The Annals of Statistics
2003, Vol. 0, No. 0, 1–5

DISCUSSION

BY OLIVER B. DOWNS[1]

Microsoft Research

The nonnegative Boltzmann machine (NNBM) is a recurrent neural network model that can describe multimodal nonnegative data. Application of maximum likelihood estimation to this model gives a learning rule that is analogous to that of the binary Boltzmann machine. While the model itself is analytically intractable an efficient stochastic version of the learning rule can be obtained using *reflective slice sampling*, since the slice boundaries can be determined analytically from the model. We compare this with the use of advanced mean field theory to learn a generative model for face image data.

1. Introduction. The multivariate Gaussian is the most elementary distribution used to model generic data. It represents the *maximum entropy* distribution under the constraint that the mean and covariance matrix of the distribution match that of the data. For the case of binary data, the maximum entropy distribution that matches the first- and second-order statistics of the data is given by the Boltzmann machine [Hinton and Sejnowski (1983)].

The Boltzmann machine can be generalized to continuous and nonnegative variables [Downs, MacKay and Lee (2000)]. In this case, the maximum entropy distribution for nonnegative data with known first- and second-order statistics is described by the nonnegative Boltzmann distribution (NNBD),

$$P(x) = \begin{cases} \dfrac{1}{Z}\exp[-E(x)], & \text{if } x_i \geq 0\ \forall i, \\ 0, & \text{if any } x_i < 0, \end{cases} \tag{1}$$

where the energy function $E(x)$ and normalization constant Z are

$$E(x) = \beta x^T Ax - b^T x, \tag{2}$$

$$Z = \int_{x\geq 0} dx\ \exp[-E(x)]. \tag{3}$$

The properties of the NNBD differ quite substantially from the Gaussian distribution which would arise for continuous, unbounded data. In particular, the presence of the nonnegativity constraints allows the distribution to have multiple modes, confined to the rectifying axes, since A can be nonpositive definite. Such a distribution would be poorly modeled by a single Gaussian. Here, we describe how a multimodal NNBD can be learned from nonnegative data.

[1] Supported in part by Microsoft Research Bell Laboratories (Lucent Technologies) and Princeton University Procter Fellowship.

2. Maximum likelihood. The learning rule for the NNBM can be derived by maximizing the log probability of the observed data under (1). Given a set of nonnegative vectors $\{x^{\mu}\}$, where $\mu = 1 \cdots M$ indexes the different examples, the log probability is

$$L = \frac{1}{M}\sum_{\mu=1}^{M} \log P(x^{\mu}) = -\frac{1}{M}\sum_{\mu=1}^{M} E(x^{\mu}) - \log Z. \tag{4}$$

Taking the derivatives of (4) with respect to the parameters A and b gives

$$\frac{\partial L}{\partial A_{ij}} = \langle x_i x_j \rangle_{\mathrm{f}} - \langle x_i x_j \rangle_{\mathrm{c}}, \tag{5}$$

$$\frac{\partial L}{\partial b_i} = \langle x_i \rangle_{\mathrm{c}} - \langle x_i \rangle_{\mathrm{f}}, \tag{6}$$

where the subscript "c" denotes a "clamped" average over the data, and the subscript "f" denotes a "free" average over the NNBM distribution

$$\langle f(x) \rangle_{\mathrm{c}} = \frac{1}{M}\sum_{\mu=1}^{M} f(x^{\mu}), \qquad \langle f(x) \rangle_{\mathrm{f}} = \int_{x \geq 0} dx\, P(x) f(x). \tag{7}$$

These derivatives are used to define a gradient ascent learning rule for the NNBM. The contrast between the clamped and free covariance matrix is used to update the interactions A, while the difference between the clamped and free means is used to update the local biases b.

3. Mean-field theories. It is possible to obtain approximations to the statistics of the model and to learn approximate parameters from data using concepts from mean-field theory. In (2) the (inverse) temperature parameter, β, controls the influence of correlations in the model. It is possible to approximate expectations under the model distribution in the limit that these correlations are assumed weak, the "high temperature" limit, at which $\beta = 0$. Since $E(x)$ is linear in this limit, integrals over the NNBD are tractable; we then accomodate small nonzero β by Taylor expansion about $\beta = 0$ [Downs (2001)].

To first order in β this approach returns the "naive" mean-field approximation, equivalent to approximating the NNBD with a factorized product of one-dimensional exponential distributions, with means matching that of the data. This replaces the "free" correlations in (5) with

$$\langle x_i x_j \rangle_{\mathrm{f}} = (1 + \delta_{ij}) \langle x_i \rangle_{\mathrm{c}} \langle x_j \rangle_{\mathrm{c}}. \tag{8}$$

Then expanding second-order in β we obtain a TAP-like [Kappen and Rodriguez (1998)] correction to this approximation,

$$\Delta \langle x_i x_j \rangle_{\mathrm{f}} = -\frac{\beta}{2}\sum_{kl} \alpha_{ijkl} A_{ij} A_{kl} \langle x_i \rangle_{\mathrm{c}} \langle x_j \rangle_{\mathrm{c}} \langle x_k \rangle_{\mathrm{c}} \langle x_l \rangle_{\mathrm{c}}. \tag{9}$$

4. Monte Carlo sampling. A direct approach to calculating the "free" averages in (5) and (6) is to numerically approximate them. This can be accomplished by generating samples from the NNBD using a Markov chain Monte Carlo method. Such methods employ an iterative stochastic dynamics whose equilibrium distribution converges to that of the desired distribution [MacKay (1998)].

Gibbs sampling from such a distribution requires repeated evaluation of the error function $\operatorname{erf}(z)$, and hence is prone to numerical error or high computational cost. The method of reflective slice sampling [Neal (1997)] circumvents this.

The basic idea of the reflective slice sampling algorithm is shown in Figure 1. Given a sample point x_i, a random $y \in [0, P^*(x_i)]$ [where $P^*(x)$ is the unnormalized density] is first chosen uniformly. Then a slice S is defined as the connected set of points $(x \in S \mid P^*(x) \geq y)$ including x_i, and the new point $x_{i+1} \in S$ is chosen randomly from this slice.

In order to efficiently choose a new point within a particular multidimensional slice, reflective "billiard ball" dynamics are used. A random initial direction in the slice, p is chosen, and the new point is evolved by traveling a predefined distance from the current point along the path specularly reflecting from the boundaries of the slice.

For the NNBM, solving the boundary points along a particular direction in a given slice is quite simple, since it only involves solving the roots of a quadratic equation,

$$\beta x^T A x - b^T x - \log y = 0, \tag{10}$$

along the line defined by the momentum vector, p, bounding the points by the

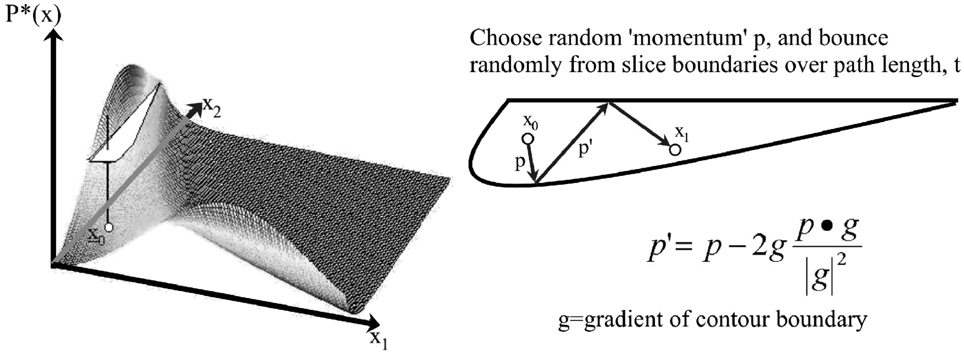
P*(x)
x2
x0
x1
Choose random 'momentum' p, and bounce randomly from slice boundaries over path length, t
x0
p
p'
x1
p'= p − 2g (p • g)/|g|^2
g=gradient of contour boundary

FIG. 1. *Reflective slice sampling in two dimensions. Given the current sample point, x_0, a height $y \in [0, P^*(x_0)]$ is randomly chosen. This defines a slice $(x \in S \mid P^*(x) \geq y)$ in which a new x_1 is chosen, using billiard-ball dynamics with specular reflections from the interior boundaries of the slice.*

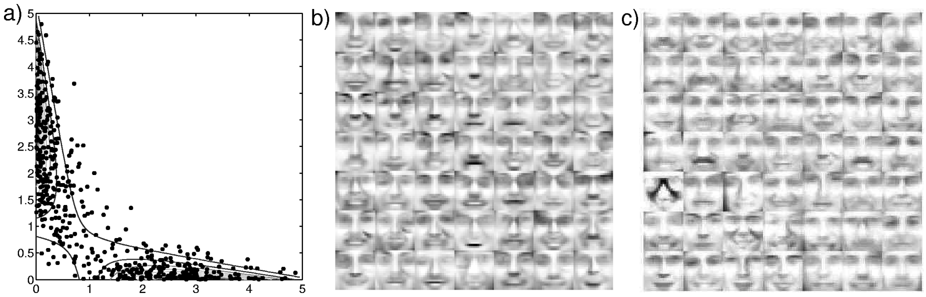

FIG. 2. (a) *Contours of the two-dimensional "competitive" NNBD overlaid by* 500 *reflective slice samples from the distribution.* (b) *Prototype face images generated from a mean-field NNBM.* (c) *Prototype face images generated from an NNBM learned via reflective slice sampling.*

rectifying axes,

$$x_{\text{boundary}} = \left[\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}\right]^+, \tag{11}$$

where $^+$ denotes rectification. The reflected momentum direction is a simple function of the incident momentum and the gradient of the slice boundary, g,

$$p' = p - 2g\frac{p \cdot g}{|g|^2} \qquad \text{with } g = 2\beta Ax - b. \tag{12}$$

The distribution of x_n for large n can be shown to converge to the desired density $P(x)$. Figure 2a demonstrates this for a two-dimensional NNBD.

5. Generative model for faces. We have used the NNBM to learn a generative model for images of human faces. The NNBM is used to model the correlations in the coefficients of the nonnegative matrix factorization (NMF) of the face images [Lee and Seung (1999)]. NMF reduces the dimensionality of nonnegative data by decomposing the face images into parts correponding to eyes, noses, ears, etc. Since the different parts are coactivated in reconstructing a face, the activations of these parts contain significant correlations that need to be captured by a generative model. Here we briefly demonstrate how the NNBM is able to learn these correlations, comparing the mean-field and reflective slice sampling approaches described above.

Sampling from the learned NNBD stochastically generates coefficients which can graphically be displayed as face images. Figure 2b and 2c compare prototype face images from NNBDs learned using advanced mean-field theory and reflective slice sampling, respectively. We suggest that the latter produces the more plausible prototypes.

6. Discussion. Here we have demonstrated the application of reflective slice sampling to NNBM learning. Since the NNBM learning rule is generally intractable, approximations are required to enable efficient learning. The use of reflective slice sampling is seen to be a reasonable alternative to mean-field approaches for learning the model, and can be implemented efficiently since slice boundaries and reflections can be determined analytically.

Extensions to the present work include incorporating hidden units into the recurrent network, which would imply modeling higher-order statistics of the data.

Acknowledgments. The author is especially grateful to David MacKay and Daniel Lee for their collaboration on the original ideas for this work, to Hagai Attias for comments on this manuscript and to John Hopfield and David Heckerman for their on-going support.

REFERENCES

DOWNS, O. B. (2001). High-temperature expansions for learning models of nonnegative data. *Adv. in Neural Inform. Processing Systems* **13** 465–471.

DOWNS, O. B., MACKAY, D. J. C. and LEE, D. D. (2000). The nonnegative Boltzmann machine. *Adv. in Neural Inform. Processing Systems* **12** 428–434.

HINTON, G. E. and SEJNOWSKI, T. J. (1983). Optimal perceptual learning. In *IEEE Conference on Computer Vision and Pattern Recognition* 448–453. Washington, DC.

KAPPEN, H. J. and RODRIGUEZ, F. B. (1998). Efficient learning in Boltzmann machines using linear response theory. *Neural Comput.* **10** 1137–1156.

LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature* **401** 788–791.

MACKAY, D. J. C. (1998). Introduction to Monte Carlo methods. In *Learning in Graphical Models* 175–204. Kluwer, Dordrecht.

NEAL, R. M. (1997). Markov chain Monte Carlo methods based on “slicing” the density function. Technical Report 9722, Dept. Statistics, Univ. Toronto.

MICROSOFT RESEARCH
ONE MICROSOFT WAY
REDMOND, WASHINGTON 98052
E-MAIL: t-odowns@microsoft.com