Lecture 4: Variational Inference in Exponential Family Graphical Models

Tianpei Xie

Aug. 25th., 2022

Contents

1 Background knowledge

Recall the formulation of Bayesian network and Markov network

- Given directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $(s,t) \neq (t,s)$, the **directed graphical model** factorizes the joint distribution into a set of *factors* $\{p_s(x_s|x_{\pi(s)}) : s \in \mathcal{V}\}$ according to the ancestor relations defined in $\mathcal{G}$

$$p(x_1, \ldots, x_m) = \prod_{s \in \mathcal{V}} p_s(x_s|x_{\pi(s)}). \tag{1}$$

  This class of models are also referred as ***Bayesian networks*** [Koller and Friedman, 2009].

- Given undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $(s,t) = (t,s)$, the joint distribution of **Markov random fields** (**Markov network**) *factorize* as

$$p(x_1, \ldots, x_m) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C), \tag{2}$$

  where Z is a constant chosen to ensure that the distribution is normalized. The set $\mathcal{C}$ is often taken to be the *set of all* ***maximal cliques*** *of the graph*, i.e., the set of cliques that are *not* properly contained within any other clique. Note that any representation based on nonmaximal cliques can always be converted to one based on maximal cliques by redefining the compatibility function on a maximal clique to be the *product* over the compatibility functions on the *subsets* of that clique.

- The canonical representation of ***exponential famlity*** of distribution has the following form

$$\begin{aligned} p(x_1, \ldots, x_m) = p(\boldsymbol{x}; \boldsymbol{\eta}) &= \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x}) \rangle - A(\boldsymbol{\eta})\right) h(\boldsymbol{x}) \nu(d\boldsymbol{x}) \\ &= \exp\left(\sum_{\alpha} \eta_\alpha \phi_\alpha(\boldsymbol{x}) - A(\boldsymbol{\eta})\right) \end{aligned} \tag{3}$$

  where ϕ is a feature map and $\boldsymbol{\phi}(\boldsymbol{x})$ defines a set of ***sufficient statistics*** (or ***potential functions***). The normalization factor is defined as

$$A(\boldsymbol{\eta}) := \log \int \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x}) \rangle\right) h(\boldsymbol{x}) \nu(d\boldsymbol{x}) = \log Z(\boldsymbol{\eta})$$

  $A(\boldsymbol{\eta})$ is also referred as ***log-partition function*** or *cumulant function*. The parameters $\boldsymbol{\eta} = (\eta_\alpha)$ are called ***natural parameters*** or *canonical parameters*. The canonical parameter $\{\eta_\alpha\}$ forms a **natural (canonical) parameter space**

$$\Omega = \left\{ \boldsymbol{\eta} \in \mathbb{R}^d : A(\boldsymbol{\eta}) < \infty \right\} \tag{4}$$

- The exponential family is the unique solution of ***maximum entropy estimation*** problem:

$$\min_{q \in \Delta} \quad \mathbb{KL}\left(q \,\|\, p_0\right) \tag{5}$$

$$\text{s.t.} \quad \mathbb{E}_q\left[\phi_\alpha(X)\right] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I} \tag{6}$$

where $\mathbb{KL}(q \parallel p_0) = \int \log(\frac{q}{p_0}) q dx = \mathbb{E}_q \left[\log \frac{q}{p_0}\right]$ is the relative entropy or the Kullback-Leibler divergence of q w.r.t. p_0.

Here $\boldsymbol{\mu} = (\mu_\alpha)_{\alpha \in \mathcal{I}}$ is a set of ***mean parameters***. The space of mean parameters $\mathcal{M}$ is a *convex polytope* spanned by potential functions $\{\phi_\alpha\}$.

$$\mathcal{M} := \left\{ \boldsymbol{\mu} \in \mathbb{R}^d : \exists q \text{ s.t. } \mathbb{E}_q [\phi_\alpha(X)] = \mu_\alpha \quad \forall \alpha \in \mathcal{I} \right\} = \text{conv}\{\phi_\alpha(x),\ x \in \mathcal{X},\ \alpha \in \mathcal{I}\} \tag{7}$$

- Note that $A(\boldsymbol{\eta})$ is a convex function and its gradient $\nabla A : \Omega \to \mathcal{M}^\circ$ is a bijection between the natural parameter space Ω and the **interior** of $\mathcal{M}$, $\mathcal{M}^\circ$; $\nabla A(\boldsymbol{\eta}) = \boldsymbol{\mu}$ based on the following equation

$$\frac{\partial A}{\partial \eta_\alpha} = \mathbb{E}_{\boldsymbol{\eta}} [\phi_\alpha(X)] := \int_{\mathcal{X}^m} \phi_\alpha(\boldsymbol{x}) q(\boldsymbol{x}; \boldsymbol{\eta}) d\boldsymbol{x} = \mu_\alpha \tag{8}$$

- Moreover $A(\boldsymbol{\eta})$ has a variational form

$$A(\boldsymbol{\eta}) = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \{\langle \boldsymbol{\eta}\,,\, \boldsymbol{\mu} \rangle - A^*(\boldsymbol{\mu})\} \tag{9}$$

where $A^*(\boldsymbol{\mu})$ is the conjugate dual function of A and it is defined as

$$A^*(\boldsymbol{\mu}) := \sup_{\boldsymbol{\eta} \in \Omega} \{\langle \boldsymbol{\mu}\,,\, \boldsymbol{\eta} \rangle - A(\boldsymbol{\eta})\} \tag{10}$$

It is shown that $A^*(\boldsymbol{\mu}) = -H(q_{\boldsymbol{\eta}(\boldsymbol{\mu})})$ for $\boldsymbol{\mu} \in \mathcal{M}^\circ$ which is the negative entropy. $A^*(\boldsymbol{\mu})$ is also the optimal value for the **maximum likelihood estimation** problem on p. The exponential family can be reparameterized according to its mean parameters $\boldsymbol{\mu}$ via backward mapping $(\nabla A)^{-1} : \mathcal{M}^\circ \to \Omega$, called **mean parameterization**.

- We can formulate the **KL divergence** between two distributions in exponential family Ω using its primal and dual form
  - **Primal-form**: given $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \Omega$

$$\begin{aligned} \mathbb{KL}\left(p_{\boldsymbol{\eta}_1} \parallel p_{\boldsymbol{\eta}_2}\right) \equiv \mathbb{KL}(\boldsymbol{\eta}_1 \parallel \boldsymbol{\eta}_2) &= A(\boldsymbol{\eta}_2) - A(\boldsymbol{\eta}_1) - \langle \boldsymbol{\mu}_1\,,\, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle \\ &\equiv A(\boldsymbol{\eta}_2) - A(\boldsymbol{\eta}_1) - \langle \nabla A(\boldsymbol{\eta}_1)\,,\, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle \end{aligned} \tag{11}$$

  - **Primal-dual form**: given $\boldsymbol{\mu}_1 \in \mathcal{M}, \boldsymbol{\eta}_2 \in \Omega$

$$\mathbb{KL}(\boldsymbol{\mu}_1 \parallel \boldsymbol{\eta}_2) = A(\boldsymbol{\eta}_2) + A^*(\boldsymbol{\mu}_1) - \langle \boldsymbol{\mu}_1\,,\, \boldsymbol{\eta}_2 \rangle \tag{12}$$

  - **Dual-form**: given $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{M}$

$$\begin{aligned} \mathbb{KL}(\boldsymbol{\mu}_1 \parallel \boldsymbol{\mu}_2) &= A^*(\boldsymbol{\mu}_1) - A^*(\boldsymbol{\mu}_2) - \langle \boldsymbol{\eta}_2\,,\, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \rangle \\ &\equiv A^*(\boldsymbol{\mu}_1) - A^*(\boldsymbol{\mu}_2) - \langle \nabla A^*(\boldsymbol{\mu}_2)\,,\, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \rangle \end{aligned} \tag{13}$$

- The exact inference on **tree-based models** makes use of the **decomposition** property of tree, $\mathcal{T} = (\mathcal{V}, \mathcal{E}_\mathcal{T})$. Specifically, define the neighborhood of node s as

$$\mathcal{N}(s) = \{t \in \mathcal{V} : (s,t) \in \mathcal{E}_\mathcal{T}\}$$

For each $u \in \mathcal{N}(s)$, let $\mathcal{T}_u = (\mathcal{V}_u, \mathcal{E}_u)$ be the subgraph formed by the set of nodes (and edges joining them) that *can be reached from* u by **paths** that ***do not pass*** through node s. An **important property** for tree is that the subgraph $\mathcal{T}$ is a tree and for any $u \neq t, \forall u, t \in \mathcal{N}(s)$, $\mathcal{T}_u \cap \mathcal{T}_t = \emptyset$. This means that we can **decompose** the tree $\mathcal{T}_s$ rooted at s, by **removing** s from the tree. It then form a collection or non-overlapping subtrees $\{\mathcal{T}_t, t \in \mathcal{N}(s)\}$.

- The tree-based Markov random field is defined as

$$p(x_1, \ldots, x_m; \mathcal{T}) = \frac{1}{Z} \prod_{s \in \mathcal{V}} \psi_s(x_s) \prod_{(s,t) \in \mathcal{E}_\mathcal{T}} \psi_{s,t}(x_s, x_t), \tag{14}$$

- The ***Belief Propagation algorithm*** is based on *dynamic programming*. The key is to pass *messages* from one node to its neighborhood carrying marginalized probability from other nodes. ***Message*** is a value function, which is defined as

$$M^*_{t \to s}(x_s) := M^*_{t,s}(x_s) = \sum_{\boldsymbol{x}_{\mathcal{V}_t}} p_t(\boldsymbol{x}_{\mathcal{V}_t} | x_s; \mathcal{T}_t) = \sum_{\boldsymbol{x}_{\mathcal{V}_t}} \psi_{s,t}(x_s, x_t) p(\boldsymbol{x}_{\mathcal{V}_t}; \mathcal{T}_t) \tag{15}$$

The target is to find marginal distribution at x_s

$$\begin{aligned} \mu_s(x_s) &:= \sum_{\boldsymbol{x}_{-s}} p(x_s, \boldsymbol{x}_{-s}) \\ &= \psi_s(x_s) \prod_{t \in \mathcal{N}(s)} M^*_{t,s}(x_s). \end{aligned} \tag{16}$$

The sum-product algorithm is based on the following ***Bellman equation***:

$$\begin{aligned} M^*_{t,s}(x_s) &= \sum_{\boldsymbol{x}_{\mathcal{V}_t}} \psi_{s,t}(x_s, x_t) p(\boldsymbol{x}_{\mathcal{V}_t}; \mathcal{T}_t) \\ &= \kappa \sum_{x_t \in \mathcal{X}} \left\{ \psi_{s,t}(x_s, x_t) \psi_t(x_t) \prod_{u \in \mathcal{N}(t) - \{s\}} M^*_{u,t}(x_t) \right\}, \forall t \in \mathcal{N}(s), s \in \mathcal{V} \end{aligned} \tag{17}$$

where $\kappa > 0$ again denotes a *normalization constant*.

2 Approximate inference via variational methods

The message passing algorithms in (17) are mainly for tabular-based graphical model. For graphical model using function representation, such as those in exponential families, exact inference task is challenging. In these graphical models, two fundamental difficulties associated with:

- the nature of the **constraint set** $\mathcal{M}$, which is high dimensional with many facets and extreme points;
- the lack of an explicit form for the **dual function** A^*, esp. **the negative entropy** on joint distribution cannot be decomposed according to graph $\mathcal{G}$.

Therefore, we focus on **approximate Inference methods** of exponential family graphical model. There are two main branches:

- ***Monte Carlo sampling methods***: Many graphical models have explicit function form for *conditional probability* (e.g. *Gaussian graphical model*, *LDA*, *RBMs* etc.). This allows us to apply sampling techiques such as *importance sampling*, *Gibbs sampling*, *stimulated annealing*, *hybrid Monte Carlo* etc. [Liu and Liu, 2001, Shapiro, 2003]. The **key idea** behind sampling based inference methods is that the high dimensional integration can be ***approximated*** via ***sample average*** of some statistics. Each factor can be used as a generator on a small number of variables based on local information. The **drawbacks** include **randomness of solutions**, **slow convergence**, **high variance** etc.
- ***Variational inference methods***: Variational inference methods are based on the ***variational principle*** of log-partition function (9) and conjugate duality between A and A^*. Unlike the *stochastic methods* above, these methods are ***deterministic***. We mainly discuss two classes of metheds
  - The *belief propagation* based on ***Bethe variational problem (BVP)***. It relaxes the constraint set from $\mathcal{M}(\mathcal{G})$ to a *convex* polytope $\mathcal{L}(\mathcal{G})$ and then approximates the entropy using *Bethe entropy approximation*. Both approximations allow the problem to be factorized according to graph topology.
  - The ***mean field methods***. These methods limit the distributions within an *non-convex set* $\mathcal{M}_{\mathcal{F}} \subseteq \mathcal{M}$ on *tractable graph* $\mathcal{F}$, e.g. fully factorized. They find the closet tractable model in exponential family that satisfy the mean matching conditions.

  These methods reach the *entropy decomposition* at the expense of ***losing convexity*** in the problem formulations. For BVP, the objective function is non-convex. For the mean field method, the feasible region is non-convex.

3 Belief propagation for exponential families

Unlike tabular methods, exponential families are defined ***globally*** by joint distributions. In order for the belief propagation to work, we need to break it down into a set of local factors and consider the local objective functions on each factor. In this section, we will

- **Relax** the constraint set from (global) marginal polytope $\mathcal{M}(\mathcal{G})$ to $\mathcal{L}(\mathcal{G})$ as a set of **pseudo-marginal** distributions within each factors

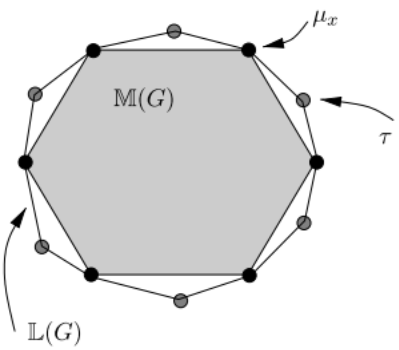

Fig. 4.2 Highly idealized illustration of the relation between the marginal polytope $\mathbb{M}(G)$ and the outer bound $\mathbb{L}(G)$. The set $\mathbb{L}(G)$ is always an outer bound on $\mathbb{M}(G)$, and the inclusion $\mathbb{M}(G) \subset \mathbb{L}(G)$ is strict whenever G has cycles. Both sets are polytopes and so can be represented either as the convex hull of a finite number of extreme points, or as the intersection of a finite number of half-spaces, known as facets.

Figure 1: The marginal polytope $\mathcal{M}(\mathcal{G})$ and its outer bound relaxation $\mathcal{L}(\mathcal{G})$.

- **Approximate** the global entropy $-A^*(\tau)$ by **Bethe entropy** which can be **decomposed** into entropies on each local factors

3.1 Approximate $\mathcal{M}(\mathcal{G})$ via pseudo-marginal distributions

Consider the pairwise Markov random fields with binary indicator potential functions:

$$\phi_{s;j}(x_s) = \mathbb{1}\{x_s = j\} \tag{18}$$

Moreover, for each edge (s,t) and pair of values $(j,k) \in \mathcal{X} \times \mathcal{X}$, define the sufficient statistics

$$\phi_{st;jk}(x_s, x_t) = \mathbb{1}\{x_s = j \wedge x_t = k\} \tag{19}$$

The joint distribution is

$$p(x_1, \ldots, x_m; \boldsymbol{\eta}) = \exp\left(\sum_{s \in \mathcal{V}} \sum_{j \in \mathcal{X}} \eta_{s;j}\, \phi_{s;j}(x_s) + \sum_{(s,t) \in \mathcal{E}} \sum_{(j,k) \in \mathcal{X} \times \mathcal{X}} \eta_{st;jk}\, \phi_{st;jk}(x_s, x_t) - Z(\boldsymbol{w}) \right), \tag{20}$$

As discussed before, the mean parameter space $\mathcal{M}(\mathcal{G})$ is the ***marginal polytope*** over $\mathcal{G}$ since

$$\begin{aligned} \mathcal{M}(\mathcal{G}) &:= \left\{ \boldsymbol{\mu} \in \mathbb{R}^d : \exists q \text{ s.t. } \mathbb{E}_q[\phi_\alpha(X)] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I} \right\} \\ \text{where } \mathbb{E}_p[\mathbb{1}\{x_s = j \wedge x_t = k\}] &= \mu_{st;jk} = \mathbb{P}(X_s = j \wedge X_t = k), \quad \forall s,t,j,k \\ \mathbb{E}_p[\mathbb{1}\{x_s = j\}] &= \mu_{s;j} = \mathbb{P}(X_s = j), \quad \forall s,j \end{aligned}$$

defines a set of matching constraints on the marginal distribution of p within each factor.

As a convex set, $\mathcal{M}(\mathcal{G})$ can be represented as *intersection of a finite number of half spaces*, which usually does not have explicit form. Therefore we consider an **outer bound set** $\mathcal{L}(\mathcal{G})$ as its **relaxation**:

$$\mathcal{L}(\mathcal{G}) = \{\boldsymbol{\tau} > 0 : \tau_s \text{ satisfies (22) } \forall s \in \mathcal{V},\ \tau_{s,t} \text{ satisfies (23) } \forall (s,t) \in \mathcal{E}\} \tag{21}$$

$$\sum_{x_s \in \mathcal{X}} \tau_s(x_s) = 1, \quad \forall s \in \mathcal{V} \tag{22}$$

$$\sum_{x'_t \in \mathcal{X}} \tau_{s,t}(x_s, x'_t) = \tau_s(x_s), \quad \forall x_s, x_t \in \mathcal{X},\ (s,t) \in \mathcal{E} \tag{23}$$

$$\sum_{x'_s \in \mathcal{X}} \tau_{s,t}(x'_s, x_t) = \tau_t(x_t).$$

The $\mathcal{L}(\mathcal{G})$ are set of marginal distributions that fit the ***local consistent requirement*** since they only consider the distribution***within each factor*** but the entire joint distribution. Note that $\mathcal{L}(\mathcal{G})$ is defined by $O(|\mathcal{V}| + |\mathcal{E}|)$ linear constraints. On the other hand, $\mathcal{M}(\mathcal{G})$ has a total of $|\mathcal{X}^m|$ extreme points.

The construction of $\mathcal{L}(\mathcal{G})$ follows a ***bottom-up*** approach, i.e. focusing on finding consistent marginal distributions just ***within each local factor***. On the other hand, the marginal polytope $\mathcal{M}(\mathcal{G})$ is constructed via a ***top-down approach***, i.e. it begins with **joint distribution** p and then *marginalize* over other variables to obtain the marginal distribution within each factor. It is easy to see that $\mathcal{M}$ follows the local consistency conditions in $\mathcal{L}$ but not vice versa.

Proposition 3.1 *The inclusion $\mathcal{M}(\mathcal{G}) \subseteq \mathcal{L}(\mathcal{G})$ holds for any graph. For a tree-structured graph $\mathcal{T}$, the marginal polytope $\mathcal{M}(\mathcal{T})$ is equal to $\mathcal{L}(\mathcal{T})$.*

For tree structures $\mathcal{L}(\mathcal{G}) = \mathcal{M}(\mathcal{G})$, $\boldsymbol{\tau} \in \mathcal{L}(\mathcal{G})$ defined a valid marginal distribution. But for graphs with circles, they are not *valid* marginal distribution, i.e. $\tau_{s,t} \neq \sum_{-\{s,t\}} p(\boldsymbol{x})$. In this case, they are called ***pseudo-marginals***.

3.2 Bethe entropy approximation

The entropy $-A^*$ usually does not have closed-form expression. One exception is tree-structured graph $\mathcal{T}$.

$$\begin{aligned} -A^*(\boldsymbol{\mu}) = H(p_{\boldsymbol{\mu}}) &= \mathbb{E}_{\boldsymbol{\mu}}\left[-\log p_{\boldsymbol{\mu}}(X)\right] \\ &= \sum_{s \in \mathcal{V}} H_s(\mu_s) - \sum_{(s,t) \in \mathcal{E}} I(\mu_{s,t}) \end{aligned} \tag{24}$$

where μ_s and $\mu_{s,t}$ are set of marginal distributions defined in $\mathcal{M}(\mathcal{T})$. The total entropy $-A^*$ is decomposed into ***node entropies*** and ***edge mutual information*** within each factor. These quantities are defined as below:

$$H_s(\mu_s) = \mathbb{E}_{\mu_s}\left[-\log \mu_s\right] = \int_{\mathcal{X}} -\mu_s(x_s) \log \mu_s(x_s) dx_s, \quad \forall s \in \mathcal{V} \tag{25}$$

$$I(\mu_{s,t}) = \mathbb{E}_{\mu_{s,t}}\left[\log \frac{\mu_{s,t}}{\mu_s \mu_t}\right] = \int_{\mathcal{X} \times \mathcal{X}} \mu_{s,t}(x_s, x_t) \log\left(\frac{\mu_{s,t}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}\right) dx_s dx_t,\ \forall (s,t) \in \mathcal{E} \tag{26}$$

For graphs with cycles, the decomposition (24) is seen as an approximation of $H(p_{\boldsymbol{\mu}})$, called ***Bethe entropy approximation***, i.e.

$$H(\tau) \approx H_{Bethe}(\tau) := \sum_{s \in \mathcal{V}} H_s(\tau_s) - \sum_{(s,t) \in \mathcal{E}} I(\tau_{s,t}) \tag{27}$$

An important fact, **central** in the derivation of the sum-product algorithm, is that this approximation (27) can be evaluated for any set of ***pseudomarginals*** $\{\tau_s, s \in \mathcal{V}\}$ and $\{\tau_{s,t}, (s,t) \in \mathcal{E}\}$ that belong to $\mathcal{L}(\mathcal{G})$.

Now we can formulate the ***Bethe variational problem (BVP)***:

$$\max_{\boldsymbol{\tau} \in \mathcal{L}(\mathcal{G})} \langle \boldsymbol{\tau}\,,\, \boldsymbol{\eta} \rangle + \sum_{s \in \mathcal{V}} H_s(\tau_s) - \sum_{(s,t) \in \mathcal{E}} I(\tau_{s,t}) \tag{28}$$

The objective function is called ***Bethe free energy***. This problem has a very simple structure: the cost function is given in ***closed form***, it is ***differentiable***, and the constraint set $\mathcal{L}(\mathcal{G})$ is a polytope specified by a *small number of constraints*. Due to its simple form and the fact that $H_{Bethe}(\tau)$ factorizes over $\mathcal{G}$, we can solve (28) efficiently via **belief propagation** (*sum-product algorithm*).

3.3 Sum-product for Bethe variational problem

Let us formulate the (partial) Lagrangian function for BVP:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\tau}, \boldsymbol{\lambda}; \boldsymbol{\eta}) := \langle \boldsymbol{\tau}\,,\, \boldsymbol{\eta} \rangle + \sum_{s \in \mathcal{V}} H_s(\tau_s) - \sum_{(s,t) \in \mathcal{E}} I(\tau_{s,t}) + \sum_{s \in \mathcal{V}} \lambda_{ss} \left[\sum_{x_s \in \mathcal{X}} \tau_s(x_s) - 1 \right] \\ + \sum_{(s,t) \in \mathcal{E}} \left\{ \sum_{x_s} \lambda_{t,s}(x_s) \left[\tau_s(x_s) - \sum_{x_t \in \mathcal{X}} \tau_{s,t}(x_s, x_t) \right] + \sum_{x_t} \lambda_{s,t}(x_t) \left[\tau_t(x_t) - \sum_{x_s \in \mathcal{X}} \tau_{s,t}(x_s, x_t) \right] \right\}\end{aligned} \tag{29}$$

We have the following theorem which tells that sum-product is the algorithm to solve (29).

Theorem 3.2 *(**Sum-Product** and the **Bethe Problem**) [Wainwright et al., 2008].*
*The sum-product updates are a Lagrangian method for attempting to solve the **Bethe variational problem**:*

1. *For any graph $\mathcal{G}$, any fixed point of the sum-product updates specifies a pair $(\boldsymbol{\tau}^*, \boldsymbol{\lambda}^*)$ such that*

$$\nabla_{\boldsymbol{\tau}} \mathcal{L}(\boldsymbol{\tau}^*, \boldsymbol{\lambda}^*; \boldsymbol{\eta}) = \mathbf{0} \quad \textit{and} \quad \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\tau}^*, \boldsymbol{\lambda}^*; \boldsymbol{\eta}) = \mathbf{0}. \tag{30}$$

2. *For a **tree-structured** Markov random field (MRF), the Lagrangian Equations* (30) *have a **unique solution** $(\boldsymbol{\tau}^*, \boldsymbol{\lambda}^*)$, where the elements of $\boldsymbol{\tau}^*$ correspond to the **exact*** singleton *and* pairwise marginal distributions *of the MRF. Moreover, the **optimal value** of the BVP is equal to the cumulant function $A(\boldsymbol{\eta})$.*

Proof: We just prove the first part. The second part see [Wainwright et al., 2008]. Computing $\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\tau}^*, \boldsymbol{\lambda}^*; \boldsymbol{\eta}) = \mathbf{0}$ results in the marginalization and normalization constraint (23) and (22). Computing $\nabla_{\boldsymbol{\tau}} \mathcal{L}(\boldsymbol{\tau}^*, \boldsymbol{\lambda}^*; \boldsymbol{\eta}) = \mathbf{0}$ results in

$$0 = \nabla_{\boldsymbol{\tau}} \mathcal{L}(\boldsymbol{\tau}^*, \boldsymbol{\lambda}^*; \boldsymbol{\eta})\big|_{s,x_s} = \eta_s(x_s) - \log \tau_s(x_s) + \lambda_{ss} + \sum_{t \in \mathcal{N}(s)} \lambda_{t,s}(x_s)$$

$$0 = \nabla_{\boldsymbol{\tau}} \mathcal{L}(\boldsymbol{\tau}^*, \boldsymbol{\lambda}^*; \boldsymbol{\eta})\big|_{(s,t),x_s,x_t} = \eta_{s,t}(x_s, x_t) - \log \tau_{s,t}(x_s, x_t)$$

$$+ \log \tau_s(x_s) + \log \tau_t(x_t) - \lambda_{t,s}(x_s) - \lambda_{s,t}(x_t)$$

Note that constant terms are absorbed by the slack variable for the positiveness constraint. Thus

$$\log \tau_s(x_s) = \eta_s(x_s) + \lambda_{ss} + \sum_{t \in \mathcal{N}(s)} \lambda_{t,s}(x_s) \tag{31}$$

$$\log \tau_{s,t}(x_s, x_t) = \eta_{s,t}(x_s, x_t) + \log \tau_s(x_s) + \log \tau_t(x_t) - \lambda_{t,s}(x_s) - \lambda_{s,t}(x_t) \tag{32}$$

Substitute (31) into (32), we have

$$\begin{aligned} \log \tau_{s,t}(x_s, x_t) = \eta_{s,t}(x_s, x_t) + \eta_s(x_s) + \eta_t(x_t) + \lambda_{ss} + \lambda_{tt} \\ + \sum_{u \in \mathcal{N}(s) - \{t\}} \lambda_{u,s}(x_s) + \sum_{u \in \mathcal{N}(t) - \{s\}} \lambda_{u,t}(x_t) \end{aligned} \tag{33}$$

Let us define, for each directed edge $t \to s$, an r_s-vector of **messages**

$$M_{t \to s}(x_s) := M_{t,s}(x_s) = \underline{\exp\left(\lambda_{t,s}(x_s)\right)}, \quad (s,t) \in \mathcal{E}, s \in \mathcal{V}, x_s \in \mathcal{X} \tag{34}$$

Then from (31), we have

$$\tau_s(x_s) = \kappa \exp\left(\eta_s(x_s)\right) \prod_{t \in \mathcal{N}(s)} M_{t \to s}(x_s) \tag{35}$$

$$\begin{aligned} \tau_{s,t}(x_s, x_t) = \kappa' \exp\left(\eta_{s,t}(x_s, x_t) + \eta_s(x_s) + \eta_t(x_t)\right) \\ \times \prod_{u \in \mathcal{N}(s) - \{t\}} M_{u \to s}(x_s) \times \prod_{u \in \mathcal{N}(t) - \{s\}} M_{u \to t}(x_t) \end{aligned} \tag{36}$$

Here κ and κ' are positive constants dependent on λ^*_{ss} and λ^*_{tt} so that the pseudo-marginals satisfy normalization conditions. Note that τ_s and $\tau_{s,t}$ so defined are nonnegative.

To conclude, we need to adjust the Lagrange multipliers or messages so that the constraint $\tau_s(x_s) = \sum_{x_t \in \mathcal{X}} \tau_{s,t}(x_s, x_t)$ is satisfied for every edge. Combining (35) and (36) into the marginalization conditions, we can obtain the **Bellman equation** under Bethe variational problem

$$M_{t \to s}(x_s) = \kappa'' \sum_{x_t \in \mathcal{X}} \left\{ \exp\left(\eta_{s,t}(x_s, x_t) + \eta_t(x_t)\right) \prod_{u \in \mathcal{N}(t) \setminus \{s\}} M_{u \to t}(x_t) \right\} \tag{37}$$

Note that (37) is equivalent to (17). ■

From equation (34), we can see that the role of ***message*** in the belief propagation is to **enforce** the ***local consistency*** in the marginal distribution x_s among all pairwise potentials that cover x_s. This can be seen from the Lagrangian multiplier $\lambda_{t,s}$ on local consistency contraint (23).

It should be noted, however, that the above connection between sum-product and the Bethe problem in itself provides **no guarantees on the convergence** of the sum-product updates on graphs with cycles. On the other hand, it provides a principled basis for applying the sum-product algorithm for graphs with cycles, namely as a particular type of *iterative method* for attempting to satisfy *Lagrangian conditions*. The convergence of the algorithm is determined by the **potential strengths** as well as the **topology** of network. In the standard scheduling of the messages, each node applies Equation (37) ***in parallel***.

With the exception of trees and other special cases, the Bethe variational problem (**??**) is usually a **nonconvex problem**, in that $H_{Bethe}(\tau)$ fails to be concave. As a consequence, there are frequently local optima, so that even when using a convergent algorithm, there are no guarantees that it will find the global optimum. For **convexified** Bethe variational problem, see [Wainwright et al., 2008] chapter 7 which introduced the **tree-reweighted Bethe** and corresponding *sum-product algorithm.*

3.4 Reparameterization

Proposition 3.3 *(**Reparameterization Properties of Bethe Approximation**) [Wainwright et al., 2008].*
Let $\boldsymbol{\tau}^* = \left\{\tau_s^*, s \in \mathcal{V}, \tau_{s,t}^*, (s,t) \in \mathcal{E}\right\}$ *denote any optimum of the Bethe variational principle defined by the distribution* $p_{\boldsymbol{\eta}}$*. Then the distribution defined by the fixed point as*

$$p_{\boldsymbol{\tau}^*}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\tau}^*)} \prod_{s \in \mathcal{V}} \tau_s^*(x_s) \prod_{(s,t) \in \mathcal{E}} \frac{\tau_{s,t}^*(x_s, x_t)}{\tau_s^*(x_s)\tau_t^*(x_t)} \tag{38}$$

is a reparameterization of the original distribution $p_{\boldsymbol{\eta}}$ *– that is,* $p_{\boldsymbol{\tau}^*}(\boldsymbol{x}) = p_{\boldsymbol{\eta}}(\boldsymbol{x})$ *for all* $\boldsymbol{x} \in \mathcal{X}^m$*.*

Note that this type of reparameterization is possible only because the exponential family is defined by an *overcomplete* set of sufficient statistics, involving the indicator functions (18) and (19). Moreover, the reparameterization viewpoint provides some insight into the ***approximation error***: that is, the difference between the exact marginals μ_s of $p_{\boldsymbol{\eta}}(\boldsymbol{x})$ and the approximations τ_s^* computed by the sum-product algorithm.

We can show that the reparameterization characterization (38) enables us to specify, for ***any pseudo-marginal*** $\boldsymbol{\tau}$ in the interior of $\mathcal{L}(\mathcal{G})$, a distribution $p_{\boldsymbol{\eta}}(\boldsymbol{x})$ for which $\boldsymbol{\tau}$ is a fixed point of the sum-product algorithm.

4 Expectation propagation

Expectation Propagation (EP) [Seeger, 2005, Wainwright et al., 2008, Koller and Friedman, 2009] provides a general-purpose framework for approximating ***posterior beliefs*** by exponential family distributions.

The idea of *expectation propagation* is close to a mixture of *belief propagation* and *mean field methods*. Expectation propagation *approximates* the entropy $-A^*$ based on ***term decoupling***. It split the factors in graphical model into **tractable** and **intractable parts**. For tractable parts, it is all *singletons* which is close to mean field method settings, while for the intractable part it is learned via message passing. During the message passing, **the moment matching conditions** in $\mathcal{M}$ are always satisfied, hence called expectation propagation.

See [Seeger, 2005, Wainwright et al., 2008, Koller and Friedman, 2009] for detailed discussions. Also see [Minka, 2013] on the expectation propagation for approximate Bayesian Inference.

5 Mean field approximation

As discussed in Section 3, there are two fundamental difficulties associated with the **variational principle** (9):

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

The **core idea** of ***mean field approaches*** is simple: let us limit the optimization to a subset of distributions for which both $\mathcal{M}$ and A^* are relatively easy to characterize; e.g., perhaps they correspond to a graph with small treewidth. Throughout this section, we refer to any such distribution as "**tractable**." The simplest choice is the family of *product* distributions, which gives rise to the *naive mean field* method.

5.1 Tractable families

We formally define the tractable family: consider an exponential family with a collection $\boldsymbol{\phi} = (\phi_\alpha, \alpha \in \mathcal{I})$ of sufficient statistics associated with the cliques of $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Given a subgraph $\mathcal{F}$, let $\mathcal{I}(\mathcal{F}) \subseteq \mathcal{I}$ be the subset of sufficient statistics associated with cliques of $\mathcal{F}$. The set of all distributions that are *Markov* ***with respect to*** $\mathcal{F}$ is a *sub-family* of the full ϕ-exponential family; it is parameterized by the ***subspace*** *of canonical parameters*

$$\Omega(\mathcal{F}) := \{\boldsymbol{\eta} \in \Omega : \eta_\alpha = 0, \ \forall\, \alpha \in \mathcal{I} \setminus \mathcal{I}(\mathcal{F})\}. \tag{39}$$

By definition, the exponential family in tractable family has some of cliques removed. Given the ϕ-exponential family on graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we can find some of examples of **tractable families**:

- The ***fully factorized*** or ***product form*** distributions on the sub-graph $\mathcal{F}_0 = (\mathcal{V}, \emptyset)$ with all nodes but ***no edges***.

$$p_{\mathcal{F}_0}(\boldsymbol{x}; \boldsymbol{\eta}) = \prod_{s \in \mathcal{V}} p(x_s; \eta_s). \tag{40}$$

  The parameter space of this family of distributions is

$$\Omega(\mathcal{F}_0) := \{\boldsymbol{\eta} \in \Omega : \eta_{s,t} = 0, \ \forall\, (s,t) \in \mathcal{E}\}, \tag{41}$$

  This model is the basis for naive mean field algorithm.

- The ***tree-structured*** distributions on ***spanning tree*** of $\mathcal{T}(\mathcal{G}) = (\mathcal{V}, \mathcal{E}(\mathcal{T}))$:

$$p_{\mathcal{T}}(\boldsymbol{x}; \boldsymbol{\eta}) = \prod_{s \in \mathcal{V}} p(x_s; \eta_s) \prod_{(s,t) \in \mathcal{E}(\mathcal{T})} p(x_s, x_t; \eta_{s,t}). \tag{42}$$

  The parameter space of this family of distributions is

$$\Omega(\mathcal{T}(\mathcal{G})) := \{\boldsymbol{\eta} \in \Omega : \eta_{s,t} = 0, \ \forall\, (s,t) \in \mathcal{E} \setminus \mathcal{E}(\mathcal{T})\}, \tag{43}$$

  which set the canonical parameters to be zero for any edge not in the tree.

Denote the mean parameter space $\mathcal{M}(\mathcal{G}) := \mathcal{M}(\mathcal{G}; \boldsymbol{\phi})$ for ϕ-exponential family on graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Given the tractable sub-graph $\mathcal{F}$, these mean field methods are based on optimizing over the ***subset*** of mean parameters that can be obtained by the *subset* of exponential family densities

$$\mathcal{M}_{\mathcal{F}}(\mathcal{G}) := \mathcal{M}_{\mathcal{F}}(\mathcal{G}; \boldsymbol{\phi}) := \left\{ \boldsymbol{\mu} \in \mathbb{R}^d, \ \ \mathbb{E}_{\boldsymbol{\eta}}\left[\boldsymbol{\phi}(X)\right] = \boldsymbol{\mu}, \quad \text{for some } \boldsymbol{\eta} \in \Omega(\mathcal{F}) \right\} \tag{44}$$

$$= \nabla A(\Omega(\mathcal{F})) \tag{45}$$

The interior of $\mathcal{M}(\mathcal{F}; \boldsymbol{\phi})$ is a subset of interior of $\mathcal{M}(\mathcal{G}; \boldsymbol{\phi})$:

$$\mathcal{M}_{\mathcal{F}}(\mathcal{G}; \boldsymbol{\phi})^{\circ} \subseteq \mathcal{M}(\mathcal{G}; \boldsymbol{\phi})^{\circ}, \quad \forall \mathcal{F} \subseteq \mathcal{G}.$$

$\mathcal{M}_{\mathcal{F}}$ is an ***inner approximation*** to the set $\mathcal{M}$ of realizable mean parameters.

5.2 Mean field lower bound and the problem formulation

By *Fenchels inequality* for conjugate duals, we have the following propositions

Proposition 5.1 ***(Mean Field Lower Bound).***
*Any mean parameter $\boldsymbol{\mu} \in \mathcal{M}^{\circ}$ yields a **lower bound** on the cumulant function:*

$$A(\boldsymbol{\eta}) \geq \langle \boldsymbol{\eta}, \boldsymbol{\mu} \rangle - A^*(\boldsymbol{\mu}) \tag{46}$$

Moreover, equality holds if and only if $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ are dually coupled (i.e., $\boldsymbol{\mu} = \mathbb{E}_{\boldsymbol{\eta}}\left[\boldsymbol{\phi}(X)\right]$).

Normally, A^* is not tractable and lacks explicit form. But restricting to mean field distribution A^* can be factorized by the graph and has explicit form. Define $A^*_{\mathcal{F}} := A^*\big|_{\mathcal{M}_{\mathcal{F}}(\mathcal{G})}$ corresponding to the dual function restricted to the set $\mathcal{M}_{\mathcal{F}}(\mathcal{G})$. This way we can compute the lower bound of log-partition function.

Thus the optimization of ***mean field method*** is to *maximize the lower bound*:

$$\max_{\boldsymbol{\mu} \in \mathcal{M}_{\mathcal{F}}(\mathcal{G})} \ \langle \boldsymbol{\eta}, \boldsymbol{\mu} \rangle - A^*_{\mathcal{F}}(\boldsymbol{\mu}) \tag{47}$$

The corresponding value of $\boldsymbol{\mu} \in \mathcal{M}_{\mathcal{F}}(\mathcal{G})$ is defined to be the **mean field approximation** to the true mean parameters.

Unlike the distribution from Bethe variation problem (28), the distribution obtained from (47) still satisfies the moment matching conditions, hence called "*mean*" field.

5.3 Variational representation of mean field

We can reformulate the mean field optimization (47) using KL divergence

$$\min_{q_{\boldsymbol{\mu}} \in \Delta, \boldsymbol{\mu} \in \mathcal{M}_{\mathcal{F}}(\mathcal{G})} \ \mathbb{KL}\left(q_{\boldsymbol{\mu}} \,\|\, p_{\boldsymbol{\eta}}\right) \tag{48}$$

As compared to (5), the mean field approximation ***projects*** the prior distribution $p_{\boldsymbol{\eta}}$ in exponential families ***into the tractable family*** $\mathcal{M}_{\mathcal{F}}(\mathcal{G})$ by minimizing the KL divergence.

Recall the primal-dual formulation of KL divergence (12), we have the ***variational representation*** of mean field

$$\min_{\boldsymbol{\mu} \in \mathcal{M}_{\mathcal{F}}(\mathcal{G})} \ A(\boldsymbol{\eta}) + A^*_{\mathcal{F}}(\boldsymbol{\mu}) - \langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle . \tag{49}$$

5.4 Naive mean field algorithms

In naive mean field algorithm, we focuse on fully factorized *product model* (40) on two graphical models: the Ising model and the Gaussian graphical model.

- The **Ising model** is a pariwise Markov random field

$$p(x_1, \ldots, x_m; \boldsymbol{\eta}) = \exp\left(\sum_{s\in\mathcal{V}} \eta_s\, x_s + \sum_{(s,t)\in\mathcal{E}} \eta_{s,t}\, x_s\, x_t - A(\boldsymbol{\eta})\right) \nu(d\boldsymbol{x}), \tag{50}$$

  where the base measure ν is the counting measure restricted to binary variables $\{0,1\}^m$.

  The potentials are indicators

$$\begin{aligned}\phi_{s;j}(x_s) &= \mathbb{1}\left\{x_s = j\right\}, \forall s \in \mathcal{V}, j \in \{0,1\}\\ \phi_{st;jk}(x_s, x_t) &= \mathbb{1}\left\{x_s = j \wedge x_t = k\right\}, \forall (s,t) \in \mathcal{E}, j, k \in \{0,1\}\end{aligned}$$

  so mean parameters are marginals

$$\begin{aligned}\mu_{st} &= \mathbb{P}(X_s = 1 \wedge X_t = 1) = \mathbb{E}\left[X_s = 1, X_t = 1\right] = \mathbb{E}\left[X_s X_t\right], \quad \forall s, t\\ \mu_s &= \mathbb{P}(X_s = 1) = \mathbb{E}\left[X_s\right], \quad \forall s\end{aligned}$$

  and $\mathcal{M}(\mathcal{G})$ is the marginal polytope. Let $\mathcal{F}_0$ be the fully disconnected subgraph so the tractable space

$$\mathcal{M}_{\mathcal{F}_0}(\mathcal{G}) = \left\{\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{V}|+|\mathcal{E}|} : \mu_s \in [0,1], \forall s \in \mathcal{V}, \text{ and } \mu_{st} = \mu_t \mu_t, \forall (s,t) \in \mathcal{E}\right\}, \tag{51}$$

  where the constraints $\mu_{st} = \mu_t \mu_t$ arise from the **product** nature of *any* distribution that is Markov with respect to $\mathcal{F}_0$.

  For any product distributions, the dual function (i.e. the negative entropy) $A^*_{\mathcal{F}}$ are fully decomposed into entropies in singletons $\{\mu_s, s \in \mathcal{V}\}$.

$$\begin{aligned}-A^*_{\mathcal{F}} &= -\sum_{s\in\mathcal{V}} \left[\mu_s \log \mu_s + (1-\mu_s)\log(1-\mu_s)\right] \\ &= \sum_{s\in\mathcal{V}} H(\mu_s)\end{aligned} \tag{52}$$

  Combining (51) and (52), we have the **naive mean field** problem

$$A(\boldsymbol{\eta}) \geq \max_{\boldsymbol{\mu}\in[0,1]^m} \sum_{s\in\mathcal{V}} \eta_s\, \mu_s + \sum_{(s,t)\in\mathcal{E}} \eta_{s,t}\, \mu_s\, \mu_t + \sum_{s\in\mathcal{V}} H(\mu_s) \tag{53}$$

  For any $s \in \mathcal{V}$, when all other $\mu_t, t \neq s$ are fixed, this objective function is ***strictly concave*** w.r.t. μ_s. Moreover, the optimal solution $\mu^*_s \in (0,1)$. Take the derivative of the objective function w.r.t. μ_s equal to zero yields the **update**:

$$\mu_s \leftarrow \sigma\left(\eta_s + \sum_{t\in\mathcal{N}(s)} \eta_{s,t}\, \mu_t\right), \forall s \in \mathcal{V} \tag{54}$$

  where $\sigma(x) = (1+\exp(-x))^{-1}$ is the logistic function. We can apply ***coordinate ascent*** for each $s \in \mathcal{V}$. It can be shown that any sequence $\{\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots\}$ generated by the updates (54) is guaranteed to converge to a **local optimum** of the naive mean field problem.

- The **Gaussian graphical model** under *mean parameterization* $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}] \in \mathbb{R}^m$ and $\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^T\right] \in \mathcal{S}_+^m$ is

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^m}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \left(\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^T\right)^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) - \frac{1}{2}\log\det\left|\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^T\right|\right) \tag{55}$$

$$\Leftrightarrow p(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\Theta}) = \exp\left(\langle \boldsymbol{\theta}\,,\, \boldsymbol{x}\rangle - \frac{1}{2}\left\langle \boldsymbol{\Theta}\,,\, \boldsymbol{x}\boldsymbol{x}^T\right\rangle - A(\boldsymbol{\theta}, \boldsymbol{\Theta})\right) \tag{56}$$

Under native mean field approximation, all variables are independent. This is equivalent to say that the covariance matrix $\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^T$ is a diagonal matrix.

$$\mathcal{M}_{\mathcal{F}_0}(\mathcal{G}) = \left\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^m \times \mathcal{S}_+^m : (\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^T) = \operatorname{diag}\left(\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^T\right) \succeq \mathbf{0}\right\}, \tag{57}$$

For any such product distribution, the entropy (negative dual function) has the form:

$$\begin{aligned} -A_{\mathcal{F}}^* &= \frac{m}{2}\log(2\pi e) + \frac{1}{2}\log\det\left|\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^T\right| \\ &= \frac{m}{2}\log(2\pi e) + \frac{1}{2}\sum_{s=1}^m \log(\Sigma_{s,s} - \mu_s^2) \end{aligned} \tag{58}$$

Combining (57) and (58), we have the **native mean field** for GGM

$$\begin{aligned} A(\boldsymbol{\theta}, \boldsymbol{\Theta}) \geq \max_{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^m \times \mathcal{S}_+^m} & \langle \boldsymbol{\theta}\,,\, \boldsymbol{\mu}\rangle - \frac{1}{2}\langle \boldsymbol{\Theta}\,,\, \boldsymbol{\Sigma}\rangle + \frac{1}{2}\sum_{s=1}^m \log(\Sigma_{s,s} - \mu_s^2) + \frac{m}{2}\log(2\pi e) \\ \text{s.t. } & \Sigma_{s,s} - \mu_s^2 > 0, \quad s \in \mathcal{V} \\ & \Sigma_{s,t} = \mu_s \mu_t, \quad (s,t) \in \mathcal{E} \end{aligned} \tag{59}$$

Substituting the condition $\Sigma_{s,t} = \mu_s\mu_t$ into the inner product $\langle \boldsymbol{\Theta}\,,\, \boldsymbol{\Sigma}\rangle = \sum_{s,t} \Theta_{s,t}\Sigma_{s,t}$ and taking the derivative of the objective function w.r.t. μ_s and $\Sigma_{s,s}$ to zero, we have equations

$$\begin{aligned} \frac{1}{2(\Sigma_{s,s} - \mu_s^2)} &= \Theta_{s,s} \\ \frac{\mu_s}{2(\Sigma_{s,s} - \mu_s^2)} &= \theta_s - \sum_{t \in \mathcal{N}(s)} \Theta_{s,t}\mu_t \end{aligned} \tag{60}$$

Thus we have the **naive mean field updates** for Gaussian graphical model:

$$\mu_s \leftarrow \frac{1}{\Theta_{s,s}}\left(\theta_s - \sum_{t \in \mathcal{N}(s)} \Theta_{s,t}\mu_t\right), \forall s \in \mathcal{V} \tag{61}$$

The updates (61) are equivalent, depending on the particular ordering used, to either the ***Gauss-Jacobi*** or the ***Gauss-Seidel methods*** [Golub and Van Loan, 2013] for solving the normal equations $\boldsymbol{\mu} = \boldsymbol{\Theta}^{-1}\boldsymbol{\theta}$.

Note that the actual condition mean based on the Gaussian graphical model (56) on $\mathcal{G}$:

$$\boldsymbol{\mu}_{S|\boldsymbol{x}_{\mathcal{N}}} = \boldsymbol{\Theta}_S^{-1}\left(\boldsymbol{\theta}_S - \boldsymbol{\Theta}_{S,\mathcal{N}}\boldsymbol{x}_{\mathcal{N}}\right),$$

which is the similar as above.

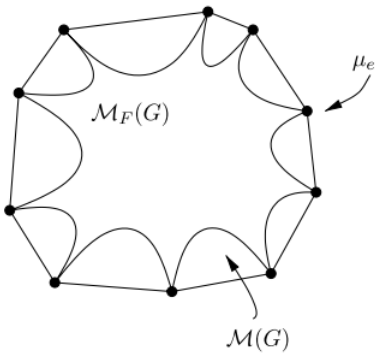

Fig. 5.3 Cartoon illustration of the set $\mathcal{M}_F(G)$ of mean parameters that arise from tractable distributions is a nonconvex inner bound on $\mathcal{M}(G)$. Illustrated here is the case of discrete random variables where $\mathcal{M}(G)$ is a polytope. The circles correspond to mean parameters that arise from delta distributions, and belong to both $\mathcal{M}(G)$ and $\mathcal{M}_F(G)$.

Figure 2: The mean field parameter space $\mathcal{M}_{\mathcal{F}}(\mathcal{G})$ is an non-convex inner bound for $\mathcal{M}(\mathcal{G})$.

5.5 Nonconvexity of mean field

An important fact about the **mean field** approach is that the variational problem (47) may be ***nonconvex***, so that there may be **local minima**, and the mean field updates can have multiple solutions. The source of this nonconvexity can be understood in different ways, depending on the formulation of the problem. See [Wainwright et al., 2008] for detailed discussions.

The geometric perspective on the set $\mathcal{M}(\mathcal{G})$ and its inner approximation $\mathcal{M}_{\mathcal{F}}(\mathcal{G})$ reveals that more generally, mean field optimization is ***always nonconvex*** for ***any exponential family*** in which the state space $\mathcal{X}^m$ is finite. This is becase $\mathcal{M}(\mathcal{G}) = \text{conv}(\phi_\alpha, \alpha \in \mathcal{I})$ is a ***convex hull*** spanned by potentials and $\mathcal{M}_{\mathcal{F}}(\mathcal{G}) \subset \mathcal{M}(\mathcal{G})$ is a strict subset of $\mathcal{M}(\mathcal{G})$. So it must be a nonconvex set. Figure 2 illustrates their relationship.

Consequently, *nonconvexity* is an *intrinsic property* of mean field approximations.

References

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Jun S Liu and Jun S Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

Thomas P Minka. Expectation propagation for approximate bayesian inference. *arXiv e-prints*, pages arXiv–1301, 2013.

Matthias Seeger. Expectation propagation for exponential families. Technical report, 2005.

Alexander Shapiro. Monte carlo sampling methods. *Handbooks in operations research and management science*, 10:353–425, 2003.

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.