Advanced Mean Field Methods
Theory and Practice
Manfred Opper and David Saad, Eds.
MIT Press, Cambridge, MA, 2001.
300 pp.
ISBN 0-262-15054-9
$40.00

Reviewed by: James M Hogan
School of Computing Science and
Software Engineering,
Queensland University of Technology
GPO Box 2434,
Brisbane, Qld 4001
AUSTRALIA
j.hogan@qut.edu.au

The proliferation of specialised workshops associated with the Advances in Neural Information Processing series of conferences has been of almost unqualified benefit to the scientific community. In keeping with the plurality which has characterised the NIPS community throughout its history, well-chosen workshop themes have nurtured extended interdisciplinary discussion of issues in neural computation, providing just the sort of intellectual scaffolding necessary if the interchange of ideas is to deliver upon its considerable promise.

At this level, the 1999 NIPS Workshop on Advanced Mean Field Methods – organized by Manfred Opper and David Saad - was an outstanding success. The workshop was grounded upon the developing similarity in the problems confronting statistical physicists, computer scientists and statisticians in their exploration of large, coupled, probabilistic systems, and upon a rapidly increasing commonality in their approach. While we shall consider the problems and the range of approximate solutions later, all contributors were in principle concerned with the accurate approximation of quantities derived from multivariate probability distributions over an enormous number of random variables. As the random variables typically exhibit some non-trivial correlation, direct computation of the desired outcomes is intractable for large-scale problems, and the physicists and information scientists share a commitment to replacing the exact joint distribution by one which admits some simplification.

Historically, such problems have arisen in attempts to explain the macroscale behaviour of a substance in terms of the microscale interactions of a massive number of its constituent molecules - in particular through attempts to identify the critical temperatures at which a phase transition may occur. Evidently, explicit calculation of the interactions between molecules cannot be contemplated on such a scale and physicists have long modeled the mutual influence as an effective field, acting independently upon each molecule and thus allowing tractable averaging across the system. This formulation is known as the *classical* or *naïve mean field theory* for the system – the somewhat pejorative label chosen to distinguish the model from the more elaborate, higher order approximations which form the bulk of the subject matter of this book.

More recently, mean field methods have appeared in the context of probabilistic graphical models, described neatly by Michael Jordan as a “general framework for associating joint probability distributions with graphs and for exploiting the structure of the graph in the computation of marginal probabilities and expectations”. While the calculation of such quantities - an aggregation operation over other variables - is straightforward over tree and chain-like structures, exact calculations are intractable in the general case, and approximate methods are again required. While the naïve mean field approach may be employed, probabilistic graphical models differ from the usual physical systems in two key respects - through the focus upon microscale outcomes such as the marginal corresponding to a particular mode; and through a greater inhomogeneity among the random variables of the system - both of which serve to limit the usefulness of the naïve theory. In consequence, and perhaps reflecting the computational background of many of the researchers in this area, a number of dynamic programming and optimization theoretical influences have been brought to bear, leading to the development of belief propagation algorithms and higher order variational approximations to the joint distribution.

These developments have coincided with a wave of advanced mean field studies in physics, commonly based around the Thouless-Anderson-Palmer or TAP approximation, which provides a second order correction to the naïve theory. While superficial mathematical similarities between the work of the physicists and information scientists have been apparent for some time, it appears that the depth of these linkages has been masked to some degree by disparate terminology, notation and even performance criteria. While other questions were posed by the workshop organisers, identification of the precise relationship between these independent results was a key prerequisite for progress, and this workshop an important step in the right direction.

Happily, the efforts of Opper and Saad were rewarded by an outstanding collection of presentations, and as a proceedings volume the book cannot be faulted. In its additional role as a self-contained tutorial, the work is less successful, although even here such flaws as there are lie more in the realm of missed opportunities than gross deficiencies, and the editors have made an admirable attempt to accommodate the novice reader through the inclusion of tutorial material on the TAP approaches and variational methods in graphical models. Nevertheless, while previous exposure to naïve mean field methods is not essential, the material assumes a fair degree of mathematical sophistication, and the title should doubtless be taken as a health warning by those without this background.

The difficult task of orienting the non-specialist to the developing convergence of physical and information science approaches is superbly handled by Michael Jordan’s Foreword, and it is difficult to imagine a better distillation of the core ideas underlying the book. While Jordan is careful to recognize the important contributions from each camp, he nonetheless cannot entirely disguise his loyalties - noting that certain approaches “may have appeal to the physicist, particularly the physicist contemplating unemployment in the modern ‘information economy’…”. More significantly, Professor Jordan provides the key insight that advanced mean field methods have undergone a substantial change in job description, from being a front-line weapon in the struggle for analytic solutions and the associated “hunt for phase transitions” to a position at the core of a new computational methodology. Such a shift in focus presents substantial opportunities for the researcher, through the consequent relaxation or abandonment of some of the more restrictive of the assumptions traditional in statistical mechanics, and the application of optimisation strategies novel in the present domain.

With the stage thus set for integration of the approaches and a focus on the computational consequences of each approximation, the subsequent organization of the material is a little puzzling, with chapters split broadly according to their field of origin. Following the general introduction of chapter 1, chapters 2-9 contain contributions whose roots lie in the statistical physics community, and chapters 10-17 those emerging from the information sciences. While there is some logic in this view, in that the early chapters share the thread of the TAP approximation, the split unnecessarily hinders appreciation of linkages between the approaches. Similarly, there are good arguments for relocating the tutorial on variational methods and graphical models (chapter 10) to follow the statistical physics tutorials of chapters 2 and 3.

Such quibbles notwithstanding, the structure within each half of the book works well, aside from occasional premature assumptions about the reader's knowledge of graphical models. The introduction to the naïve and TAP mean field theories (chapter 2; Opper and Winther) is clear and concise, and linked nicely to subsequent treatments through a focus on Ising spin or Boltzmann machine models (see for example Hertz, Krogh and Palmer (1991)). Two derivations of the naïve theory are presented, each of some importance as a basis for subsequent contributions:

- The variational approach: the intractable joint distribution P is replaced by an approximation Q, drawn from a class of factorisable distributions and chosen so that Q minimises the Kullback-Leibler divergence between Q and P. Under the assumption that P is a Boltzmann-like exponential of some energy function over the spins, the problem reduces to one of minimising the variational free energy.
- The field theoretic approach: expectations involving summation over a large number of discrete variables are replaced by integrations over auxiliary field variables, leading to approximations via Laplace or saddle point methods applied to the integrand.

Similarly thorough treatment is provided of the TAP results, Opper and Winther providing two alternative derivations – with the latter expansion especially useful in subsequent linkages with graphical models:

- The cavity approach: the approximation for the marginal for a particular variable S_i is derived through consideration of the joint distribution which results when this spin is deleted from the system. The resulting TAP equations differ from the naïve theory through the introduction of the *Onsager Reaction Term*, a correction accounting for the reaction of neighbouring spins to the presence of $S_{i.}$
- Plefka's expansion: as in the variational development of the naïve theory, the problem reduces to the minimisation of the variational free energy – only this time Q is not restricted to the class of product distributions, but rather is constrained to deliver some fixed vector m of expectations of the spin variables. In this light, the *Gibbs Free Energy G(m)* may be identified as the constrained minimum of the variational free energy with respect to Q, with minimisation of G with respect to m delivering exact expectations $m=<S>$. Through an elegant perturbation of G(m), the Plefka expansion allows recovery of the naïve theory at first order, and the TAP approximation when truncated at second order.

A number of subsequent chapters consider alternative assumptions about the distribution of couplings between the spins, resulting in novel TAP equations for the model systems. In chapters 5 and 6 (Kabashima and Saad; Saad, Kabashima and Vicente) develop a TAP framework applicable to both intensively and extensively connected systems and explore its application in the context of error-correcting codes. In chapter 7, Opper and Winther provide an adaptive TAP approach, in which the Onsager correction is revised in the light of successive concrete observations of the interactions.

The cavity method is used in Chapter 8 (Wong, Lee and Luo) in the derivation of a general framework for the analysis of batch learning systems, complementing earlier successes by physicists in the analysis of on-line learning systems (Saad, 1998).

Jonathan Yedidia's good-humoured "Idiosyncratic Journey Beyond Mean Field Theory" (chapter 3) explores the physicist's ground with more explicit linkages to graphical models, using the vehicle of a pair-wise Markov network of N nodes. Here, the probability distribution is comprised of a normalised product of two-parameter 'compatibilities' between nodes, and the 'evidence' values for each individual node. In Yedidia's illustration of a medical diagnosis system, the nodes represent symptons and diseases and the 'compatibilities' the statistical dependencies between them. Given the evidence associated with a particular patient, our task might be to infer the probability that the patient has a specific disease. If an approximate value is computed, the relevant marginal probability is usually termed a 'belief', although the latter is similarly constrained.

Within the Markov network, beliefs at a particular node *i* may be characterised as though the node is in receipt of *messages* about its appropriate state from all nodes within a local neighbourhood, the aggregated messages being combined with the independent evidence associated with *i*. Similarly, the *joint* beliefs of two nodes *i and j* may be described in terms of messages from the two surrounding neighbourhoods, combined again with the evidence and the compatibility associated with *i and j*. Algorithms for reasoning within such a framework were first presented by Pearl (1988).

Yedidia shows that such probabilistic structures may be re-cast readily within the statistical physics framework, and a mean field theory obtained through a variational approximation to the Gibbs free energy. Moreover, he reports joint work showing a deep connection – which holds for general Markov networks - between the stationarity conditions for the Bethe approximation to the Gibbs free energy (Bethe, 1935), and belief propagation. Similar ground is covered by Weiss (chapter 15), who makes the important observation that the superiority of belief propagation over naïve mean field methods may be due not only to the sophistication of the free energy, but also to the effectiveness of the algorithm in avoiding the local minima which plague the latter approach. This chapter also treats the – strictly invalid – application of the belief propagation algorithms to loopy graphs, a matter addressed through message attenuation in chapter 14 (Frey and Koetter).

Saddle-point methods are used to deal with intractable belief networks in chapter 9 (Pineda, Resch and Wang) - leading to a novel second order Gaussian approximation – and in chapter 13 (Barber) - in which the message calculations are represented as one dimensional Fourier integrals. This latter representation has substantial computational advantages for directed propagation, whose complexity scales exponentially with the number of parents of a node.

A more elaborate discussion of graphical models is provided in chapter 10 (Jaakkola), with detailed discussion of more general topologies and the two node compatibilities encountered earlier being replaced by potentials defined over each clique. Jaakkola's tutorial is especially valuable, providing a superb introduction to variational approximations in a number of graphical contexts, and leading elegantly into the extensions of subsequent chapters. In particular, variational methods for Bayesian inference are considered at some length in chapters 11 (Ghahramani and Beal) and 12 (Humphreys and Titterington), the former presenting results for classes of exponential models, and the latter using recursive methods to reduce the problem's considerable computational burden.

While all papers are of high quality, perhaps the outstanding evidence of the mutual benefit to be obtained from this interdisciplinary work is provided by the contribution of Kappen and Wiegerinck, (chapter 4) in which an information theoretical approach is used to devise second and higher order approximations for graphical models without the need for a free energy – in essence without the restriction to Boltzmann-Gibbs probability distributions. There are strong linkages between this work and the information geometry approaches of Amari, Ikeda and Shimokawa (chapter 16) and the unified variational treatment of Tanaka (chapter 17), and this area promises a valuable framework for further progress.

In summary, Advanced Mean Field Methods is an excellent collection, providing good value for money and a rich vein of material to be mined and mined again.

References:

Bethe, H.A.,
"Statistical Theory of Superlattices".
Proc. Roy. Society of London, Series A, **151,** 552-575, 1935.

Hertz, J., Krogh, A. and Palmer, R.G.,
An Introduction to the Theory of Neural Computation.
Redwood City, CA. Addison Wesley,1991.

Pearl, J.
Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.
San Francisco, CA. Morgan Kauffman, 1988.

Saad, D., (Ed.),
On-line Learning in Neural Networks.
Cambridge, UK. Cambridge University Press, 1998.