Field-theoretic methods for intractable probabilistic models

Dennis Lucarelli[*]   Cheryl Resch[*]   I-Jeng Wang[*]   Fernando J. Pineda[†]

Abstract

We describe a general technique for estimating the intractable quantities that occur in a wide variety of large-scale probabilistic models. The technique transforms intractable sums into integrals which are subsequently approximated via saddle point methods. When applied to sigmoid and noisy-OR networks, the technique yields a generic mean-field approximation as well as a second order Gaussian approximation that accounts for the pairwise correlations between random variables in the network. In two example models, we observe that our lowest order approximation is identical to expressions obtained using Plefka's approach for deriving the TAP equations.

1 Introduction

It is well known that large-scale probabilistic modeling involves the computation of expected values or posterior probabilities which are generally intractable for sufficiently large systems due to the occurrence of terms that involve summations over all configurations of the relevant random variables. In this paper we describe the use of a generalized mean-field theory for intractable probabilistic models. The mathematical formalism was originally derived in [1]. In this approach, expectations over binary random variables are represented as integrals over real continuous variables. The integral representation is desirable because it lends itself to approximation via saddle point methods. When applied to probabilistic models, the method yields a precise second order approximation as well as a mean-field approximation. We show how these techniques both complement and extend variational methods. Mean-field methods [2] have recently become very popular for approximate inference and learning in Bayesian belief networks [3]. Upper [4] and lower [5] bounds on the marginal probability of the observed data can be obtained by variational methods. In these methods, one approximates the true joint probability function with a tractable one that can be optimized with respect to adjustable pa-

rameters. We found that in our numerical experiments our mean-field approximation is nearly identical to the variational upper bound [4]. Although we have not proven that our mean-field approximation is indeed an upper bound, it appears that in some circumstances this may be true. Higher order techniques from statistical physics that account for correlations between spins have recently been applied to probabilistic data models. Plefka's [6] method for obtaining the TAP equations [7] has been extended to the Bayesian belief network framework [8]. As observed in [8], saddle-point methods are an alternative way of obtaining their lowest order approximation. Indeed, our mean-field approximation is identical to their lowest order approximation.

2 Preliminaries

To begin, consider an unspecified probabilistic model whose state vector consists of n binary random variables, $\mathbf{s} = (s_1, \cdots, s_n)$. Without loss of generality we can write the joint distribution of these state variables in a Boltzmann-like form $P(\mathbf{s}) = \exp\left(-\beta E(\mathbf{s})\right)$ where the energy-like term, $E(\mathbf{s})$, has a form defined by the particular model in question. We need only assume that $E(\mathbf{s})$ is bounded from below over the unit hypercube $[0,1]^n$ and that it has finite first and second order derivatives. The inverse temperature β is useful for subsequent theoretical analysis, but for actual inference calculations it is set equal to 1.

Assume the set of binary variables (spins) is partitioned into a set V, of n_V instantiated (or visible) variables and a set H, of n_H uninstantiated (or hidden) variables. The requisite computational task is to compute the expected value of an arbitrary function $f(\mathbf{s})$ over all configurations of the variables in a subset H

$$(2.1) \quad E\left\{f(\mathbf{s})|V\right\} \equiv \sum_{\{s_i|s_i \in H\}} f(s) \exp\left(-\beta E(\mathbf{s})\right).$$

This form appears generically in calculations with probabilistic models. For example, if $f(\mathbf{s})$ is unity and V is the empty set, then $E\{f(\mathbf{s})|V\}$ is just the partition function of the system. On the other hand, if $f(\mathbf{s})$ is unity and the values in V are determined by some data, then $E\{f(\mathbf{s})|V\}$ is the likelihood of the data, $P(V)$. Finally, if $f(\mathbf{s}) = s_\mu s_\nu$, then $E\{f(\mathbf{s})|V\}$ is just the average correlation between the two spins s_μ and s_ν.

---
[*]Research and Technology Development Center, The Johns Hopkins University Applied Physics Laboratory

[†]Department of Molecular Microbiology and Immunology, The Johns Hopkins Bloomberg School of Public Health

To proceed further, we simplify the notation by defining a permutation $\Omega \mathbf{s} = \mathbf{s}'$ that sorts the components of the state vector $\mathbf{s}$, such that the hidden variables come first, i.e. have indices $\{1, \cdots, n_H\}$. This transformation induces corresponding transformations over all other quantities in the model, e.g. if $g(\cdot)$ is an arbitrary function of the state vector $\mathbf{s}$, then $\Omega \colon g(\mathbf{s}) \to g'(\mathbf{s}')$. In the transformed coordinate system, summations of arbitrary functions over configurations of hidden units take the form

$$(2.2) \qquad \sum_{\{s_i | s_i \in H\}} g(\mathbf{s}) \to \sum_{s_1' = 0}^{1} \cdots \sum_{s_{n_H}' = 0}^{1} g'(\mathbf{s}').$$

In effect, we define a transformed probabilistic model with n_H variables, and whose model parameters depend on the values of the visible units in the original model. With these definitions and conventions, our generic calculational task reduces to evaluation of

$$(2.3)$$
$$\begin{aligned} Z &\equiv E\{f(\mathbf{s}) | V\} \\ &= \sum_{s_1' = 0}^{1} \cdots \sum_{s_{n_H}' = 0}^{1} \exp\left(-\beta E'(\mathbf{s}') + \log(f'(\mathbf{s}'))\right). \end{aligned}$$

We henceforth denote generic sums of the form (2.3) by the symbol Z.

3 Integral Representation

The key to exploiting field theory techniques is to transform the summations in discrete models into equivalent integral representations in field theories. In particular we apply the transformation

$$(3.4) \quad \sum_{s_1' = 0}^{1} \cdots \sum_{s_{n_H}' = 0}^{1} \exp\left(-\beta E'(\mathbf{s}') + \log(f'(\mathbf{s}'))\right)$$
$$= \frac{1}{(2\pi)^{n_H}} \int D\{\mathbf{x}\} D\{\hat{\mathbf{x}}\} \exp\left(-\beta \tilde{F}'(\mathbf{x}, \hat{\mathbf{x}})\right)$$

where
$$(3.5)$$
$$\tilde{F}'(\mathbf{x}, \hat{\mathbf{x}}) \equiv E'(\mathbf{x}) - \beta^{-1} \log(f'(\mathbf{x})) - \beta^{-1} \boldsymbol{\Sigma}(\mathbf{x}, \hat{\mathbf{x}})$$

and

$$(3.6) \qquad \boldsymbol{\Sigma}(\mathbf{x}, \hat{\mathbf{x}}) = \sum_\mu \left[i x_\mu \hat{x}_\mu + \ln(1 + \exp(-i \hat{x}_\mu)) \right].$$

To obtain this transformation, observe that any function $g(s)$, satisfies the identity $\sum_{s=0}^{1} g(s) = \int_{-\infty}^{\infty} g(x) \{\delta(x) + \delta(x - 1)\} dx$, where $\delta(\cdot)$ is the Dirac-δ function. Substituting this identity into (2.3) and using the Fourier

representation of the Dirac-δ function [9], allows us to explicitly perform the summation over spins, thereby recovering transformation (3.4) where $\int D\{\mathbf{x}\} D\{\hat{\mathbf{x}}\} \equiv \{\prod_\mu \int_{-\infty}^{\infty} dx_\mu\} \{\prod_\nu \int_{-\infty}^{\infty} d\hat{x}_\nu\}$. In the resulting representation for Z, no trace remains of the original spin variables, each of which has been replaced by a conjugate-pair of continuous real fields $(x_\mu, \hat{x}_\mu)$. These fields are coordinates in a continuous phase-space $\Re^{n_H} \times \Re^{n_H}$. Observe that if $\beta E'(\mathbf{x}) + \log(f'(\mathbf{x}))$ is a positive definite quadratic function of $\mathbf{x}$, the RHS of (3.4) can be integrated exactly with respect to $\mathbf{x}$, in which case the generalized transformation reduces to the textbook Gaussian transformation where $\hat{\mathbf{x}}$ is the usual auxillary real field.

To proceed further, we appeal to the saddle-point method which is an asymptotic approximation method which becomes exact in the $\beta \to \infty$ limit. Although a detailed discussion of this approximation is beyond the scope of this manuscript, the interested reader may read more details in Erdélyi [10]. The intuition behind the method is based on the observation that by a suitable choice of the integration path, the values of the integrand which are far from the stationary point of the integrand make arbitrarily small contributions to the integral as β becomes arbitrarily large. In other words, the integral is insensitive to the shape of the tails. This motivates the saddle point approximation wherein $\tilde{F}'(\mathbf{x}, \hat{\mathbf{x}})$ is replaced by a second-order Taylor series expansion about the stationary point. The exact and approximate integrands only deviate away from the stationary point where the contribution to the integral is small. As we shall see below, this leads to two approximations: a zeroth and a second order approximation. The simplest saddle point approximation method assumes that $\tilde{F}'(\mathbf{x}, \hat{\mathbf{x}})$ has a unique stationary point in phase-space. In general this assumption is false, but it is useful to derive the approximation under this assumption and then to revisit the assumption later. Finally, it is worthwhile to point out that there exists a dichotomy in the literature concerning the term "saddle point approximation." The statistics community uses the term to refer to second order approximations [11], whereas the physics community usually uses the term to refer to zeroth order approximations [2]. We adopt the physics terminology and use the terms "mean-field" and "Gaussian" approximations to refer to the zeroth and second order approximations respectively.

To carry out the saddle point calculation, one must first find the saddle point. This requires that we solve the two stationarity conditions $\nabla_{x_\mu} \tilde{F}' = \nabla_{\hat{x}_\mu} \tilde{F}' = 0$. These $2n_H$ conditions can be reduced to just n_H conditions by eliminating $\hat{\mathbf{x}}$ which leads to

$$(3.7) \qquad \nabla_{x_\mu} F'(\mathbf{x}) = 0$$

where F' has the form

$$(3.8) \quad F'(\mathbf{x}) \equiv E'(\mathbf{x}) - \beta^{-1}\ln\left(f'(\mathbf{x})\right) - \beta^{-1}S(\mathbf{x})$$

and $S(\mathbf{x}) \equiv -\sum_\mu x_\mu \log(x_\mu) - \sum_\mu (1-x_\mu)\log(1-x_\mu)$ is the familiar binary entropy. Note that $S(\mathbf{x})$ is only defined over the unit hypercube and that within the unit hypercube, $F'(\mathbf{x})$ is bounded from below. Moreover, the contribution to $F'(\mathbf{x})$ due to $S(\mathbf{x})$ causes the gradient $\nabla_{x_\mu} F'(\mathbf{x})$ to diverge to $+\infty$ at the boundaries of the unit hypercube. In short, the saddle point is found by performing gradient descent on $F'(\mathbf{x})$, starting from an interior point of the unit hypercube. Note that since $S(\mathbf{x})$ confines the gradient descent algorithm to the interior of the n_H-dimensional hypercube, it plays the role of a barrier function. Essentially, we find the saddle point by employing an interior point method [12] with an entropic, rather than a logarithmic barrier function. We note that with $f'(x) \equiv 1$, the expression (3.8) is identical to the expression appearing in the lowest order approximation of Bhattacharyya and Keerthi [8].

Armed with the saddle point, we expand $\tilde{F}'(\mathbf{x}, \hat{\mathbf{x}})$ to second order, and substitute the result into the RHS of (3.4) thereby obtaining

$$(3.9)$$
$$Z^{(2)} \equiv Z^{(0)} \cdot \frac{1}{(2\pi)^{n_H}} \int D\{\mathbf{x}\} D\{\hat{\mathbf{x}}\} \exp\left(-\beta \tilde{F}'^{(2)}(\mathbf{x}, \hat{\mathbf{x}})\right),$$

where

$$(3.10) \quad Z^{(0)} \equiv \exp\left(-\beta F'(x^o)\right).$$

$Z^{(0)}$ is the zeroth order approximation for the desired sum, while $Z^{(2)}$ is the second order approximation for the desired sum. $Z^{(2)}$ is easily calculated because the integral is Gaussian and can be integrated explicitly. The resulting closed form expression for $Z^{(2)}$ is

$$(3.11) \quad Z^{(2)} = \frac{\exp\left(-\beta F'(x^o)\right)}{\sqrt{\det\left(\mathbf{I} + \mathbf{H}^{(xx)}\mathbf{H}^{(\hat{x}\hat{x})}\right)}}$$

where the submatrices $\mathbf{H}^{(xx)}$ and $\mathbf{H}^{(\hat{x}\hat{x})}$ are respectively, the configuration-space and Fourier-space projections of the curvature matrix, $\beta \nabla^2 \hat{F}'(\mathbf{x}, \hat{\mathbf{x}})$, i.e.

$$(3.12) \quad \nabla^2 \tilde{F}'\Big|_{x=x^o, \hat{x}=\hat{x}^o} = \beta^{-1}\begin{pmatrix} \mathbf{H}^{(xx)} & -i\mathbf{I} \\ -i\mathbf{I} & \mathbf{H}^{(\hat{x}\hat{x})} \end{pmatrix}.$$

Our mean-field algorithm (FT-0) involves a numerical minimization in an n_H dimensional space while the variational methods involve minimizations in $2n_H$ dimensional spaces. The saddle-point approximation algorithm (FT-2) requires no additional minimization beyond the FT-0 algorithm but does require an additional $O\left(n_H^3\right)$ determinant evaluation. Despite the increased computational cost, in our numerical experiments the second-order saddle-point method proved to be a more precise approximation.

4 Revisiting Convexity

The discussion up to this point assumed a convex $F'(\mathbf{x})$. For $\beta = 0$ it is easily demonstrated that $F'(\mathbf{x})$ is convex. For nonzero β however, the convexity of $F'(\mathbf{x})$ depends on the details of the particular probabilistic model. Clearly, one can construct a model with convex $F'(\mathbf{x})$. In general, however, the question naturally arises: how large can we make β and still have a convex optimization? Although no general answer is possible, for any particular model it is possible to establish an upper bound on the inverse temperature below which we can guarantee convexity. In particular, since we perform inference with $\beta = 1$, we can guarantee convexity if the upper bound exceeds 1. A necessary and sufficient condition for strict convexity is for the eigenvalues of $\nabla^2 F'(\mathbf{x})$ to be positive. For this purpose it is useful to localize the eigenvalues of $\nabla^2 F'(\mathbf{x})$ via the Gershgorin circle theorem [13]. In the absence of guaranteed convexity, one can appeal to heuristic methods. For example, deterministic annealing is a heuristic method, wherein one first solves for the unique minimum at $\beta < \beta^*$ and then tracks this minimum as one gradually increases β until the value $\beta = 1$ is reached. Deterministic annealing has a successful record on a broad spectrum of related practical problems, e.g. Boltzmann machines [14], elastic nets [15] and travelling salesman problems [16].

5 Examples

To demonstrate the utility of our methods we applied them to the task of inference with sigmoid belief networks [17] and noisy-OR networks [18]. The required sum in this case is the likelihood of the data, which is simply the marginal of the joint distribution taken over the configurations of the hidden units, $P(V) = \sum_H P(H, V)$. The energy function in the original spin variables takes the form

$$(5.13)$$
$$E(\mathbf{s}) = -\sum_i^N \left(s_i \log G(\mathbf{w}_i) + (1 - s_i)\log(1 - G(\mathbf{w}_i))\right)$$

where $\mathbf{w}_i = \sum_{j \in \pi(s_i)} J_{ij} s_j + h_i$ and G is the activation function which depend on the specific probabilistic model. For sigmoid and noisy-OR networks, the activation function is given by $\sigma(x) = (1 + \exp(-x))^{-1}$ and $\rho(x) = 1 - \exp(-x)$ respectively.

5.1 Sigmoid Network Technical details for the sigmoid network may be found in [19]. Here we simply summarize numerical results for the case of the sigmoid three layer ($2 \times 4 \times 6$) network used by Saul et al. [5] to illustrate the variational lower bound algorithm and the

sigmoid bipartite (8×8) network to illustrate the variational upper bound derived by Jaakkola and Jordan [4].

We define the relative approximation error to be

$$(5.14) \qquad \varepsilon = \frac{\ln\left(P^{(\lambda)}(V)\right) - \ln\left(P(V)\right)}{|\ln\left(P(V)\right)|}$$

where $P^{(\lambda)}(V)$ is the likelihood of the data computed for one realization of a sigmoid belief network with the λ-th approximation algorithm. The distribution of relative approximation errors, over an ensemble of sigmoid belief networks, is $P^{(\lambda)}(\varepsilon)$ and is estimated numerically. Statistics were accumulated over 10000 realizations of $\mathbf{J}$ and $\mathbf{h}$. The elements of these were sampled from a uniform distributed between -1 and 1. Of these 10000 realizations, all were convex, i.e. there were no realizations that violated the convexity criterion. We calculated the likelihood of the event that all the units in the bottom layer were instantiated to zero. For this event we compute four quantities: 1) the exact, 2) the variational lower-bound, 3) the saddle point mean-field approximation, and 4) the field-theory saddle-point approximation. The variational upper bound is not defined for this network.

For the bipartite network, statistics were accumulated over 2494 realizations of $\mathbf{J}$ and $\mathbf{h}$. Approximately, 21% of these satisfied the Gershgorin convexity criterion. The weights are sampled from a Gaussian distribution with zero mean and standard deviation 1. For this network, we calculate the variational upper bound for bipartite networks [4] in addition to the four quantities above. A comparison of the average error of the four algorithms is displayed in Table 1.

Table 1 Sigmoid network

Algorithm	Expected % error (8×8)	Expected % error ($2 \times 4 \times 6$)
VLB	-3.2%	-1.58%
VUB	15.2%	n/a
FT-0	15.1%	4.0%
FT-2	.18%	0.033%

Table 1: Average error for four approximation algorithms: variational lower and upper bounds (VLB and VUB respectively) and field theory zeroth and second order algorithms (FT-0 and FT-2 respectively).

5.2 Noisy-OR Network In a noisy-OR belief network, variable dependencies are modeled as binary valued noisy OR logic gates. The matrix element q_{ij} is the probability that an active parent node s_j will activate

the node s_i. The conditional probability is given by

$$(5.15)$$
$$P(s_i|\pi(s_i)) = \left(1 - \exp\left(-\sum_{j \in \pi(s_i)}^{n} J_{ij}s_j + h_i\right)\right)^{s_i} \cdot$$
$$\left(\exp\left(-\sum_{j \in \pi(s_i)}^{n} J_{ij}s_j + h_i\right)\right)^{1-s_i}$$

where $J_{ij} = -\log(1 - q_{ij})$, $\pi(s_i)$ denotes the set of parents of the variable s_i, and h_i defines the prior distribution over top layer hidden nodes. Substituting expression (5.15) into the joint distribution $P(\mathbf{s}) \equiv \prod_i^N P(s_i|\pi(s_i))$ and rearranging, we obtain the energy function (5.13) in the simple form

$$(5.16)$$
$$E(\mathbf{s}) = \sum_i^N \left(\mathbf{w}_i(\mathbf{s}) - s_i \log\left(\exp(\mathbf{w}_i(\mathbf{s}) - 1)\right)\right).$$

To compute the likelihood of the data $P(V)$, we construct the integral representation (3.4) and solve the stationarity conditions $\nabla_{x_\mu}\tilde{F}'(\mathbf{x}) = 0$ where

$$(5.17)$$
$$\tilde{F}'(\mathbf{x}) = \sum_{\mu \in H}(\mathbf{w}_\mu(\mathbf{x}) - x_\mu \log(\exp(\mathbf{w}_\mu(\mathbf{x}) - 1))$$
$$+ \sum_{j \in V}(\mathbf{w}_j(\mathbf{x}) - v_j \log(\exp(\mathbf{w}_j(\mathbf{x}) - 1)) - \beta^{-1}S(\mathbf{x})$$

and v_j is an visible node.

To test the performance of the algorithms when applied to noisy-OR networks, we employed a strategy similar to the one used in the sigmoid case. For both topologies, statistics were accumulated from 10000 realizations of $\mathbf{J}$, sampled uniformly from the interval $[.4, .8]$. We calculated the worst case marginal, *i.e* when all evidence nodes are set to 1. In the bipartite case, we found no samples that violated the convexity criterion. However, in the three-layer network less than 1% of the samples were found to be convex. The results indicate that the approximation performs reasonably well even when we do not have a convex minimization problem.

6 Discussion

To conclude, we described how to use integral representations in large scale probabilisitic models to approximate intractable summations. The mathematical techniques described here are quite general and we expect the approach described in this paper to be useful in

Table 2 Noisy-OR network

Algorithm	Expected % error (8×8)	Expected % error $(2 \times 4 \times 6)$
VUB	17.3%	n/a
FT-0	17.3%	18.2%
FT-2	2.18%	3.15%

Table 2: Average error for three approximation algorithms: variational upper bound (VUB) and field theory zeroth and second order algorithms (FT-0 and FT-2 respectively).

a very broad range of application domains, (e.g. neural modeling, computational biology, computational finance, computer-aided diagnosis and decision support) where complex large-scale probabilistic models are the norm rather than the exception.

A direct comparison of our second-order saddle-point approximation to the second-order methods of Bhattacharyya and Keethi [8] is difficult to make since they choose to approximate the unclamped partition function as opposed to the marginal of the joint probability considered here. In addition, in our examples we have chosen the visible nodes to correspond to the *worst case* marginal where they have chosen to simulate the cases that maximize and minimize the value of the partition function. Despite these differences, an inspection of the numerical results obtained in the two approaches, especially in the sigmoid case, shows that they are of comparable precision. As mentioned above, one can show analytically that the lowest order approximations are identical. Moreover, in the two-layer case the numerical results from our lowest order approximation coincide with the variational upper bound of Jaakkola and Jordan [4].

In the case of sigmoid and noisy-OR belief networks, we demonstrated a mean-field approximation and a Gaussian approximation which, when combined with the variational mean-field algorithm, form a suite of three algorithms for approximate inference with general sigmoid and noisy-OR belief networks.

Acknowledgements

References

[1] F. J. Pineda and I-J. Wang, *Field theory for intractable probabilistic models*, RTDC Technical Report No. RSI-99-003, (1999).

[2] G. Parisi, *Statistical Field Theory*, Addison Wesley, Redwood City, CA, 1988.

[3] M. Opper and D. Saad eds. *Advanced Mean Field Methods: Theory and Practice*, M.I.T Press, Cambridge, MA, 2001.

[4] T. Jaakkola and M.I. Jordan, *Computing upper and lower bounds on likelihoods in intractable networks* Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence, (1996), pp. 340-348.

[5] L. Saul, T. Jaakkola, and M. I. Jordan, *Mean field theory for sigmoid belief networks*, Journal of Artificial Intelligence Research, 4 (1996), pp. 61-76.

[6] T. Plefka, *Convergence condition of the tap equation for the infinite-ranged Ising glass model*, Journal of Physics A: Mathematical and General, 15 (1982), pp. 1971-1978.

[7] D. J. Thouless, P. W. Anderson, and R. G. Palmer, *Solution of 'Solvable model of a spin glass'*, Philosophical Magazine A, 35 (1977) pp. 593-601.

[8] C. Bhattacharyya and S. Keerthi, *Mean-field methods for a special class of Belief Networks*, Journal of Artificial Intelligence, 15, (2001), pp. 91-114.

[9] M. J. Lighthill, *Introduction to Fourier Analysis and Generalized Functions*, Cambridge University Press, London, England, 1958.

[10] A. Erdélyi, *Asymptotic Expansions*, Dover, New York, NY, 1956.

[11] N. Reid, *Saddlepoint Methods and Statistical Inference*, Statistical Science, 3 (1998), pp. 213-238.

[12] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.

[13] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, Baltimore, MD, 1996.

[14] C. Peterson and B. Söderberg, *A new method for mapping optimization problems onto neural networks*, International Journal of Neural Systems, 1 (1989), pp. 3-22.

[15] R. Durbin and D. Wilshaw, *An analog approach to the traveling salesman problem using an elastic net method*, Nature, 326, 1987, pp. 689-691.

[16] C. Peterson, *Parallel distributed approaches to combinatorial optimization problems – benchmark studies on T. S.P.*, Neural Computation, 2 (1990), pp. 261-270.

[17] R. M. Neal, *Connectionist learning of belief networks*, Artificial Intelligence, 56 (1992), pp.71-113.

[18] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kauffman, San Mateo, CA, 1988.

[19] F. J. Pineda, C. Resch, and I-J. Wang , *Saddle-point methods for intractable graphical models*, in *Advanced Mean Field Methods: Theory and Practice*, edited by M. Opper and D. Saad, M.I.T. Press, Cambridge, MA, 2001.