REVISITING BOLTZMANN LEARNING: PARAMETER ESTIMATION IN MARKOV RANDOM FIELDS

Lars Kai Hansen, Lars Nonboe Andersen, Ulrik Kjems, and Jan Larsen

CONNECT, Electronics Institute, building 349
Technical University of Denmark, DK-2800 Lyngby, Denmark
emails: lkhansen, nonboe, kjems, jlarsen@ei.dtu.dk

ABSTRACT

This contribution concerns a generalization of the Boltzmann Machine that allows us to use the learning rule for a much wider class of maximum likelihood and maximum a posteriori problems, including both supervised and unsupervised learning. Furthermore, the approach allows us to discuss regularization and generalization in the context of Boltzmann Machines. We provide an illustrative example concerning parameter estimation in an inhomogeneous Markov Field.

1. INTRODUCTION

Boltzmann Learning is conventionally formulated in terms of Gibbs distributions for *spin models*, for a review see, e.g., [1]. The learning rule is formulated for a general stochastic signal x. Let the "true" distribution of x be denoted $P(x)$, and let us assume that we have at our disposal a family H of model distributions parametrized by parameters w: $P(x|w,H)$. The objective of the conventional Boltzmann Machine learning rule is to choose w so as to minimize the Kullback information distance between the "true" distribution $P(x)$, (of which the *training set*: $D = \{x_k | k = 1, \cdots, p\}$ is a finite sample), and the model distribution $P(x|w,H)$. In this work we replace the true distribution by the *empirical distribution*

$$P_e(x) = 1/p \sum_k \delta(x - x_k). \tag{1}$$

The cost function for selection of w is then given by

$$\Delta[P_e(x), P(x|w)] = \int dx\, P_e(x) \log \frac{P_e(x)}{P(x|w)}. \tag{2}$$

The derivation of the Boltzmann learning rule assumes that the model distribution can be written in the *Gibbs* form:

$$P(x|w) = Z^{-1} \exp\left(-\epsilon(x|w)\right), \tag{3}$$

To appear in Proceedings of IEEE ICASSP-96, Atlanta, GA, May 7-10, 1996.

with $\epsilon(x|w)$ being a smooth function of w. Z is a (finite) normalization constant. The learning algorithm (estimation of w) can then be derived by gradient descent minimization of (2). The resulting recursive learning algorithm, with n being the iteration index, reads

$$w^{(n+1)} - w^{(n)} = \eta \left[\left\langle \frac{\partial \epsilon}{\partial w}(x) \right\rangle_{\text{clamp}} - \left\langle \frac{\partial \epsilon}{\partial w}(x) \right\rangle_{\text{free}} \right]. \tag{4}$$

$\langle \cdot \rangle_{\text{clamp}}$ indicates that the average is performed with respect to the empirical distribution. Similarly $\langle \cdot \rangle_{\text{free}}$ is the average computed with the model distribution based on the current set of parameters $w^{(n)}$. In brief we can characterize the learning process as follows: The parameters w are adjusted to minimize the difference between the conjugate operator averages in situations with and without grounding by observed data.

We further make the observation that the cost function $\Delta[P_e(x), P(x|w)]$, apart from a additive constant, is identical to the negative log likelihood of model distribution parameters:

$$\Delta[P_e, P(x|w)] = -\sum_k \log P(x_k|w) + \text{const.} \tag{5}$$

Using Bayes formula we find the posterior distribution of the parameters (see e.g. [4] for a discussion in contexts of neural nets),

$$-\log P(w|D) = \Delta[P_e(x), P(x|w)] - \log P(w) + \text{const.} \tag{6}$$

Where $P(w)$ is a prior. Minimization of the log posterior by gradient descent leads to the generalized Boltzmann Machine learning rule:

$$w^{(n+1)} - w^{(n)} = \eta \left[\left\langle \frac{\partial \epsilon}{\partial w}(x) \right\rangle_{\text{clamp}} - \left\langle \frac{\partial \epsilon}{\partial w}(x) \right\rangle_{\text{free}} \right] + \eta \frac{\partial \log P(w)}{\partial w}. \tag{7}$$

2. MARKOV FIELD BASED IMAGE SEGMENTATION

The Markov Field approach has found ample use in signal processing [2]. While most previous work has been based on homogeneous models, many image processing applications, in fact, involve inhomogeneous visual fields. In particular, sonar or radar imagery where the two image-directions implement deflection and range respectively. In this note we formulate a simple inhomogeneous Markov model involving a space-variant parametrization of the noise process. Our model is a simplified version of the model considered by Geman and Geman [2]. While the Markov Field model of [2] was based on Metropolis sampling from the Gibbs distribution of the Markov model, we compute averages using the *Mean Field* approximation [1]. Our Mean Field averaging is in turn implemented in the form of a cellular neural net, see [3] for more details.

Segmentation is important to most computer vision systems, however, even in its simplest form: Binarization of a grey-scale image there exist no established standard practice. We use the Markov Field approach to solve the segmentation problem. The target signal is a smooth binarization of a grey-scale image d_{ij}, in terms of two-valued pixels $S_{ij} \in \{-1, +1\}$. A Gibbs distribution is designed for which the appropriate averages are useful binarizations,

$$\begin{aligned} P(S|w, \alpha, d) = \\ & Z^{-1} \exp\left(-\frac{1}{2} \sum_{i,j=1}^{N} \sum_{i'j'=1}^{N} M_{ij;i'j'} (S_{ij} - S_{i'j'})^2 \right. \\ & \left. - \frac{1}{2} \sum_{ij=1}^{N} \alpha_j (S_{ij} - d_{ij})^2 \right). \end{aligned} \tag{8}$$

$M_{ij;ij'}$ defines the connectivity of the Markov Field. In the demonstration we use nearest neighbor coupling with unit strength. The derivation of the distribution (8) is further motivated in [3]. The distribution can be shown to follow if the signal degradation consists in addition of Gaussian white noise with spatially varying variance. We want to approach real data for which the noise statistics, here represented by the spatially varying noise variance α_j[1] is unknown, hence, has to estimated as part of the learning process.

[1]For Gaussian noise we have $\alpha_j = 1/\sigma_j^2$. We assume that the noise varies in strength in the "range" (horizontal) direction in the image.

3. EXPERIMENTAL EVALUATION AND CONCLUDING REMARKS

We illustrate the performance of the procedure on the artificially generated multi-object scene shown in the upper panel of figure 1. To simulate radar or sonar imagery, Gaussian white noise, with an increasing noise variance along the horizontal range direction, was added to produce the degraded image in the lower panel of figure 1.

To produce the outputs of a training set for supervised learning a "teacher network" with spatially varying noise parameter, α_j, where j index the horizontal range direction, was applied to the noisy scene. The inhomogeneous teacher network produced the lower image in figure 2. First, a "student network" was trained in supervised mode with output as produced by the teacher. The student network was initialized with uniform noise parameters, $\alpha_j = 0.5$. The resulting range dependence of the student network's variance parameter, after 100 iterations of the Boltzmann learning rule, is shown in figure 3, upper panel. Secondly, we included an additional "hyperparameter" *prior* enforcing smoothness of the spatially varying noise parameter,

$$P(\alpha) \propto \exp\left(-\frac{1}{2} \sum_j \mu(\alpha_{j+1} - \alpha_j)^2 \right) \tag{9}$$

where μ controls the smoothness. 100 iterations of the regularized training procedure leads to smooth adaptation of the inhomogeneous network as shown in the right panel of figure 3. The regularized adaptation produces a parameter set that closely resembles the teacher parameters, hence, will produce segmentations that closely reproduce those of the inhomogeneous teacher network. In a practical application of this learning procedure, the training set could, e.g., be produced by an expert. By inspection of a noisy scene representing a known environment, the expert could indicate a valid segmentation for the student network to learn. The procedure is evidently not confined to segmentation but has a wide range of applications in image processing, such as texture identification. Texture statistics are often inhomogeneous because of perspective distortions.

4. CONCLUSION

In conclusion we have presented a generalization of the Boltzmann Machine that allows us to discuss the learning rule for a wide class of models involving maximum likelihood and penalized maximum likelihood estimation (a posteriori estimation). We have shown

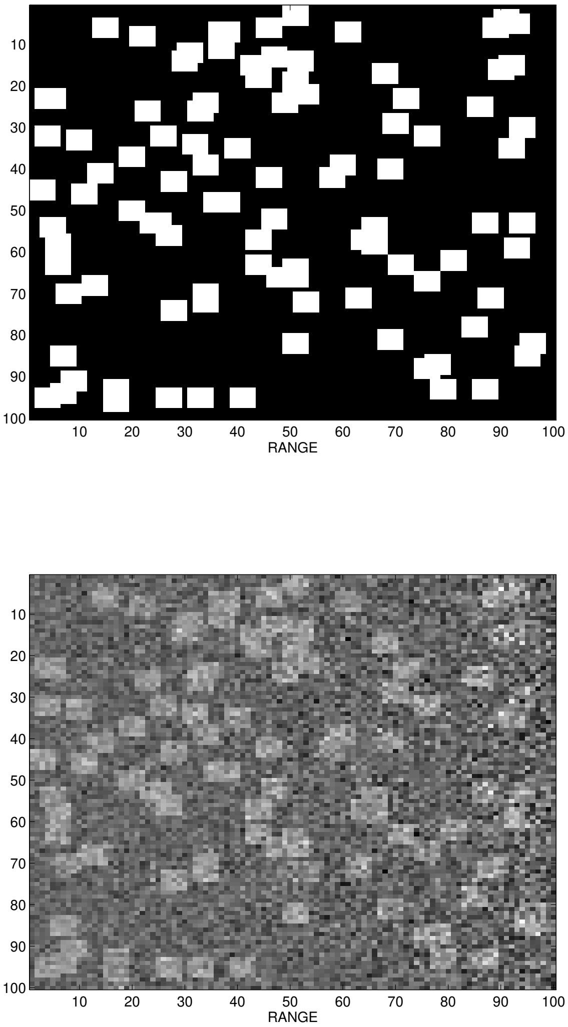


Figure 1: Upper: Multiple object scenery. Lower: Scene degraded by additive white noise of spatially varying variance (the noise variance increases in the horizontal "range" direction).

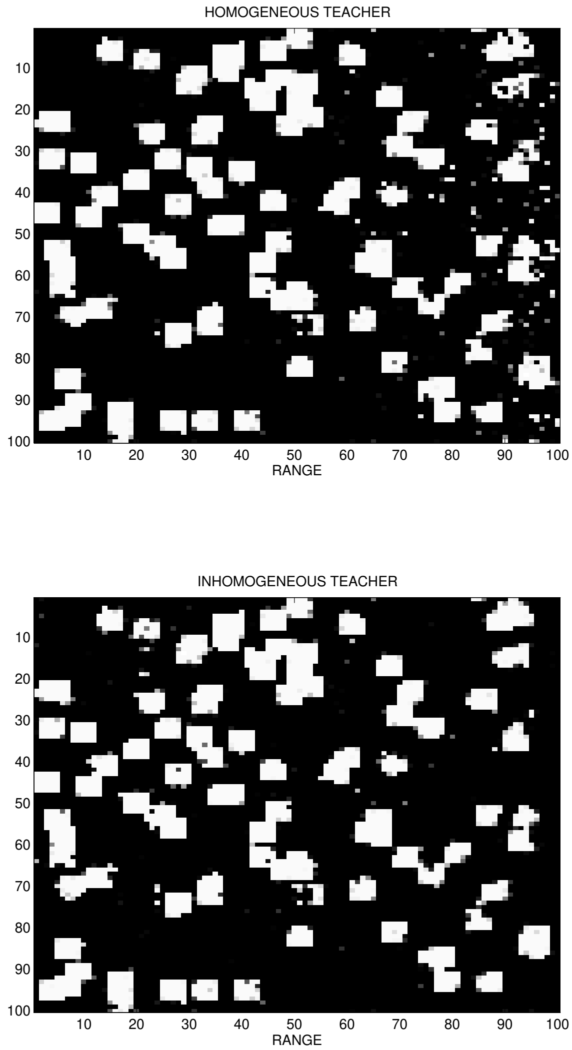


Figure 2: Upper: Multiple object scene reconstructed by a "teacher net" with noise variance parameter α that is constant in the range direction. Since the teacher does not match the statistics of the white noise process the reconstruction is increasingly poor in the right part of the scene. Lower: Reconstructed scene by a teacher net with spatially variant noise parameter (decreasing linearly in the range direction). Hence, the noise variance used by this teacher network matches the process generating the white noise in the input image, and as a result we see that the reconstruction is clearly improved in the noisy, right, part of the scene.

the viability of the approach for a particular example namely parameter estimation in an inhomogeneous Markov Field model.

With an inhomogeneous parametrization it is not possible to hand-tune the parameters as is custom in most applications of Markov Field models. However, we demonstrated that the Boltzmann Machine learning rule can be used to adapt the inhomogeneous network and, furthermore, may be generalized to incorporate prior information on the parameters of the associated Gibbs distribution.

5. ACKNOWLEDGMENTS

This research is supported by the Danish Research Councils for the Natural and Technical Sciences through the Danish Computational Neural Network Center. JL acknowledge the Radio Parts Foundation for financial support.

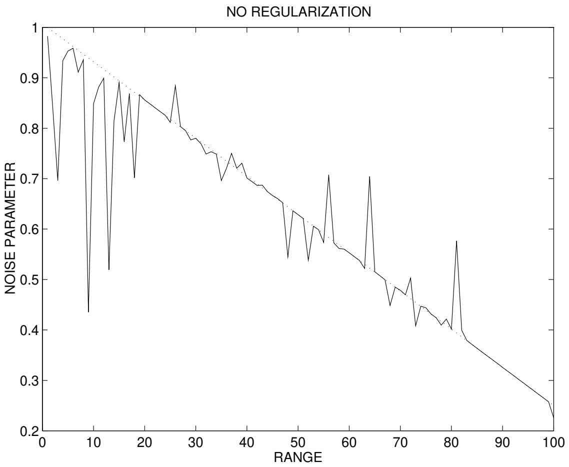
6. REFERENCES

[1] J. Hertz, A. Krogh and R.G. Palmer: *Introduction to the Theory of Neural Computation*, Addison Wesley, New York, 1991.

[2] D. Geman and S. Geman: "Stochastic Relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, 721–741, 1984.

[3] L. Nonboe Andersen and L.K. Hansen: "Inhomogeneous Markov Field Models," CONNECT technical report 1994, available by anonymous ftp from the server eivind.ei.dtu.dk in postscript form: /dist/1995/nonboe.inhomogeneous.ps.Z.

[4] D. MacKay: "Bayesian Interpolation," *Neural Computation*, vol. 4, 415–447, 1992.



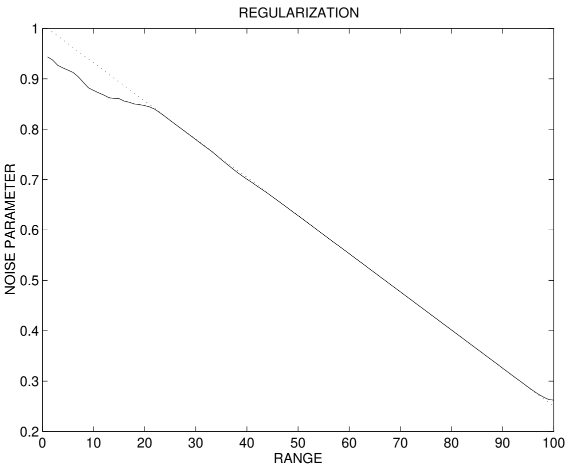


Figure 3: Upper panel: Result of Boltzmann Machine learning for a "student" network based on a training set of examples produced by the teacher network with spatially variant noise parameter. The linearly decreasing teacher network noise variance parameter, α_j, is marked with the dotted line. In this experiment the student adapted its weights without regularization, hence the relatively noisy identification of parameters. Lower panel: Result of Boltzmann Machine learning for a student network but here with a regularization term enforcing a smoothly varying noise parameter. As above we indicate the teacher networks noise variance parameter α_j by a dotted line). The regularized estimate is much less noisy and the student network we produce segmentations very similar to those of the "teacher" network.