Variational Methods for Inference

Dexiong Chen, Chia-Man Hung, Baoyang Song

Probabilistic Graphical Models, Master 2 MVA, ENS Cachan

In a Nutshell

We present the variational methods for inference in graphical models, a class of approximation techniques arising from the calculus of variations and convex analysis to the optimization-based formulations of problems. We study and build on existing links between variational analysis and exponential families. Based on these basic links, we present different formulations for variational inference.

Motivation

Probabilistic inference 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. 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 [1].

An alternative methodology for statistical inference is based on variational methods. 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..

Fundamental Theorem for Inference

Assume that the distribution of interest p is in an exponential family $q_\theta(x) = \exp(\theta^T\phi(x) - A(\theta))$ represented as a graphical model G. The convexity of the log partition function A provides the following connection between A and its conjugate dual function A^*

- The log partition function has the variational representation
$$A(\theta) = \sup_{\mu\in\mathcal{M}} (\theta^T\mu - A^*(\mu)), \tag{1}$$
where $\mathcal{M}$ is the marginal polytope.
- 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
$$\mu = \mathbb{E}_\theta[\phi(x)], \tag{2}$$
which is the goal of inference.

Main difficulties:

- The nature of the constraint set $\mathcal{M}$.
- The lack of an explicit form for the dual function A^*.

Loopy Belief Propagation

- Replace $\mathcal{M}$ by a larger set $\mathcal{L}$, the set of *locally consistent* marginal distributions
$$\mathcal{L}(G) = \{\tau \geq 0 \,|\, \sum_{x_i} \tau_i(x_i) = 1,\ \forall i \in V \text{ and}$$
$$\sum_{x'_j} \tau_{ij}(x_i, x'_j) = \tau_i(x_i)\ \forall x_i \quad \sum_{x'_i} \tau_{ij}(x'_i, x_j) = \tau_j(x_j)\ \forall x_j$$
$$\forall (i,j) \in E\} \tag{3}$$
- Replace A^* by the negative Bethe entropy approximation
$$-A^*(\tau) = H_{\text{Bethe}}(\tau) = \sum_{i\in V} H_i(\tau_i) - \sum_{(i,j)\in E} I_{ij}(\tau_{ij}). \tag{4}$$
- By combining the above two ingredients, 1 becomes the Bethe variational problem (BVP)
$$\max_{\tau\in\mathcal{L}(G)} \theta^T\tau + \sum_{i\in V} H_i(\tau_i) - \sum_{(i,j)\in E} I_{ij}(\tau_{ij}). \tag{5}$$
- By writing its optimal condition, we obtain the sum-product updates.

Mean Field Methods

- **Idea.** Limit the optimization in a tractable subset of distributions in such a way that $\mathcal{M}$ and A^* are easy to characterize.
- **Tractable subgraph.** A subgraph F of the graph G is tractable if it is feasible to perform exact calculations over it.
- **Naive mean field algorithm.** In this case, F is the fully disconnected subgraph, which contains all the vertices but none of the edges.
For the 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
$$\mu_i = \mathbb{E}[X_i], \quad \mu_{ij} = \mathbb{E}[X_i X_j]. \tag{6}$$
And we can now compute the explicit form of $\mathcal{M}$ and A^*.
$$\mathcal{M}_F(G) = \{\mu \in \mathbb{R}^{|V|+|E|} \,|\, 0 \leq \mu_i \leq 1\ \forall i \in V, \text{ and}$$
$$\mu_{ij} = \mu_i\mu_j\ \forall (i,j) \in E\}. \tag{7}$$
And
$$A^*_F(\mu) = \sum_{i\in V} [\mu_i \log \mu_i + (1-\mu_i)\log(1-\mu_i)]. \tag{8}$$
Then we can optimize the approximate problem and obtain the following update
$$\mu_i \leftarrow \sigma(\theta_i + \sum_{j\in N(i)} \theta_{ij}\mu_j). \tag{9}$$
- **Structured mean field.** More specific and structural choice of subgraph F leads to structured mean field.

Simulation and Results

We compare different methods for Ising model on a graph $G = (V, E)$ consisting of the densities
$$p_\theta(x) = \exp(\sum_{i\in V} \theta_i x_i + \sum_{(i,j)\in E} \theta_{ij} x_i x_j - A(\theta)), \tag{10}$$
with X a binary random variable. For simplicity, we suppose $\theta_i = \theta_{ij} = \theta_{ji} := \theta$ in all our simulations, but our method is not limited to such case.

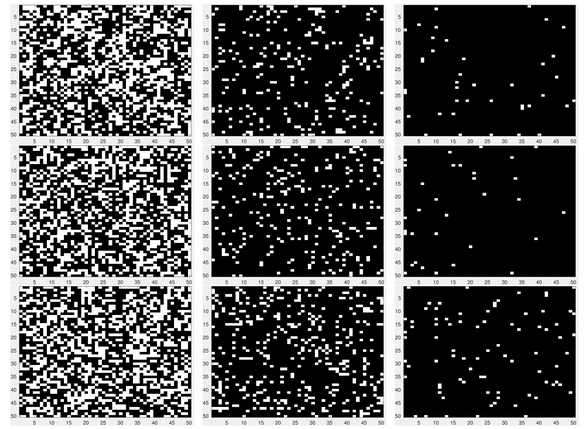

Figure: Gibbs sampling, mean field and loopy belief propagation for $\theta = 0.1, 0.5$ and 0.9

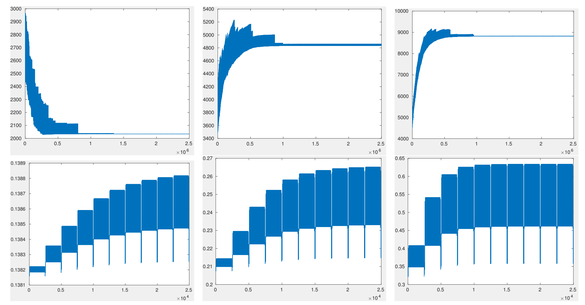

Figure: Convergence of A. Top: mean field method; Bottom: loopy belief propagation. From left to right: $\theta = 0.1, 0.5$ and 0.9.

Conclusion

Variational methods can be successfully applied to Ising model and provide an efficient computation of marginal probability in practice. Besides the probabilistic inference problem, it can also be used to solve learning problems, under a setting of Bayesian inference, where parameters are viewed as random parameters [2].

References

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

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