Graphs with Cycles:
Ph.D. Thesis of M. Wainwright

Constantine Caramanis

Main Problem and Approach

- Given a distribution $p(\boldsymbol{x})$ on N variables, find marginals.
- Why is this hard?
- When can we solve this problem?

Trees: Marginalization is Easy

- Why is it easy?

- How far does the idea extend?

- How can we extend it farther?

OUTLINE

1. Geometry and Convex Analysis,

2. Tree Reparameterization,

3. Extensions to Hypergraphs and Node Clustering,

4. Bounds on the so–called Log Partition Function.

Factorization

- Hammersley Clifford Theorem:

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C).$$

- Tree Factorization:

$$p(\boldsymbol{x}) = \prod_{s \in V} P_s(x_s) \prod_{(s,t) \in E} \frac{P_{st}(x_s, x_t)}{P_s(x_s) P_t(x_s)}.$$

- Consequence of Tree Factorization:

Local Constraints $\Leftrightarrow$ **Global Constraints**.

Exponential Families and Riemannian Geometry

$$
\begin{aligned}
p(\boldsymbol{x};\theta) &= \exp\{\sum_{\alpha}\theta_{\alpha}\phi_{\alpha}(\boldsymbol{x}) - \Phi(\theta)\}, \\
\Phi(\theta) &= \log\left(\sum_{\boldsymbol{x}}\exp\{\sum_{\alpha}\theta_{\alpha}\phi_{\alpha}(\boldsymbol{x})\}\right).
\end{aligned}
$$

- Minimal Families: $\{\phi_{\alpha}\}$ linearly independent.
- Overcomplete Families: Also will be useful.

Log Partition Function

- Key properties:

$$\begin{aligned}
\Phi(\theta) &= \log \left(\sum_{\boldsymbol{x}} \exp\{\sum_{\alpha} \theta_\alpha \phi_\alpha(\boldsymbol{x})\} \right) \\
\frac{\partial \Phi}{\partial \theta_\alpha}(\theta) &= \mathbb{E}_\theta[\phi_\alpha] \\
\frac{\partial^2 \Phi}{\partial \theta_\alpha \partial \theta_\beta}(\theta) &= \mathrm{cov}_\theta\{\phi_\alpha, \phi_\beta\} = \mathbb{E}_\theta\left[(\phi_\alpha - \mathbb{E}_\theta[\phi_\alpha])(\phi_\beta - \mathbb{E}_\theta[\phi_\beta])\right].
\end{aligned}$$

- Therefore $\Phi(\theta)$ is **convex**.

Legendre Transform and Duality

- We define the Legendre dual of the log partition function:

$$\Psi(\eta) := \sum_{\theta} \{\eta^T \theta - \Phi(\theta)\}.$$

- By convexity, we have stationary conditions for optimum:

$$\eta_\alpha = \eta_\alpha(\widehat{\theta}) = \mathbb{E}_{\widehat{\theta}}[\phi_\alpha],$$

  and thus

$$\Psi(\eta(\widehat{\theta})) = -H(p).$$

- This gives a map $\Lambda : \theta \to \eta$ to dual coordinates:

$$[\Lambda(\theta)]_\alpha = \frac{\partial \Phi(\theta)}{\partial \theta_\alpha} = \mathbb{E}_\theta[\phi_\alpha].$$

Example: **Ising Model**

- $p(\boldsymbol{x};\theta) = \exp\{\sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t - \Phi(\theta)\}$.

- The potential functions are

$$\phi_\alpha(\boldsymbol{x}) = \begin{cases} x_s & \alpha = s \\ x_s x_t & \alpha = (s,t) \in E, \end{cases}$$

- The dual variables are given by:

$$\begin{aligned} \eta_s &= \mathbb{E}_\theta[x_s] = p(x_s = 1;\theta) \\ \eta_{st} &= \mathbb{E}_\theta[x_s x_t] = p(x_s = 1, x_t = 1;\theta). \end{aligned}$$

Flat Manifolds

- An e-flat manifold is the image of an affine set of θ's:

$$\mathcal{M}_e = \{p(\boldsymbol{x};\theta) \,|\, A\theta = b\}.$$

- An m-flat manifold is the image of an affine set of η's:

$$\mathcal{M}_e = \{p(\boldsymbol{x};\theta) \,|\, B\eta = 0\}.$$

I–Projections onto Flat Manifolds

- Kullback–Leibler divergence:

$$D(\theta||\theta^*) = \sum_{\boldsymbol{x}} p(\boldsymbol{x};\theta)[\log p(\boldsymbol{x};\theta) - \log p(\boldsymbol{x};\theta^*)],$$

  becomes using the exponential and mean parameters,

$$\begin{aligned} D(\theta||\theta^*) &= \eta^T(\theta - \theta^*) + \Phi(\theta^*) - \Phi(\theta) \\ &= (\theta^*)^T(\eta^* - \eta) + \Psi(\eta) - \Psi(\eta^*). \end{aligned}$$

- $D(\theta||\theta^*)$ is **convex** in second argument.

I–Projections onto Flat Manifolds (cont'd)

- I-Projection onto an e-flat manifold $\mathcal{M}_e$:

$$\begin{aligned}&\min_\theta D(\theta^*||\theta)\\ &\text{s.t. } \theta \in \mathcal{M}_e.\end{aligned}$$

- Convex objective, subject to linear constraints.

- Gradient condition: $\nabla_\theta D(\theta^*||\theta) = \eta - \eta^*$, which implies

$$[\eta^* - \widehat{\eta}]^T[\theta - \widehat{\theta}] = 0.$$

I–Projections onto Flat Manifolds (cont'd)

- Consider an m-geodesic joining θ and θ^*:

$$\theta(t) = \Lambda^{-1}(\widehat{\eta} + t[\eta^* - \widehat{\eta}]).$$

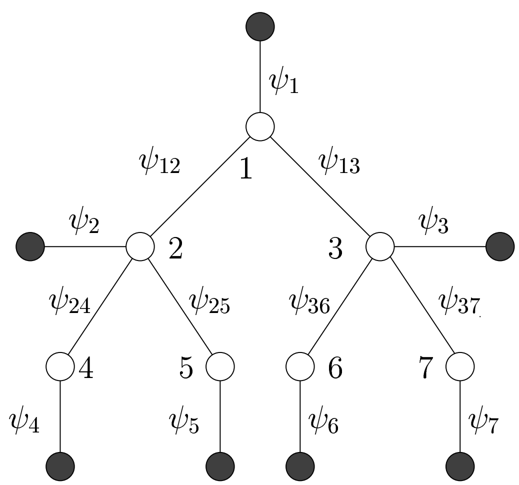
- Meets $\mathcal{M}$ orthogonally:

$$\langle[\theta - \widehat{\theta}], G^{-1}(\widehat{\eta})[\eta^* - \widehat{\eta}]\rangle_{G(\widehat{\theta})} = [\eta^* - \widehat{\eta}]^T[\theta - \widehat{\theta}] = 0.$$

- Also, have Pythagorean relation:

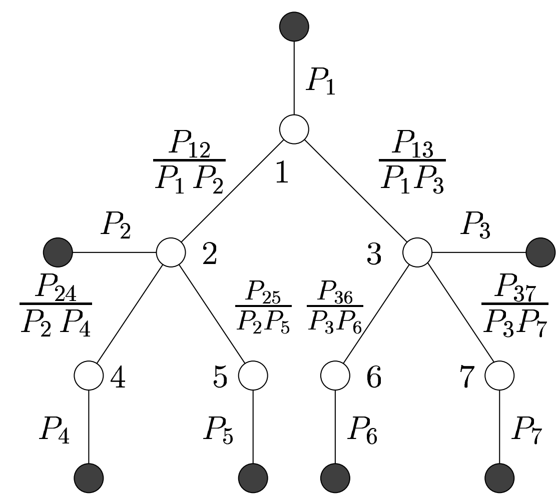
$$D(\theta^*||\theta) = D(\theta^*||\widehat{\theta}) + D(\widehat{\theta}||\theta).$$

Tree Reparameterization



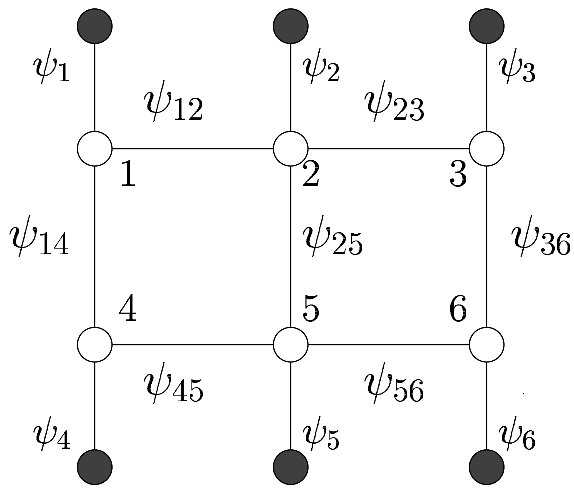
$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t)$$

Tree Reparameterization



$$p(\boldsymbol{x}) = \prod_{s \in V} P_s(x_s) \prod_{(s,t) \in E} \frac{P_{st}(x_s, x_t)}{P_s(x_s) P_t(x_t)}.$$

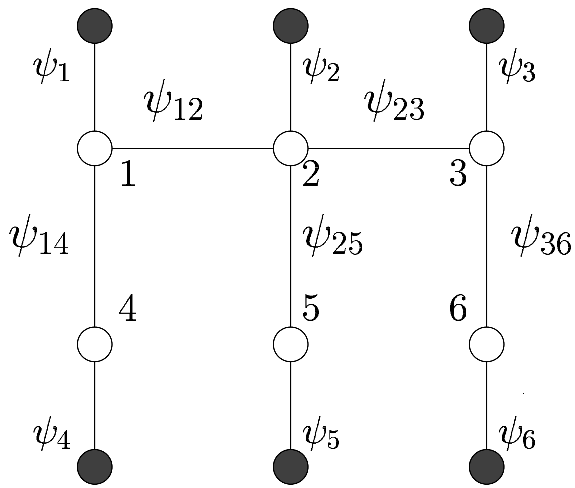
Loopy Graph Reparameterization



- By Hammersley–Clifford we have the factorization:

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t).$$

Find Tree Structured Subgraph



$$\begin{aligned} p(\boldsymbol{x}) &= \frac{1}{Z}\left(\prod_{s=1}^{6}\psi_s(x_s)\prod_{(s,t)\in\mathcal{T}}\psi_{st}(x_s,x_t)\right)\psi_{45}(x_4,x_5)\psi_{56}(x_5,x_6) \\ &= \frac{1}{Z}p^{\mathcal{T}}(\boldsymbol{x})r^{\mathcal{T}}(\boldsymbol{x}), \end{aligned}$$

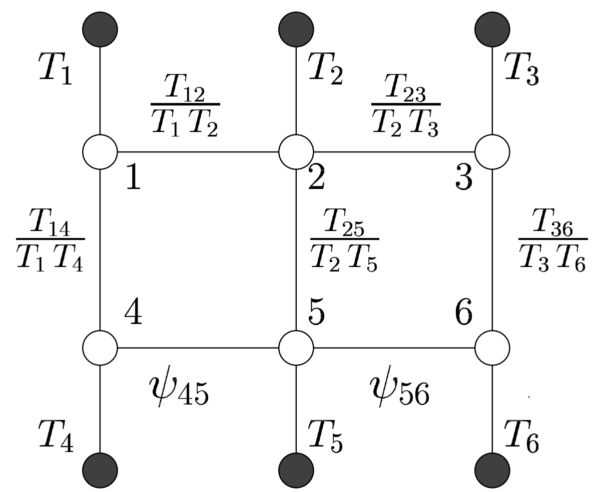
Factor Tree into Pseudomarginals

$$\begin{aligned} p^{\mathfrak{T}}(\boldsymbol{x}) &\propto \prod_{s=1}^{6} \psi_s(x_s) \prod_{(s,t)\in\mathfrak{T}} \psi_{st}(x_s, x_t) \\ &\propto \prod_{s=1}^{6} T_s \prod_{(s,t)\in\mathfrak{T}} \frac{T_{st}}{T_s T_t}, \end{aligned}$$

and therefore, we have:

$$p(\boldsymbol{x}) = \frac{1}{Z} \left(\prod_{s=1}^{6} T_s \prod_{(s,t)\in\mathfrak{T}} \frac{T_{st}}{T_s T_t} \right) \psi_{45}\psi_{56}.$$

Factor Tree into Pseudomarginals



- Next step: Find different subgraph with tree structure, and repeat.

Exponential Families

- We choose an overcomplete family:

$$\mathcal{A} := \{(s;j), (st;jk) \,|\, s,t \in V, (s,t) \in E, \quad 1 \leq j,k \leq m\}.$$

- Then we can define,

$$\begin{aligned} \phi_\alpha(\boldsymbol{x}) &= \delta(x_s = j), \text{ for } \alpha = (s;j) \\ \phi_\alpha(\boldsymbol{x}) &= \delta(x_s = j)\delta(x_t = k), \text{ for } \alpha = (st;jk). \end{aligned}$$

- We have mean (dual) parameters:

$$\begin{aligned} \eta_{s;j} &= P_{s;j} = \mathbb{E}_\theta[\delta(x_s = j)], \\ \eta_{st;jk} &= P_{st;jk} = \mathbb{E}_\theta[\delta(x_s = j)\delta(x_t = k)]. \end{aligned}$$

Some Remarks

- We therefore want to compute:

$$\eta = \Lambda(\theta).$$

  We will see this is easy for trees.

- Given the distribution, $p(\boldsymbol{x};\theta^0)$, then

$$\mathcal{M}(\theta^0) := \{\theta \,|\, p(\boldsymbol{x};\theta) = p(\boldsymbol{x};\theta^0)\},$$

  is an e-flat manifold.

Geometric Interpretation

- We cannot compute $\Lambda(\theta)$ exactly for loopy graphs.
- We can compute it for graph structures.
- Therefore: Reparameterizing according to a tree corresponds to using the pseud–marginals $\{T_s, T_{st}\}$ to approximate $\{\eta_s, \eta_{st}\}$.

Geometric Interpretation

- The marginals must satisfy local consistency constraints:

$$\begin{aligned} \mathsf{TREE}(G) \;&:=\; \{\boldsymbol{T} \,|\, \boldsymbol{T} \in (0,1)^{d(\theta)}, \sum_j T_{s;j} = 1 \;\; \forall s \in V; \\ &\qquad \sum_k T_{st;jk} = T_{s;j}, \;\; \forall (s,t) \in E\}. \end{aligned}$$

- The pseudo–marginals $\boldsymbol{T}$ obtained from reparameterizing w.r.t tree $\mathcal{T}^i$ must be in the set:

$$\begin{aligned} \mathsf{TREE}^i(G) \;&:=\; \{\boldsymbol{T} \,|\, \boldsymbol{T} \in (0,1)^{d(\theta)}, \sum_j T_{s;j} = 1 \;\; \forall s \in V; \\ &\qquad \sum_k T_{st;jk} = T_{s;j}, \;\; \forall (s,t) \in \mathcal{T}^i\}. \end{aligned}$$

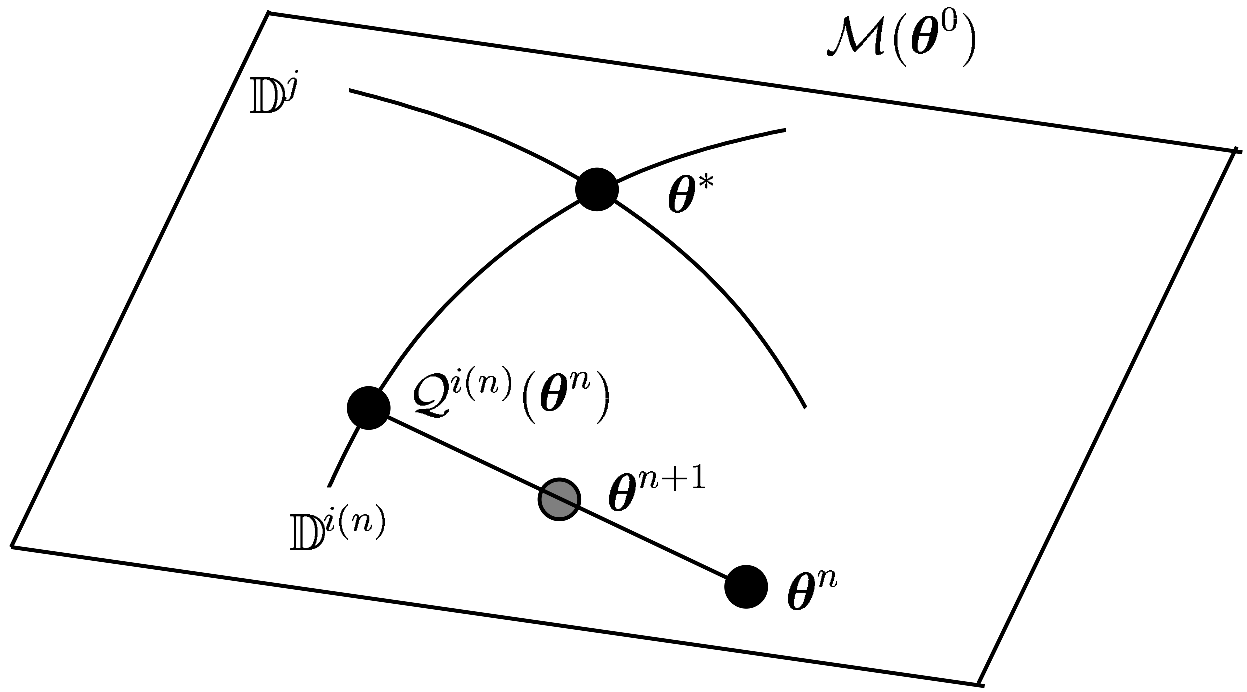
E-Flat and M-Flat Manifolds

- EXPTREE$^i(G)$ and therefore EXPTREE(G) are m-flat.

- Reparameterization according to trees does not change the distribution: Therefore we stay within the e-flat manifold.

- Reparameterizing according to tree $\mathcal{T}^i$, moves within the e–flat manifold, to the m–flat manifold EXPTREE$^i(G)$.

Moving in E-Flat Manifold to M-Flat Manifolds



Projections onto M-Flat Manifolds

- The point in EXPTREE$^i(G)$ to which TRP moves has a projection interpretation.

$$
\begin{aligned}
\mathcal{G}_s(T_s;\theta_s) &= \sum_j T_{s;j}[\log T_{s;j} - \theta_{s;j}] \\
\mathcal{G}_{st}(T_{st};\theta_{st}) &= \sum_{j,k} T_{st;jk}\left\{\log[T_{st;jk}/(\sum_j T_{st;jk})(\sum_k T_{st;jk})] - \theta_{st;jk}\right\} \\
\mathcal{G}(\boldsymbol{T};\theta) &= \sum_{s\in V} \mathcal{G}_s(T_s;\theta_s) + \sum_{(s,t)\in E} \mathcal{G}_{st}(T_{st};\theta_{st}) \\
\mathcal{G}^i(\Pi^i(\boldsymbol{T});\Pi^i(\theta)) &= \sum_{s\in V} \mathcal{G}_s(T_s;\theta_s) + \sum_{(s,t)\in E^i} \mathcal{G}_{st}(T_{st};\theta_{st}).
\end{aligned}
$$

Variation Functional

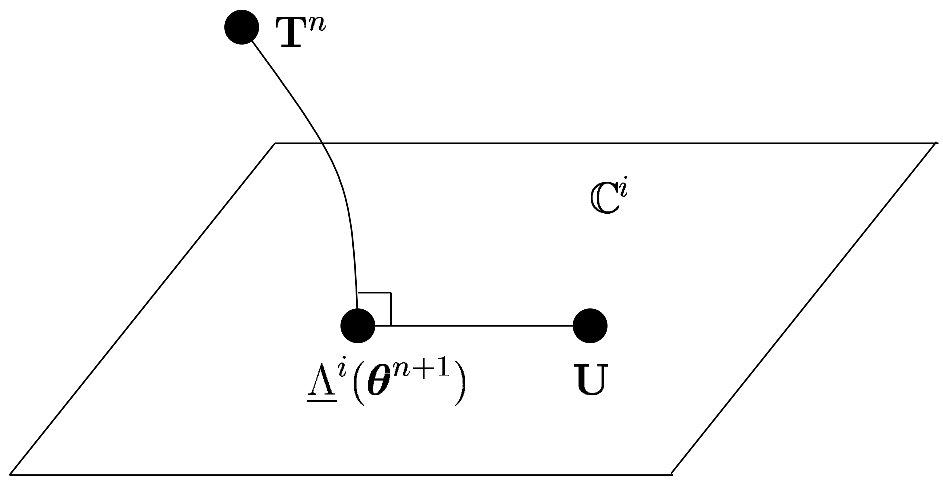
- Minimizing the KL divergence $D(\Theta^i(\Pi^i(\boldsymbol{T}))||\Pi^i(\theta))$ subject to $\boldsymbol{T} \in \text{EXPTREE}^i(G)$ yields correct marginals for the tree.

- Minimizing $\mathcal{G}^i(\Pi^i(\boldsymbol{T}); \Pi^i(\theta))$ can be seen to be equivalent, over the tree.

- For the full graph G with cycles, the KL divergence and $\mathcal{G}$ are **not** equivalent.

- Like the Bethe approximation, $\mathcal{G}$ approximates KL divergence.

Variational Formulation



- Tree reparameterization is a projection with respect to the functional $\mathcal{G}$:

$$\mathcal{G}(\boldsymbol{U};\theta^{n+1}) = \mathcal{G}(\boldsymbol{U};\theta^{n+1}) + \lambda^n \mathcal{G}(\underline{\Lambda}^i(\theta^{n+1});\theta^n).$$

Fixed Point Characterization

- Fixed points always exist and coincide with fixed points of BP.

- A fixed point θ^* corresponds to a pseudo–marginal vector $\boldsymbol{T}^* \in \mathrm{TREE}(G)$.

- That vector $\boldsymbol{T}^*$ satisfies stationary conditions w.r.t $\mathcal{G}$:

$$\sum_{\alpha \in \mathcal{A}} \frac{\partial \mathcal{G}}{\partial T_\alpha}(\boldsymbol{T}^*; \theta^0)[\boldsymbol{U} - \boldsymbol{T}^*]_\alpha = 0.$$

A Blackboard Detour: Extensions to Hypergraphs

Bounds on Partition Function

- $\boldsymbol{\theta} = \{\theta(\mathcal{T}) | \mathcal{T} \in \mathfrak{T}\}$

- $\mathcal{A}(\theta^*) := \{(\boldsymbol{\theta}, \mu) \,|\, \mathbb{E}_\mu(\theta(\mathcal{T})) = \theta^*\}$.

- By convexity of $\Phi(\theta)$, and Jensen's inequality:

$$\Phi(\theta^*) \leq \mathbb{E}_\mu[\Phi(\theta(\mathcal{T}))] = \sum_{\mathcal{T} \in \mathfrak{T}} \mu(\mathcal{T}) \Phi(\theta(\mathcal{T})).$$

Optimizing Over θ Parameter

- Fixing μ, we have a convex optimization problem:

$$\begin{cases} \min_{\boldsymbol{\theta}} \mathbb{E}_\mu[\Phi(\theta(\mathcal{T}))] \\ \text{s.t.} \quad \mathbb{E}_\mu[\boldsymbol{\theta}] = \theta^*. \end{cases}$$

- Form the dual, via the Lagrangian:

$$\begin{cases} \max_\lambda \mathcal{Q}(\lambda; \mu; \theta^*) = -\mathbb{E}_\mu[\Psi(\Pi^{\mathcal{T}}(\lambda))] + \sum_\alpha \lambda_\alpha \theta^*_\alpha \\ \text{s.t.} \quad \lambda \in \mathbb{L}(G), \end{cases}$$

where

$$\mathbb{L}(G) = \{\lambda \,|\, 0 \leq \lambda_{st} \leq \lambda_s \leq 1;\ \lambda_s + \lambda_t \leq 1 + \lambda_{st}\}.$$

Optimizing Over μ Parameter

- We can decompose entropy over a tree distribution:

$$\Psi(\Pi^{\mathcal{T}}(\lambda)) = -\sum_{s \in V} H_s(\lambda) + \sum_{(s,t) \in E(\mathcal{T})} I_{st}(\lambda).$$

- Therefore, the expectation becomes

$$\begin{aligned} \mathbb{E}_{\mu}[\Psi(\Pi^{\mathcal{T}}(\lambda))] &= \sum_{\mathcal{T} \in \mathfrak{T}} \mu(\mathcal{T}) \left\{ -\sum_{s \in V} H_s(\lambda) + \sum_{(s,t) \in E(\mathcal{T})} I_{st}(\lambda) \right\} \\ &= -\sum_{s \in V} H_s(\lambda) + \sum_{(s,t) \in E} \mu_{st} I_{st}(\lambda). \end{aligned}$$

Optimal Bounds for $\Phi(\theta^*)$

- $\text{MARG}(G) := \{\boldsymbol{\mu_e} \,|\, \mu_e = \mathbb{E}_\mu[\delta(e \in \mathcal{T})] \;\text{ for some } \mu;\; \forall e \in E\}$.

- $\mathcal{F}(\lambda; \boldsymbol{\mu_e}; \theta^*) := -\sum_{s \in V} H_s(\lambda) + \sum_{(s,t)\in E} \mu_{st} I_{st}(\lambda) - \sum_\alpha \lambda_\alpha \theta^*_\alpha$.

- $\mathcal{H}(\boldsymbol{\mu_e}; \theta^*) := \min_{\lambda \in \mathbb{L}(G)} \{\mathcal{F}(\lambda; \boldsymbol{\mu_e}; \theta^*)\}$.

- $\Phi(\theta^*) \leq -\max_{\boldsymbol{\mu_e} \in \text{MARG}(G)} \mathcal{H}(\boldsymbol{\mu_e}; \theta^*)$.

Conclusion

- Tree Reparameterization is equivalent to BP.
- Convexity of log partition function: links mean parameters and exponential parameters via duality.
- Geometrical interpretation: Reparameterization confines us to an e-flat manifold, and we "project" onto m-flat manifolds.
- Fixed points always exist.
- Extensions to Hypergraphs.

- Upper bounds: solving a convex optimization gives upper bounds to $\Phi(\theta)$.

BP and Bethe Free Energy

- Boltzmann: $p(\boldsymbol{x}) = \frac{1}{Z}\exp(-E(\boldsymbol{x})/T)$.

- Minimizing the Gibbs free energy functional

$$G(b(\boldsymbol{x})) = \sum_{\boldsymbol{x}} b(\boldsymbol{x})E(\boldsymbol{x}) + \sum_{\boldsymbol{x}} b(\boldsymbol{x}) \log b(\boldsymbol{x}),$$

  yields the correct distribution. (Note relation to KL).

- The Bethe free energy approximates the Gibbs free energy with a tree approximation:

$$\begin{aligned} U_{Bethe} &= -\sum_{ij} b_{ij} \log \psi_{ij} - \sum_{i} b_i \psi_i \\ H_{Bethe} &= \sum_{i} H(b_i) + \sum_{ij} I(b_i, b_j). \end{aligned}$$