Products of Experts

Max Welling
Donald Bren School of Information and Computer Science
University of California Irvine
Irvine, CA 92697-3425 USA

August 8, 2007

1 The Product of Experts Model

A **Product of Experts** model (PoE) [5] combines a number of individual component models (the experts) by taking their product and normalizing the result. Each expert is defined as a possibly unnormalized probabilistic model $f(x)$ over its input space.

$$P(x|\{\theta_j\}) = \frac{1}{Z}\prod_{j=1}^{M} f_j(x|\theta_j) \tag{1}$$

with

$$Z = \int \mathrm{d}x \prod_{j=1}^{M} f_j(x|\theta_j) \tag{2}$$

PoEs stand in contrast to *Mixture Models* which combine expert models additively,

$$P(x|\{\theta_j\}) = \sum_{j=1}^{M} \alpha_j p_j(x|\theta_j) \tag{3}$$

where each component model $p(x)$ is normalized over x and

$$\sum_{j=1}^{M} \alpha_j = 1 \tag{4}$$

Note that *Mixture of Expert Models* are usually associated with conditional models where the experts are of the form $p(y|x)$ and the mixture coefficients (known as gating functions) may depend on x as well, $\alpha(x)$. Conditional PoEs may be defined as well.

One can qualitatively understand the difference between mixtures and products by observing that a mixture distribution can have high probability for event x when only a single expert assigns high probability to that event. In contrast, a product can only have high probability for an event x when all experts assign high probability to that event. Hence, metaphorically speaking, a single expert in a mixture has the power to pass a bill while a single expert in a product has the power to veto it.

Put another way, each component in a product represents a soft *constraint*, while each expert in a mixture represents a soft template or prototype. For an event to be likely under a product model, all constraints must

be (approximately) satisfied, while an event is likely under a mixture model if it (approximately) matches with a single template. Hence, the joint probability distribution under a mixture is sculpted by *adding* clumps of probability to the existing distribution and renormalizing, while the joint probability of a product is sculpted by cutting away bits and pieces from the distribution and renormalizing. This essential difference has been observed to result in much sharper boundaries especially for high dimensional input spaces [5]

2 Training a Product of Experts

Given data $\{x_n\},\ n = 1..N$ with $x \in R^d$, one can use the log-likelihood as an objective function to train a PoE,

$$L(\{\theta_j\}|\{x_n\}) = \sum_{n=1}^{N}\sum_{j=1}^{M} \log f_j(x_n|\theta_j) - N \log Z \tag{5}$$

Denoting the gradient of the objective w.r.t. θ_j with $\nabla_j L$ one can compute the following gradient,

$$\nabla_j L = \sum_{n=1}^{N} \nabla_j \log f_j(x_n) - N\langle \nabla_j \log f_j(x)\rangle_{P(x)} \tag{6}$$

where $\langle\cdot\rangle_{P(x)}$ denotes taking the average w.r.t. $P(x)$. Learning is achieved by changing the parameters incrementally according to the following update rule,

$$\theta_j \rightarrow \theta_j + \eta \nabla_j L \tag{7}$$

where η represents the learning rate. Learning efficiency can usually be improved by using a stochastic approximation of the full gradient based on a single data-case or a small mini-batch of data-cases and by including a momentum term in the gradient update.

The first term of the gradient (Eqn.6) can be interpreted as *increasing* the probability of expert i on the dataset. The second term on the other hand can be interpreted as *decreasing* the probability of expert i in regions of input space where the model assigns high probability. When these terms balance, learning has converged to a local maximum of the log-likelihood.

2.1 Contrastive Divergence learning

The simplicity of the gradient in eqn.6 is deceptive: it requires the evaluation of an intractable average over $P(x)$. For most interesting models this average requires approximate methods like MCMC sampling to approximate it. But MCMC sampling is computationally expensive and results in high-variance estimates of the required averages. A cheaper, lower-variance alternative was proposed by [6] under the name *contrastive divergence* (CD). The idea is to run N samplers in parallel, one for each data-case in the (mini-)batch. These samplers must be initialized at the respective data-cases and will move towards equilibrium by applying of the MCMC kernel. However, even after a few steps of sampling there is sufficient signal in the population of samples to change the parameters. A surrogate learning rule can be derived by replacing the log-likelihood with a new objective: the contrastive divergence $KL(P_{data}||P_{model}) - KL(P_{data}||P_k)$ where P_{data} is the empirical distribution $P_{data} = \frac{1}{N}\sum_n \delta(x - x_n)$, P_{model} is the current estimate of the model distribution and P_k is the distribution based on k steps of sampling. Taking gradients and ignoring a term which is usually very small, the new learning rule is almost identical to the one based on the log-likelihood, but replacing,

$$\langle \log f_j(x)\rangle_{P(x)} \approx \frac{1}{N}\sum_n f_j(x_n^k) \tag{8}$$

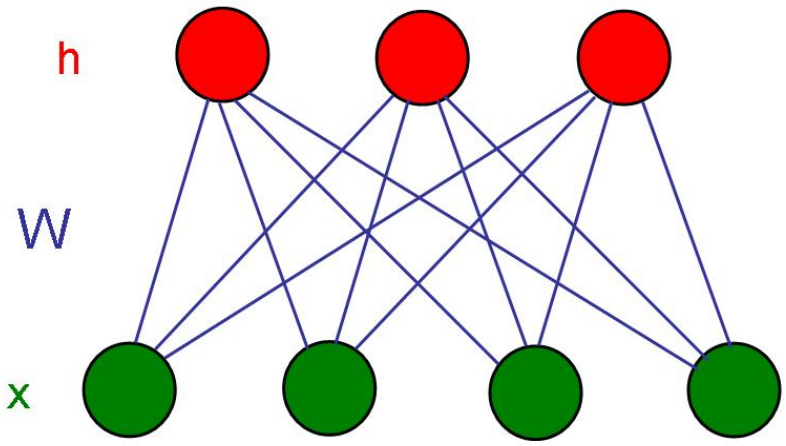

Figure 1: Graphical model representation of the *restricted Boltzmann machine* (RBM) and its generalization the *exponential family harmonium* (EFH). Top layer nodes represent hidden variables while filled bottom layer nodes represent observed variables. The architecture is that of a bipartite Markov Random Field (MRF).

where x_n^k is the sample obtained from MCMC sampler n after k steps of sampling.

One can view this approximation as trading variance for bias. Thus, at convergence, we do not expect that the estimates of the parameters are equal to those of maximum likelihood learning, but will be slightly biased. To correct this, one can increase k close to convergence [2].

3 Restricted Boltzmann Machines and Exponential Family Harmoniums

Perhaps the simplest PoE is given by a restricted Boltzmann machine (see figure 1). In this model there are two layers of binary (0/1) variables where the bottom layer is observed while the top layer remains unobserved or hidden. The joint probability distribution over hidden and observed variables is given as,

$$P(x,h) = \frac{1}{Z}\exp\left(\sum_i \alpha_i x_i + \sum_j \beta_j h_j + \sum_{ij} W_{ij} x_i h_j\right) \tag{9}$$

where the undirected edges in the graphical model in figure 1 are representing $\{W_{ij}\}$. The bias terms are parameterized by $\{\alpha_i, \beta_j\}$. Marginalizing over $\{h_j\}$ the PoE structure becomes evident,

$$P(x) = \frac{1}{\tilde{Z}} \prod_i \exp(\alpha_i x_i) \prod_j \left(1 + \exp(\beta_j + \sum_i W_{ij} x_i)\right) \tag{10}$$

where elements in the first product represent single-variable experts and elements in the second product represent constraints between the input variables.

The conditional Bernoulli expert distributions can be generalized to distributions in the exponential family . The resulting joint model is called an **exponential family harmonium** (EFH) [14]. The joint distribution can be obtained by replacing $x_i \rightarrow f_i(x_i)$ and $h_j \rightarrow g_j(h_j)$. where $f(\cdot)$ and $g(\cdot)$ are the features for the corresponding exponential family distribution.

The special bipartite structure of the RBM and EFH results in a very efficient Gibbs sampler that alternates between sampling all hidden variables independently given values for the observed variables and vice versa sampling all visible variables independently given values for the hidden variables. The efficient Gibbs sampler directly translates into an efficient *contrastive divergence learning algorithm* (see section 2.1).

4 Relation to Independent and Extreme Components Analysis

Noiseless **Independent Components Analysis** (ICA) [3] with an equal number of input dimensions and source distributions can be written as a PoE model as follows,

$$P(x|\{w_j\}) = |\det(W)| \prod_{j=1}^{M} p_j \left(\sum_i w_{ji} x_i\right) \tag{11}$$

where W is the matrix with elements w_{ji}. Note that in this case each expert is defined on a 1 dimensional projection of the input space. Unless W is rank deficient, the product is a well defined distribution over the entire input space.

Choosing the heavy tailed Student-T distributions as the experts one obtains the general form of the "Products of Student-T" distribution (PoT) [13]. The PoT can be represented with the help of auxiliary variables (taking the role of hidden variables) as follows,

$$P(x,h) = \frac{1}{Z} \prod_{j=1}^{M} \exp\left(-h_j[1 + \frac{1}{2}(\sum_i w_{ji} x_i)^2] + (1-\alpha_j)\log h_j\right) \tag{12}$$

where $P(x|h)$ is a full covariance Gaussian distribution and $P(h|x)$ a product of Gamma distributions.

The PoT becomes different from ICA if one chooses the number of experts to be larger than the number of input dimensions (a.k.a. an *over-complete representation*). In this case *marginal independence* between the hidden variables is lost, but *conditional independence* between the hidden variables is retained. Over-complete variants of ICA that retain marginal independence have also been proposed [7]. This model has conditional dependencies between the hidden variables known as *explaining away* which makes inference difficult. In contrast, for the over-complete PoT model inference over the hidden variables given observations is trivial due to the absence of such conditional dependencies [11].

Instead of the non-Gaussian experts used for ICA, one can also choose an *under-complete* ($M < d$) set of 1-dimensional *Gaussian* experts, i.e. $p_j\left(\sum_i w_{ji} x_i\right)$ with $p_j(\cdot)$ Gaussian. In addition, an isotropic Gaussian noise model with variance σ^2 can be chosen to fill up the remaining dimensions of input space.

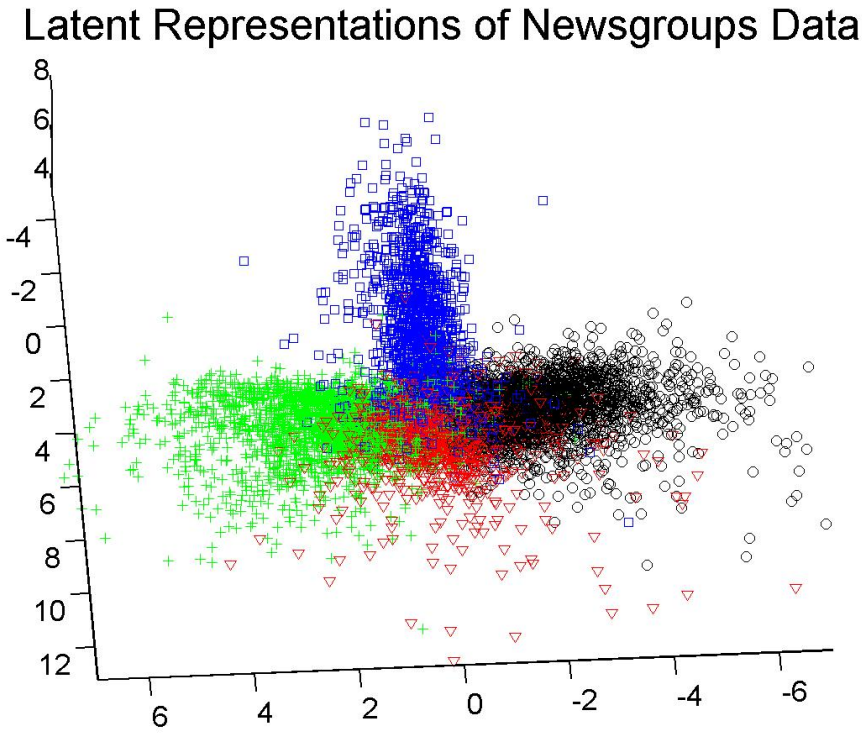

Figure 2: Latent representation for a exponential family harmonium fit to text data. Each point represents a document while its color codes for the hand labelled topic of that document. Each dimension in latent space corresponds to the "activity" of a latent variable. The EFH did not see the labels but managed to organize the documents according to their topics.

Optimizing the log-likelihood over W, σ corresponds to choosing the optimal combination of principal and minor components in the spectrum of the sample covariance matrix. The probabilistic model, known as "eXtreme Components Analysis" (XCA), is described in [12].

5 Applications of PoEs

Variants of PoEs have been applied under different names to various data-domains: for example the *rate-coded RBM* to face recognition [10], the *dual wing harmonium* to video-track data [15], the *rate adapting Poisson* model (see figure 1) to text and image data [4], the *product of HMMs* model to language data [1], *hierarchical* versions of PoE to digits, text and collaborative filtering data [8, 9].

References

[1] A. Brown and G. Hinton. Products of hidden markov models. In *Proceedings of the Conference on Artifcial Intelligence and Statistics*, 2001.

[2] M. Carreira-Perpinan and G.E. Hinton. On contrastive divergence learning. In *Tenth International Workshop on Artificial Intelligence and Statistics*, Barbados, 2005.

[3] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[4] P. V. Gehler, A. D. Holub, and M. Welling. The rate adapting poisson model for information retrieval and object recognition. ACM, 06 2006.

[5] G.E. Hinton. Products of experts. In *Proc. of the Int'l Conf. on Artificial Neural Networks*, volume 1, pages 1–6, Edinburgh, GB, 1999.

[6] G.E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

[7] M.S. Lewicki and T.J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:p.337–365, 2000.

[8] G. Mayraz and G. Hinton. Recognizing hand-written digits using hierarchical products of experts. In *NIPS*, volume 13, 2000.

[9] R.R. Salakhutdinov, A. Mnih, and G.E. Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 21st International Conference on Machine Learning*, 2007.

[10] Y. W. Teh and G. E. Hinton. Rate-coded restricted Boltzmann machines for face recognition. In *Advances in Neural Information Processing Systems*, volume 13, 2001.

[11] Y.W. Teh, M. Welling, S. Osindero, and G.E. Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research - Special Issue on ICA*, 4:1235–1260, 2003.

[12] M. Welling, F. Agakov, and C.K.I. Williams. Extreme components analysis. In *NIPS*, volume 16, Vancouver, Canada, 2003.

[13] M. Welling, G.E. Hinton, and S. Osindero. Learning sparse topographic representations with products of student-t distributions. In *NIPS*, volume 15, Vancouver, Canada, 2002.

[14] M. Welling, M. Rosen-Zvi, and G.E. Hinton. Exponential family harmoniums with an application to information retrieval. In *NIPS*, volume 17, Vancouver, Canada, 2004.

[15] E. Xing, R. Yan, and A. Hauptman. Mining associated text and images with dual-wing harmoniums. In *UAI*, 2005.