A Study Report on Variational Principle for Graphical Models [3]*

Xinyuan Zheng xz2906

December 2020

*Final project for STAT5293 Probablistic Graphical Models, instructor: Prof. Ori Shental, 2020 Fall. This report is a study for Wainwright and Jordan's work.[3]

1 Introduction

Probabilistic graphical models are powerful tools for multivariate modelling. To solve the computation problems for graphical models, a class of methods including belief propagation and mean-field algorithms based on variational principle are developed.

The aim of the this report is to provide a mathematical framework to this class of algorithms, specially, this report formalises the likelihood calculation as an optimization problem by specifying the cost functions and constraints in the setting of exponential graphical models.

The report is organized as follows: The second section discusses graphical model and its exponential form. Section 3 introduces the variational principle. Section 4 provides an example to illustrate exact inference in variational form and Section 5 focuses on approximate inference.

2 Graphical Models in Exponential Form

2.1 Graphical models

A graphical model is a probabilistic model which has a underlying graph defining its variables' conditional dependence, the model then can be factorized according to the graph. Consider a graphical model G = (V, E). The collection of vertices is denoted as V, and the collection of edges E. For each vertex, there is an attached random variable $x_s \in X_s$, whose dependence structure is determined by E.

For an undirected graphical model (Markov random field, MRF), it can be factorised based on its fully-connected subsets, i.e. cliques.

$$p(\mathbf{x}) = \frac{1}{Z} \prod \psi_C(x_C) \tag{1}$$

where ψ_C is the compatibility function for each clique C, Z is the normalizing constant.

2.2 Inference problems

The problems we care about for graphical models, are usually likelihood computation, conditional density computation and finding the mode over the whole distribution.

In our class, to find exact answers to this problems, we discussed sum-product and max-product algorithms. Later we will show that these inference problems can be re-pharsed in a more generalised setting and the algorithms can be understood from an optimization perspective. To see that, the next subsection formally introduces exponential family.

2.3 Exponential family in generality

An exponential family is a special set of probability distributions which can be written in the following form:

$$p(\mathbf{x};\theta) = exp\{\langle\theta, \phi(\mathbf{x})\rangle - A(\theta)\} \tag{2}$$

where the the quantity A, known as the cumulant generating function or the log partition function, is defined by:

$$A(\theta) = log \int_{X^n} exp\langle\theta, \phi(x)\rangle\nu(d\mathbf{x}) \tag{3}$$

θ is called the canonical parameter, and ϕ is called the sufficient statistics. Presuming $\int_{X^n} exp\langle\theta, \phi(x)\rangle\nu(d\mathbf{x})$ is finite, we ensure $p(\mathbf{x};\theta)$ (2) is a valid probability density:

$$p(\mathbf{x};\theta) = exp\{\langle\theta, \phi(\mathbf{x})\rangle - log \int_{X^n} exp\langle\theta, \phi(x)\rangle\nu(d\mathbf{x})\} = \frac{exp\langle\theta, \phi(\mathbf{x})\rangle}{\int_{X^n} exp\langle\theta, \phi(x)\rangle\nu(d\mathbf{x})}$$

$$\Longrightarrow \int_{X^n} p(\mathbf{x};\theta)\nu(d\mathbf{x}) = 1$$

Lemma 2.1. The cumulant generating function A is a convex function in terms of θ, A is infinitely differentiable on $\Theta := \{\theta \in \mathbb{R}^d | A(\theta) < \infty\}$ and its derivatives correspond to cumulants.

$$\frac{\partial A}{\partial \theta_\alpha} = \int_{X^n} \phi_\alpha(\mathbf{x}) p(\mathbf{x};\theta)\nu(d\mathbf{x}) = \mathbb{E}_\theta[\phi_\alpha(\mathbf{x})] \tag{4}$$

We call $\mu := \mathbb{E}_\theta[\phi_\alpha(\mathbf{x})]$ the mean parameters.

Example 2.1. Here we find out that multinomial Markov random fields (MRF), such as the examples shown in class note 3, can be written in the form of equation (2) and therefore belong to the exponential family:

For a MRF over $\mathbf{x} = \{x_s | s \in V\}$, each x_s is a multinomial random variable.

Let $\mathbb{I}_j(x_s)$ and $\mathbb{I}_{jk}(x_s, x_t)$ be indicator functions for the events $\{x_s = j\}$ and $\{(x_s, x_t) = (j, k)\}$ respectively. The set of sufficient statistics then is $\{\mathbb{I}_j(x_s) | s \in V, j \in X_s\} \cup \{\mathbb{I}_j(x_s)\mathbb{I}_k(x_t) | (s,t) \in E, (j,k) \in X_s \times X_t\}$; and the corresponding canonical parameter has the form $\theta = \{\theta_{s;j} | s \in V, j \in X_s\} \cup \{\theta_{st;jk} | (s,t) \in E, (j,k) \in X_s \times X_t\}$.

The multinomial MRF then can be written as:

$$p(\mathbf{x};\theta) = exp\{\sum_{s\in V}\sum_{j\in X_s} \theta_{s;j}\mathbb{I}_j(x_s) + \sum_{(s,t)\in E}\sum_{(j,k)\in X_s\times X_t} \theta_{st;jk}\mathbb{I}_j(x_s)\mathbb{I}_k(x_t) - A(\theta)\} \tag{5}$$

where

$$A(\theta) := log \sum_{\mathbf{x}\in X^n} exp\{\sum_{s\in V}\sum_{j\in X_s} \theta_{s;j}\mathbb{I}_j(x_s) + \sum_{(s,t)\in E}\sum_{(j,k)\in X_s\times X_t} \theta_{st;jk}\mathbb{I}_j(x_s)\mathbb{I}_k(x_t)\}$$

In particular, the equation (5) that $X_s = \{0, 1\}$ for all $s \in V$, is known as the Ising model.

3 Exact Variational Principle

With the setup in Section 2.3, we can rewrite some of the graphical models in exponential form, thereby re-phrasing the inference problems in Section 2.2 as:

- Computing $A(\theta)$
- Computing $\mu := \mathbb{E}_\theta[\phi(\mathbf{x})]$

In particular, one can see that in Example 2.1, to compute the mean parameters is equivalent to compute the marginal distributions. These two problems will be represented as optimization problems in this section and investigated for the rest of the report.

3.1 Conjugate duality

The theory of conjugate duality for convex optimization problems [1] will be useful, this subsection very briefly introduces it:

Let f be a proper convex function $f : \mathbb{R}^d \to \mathbb{R}_*$, $\mathbb{R}_* := \mathbb{R} \cup \{+\infty\}$.
Then $f^* : \mathbb{R}^d \to \mathbb{R}_*$ is a convex conjugate of f, defined as:

$$f^*(y) := sup_{x\in\mathbb{R}^d}\{\langle y, x\rangle - f(x)\}. \tag{6}$$

An important note on the dual function is, if f meets the technical condition of lower semi-continuous, taking the dual of a dual recovers itself, that is:

$$f(x) = sup_{x\in\mathbb{R}^d}\{\langle y, x\rangle - f^*(x)\}. \tag{7}$$

3.2 Realizable mean parameters

Apply the conjugacy discussed in the last subsection to the cumulant generating function A (3), we have

$$A^*(\mu) := sup_{\theta\in\Theta}\{\langle\theta, \mu\rangle - A(x)\}. \tag{8}$$

Take the derivative of equation (8) with respect to θ and set it to 0, we get

$$\mu = \nabla A(\theta) \tag{9}$$

By Lemma 2.1, we know that

$$\nabla A(\theta) = \mathbb{E}_\theta[\phi(\mathbf{x})] \tag{10}$$

Therefore,

$$\mu = \mathbb{E}_\theta[\phi(\mathbf{x})] \tag{11}$$

and we want to determinate the set of $\mu \in \mathbb{R}^d$ satisfying equation (11). Consequently, we define the set of *globally realizable mean parameter* as

$$\mathcal{M} := \{\mu \in \mathbb{R}^d | \; \exists \, p(\cdot) \; such \; that \int \phi(\mathbf{x}) p(\mathbf{x}) \nu(d\mathbf{x}) = \mu\} \tag{12}$$

3.3 Entropy

The form of conjugate dual is related to the entropy function. Consider a density function p, its entropy be definition is:

$$H(p) = -\int_{X^n} p(\mathbf{x}) log[p(\mathbf{x})] \nu(d\mathbf{x}) = -\mathbb{E}_p[log \, p(\mathbf{x})] \tag{13}$$

If $\mu \in \mathcal{M}$, then

$$\exists \, p(\cdot) \; such \; that \int \phi(\mathbf{x}) p(\mathbf{x}) \nu(d\mathbf{x}) = \mu$$

$$\implies \; \exists \, \theta(\mu) \in \Theta \; such \; that \; \mathbb{E}_p[\phi(\mathbf{x})] = \mu$$

We say μ and $\theta(\mu)$ are dually coupled, and by substituting $\mathbb{E}_p[\phi(\mathbf{x})] = \mu$ into the dual of cumulant generating function (8), we have

$$A^*(\mu) = \langle \mu, \theta(\mu) \rangle - A(\theta(\mu)) = \mathbb{E}_p[log \, p(\mathbf{x}; \theta(\mu))].$$

$$\implies A^*(\mu) = -H(p), \; \mu \in \mathcal{M}$$

Remark. The intuition behind the optimization problem can be thought as maximum entropy: Assume that the expected values $\mathbb{E}[\phi_\alpha(\mathbf{x})] = \mu_\alpha$ is known for a collection of functions ϕ_α, to infer a full probability distribution based on the observed means, we wish to choose the distribution satisfying the observed means while maximsing the uncertainty. In mathematical terms:

$$p_{ME} := arg \, max_{p \in \mathcal{P}} H(p) \qquad subject \; to \; constraints \quad \mathbb{E}[\phi_\alpha(\mathbf{x})] = \mu_\alpha.$$

3.4 Exact variational principle

To summarise the above findings, the dual function of A can be written as

$$A^*(\mu) = \begin{cases} -H(p(\mathbf{x}; \theta(\mu))) & \mu \in \mathcal{M} \\ +\infty & \mu \notin \mathcal{M} \end{cases} \tag{14}$$

Give this form of A^*, by the property of dual described in equation (7), we are now able to express the cumulant generating function A as an optimization problem

$$A(\theta) := sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \tag{15}$$

Remark. The supremum is taken over the set $\mathcal{M}$ of realizable means.
The nature of this framework guarantees the optimum is achieved at $\mu = \mathbb{E}_\theta[\phi(\mathbf{x})]$, therefore the full set of mean parameters are solved simultaneously while solving for the cumulant generating function.

4 Exact Inference in Variational Form

This section considers one important case, namely multinomial MRF on a tree, using the variational principle derived above.

The sum-product algorithm for tree-structured problems is discussed.

4.1 Junction tree representation

In class, we define a *tree* as a singly connected undirected graph, and a *clique tree* as an acyclic graph whose nodes are defined over several variables, holding family preserving and running intersection property.

Here in the following context, we particularly define a *clique tree* as an acyclic graph with nodes formed by cliques of a graph, and a *junction tree* as a *clique tree* which satisfies the running intersection property.

For a junction tree, let $\mathcal{C}$ denote the set of maximal cliques and $\mathcal{S}$ denote the set of separator sets. The distribution of a junction tree then can be factorized in terms of its marginals:

$$p(\mathbf{x}) = \frac{\prod_{C\in\mathcal{C}} \mu_C(x_C)}{\prod_{S\in\mathcal{S}}[\mu_S(x_S)]^{d(S)-1}} \tag{16}$$

4.2 Exact inference in trees

Consider a multinomial Markov random field constructed in Example 2.1. The mean parameters in this example are local marginal probabilities with the choice of sufficient statistics $\{\mathbb{I}_j(x_s)|s \in V, j \in X_s\} \cup \{\mathbb{I}_j(x_s)\mathbb{I}_k(x_t)|(s,t) \in E, (j,k) \in X_s \times X_t\}$.

Formally,

$$\begin{aligned} &\mu_{s;j} := p(x_s = j;\theta) \quad \forall s \in V; \\ &\mu_{st;jk} := p((x_s, x_t) = (j,k);\theta) \quad \forall (s,t) \in E \end{aligned} \tag{17}$$

Define the above mean parameters in functional forms,

$$\begin{aligned} \mu_s(x_s) &:= \sum_{j\in X_s} \mu_{s;j}\mathbb{I}_j(x_s); \\ \mu_{st}(x_s,x_t) &:= \sum_{(j,k)\in X_s\times X_t} \mu_{st;jk}\mathbb{I}_{jk}(x_s,x_t) \end{aligned} \tag{18}$$

By definition of the set $\mathcal{M}$, it consists of all realizable singleton marginals μ_s and pairwise marginals μ_{st} on X^n. Since in this case n is a finite number, $\mathcal{M}$ is the convex hull of finite vectors and therefore a *polytope*. As $\mathcal{M}$ corresponds to a marginal polytope, we denote it by MARG(G) in discrete case.

Remark. Characterizing the marginal polytope exactly and solving a linear program over marginal polytope can be intractable computationally in general.

We also define the local constraint set LOCAL(G) for the mean parameters meeting the non-negative, normalization and pairwise marginalization conditions:

$$LOCAL(G) := \{\mu \geq 0 \mid \sum_{x_s} \mu_s(x_s) = 1, \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s) \ \forall (s,t) \in E\} \tag{19}$$

As the marginals must statisfy the above conditions,

$$MARG(G) \subseteq LOCAL(G) \ \forall \ G$$

Remark. When T is a tree-structured graph, MARG(T) = LOCAL(T). See Wainwright and Jordan [2] for a proof.

Consider the junction tree framework (16), T can be factorized in the following way:

$$p(\mathbf{x}; \mu) = \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}$$

Based on above, the dual function A^* of T can be written as a function of mean parameters:

$$-A^*(\mu) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}) \tag{20}$$

where the singleton entropy $H_s(\mu_s)$ is defined as $H_s(\mu_s) := -\sum_{x_s} \mu_s(x_s) log \mu_s(x_s)$, and the mutual information $I_{st}(\mu_{st}) := \sum_{x_s . x_t} \mu_{st}(x_s, x_t) log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}$.

Applying the property of dual function (7), we come to the form:

$$A(\theta) = max_{\mu \in LOCAL(T)} \{\langle \theta, \mu \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st})\} \tag{21}$$

4.3 Sum-product Algorithm

Sum-product algorithm, also known as belief propagation, is a message-passing algorithm for graphical models.

In class, the sum-product algorithm and its performance on clique trees are discussed, we revisit it here: Consider a tree $T = (V, E)$. For any $s \in V$, the

set of its neighbors is defined as $\mathcal{N}(s) = \{u \in V | (s,u) \in E\}$. For any $u \in \mathcal{N}(s)$, $T_u = (V_u, E_u)$ is the subgraph formed by the nodes which can be reached from u without passing through s. Also the nodes in a subtree V_t are denoted by $\{x_u | u \in V_t\}$.

Using these notations, the sum-product algorithm in lecture note 5 can be described in the following way:

$$\mu_s(x_s) \propto \psi_s(x_s) \prod_{t \in \mathcal{N}(s)} \sum_{\{x'_{T_t} | x'_s = x_s\}} \psi_{st}(x_s, x'_t) p(x'_{T_t}; T_t). \tag{22}$$

Denote $\sum_{\{x'_{T_t} | x'_s = x_s\}} \psi_{st}(x_s, x'_t) p(x'_{T_t}; T_t)$ as $M^*_{ts}(x_s)$, the update rule can be written as:

$$M_{ts}(x_s) \leftarrow \kappa \sum_{x'_t} \{\psi_{st}(x_s, x'_t) \psi_t(x'_t) \prod_{u \in \mathcal{N}(t)/s} M_{ut}(x'_t)\}, \tag{23}$$

where κ is a constant chosen to satisfy normalization conditions.

5 Approximate Inference in Variational Form

The findings on exact variational formulation discussed above implicitly make two assumptions:

- The set of globally realizable mean parameters $\mathcal{M}$ can be characterized;
- and the dual function A^* has an explicit form.

However, in real life neither these two assumptions hold generally, which motivates the approximations to $\mathcal{M}$ and A^*. A class of algorithms are based on this strategy. For instance, the mean field method uses a non-convex inner bound on $\mathcal{M}$ to approximate $\mathcal{M}$; loopy belief propagation (sum-product algorithm) makes use of polyhedral outer bound. This section will focus on the loopy form of sum-product algorithm.

5.1 Bethe entropy approximation

Although a tree-structured distribution has a closed form expression (20) for the dual function A^*, a graph with cycles does not have this property. This subsection introduces Bethe approximation to the entropy on a graph with cycles based on the expression (20):

$$-A^*(\mu) \approx H_{Bethe}(\mu) := \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}) \tag{24}$$

Here $H_{Bethe}(\mu)$ provides an approximation for $-A^*(\mu)$.

Consider the constraint set (19), which is the marginal polytope for a tree. Denote the set of marginals in the multinomial MRF from Example 2.1 as $\tau_s(x_s)$ and $\tau_{st}(x_s, x_t)$, we apply (19) as an outer bound for general graphs with cycles, e.g. a multinomial MRF:

$$LOCAL(G) = \{\tau \geq 0 \mid \sum_{x_s} \tau_s(x_s) = 1, \sum_{x_s} \tau_{st}(x_s, x_t) = \tau_t(x_t)\}. \tag{25}$$

Therefore we define the *Bethe variational problem* (BVP) over the constraint set LOCAL(G):

$$max_{\tau \in LOCAL(G)}\{\langle \theta, \tau \rangle + \sum_s H_s(\tau_s) - \sum_{(s,t)\in E} I_{st}(\tau_{st})\} \tag{26}$$

Remark. The optimum τ^* of the above expression may not be consistent with any distribution, as LOCAL(G) is a strict outer bound on MARG(G). As a result, we will see the sum-product algorithm based on BVP is approximating the entropy function by enlarging the constraint set.

To solve the problem of (26), the method of Lagrange multipliers is used. For any variable $x_s \in X_s$, associated with constraint $C_{ts}(x_s) = 0$, where $C_{ts}(x_s) := \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s, x_t)$, a Lagrange multiplier is denoted by $\lambda_{st}(x_s)$. Given this notation, to solve BVP is equivalent to solve:

$$\mathcal{L}(\tau; \lambda) := \langle \theta, \tau \rangle + H_{Bethe}(\tau) + \sum_{(s,t)\in E} [\sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t)] \tag{27}$$

With the above setup, the sum-product algorithm can be written as a Lagrangian method to solve BVP [2]:

Any fixed point of the sum-product updates specifies a pair (τ^*, λ^*) s.t.

$$\nabla_\tau \mathcal{L}(\tau^*; \lambda^*; \theta) = 0, \quad \nabla_\lambda \mathcal{L}(\tau^*; \lambda^*; \theta) = 0 \tag{28}$$

For a tree-structured MRF, it can be shown that (28) has a unique solution (τ^*, λ^*), where τ^* corresponds to the exact singleton and pairwise marginal distributions and the cumulant generating function $A(\theta)$ is found while the optimal value of BVP is attained.

Computing the partial derivatives in (28) yields,

$$\log \tau_s(x_s) = \lambda_{ss} + \theta_s(x_s) + \sum_{t \in N(s)} \lambda_{ts}(x_s) \tag{29a}$$

$$\log \frac{\tau_{st}(x_s, x_t)}{\sum_{x_t} \tau(x_s, x_t) \sum_{x_s} \tau(x_t, x_s)} = \theta_{st}(x_s, x_t) - \lambda_{ts}(x_s) - \lambda_{st}(x_t) \tag{29b}$$

and

$$\mathcal{C}_{ts}(x_s;\tau) = 0; \quad \mathcal{C}_{ss}(\tau) = 0 \tag{30}$$

Using (29b) and (30), (29a) can be rearranged as:

$$\begin{aligned}\log \tau_{st}(x_s, x_t) = \lambda_{ss} + \lambda_{tt} + \theta_{st}(x_s, x_t) + \theta_s(x_s) + \theta_t(x_t) \\ + \sum_{u\in N(s)\setminus t} \lambda_{us}(x_s) + \sum_{u\in N(t)\setminus s} \lambda_{ut}(x_t).\end{aligned} \tag{31}$$

To see the connection with (23), define the message for edge $t \to s$,

$$M_{ts}(x_s) := \exp(\lambda_{ts}(x_s)).$$

Equation (29a) and (31) can be written as:

$$\tau_s(x_s) = \kappa \exp(\theta_s(x_s)) \prod_{t\in N(x)} M_{ts}(x_s). \tag{32}$$

$$\begin{aligned}\tau_{st}(x_s, x_t) =& \kappa' \exp(\theta_{st}(x_s, x_t) + \theta_s(x_s) + \theta_t(x_t)) \\ &\times \prod_{u\in N(s)\setminus t} M_{us}(x_s) \prod_{u\in N(t)\setminus s} M_{ut}(x_t).\end{aligned} \tag{33}$$

where κ, κ' are normalization constants.

Using and rearranging (32) and (33) yields,

$$M_{ts}(x_s) \propto \sum_{x_t} [\exp\{\theta_{st}(x_s, x_t) + \theta_t(x_t)\} \prod_{u\in N(t)\setminus s} M_{ut}(x_t)] \tag{34}$$

which is equivalent to the sum-product update rule.

Remark. The sum-product algorithm does not guarantee to converge on graphs with cycles.

5.2 Kikuchi and hypertree-based methods[2]

This section will introduce hypergraphs and hypertrees, and briefly touch on hypertree-based methods.

5.2.1 Hypertrees

A hypergraph $G = (V, E)$ is a generalization of a graph, in which each *hyperedge* E connects a set of vertices in V. In other words, normal graph edges connect a pair of nodes while hyperedges in hypergraphs can be arbitrary sets of nodes. Hypertrees are acyclic hypergraphs.

With the background of hypertrees, another factorization for any hypertree-structed graph can be provided:

Let $\mu = (\mu_h, h \in E)$ denote the set of marginals associated with hyperedges E, $\omega : E \times E \to \mathbb{R}$ denote a Mobius function.

$$\log \varphi_h(x_h) := \sum_{g \subseteq h} \omega(g, h) \log \mu_g(x_g)$$

$$\implies \log \mu_h(x_h) = \sum_{g \subseteq h} \log \varphi_g(x_g) \tag{35}$$

Particularly, the distribution of a hypertree can be factorized as below, if E contains the intersections between hyperedges which are not contained within other hyperdeges:

$$p_\mu(x) = \prod_{h \in E} \varphi_h(x_h; \mu). \tag{36}$$

Also define the hyperedge entropy and multi-information as the following respectively:

$$H_h(\mu_h) := \sum_{x_h} \mu_h(x_h) \log \mu_h(x_h); \tag{37}$$

$$I_h(\mu_h) := \sum_{x_h} \mu_h(x_h) \log \varphi_h(x_h) \tag{38}$$

As a result, the entropy follows from (36) and (38) can be expressed as:

$$H_{hyper}(\mu) = -\sum_{h \in E} I_h(\mu_h) \tag{39}$$

5.2.2 Kikuchi method

The basic idea of Kikuchi method is to extend the tree-based Bethe approximation using hypertrees. Now consider a MRF which has a underlying hypergraph $G = (V, E)$:

$$p_\theta(x) \propto \exp\{\sum_{h \in E} \theta_h(x_h)\} \tag{40}$$

Similarly, let $\tau = \tau_h$ be the set of local marginals. The constraint set for the mean parameters meeting the non-negative, normalization and marginalization conditions is defined:

$$LOCAL_t(G) = \{\tau \geq 0 \mid \sum_{x'_h} \tau_h(x'_h) = 1 \ \forall h, \sum_{\{x'_h | x'_g = x_g\}} \tau_h(x'_h) = \tau_g(x_g) \ \forall g \subset h\} \tag{41}$$

which is a natural generalization of (19).

Also, the hypertree-based approximation to the entropy

$$H_{Kikuchi}(\tau) = \sum_{g \in E} \sum_{f \supseteq g} \omega(g, f) H_g(\tau_g) \tag{42}$$

is an analogy to Bethe entropy approximation.

Remark. $c(g) := \sum_{f \supseteq g} \omega(g, f)$ is called the *overcounting number*.

Based on (42), we have:

$$max_{\tau \in LOCAL_t(G)} \{\langle \theta, \tau \rangle + H_{Kikuchi}(\tau)\}. \tag{43}$$

which is a hypertree-based generalization to BVP.

Remark. To solve variational problem of equation (43), a class of Lagrangian-based message-passing methods called generalized belief propagation (GBP) are developed.[4] These approaches can be thought as a generalization of the ordinary sum-product updates.[2]

6 Discussion

This report summaries Wainwright and Jordan's work on "A Variational Principle for Graphical Models"[3], illustrates it on the multinomial MRF case, discusses sum-product algorithm and Bethe and Kikuchi methods.

The methods discussed in this report are exclusively applicable to the exponential family of models. Further work could explore the study of variational principle for the non-exponential families, e.g., nonparametric Bayesian distributions.[2]

The research on variational principle is interesting both in theory and in practice, however due to limited knowledge of the author, this report covers the topic in a superficial manner. Also due to space and time constraints, proof is largely skipped, I would like to refer the reader to the well-written papers in the reference in the end.

References

[1] J. Hiriart-Urruty and C. Lemarechal. *Convex analysis and minimization algorithms*, volume 1. Springer-Verlag, New York, 1993.

[2] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Technical report, UC Berkeley Department of Statistics*, 2003.

[3] M. J. Wainwright and M. I. Jordan. A variational principle for graphical models. In *New Directions in Statistical Signal Processing*. MIT Press, 2005.

[4] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. pages 668–674. NIPS 13, MIT Press, 2000.