Variational Methods for Inference

Dexiong Chen Chia-Man Hung Baoyang Song

January 20, 2017

Abstract

Probabilistic inference problem is a major problem in graphical models. Such problem of computing marginal probabilities and conditional probabilities can be computationally intractable when the structure of the graphical models gets complex. In this report, We explain how a number of algorithms — among them the Loopy Belief Propagation, the Mean Field methods and Variational Bayes — can be understood in terms of approximate forms of these variational representations. Furthermore, we apply variational methods to an Ising model in comparison with Gibbs sampling, a Markov Chain Monte Carlo method.

1 Introduction

One of the most important problems in graphical models is the probabilistic inference. It consists of deducing and computing properties including marginal probabilities or conditional probabilities of an underlying distribution represented as a graphical model. For graphs with simple structure such as trees, the inference problem can be exactly solved by message-passing algorithms taking form of sum-product or max-product algorithm, which explore the conditional independence properties present in the graph and has only linear complexity in the number of nodes. These algorithms can also be extended to arbitrary graphs by using junction tree representation. However, the time complexity will also be increased to be exponential in the size of the maximal clique in the junction tree, which makes the exact computation intractable. Thus, a variety of approximation procedures have been developed and studied. One of the fundamental approaches is to design algorithms involving Monte Carlo methods, referred as Markov Chain Monte Carlo (MCMC). The idea is simply sampling a Markov Chain that converges to the distribution of interest. These approaches possess theoretical guarantee and simple implementation. Nevertheless, sampling methods can be very slow to converge and lack stopping criterion [2].

An alternative methodology for computing approximations of marginal probabilities is based on variational principle, which generally converts a complex problem to a simpler problem by including additional parameters following convex duality and conjugate. The general idea of this approach is to express an intractable quantity as the solution of an optimization problem, then relaxing the optimization problem can simplify the original problem. Various manners of relaxing the optimization problem, approximating either the objective function or the set over which the optimization takes place, lead to different formulations of variational inference, including mean field, loopy sum-product or belief propagation, structured mean field etc.. In this project, we will study these formulations and compare them with the MCMC methods such as Gibbs sampling in an Ising model.

2 Background and Problem Formulation

The principle of maximum entropy [3] convinces us that the majority of distributions belong to the exponential families. Thus, in the following of the section, we will limit our scope to the distributions in an exponential family and establish the variational formulation for this particular family of distributions.

2.1 Notations and Exponential Families

We denote by $G = (V, E)$ the graph representation for a graphical model. Let X be a random vector in $\mathcal{X}$. For a given vector of sufficient statistics $\phi = (\phi_\alpha, \alpha \in \mathcal{I}) \in \mathbb{R}^d$ indexed by $\mathcal{I}$ and an associated vector of canonical parameters, the exponential family associated with ϕ is defined as the following parameterized collection of density functions

$$p(x)d\mu(x) = \exp(\theta^T \phi(x) - A(\theta))d\mu(x), \tag{1}$$

where A is the log-partition function and $\theta \in \Theta$ the canonical parameter. We recall that A is a convex function. We denote by μ the mean parameter or moment vector associated with the sufficient statistics ϕ defined, with respect to a density function p in an exponential family parameterized by θ, by the expectation

$$\mu(\theta) = \mathbb{E}_\theta[\phi(X)]. \tag{2}$$

We denote by $\mathcal{M}$ the marginal polytope consists of all realizable mean parameters

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \,|\, \exists \theta \text{ s.t. } \mathbb{E}_\theta[\phi(X)] = \mu\}. \tag{3}$$

We also recall that if the exponential representation is minimal, i.e. the components of the sufficient statistics are affinely independent, then the gradient map ∇A is one-to-one onto the interior of $\mathcal{M}$, denoted by $\mathcal{M}^o$. By consequence, for a given $\mu \in \mathcal{M}^o$, there exists a unique vector of canonical parameters $\theta(\mu)$ satisfying

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu. \tag{4}$$

2.2 Conjugate Duality

In order to formulate the variational problem for the probabilistic inference, we begin with defining the conjugate duality and exploring the link between the log-partition function and its conjugate dual function. As conjugate duality is a crucial notion in convex analysis, it also helps establish variational representation for the inference. Given a function A, the conjugate dual function of A is defined as

$$A^*(\mu) = \sup_{\theta \in \Theta} (\mu^T \theta - A(\theta)), \tag{5}$$

where $\mu \in \mathbb{R}^n$ is called the dual variable. In a formal way, the following theorem shows the relationship between A and A^* as well as a close expression of A^* when $\mu \in \mathcal{M}^o$

Theorem 2.1. *1. For any $\mu \in \mathcal{M}^o$, denote by $\theta(\mu)$ the unique canonical parameter satisfying the dual matching condition (4). The conjugate dual function A^* takes the form*

$$A^*(\mu) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^o, \\ +\infty & \text{if } \mu \notin \overline{M}. \end{cases} \tag{6}$$

Boundary points are limits of the interior points. Here H is the entropy.

2. *The log-partition function A possesses the variational representation*

$$A(\theta) = \sup_{\mu \in \mathcal{M}} (\theta^T \mu - A^*(\mu)). \tag{7}$$

3. *For all θ, the supremum of this equation is attained uniquely at the vector $\mu \in \mathcal{M}^o$ specified by the moment matching condition*

$$\mu = \mathbb{E}_\theta[\phi(x)]. \tag{8}$$

The formula (7) connects the goal of inference and the conjugate dual function of the log-partition, which corresponds to the standard formulation of the variational inference.

However, the main computational difficulties of this formulation are the lack of an explicit form for the dual function A^* and the nature of the constraint set $\mathcal{M}$. In order to make the problem tractable, we consider either specifying the marginal polytope, the expression of A^* or even restricting the set of distribution. This leads to various formulations for inference problem, some of which will be detailed in the following paragraphs.

3 Variational Methods

3.1 Loopy Belief Propagation

The key idea of loopy belief propagation is to replace $\mathcal{M}$ by a larger set $\mathcal{L}$, the set of *locally consistent* marginal distributions defined by

$$\begin{aligned}\mathcal{L}(G) = \{\tau \geq 0 \,|\, &\sum_{x_i} \tau_i(x_i) = 1, \ \forall i \in V \text{ and} \\ &\sum_{x'_j} \tau_{ij}(x_i, x'_j) = \tau_i(x_i) \ \forall x_i \quad \sum_{x'_i} \tau_{ij}(x'_i, x_j) = \tau_j(x_j) \ \forall x_j \forall (i,j) \in E\}.\end{aligned} \tag{9}$$

On the other hand, we also replace A^* by the negative Bethe entropy approximation

$$-A^*(\tau) \approx H_{\text{Bethe}}(\tau) = \sum_{i \in V} H_i(\tau_i) - \sum_{(i,j) \in E} I_{ij}(\tau_{ij}), \tag{10}$$

corresponding to the entropy of a tree-structured Markov random field. Here, H_i refers to the singleton entropy and I_{ij} refers to the mutual information.

Combining the above two ingredients conducts the exact variational principle (7) to the Bethe variational problem (BVP)

$$\max_{\tau \in \mathcal{L}(G)} \theta^T \tau + \sum_{i \in V} H_i(\tau_i) - \sum_{(i,j) \in E} I_{ij}(\tau_{ij}). \tag{11}$$

By writing the Lagrangian of this optimization problem and the Karush–Kuhn–Tucker (KKT) conditions, we find the sum-product updates. Consequently, sum-product algorithm can be applied to graphs with cycles thanks to this formulation. Furthermore, as the Bethe variational problem is non-convex, there are generally no guarantees that the sum-product algorithm will find the global optimum. Another fruitful consequence is that this connection provides a number of tracks for improving upon the ordinary sum-product algorithm.

3.2 Mean Field Methods

Similar to the Bethe variational approximation, the principle of mean field methods is to limit the optimization in a tractable subset of distributions in such a way that $\mathcal{M}$ and A^* are easy to characterize. Following this idea, we begin with defining the tractable subgraph as follows: a subgraph F of the graph G is tractable if it is feasible to perform exact calculations over it. A simple example is the fully connected subgraph $F_0 = (V, \emptyset)$. For a given tractable subgraph F, we are interested in the collection of all distributions that are Markov with respect to F, i.e. its canonical parameters are restricted only to the cliques of F. We denote this restricted set of the canonical parameters by $\Theta(F)$. The mean field methods are based on replacing $\mathcal{M}$ by the subset of mean parameters associated with the previous subset of canonical parameters $\Theta(F)$, given by

$$M_F(G) = \{\mu \in \mathbb{R}^d \,|\, \exists \theta \in \Theta(F) \quad \mu = \mathbb{E}_\theta[\phi(x)]\}. \tag{12}$$

As a consequence, the variational formulation (7) becomes

$$\max_{\mu \in \mathcal{M}_F(G)} \theta^T \mu - A_F^*(\mu), \tag{13}$$

where $A_F^* = A^*|_{\mathcal{M}_F(G)}$ is the conjugate dual function restricted to the set $\mathcal{M}_F(G)$. The choice of the tractable subgraph yields different formulations of mean field methods.

Naive Mean Field. In this case, F is chosen to be the fully disconnected subgraph. In order to illustrate how mean field works in practice, we concentrate on an Ising model here. For an Ising model, the sufficient statistics are $(x_i, i \in V)$ and $(x_i x_j, (i,j) \in E)$. The associated mean parameters are

$$\mu_i = \mathbb{E}[X_i], \quad \mu_{ij} = \mathbb{E}[X_i X_j]. \tag{14}$$

And we can now compute the explicit form of $\mathcal{M}$ and A^*.

$$\mathcal{M}_F(G) = \{\mu \in \mathbb{R}^{|V|+|E|} \,|\, 0 \leq \mu_i \leq 1 \ \forall i \in V, \text{ and } \mu_{ij} = \mu_i \mu_j \ \forall (i,j) \in E\}. \tag{15}$$

And

$$A_F^*(\mu) = \sum_{i \in V} [\mu_i \log \mu_i + (1 - \mu_i) \log(1 - \mu_i)]. \tag{16}$$

By setting the gradient of the optimization problem equal to zero, we obtain the following coordinate descent update

$$\mu_i \leftarrow \sigma \left(\theta_i + \sum_{j \in N(i)} \theta_{ij} \mu_j \right). \tag{17}$$

Structured Mean Field. More specific and structural choice of subgraph F leads to structured mean field.

3.3 Variational Bayes

In Bayesian formulation, parameters are considered as random variables. As a consequence, they can be estimated via variational inference approach and all formulations we presented previously can be applied successfully to the Bayesian inference [1].

4 Experiments

In this section, we consider applying variational methods for the square-lattice Ising model, consists of the densities

$$p_\theta(x) = \exp\left(\sum_{i\in V}\theta_i x_i + \sum_{(i,j)\in E}\theta_{ij}x_i x_j - A(\theta)\right), \tag{18}$$

where X is a random binary variable. While deriving updating formulas, we implicitly suppose $X \in \{0,1\}$. However, in order to be consistent with statistical physics and also to observe (anti)ferromagnetism, it is important to take $X \in \{-1,1\}$. For simplification, we suppose that $\mu_i j = \mu$ for all $(i,j) \in E$.

Thus far, we studied three methods, namely Gibbs sampling, mean field and Loopy belief propagation. We shall compare them in various ways.

Simulation Results. Figure 3 at the end of report shows states of three models after about 100 iterations ($250^2 \times 100$ steps), from antiferromagnetic to noninteracting and to ferromagnetic. We see a clear *phase transition* from disordered to ordered phase when increasing the intensity of interaction μ_{ij}.

We observe that Gibbs sampling may usually generate some noise in the simulation while variational approaches are generally cleaner without much noise within each region of spins. Another remark is that while increasing the value of the correlation between neighbors, we can more easily observe the interaction between neighbor spins, which suggest that larger mu encourages neighbor spins to align.

Convergence of Damping LBP. Here we focus on the role of the damping parameter in terms of the convergence. Figure 1 shows the approximate log-partition value (11) versus iterations for different values of damping. We observe that when damp $= 1$, the standard LBP oscillates without convergence while the LBP with $damp = 0.9$ converges properly. For smaller damp, the LBP may suffer from the local optimum. These observations are conform to the fact that the BVP is non-convex. Additionally, adding the damp parameter is exactly a common technique used in stochastic descent method in order to accelerate and stabilize the algorithm, which is referred to as the momentum.

Comparison of Mean Field and LBP. Figure 2 compares the convergence of mean field and LBP in terms of (approximate) log-partition function. As expected (because of equation (7)), for the mean field method, the log-partition increases and converges rather quickly. The approximate log-partition of loopy belief propagation, on the other hand, has very little fluctuation.

5 Conclusion

In this report, we have investigated the variational method, an alternative to MCMC for computing approximations of marginal probabilities over complex distributions. Based on a fundamental formulation, various variational formulations have been studied, including LBP, damping LBP, mean field and structured mean field. These formulations differ from the choice of the constraint set and the form of the conjugate

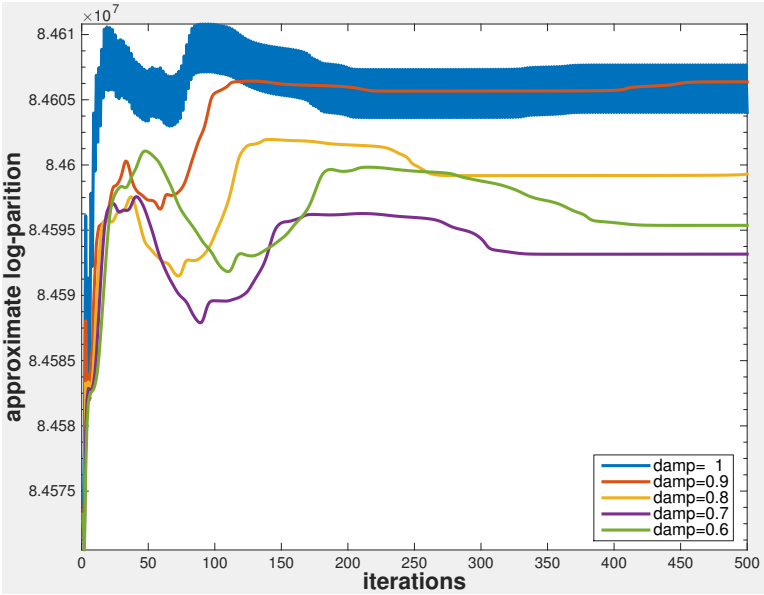

Figure 1: Convergence of damping LBP with different values of damp.

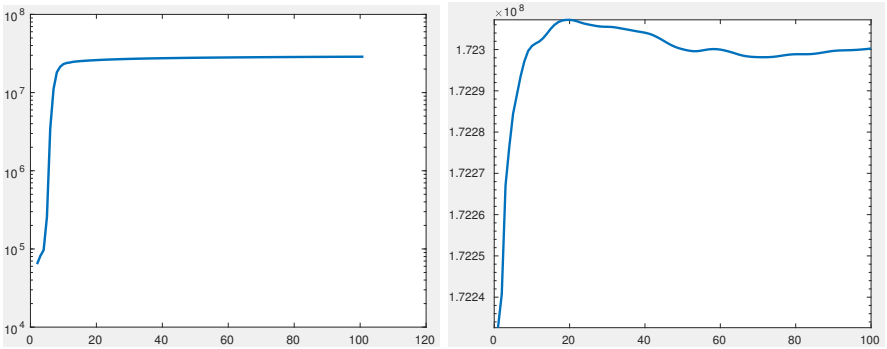

Figure 2: Convergence of mean field and LBP (damp= 0.8). Left: mean field, right: LBP.

dual function of the log-partition. We have applied these methods to simulate the Ising model. We have studied the convergence of each method as well as the role of the damping parameter in LBP.

References

[1] M. J. Beal, Z. Ghahramani, et al. Variational bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 1(4):793–831, 2006.

[2] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[3] N. Wu. *The maximum entropy method*, volume 32. Springer Science & Business Media, 2012.

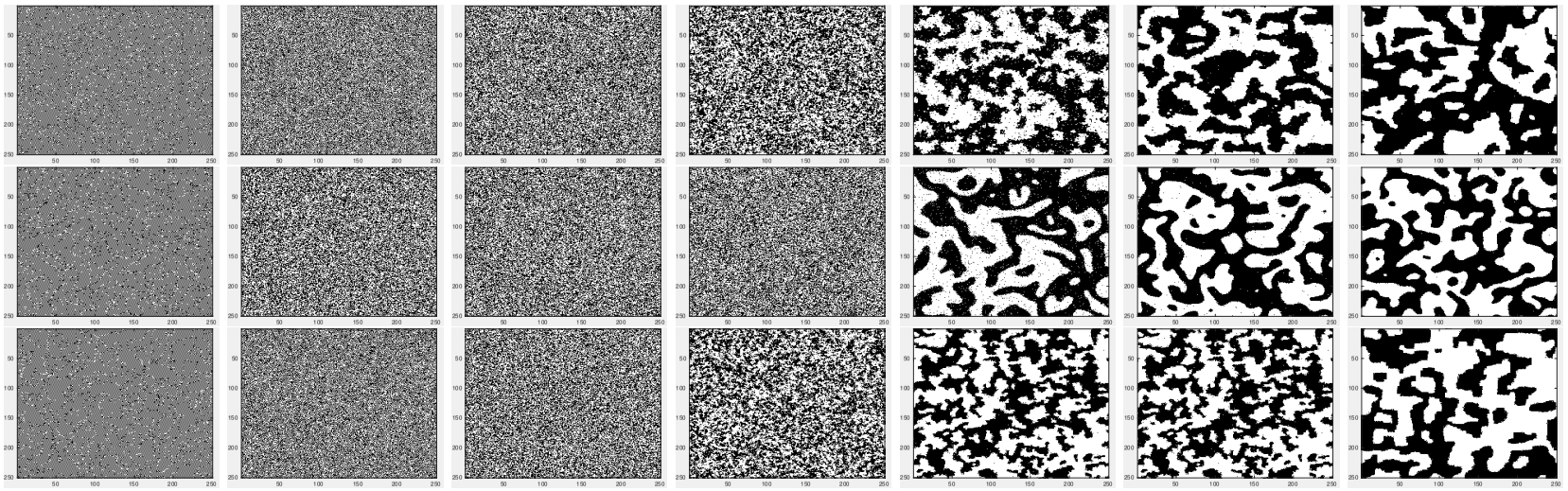

Figure 3: Ising model on a 250×250 grid. From top to bottom: Gibbs sampling, Mean-field, LBP (damp $= 0.8$). From left to right: $\mu_{ij} = -1, -0.5, 0, 0.5, 1, 1.5, 2$.