Undirected Graphical Models

Chris Williams

Neural Information Processing
School of Informatics, University of Edinburgh

January 15, 2018

Overview

- Boltzmann Machines
- Learning Rule for Boltzmann Machines
- Sparse deep belief net model for visual area V2
- Energy Models for Natural Scenes
- Summary of Higher-Order Statistical Models

Boltzmann Machines

Dayan and Abbott §7.6

- Energy function of 0/1 units

$$E(\mathbf{u}) = -\mathbf{h} \cdot \mathbf{u} - \frac{1}{2}\mathbf{u}^T M \mathbf{u}$$

  note: M can be taken as symmetric

- Boltzmann distribution

$$P(\mathbf{u}) = \frac{1}{Z} \exp -E(\mathbf{u}), \qquad \text{where } Z = \sum_{\mathbf{u}} \exp -E(\mathbf{u})$$

- Glauber dynamics/Gibbs sampling

$$p(u_a(t+1) = 1 | \mathbf{u}(t)) = \frac{1}{1 + \exp(-I_a(t))}$$

  where

$$I_a(t) = h_a + \sum_{b=1}^{N_u} M_{ab} u_b(t)$$

- Asynchronous updates define a Markov chain whose equilibrium distribution is the Boltzmann distribution

Learning Rule

- Log likelihood $L(M) = \langle \log p(\mathbf{u}|M) \rangle_{p(\mathbf{u})}$

$$\frac{\partial L}{\partial M_{ab}} = \langle u_a u_b \rangle_{p(\mathbf{u})} - \sum_{\mathbf{u}} p(\mathbf{u}|M) u_a u_b$$
$$\stackrel{def}{=} \langle u_a u_b \rangle^+ - \langle u_a u_b \rangle^-$$

- Learning stops when statistics are the same in clamped $(+)$ and unclamped $(-)$ phases (aka wake and sleep phases)
- Note Hebbian and anti-Hebbian terms
- Generally the expectation in the negative phase is intractable and is approximated by sampling

Boltzmann Machines with Hidden Units

see Dayan and Abbott §8.4 pp 322-326

- Hidden units denoted by $\mathbf{v}$
- Energy Function

$$E(\mathbf{u}, \mathbf{v}) = -\mathbf{v}^T W \mathbf{u} - \frac{1}{2} \mathbf{v}^T M \mathbf{v}$$

$$P(\mathbf{u}|W, M) = \frac{1}{Z} \sum_{\mathbf{v}} \exp -E(\mathbf{u}, \mathbf{v})$$

- Learning rule (Ackley, Hinton and Sejnowski, 1985)

$$\frac{\partial \log p(\mathbf{u}^n|W, M)}{\partial W_{ab}} = \sum_{\mathbf{v}} p(\mathbf{v}|\mathbf{u}^n, W, M) v_a u_b^n - \sum_{\mathbf{u},\mathbf{v}} p(\mathbf{v}, \mathbf{u}|W, M) v_a u_b$$
$$= \langle v_a u_b^n \rangle^+ - \langle v_a u_b \rangle^-$$

Restricted Boltzmann Machines and Deep Learning

- If $M = 0$ then $p(\mathbf{v}|\mathbf{u}) = \prod_a p(v_a|\mathbf{u})$ (show this)
- This architecture is known as a *restricted* Boltzmann Machine (RBM)
- RBMs can be stacked to carry out “deep learning”; after learning one hidden layer, the activity vectors of the hidden units, when they are being driven by the real data, can be treated as "data" for training another RBM (Hinton et. al., 2006). This is called a “deep belief network” (DBN)

Sparse deep belief net model for visual area V2

Lee, Ekanadham and Ng (2008)

- Consider an RBM with Gaussian visible units

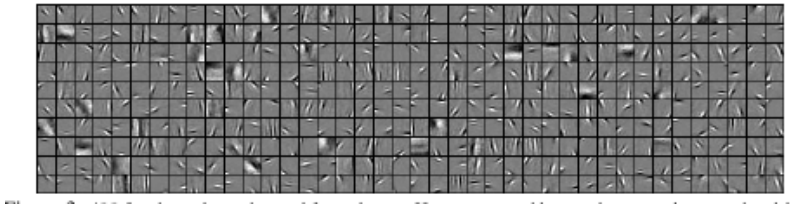
$$E(\mathbf{u}, \mathbf{v}) = \frac{1}{2\sigma^2} \sum_i u_i^2 - \frac{1}{\sigma^2} \left(\sum_i c_i u_i + \sum_j b_j v_j + \sum_{i,j} u_i v_j w_{ij} \right)$$

- $p(u_i|\mathbf{v}) \sim N(c_i + \sum_j w_{ij} v_j, \sigma^2)$
- Also impose a *sparsity prior* on the hidden units, with target sparseness p

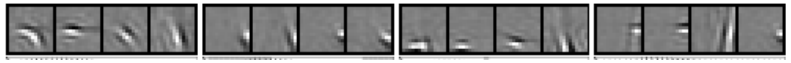
$$\sum_j ||p - \frac{1}{m} \sum_{k=1}^{m} \mathbb{E}[v_j^{(k)}|\mathbf{u}^{(k)}]||^2$$

- Layer 2 trained after layer 1 has learned (DBN)

First layer filters



Second layer: each unit "looks at" a small number of first layer units, e.g.



The leftmost patch in each group is a visualization of the model V2 basis, obtained by taking a weighted linear combination of the first layer bases to which it is connected.

Figure credits: Lee, Ekanadham and Ng (2008)

Properties of "V2" units can be compared to neural data

Energy Models for Natural Scenes

Osindero, Welling and Hinton (2005)

- Energy function

$$E(\mathbf{u}) = \sum_{i=1}^{M} \alpha_i \log\left(1 + \frac{1}{2}(W_i\mathbf{u})^2\right)$$

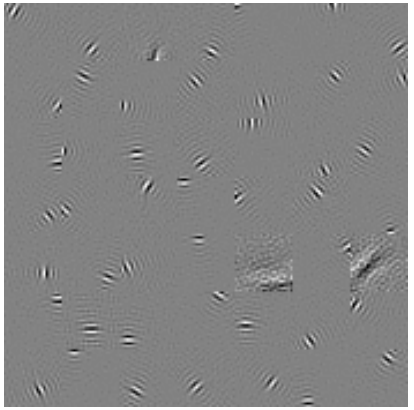
  Note: $\mathbf{u}$ is real-valued. W_i is ith row of W
- Product of Student-t distributions (PoT)

$$p(\mathbf{u}) = \frac{1}{Z}\prod_{i=1}^{M} \frac{1}{(1 + \frac{1}{2}(W_i\mathbf{u})^2)^{\alpha_i}}$$

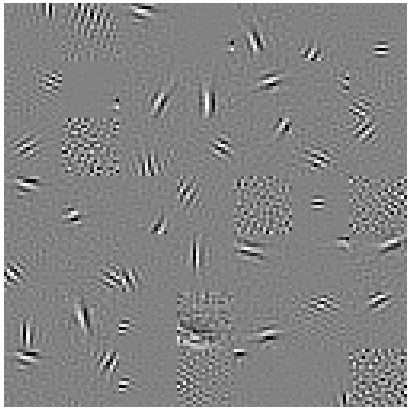
  a specific example of the Product of Experts formalism

- If $M = N_u$ then this model is equivalent to noiseless ICA with Student-t priors on the hidden units
- For the overcomplete case $M > N_u$ this model differs from the sparse coding approach.
- This model is discused in HHH §13.1.5-6
- As a t distribution can be represented as a GSM, we can set up the PoT models as a two-layer network; this gives one approach to learning W. In this case the energy function for the hidden variables corresponds to a Gamma distribution
- Roth and Black (2005) generalized the PoT model from image patches to whole images as a "field of experts" (convolutional architecture)

Learned filters shown in raw data space



complete



1.7× overcomplete

[Osindero, Welling and Hinton, 2005]

Hierarchical PoT models

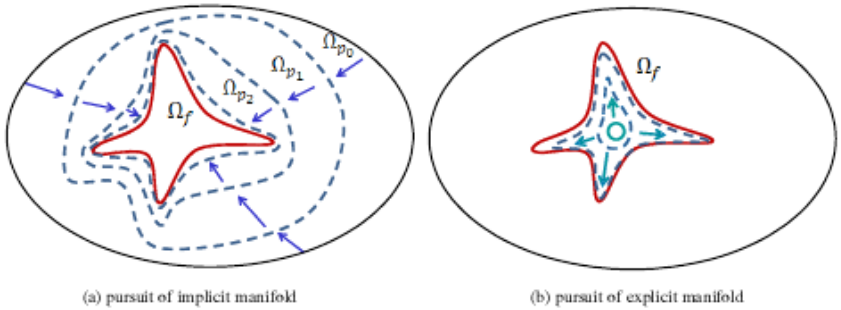
$$E(\mathbf{u}) = \sum_{i=1}^{M} \alpha_i \log \left(1 + \frac{1}{2} \sum_{j=1}^{K} C_{ij} (W_i \, \mathbf{u})^2 \right)$$

- C has non-negative weights
- As in subspace ICA and topographic ICA the C weights model the dependencies between the $\mathbf{v}$ units (where $\mathbf{v} = W\mathbf{u}$)
- Can obtain results similar to subspace ICA and topographic ICA

Discussion of Undirected Models

- Undirected models can also give rise to simple-cell like receptive fields (note they are receptive rather than projective fields)
- Lee et al (2008) demonstrate how a second layer can learn combinations of "simple cell" type units
- HPoT can give rise to layouts of retinotopy, phase, spatial frequency and orientation similar to cortical maps

Directed and Undirected Models



Zhu, Shi and Si, 2009

- Undirected: impose constraints
- Directed: model directions of variation
- Hybrid: combination of the above

Summary of Higher-Order Statistical Models

- Sparse coding and ICA models give rise to Gabor patches, as do energy models, and slow feature analysis (Berkes and Wiskott, 2005)
- Extensions can give rise to subspaces (modelling complex cells) and topographic organization of units. See also spatio-temporal bubbles (Hyvärinen et al, 2003)
- Convolutional architecture can deal with translation invariance (in space and/or time)
- Comparison of models ...

Recall the discussion from Dayan and Abbott (2001) p. 382

> *... structure in images arising from more complex objects than bars and gratings. It is unlikely that this higher-order structure can be extracted by a model with only one set of causes. It is more natural to think of causes in a hierarchical manner, with causes at a higher level accounting for structure in the causes at a lower level. The multiple representations in areas along the visual pathway suggest such a hierarchical scheme, but the corresponding models are still in the rudimentary stages of development.*

References

- Ackley, D. H., Hinton, G. E., Sejnowski, T. J. A learning algorithm for Boltzmann machines, *Cognitive Science* 9, 147-169, (1985)
- Berkes, P. and Wiskott, L. Slow feature analysis yields a rich repertoire of complex cell properties, *Journal of Vision*, 5(6):579-602 (2005)
- Hinton, G. E, Osindero, S., and Teh, Y. W. A fast learning algorithm for deep belief nets. Neural Computation, 18:1527-1554 (2006)
- Lee, H., Ekanadham, C., Ng, A. Y. Sparse deep belief net model for visual area V2. Advances in Neural Information Processing Systems 20 (2008).
- Osindero, S., Welling, M., Hinton, G. E. Topographic Product Models Applied to Natural Scene Statistics, *Neural Computation*, 18 (2), 2006
- Roth, S., Black, M. J. Fields of Experts: A Framework for Learning Image Priors, CVPR 2005