**10-708: Probabilistic Graphical Models 10-708, Spring 2017**

# 14 : Theory of Variational Inference: Inner and Outer Approximation

*Lecturer: Eric P. Xing* *Scribes: Maria Ryskina, Yen-Chia Hsu*

## 1 Introduction

What are variational methods? It means optimization based formulations. When a optimization problem is intractable, we can approximate the solution by relaxing the optimization problem. For example, when solving a linear system of equations $Ax = b$, one can obtain the solution by computing the inverse of matrix $A$ and get $x = A^{-1}b$. However, it may be hard to inverse matrix $A$. Thus, we can use a relaxed version of the problem $x = \text{argmin}_x\big((1/2)x^T Ax - b^T x\big)$. For arbitrary optimization problems, one can always start from a heuristic method without deriving the close form.

The same concept can be applied to graphical models. When performing inference on graphical models, we do not always get the close form. For example, consider an undirected graphical model (Markov Random Field, MRF), we want to compute the marginal distributions and the normalization constant (the partition function). We can use two techniques to approximate them: exponential families and convex conjugate.

## 2 Exponential Families

Recall that exponential families can be expressed in the following form:

$$p_\theta(x_1...x_m) = \exp \theta^T \phi(x) - A(\theta)$$

where $\theta$ are canonical parameters, $\phi(x)$ are sufficient statistics, and $A(\theta)$ is the log partition function. The is called canonical parameterization. Recall that the probability density function (PDF) of an undirected graphical model is:

$$p(x;\theta) = \frac{1}{Z(\theta)} \prod_C \psi(x_C, \theta_C)$$

where $\psi(x_C, \theta_C)$ is the clique potential function. The above PDF can be rewritten in the form of exponential families:

$$p(x;\theta) = \exp\left(\sum_C \log \psi(x_C, \theta_C) - \log Z(\theta)\right)$$

For example, Gaussian MRF can be expressed as:

$$p(x) = \exp\left(\frac{1}{2}\langle \Theta, xx^T\rangle - A(\Theta)\right)$$

where $\Theta = -\Sigma^{-1}$ and $\Sigma$ is the covariance matrix. The Gaussian MRF can be learned via Graphical Lasso.

Another example is discrete MRF (see Figure 1), which can be expressed as:

$$p(x;\theta) \propto \exp\left(\sum_{s\in V}\sum_j \theta_{s;j} I_j(x_s) + \sum_{(s,t)\in E} \theta_{st;jk} I_j(x_s) I_k(x_t)\right)$$

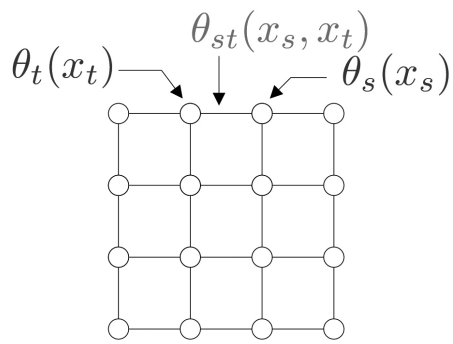


Figure 1: The discrete Markov random field.

where $I_j(x_s)$ is the indicator function that has value 1 if $x_s = j$ and 0 otherwise, $\theta_s$ is the node potential, and $\theta_{st}$ is the edge potential. The reason why we want to represent the PDF in the form of exponential families is because it gives nice forms to solve the inference problem (parameter estimation). Computing the expectation of sufficient statistics given the canonical parameters yields the marginals.

$$\mu_{s;j} = E_p[I_j(X_s)] = P[X_s = j] \forall j \in X_s$$

$$\mu_{st;jk} = E_p[I_{st;jk}(X_s, X_t)] = P[X_s = j, X_t = k] \forall (j,k) \in X_s \in X_t$$

The expectation of sufficient statistics is also called mean parameters. Then computing the normalizer yield the log partition function.

$$\log Z(\theta) = A(\theta)$$

Let us now consider a Bernoulli distribution example in the following form:

$$p(x;\theta) = \exp\big(\theta x - A(\theta)\big)$$

where $x \in 0, 1$ and $A(\theta) = \log(1 + e^\theta)$. Now we compute its mean parameter in a variational manner.

$$\mu(\theta) = E_\theta[X] = 1 \cdot p(X = 1, \theta) + 0 \cdot p(X = 0, \theta) = \frac{e^\theta}{1 + e^\theta}$$

# 3 Convex Conjugate

The convex conjugate dual function of any given function $f(x)$ is:

$$f^*(\mu) = \sup_\theta \big(\langle \theta, \mu \rangle - f(\theta)\big)$$

Figure 2 shows the geometry of the conjugate dual function. We can think of it as the lower bound of slope of the original function. Notice that the dual function is always convex. In addition, the dual of the dual is the original under the situation that the original $f(\theta)$ is convex and lower semi-continuous. Now let us use this concept to write the log partition function $A(\theta)$ in the following form:

$$A(\theta) = \sup_\mu \big(\langle \mu, \theta \rangle - A^*(\mu)\big)$$

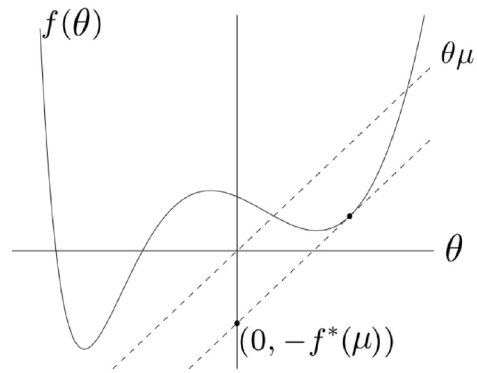


Figure 2: The conjugate dual function.

The primal variable is $\theta$ and the dual variable is $\mu$. Now the dual variable can be interpreted as the mean parameter. Let us now go back to the Bernoulli example. First we can express the conjugate dual function of Bernoulli as:

$$A^*(\mu) = \sup_{\theta \in R} \Big( \mu\theta - \log\big(1 + \exp(\theta)\big) \Big)$$

Taking the derivative with respect to $\theta$ and settign it to zero yields

$$\mu = \frac{e^\theta}{1 + e^\theta}$$

Then we can solve $\theta$ and obtain

$$\theta = \log\left(\frac{\mu}{1-\mu}\right)$$

Plug $\theta$ back into $A^*(\mu)$ and we get

$$A^*(\mu) = \begin{cases} \mu \log \mu + (1-\mu)\log(1-\mu) & \text{if } \ \mu \in [0,1] \\ +\infty & \text{otherwise} \end{cases}$$

Recall that the variational form of $A(\theta)$ is

$$A(\theta) = \max_{\mu \in [0,1]} \big( \mu^T \theta - A^*(\mu) \big)$$

Taking the derivative with respect to $\mu$ and setting it to zero yields

$$\mu = \frac{e^\theta}{1 + e^\theta}$$

which is exactly the mean parameter that we computed before. Notice that the dual function $A^*(\mu)$ is the negative entropy of a Bernoulli distribution. This is true in general. Now let us consider a generalized case for arbitrary exponential families in the following form

$$p(x_1, ..., x_m; \theta) = \exp\left( \sum_i \theta_i \phi_i(x) - A(\theta) \right)$$

By definition, the dual function of $A(\theta)$ is

$$A^*(\mu) = \sup_\theta \big( \langle \theta, \mu \rangle - A(\theta) \big)$$

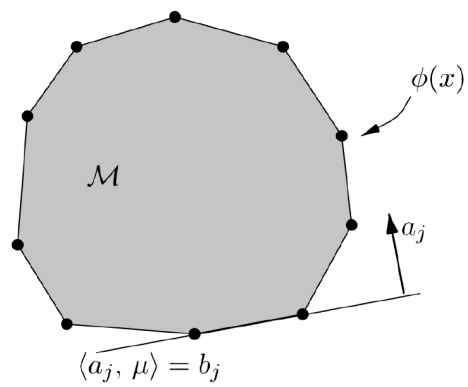


Figure 3: The marginal polytope.

Taking the derivative with respect to $\theta$ and setting it to zero yields

$$\mu = \nabla A(\theta) = E_\theta[\phi_i(X)] = \int \phi_i(x) p(x;\theta) dx$$

Recall that $\theta$ is actually $\theta(\mu)$, a function of $\mu$. Assume that a solution $\theta$ exists such that $\mu = E_\theta[\phi(X)]$. The dual function has the following form:

$$\begin{aligned} A^*(\mu) &= \langle \theta, \mu \rangle - A(\theta) \\ &= \langle \theta, E_\theta[\phi(X)] \rangle - A(\theta) \\ &= E_\theta[\langle \theta, \phi(X) \rangle - A(\theta)] \\ &= E_\theta[\log p(X;\theta)] \\ &= \int \log p(X;\theta) p(x;\theta) dx \\ &= -H(p(x;\theta)) \end{aligned}$$

Now we can see that the dual function is the negative entropy of the probability density distribution. However, computing the entropy function is generally intractable and the domain of the mean parameter is hard to define. Therefore, we need approximation methods. Now the next question is: how can we setup the domain of the mean parameter $\mu$. When $p(x)$ is an exponential family, the set of all realizable mean parameters has the following form:

$$\mathcal{M} = \left\{ \mu \in R^d \ \mid \ \exists p \ \text{ s.t. } \ E_p[\phi(X)] = \mu \right\}$$

For discrete exponential families, the domain is actually a marginal polytope. It is a convex set. According to the Minkowski-Weyl Theorem that any non-empty convex polytope can be characterized by a finite collection of linear inequality constraints, we can obtain a half-plane representation (see Figure 3):

$$\mathcal{M} = \left\{ \mu \in R^d \ \mid \ a_j^T \mu \leq b_j, \forall j \in J \right\}$$

# 4 Variational principle

The dual function takes the form:

$$A^*(\theta) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^o \\ +\infty & \text{if } \mu \in \bar{\mathcal{M}} \end{cases}$$

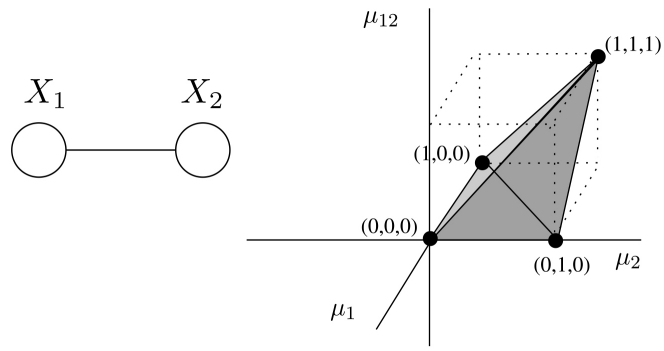


Figure 4: The two-node Ising model.

where $\mathcal{M}^o$ and $\bar{\mathcal{M}}$ are the interior and closure of $\mathcal{M}$ respectively. Then we can define the log partition function as a variational problem:

$$A(\theta) = \sup_{\mu \in \mathcal{M}^o} \{\theta^T \mu - A^*(\mu)\} \tag{1}$$

Theorem 3.4 in Wainwright et al. (2008) states that the unique solution of this problem is given by $\mu(\theta) \in \mathcal{M}^o$ that satisfies

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

This means that once we solve the optimization problem for $A(\theta)$, we would already have solved the parameter estimation problem for $\mu(\theta)$. So we can limit ourselves to just solving the problem stated Equation 1.

**Example: two-node Ising model**

The distribution for the model (see Figure 4) is $p(x;\theta) \propto \exp\{\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2\}$; the sufficient statistics are $\phi(x) = \{x_1, x_2, x_1 x_2\}$

Marginal polytope is characterized by the following constraints: $\begin{cases} \mu_1 \geq \mu_{12} \\ \mu_2 \geq \mu_{12} \\ \mu_{12} \geq 0 \\ 1 + \mu_{12} \geq \mu_1 + \mu_2 \end{cases}$

We can write the dual function explicitly:

$$A^*(\theta) = \mu_{12} \log \mu_{12} + (\mu_1 - \mu_{12}) \log(\mu_1 - \mu_{12}) + (\mu_2 - \mu_{12}) \log(\mu_2 - \mu_{12}) + (1 + \mu_{12} - \mu_1 - \mu_2) \log(1 + \mu_{12} - \mu_1 - \mu_2)$$

Then the variational problem takes the form

$$A(\theta) = \max_{\{\mu_1, \mu_2, \mu_{12}\} \in \mathcal{M}} \{\theta_1 \mu_1 + \theta_2 \mu_2 + \theta_{12} \mu_{12} - A^*(\theta)\}$$

And its optimum is attained at

$$\mu_1(\theta) = \frac{\exp\{\theta_1\} + \exp\{\theta_1 + \theta_2 + \theta_{12}\}}{1 + \exp\{\theta_1\} + \exp \theta_2 + \exp\{\theta_1 + \theta_2 + \theta_{12}\}}, \text{ etc.}$$

This is a very simple example where we can solve the optimization problem exactly. In an arbitrary complex model we can encounter difficulties, such as:

- marginal polytope $\mathcal{M}$ is difficult to characterize
- negative entropy $A^*$ has no explicit form

In that case, we would use the exact problem as a starting point and actually solve an approximate problem. The two approximations we are going to discuss are:

- Mean field method: *non-convex inner bound* on $\mathcal{M}$ and the *exact form* of entropy
- Bethe approximation and loopy belief propagation: *polyhedral outer bound* on $\mathcal{M}$ and *non-convex Bethe approximation* for entropy

## 4.1 Mean Field Method

For an exponential family with sufficient statistics $\phi$ defined on an arbitrary graph $G$, the original realizable mean parameter set is:

$$\mathcal{M}(G;\phi) := \{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\}$$

The idea of mean field is to tighten the constraints and solve the problem on $\mathcal{M}_F \subseteq \mathcal{M}$. $G$ can be intractable, but we can restrict $p$ to a subset of distributions associated with a tractable subgraph.

For example, instead of searching the original parameter space $\Omega = \{\theta \in \mathbb{R}^d | A(\theta) < +\infty\}$ we can:

- Delete all edges from $G$: $\Omega_{F_0} = \{\theta \in \Omega | \theta_{(s,t)} = 0,\ \forall (s,t) \in E\}$.
- Delete some edges from $G$ to turn it into a tree: $\Omega_T = \{\theta \in \Omega | \theta_{(s,t)} = 0,\ \forall (s,t) \notin E(T)\}$.

For a given tractable subgraph $F$, the subset of canonical parameters is:

$$\mathcal{M}(F;\phi) := \{\tau \in \mathbb{R}^d | \tau = \mathbb{E}_\theta[\phi(X)] \text{ for some } \theta \in \Omega(F)\}$$

We are using the inner approximation:

$$\mathcal{M}(F;\theta)^o \subseteq \mathcal{M}(G;\theta)^o$$

Mean field solves the relaxed problem:

$$\max_{\tau \in \mathcal{M}_F(G)} \{\langle \tau, \theta \rangle - A_F^*(\tau)\},$$

where $A_F^*(\tau) = A^*|_{\mathcal{M}_F}$ is the *exact* dual function restricted to $\mathcal{M}_F$.

**Example: Naive mean field for Ising model**

Ising model in $\{0,1\}$ representation:

$$p(x) \propto \exp\left\{\sum_{s \in V} x_s \theta_s + \sum_{(s,t) \in E} x_s x_t \theta_{st}\right\}$$

Its mean parameters are:

$$\mu_s = \mathbb{E}_p[X_s] = \mathbb{P}[X_s = 1] \text{ for all } s \in V$$

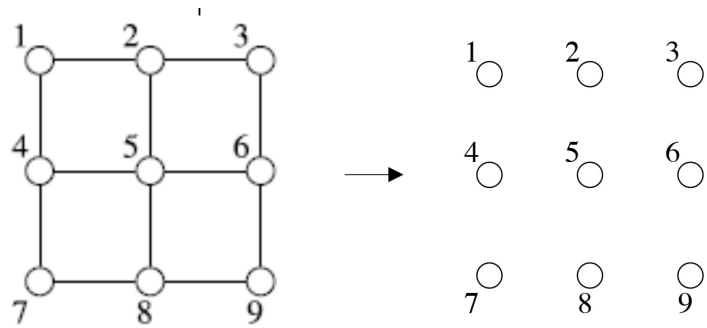


Figure 5: The naive mean field for Ising model.

$$\mu_{st} = \mathbb{E}_p[X_s X_t] = \mathbb{P}[(X_s, X_t) = (1,1)] \text{ for all } (s,t) \in E$$

Now let us remove all the edges (see Figure 5). For the fully disconnected graph $F$,

$$\mathcal{M}_F(G) = \{\tau \in \mathbb{R}^{|V|+|E|} | \, 0 \leq \tau_s \leq 1, \forall s \in V, \, \tau_{st} = \tau_s \tau_t, \forall (s,t) \in E\}$$

Now the dual decomposes into a sum:

$$A_F^*(\tau) = \sum_{s \in V} [\tau_s \log \tau_s + (1 - \tau_s) \log(1 - \tau_s)]$$

And the mean field variational problem is:

$$A(\theta) = \max_{\{\tau_1, \ldots \tau_m\} \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \tau_s + \sum_{(s,t) \in E} \theta_{st} \tau_s \tau_t - A_F^*(\tau) \right\},$$

which is the same objective function as in the free energy-based approach.

The naive mean field update equations are

$$\tau_s \leftarrow \sigma \left( \theta_s + \sum_{t \in N(s)} \theta_s \tau_t \right)$$

**Geometry of mean field.**

Mean field optimization is always non-convex for any exponential family in which state space $\mathcal{X}^m$ is finite. To demonstrate it, let us remember that marginal polytope is a convex hull:

$$\mathcal{M}(G) = \text{conv}\{\phi(e); e \in \mathcal{X}^m\}$$

$\mathcal{M}_{MF}$ necessarily contains all of its extreme points, so if it is a *strict* subset, it must be non-convex. This is illustrated by Figure 6.

As an example, let us consider two-node Isisng model again. Its naive mean field marginal polytope approximation

$$\mathcal{M}_F(G) = \{0 \leq \tau_1 \leq 1, 0 \leq \tau_2 \leq 1, \tau_{12} = \tau_1 \tau_2\}$$

has a parabolic cross-section along $\tau_1 = \tau_2$, which means it is non-convex.

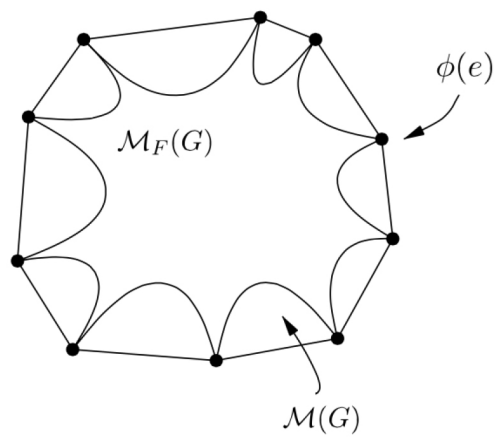


Figure 6: Mean field inner approximation on marginal polytope.

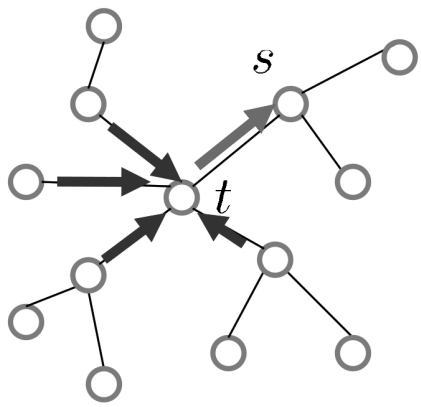


Figure 7: The sum-product algorithm for belief propagation.

## 4.2 Bethe Approximation and Sum-Product

Let us remember the sum-product algorithm (see Figure 7) that we defined for trees. The message passing rule has the form:

$$M_{ts}(x_s) \leftarrow \kappa \sum_{x'_t} \left\{ \psi_{st}(x_s, x'_t)\psi_t(x'_t) \prod_{u \in N(t)\setminus s} M_{ut}(x_u, x'_t) \right\},$$

and the marginals are:

$$\mu_s(x_s) = \kappa \psi_s(x_s) \prod_{t \in N(s)} M^*_{ts}(x_s).$$

Sum-product algorithm is exact on trees, but becomes approximate for loopy graphs (*loopy belief propagation*).

### 4.2.1 Exact inference on trees

Let us first consider a tree $T(V, E)$ where the nodes are discrete variables $X_s \in \{0, 1, \ldots, m_s - 1\}$. The sufficient statistics are given by:

$$\mathbb{I}_j(x_s) \text{ for } s = 1, \ldots n, j \in \mathcal{X}_s$$
$$\mathbb{I}_{jk}(x_s, x_t) \text{ for } (s,t) \in E, (j,k) \in \mathcal{X}_s \times \mathcal{X}_t$$

With these sufficient statistics, we define the exponential family of the form:

$$p(x;\mu) \propto \exp\left\{\sum_{s\in V} \theta_s(x_s) + \sum_{(s,t)\in E} \theta_{st}(x_s, x_t)\right\},$$

where $\theta_s(x_s) := \sum_{j\in\mathcal{X}_s} \theta_{s;j}\mathbb{I}_j(x_s)$ and $\theta_{st}(x_s, x_t) := \sum_{(j,k)\in\mathcal{X}_s\times\mathcal{X}_t} \theta_{st;jk}\mathbb{I}_{jk}(x_s, x_t)$.

The associated mean parameters correspond to marginal probabilities:

$$\mu_{s;j} = \mathbb{E}_p[\mathbb{I}_j(X_s)] = \mathbb{P}[X_s = j] \ \forall j \in \mathcal{X}_s$$

$$\mu_{st;jk} = \mathbb{E}_p[\mathbb{I}_j k(X_s, X_t)] = \mathbb{P}[X_s = j, X_t = k] \ \forall (j,k) \in \mathcal{X}_s \times \mathcal{X}_t$$

Let us define the following notation:

$$\mu_s(x_s) := \sum_{j\in\mathcal{X}_s} \mu_{s;j}\mathbb{I}_j(x_s) = \mathbb{P}(X_s = x_s)$$

$$\mu_{st}(x_s, x_t) := \sum_{(j,k)\in\mathcal{X}_s\times\mathcal{X}_t} \mu_{st;jk}\mathbb{I}_j k(x_s, x_t) = \mathbb{P}(X_s = x_s, X_t = x_t)$$

Now let us define the marginal polytope for trees.

$$\mathcal{M}(G) = \{\mu \in \mathbb{R}^d | \exists p \text{ with marginals } \mu_{s;j}, \mu_{st;jk}\}$$

By junction tree theorem, it takes the form

$$\mathcal{M}(T) = \left\{\mu \geq 0 | \sum_{x_s} \mu_s(x_s) = 1, \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s)\right\}$$

In particular, if $\mu \in \mathcal{M}(T)$

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

has the corresponding marginals.

For trees, the entropy can be decomposed as

$$H(p(x;\mu)) = -\sum_x p(x;\mu)\log p(x;\mu) =$$
$$= \sum_{s\in V}\left(-\sum_{x_s} \mu_s(x_s)\log\mu_s(x_s)\right) - \sum_{(s,t)\in E}\left(-\sum_{x_s,x_t} \mu_{st}(x_s, x_t)\log\frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}\right) =$$
$$= \sum s \in V H_s(\mu_s) - \sum_{(s,t)\in E} I_{st}(\mu_{st})$$

The dual function has the explicit form

$$A^*(\mu) = -H(p(x;\mu))$$

The variational formulation is

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

Let us construct a Lagrangian for this optimization problem. The constraints are:

$$C_{ss}(\mu) := 1 - \sum_{x_s} \mu_s(x_s) = 0, \;\; C_{st}(\mu) := \mu_s(x_s) - \sum_{x_t} \mu_{st}(x_s, x_t) = 0$$

Then the Lagrangian has the form:

$$\mathcal{L}(\mu, \lambda) = \langle \theta, \mu \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}) + \sum_{s \in V} \lambda_{ss} C_{ss}(\mu) + \\ + \sum_{(s,t) \in E} [\sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t) + \sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s)]$$

Taking the derivatives of the Lagrangian w.r.t. $\mu_s$ and $\mu_{st}$:

$$\frac{\partial \mathcal{L}}{\partial \mu_s(x_s)} = \theta_s(x_s) - \log \mu_s(x_s) + \sum_{t \in N(s)} \lambda_{ts}(x_s) + C$$

$$\frac{\partial \mathcal{L}}{\partial \mu_{st}(x_s, x_t)} = \theta_{st}(x_s, x_t) - \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} - \lambda_{ts}(x_s) - \lambda_{st}(x_t) + C'$$

Setting them to zeros yields

$$\mu_s(x_s) \propto \exp\{\theta_s(x_s)\} \prod_{t \in N(s)} \underbrace{\exp\{\lambda_{ts}(x_s)\}}_{M_{ts(x_s)}}$$

$$\mu_{st}(x_s, x_t) \propto \exp\{\theta_s(x_s) + \theta_t(x_t) + \theta_{st}(x_s, x_t)\} \prod_{u \in N(s) \setminus t} \exp\{\lambda_{us}(x_s)\} \prod_{v \in N(t) \setminus s} \exp\{\lambda_{vt}(x_t)\}$$

Finally, adjusting the Lagrange multipliers or messages to enforce $C_{ts}(x_s; \mu) = 0$ yields:

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

We have just demonstrated that the message passing updates are a Lagrange method to solve the stationary condition of the variational formulation.

### 4.2.2 BP on arbitrary graphs

Now let us come back to the two main difficulties of variational formulation given in Equation 1:

- The marginal polytope $\mathcal{M}$ is hard to characterize, so let us approximate it with the tree-based outer bound:

  $$\mathbb{L}(G) = \left\{ \tau \geq 0 | \sum_{x_s} \tau_s(x_s) = 1, \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s) \right\}$$

  These locally consistent vectors $\tau$ are called *pseudo-marginals*.

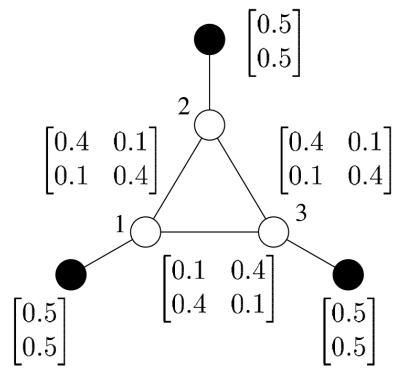


Figure 8: An example of belief propagation with pseudo-marginals.

- Exact entropy $-A^*(\mu)$ cannot be written explicitly, so we approximate it with *Bethe enthropy*, which would also be exact in case of trees:

$$-A^*(\tau) \approx H_{Bethe}(\tau) := \sum_{s\in V} H_s(\tau_s) - \sum_{(s,t)\in E} I_{st}(\tau_{st})$$

These two approximations combined define the *Bethe variational problem* (BVP):

$$\max_{\tau\in\mathbb{L}(G)} \left\{ \langle\theta,\mu\rangle + \sum_{s\in V} H_s(\tau_s) - \sum_{(s,t)\in E} I_{st}(\tau_{st}) \right\}$$

BVP is a simple structured problem: the objective is differentiable and the constraint set defines a simple convex polytope. Loopy belief propagation can be derived as an iterative method for solving a Lagrangian formulation of the BVP (Theorem 4.2 in Wainwright et al. (2008)), the same way we derived it for tree graphs.

**Geometry of BP**. Let us consider a three-node cycle graph (see Figure 8) with binary variables $X \in \{0,1\}^3$ and define the following family of pseudo-marginals (in matrix form):

$$\tau_s = \begin{bmatrix}0.5\\0.5\end{bmatrix}, \tau_{st} = \begin{bmatrix}0.4 & 0.1\\0.1 & 0.4\end{bmatrix}, \ \forall s,t \in \{1,2,3\}$$

We can easily verify that $\tau \in \mathbb{L}(G)$, but with some effort we can show that $\tau \notin \mathcal{M}(G)$.

For any graph $\mathbb{L}(G) \subseteq \mathcal{M}(G)$, but the equality holds if an only if the graph is a tree. A solution of BVP can fall into the gap: for any element of outer bound (see Figure 9) $\mathbb{L}(G)$ it is possible to construct a distribution with it as a fixed point (Wainwright et al., 2008).

**Inexactness of Bethe entropy approximation**. Let us consider a fully connected graph with 4 nodes and the following family of marginals:

$$\tau_s = \begin{bmatrix}0.5\\0.5\end{bmatrix}, \tau_{st} = \begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix}, \ \forall s,t \in \{1,2,3,4\}$$

These marginals are globally valid, realized by the distribution that places mass 1/2 on each configuration $(0,0,0,0)$ and $(1,1,1,1)$.

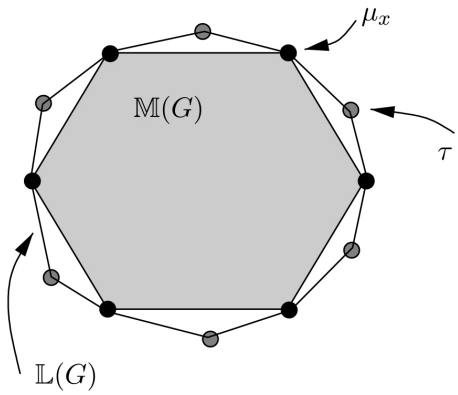


Figure 9: The geometry of belief propagation.

Then Bethe entropy for this case is $H_{Bethe}(\mu) = 4\log 2 - 6\log 2 = -2\log 2 < 0$, which cannot be the true entropy. The true entropy in this example is given by $-A^*(\mu) = \log 2 > 0$.

We have demonstrated the principled basis for applying sum-product algorithm for loopy graphs. However, the algorithm does not necessarily converge to the desired point:

- Even though there is always a fixed point of loopy BP, there is no guarantees on the convergence of the algorithm on loopy graphs;
- BVP is usually non-convex, so there are no guarantees on the global optimum;
- Generally, there are no guarantees that $A_{Bethe}(\theta)$ is a lower bound of $A(\theta)$.

Nevertheless, this connection suggests a number of possible improvements of ordinary sum-product algorithm via progressively better approximations to entropy function and tighter outer bounds on $\mathcal{M}$ (Kikuchi clustering).

# 5 Summary

In this lecture, we have discussed the foundations of variational methods. Variational methods in general turn inference into an optimization problem via exponential families and convex duality. The exact variational principle is intractable to solve, so instead we solve an approximation consisting of two components: inner or outer bound on marginal polytope and various approximations to the entropy function.

We have briefly described the following methods:

- Mean field: non-convex inner bound and exact form of entropy
- BP: polyhedral outer bound and non-convex Bethe approximation for entropy
- Kikuchi and variants: tighter polyhedral outer bounds and better entropy approximations (Yedidia et al., 2005).

## References

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

Jonathan S Yedidia, William T Freeman, and Yair Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, 2005.