}) = \frac{1}{(j2\pi)^M}\int_C \exp\left(-\boldsymbol{\lambda}'\mathbf{z}\right) \;\; g_z(\boldsymbol{\lambda}) \; \mathrm{d}\boldsymbol{\lambda}$, where $j = \sqrt{-1}$. The contour C is parallel to the imaginary axis in each of the M dimensions of $\boldsymbol{\lambda}$. The joint cumulant generating function (CGF) is $c_z(\boldsymbol{\lambda}) = \log g_z(\boldsymbol{\lambda})$. For a specified $\mathbf{z}$, the Saddle-point is that real point $\hat{\boldsymbol{\lambda}}(\mathbf{z})$ where all M partial derivatives satisfy $\left.\frac{\partial c_z(\boldsymbol{\lambda})}{\partial \lambda_i}\right|_{\hat{\boldsymbol{\lambda}}} = z_i, \quad 1 \leq i \leq M$. The Saddlepoint may be found iteratively using the recursion $\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n + \mathbf{C}_z^{-1}(\boldsymbol{\lambda}_n) \;\left(\mathbf{z} - \mathbf{c}_z^{\lambda}(\boldsymbol{\lambda}_n)\right)$, where $\mathbf{c}_z^{\lambda}(\boldsymbol{\lambda})$ is the gradient vector of $c_z(\boldsymbol{\lambda})$ w/r to $\boldsymbol{\lambda}$, and $\mathbf{C}_z(\lambda)$ is the $M \times M$ matrix of second partial derivatives $\mathbf{C}_z(\lambda) \triangleq \left[\frac{\partial^2 c_z(\lambda)}{\partial \lambda_l \partial \lambda_m}\right]$.

Once the saddle-point $\hat{\boldsymbol{\lambda}}(\mathbf{z})$ is found, the saddle-point approximation is given by

$$p_z(\mathbf{z}) \simeq \frac{\exp\left\{c_z(\hat{\boldsymbol{\lambda}}) - \hat{\boldsymbol{\lambda}}'\mathbf{z}\right\}}{(2\pi)^{M/2}\left[\det\left(\mathbf{C}_z(\hat{\boldsymbol{\lambda}})\right)\right]^{1/2}}, \quad \hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}(\mathbf{z}). \tag{13}$$

B. Saddle point approximation for linear function of independent uniform RV

Let $\mathbf{x}$ be a set of N independent uniform-distributed RV in $[0, 1]$ Let $\mathbf{A}$ be an N-by-M full-rank matrix and let $\mathbf{z}$ be the $M \times 1$ feature vector $\mathbf{z} = \mathbf{A}'\mathbf{x}$.

When $\mathbf{x} \in \mathcal{R}^N$ is uniformly distributed, the CGF is given by

$$c_z(\boldsymbol{\lambda}) = \sum_{n=1}^{N} \log\left(\frac{\exp\left(\sum_{i=1}^{M} \lambda_i A_{n,i}\right) - 1}{\sum_{i=1}^{M} \lambda_i A_{n,i}}\right).$$

For conciseness, define $w_n = \sum_{i=1}^{M} \lambda_i A_{n,i}$. Then, $c_z(\boldsymbol{\lambda}) = \sum_{n=1}^{N} \log\left(\frac{e^{w_n} - 1}{w_n}\right)$, $\frac{\partial c_z(\boldsymbol{\lambda})}{\partial \lambda_i} = \sum_{n=1}^{N}\left(\frac{e^{w_n}}{e^{w_n} - 1} - \frac{1}{w_n}\right) A_{n,i}$, and

$$\frac{\partial^2 c_z(\boldsymbol{\lambda})}{\partial \lambda_i \partial \lambda_j} = \sum_{n=1}^{N}\left(\frac{e^{w_n}}{e^{w_n} - 1} - \frac{(e^{w_n})^2}{(e^{w_n} - 1)^2} + \frac{1}{w_n^2}\right) A_{n,i} A_{n,j}.$$

From these expressions, we may find the saddle-point approximation (13).

References

[1] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," in *Neural Computation 2006*, 2006.

[2] R. Salakhutdinov and I. Murray, "On the quantitative analysis of deep belief networks," *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008.

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

[4] K. Cho, A. Ilin, and T. Raiko, "Improved learning of gaussian-bernoulli restricted boltzmann machines," in *ICANN 2011*, pp. 10–17, 2011.

[5] P. M. Baggenstoss, "Uniform manifold sampling (UMS): Sampling the maximum entropy pdf," *IEEE Transactions on Signal Processing*, Jan. 2017.

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

[7] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. London: Chapman and Hall, 1974.

[8] S. M. Kay, A. H. Nuttall, and P. M. Baggenstoss, "Multidimensional probability density function approximation for detection, classification and model order selection," *IEEE Trans. Signal Processing*, pp. 2240–2252, Oct 2001.

[9] A. H. Nuttall and P. M. Baggenstoss, "The joint distributions for two useful classes of statistics with applications to classification and hypothesis testing," *NUWC Report, DTIC (http://dtic.mil/dtic/) document number ADA477219*, 2002.

[10] P. M. Baggenstoss, "The class-specific classifier: Avoiding the curse of dimensionality (tutorial)," *IEEE Aerospace and Electronic Systems Magazine, special Tutorial addendum*, vol. 19, pp. 37–52, January 2004.

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

[12] J. LeCun, "Mnist database," 2014.