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

$$
\mathcal{M}(F) := \{ \mu \in \mathbb{R}^{I(F)} \mid \exists p \text{ s.t. } \mu_\alpha = \mathbb{E}_p[\phi(X)] \ \forall \alpha \in I(F) \}.
$$

- $\mu \mapsto \mu(F)$: represents the coordinate projection mapping from the full space $I$ to the subset $I(F)$ of indices associated with $F$.
- Sub-graph $F$ extracts a subset of indices $I(F)$ from the full index set $I$ of potential functions.
We have these bounds on the dual function and entropy

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

\[ H(\mu(F)) = \sup_{\theta \in \mathbb{R}^d} \{ \langle \mu(F), \theta(F) \rangle - A(\theta(F)) \} \geq H(\mu). \]

**Proof.**

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

\[ A^*(\mu(F)) = \sup_{\substack{\theta \in \mathbb{R}^d, \\ \theta_x = 0 \forall x \notin \mathcal{I}(F)}} \{ \langle \mu, \theta \rangle - A(\theta) \}, \]
Generic Convex Combinations and Surrogates

➢ For convex combination of $F$

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

➢ 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 $\mathcal{M}$

\[ \mathcal{L}(G; \mathcal{D}) := \{ \tau \in \mathbb{R}^d \mid \tau(F) \in \mathcal{M}(F) \quad \forall F \in \mathcal{D} \}. \]
Generic Convex Combinations and Surrogates

- Final approximate variational principle

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

- Note the objective function is concave.
- The constraint set \( \cap_F \mathcal{M}_F \) is convex.
- \( B_\mathcal{D}(\theta; \rho) \) is convex surrogate for \( A \)
Tree-reweighted Sum-Product and Bethe

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

$$p_0(x) \propto \exp \left\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right\},$$

Let the tractable class $\mathcal{D}$ be the set $\mathcal{T}$ 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.

$\rho :$ 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

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

$$\rho_{st} = \mathbb{E}_{\rho} \left[ \mathbb{I} \left[ (s,t) \in E(T) \right] \right] \quad \text{Edge appearance prob.}$$
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 \(\theta\) is upper bounded by the solution of the tree-reweighted Bethe variational problem (BVP):

\[
B_{\overline{\lambda}}(\theta; \rho_e) := \max_{\tau \in \mathbb{L}(G)} \left\{ \langle \tau, \theta \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}) \right\}.
\]

(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

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

where \(\varphi_{st}(x_s, x'_t) := \exp \left( \frac{1}{\rho_{st}} \theta_{st}(x_s, x'_t) + \theta_t(x'_t) \right)\). The updates (7.12) have a unique fixed point under the assumptions of part (a).
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 $I(t)$ of width less than or equal to $t$

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

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

$$A(\theta) \leq B_{\mathcal{T}(t)}(\theta; \rho) := \max_{\tau \in \mathbb{L}(G)} \{ \langle \tau, \theta \rangle + \mathbb{E}_\rho[H(\tau(T))] \}.$$
\[
\varphi_{1245} = \frac{\tau_{1245}}{\varphi_{25} \varphi_{45} \varphi_{5} \varphi_{1}} = \frac{\tau_{1245}}{\tau_{25} \tau_{45} \tau_{5} \tau_{1}} = \frac{\tau_{1245} \tau_{5}}{\tau_{25} \tau_{45} \tau_{7}}.
\]

\[
p_{\tau(T1)}(x) = \left[ \frac{\tau_{1245} \tau_{5}}{\tau_{25} \tau_{45} \tau_{7}} \right] \left[ \frac{\tau_{2356} \tau_{5}}{\tau_{25} \tau_{45} \tau_{7}} \right] \left[ \frac{\tau_{4578} \tau_{5}}{\tau_{25} \tau_{45} \tau_{7}} \right] \times \left[ \frac{\tau_{26}}{\tau_{5}} \right] \left[ \frac{\tau_{45}}{\tau_{5}} \right] \left[ \frac{\tau_{56}}{\tau_{5}} \right] \left[ \frac{\tau_{58}}{\tau_{5}} \right] \left[ \frac{\tau_{7}}{\tau_{1}} \right] \left[ \frac{\tau_{3}}{\tau_{7}} \right] \left[ \frac{\tau_{9}}{\tau_{7}} \right].
\]

\[
\sum_{i=1}^{4} \frac{1}{4} A^*(\tau(T^i)) = \sum_{h \in E_4} \sum_{x_h} \tau_h(x_h) \log \varphi_h(x_h) + \sum_{s \in \{2, 4, 6, 8\}} \sum_{x_{s5}} \tau_{s5}(x_{s5}) \log \frac{\tau_{s5}(x_{s5})}{\tau_{5}(x_{s5})} + \sum_{s \in \{1, 3, 5, 7, 9\}} \sum_{x_s} \tau_s(x_s) \log \tau_s(x_s).
\]

\[
= \frac{3}{4} [H_{1245} + H_{2356} + H_{5689} + H_{4578}] - \frac{1}{2} [H_{25} + H_{45} + H_{56} + H_{58}] + \frac{1}{4} [H_1 + H_3 + H_7 + H_9].
\]

\[E_4 = \{(1245), (2356), (5689), (4578)\}\]
Algorithm Stability

- Let $\tau(\theta)$ denote the output when some variational method is applied to the model $p_\theta$
- 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.
Algorithm Stability

- Consider general variational method

\[ B(\theta) = \sup_{\tau \in \mathcal{L}} \{ \langle \theta, \tau \rangle - B^*(\tau) \}, \]

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

- When \( B \) is strictly convex and \( B^* \) is strongly convex with parameter \( 1/\mathcal{L} \) then the output

\[ \tau(\theta) = \nabla B(\theta) \]

is Lipschitz stable with parameter \( \mathcal{L} \).
Convex Surrogate in Parameter Estimation

- Maximum likelihood for exponential family

\[ \ell(\theta; X_1^n) = \ell(\theta) = \langle \theta, \hat{\mu} \rangle - A(\theta) \]
\[ \hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} \phi(X^i) \]
\[ \nabla \ell(\theta) = \hat{\mu} - \mu(\theta) \]

- Surrogate Likelihood:

\[ \ell_B(\theta; X_1^n) = \ell_B(\theta) := \langle \theta, \hat{\mu} \rangle - B(\theta) \]

- Surrogate likelihood gives a lower bound on likelihood.

\[ \tilde{\theta}_B := \arg \max_{\theta \in \Omega} \ell_B(\theta; X_1^n) \]
Convex Surrogate in Parameter Estimation

- Optimizing surrogate likelihood

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

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

- Since the objective is concave, a standard coordinate ascent can be used to compute $\tilde{\theta}_B$

- But for some ML surrogate, there is closed form.

- For Tree-reweighted Bethe surrogate:

$$\forall s \in V, j \in \mathcal{X}_s, \quad \tilde{\theta}_{s;j} = \log \hat{\mu}_{s;j}, \quad \text{and} \quad \hat{\mu}_{s;j} = \mathbb{E} [ f_j(X_s)]$$

$$\forall (s,t) \in E, (j,k) \in \mathcal{X}_s \times \mathcal{X}_t, \quad \tilde{\theta}_{st;jk} = \rho_{st} \log \frac{\hat{\mu}_{st;jk} \hat{\mu}_{s;j} \hat{\mu}_{t;k}}{\hat{\mu}_{s;j} \hat{\mu}_{t;k}}.$$
Convex Surrogate in Parameter Estimation

- Penalized surrogate likelihood
  \( \tilde{\ell}(\theta; \lambda) := \ell(\theta) - \lambda R(\theta) \)

- \( R \): convex but not necessarily differentiable

- Regularized surrogate likelihood (RSL)
  \( \tilde{\ell}_B(\theta; \lambda) := \ell_B(\theta) - \lambda R(\theta) \)

- Could be solved by using standard methods
Convex Surrogate in Parameter Estimation

- Alternative formulation for RSL

\[
\inf_{\theta \in \Omega} \left\{ -\langle \theta, \hat{\mu} \rangle + B(\theta) + \lambda R(\theta) \right\} = \inf_{\theta \in \Omega} \left\{ -\langle \theta, \hat{\mu} \rangle + \sup_{\tau \in \mathcal{L}} \{ \langle \theta, \tau \rangle - B^*(\tau) \} + \lambda R(\theta) \right\} = \inf_{\theta \in \Omega} \sup_{\tau \in \mathcal{L}} \{ \langle \theta, \tau - \hat{\mu} \rangle - B^*(\tau) + \lambda R(\theta) \}.
\]

- Under some regularity conditions

\[
\inf_{\theta \in \Omega} \{ -\ell_B(\theta) + \lambda R(\theta) \} = \sup_{\tau \in \mathcal{L}} \inf_{\theta \in \Omega} \{ \langle \theta, \tau - \hat{\mu} \rangle - B^*(\tau) + \lambda R(\theta) \} = \sup_{\tau \in \mathcal{L}} \left\{ -B^*(\tau) - \lambda \sup_{\theta \in \Omega} \left\{ \left\langle \theta, \frac{\tau - \hat{\mu}}{\lambda} \right\rangle - R(\theta) \right\} \right\} = \sup_{\tau \in \mathcal{L}} \left\{ -B^*(\tau) - \lambda R^*_\Omega \left( \frac{\tau - \hat{\mu}}{\lambda} \right) \right\},
\]

(7.29)

- \( R^*_\Omega \) is conjugate dual of \( R(\theta) + \Pi_\Omega(\theta) \)
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!