Tractable Inference for Probabilistic Data Models

Lehel Csató and Manfred Opper
Neural Computing Research Group,
School of Engineering and Applied Science,
Aston University, Birmingham B4 7ET, United Kingdom

Ole Winther
Digital Signal Processing, Informatics and Mathematical Modelling
Technical University of Denmark, B208, 2800 Lyngby, Denmark

April 16, 2003

Abstract

Based on ideas from statistical physics, we present an approximation technique for probabilistic data models with a large number of hidden variables. We give examples for two non–trivial applications.

1 Introduction

The increasing amount of complex multivariate data which are produced in many modern scientific areas (eg. bioinformatics or biomedical research) makes the development of sensible models which help to interprete the data and allow to predict new data instances an important problem. Realistic models usually need to be flexible to adapt to complex data, they can be highly nonlinear and should also allow to incorporate the effects of noise and other uncertainties. Adaptive probabilistic data models provide a conceptually simple but highly flexible framework for explaining observed data by a set of hidden, unobserved causes which are modeled as random variables. Examples of such models are: Bayes belief networks [19] (used eg as trainable expert systems), independent component analysis [8, 1] (abbreviated ICA, which detects independent sources in nonlinear signal processing) and Gaussian process models [13] (modeling hidden spatial structures by random fields). Based on the joint distribution of all variables one can assign

plausible numerical values to the hidden causes from suitable conditional averages over the hidden variables.

Unfortunately, the flexibility of these models comes with a serious drawback. Except for a few simple cases (when the graph of dependencies between random variables has the structure of a tree, or, when the joint distribution is Gaussian) exact inference with probabilistic models becomes intractable in realistic cases when the number of variables is large. Hence, finding fast but reliable techniques for *approximate* inference with probabilistic models is an important and nontrivial task.

In recent years, a variety of approximation techniques have been imported from the field of statistical physics. One of the simplest methods is the well known mean field (MF) approximation which approximates the joint distribution of variables by a factorizing one. To take the neglected correlations at least partly into account, corrections to the MF method such as the Bethe/Kikuchi approximation [24] and the TAP approach, see e.g.[9, 23, 12, 7, 21, 13, 6] and references in [11] have become popular.

The TAP method was introduced by Thouless, Anderson & Palmer [3] to treat disordered materials with infinite ranged *random* interactions. We can view data models as disordered systems because the observed random data are parameters in the conditional distributions for the hidden variables. Unfortunately, the TAP method requires the exact knowledge of the distribution of the disorder, which for statistical physics models is usually assumed to be known, but for *real data* typically not. In order to make the method a general tool for practical applications, we have recently developed a version of the TAP approach [15, 14] which no longer requires the knowledge of the underlying distribution but adapts to the concrete observations. In this paper we present a simple derivation of our *adaptive* TAP (ADATAP) method and demonstrate its applications to a model for classification and an ICA model.

2 Probabilistic Models: Two Examples

2.1 Gaussian Process Models for Classification

Gaussian process (GP) models provide nonparametric approaches for supervised learning [22, 4, 13]. Take e.g., a binary classification problem, where we would like to classify input features $x \in R^D$ (which might be the vectors of pixel values for digitized handwritten characters) into two classes $y = \pm 1$ (say, the handwritten digit "4" against all other digits). A prob-

abilistic model could assume that the observed class labels are generated as $y = \text{sign}[f(x) + \xi]$ with an unknown function f, and where ξ is a zero mean noise process. The statistics of the noise can be encoded in the *likelihood* $P(y|f(x))$ of observing a label y, based on *knowing* the function value $f(x)$. In a Bayesian probabilistic model also the unknown function f becomes a random variable. We will encode a vague prior knowledge about the variability of functions f with their arguments x by modelling them as realizations of a Gaussian random field. For such *Gaussian processes* (with zero mean) the entire distribution $P_0[f]$ over function space is determined by its correlation function (or *kernel*) $K(x, x')$ which has to be supplied by the user of the algorithm. When a dataset of N correctly classified input/label pairs $D = (x_1, y_1), \ldots, (x_N, y_N)$ is available for training, one can use Bayes' theorem of probability to convert the prior distribution P_0 together with the likelihood into a posterior distribution over functions

$$P[f|D] = \frac{1}{Z} P_0[f] \cdot \prod_{i=1}^{N} P(y_i|f_i) \ , \tag{1}$$

where $f_i \equiv f(x_i)$ and Z acts a a normalizer. With an increasing number of training data one expects that the posterior distribution (1) becomes more and more concentrated around the function f which optimally classifies the data. Good predictions on novel test inputs x could then be based naturally on the the average $\langle f(x) \rangle$ over the distribution (1) which could be used to classify the new inputs x as $y = \text{sign}[\langle f(x) \rangle]$. One can show [13] that the computation of $\langle f(x) \rangle$ can be reduced to averages with respect to the joint distribution of function values at the training inputs x_i, which is

$$p(f_1, \ldots, f_N|D) \propto \exp\left[-\frac{1}{2} \sum_{i,j} f_i (K^{-1})_{ij} f_j \right] \cdot \prod_{i=1}^{N} P(y_i|f_i) \tag{2}$$

and the matrix K is defined through the kernel via $K_{ij} = K(x_i, x_j)$.

This probabilistic, *nonparametric* approach has the advantage over *parametric* techniques (e.g. neural networks) that the effective complexity of the model is not fixed beforehand but will effectively adapt to the dataset. Also the applicability of *kernel machines* to various non-trivial problems has been increased by the development of new types of kernels which are especially designed for classifying complex types of objects such as texts or protein strings [5]. The *disadvantage* comes from the fact that the necessary mathematical operations can not be performed exactly in an efficient way. Besides

the problem of analytically intractable distributions (2), the high dimensionality of correlation matrices K_{ij} make computations inefficient, when the size N of the training data sets becomes large.

2.2 Probabilistic Independent Component Analysis (ICA)

ICA is a widely applicable approach [8, 1, 17] in nonlinear signal processing and data analysis which aims at decomposing signals obtained from different sensors into a set of statistically independent *sources*. This finds a variety of applications for images, sound, text and telecommunication problems [8, 1] and also in the analysis of biomedical data, where one tries to separate an "interesting" part of the signal from other statistically independent contributions. In the simplest *probabilistic* formulation of ICA (for other approaches, see [8, 1]), one assumes that the vector $\mathbf{X}_t$ of signals at time t is an instantaneous linear mixing of sources $\mathbf{S}_t$ corrupted by additive Gaussian noise Γ_{ij}. We can write

$$\mathbf{X}_t = \mathbf{A}\mathbf{S}_t + \mathbf{\Gamma}_t \ , \tag{3}$$

where $\mathbf{A}$ is an unknown (but time independent) mixing matrix and the noise vector is assumed to be without temporal correlations having a time independent covariance matrix $\mathbf{\Sigma}$. The total probability of the temporal signal is assumed to be factorizing in time, ie. $P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma},\mathbf{S}) = \prod_t P(\mathbf{X}_t|\mathbf{A},\mathbf{\Sigma},\mathbf{S}_t)$. The aim of independent component analysis is to recover the unknown quantities which are the mixing matrix $\mathbf{A}$, the noise covariance $\mathbf{\Sigma}$ and the unknown sources $\mathbf{S}$ from the observed data. The crucial assumption is that of *statistical independence* of the sources (the hidden variables) at each time t, i.e. $P(\mathbf{S}_t) = \prod_i P(S_{it})$. A suitable functional form (which has to be non-Gaussian) for the source distribution $P(S)$ which incorporates e.g. known constraints (such as positivity or sub-Gaussian tails) must be chosen for each individual application. Alternatively, the source distribution can also adapt to the data, see e.g. Ref. [16].

Again, we can get plausible values for the unobserved sources by averaging the random values S_{it} over the posterior distribution computed from the prior distribution $P(\mathbf{S}_t)$ and the likelihood of the observations derived from equation (3). Although the *prior* distribution assumed independent sources, the posterior will obviously have *correlations* between different sources, which again makes computations of averages non–trivial. In addition, we must also learn the the mixing matrix $\mathbf{A}$ and the noise covariance $\mathbf{\Sigma}$ in parallel. These can be estimated from the training data by

the method of *maximum likelihood* (ML) [17], i.e. by maximizing the total probability of the observations

$$P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma}) = \int d\mathbf{S} P(\mathbf{X}|\mathbf{A},\mathbf{\Sigma},\mathbf{S})\, P(\mathbf{S}) \tag{4}$$

under the statistical assumptions.

3 A canonical Model

It is not hard to show that the two previous examples of probabilistic models (and, in fact, many others) require the computation of averages over posterior distributions of hidden variables which are of the type

$$P(\mathbf{S}) = \frac{\rho(\mathbf{S})}{Z} \exp\left[\frac{1}{2}\sum_{i,j} S_i J_{ij} S_j\right] \ . \tag{5}$$

The set of couplings J_{ij}'s encodes pairwise dependencies between the hidden random variables $\mathbf{S} = (S_1, \ldots, S_N)$. The factorizing term $\rho(\mathbf{S}) = \prod_j \rho_j(S_j)$ (called *likelihood* in the following) usually contains local observations at sites j, but can also incorporate additional local prior information about the variables S_i. E.g., by proper choices of the ρ_j's we can include both *discrete and continuous* random variables in the same model (5). The normalizing partition function Z is often (within a constant) equal to the probability that the model gives to the observed variables, which can be used as a yardstick for comparing different models or optimizing their hyperparameters. In the rest of the paper we will describe a simple and computationally efficient method for approximating marginal moments and correlation functions for the distribution (5) which enables us to deal with a variety of probabilistic models on real data.

4 The Gibbs Free Energy

Our approximation scheme is based on a *Gibbs Free Energy* G. It is an entropic quantity which allows us to compute moments of the distribution P, eq. (5) as well as the log of the normalization, $-\ln Z$ within the same approach. G is defined by a constrained minimization of a relative entropy measure $D(Q||P) \equiv \int d\mathbf{S}\ Q(\mathbf{S}) \ln \frac{Q(\mathbf{S})}{P(\mathbf{S})}$ between a distribution Q and the

posterior distribution P, where a set of relevant marginal moments are fixed. To be precise, we define

$$G(\mathbf{m}, \mathbf{M}) = \min_{Q} \Big\{ D(Q||P) \mid \langle \mathbf{S} \rangle_Q = \mathbf{m}, \langle \mathbf{S}^2 \rangle_Q = \mathbf{M} \Big\} - \ln Z \ , \qquad (6)$$

where the brackets denote expectations with respect to the variational distribution Q. $\langle \mathbf{S}^2 \rangle_Q$ is shorthand for a vector with elements $\langle S_i^2 \rangle_Q$. Minimizing G with respect to all arguments obviously leads to $\min_{\mathbf{m},\mathbf{M}} G(\mathbf{m}, \mathbf{M}) = -\ln Z$, where the total the minimizer is just $Q = P$. Hence, the moments of the distribution P are obtained as $\langle \mathbf{S} \rangle, \langle \mathbf{S}^2 \rangle = \mathrm{argmin}_{\mathbf{m},\mathbf{M}} G(\mathbf{m})$. Unfortunately, an exact calculation of (6) is as complicated as computing averages with respect to the distribution (5). To approximate the Gibbs free energy, we split G into two terms $G = G^0 + \Delta G$, where G^0 is the Gibbs free energy for the distribution (5), but where all couplings J_{ij} between the random variables are set to zero. The computation of the corresponding Gibbs free energy G^0 for such a "free" model is easy. Previous versions of the TAP approximation have been obtained by truncating a power series expansion of ΔG with respect to the interactions J_{ij} at second order [20, 21]. In contrast, our ADATAP approximation (motivated by the treatment of Parisi and Potters [18] of an Ising model with random orthogonal coupling matrix) will include terms of arbitrary order in the interactions. It will be defined in such a way that ΔG becomes exact when (5) is a Gaussian distribution. It can be shown for the Gaussian case that the interaction part ΔG^g (and the optimizing Gaussian distributions in (6)) comes out *independent* of the actual Gaussian likelihood chosen to compute G. It is only a function of the moments $\mathbf{m}$ and $\mathbf{M}$ and equals

$$\Delta G^g(\mathbf{m}, \mathbf{M}) = \max_{\mathbf{\Lambda}} \left\{ \frac{1}{2} \ln \det(\mathbf{\Lambda} - \mathbf{J}) - \frac{1}{2} \mathbf{m}^T \mathbf{J} \mathbf{m} - \frac{1}{2} \sum_i \Lambda_i \chi_{ii} \right\} + \frac{1}{2} \sum_i \ln \chi_{ii} + \frac{N}{2} \ , \qquad (7)$$

where $\mathbf{\Lambda}$ is a diagonal matrix with entries Λ_i and $\chi_{ii} \equiv M_i - m_i^2$. The *universal form* (7) will provide an approximation to G for *arbitrary likelihoods* via

$$G \approx G^0 + \Delta G^g \ . \qquad (8)$$

Hence, the problem of computing certain averages with the distribution (5) has been reduced to an optimization problem. We have developed a *message passing algorithm* for finding local minima (based on an earlier idea of T. Minka [10]) which is found to perform efficiently in practice.

5 Applications

5.1 Sparse approximation for Gaussian process classifiers

A straightforward application of the ADATAP approximation for computing predictions with the GP classifier model (2) [13] becomes impractical for datasets of several thousand training examples. Hence, a further approximation introducing *sparsity* is necessary. The idea is to replace the distribution P (more precisely, a Gaussian approximation Q^* which is implicitly computed with ΔG^g) by another one having a likelihood which depends only on a smaller subset of variables called "basis vectors" (BV) of size $n << N$ [2]. In order to minimize the loss of information caused by sparsity, the new distribution $\hat{Q}$ with a sparse likelihood is optmized by minimizing the relative entropy $D(\hat{Q}||Q^*)$ which can be done in closed form. We have applied the sparse ADATAP algorithm on the USPS dataset[1] of gray-scale handwritten digit images of size 16×16. We used an RBF kernel $K(x, x') = a_K \exp(-\|x - x'\|^2/(m\sigma_K^2))$, m being the dimension of the input vectors (256 in this case), and a_K and σ_K are parameters. The task was to classify the digits into fours and non-fours. Figure 1 shows the percentage of errors of the classifier on 2007 test patterns for different sizes n of the BV set (with $N = 7000$ training examples). The result shows a saturation of errors with increasing BV set suggesting that the sparse approximation extracts sufficient information from the data. Multiple sweeps of the algorithm through the dataset (averaged over different permutations of examples in the sequence) diminish the fluctuations caused by different orders of presentations.

5.2 Independent Component Analysis

We will present an application of ICA to feature extraction in hand-written digits [17]. We assume positive components of $\mathbf{A}$ (enforced by Lagrange multipliers) and a positive exponential prior on the sources $P(S_{it}) = \Theta(S_{it}) \exp(-S_{it})$ As in [17] we used 500 handwritten '3's which are assumed to be generated by 25 hidden images. Enforcing positivity (i.e. the images are generated by positive additions) will force the solution to become sparse, i.e. with many zeros in $\mathbf{A}$ and $\langle \mathbf{S} \rangle$. We obtained the statistically independent stroke styles of figure 2. This can be compared to the 25 components with largest eigenvalues obtained from a standard *principal component analysis* (PCA)

[1] Available from `http://www.kernel-machines.org/data/`

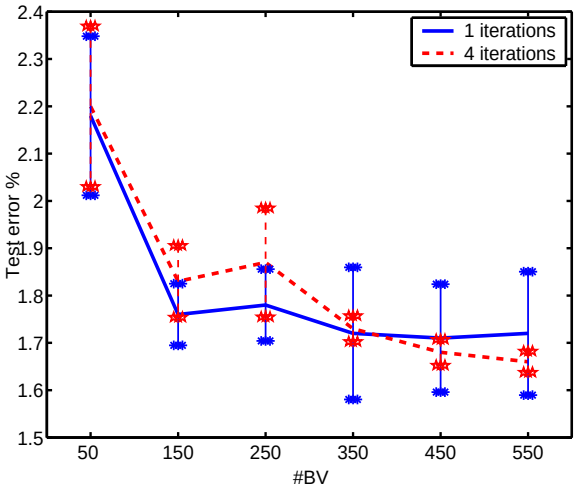
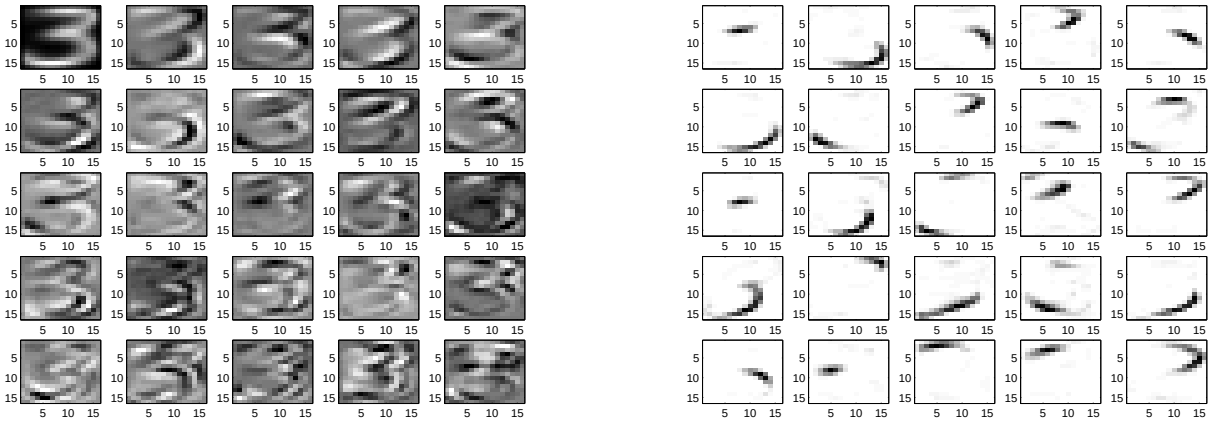


Figure 1: Test errors for classification with different BV sizes (x-axis) and multiple sweeps through the data.

which exhibit the typical "shadow effects" that occur when both negative and positive values are possible. The basis set found by the ICA algorithm can be seen as a statistically more reasonable representation of the components of images than the one found by PCA since it models more closely the true generative process of handwriting. Projection on this basis can be a powerful preprocessing step for hand-written digit classifiers.



Feature extraction for hand-written digits: The top plot show the first 25 principal components ordered according to eigenvalues. The bottom plot shows the 25 mean images (sources) for ICA with positive mixing matrix $\mathbf{A}$ and exponential (positive) source prior.

6 Conclusion and Outlook

We have demonstrated how approximation techniques from statistical physics can help to solve problems in data modelling. We expect that our ADATAP approximation will become a practical tool for inference with a variety of probabilistic data models. In fact, we are presently developing program packages both for ICA, Gaussian processes and general model of the type

(5) that will be made available online[2].

An important future direction of research will be the development of systematic improvements of the approximation. This will not only be of interest from a theoretical point of view but could also provide a user of the method with a measure of how well the final result can be trusted.

Acknowledgments

The work was supported by EPSRC grant no. GR/M81601.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[2] L. Csató and M. Opper. Sparse gaussian processes. *Neural Computation*, 14:641 – 668, 2002.

[3] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of a 'solvable model of a spin glass'. *Phil. Mag.*, 35:593, 1977.

[4] M. N. Gibbs. *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, Cambridge University, 1997.

[5] T. Jaakkola, M. Diekhans, and D. Haussler. Using the fisher kernel method to detect remote protein homologies. In *The Seventh International Conference on Intelligent Systems for Molecular Biology*, 1999.

[6] Y. Kabashima and D. Saad. Belief propagation vs. tap for decoding corrupted messages. *Euro. Phys. Lett.*, 44:668, 1998.

[7] H. J. Kappen and F. B. Rodríguez. Efficient learning in boltzmann machines using linear response theory. *Neural Computation*, 10:1137, 1998.

[8] T.-W. Lee. *Independent Component Analysis*. Kluwer Academic Publishers, Boston, 1998.

[9] M. Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *J. Phys. A (Math. Gen.)*, 22:2181, 1989.

[2]Available from `http://isp.imm.dtu.dk/staff/winther/`.

[10] T. P. Minka. *Expectation propagation for approximate Bayesian inference*. PhD thesis, Dep. of Electrical Eng. and Comp. Sci.; MIT, 2000.

[11] M. Opper and eds D. Saad. *Advanced Mean Field Methods, Theory and Practice*. MIT Press, 2001.

[12] M. Opper and O. Winther. A mean field approach to bayes learning in feed-forward neural networks. *Phys. Rev. Lett.*, 76:1964, 1996.

[13] M. Opper and O. Winther. Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12:2655, 2000.

[14] M. Opper and O. Winther. Adaptive and self-averaging thouless-anderson-palmer mean field theory for probabilistic modeling. *Phys. Rev. E*, 64:056131, 2001.

[15] M. Opper and O. Winther. Tractable approximations for probabilistic models: The adaptive thouless-anderson-palmer mean field approach. *Phys. Rev. Lett.*, 86:3695–3699, 2001.

[16] P. A. d. F. R. Højen-Sørensen, O. Winther, and L. K. Hansen. Analysis of functional neuroimages using ica adaptive binary sources. *to appear in Neurocomputing*, 2002.

[17] P. A. d. F. R. Højen-Sørensen, O. Winther, and L. K. Hansen. Mean field approaches to independent component analysis. *Neural Computation*, 14:3695–3699, 2002.

[18] G. Parisi and M. Potters. Mean-field equations for spin models with orthogonal interaction matrices. *J. Phys. A (Math. Gen.)*, 28:5267, 1995.

[19] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, 1988.

[20] T. Plefka. Convergence condition of the tap equations for the infinite-ranged ising spin glass model. *J. Phys. A*, 15:1971, 1982.

[21] T. Tanaka. Mean-field theory of boltzmann machine learning. *Phys. Rev. E*, 58:2302, 1998.

[22] C. K. I. Williams and C. E. Rasmussen. Gaussian proceses for regression. In *Advances in Neural Information Processing Systems*, number 8, pages 514–520, 1996.

[23] K. Y. M. Wong. Microscopic equations and stability conditions in optimal neural networks. *Europhys. Lett.*, 30:245, 1995.

[24] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In T.G. Dietterich T.K. Leen and V. Tresp, editors, *Advances in Neural Information Processing Systems*, number 13, pages 689–695. MIT Press, 2001.