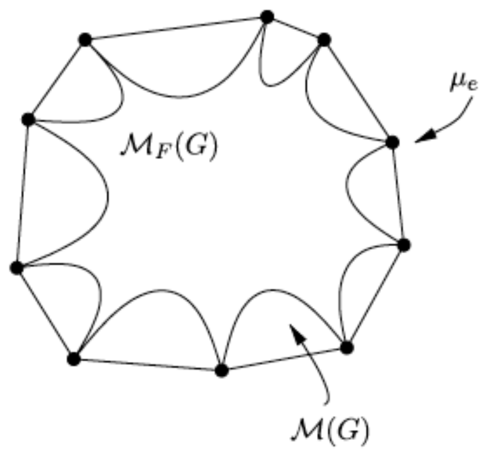
Convex Relaxation and Upper Bounds

REZA BABANEZHAD
UBC
SEP 02 2015

Motivation

- ➢Mean Field methods provide mean approximation and lower bound for the partition function
- ➢Bethe type methods just provide approximation
- ➢Both are non-convex
  - ➢In Mean filed: the approximation to the mean set is non convex
  - ➢For Bethe type: the objective function is non convex
- ➢Consequences: multiple optima, sensitivity to the problem parameters, convergence issue, and dependence on initialization.



Motivation

- But the underlying exact variational principle is convex

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

- So the goal is :
  - Approximating the set $\mathcal{M}$ with a convex set
  - Replacing the dual function A^* with a convex function

Generic Convex Combinations and Surrogates

- Computing mean parameters is tractable for some sub-graph F of G.
  - E.g. Spanning tree and Planar graph
- $$\mathcal{M}(F) := \{\mu \in \mathbb{R}^{|\mathcal{I}(F)|} \mid \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi(X)] \;\forall\, \alpha \in \mathcal{I}(F)\}.$$
- $\mu \mapsto \mu(F)$: represents the coordinate projection mapping from the full space $\mathcal{I}$ to the subset $\mathcal{I}(F)$ of indices associated with F.
- Sub-graph F extracts a subset of indices $\mathcal{I}(F)$ from the full index set $\mathcal{I}$ of potential functions.

Generic Convex Combinations and Surrogates

- We have these bounds on the dual function and entropy

$$A^*(\mu(F)) \leq A^*(\mu),$$

$$H(\mu(F)) \geq H(\mu).$$

- Proof.

$$A^*(\mu) = \sup_{\theta \in \mathbb{R}^d} \{\langle \mu, \theta \rangle - A(\theta)\}.$$

$$A^*(\mu(F)) = \sup_{\theta(F) \in \mathbb{R}^{d(F)}} \{\langle \mu(F), \theta(F) \rangle - A(\theta(F))\}.$$

$$A^*(\mu(F)) = \sup_{\substack{\theta \in \mathbb{R}^d, \\ \theta_\alpha = 0 \ \forall \alpha \notin \mathcal{I}(F)}} \{\langle \mu, \theta \rangle - A(\theta)\},$$

Generic Convex Combinations and Surrogates

- For convex combination of F's

$$H(\mu) \leq \mathbb{E}_\rho[H(\mu(F))] := \sum_{F \in \mathfrak{D}} \rho(F) H(\mu(F)).$$

- We found upper bound for entropy, now finding outer bound for $\mathcal{M}$
  - Main constrain:

$$H(\mu(F)) = -A^*(\mu(F))$$

  - Convex (each $\mathcal{M}(F)$ is convex) outrebound on $\mathcal{M}$

$$\mathcal{L}(G; \mathfrak{D}) := \{\tau \in \mathbb{R}^d \mid \tau(F) \in \mathcal{M}(F) \quad \forall F \in \mathfrak{D}\}.$$

Generic Convex Combinations and Surrogates

- Final approximate variational principle

$$B_{\mathfrak{D}}(\theta;\rho) := \sup_{\tau \in \mathcal{L}(G;\mathfrak{D})} \left\{ \langle \tau, \theta \rangle + \sum_{F \in \mathfrak{D}} \rho(F) H(\tau(F)) \right\}.$$

- Note the objective function is concave.
- The constraint set $\cap_F \mathcal{M}_F$ is convex,
- $B_{\mathfrak{D}}(\theta;\rho)$ is convex surrogate for A

Tree-reweighted Sum-Product and Bethe

- For a given G=(V,E) consider pairwise MRF

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

- Let the tractable class $\mathfrak{D}$ be the set $\mathfrak{T}$ of all spanning trees $T = (V, E(T))$
- A spanning tree of a graph is a tree-structured sub-graph whose vertex set covers the original graph.
- ρ : prob. dist. Over T $H(\mu) \leq \sum_T \rho(T) H(\mu(T))$
- For tree-structured entropies: they decompose additively in terms of entropies associated with the vertices and edges of the tree

$$H(\mu) \leq \sum_{s\in V} H_s(\mu_s) - \sum_{(s,t)\in E} \rho_{st} I_{st}(\mu_{st}).$$

$$\rho_{st} = \mathbb{E}_\rho\left[\mathbb{I}\left[(s,t) \in E(T)\right]\right]$$ Edge appearance prob.

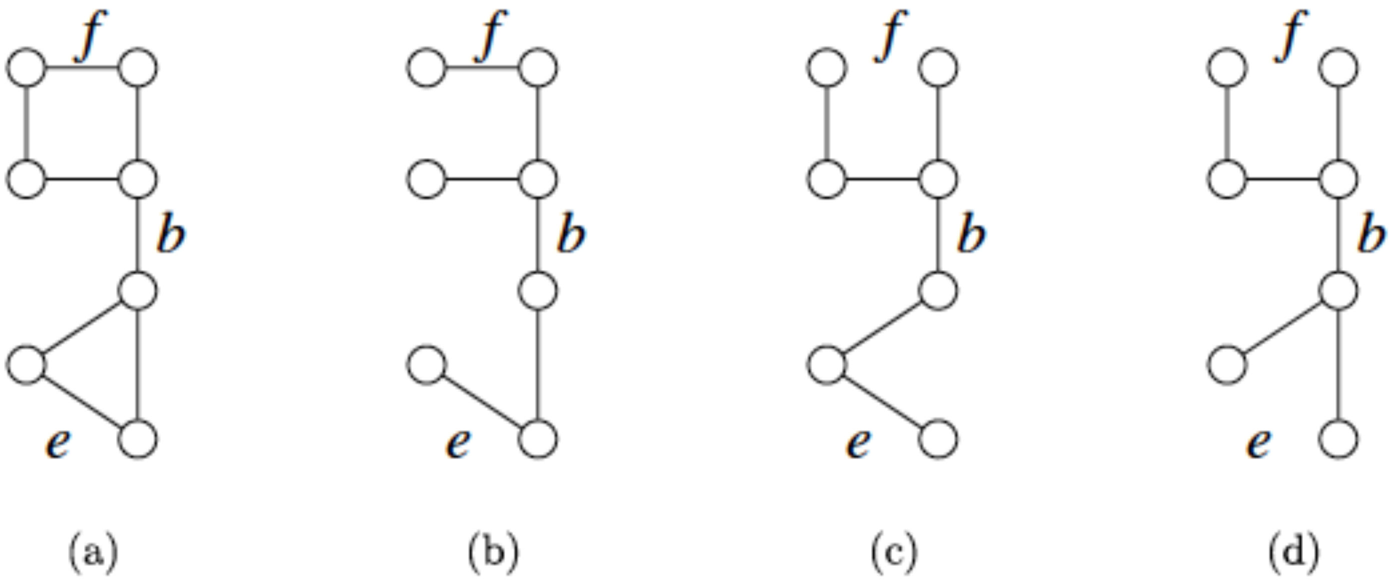


Fig. 7.1 Illustration of valid edge appearance probabilities. Original graph is shown in panel (a). Probability 1/3 is assigned to each of the three spanning trees $\{T_i \mid i = 1,2,3\}$ shown in panels (b)–(d). Edge b appears in all three trees so that $\rho_b = 1$. Edges e and f appear in two and one of the spanning trees, respectively, which gives rise to edge appearance probabilities $\rho_e = 2/3$ and $\rho_f = 1/3$.

Theorem 7.2 (Tree-Reweighted Bethe and Sum-Product).

(a) For any choice of edge appearance vector $(\rho_{st},\ (s,t) \in E)$ in the spanning tree polytope, the cumulant function $A(\theta)$ evaluated at θ is upper bounded by the solution of the tree-reweighted Bethe variational problem (BVP):

$$B_{\mathfrak{T}}(\theta; \rho_e) := \max_{\tau \in \mathbb{L}(G)} \Big\{ \langle \tau, \theta \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}) \Big\}. \tag{7.11}$$

For any edge appearance vector such that $\rho_{st} > 0$ for all edges (s,t), this problem is strictly convex with a unique optimum.

(b) The tree-reweighted BVP can be solved using the tree-reweighted sum-product updates

$$M_{ts}(x_s) \leftarrow \kappa \sum_{x'_t \in \mathcal{X}_t} \varphi_{st}(x_s, x'_t) \frac{\prod_{v \in N(t) \backslash s} \big[M_{vt}(x'_t)\big]^{\rho_{vt}}}{\big[M_{st}(x'_t)\big]^{(1-\rho_{ts})}}, \tag{7.12}$$

where $\varphi_{st}(x_s, x'_t) := \exp\Big(\frac{1}{\rho_{st}} \theta_{st}(x_s, x'_t) + \theta_t(x'_t)\Big)$. The updates (7.12) have a unique fixed point under the assumptions of part (a).

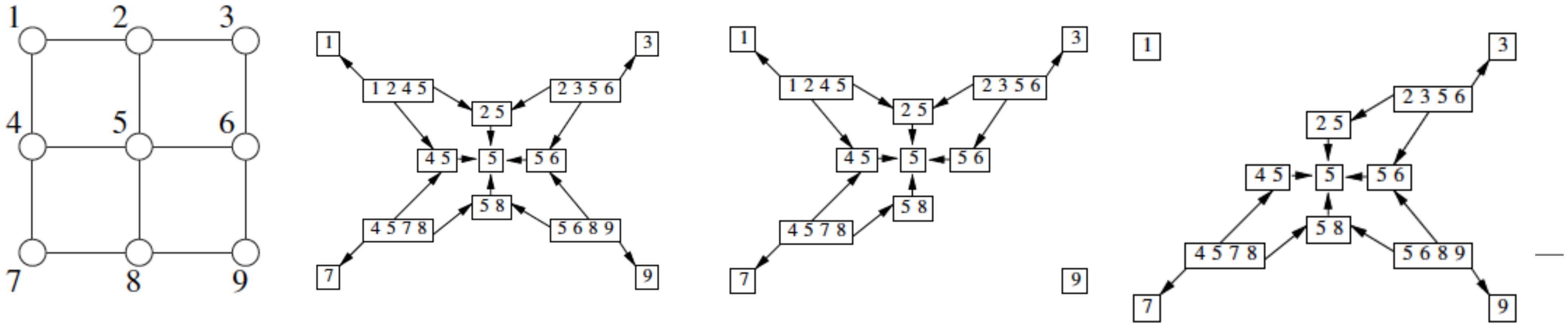
Reweighted Kikuchi Approximations

- The transition from the Bethe to Kikuchi variational problems is to take convex combinations of hypertrees
- For a given treewidth t, consider the set of all hypertrees of width $\mathfrak{T}$(t) of width less than or equal to t

$$\rho = (\rho(T),\, T \in \mathfrak{T}(t))$$

$$H(\mu) \leq \mathbb{E}_\rho[H(\mu(T))] = -\sum_T \rho(T) H(\mu(T)).$$

$$A(\theta) \leq B_{\mathfrak{T}(t)}(\theta;\rho) := \max_{\tau \in \mathbb{L}(G)} \{\langle \tau, \theta \rangle + \mathbb{E}_\rho[H(\tau(T))]\}.$$



$$\varphi_{1245} = \frac{\tau_{1245}}{\varphi_{25}\ \varphi_{45}\ \varphi_5 \varphi_1} = \frac{\tau_{1245}}{\frac{\tau_{25}}{\tau_5}\frac{\tau_{45}}{\tau_5}\tau_5\tau_1} = \frac{\tau_{1245}\ \tau_5}{\tau_{25}\tau_{45}\tau_1}.$$

$$p_{\tau(T^1)}(x) = \left[\frac{\tau_{1245}\ \tau_5}{\tau_{25}\tau_{45}\tau_1}\right]\left[\frac{\tau_{2356}\ \tau_5}{\tau_{25}\tau_{56}\tau_3}\right]\left[\frac{\tau_{4578}\ \tau_5}{\tau_{45}\tau_{58}\tau_7}\right] \times \left[\frac{\tau_{25}}{\tau_5}\right]\left[\frac{\tau_{45}}{\tau_5}\right]\left[\frac{\tau_{56}}{\tau_5}\right]\left[\frac{\tau_{58}}{\tau_5}\right][\tau_1][\tau_3][\tau_5][\tau_7][\tau_9].$$

$$\begin{aligned}\textstyle\sum_{i=1}^4 \frac{1}{4} A^*(\tau(T^i)) &= \frac{3}{4}\sum_{h\in E_4}\sum_{x_h}\tau_h(x_h)\log\varphi_h(x_h) + \sum_{s\in\{2,4,6,8\}}\sum_{x_{s5}}\tau_{s5}(x_{s5})\log\frac{\tau_{s5}(x_{s5})}{\tau_5(x_5)} \\ &\quad + \sum_{s\in\{1,3,5,7,9\}}\sum_{x_s}\tau_s(x_s)\log\tau_s(x_s). \\ &= \frac{3}{4}[H_{1245} + H_{2356} + H_{5689} + H_{4578}] - \frac{1}{2}[H_{25} + H_{45} + H_{56} + H_{58}] \\ &\quad + \frac{1}{4}[H_1 + H_3 + H_7 + H_9]. \qquad (7.18)\end{aligned}$$

$$E_4 = \{(1245), (2356), (5689), (4578)\}$$

Algorithm Stability

- Let $\tau(\theta)$ denote the output when some variational method is applied to the model p_θ
- Globally Lipschitz stable condition: $\|\tau(\theta) - \tau(\theta')\|_a \leq L\|\theta - \theta'\|_b,$
- Algorithmic Stability: When $\|\theta - \theta'\|_b$ is small, $\|\tau(\theta) - \tau(\theta')\|_a$ is small too.

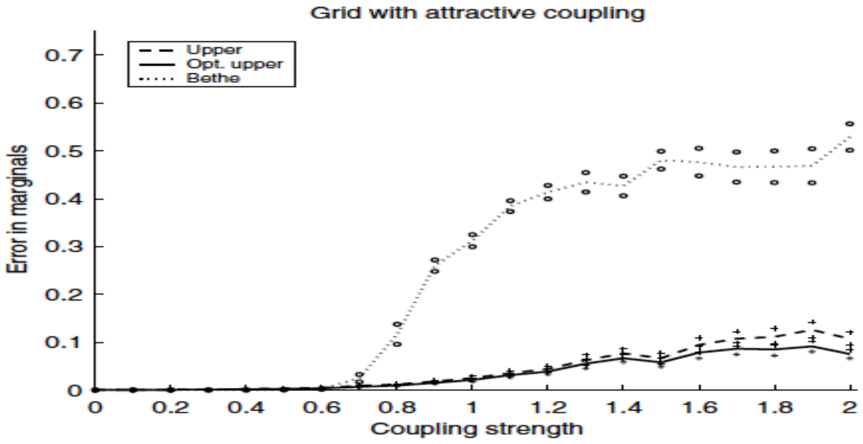


Fig. 7.3 Contrast between the instability of ordinary sum-product and stability of tree-reweighted sum-product [246]. Plots show the error between the true marginals and correct marginals versus the coupling strength in a binary pairwise Markov random field. Note that the ordinary sum-product algorithm is very accurate up to a critical coupling (≈ 0.70), after which it degrades rapidly. On the other hand, the performance of TRW message-passing varies smoothly as a function of the coupling strength. The plot shows two versions of the TRW sum-product algorithm, either based on uniform edge weights $\rho_{st} = 1/2$, or edge weights optimized to minimize the upper bound.

Algorithm Stability

- Consider general variational method $$B(\theta) = \sup_{\tau \in \mathcal{L}} \{\langle \theta, \tau \rangle - B^*(\tau)\},$$

- B is convex surrogate for $\mathfrak{A}$ and $\mathcal{L}$ is a convex outrebound for $\mathcal{M}$.

- When B is strictly convex and $B*$ is strongly convex with parameter $1/\boldsymbol{L}$ then the output

$$\tau(\theta) = \nabla B(\theta)$$

is Lipschitz stable with parameter $\boldsymbol{L}$.

Convex Surrogate in Parameter Estimation

- Maximum likelihood for exponential family

$$\ell(\theta; X_1^n) = \ell(\theta) = \langle \theta, \widehat{\mu} \rangle - A(\theta)$$

$$\widehat{\mu} := \frac{1}{n}\sum_{i=1}^n \phi(X^i)$$

-

$$\nabla \ell(\theta) = \widehat{\mu} - \mu(\theta)$$

- Surrogate Likelihood:

$$\ell_B(\theta; X_1^n) = \ell_B(\theta) := \langle \theta, \widehat{\mu} \rangle - B(\theta)$$

- Surrogate likelihood gives a lower bound on likelihood.

$$\widetilde{\theta}_B := \arg\max_{\theta \in \Omega} \ell_B(\theta; X_1^n)$$

Convex Surrogate in Parameter Estimation

- Optimizing surrogate likelihood

$$\nabla \ell_B(\theta) = \widehat{\mu} - \tau(\theta)$$
$$\tau(\theta) = \nabla B(\theta)$$

- Since the objective is concave, a standard coordinate ascent can be used to compute $\widetilde{\theta}_B$
- But for some ML surrogate, there is closed form.
- For Tree-reweighted Bethe surrogate:

$$\begin{aligned} \forall\, s \in V,\ j \in \mathcal{X}_s, \qquad \widetilde{\theta}_{s;j} &= \log \widehat{\mu}_{s;j}, \quad \text{and} \\ \forall\, (s,t) \in E,\ (j,k) \in \mathcal{X}_s \times \mathcal{X}_t, \qquad \widetilde{\theta}_{st;jk} &= \rho_{st} \log \frac{\widehat{\mu}_{st;jk}}{\widehat{\mu}_{s;j}\widehat{\mu}_{t;k}}. \end{aligned}$$

$$\widehat{\mu}_{s;j} = \widehat{\mathbb{E}}[\mathbb{I}_j(X_s)]$$

Convex Surrogate in Parameter Estimation

- Penalized surrogate likelihood $\widetilde{\ell}(\theta; \lambda) := \ell(\theta) - \lambda R(\theta)$
- R: convex but not necessarily differentiable
- Regularized surrogate likelihood(RSL) $\widetilde{\ell}_B(\theta; \lambda) := \ell_B(\theta) - \lambda R(\theta)$
- Could be solved by using standard methods

Convex Surrogate in Parameter Estimation

- Alternative formulation for RSL

$$\begin{aligned}\inf_{\theta\in\Omega}&\{-\langle\theta,\widehat{\mu}\rangle + B(\theta) + \lambda R(\theta)\}\\ &= \inf_{\theta\in\Omega}\Big\{-\langle\theta,\widehat{\mu}\rangle + \sup_{\tau\in\mathcal{L}}\{\langle\theta,\tau\rangle - B^*(\tau)\} + \lambda R(\theta)\Big\}\\ &= \inf_{\theta\in\Omega}\sup_{\tau\in\mathcal{L}}\{\langle\theta,\tau-\widehat{\mu}\rangle - B^*(\tau) + \lambda R(\theta)\}.\end{aligned}$$

- Under some regularity conditions

$$\begin{aligned}\inf_{\theta\in\Omega}\{-\ell_B(\theta)+\lambda R(\theta)\} &= \sup_{\tau\in\mathcal{L}}\inf_{\theta\in\Omega}\{\langle\theta,\tau-\widehat{\mu}\rangle - B^*(\tau) + \lambda R(\theta)\}\\ &= \sup_{\tau\in\mathcal{L}}\Big\{-B^*(\tau) - \lambda\sup_{\theta\in\Omega}\Big\{\langle\theta,\frac{\tau-\widehat{\mu}}{\lambda}\rangle - R(\theta)\Big\}\Big\}\\ &= \sup_{\tau\in\mathcal{L}}\Big\{-B^*(\tau) - \lambda\, R^*_\Omega\big(\frac{\tau-\widehat{\mu}}{\lambda}\big)\Big\}, \qquad (7.29)\end{aligned}$$

- R^*_Ω is conjugate dual of $R(\theta) + \mathbb{I}_\Omega(\theta)$

Conclusion

- How we can convexify the variational approximations in general
- Two examples: Bethe and Kukuichi methods
- Stability of the methods
- Parameter estimation by using convex surrogate

Thank you!