On the Relationship Between the Minimum of the Bethe Free Energy Function of a Factor Graph and Sum-Product Algorithm Fixed Points

Yuwen Huang and Pascal O. Vontobel

Department of Information Enginerring
The Chinese University of Hong Kong
yuwen.huang@ieee.org, pascal.vontobel@ieee.org

Outline

Outline

- **Overview of the main results**

  Standard normal factor graphs (S-NFGs)

  The sum-product algorithm (SPA)

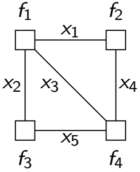
  The primal and dual formulations of the Bethe partition function

  Comparing different dualizations

  Comparison of Yedidia et al.'s results and our results

Overview of standard factor graphs (S-FGs)

- The standard factor graph (S-FG) N consists of
  1. **nonnegative-valued** local functions $f_1, \ldots, f_4$;
  2. edges $1, \ldots, 5$;
  3. alphabets $\mathcal{X}_1, \ldots, \mathcal{X}_5$ for variables $x_1, \ldots, x_5$, respectively.



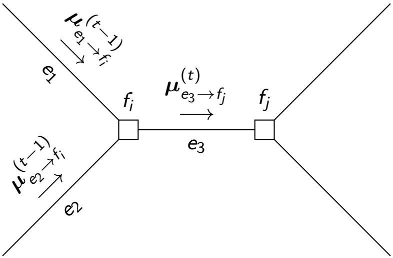
- The global function for N:

$$g(x_1, \ldots, x_5) \triangleq f_1(x_1, x_2, x_3) \cdot f_2(x_1, x_4) \cdot f_3(x_2, x_5) \cdot f_4(x_3, x_4, x_5).$$

- We want to approximate the **partition function** of N:

$$Z(\mathsf{N}) \triangleq \sum_{x_1 \in \mathcal{X}_1, \ldots, x_5 \in \mathcal{X}_5} g(x_1, \ldots, x_5).$$

Overview of the sum-product algorithm (SPA)



Let $e_3 = (f_i, f_j) \in \mathcal{E}$. The message $\boldsymbol{\mu}^{(t)}_{e_3 \to f_j}$ is updated based on

$$\mu^{(t)}_{e_3 \to f_j}(x_{e_3}) \propto \sum_{x_{e_1}, x_{e_2}} f_i(x_{e_1}, x_{e_2}, x_{e_3}) \cdot \mu^{(t-1)}_{e_1 \to f_i}(x_{e_1}) \cdot \mu^{(t-1)}_{e_2 \to f_i}(x_{e_2}).$$

Overview of the main results

Prior work by Yedidia *et al.* in [Yedidia et al., 2005]:

1. For standard factor graph (S-FG) with **positive-valued** local functions only, all **local minima** of the Bethe free energy function correspond to **SPA fixed points**.

Our work:

1. By slightly modifying the S-FG with **nonnegative-valued** local functions if necessary, we relate the **global minimum** of the Bethe free energy function to **an SPA fixed point**.

2. The result is mainly based on a **dual** formulation of the Bethe partition function.

Outline

Introduction to S-NFGs

- Global multivariate function **factors** into a product of local functions.
- Many inference problems can be formulated as computing the **marginals** and **partition function** of the global functions.
- S-NFGs are used to visualize the **factorizations** of the **nonnegative-valued** global functions.
- **Efficient** algorithms take advantage of such factorization.
  - The word "normal" means that the variables are arguments of only **one or two** local functions.

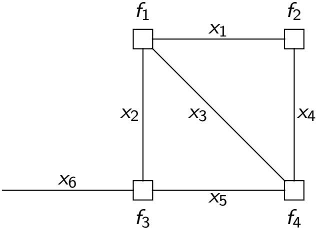
The definition of S-NFGs

The S-NFG N$(\mathcal{F}, \mathcal{E}, \mathcal{X})$ consists of:

1. the graph $(\mathcal{F}, \mathcal{E})$, where an $f \in \mathcal{F}$ denotes a function node and the associated local function;
2. the alphabet $\mathcal{X} \triangleq \prod_{e \in \mathcal{E}} \mathcal{X}_e$.

An S-NFG consists of two kinds of edges:

1. full edges;
2. half edges.



The definition of S-NFGs

Given $\mathsf{N}(\mathcal{F}, \mathcal{E}, \mathcal{X})$, define

1. the local function: $f : \prod_{e \in \partial f} \mathcal{X}_e \to \mathbb{R}_{\geq 0}$;
2. the global function: $g(\boldsymbol{x}) \triangleq \prod_{f \in \mathcal{F}} f(\boldsymbol{x}_f)$;
3. the partition function: $Z(\mathsf{N}) \triangleq \sum_{\boldsymbol{x}} g(\boldsymbol{x})$;
4. the probability mass function (PMF): $p(\boldsymbol{x}) \triangleq g(\boldsymbol{x})/Z(\mathsf{N})$;
5. the marginal:

$$p_{\mathcal{I}}(\boldsymbol{x}_{\mathcal{I}}) \triangleq \sum_{\boldsymbol{x}_{\mathcal{I}^{\mathrm{c}}}} p(\boldsymbol{x}), \qquad \boldsymbol{x}_{\mathcal{I}} \in \mathcal{X}_e^{|\mathcal{I}|}, \mathcal{I} \subseteq \mathcal{E}(\mathsf{N}).$$

From Factor Graph to Normal Factor Graph

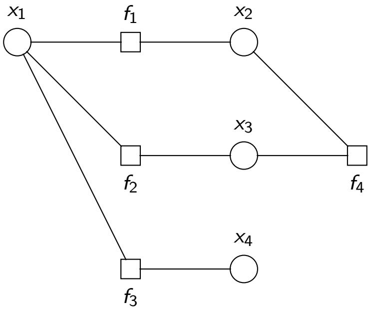


Figure: The factor graph.

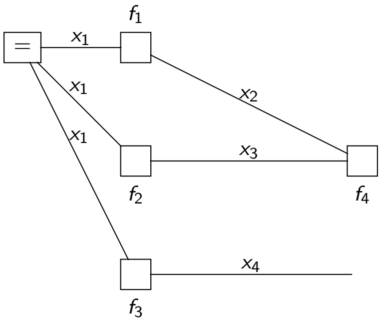


Figure: The associated normal factor graph.

Consider a global function

$$g(x_1, \ldots, x_4) = f_1(x_1, x_2) \cdot f_2(x_1, x_3) \cdot f_3(x_1, x_4) \cdot f_4(x_2, x_3)$$

The partition function and the marginals are **unchanged**.

From NFG with half edges to NFG with full edges

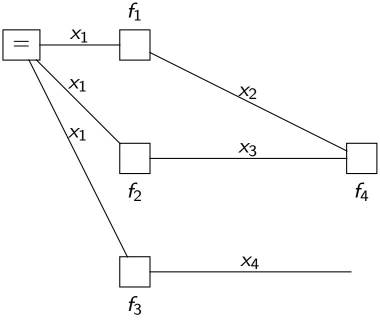


Figure: The normal factor graph with a half edge.

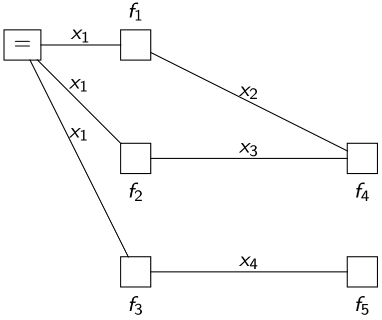


Figure: The normal factor graph with full edges only.

The auxiliary function is defined to be

$$f_5(x_4) \triangleq 1, \qquad x_4 \in \mathcal{X}_4.$$

The partition function and the marginals are **unchanged**.

Outline

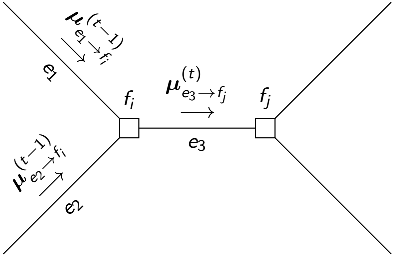
Introduction of the sum-product algorithm (SPA)

- The sum-product algorithm (SPA) is also known as **loopy belief propagation (LBP)**.

- The SPA is a **practical and powerful** way to approximately compute the marginals and the partition function.

- The SPA decoding of **low-density parity-check (LDPC)** codes appears in the 5G telecommunications standard.

The sum-product algorithm (SPA)

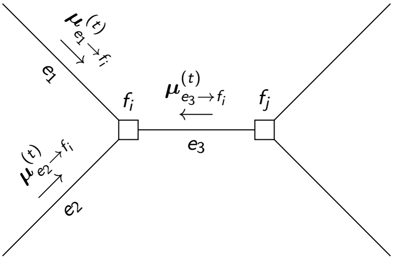


Let t be the iteration index.

1. For $t = 0$, we randomly generate $\boldsymbol{\mu}^{(0)}_{e\to f} \in [0,1]^{|\mathcal{X}_e|} \setminus \{\mathbf{0}\}$.
2. For $t \in \mathbb{Z}_{>0}$ and $e = (f_i, f_j)$, the message from e to f_j is updated according to

$$\mu^{(t)}_{e\to f_j}(x_e) \propto \sum_{\boldsymbol{z}_{f_i}:\, z_e = x_e} f_i(\boldsymbol{z}_{f_i}) \cdot \prod_{e' \in \partial f_i \setminus \{e\}} \mu^{(t-1)}_{e'\to f_i}(z_{e'}) \in \mathbb{R}_{\geq 0}.$$

Evaluate the belief using the messages



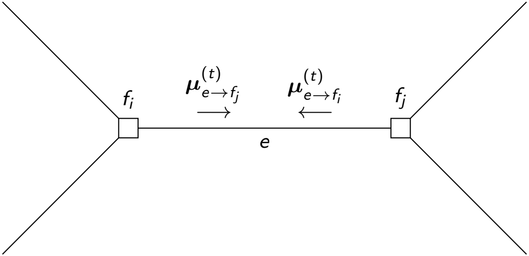
For each $f \in \mathcal{F}$, the belief (a.k.a. pseudo-marginal) is

$$\beta_f^{(t)}(\boldsymbol{x}_f) \triangleq \frac{1}{Z_f(\boldsymbol{\mu}^{(t)})} \cdot f(\boldsymbol{x}_f) \cdot \prod_{e \in \partial f} \mu^{(t)}_{e \to f}(x_e),$$

where the normalization constant is given by

$$Z_f(\boldsymbol{\mu}^{(t)}) \triangleq \sum_{\boldsymbol{x}_f} f(\boldsymbol{x}_f) \cdot \prod_{e \in \partial f} \mu^{(t)}_{e \to f}(x_e).$$

Evaluate the belief using the messages



For each $e = (f_i, f_j)$, the belief (a.k.a. pseudo-marginal) is defined to be

$$\beta_e^{(t)}(x_e) \triangleq \frac{1}{Z_e(\boldsymbol{\mu}^{(t)})} \cdot \mu^{(t)}_{e\to f_i}(x_e) \cdot \mu^{(t)}_{e\to f_j}(x_e),$$

where the normalization constant Z_e is given by

$$Z_e(\boldsymbol{\mu}^{(t)}) \triangleq \sum_{x_e} \mu^{(t)}_{e\to f_i}(x_e) \cdot \mu^{(t)}_{e\to f_j}(x_e).$$

The Sum Product Algorithm (SPA)

Given $\boldsymbol{\mu}^{(t)}$ such that

$$Z_e(\boldsymbol{\mu}^{(t)}) > 0, \qquad e \in \mathcal{E},$$

the approximation of the partition function is defined to be

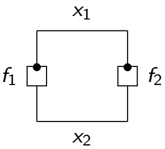
$$Z_{\mathrm{SPA}}(\boldsymbol{\mu}^{(t)}) \triangleq \frac{\prod_f Z_f(\boldsymbol{\mu}^{(t)})}{\prod_e Z_e(\boldsymbol{\mu}^{(t)})}.$$

- For a **cycle-free** S-NFG, the SPA fixed point provides **exact** marginals and partition function.
  - By the **factorization** of the global function, the SPA **reduces the complexity** in computing the marginals and partition function.
- For an S-NFG from **certain** classes of S-NFGs **with cycles**, the SPA fixed-point messages give **good approximations**.

The SPA on an example S-NFG

We associate the matrices $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ with local functions f_1 and f_2, respectively.

$$\boldsymbol{f}_1 \triangleq \Big(f_1(x_1, x_2)\Big)_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2} = \begin{pmatrix} f_1(1,1) & \cdots & f_1(1,|\mathcal{X}_2|) \\ \vdots & \ddots & \vdots \\ f_1(|\mathcal{X}_1|,1) & \cdots & f_1(|\mathcal{X}_1|,|\mathcal{X}_2|) \end{pmatrix},$$



$$\boldsymbol{f}_2 \triangleq \Big(f_2(x_1, x_2)\Big)_{x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2} = \begin{pmatrix} f_2(1,1) & \cdots & f_2(1,|\mathcal{X}_2|) \\ \vdots & \ddots & \vdots \\ f_2(|\mathcal{X}_1|,1) & \cdots & f_2(|\mathcal{X}_1|,|\mathcal{X}_2|) \end{pmatrix},$$

$$\boldsymbol{M} \triangleq \boldsymbol{f}_1 \cdot \boldsymbol{f}_2^{\mathsf{T}}.$$

The partition function equals

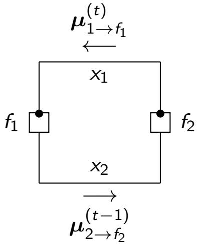
$$Z(\mathsf{N}) = \sum_{x_1, x_2} f_1(x_1, x_2) \cdot f_2(x_1, x_2) = \mathrm{tr}\Big(\boldsymbol{f}_1 \cdot \boldsymbol{f}_2^{\mathsf{T}}\Big) = \mathrm{tr}(\boldsymbol{M}).$$

The SPA on an example S-NFG

The SPA update rule of the message $\mu_{1\to f_1}^{(t)}$:

$$\boldsymbol{\mu}_{1\to f_1}^{(t)} \propto \boldsymbol{f}_2 \cdot \boldsymbol{\mu}_{2\to f_2}^{(t-1)},$$

$$\mu_{1\to f_1}^{(t)}(x_1) = \frac{1}{C_{1\to f_1}^{(t)}} \cdot \sum_{x_2} f_2(x_1, x_2) \cdot \mu_{2\to f_2}^{(t-1)}(x_2),$$



where the normalization constant is given by

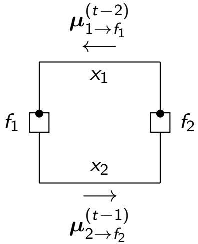
$$C_{1\to f_1}^{(t)} = \sum_{x_1, x_2} f_2(x_1, x_2) \cdot \mu_{2\to f_2}^{(t-1)}(x_2).$$

The SPA on an example S-NFG

The SPA update rule of $\mu^{(t-1)}_{2\to f_2}$:

$$\boldsymbol{\mu}^{(t-1)}_{2\to f_2} \propto \boldsymbol{f}_1^{\mathsf{T}} \cdot \boldsymbol{\mu}^{(t-2)}_{1\to f_1},$$

$$\mu^{(t-1)}_{2\to f_2}(x_2) = \frac{1}{C^{(t-1)}_{2\to f_2}} \cdot \sum_{x_1} f_1(x_1, x_2) \cdot \mu^{(t-2)}_{1\to f_1}(x_1),$$



where the normalization constant is given by

$$C^{(t-1)}_{2\to f_2} = \sum_{x_1, x_2} f_1(x_1, x_2) \cdot \mu^{(t-2)}_{1\to f_1}(x_1).$$

The SPA on an example S-NFG

1. The SPA update rule of $\boldsymbol{\mu}^{(t)}_{1\to f_1}$ is equivalent to applying the **power method** for the matrix $\boldsymbol{M}$:

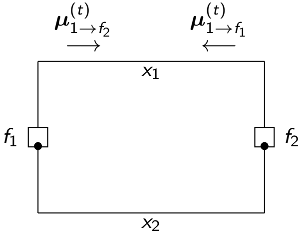
$$\boldsymbol{\mu}^{(t)}_{1\to f_1} \propto \boldsymbol{M}^{\mathsf{T}} \cdot \boldsymbol{\mu}^{(t-2)}_{1\to f_1}, \qquad \boldsymbol{M}^{\mathsf{T}} = \boldsymbol{f}_2 \cdot \boldsymbol{f}_1^{\mathsf{T}}.$$

2. At an SPA fixed point $\boldsymbol{\mu}^{(t)}$:

$$\boldsymbol{\mu}^{(t)}_{1\to f_1} \propto \boldsymbol{M}^{\mathsf{T}} \cdot \boldsymbol{\mu}^{(t)}_{1\to f_1}, \qquad \boldsymbol{\mu}^{(t)}_{1\to f_2} \propto \boldsymbol{M} \cdot \boldsymbol{\mu}^{(t)}_{1\to f_2}.$$

3. The SPA fixed point messages are the **left and right eigenvectors**.

The SPA on an example S-NFG



Belief on edge 1:

$$\beta_1^{(t)}(x_1) = \frac{1}{Z_1(\boldsymbol{\mu}^{(t)})} \cdot \mu^{(t)}_{1\to f_1}(x_1) \cdot \mu^{(t)}_{1\to f_2}(x_1),$$

where the normalization constant Z_1 is given by

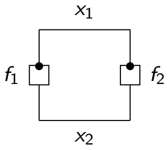
$$Z_1(\boldsymbol{\mu}^{(t)}) = \left(\boldsymbol{\mu}^{(t)}_{1\to f_1}\right)^{\mathsf{T}} \cdot \boldsymbol{\mu}^{(t)}_{1\to f_2}.$$

The SPA on an example S-NFG

Consider specific $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$:

$$\boldsymbol{f}_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{f}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\boldsymbol{M} = \boldsymbol{f}_1 \cdot \boldsymbol{f}_2^{\mathsf{T}} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$



- The largest eigenvalue is **degenerate**.
- The SPA fixed-point messages on edge 1:
$$\boldsymbol{\mu}^{(t)}_{1\to f_1} = (0,\, 1)^{\mathsf{T}}, \quad \boldsymbol{\mu}^{(t)}_{1\to f_2} = (1,\, 0)^{\mathsf{T}}.$$
- With that, the normalization constant equals
$$Z_1(\boldsymbol{\mu}^{(t)}) = \left(\boldsymbol{\mu}^{(t)}_{1\to f_1}\right)^{\mathsf{T}} \cdot \boldsymbol{\mu}^{(t)}_{1\to f_2} = 0.$$
- This poses **a significant issue** when generalizing the results by Yedidia *et al.* [Yedidia et al., 2005].

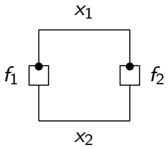
The SPA on an example S-NFG

To address the previous issue, we consider specific $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ such that

$$\boldsymbol{M} = \begin{pmatrix} 1+\delta_2(r) & 1 \\ \delta_1(r) & 1 \end{pmatrix},$$

$$r > 0, \quad \delta_1(r) > 0, \quad \delta_2(r) > 0,$$

$$\lim_{r \downarrow 0} \delta_1(r) = \lim_{r \downarrow 0} \delta_2(r) = 0.$$



- **Perron–Frobenius theory** can be used to show that at the SPA fixed point,
$$\beta_1(x_1) > 0, \quad \forall x_1, \qquad Z_1(\boldsymbol{\mu}^{(t)}) > 0.$$
- Set $r \to 0$. **Different** $\delta_1(r)/\delta_2(r)$ results in **different** SPA fixed-point messages and **different** beliefs $\beta_1(x_1)$.

Outline

The primal formulation

Given $\mathsf{N}(\mathcal{F}, \mathcal{E}, \mathcal{X})$, the **local marginal polytope (LMP)** $\mathcal{B}(\mathsf{N})$ is a collection of vectors

$$\boldsymbol{\beta} \triangleq \left(\{\boldsymbol{\beta}_e\}_{e \in \mathcal{E}}, \{\boldsymbol{\beta}_f\}_{f \in \mathcal{F}} \right)$$

satisfying

1. for $f \in \mathcal{F}$, $\sum_{\boldsymbol{x}_f} \beta_f(\boldsymbol{x}_f) = 1$ **(normalization)**;
2. for $f \in \mathcal{F}$, $\beta_f(\boldsymbol{x}_f) \in \mathbb{R}_{\geq 0}$ **(nonnegativity)**;
3. for $e = (f_i, f_j)$, $\sum_{\boldsymbol{x}_{f_i} : x_e = z_e} \beta_{f_i}(\boldsymbol{x}_{f_i}) = \beta_e(z_e) = \sum_{\boldsymbol{x}_{f_j} : x_e = z_e} \beta_{f_j}(\boldsymbol{x}_{f_j})$ **(local consistency)**.

$\boldsymbol{\beta} \in \mathcal{B}(\mathsf{N})$ is called a collection of **beliefs (a.k.a. pseudo-marginals)**.

The primal formulation

The **Bethe free energy function** is defined to be

$F_{\mathrm{B,p,N}} : \ \mathcal{B}(\mathsf{N}) \to \mathbb{R} \cup \{+\infty\}$

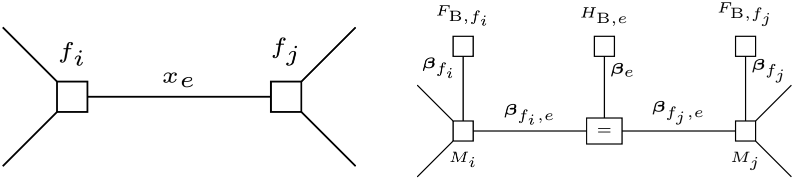
$$\boldsymbol{\beta} \mapsto -\sum_f \underbrace{\Bigg(\underbrace{\sum_{\boldsymbol{x}_f} \beta_f(\boldsymbol{x}_f) \cdot \log f(\boldsymbol{x}_f)}_{U_{\mathrm{B},f}(\boldsymbol{\beta}_f)} + \underbrace{\sum_{\boldsymbol{x}_f} \beta_f(\boldsymbol{x}_f) \cdot \log \beta_f(\boldsymbol{x}_f)}_{-H_{\mathrm{B},f}(\boldsymbol{\beta}_f)} \Bigg)}_{F_{\mathrm{B},f}(\boldsymbol{\beta}_f)}$$

$$- \sum_e \underbrace{\sum_{x_e} \beta_e(x_e) \cdot \log \beta_e(x_e)}_{H_{\mathrm{B},e}(\boldsymbol{\beta}_e)} .$$

The **Bethe approximation of the partition function** $Z(\mathsf{N})$, called the Bethe partition function, is defined to be

$$Z^*_{\mathrm{B,p,N}} \triangleq \exp\Big(- \min_{\boldsymbol{\beta} \in \mathcal{B}(\mathsf{N})} F_{\mathrm{B,p,N}}(\boldsymbol{\beta}) \Big).$$

Factor graphs of the primal formulation



- LHS: part of an S-NFG of interest.
- RHS: part of an NFG whose global function is equal to the **Bethe free energy** function.
  - The global function of this NFG equals the **sum (not the product)** of the local functions.

The primal formulation

When the S-NFG N is **cycle-free**,

1. the function $F_{\mathrm{B,p,N}}(\boldsymbol{\beta})$ is **convex** [Heskes, 2004, Corollary 1];

2. the Bethe partition function $Z^{*}_{\mathrm{B,p,N}}$ satisfies

$$Z^{*}_{\mathrm{B,p,N}} = \exp\left(- \min_{\boldsymbol{\beta}} F_{\mathrm{B,p,N}}(\boldsymbol{\beta}) \right) = Z(\mathsf{N});$$

3. the elements in the collection of beliefs

$$\boldsymbol{\beta}^{*} \in \operatorname{argmin} F_{\mathrm{B,p,N}}(\boldsymbol{\beta})$$

are the marginals induced by N [Yedidia et al., 2005, Proposition 3].

The Primal Formulation

[Yedidia et al., 2005, Theorem 2] **Interior stationary** points of the Bethe free energy function **must be** SPA fixed points with **positive** beliefs and **vice versa**.

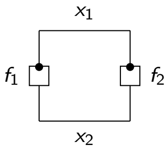
An **interior stationary** point of the Bethe free energy function satisfies two conditions.

1. The belief satisfies $\beta_f(\boldsymbol{x}_f) > 0$ for all $\boldsymbol{x}_f \in \prod_{e\in\partial f} \mathcal{X}_e$ and $f \in \mathcal{F}$.
2. The partial derivatives of the associated Lagrangian function **exist and equal zero** at this point.
   - Recall that we want to find the minimum of the Bethe free energy function over the local marginal polytope.

The primal formulation

Consider specific $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ associated with function nodes f_1 and f_2:

$$\boldsymbol{f}_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \boldsymbol{f}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$



1. To minimize $F_{\mathrm{B,p,N}}$, we set $\beta_{f_1}(0,1) = \beta_{f_2}(0,1) = \beta_{f_1}(1,0) = 0$.

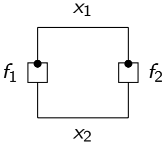
2. The collection of the beliefs that minimize $F_{\mathrm{B,p,N}}$ is **not in the interior** of the local marginal polytope (LMP).

3. We **cannot** apply Yedidia et al.'s results directly.

The primal formulation

To make use of Yedidia et al.'s result, we consider positive $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ instead.

$$\boldsymbol{f}_1 = \begin{pmatrix} 1 & 1 \\ \delta_1(r) & 1 \end{pmatrix}, \ \boldsymbol{f}_2 = \begin{pmatrix} 1 & \delta_2(r) \\ \delta_3(r) & 1 \end{pmatrix},$$



$$r > 0, \quad \delta_1(r) > 0, \quad \delta_1(r) > 0, \quad \delta_3(r) > 0,$$

$$\lim_{r \downarrow 0} \delta_1(r) = \lim_{r \downarrow 0} \delta_2(r) = \lim_{r \downarrow 0} \delta_3(r) = 0.$$

1. Apply [Yedidia et al., 2005, Theorem 3] to this modified S-NFG.
2. Let $r \to 0$.
3. Relate **the global minimum** of the Bethe free energy function to **an SPA fixed point** for the original S-NFG with $\boldsymbol{f}_1 = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\boldsymbol{f}_2 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

The dual formulation

A dual formulation of the Bethe partition function was proposed in [Yedidia et al., 2005, Walsh et al., 2006, Regalia and Walsh, 2007].

Another dual formulation was presented in [Heskes, 2003, Section 4]:

$$Z^{*}_{\mathrm{B,p,N}} = \max \min \ldots$$

- The dual formulation in [Heskes, 2003, Section 4] is not **well defined**. Heskes did not analyze the optimal values' locations.
- Our contribution is to introduce a **well-defined problem** and study the optimal value's locations in [Huang and Vontobel, 2022, Section III].

The definition of the dual formulation

For every edge $e = (f_i, f_j) \in \mathcal{E}$,

$$\boldsymbol{\lambda}_e \triangleq \Big(\lambda_e(x_e)\Big)_{x_e} \in \mathbb{R}^{|\mathcal{X}_e|}, \qquad \boldsymbol{\lambda}_{e,f_i} \triangleq \boldsymbol{\lambda}_e,\ \boldsymbol{\lambda}_{e,f_j} \triangleq -\boldsymbol{\lambda}_e,$$

$$\boldsymbol{\gamma}_e \triangleq \Big(\gamma_e(x_e)\Big)_{x_e} \in \mathbb{R}_{\geq 0}^{|\mathcal{X}_e|}, \qquad \sum_{x_e} \gamma_e(x_e) = 1.$$

Let $\mu_{e\to f}(x_e) = \exp\big(\lambda_{e,f}(x_e)\big) \cdot \sqrt{\gamma_e(x_e)}$. We define

$$Z_e(\boldsymbol{\gamma}_e) \triangleq \sum_{x_e} \underbrace{\Big(\exp\big(\lambda_{e,f_i}(x_e)\big) \cdot \sqrt{\gamma_e(x_e)}\Big)}_{\mu_{e\to f_i}} \cdot \underbrace{\Big(\exp\big(\lambda_{e,f_j}(x_e)\big) \cdot \sqrt{\gamma_e(x_e)}\Big)}_{\mu_{e\to f_j}}$$

$$= \sum_{x_e} \gamma_e(x_e).$$

For every function node $f \in \mathcal{F}$, we define

$$Z_f(\boldsymbol{\gamma}_{\partial f}, \boldsymbol{\lambda}_{\partial f}) \triangleq \sum_{\boldsymbol{x}_f} f(\boldsymbol{x}_f) \cdot \prod_{e\in\partial f} \underbrace{\Big(\exp\big(\lambda_{e,f}(x_e)\big) \cdot \sqrt{\gamma_e(x_e)}\Big)}_{\mu_{e\to f}(x_e)}.$$

The definition of the dual formulation

The dual formulation of the Bethe partition function is

$$Z_{\mathrm{B,d,N}}^{\mathrm{alt},*} \triangleq \sup_{\boldsymbol{\gamma}} \inf_{\boldsymbol{\lambda}} \prod_{f} Z_f(\boldsymbol{\gamma}_{\partial f}, \boldsymbol{\lambda}_{\partial f})$$

$$= \sup_{\boldsymbol{\gamma}} \inf_{\boldsymbol{\lambda}} \frac{\prod_f Z_f(\boldsymbol{\gamma}_{\partial f}, \boldsymbol{\lambda}_{\partial f})}{\prod_e Z_e(\boldsymbol{\gamma}_e)}, \qquad Z_e(\boldsymbol{\gamma}_e) = 1,\ e \in \mathcal{E}.$$

Recall that for SPA fixed-point messages $\boldsymbol{\mu}$, the function Z_{SPA} is

$$Z_{\mathrm{SPA}}(\boldsymbol{\mu}) = \frac{\prod_f Z_f(\boldsymbol{\mu})}{\prod_e Z_e(\boldsymbol{\mu})}, \qquad Z_e(\boldsymbol{\mu}) > 0,\ e \in \mathcal{E},$$

where

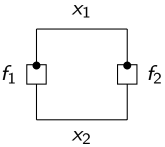
$$Z_f(\boldsymbol{\mu}) = \sum_{\mathbf{x}_f} f(\mathbf{x}_f) \cdot \prod_{e \in \partial f} \mu_{e \to f}(x_e), \qquad f \in \mathcal{F},$$

$$Z_e(\boldsymbol{\mu}) = \sum_{x_e} \mu_{e \to f_i}(x_e) \cdot \mu_{e \to f_j}(x_e), \qquad e = (f_i, f_j) \in \mathcal{E}.$$

The dual formulation for an example S-NFG

Consider specific $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ associated with function nodes f_1 and f_2:

$$\boldsymbol{f}_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \boldsymbol{f}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$



There are $\{\boldsymbol{\gamma}^{(m)}\}$ and $\{\boldsymbol{\lambda}^{(n)}\}$ such that

1. $\{\boldsymbol{\gamma}^{(m)}\}$ and $\{\boldsymbol{\lambda}^{(n)}\}$ **converges** to the location of the **optimal** value

$$Z_{\mathrm{B,d,N}}^{\mathrm{alt},*} = \sup_{\boldsymbol{\gamma}} \inf_{\boldsymbol{\lambda}} \prod_{f} Z_f(\boldsymbol{\gamma}_{\partial f}, \boldsymbol{\lambda}_{\partial f}) = Z_{\mathrm{B,p,N}}^{*} = \exp\left(- \min_{\boldsymbol{\beta}} F_{\mathrm{B,p,N}}(\boldsymbol{\beta}) \right);$$

2. an associated message sequence **converges** to a collection of **SPA fixed-point** messages.

We relate the SPA fixed point to the global minimum of $F_{\mathrm{B,p,N}}$.

Outline

Overview of the main results

Standard normal factor graphs (S-NFGs)

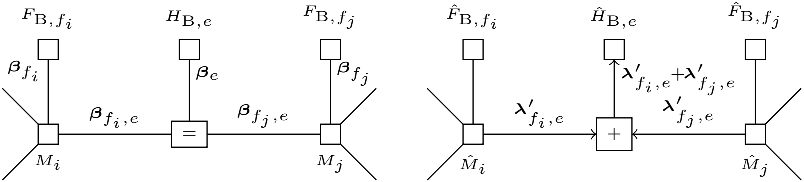
The sum-product algorithm (SPA)

The primal and dual formulations of the Bethe partition function

- **Comparing different dualizations**

Comparison of Yedidia et al.'s results and our results

The dualization by Yedidia et al.



- **Dualizing** the NFG according to [Yedidia et al., 2005, Walsh et al., 2006, Regalia and Walsh, 2007].
- The details are given in [Yedidia et al., 2005, Section VI] and [Regalia and Walsh, 2007, Section V-C].

The dualization by Yedidia et al.

$$\boldsymbol{\beta}^* \in \arg \min_{\boldsymbol{\beta} \in \mathcal{B}(\mathsf{N})} F_{\mathrm{B,p,N}}(\boldsymbol{\beta}).$$

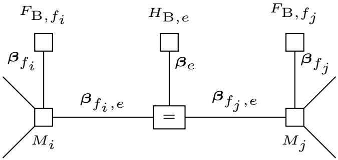
1. Construct the associated Lagrangian function L.

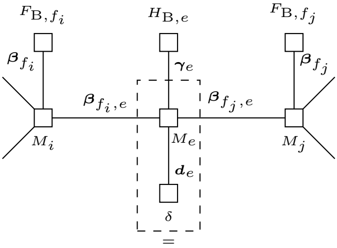
2. The set $\mathcal{B}(\mathsf{N})$ is defined by linear constraints. Thus $\boldsymbol{\beta}^*$ satisfies **the KKT conditions**. [Bertsekas, 2016]

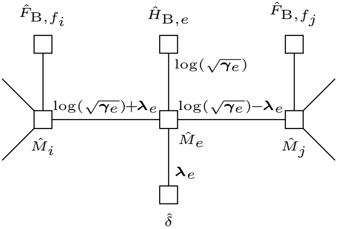
3. **Assume** that $\boldsymbol{\beta}^*$ is in the **interior** of the local marginal polytope $\mathcal{B}(\mathsf{N})$, which implies that
   - the elements in $\boldsymbol{\beta}^*$ are **positive-valued**;
   - the partial derivatives of L **exist** at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$.

4. The KKT conditions imply the dual formulation.

The dualization by Heskes







1. **Replacing** the equal-constraint function node.
2. **Dualizing** the resulting NFG.
3. The details are in [Huang and Vontobel, 2022, Appendix C].

Comparison between these two dualizations

$$\boldsymbol{\beta}^* \in \arg \min_{\boldsymbol{\beta} \in \mathcal{B}(\mathsf{N})} F_{\mathrm{B,p,N}}(\boldsymbol{\beta}).$$

The dualization by Yedidia et al.

1. Works for the S-NFG where $\boldsymbol{\beta}^*$ is **in the interior** of the local marginal polytope $\mathcal{B}(\mathsf{N})$.
2. Relates $\boldsymbol{\beta}^*$ to the SPA fixed point with **positive-valued messages only** when $\boldsymbol{\beta}^*$ is in the interior of LMP.
3. **Does not hold** for some S-NFGs where some entries in $\boldsymbol{\beta}^*$ are **zero-valued**.

The dualization by Heskes.

1. Works for **all** S-NFG N.
2. Allows us to relate $\boldsymbol{\beta}^*$ to the SPA fixed point where some entries in the messages are **zero-valued**.

Outline

Overview of the main results

Standard normal factor graphs (S-NFGs)

The sum-product algorithm (SPA)

The primal and dual formulations of the Bethe partition function

Comparing different dualizations

Comparison of Yedidia et al.'s results and our results

Comparison of the results

Prior work by Yedidia *et al.* in [Yedidia et al., 2005]:

- **Interior stationary** points of the Bethe free energy function are realted to **SPA fixed points** with **positive** beliefs and **vice versa**
- For the S-NFG with **positive-valued** local functions only, all **local minima** of the Bethe free energy function correspond to **SPA fixed points** .

Our work:

- Consider the S-NFG with **nonnegative-valued** local functions. By slightly modifying the S-NFG if necessary, we relate the **global minimum** of the Bethe free energy function to **an SPA fixed point**.

Selected References I

Bertsekas, D. P. (2016).
Nonlinear Programming.
Athena Scientific, Belmont, MA, USA, 3rd edition.

Heskes, T. (2003).
Stable fixed points of loopy belief propagation are local minima of the Bethe free energy.
In *Proc. Neural Information Processing Systems (NIPS)*, pages 359–366, Vancouver, Canada.

Heskes, T. (2004).
On the uniqueness of loopy belief propagation fixed points.
Neural Comput., 16(11):2379–2413.

Huang, Y. and Vontobel, P. O. (2022).
On the relationship between the global minimum of the Bethe free energy function of a factor graph and sum-product algorithm fixed point (extended version).

Selected References II

Regalia, P. A. and Walsh, J. M. (2007).
Optimality and duality of the turbo decoder.
Proc. IEEE, 95(6):1362–1377.

Walsh, J. M., Regalia, P. A., and Johnson, Jr, C. R. (2006).
Turbo decoding as iterative constrained maximum-likelihood sequence detection.
IEEE Trans. Inf. Theory, 52(12):5426–5437.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2005).
Constructing free-energy approximations and generalized belief propagation algorithms.
IEEE Trans. Inf. Theory, 51(7):2282–2312.

Thank you!