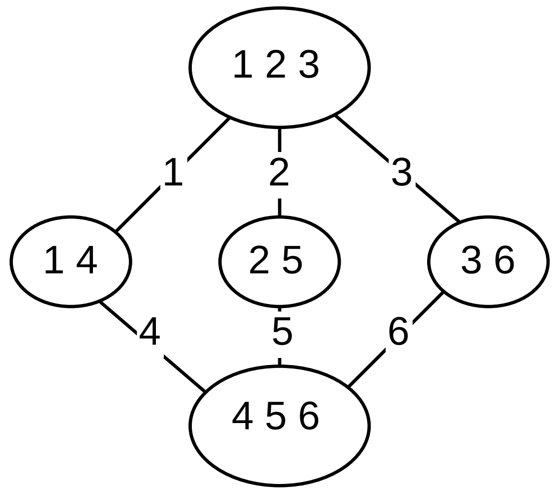
# The Generalized Distributive Law and Free Energy Minimization

**Srinivas M. Aji**
**Rainfinity, Inc.**

**Robert J. McEliece**
**Caltech**




Claude E. Shannon Symposium & Dedication
University of California–San Diego
October 16, 2001

# A Big Tip of the Hat to:

Jonathan Yedidia, William Freeman,
and Yair Weiss, the authors of

*"Bethe Free Energy, Kikuchi Approximations,
and Belief Propagation Algorithms,"*

which inspired this paper.

## Our Goals:

- To understand existing iterative (decoding or otherwise) algorithms better. In particular, *what happens when there are cycles in the underlying graph?*

- To use this understanding to design new and improved iterative algorithms.

# A General Probabilistic Inference Problem

- Variables $\{x_1, \ldots, x_n\}$, $x_i \in A = \{0, 1, \ldots, q-1\}$.
- $\mathcal{R} = \{R_1, \ldots, R_M\}$, a collection of subsets of $\{1, 2, \ldots, n\}$.
- A set of nonnegative "local kernels" $\{\alpha_R(\boldsymbol{x}_R) : R \in \mathcal{R}\}$.

- Example: $n = 4$ and $\mathcal{R} = \{\{1,2\}, \{2,3\}, \{3,4\}, \{1,4\}\}$.

$$\alpha_{\{1,2\}}(x_1, x_2) \geq 0$$
$$\alpha_{\{2,3\}}(x_2, x_3) \geq 0$$
$$\alpha_{\{3,4\}}(x_3, x_4) \geq 0$$
$$\alpha_{\{1,4\}}(x_1, x_4) \geq 0$$

# A General Probabilistic Inference Problem

- Define the "global" probability density function;

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{R \in \mathcal{R}} \alpha_R(\boldsymbol{x}_R).$$

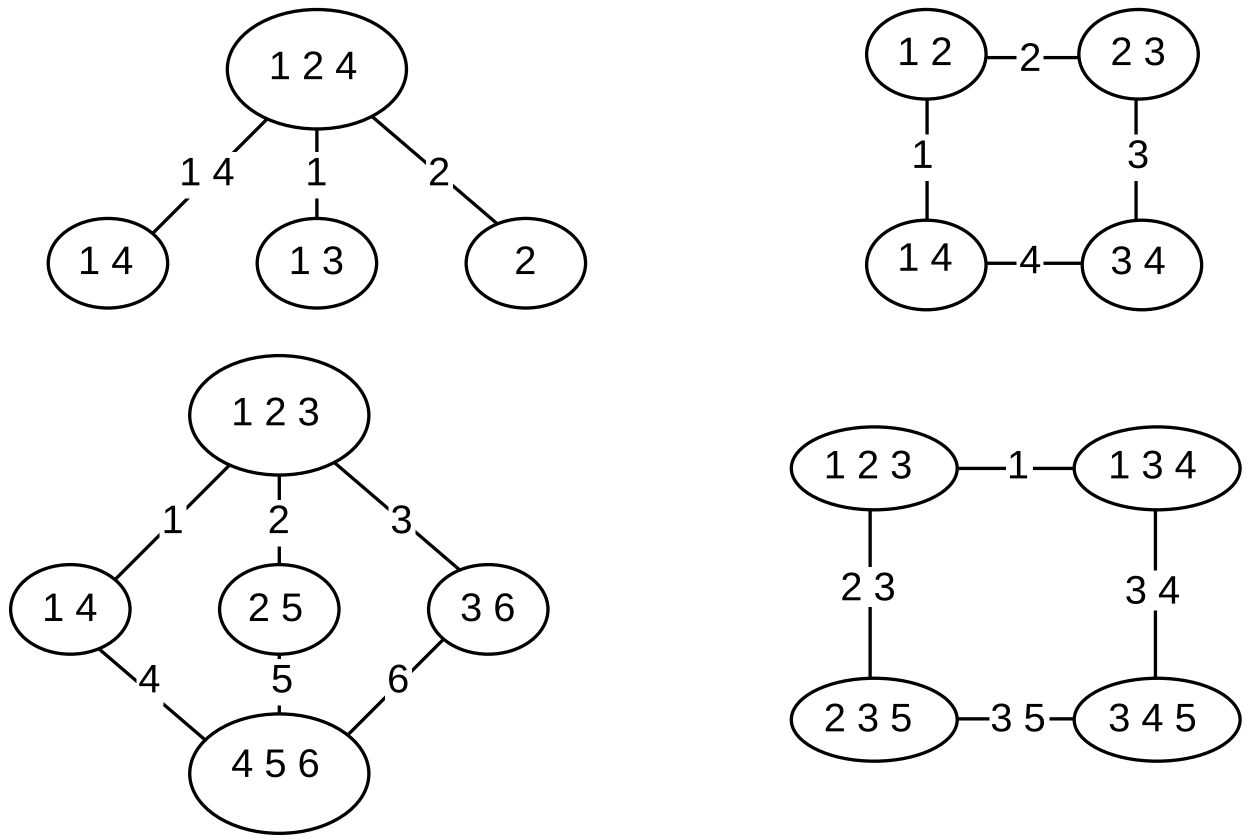
$(Z =$ *Global normalization constant*$)$.

**Problem:** Compute, *exactly or approximately,* $Z$ and some or all of the local marginal densities of the global density:

$$p_R(\boldsymbol{x}_R) = \sum_{\boldsymbol{x}_{R^c} \in A^{R^c}} p(\boldsymbol{x}), \quad \text{for } R \in \mathcal{R}$$

# Solution: Belief Propagation on Junction Graphs

## Junction Graphs $G = (V, E, L)$.

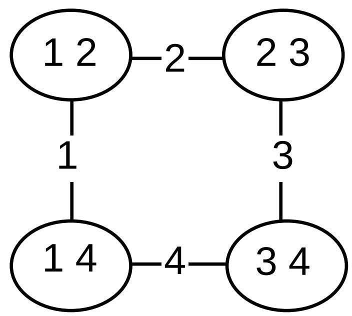



*The subgraph induced by any index* $i \in \{1, 2, \ldots, n\}$ *is a* **tree.**

# Junction Graphs for Solving the Inference Problem

- A junction graph $(V, E, L)$ is called a *junction graph for* $\mathcal{R}$ if $\mathcal{R}$ = the labels of $V$.

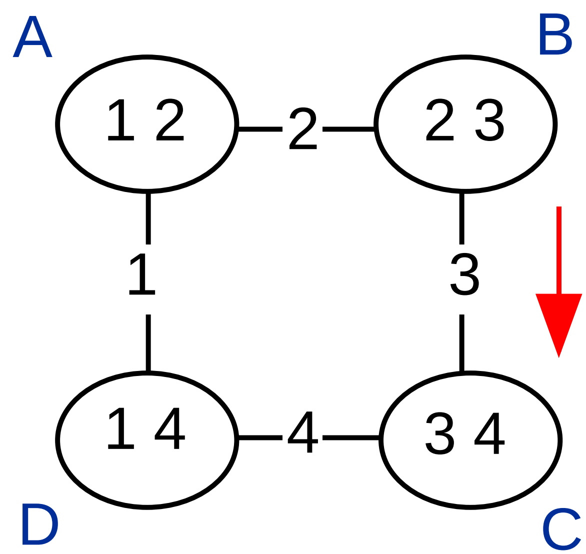
- Example:




is a junction graph for $\{\{1,2\},\{2,3\},\{3,4\},\{1,4\}\}$.

- *It is always possible to find a junction graph (but not necessarily a junction tree) for* $\mathcal{R}$.

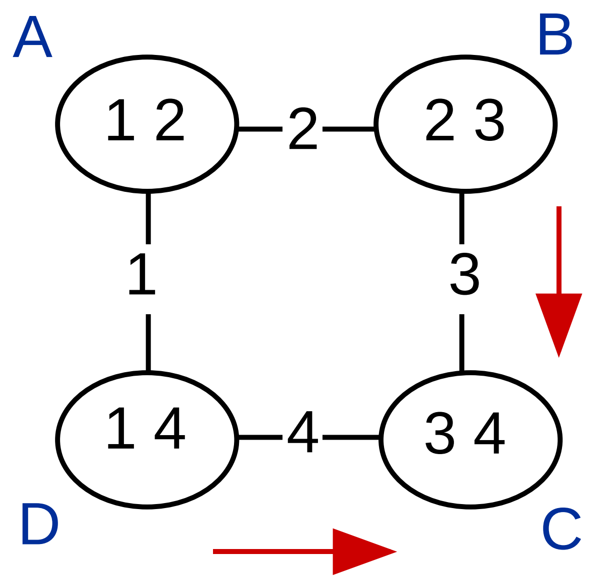
# Belief Propagation on Junction Graphs: the GDL




Example message:

$$m_{B,C}(x_3) \leftarrow K \sum_{x_2} \alpha_{\{2,3\}}(x_2, x_3) m_{A,B}(x_2).$$

# Belief Propagation on Junction Graphs




Example "belief" (approximate marginal density):

$$b_C(x_3, x_4) \leftarrow \frac{1}{Z_C} \alpha_{\{3,4\}}(x_3, x_4) m_{B,C}(x_3) m_{D,C}(x_4)$$

# Belief Propagation on Junction Graphs

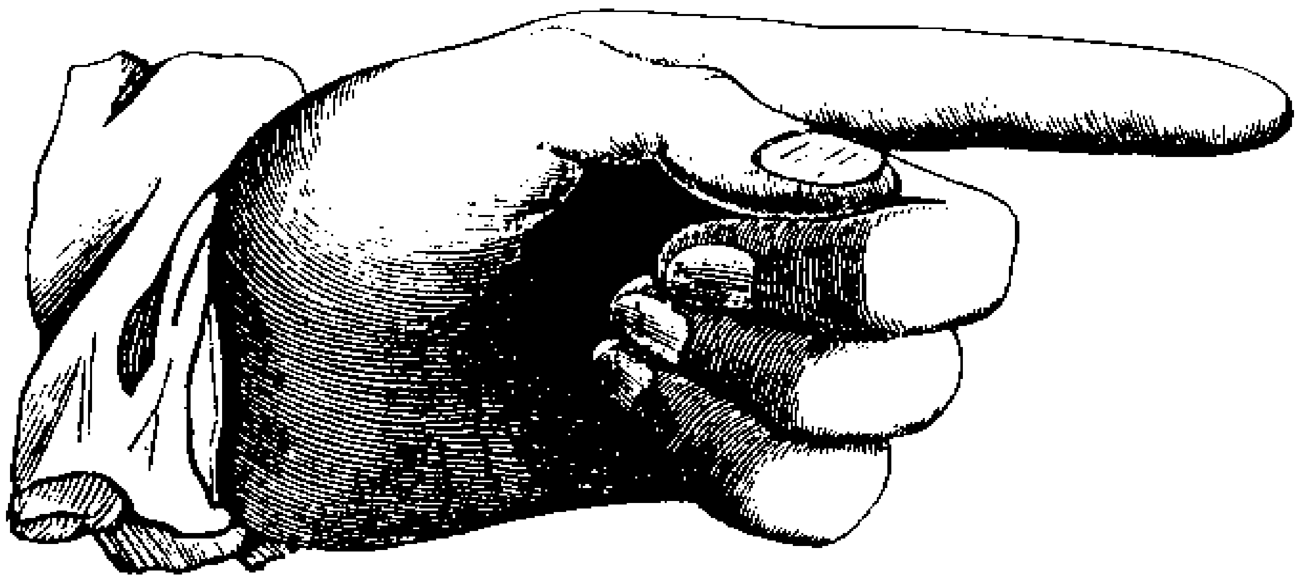
**Theorem.** *If $G$ is a tree (has no cycles), then*

$$b_v(\boldsymbol{x}_{L(v)}) = p_{L(v)}(\boldsymbol{x}_{L(v)})$$

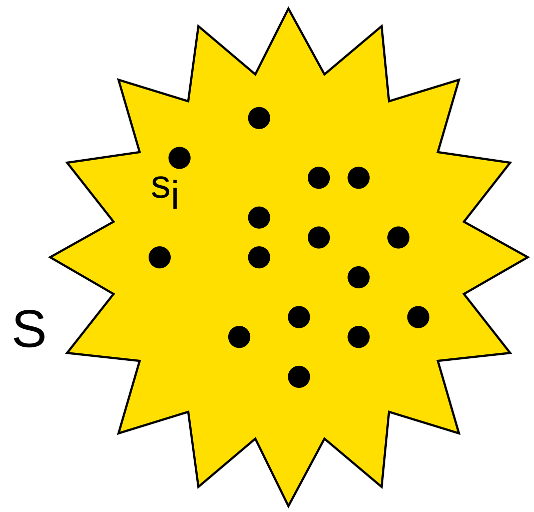
*After a finite number of steps. (In other words, the beliefs converge to the exact desired local marginal probabilities. )*

*But what if $G$ has cycles?*

# And Now, for Something Completely Different . . .

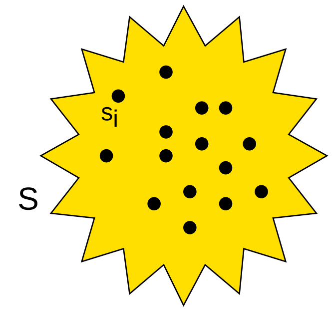


# Some statistical physics




- $S = \{s_1, \ldots, s_n\} = n$ identical particles.
- "Spin" of $s_i = x_i \in A = \{0, 1, \ldots, q-1\}$.
- $E(x_1, \ldots, x_n) =$ energy of state $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$.

# (Helmholtz) Free Energy




- Partition function:

$$Z(\beta) = \sum_{\boldsymbol{x} \in A^n} e^{-\beta E(\boldsymbol{x})}, \qquad \beta = 1/T.$$

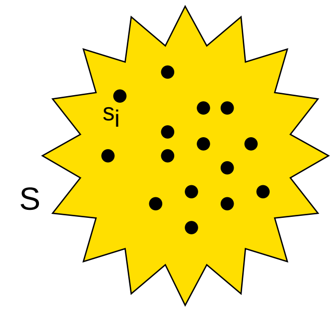
- Free energy:

$$F(\beta) = -\frac{1}{\beta} \ln Z(\beta).$$

*"All macroscopic thermodynamic properties follow from differentiating the free energy."*

- We will take $\beta = 1$.

# Variational Free Energy ($\beta = 1$)




- $p(\boldsymbol{x})$ = Prob. of state $\boldsymbol{x}$.
- Average energy: $U = \sum_{\boldsymbol{x} \in A^n} p(\boldsymbol{x}) E(\boldsymbol{x})$.
- Entropy: $H = -\sum_{\boldsymbol{x} \in A^n} p(\boldsymbol{x}) \ln p(\boldsymbol{x})$.
- Variational free energy:

$$\widetilde{F}(p) = U - H.$$

# A Famous Theorem from Statistical Mechanics

**Theorem.**

$$\widetilde{F}(p) \geq F,$$

*with equality if and only if*

$$p(\boldsymbol{x}) = p^B(\boldsymbol{x}) = \frac{1}{Z} e^{-E(\boldsymbol{x})},$$

*the Boltzmann, or equilibrium, density.*

**Corollary.**

$$F = \min_{p(\boldsymbol{x})} \widetilde{F}(p)$$

$$p^B(\boldsymbol{x}) = \arg\min_{p(\boldsymbol{x})} \widetilde{F}(p).$$

- Suggests a method for computing $F$, but ...

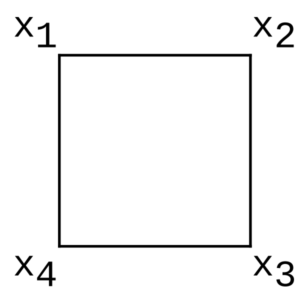
# The "Mean Field" Approximation

$$F_{\text{MF}} = \min\{\widetilde{F}(p) : p(\boldsymbol{x}) = p_1(x_1)p_2(x_2)\dots p_n(x_n)\}.$$

- In general, $F_{\text{MF}} > F$, but ...
- Feynman used this method successfully in 1955 in his paper on the polaron.
- Its use by physicists is widespread
- Too crude for our purposes

# Beyond the Mean Field: The Bethe-Kikuchi Approximation to $\widetilde{F}(p)$

- Often $E(\boldsymbol{x})$ decomposes:




$$E(x_1, x_2, x_3, x_4) = E_{1,2}(x_1, x_2) + E_{2,3}(x_2, x_3) + E_{3,4}(x_3, x_4) + E_{1,4}(x_1, x_4).$$

- In general,

$$E(\boldsymbol{x}) = \sum_{R \in \mathcal{R}} E_R(\boldsymbol{x}_R).$$

# If $E(\boldsymbol{x})$ decomposes, $p^B(\boldsymbol{x})$ factors

$$\text{General } E(\boldsymbol{x}) \Longrightarrow p^B(\boldsymbol{x}) = \frac{1}{Z} e^{-E(\boldsymbol{x})}$$

$$E(\boldsymbol{x}) = \sum_{R \in \mathcal{R}} E_R(\boldsymbol{x}_R) \Longrightarrow p^B(\boldsymbol{x}) = \frac{1}{Z} \prod_{R \in \mathcal{R}} e^{-E_R(\boldsymbol{x}_R)}$$

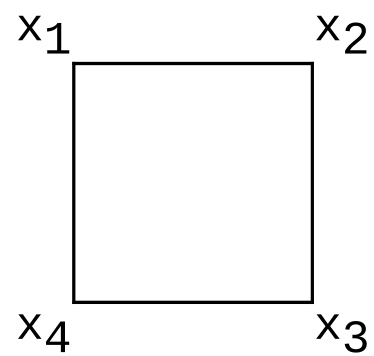
# If $E(\boldsymbol{x})$ decomposes, $p^B(\boldsymbol{x})$ factors

$$\text{General } E(\boldsymbol{x}) \Longrightarrow p^B(\boldsymbol{x}) = \frac{1}{Z} e^{-E(\boldsymbol{x})}$$

$$E(\boldsymbol{x}) = \sum_{R \in \mathcal{R}} E_R(\boldsymbol{x}_R) \Longrightarrow p^B(\boldsymbol{x}) = \frac{1}{Z} \prod_{R \in \mathcal{R}} e^{-E_R(\boldsymbol{x}_R)}$$

$$= \frac{1}{Z} \prod_{R \in \mathcal{R}} \alpha_R(\boldsymbol{x}_R)$$

**Assuming** $E(\boldsymbol{x}) = \sum_{R \in \mathcal{R}} E_R(\boldsymbol{x}_R)$




$$\widetilde{F}(p) = U - H.$$

- In this case the average energy decomposes:

$$U = \sum_R U_R,$$

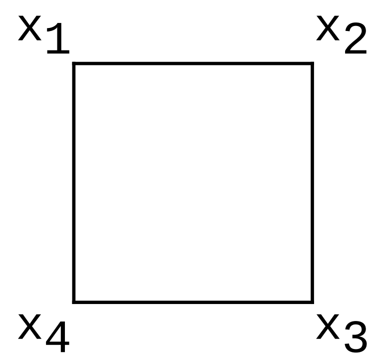
where

$$U_R = \sum_{\boldsymbol{x}_R} p_R(\boldsymbol{x}_R) E_R(\boldsymbol{x}_R).$$

E.g. $U = U_{1,2} + U_{2,3} + U_{3,4} + U_{1,4}$.

*Thus $U$ depends only on the marginals $p_R(\boldsymbol{x}_R)$, and not on the global $p(\boldsymbol{x})$.*

**Assuming** $E(\boldsymbol{x}) = \sum_{R \in \mathcal{R}} E_R(\boldsymbol{x}_R)$

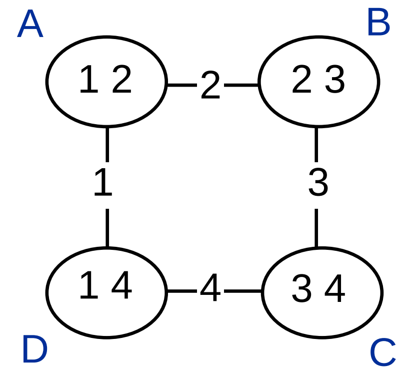



- What about $H(X_1, \ldots, X_n)$? Does it depend only on the marginals $\{p_R(\boldsymbol{x}_R)\}$? NO, but ...

**Theorem.** *If $G = (V, E, L)$ is a junction tree for $\mathcal{R}$, then at Boltzmann equilibrium (the global density):*

$$H(\boldsymbol{X}) = \sum_{v \in V} H(\boldsymbol{X}_v) - \sum_{e \in E} H(\boldsymbol{X}_e).$$

**Example when there are cycles:**




$$H(X_1, X_2, X_3, X_4) \stackrel{?}{=} \\ H(X_1, X_2) + H(X_2, X_3) + H(X_3, X_4) + H(X_1, X_4) \\ -H(X_1) - H(X_2) - H(X_3) - H(X_4).$$

No, but it may be a good approximation. (In essence, this is the BK approximation.)

## The "Bethe-Kikuchi" Approximation to $\widetilde{F}(p)$ With Respect to a Junction Graph $G = (V, E, L)$ for $\mathcal{R}$

$$\widetilde{F}_{BK}(p) = \sum_{v \in V} U_{L(v)}(p_v) - \left( \sum_{v \in V} H(p_v) - \sum_{e \in E} H(p_e) \right)$$

$$= \widetilde{F}_{BK}(\{p_v, p_e\}).$$

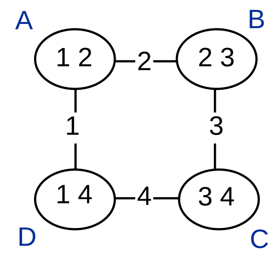
- The BK approximation to the free energy:

$$F_{BK} = \min_{\{p_v, p_e\}} \widetilde{F}_{BK}(\{p_v, p_e\}) \approx F$$

- The BK approximation to the optimizing "beliefs:"

$$\{p_v^{BK}, p_e^{BK}\} = \arg \min_{\{p_v, p_e\}} \widetilde{F}_{BK}(\{p_v, p_e\})$$

**Example:**




$$F_{BK} = \min_{\{p_v, p_e\}} \widetilde{F}_{BK}(\{p_v, p_e\}),$$

subject to:

$$\sum_{x_2} p_{1,2}(x_1, x_2) = p_1(x_1)$$

$$\sum_{x_4} p_{1,4}(x_1, x_4) = p_1(x_1)$$

$$\vdots$$

$$\sum_{x_1} p_1(x_1) = 1$$

$$\vdots$$

# The Main Result

**Theorem.** *Given an instance BP on a junction graph $G = (V, E, L)$, define a corresponding "statistical mechanics problem" via*

$$E_R(\boldsymbol{x}_R) = -\log \alpha_R(\boldsymbol{x}_R).$$

*Then if $G$ is a tree, the unique fixed point $\{b_v, b_e\}$ of BP is the unique global minimum of $\widetilde{F}_{BK}$ (which is convex); and if $G$ has cycles,*

**Any fixed point of the BP algorithm is a stationary point* of $\widetilde{F}_{BK}$ with respect to the same junction graph, and vice-versa.**

* Conjecturally, a local minimum if the fixed point is stable.

**Proof:**

- Set up a Lagrangian:

$$\mathcal{L} = \widetilde{F}_{BK}(\{b_v, b_e\})$$
$$+ \sum_{(u,v)\in E} \sum_{\boldsymbol{x}_{L(u,v)}} \lambda_{u,v}(\boldsymbol{x}_{L(u,v)}) \left( \sum_{\boldsymbol{x}_{L(u)\setminus L(u,v)}} b_v(\boldsymbol{x}_{L(v)}) - b_e(\boldsymbol{x}_{L(u,v)}) \right)$$
$$+ \sum_{v\in V} \mu_v \left( \sum_{\boldsymbol{x}_{L(v)}} b_v(\boldsymbol{x}_{L(v)}) - 1 \right)$$
$$+ \sum_{e\in E} \mu_e \left( \sum_{\boldsymbol{x}_{L(e)}} b_e(\boldsymbol{x}_{L(e)}) - 1 \right).$$

## Proof:

- Set $\frac{\partial \mathcal{L}}{\partial b_v(\boldsymbol{x}_{L(v)})} = 0$:

$$\log b_v(\boldsymbol{x}_{L(v)}) = k_v - E_{L(v)}(\boldsymbol{x}_{L(v)}) - \sum_{u \in N(v)} \lambda_{v,u}(\boldsymbol{x}_{L(u,v)})$$

- Set $\frac{\partial \mathcal{L}}{\partial b_e(\boldsymbol{x}_{L(e)})} = 0$:

$$\log b_e(\boldsymbol{x}_{L(e)}) = k_e - \lambda_{v,u}(\boldsymbol{x}_{L(e)}) - \lambda_{u,v}(\boldsymbol{x}_{L(e)})$$

**Proof:**

- Now use the "translation"

$$E_v(\boldsymbol{x}_{L(v)}) = -\ln \alpha_{L(v)}(\boldsymbol{x}_{L(v)})$$
$$\lambda_{v,u}(\boldsymbol{x}_{L(v,u)}) = -\ln m_{u,v}(\boldsymbol{x}_{L(u,v)}),$$

- With this translation, the stationarity conditions for $\widetilde{F}_{BK}$ are identical to the BP update rules. ▪

## Conclusions:

• Even when cycles are present, BP on a junction graph does something "sensible." (Beliefs converge to a stationary point of the BK approximations to the true marginal probabilities.)

• If $G$ is a tree, *or has only one cycle*, $\widetilde{F}_{BK}$ is convex.

• The junction graph methodology suggests many variations of BP for a given set of local kernels. ($\prod_{i=1}^{n} m_i^{m_i-2}$ junction graphs.)

• This may permit good BP decoding where conventional BP fails (e.g. low-rate RA codes).