Graphical Models, ExpFam, Variational Inference Chapter 3/4.1

Wittawat Jitkrittum[1]
Heiko Strathmann[1]

Gatsby Machine Learning Journal Club

16 Feb 2015

[1] Equal contribution

3.2 Basics of ExpFam (pp39)

- Random vector $(X_1, \ldots, X_m) \in \mathcal{X}^m$
- Sufficient statistics: $\phi_\alpha : \mathcal{X}^m \to \mathbb{R}$, stack to ϕ
- Associated canonical parameters $\theta = (\theta_\alpha, \alpha \in \mathcal{I})$
- Density

$$p_\theta(x_1, \ldots, x_m) = \exp\left(\langle \theta, \phi(x) \rangle - A(\theta)\right)$$

- Convex cumulant

$$A(\theta) = \log \int_{\mathcal{X}^m} \exp\langle \theta, \phi(x) \rangle \nu(dx)$$

- Convex parameter space

$$\Omega = \{\theta \in \mathbb{R}^d \mid A(\theta) < +\infty\}$$

- Minimal: No $a \in \mathbb{R}^d \backslash \{\mathbf{0}\}$ such that $\sum_{\alpha \in \mathcal{I}} a_\alpha \phi_\alpha(x)$ is constant.
- Overcomplete: Not minimal. There is affine subset of θs associated with same density. Useful for sum-product.

3.1 ExpFam via Maximum Entropy (pp37)

$$p^* := \arg\max_{p \in \mathcal{P}} H(p)$$

$$\text{subject to } \mathbb{E}_p\left[\phi_\alpha(X)\right] = \hat{\mu}_\alpha$$

$$\hat{\mu}_\alpha := \frac{1}{n}\sum_{i=1}^{n} \phi_\alpha(X^i) \text{ for all } \alpha \in \mathcal{I}$$

The optimal p^* takes the form

$$p_\theta(x) \propto \exp\left(\sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(x)\right).$$

3.1 ExpFam via Maximum Entropy (pp37)

Discrete case Lagrangian

$$\mathcal{L} = \underbrace{-\sum_j p_j \log p_j}_{H(p)} + \lambda_0 \underbrace{\left(\sum_j p_j - 1\right)}_{\text{normalisation}} + \sum_\alpha \lambda_\alpha \underbrace{\left(\sum_j p_j \phi_\alpha(X) - \hat{\mu}_\alpha\right)}_{\text{data}}$$

Differentiating

$$\frac{\partial}{\partial p_j}\mathcal{L} = -\log p_j + \lambda_0 + \sum_\alpha \lambda_\alpha \phi_\alpha(X)$$

Setting to zero

$$p_j \propto \exp\left(\sum_\alpha \lambda_\alpha \phi_\alpha(x)\right)$$

Example 3.1 Ising Model (pp41)

- Graph $G = (V, E)$
- Random variables $X_s \in \{0, 1\}$ with node $s \in V$
- $$p_\theta(x) = \exp\left(\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta)\right)$$
- Sufficient statistics

$$\phi(x) = (x_s, s \in V; x_s x_t, (s,t) \in E) \in \mathbb{R}^{|V|+|E|}$$

- Minimal

3.4 Mean Parametrisation (pp52)

- Mean parameter

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] \qquad \text{for } \alpha \in \mathcal{I}$$

- Feasible means

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d \mid \exists p \text{ s.t. } \mathbb{E}_p[\phi_\alpha(X)] = \mu_\alpha \quad \forall \alpha \in \mathcal{I} \right\}$$

(p not necessarily exponential family)

- Discrete case

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d \mid \quad \mu = \sum_{x \in \mathcal{X}^m} \phi(x) p(x) \quad \text{for some } p(x) \geq 0, \right.$$
$$\left. \sum_{x \in \mathcal{X}^m} p(x) = 1 \right\}$$

- Minkowski-Weyl

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d \mid \langle a_j, \mu \rangle \geq b_j \quad \forall j \in \mathcal{J} \right\}$$

3.4 Mean Parametrisation (pp52)

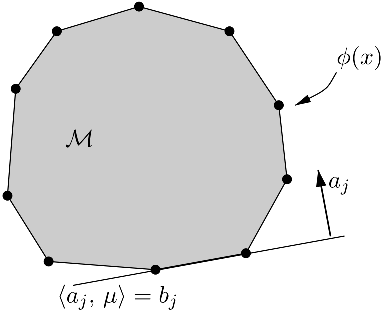


Fig. 3.5 Generic illustration of $\mathcal{M}$ for a discrete random variable with $|\mathcal{X}^m|$ finite. In this case, the set $\mathcal{M}$ is a convex polytope, corresponding to the convex hull of $\{\phi(x) \mid x \in \mathcal{X}^m\}$. By the Minkowski–Weyl theorem, this polytope can also be written as the intersection of a finite number of half-spaces, each of the form $\{\mu \in \mathbb{R}^d \mid \langle a_j, \mu\rangle \geq b_j\}$ for some pair $(a_j, b_j) \in \mathbb{R}^d \times \mathbb{R}$.

Example 3.8 Ising Mean Parameters (pp55)

- Graph $G = (V, E)$
- Random variables $X_s \in \{0, 1\}$ with node $s \in V$
- $$p_\theta(x) = \exp\left(\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta)\right)$$
- Sufficient statistics

$$\phi(x) = (x_s, s \in V; x_s x_t, (s,t) \in E) \in \mathbb{R}^{|V|+|E|}$$

- Mean parameters

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

- Feasible means / correlation polytope

$$\mathcal{M} = \text{conv}\{\phi(x) \mid x \in \{0, 1\}^m\}$$

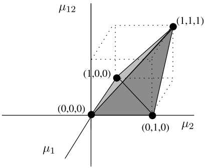
Example 3.1/3.8 Ising Mean Parameters

- Consider $V = \{X_1, X_2\}$, $E = \{(1,2),(2,1)\}$
- Feasible means

$$\begin{aligned}\mathcal{M} &= \text{conv}\left\{(x_1, x_2, x_1x_2) \mid (x_1, x_2) \in \{0,1\}^2\right\} \\ &= \text{conv}\{(0,0,0),(1,0,0),(0,1,0),(1,1,1)\}\end{aligned}$$

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_{12} \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \end{bmatrix}.$$

These four constraints provide an alternative characterization of the 3D polytope illustrated in Figure 3.6.



3.5 Properties of $A = \log \int \exp\langle\theta, \phi(x)\rangle\nu(dx)$ (pp62)

Proposition (3.1 Cumulant)

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)]$$
$$\frac{\partial^2 A}{\partial \theta_\alpha \partial \theta_\beta}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)\phi_\beta(X)] - \mathbb{E}_\theta[\phi_\alpha(X)]\mathbb{E}_\theta[\phi_\beta(X)]$$

Proof: Take derivative under condition that $\int$ and $\frac{\partial}{\partial \theta_\alpha}$ can be switched. Need minimal representation for strictly positive definite Hessian.

Proposition (3.2 Forward mapping to mean parameters)

The gradient mapping $\nabla A : \Omega \to \mathcal{M}$ is one-to-one iff exponential representation is minimal.

If overcomplete (non-identifiable), then many-to-one, affine subset of Ω

Theorem 3.3 (pp65) Moment matching

Theorem (3.3)

In a minimal exponential family, the gradient map ∇A *is onto the interior of* $\mathcal{M}$*, denoted by* $\mathcal{M}^\circ$*, i.e.*

$$\nabla A(\Omega) = \mathcal{M}^\circ.$$

Consequently, for each $\mu \in \mathcal{M}^\circ$ *, there exists some* $\theta = \theta(\mu) \in \Omega$ *such that*

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

.

Proof: Use minimal representation, and properties of convex sets

Remarkable: "All mean parameters $\mu \in \mathcal{M}^\circ$ that are realizable by some distribution can be realized by a member of the exponential family."

Conjugate Duality: $A^*(\mu) := \sup_{\theta\in\Omega} \langle \mu, \theta \rangle - A(\theta)$

Theorem (3.4, pp67)

a) Entropy form

$$A^*(\mu) = \begin{cases} -H(p_{\theta(\mu)}) & \textit{if } \mu \in \mathcal{M}^\circ \\ +\infty & \textit{if } \mu \notin \overline{\mathcal{M}} \end{cases}$$

b) Variational representation

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

c) Above supremum attained by

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

We will need b) later with $A^*(\mu)$ is replaced with a the Bethe entropy.

Conjugate Duality

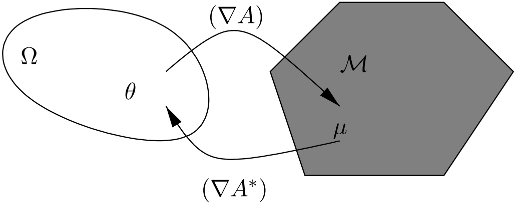


Fig. 3.8 Idealized illustration of the relation between the set Ω of valid canonical parameters, and the set $\mathcal{M}$ of valid mean parameters. The gradient mappings ∇A and ∇A^* associated with the conjugate dual pair (A, A^*) provide a bijective mapping between Ω and the interior $\mathcal{M}^\circ$.

Conjugate Duality: Bernoulli

- $X \in \{0,1\}, \quad \phi(x) = x, \quad A(\theta) = \log(1+\exp(\theta)), \quad \Omega = \mathbb{R}$
- Dual

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

- Differentiate, set to zero, get

$$\mu = \frac{\exp(\theta)}{1+\exp(\theta)}$$

- If $\mu \in (0,1)$

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

- Substitude into A^* gives negative entropy of X with mean parameter μ

$$A^*(\mu) = \mu \log \mu + (1-\mu)\log(1-\mu)$$

- Unbounded otherwise

Summary of Chaper 3

- Exponential family form
- Cannonical and mean parametrisation
- Duality via cumulant and entropy

$$\mu \longrightarrow \boxed{(\nabla A)^{-1}} \xrightarrow{\theta(\mu)} \boxed{-H(p_{\theta(\mu)})} \longrightarrow A^*(\mu)$$

- Variational representation

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

- In practice
  - constraint set $\mathcal{M}$ is hard to characterise
  - negative entropy A^* lacks and explicit form
- Now: replace $\mathcal{M}$, approximate A^*

Section 4.1: Sum-Product and Bethe Approximation

- Explore sum-product or belief propagation (BP) algorithm, an inference algorithm for finding marginals.
- Exact on a tree.
- Can be applied to a loopy graph as well, yielding loopy BP.
- Will see that loopy BP attempts to solve the so-called **Bethe variational problem**.
    - Approximate the variational representation $A(\theta) := \sup_{\mu \in \mathcal{M}} \langle \mu, \theta \rangle - A^*(\mu)$.
    - **First approximation**: approximate $\mathcal{M}$
    - **Second approximation**: approximate $A^*(\mu)$ (Bethe approximation)

4.1 Sum-Product and Bethe Approximation (pp. 76)

Notations for graph $G = (V, E)$

- Domain of X_s is $\mathcal{X}_s = \{0, 1, \ldots r_s - 1\}$. Discrete random variable.

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

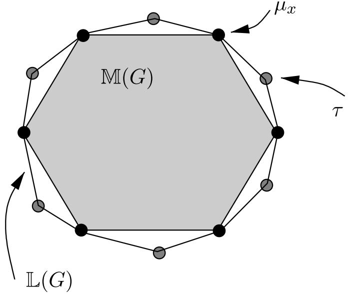
- θ_s is a vector of length r_s.
- $\theta_s(x_s) := \sum_j \theta_{s;j} I_{s;j}(x_s) = \theta_s^\top I_s$ is the k^{th} element of θ_s if $x_s = k$.
  - $I_s = (0, \ldots, 1, \ldots 0)^\top$ with 1 at the k^{th} position.
- Mean parameter $\mu_s = \mathbb{E} I_s$ is the probability vector for x_s.
- **Marginal polytope** (Eq. 4.4):

$$\mathbb{M}(G) := \left\{\mu \in \mathbb{R}^d \mid \exists p \text{ with marginals } \mu_s(x_s), \mu_{st}(x_s, x_t)\right\}$$

- $\mathbb{M}(G)$ requires global consistency i.e., marginalization of the full joint gives $\mu_s(x_s)$ and $\mu_{st}(x_s, x_t)$.

4.1.1 A Tree-Based Outer Bound to $\mathbb{M}(G)$ (pp. 77)

- $\mathbb{M}(G)$ can be written as the intersection of a finite number of half-spaces (**facets**).
- Extremely difficult to list these half-space constraints.
- **Solution**: List only subsets. Obtain a polyhedral outer bound $\mathbb{L}(G)$ on $\mathbb{M}(G)$. The first approximation.



(Misleading picture. The polytope $\mathbb{L}(G)$ has fewer facets and more vertices, but this is difficult to convey in a 2D representation.)

Specification of $\mathbb{L}(G)$ (pp. 78)

- Consider nonnegative $\tau_s(x_s)$ and $\tau_{st}(x_s, x_t)$.
  - τ plays the same role as μ
  - τ for **pseudomarginals**.
- Two constraints for $\mathbb{L}(G)$:
  1. Normalization condition: $\sum_{x_s} \tau_s(x_s) = 1$
  2. Marginalization constraints:

$$\sum_{x'_t} \tau_{st}(x_s, x'_t) = \tau_s(x_s) \text{ for all } x_s$$

$$\sum_{x'_s} \tau_{st}(x'_s, x_t) = \tau_t(x_t) \text{ for all } x_t$$

- The two constraints define $\mathbb{L}(G)$

$$\mathbb{L}(G) = \{\tau \geq 0 \mid \text{both conditions hold}\}\,.$$

- Polytope $\mathbb{L}(G)$ defines a set of **locally consistent** marginal distributions.

Proposition 4.1 $\mathbb{M}(G) \subseteq \mathbb{L}(G)$ (pp. 78-79)

Proposition (4.1)

$\mathbb{M}(G) \subseteq \mathbb{L}(G)$ *holds for any graph* G. *For a tree* T, $\mathbb{M}(T) = \mathbb{L}(T)$.

Proof

- If $\mu \in \mathbb{M}(G)$, then it must be normalized and satisfy marginalization conditions. So, $\mu \in \mathbb{L}(G)$ proving $\mathbb{M}(G) \subseteq \mathbb{L}(G)$.
- For a tree T, need to show $\mathbb{L}(G) \subseteq \mathbb{M}(G)$ by showing $\mu \in \mathbb{L}(G) \Rightarrow \mu \in \mathbb{M}(G)$.
- Assume a tree. By the junction tree theorem,

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

- By running intersection property of the junction tree, local consistency implies global consistency.
- So $\mu \in \mathbb{M}(G)$.

Example of $\tau \in \mathbb{L}(G) \backslash \mathbb{M}(G)$ (pp. 80)

- Consider a simpler example than example 4.1.
- 3 binary variables: $\{x_1, x_2, x_3\}$

$$\tau_s(x_s) := (0.5, 0.5)$$
$$\tau_{st}(x_s, x_t) := \begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}$$

- Equivalently, $p(x_s \neq x_t) = 0.5$.
- $\tau := \{\tau_s, \tau_{st}\}_{s,t}$ give locally consistent marginals i.e., in $\mathbb{L}(G)$.
- But, $p(x_1, x_2, x_3) = 0$ for all configurations.
- So, $\sum_{x_1} \sum_{x_2} p(x_1, x_2, x_3) \neq \tau_3(x_3)$ for example.
- Not globally consistent i.e., not in $\mathbb{M}(G)$.
- In fact, $\mathbb{M}(G)$ is empty (we think..).

4.1.2 Bethe Entropy Approximation (pp. 81-82) I

- Second approximation (Bethe entropy) that underlies the sum-product.
- Approximate A^* in $A(\theta) := \sup_{\mu \in \mathbb{M}(G)} \langle \mu, \theta \rangle - A^*(\mu)$.
- Consider a tree (eq. 4.8)

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

- Marginal distributions correspond to μ under the **canonical overcomplete representation** (sufficient statistics given by indicator functions).

4.1.2 Bethe Entropy Approximation (pp. 81-82) II

- Exact entropy for a tree:

$$
\begin{aligned}
H_{\text{Bethe}}(p_\mu) &= -A^*(\mu) = \mathbb{E}_\mu\left[-\log p_\mu(x)\right] \\
&= \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t)\in E} I_{st}(\mu_{st}).
\end{aligned} \tag{1}
$$

- where

$$
H_s(\mu_s) = -\sum_{x_s \in \mathcal{X}_s} \mu_s(x_s) \log \mu_s(x_s)
$$

$$
I_{st}(\mu_{st}) = \sum_{(x_s, x_t) \in \mathcal{X}_s \times \mathcal{X}_t} \mu_{st}(x_s, x_t) \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}
$$

4.1.2 Bethe Entropy Approximation (pp. 81-82) III

- **Bethe entropy approximation**: Assume H_{Bethe} in Eq. 1 anyway even on a general loopy graph.
- $H_{\text{Bethe}}(p_\mu)$ can be evaluated on $\tau = \{\tau_s, \tau_{st}\}_{s,t}$ that belongs to $\mathbb{L}(G)$.
  - Singleton and pairwise marginals are properly defined.

4.1.3 Bethe Variational Problem (BVP)

With the two ingredients

1. Set $\mathbb{L}(G)$ of locally consistent marginals (outer bound on $\mathbb{M}(G)$).
2. Bethe entropy approximation $\sum_{s\in V} H_s(\mu_s) - \sum_{(s,t)\in E} I_{st}(\mu_{st})$ to $-A^*(\mu)$.

Approximate variational representation of the log-partition

$$\text{(exact) } A(\theta) = \sup_{\mu\in\mathbb{M}(G)} \langle \mu, \theta\rangle - A^*(\mu)$$

with

$$\text{(BVP: 4.16) } A_{\text{Bethe}}(\theta) = \max_{\tau\in\mathbb{L}(G)} \langle \tau, \theta\rangle + \sum_{s\in V} H_s(\tau_s) - \sum_{(s,t)\in E} I_{st}(\tau_{st}).$$

- Solution of BVP admits the same form as the sum-product algorithm.
- $\Rightarrow$ Sum-product algorithm finds a fixed point of BVP.

Lagrangian of BVP (pp. 84)

Define constraint functions for $\mathbb{L}(G)$

$$\text{(normalization) } C_{ss}(\tau) := 1 - \sum_{x_s} \tau_s(x_s)$$

$$\text{(marginalization) } C_{ts}(x_s;\tau) := \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s, x_t)$$

- λ_{ss} := Lagrange multiplier associated with $C_{ss}(\tau) = 0$
- $\lambda_{ts}(x_s)$:= Lagrange multiplier associated with $C_{ts}(x_s;\tau) = 0$
- $\lambda_{ts}(x_s)$ is a vector indexed by x_s. In continuous case, λ_{ts} is a function.

Lagrangian

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

Thm 4.2 (Sum-Product and the Bethe Problem) (pp. 84)

Connection between sum-product and BVP is made precise by

Theorem (4.2)

The sum-product updates are a Lagrangian method for finding a fixed point of BVP.

1. *For any G, any fixed point specifies a pair (τ^*, λ^*) such that*

$$\nabla_\tau \mathcal{L}(\tau^*, \lambda^*; \theta) = 0 \text{ (stationary)}$$
$$\nabla_\lambda \mathcal{L}(\tau^*, \lambda^*; \theta) = 0 \text{ (constraint satisfaction)}$$

2. *For a tree, (τ^*, λ^*) is unique. Elements of τ^* correspond to* ***exact*** *singleton and pairwise marginal distributions. Moreover, the optimal value of BVP is equal to $A(\theta)$ (cumulant function).*

Proof of Theorem 4.2a

- Define $M_{ts}(x_s) := \exp(\lambda_{ts}(x_s))$.
- Find $\nabla_\tau \mathcal{L}(\tau^*, \lambda^*; \theta)$. Equate it to 0. Solve for τ.

$$\tau_s(x_s) = \kappa \exp(\theta_s(x_s)) \prod_{t \in N(s)} M_{ts}(x_s)$$

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

- Solving for $M_{ts}(x_s)$ by using these two equations and others gives

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

  which is the familiar sum-product message from x_t to x_s.
- λ turns out to be log of messages.

Proof of Theorem 4.2b (pp. 273)

Theorem (4.2b)

For a tree T, (τ^*, λ^*) is unique. Elements of τ^* correspond to ***exact*** *singleton and pairwise marginal distributions. Moreover, the optimal value of BVP is equal to $A(\theta)$ (cumulant function).*

- By Proposition 4.1, $\mathbb{L}(T) = \mathbb{M}(T)$.
- H_{Bethe} is exact. So, $-A^* = H_{\text{Bethe}}$.
- So there is no approximation in moving from

$$A(\theta) = \sup_{\mu \in \mathbb{M}(G)} \langle \mu, \theta \rangle - A^*(\mu)$$

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

- The value of the optimized BVP is $A(\theta)$.
- Theorem 3.4(c) implies that the optimum τ^* corresponds to the exact marginal distributions.
- Strict convexity implies that the solution is unique. $\square$

Remark 4.1 (pp. 85)

- Nonnegativity of τ is handled implicitly by logarithmic barriers (in H_s and I_{st}).
- Any optimum $\tau^* > 0$ must satisfy the Theorem 4.2a (stationarity).
- For graphical models where all configurations are given strictly positive mass (ExpFam with finite θ in particular), the sum-product messages stay bounded strictly away from zero.
  - $\Rightarrow$ There is always an optimum $\tau^* > 0$.

4.1.5 Bethe Optima and Reparameterization (pp. 91)

- Recall: the junction tree algorithm takes in a set of potential functions and returns an alternative factorization of the distribution.

$$p(x_{1:m}) = \frac{\prod_{C \in \mathcal{C}} \mu_C(x_C)}{\prod_{S \in \mathcal{S}} \left[\mu_S(x_S)\right]^{d(S)-1}} \text{ Eq 2.12 (pp. 32)}$$

- Same interpretation for the sum-product.
- Any local optimum of BVP specifies a reparameterization of the original distribution p_θ.

Proposition (4.3 Reparameterization by Bethe Approximation)

Let $\tau^* = (\tau_s^*, s \in V; \tau_{st}^*, (s,t) \in E)$ denote any optimum of the BVP defined by the distribution p_θ. At the fixed point,

$$p_{\tau^*}(x) := \frac{1}{Z(\tau^*)} \prod_{s \in V} \tau_s^*(x_s) \prod_{(s,t) \in E} \frac{\tau_{st}^*(x_s, x_t)}{\tau_s^*(x_s)\tau_t^*(x_t)} = p_\theta(x) \text{ (Eq. 4.27)}$$

Comments on the Reparameterization (pp. 92)

- Applied to a graph with cycles.
- $Z(\tau^*)$ is not 1 in general. $Z(\tau^*) = 1$ for a tree.
- Every graph has at least one such reparameterization.
  - Multiple optima of BVP $\Rightarrow$ multiple reparameterizations.
- Possible to derive the approximation error of the sum-product. Detailed not mentioned in the text.
  - Difference of exact marginals μ_s of $p_\theta(x)$ and τ_s^* from the sum-product.

Example 4.3 Fooling the Sum-Product Algorithm (pp. 92-93)

- For any pseudomarginal τ in the interior of $\mathbb{L}(G)$, possible to construct p_θ for which τ is a fixed point of the sum-product.
- The text provides an example where messages are initialized to be uniform distributions.
- Messages do not change with sum-product updates. Already a fixed point.
- Messages give the same pseudomarginals as τ.
- For any discrete MRF in ExpFam with at most one cycle, sum-product has a **unique** fixed point and always converges to it from any initialization..

4.1.6 Bethe and Loop Series Expansions (pp. 94)

- Loop series expansions provide an **exact** representation of the cumulant function $A(\theta)$ as a sum of terms.
- The first term is $A_{\mathsf{Bethe}}(\theta)$ and higher-order terms obtained by adding in so-called **loop corrections**.
- Need some definitions.
- Given an undirected graph $G = (V, E)$ and $\tilde{E} \subseteq E$, let $G(\tilde{E}) = (V(\tilde{E}), \tilde{E})$ be the induced subgraph associated with $\tilde{E}$.
- Degree of $s \in V$ w.r.t. $\tilde{E}$

$$d_s(\tilde{E}) := \left| \{ t \in V \mid (s, t) \in \tilde{E} \} \right| .$$

Generalized Loop (pp. 95)

- Define a **generalized loop** to be a subgraph $G(\tilde{E})$ for which all nodes $s \in V$ have degree $d_s(\widetilde{E}) \neq 1$.
- In other words, either $d_s(\widetilde{E}) = 0$ or $d_s(\tilde{E}) \geq 2$.

4.1 Sum-Product and Bethe Approximation 95

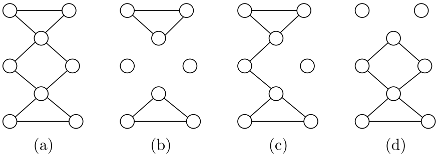


Fig. 4.3 Illustration of generalized loops. (a) Original graph. (b)–(d) Various generalized loops associated with the graph in (a). In this particular case, the original graph is a generalized loop for itself.

- A tree does not have any generalized loops because at least one node s has $d_s(\widetilde{E}) = 1$.

Proposition 4.4 Loop Series Expansion (pp. 96) I

Proposition (4.4)

Consider a pairwise binary MRF. Let $A_{Bethe}(\theta)$ *be the optimized BVP objective evaluated at a BP fixed point* τ*. The cumulant* $A(\theta)$ *is eqal to the loop series expansion:*

$$A(\theta) = A_{Bethe}(\theta) + \log\left(1 + \sum_{\emptyset \neq \widetilde{E} \subseteq E} \beta_{\widetilde{E}} \prod_{s \in V} \mathbb{E}_{\tau_s}\left[(X_s - \tau_s)^{d_s(\widetilde{E})}\right]\right)$$

$$\beta_{\widetilde{E}} := \prod_{(s,t) \in \widetilde{E}} \beta_{st} \text{ and } \beta_{st} := \frac{\tau_{st} - \tau_s \tau_t}{\tau_s(1-\tau_s)(1-\tau_t)}$$

- β_{st} defines an edge weight. $\beta_{\widetilde{E}}$ defines a subgraph weight.

Proposition 4.4 Loop Series Expansion (pp. 96) II

- Equivalently, can also sum over all subsets of E.
- Note the dth central moments of a Bernoulli variable

$$\mathbb{E}_{\tau_s}\left[(X_s - \tau_s)^d\right] = (1-\tau_s)(-\tau_s)^d + \tau_s(1-\tau_s)^d$$

- If $d_s(\widetilde{E}) = 1$ for an $s \in V$ (i.e., a tree), the associated term in the expansion vanishes.
- Only generalized loops $\widetilde{E}$ lead to nonzero terms in the expansion.
- Provide an alternative proof that $A(\theta) = A_{\text{Bethe}}(\theta)$ for a tree.

Comments on Loop Series Expansion (pp. 96,98)

- Computation of full sequence of loop corrections is intractable.
- Same problem as in computing $A(\theta)$ in a loopy graph.
- Any fully connected graph with $n \geq 5$ has more than 2^n generalized loops.
- Can improve approximation to $A(\theta)$ by accounting for a small set of loop corrections.
- The expansion has a generalization for factor graphs. Ref: [51, 224].

Summary of Section 4.1

- Sum-product and its connection to Bethe approximation.
- Sum-product tries to solve Bethe variational problem, a relaxed form of variational representation of $A(\theta)$:

(BVP: 4.16) $A_{\text{Bethe}}(\theta) = \max_{\tau \in \mathbb{L}(G)} \langle \tau, \theta \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}).$

- At a fixed point τ^*, it reparameterize the distribution p_θ

$p_{\tau^*}(x) := \frac{1}{Z(\tau^*)} \prod_{s \in V} \tau_s^*(x_s) \prod_{(s,t) \in E} \frac{\tau_{st}^*(x_s, x_t)}{\tau_s^*(x_s) \tau_t^*(x_t)} = p_\theta(x)$ (Eq. 4.27)

- The cumulant $A(\theta)$ has a loop series expansion.

- Two ways in which BVP is an approximate to the exact variational principle
  1. Only approximate the entropy or $-A^*$.
  2. Use outer bound $\mathbb{L}(G)$ instead of marginal polytop $\mathbb{M}(G)$
- The accuracy can be strengthened by improving either one.
- BVP approximation is based on trees.
- Kikuchi and Hypertree-based methods follow the same principle as BVP by using hypertrees.
  - A (hyper)edge involves more than two vertices.

References I

- https://www.eecs.berkeley.edu/~wainwrig/Papers/WaiJor08_FT