Graphical Models, Exponential Families, and Variational Inference
Editorial Scope

Foundations and Trends® in Machine Learning publishes survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians
Foundations and Trends® in Machine Learning, 2008, Volume 1, 4 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.
Graphical Models, Exponential Families, and Variational Inference

Martin J. Wainwright$^1$ and Michael I. Jordan$^2$

$^1$ Department of Statistics, and Department of Electrical Engineering and Computer Science, University of California, Berkeley 94720, USA, wainwrig@stat.berkeley.edu
$^2$ Department of Statistics, and Department of Electrical Engineering and Computer Science, University of California, Berkeley 94720, USA, jordan@stat.berkeley.edu

Abstract

The formalism of probabilistic graphical models provides a unifying framework for capturing complex dependencies among random variables, and building large-scale multivariate statistical models. Graphical models have become a focus of research in many statistical, computational and mathematical fields, including bioinformatics, communication theory, statistical physics, combinatorial optimization, signal and image processing, information retrieval and statistical machine learning. Many problems that arise in specific instances — including the key problems of computing marginals and modes of probability distributions — are best studied in the general setting. Working with exponential family representations, and exploiting the conjugate duality between the cumulant function and the entropy for exponential families, we develop general variational representations of the problems of computing likelihoods, marginal probabilities and most probable configurations. We describe how a wide variety
of algorithms — among them sum-product, cluster variational methods, expectation-propagation, mean field methods, max-product and linear programming relaxation, as well as conic programming relaxations — can all be understood in terms of exact or approximate forms of these variational representations. The variational approach provides a complementary alternative to Markov chain Monte Carlo as a general source of approximation methods for inference in large-scale statistical models.
Content
4.2 Kikuchi and Hypertree-based Methods 96
4.3 Expectation-Propagation Algorithms 107

5 Mean Field Methods 125
5.1 Tractable Families 125
5.2 Optimization and Lower Bounds 127
5.3 Naive Mean Field Algorithms 132
5.4 Nonconvexity of Mean Field 136
5.5 Structured Mean Field 140

6 Variational Methods in Parameter Estimation 147
6.1 Estimation in Fully Observed Models 147
6.2 Partially Observed Models and Expectation–Maximization 152
6.3 Variational Bayes 158

7 Convex Relaxations and Upper Bounds 165
7.1 Generic Convex Combinations and Convex Surrogates 167
7.2 Variational Methods from Convex Relaxations 170
7.3 Other Convex Variational Methods 182
7.4 Algorithmic Stability 186
7.5 Convex Surrogates in Parameter Estimation 189

8 Integer Programming, Max-product, and Linear Programming Relaxations 195
8.1 Variational Formulation of Computing Modes 195
8.2 Max-product and Linear Programming on Trees 198
8.3 Max-product For Gaussians and Other Convex Problems 202
8.4 First-order LP Relaxation and Reweighted Max-product 204
8.5 Higher-order LP Relaxations 226
Introduction

Graphical models bring together graph theory and probability theory in a powerful formalism for multivariate statistical modeling. In various applied fields including bioinformatics, speech processing, image processing and control theory, statistical models have long been formulated in terms of graphs, and algorithms for computing basic statistical quantities such as likelihoods and score functions have often been expressed in terms of recursions operating on these graphs; examples include phylogenies, pedigrees, hidden Markov models, Markov random fields, and Kalman filters. These ideas can be understood, unified, and generalized within the formalism of graphical models. Indeed, graphical models provide a natural tool for formulating variations on these classical architectures, as well as for exploring entirely new families of statistical models. Accordingly, in fields that involve the study of large numbers of interacting variables, graphical models are increasingly in evidence.

Graph theory plays an important role in many computationally oriented fields, including combinatorial optimization, statistical physics, and economics. Beyond its use as a language for formulating models, graph theory also plays a fundamental role in assessing computational
complexity and feasibility. In particular, the running time of an algorithm or the magnitude of an error bound can often be characterized in terms of structural properties of a graph. This statement is also true in the context of graphical models. Indeed, as we discuss, the computational complexity of a fundamental method known as the junction tree algorithm — which generalizes many of the recursive algorithms on graphs cited above — can be characterized in terms of a natural graph-theoretic measure of interaction among variables. For suitably sparse graphs, the junction tree algorithm provides a systematic solution to the problem of computing likelihoods and other statistical quantities associated with a graphical model.

Unfortunately, many graphical models of practical interest are not “suitably sparse,” so that the junction tree algorithm no longer provides a viable computational framework. One popular source of methods for attempting to cope with such cases is the Markov chain Monte Carlo (MCMC) framework, and indeed there is a significant literature on the application of MCMC methods to graphical models [e.g., 28, 93, 202]. Our focus in this survey is rather different: we present an alternative computational methodology for statistical inference that is based on variational methods. These techniques provide a general class of alternatives to MCMC, and have applications outside of the graphical model framework. As we will see, however, they are particularly natural in their application to graphical models, due to their relationships with the structural properties of graphs.

The phrase “variational” itself is an umbrella term that refers to various mathematical tools for optimization-based formulations of problems, as well as associated techniques for their solution. The general idea is to express a quantity of interest as the solution of an optimization problem. The optimization problem can then be “relaxed” in various ways, either by approximating the function to be optimized or by approximating the set over which the optimization takes place. Such relaxations, in turn, provide a means of approximating the original quantity of interest.

The roots of both MCMC methods and variational methods lie in statistical physics. Indeed, the successful deployment of MCMC methods in statistical physics motivated and predated their entry into
statistics. However, the development of MCMC methodology specifically designed for statistical problems has played an important role in sparking widespread application of such methods in statistics [88]. A similar development in the case of variational methodology would be of significant interest. In our view, the most promising avenue toward a variational methodology tuned to statistics is to build on existing links between variational analysis and the exponential family of distributions [4, 11, 43, 74]. Indeed, the notions of convexity that lie at the heart of the statistical theory of the exponential family have immediate implications for the design of variational relaxations. Moreover, these variational relaxations have particularly interesting algorithmic consequences in the setting of graphical models, where they again lead to recursions on graphs.

Thus, we present a story with three interrelated themes. We begin in Section 2 with a discussion of graphical models, providing both an overview of the general mathematical framework, and also presenting several specific examples. All of these examples, as well as the majority of current applications of graphical models, involve distributions in the exponential family. Accordingly, Section 3 is devoted to a discussion of exponential families, focusing on the mathematical links to convex analysis, and thus anticipating our development of variational methods. In particular, the principal object of interest in our exposition is a certain conjugate dual relation associated with exponential families. From this foundation of conjugate duality, we develop a general variational representation for computing likelihoods and marginal probabilities in exponential families. Subsequent sections are devoted to the exploration of various instantiations of this variational principle, both in exact and approximate forms, which in turn yield various algorithms for computing exact and approximate marginal probabilities, respectively. In Section 4 we discuss the connection between the Bethe approximation and the sum-product algorithm, including both its exact form for trees and approximate form for graphs with cycles. We also develop the connections between Bethe-like approximations and other algorithms, including generalized sum-product, expectation-propagation and related moment-matching methods. In Section 5 we discuss the class of mean field methods, which arise from a qualitatively
different approximation to the exact variational principle, with the added benefit of generating lower bounds on the likelihood. In Section 6, we discuss the role of variational methods in parameter estimation, including both the fully observed and partially observed cases, as well as both frequentist and Bayesian settings. Both Bethe-type and mean field methods are based on nonconvex optimization problems, which typically have multiple solutions. In contrast, Section 7 discusses variational methods based on convex relaxations of the exact variational principle, many of which are also guaranteed to yield upper bounds on the log likelihood. Section 8 is devoted to the problem of mode computation, with particular emphasis on the case of discrete random variables, in which context computing the mode requires solving an integer programming problem. We develop connections between (reweighted) max-product algorithms and hierarchies of linear programming relaxations. In Section 9, we discuss the broader class of conic programming relaxations, and show how they can be understood in terms of semidefinite constraints imposed via moment matrices. We conclude with a discussion in Section 10.

The scope of this survey is limited in the following sense: given a distribution represented as a graphical model, we are concerned with the problem of computing marginal probabilities (including likelihoods), as well as the problem of computing modes. We refer to such computational tasks as problems of “probabilistic inference,” or “inference” for short. As with presentations of MCMC methods, such a limited focus may appear to aim most directly at applications in Bayesian statistics. While Bayesian statistics is indeed a natural terrain for deploying many of the methods that we present here, we see these methods as having applications throughout statistics, within both the frequentist and Bayesian paradigms, and we indicate some of these applications at various junctures in the survey.
References

References


References


References


Full text available at: http://dx.doi.org/10.1561/2200000001
300 References


References


References


Full text available at: http://dx.doi.org/10.1561/2200000001
References

References


Full text available at: http://dx.doi.org/10.1561/2200000001
References


References


References


