School of Computer Science
Carnegie Mellon

Probabilistic Graphical Models

Theory of Variational Inference:
Inner and Outer Approximation

Junming Yin
Lecture 15, March 4, 2013

Reading: W & J Book Chapters

Fig. 3.5 Generic illustration of $M$ for a discrete random variable with $|X|_m$ finite. In this case, the set $M$ is a convex polytope, corresponding to the convex hull of $\{\phi(x) | x \in X_m\}$. By the Minkowski–Weyl theorem, this polytope can also be written as the intersection of a finite number of half-spaces, each of the form $\{\mu \in \mathbb{R}^d | \langle a_j, \mu \rangle \geq b_j\}$ for some pair $(a_j, b_j) \in \mathbb{R}^d \times \mathbb{R}$.

Example 3.8 (Ising Mean Parameters). Continuing from Example 3.1, the sufficient statistics for the Ising model are the singleton functions $(x_s, s \in V)$ and the pairwise functions $(x_s x_t, (s, t) \in E)$. The vector of sufficient statistics takes the form:

$$\phi(x) := (x_s, s \in V; x_s x_t, (s, t) \in E) \in \mathbb{R}^{|V|+|E|}. \quad (3.30)$$

The associated mean parameters correspond to particular marginal probabilities, associated with nodes and edges of the graph $G$ as $\mu_s = \mathbb{E}_{\mathbb{P}}[X_s] = \mathbb{P}[X_s = 1]$ for all $s \in V$, and (3.31a)

$$\mu_{st} = \mathbb{E}_{\mathbb{P}}[X_s X_t] = \mathbb{P}[(X_s, X_t) = (1, 1)] \quad \text{for all } (s, t) \in E. \quad (3.31b)$$

Consequently, the mean parameter vector $\mu \in \mathbb{R}^{|V|+|E|}$ consists of marginal probabilities over singletons ($\mu_s$), and pairwise marginals over variable pairs on graph edges ($\mu_{st}$). The set $M$ consists of the convex hull of $\{\phi(x) | x \in \{0, 1\}_m\}$, where $\phi$ is given in Equation (3.30).

In probabilistic terms, the set $M$ corresponds to the set of all singleton and pairwise marginal probabilities that can be realized by some distribution over $(X_1, \ldots, X_m) \in \{0, 1\}_m$. In the polyhedral combinatorics literature, this set is known as the correlation polytope, or the cut polytope [69, 187].

Roadmap

- Two families of approximate inference algorithms
  - Loopy belief propagation (sum-product)
  - Mean-field approximation

- Are there some connections of these two approaches?

- We will re-exam them from a unified point of view based on the variational principle:
  - Loop BP: outer approximation
  - Mean-field: inner approximation
Variational Methods

- "Variational": fancy name for optimization-based formulations
  - i.e., represent the quantity of interest as the solution to an optimization problem
  - approximate the desired solution by \textit{relaxing/approximating} the \textit{intractable} optimization problem

- Examples:
  - Courant-Fischer for eigenvalues: \( \lambda_{\max}(A) = \max_{\|x\|_2=1} x^T A x \)
  - Linear system of equations: \( Ax = b, A \succ 0, x^* = A^{-1} b \)
    - variational formulation:
      \[
      x^* = \arg\min_x \left\{ \frac{1}{2} x^T A x - b^T x \right\}
      \]
    - for large system, apply conjugate gradient method

Inference Problems in Graphical Models

- Undirected graphical model (MRF):
  \[
p(x) = \frac{1}{Z} \prod_{C \in C} \psi_C(x_C)
  \]

- The quantities of interest:
  - marginal distributions: \( p(x_i) = \sum_{x \neq i \neq j} p(x) \)
  - normalization constant (partition function): \( Z \)

- Question: how to represent these quantities in a variational form?
  - Use tools from (1) exponential families; (2) convex analysis
Exponential Families

- Canonical parameterization
  
  \[ p_\theta(x_1, \ldots, x_m) = \exp \left\{ \theta^T \phi(x) - A(\theta) \right\} \]

  - Log normalization constant:
    \[ A(\theta) = \log \int \exp \{\theta^T \phi(x)\} dx \]
    - it is a convex function (Prop 3.1)
    - Effective canonical parameters:
      \[ \Omega := \left\{ \theta \in \mathbb{R}^d | A(\theta) < +\infty \right\} \]

Graphical Models as Exponential Families

- Undirected graphical model (MRF):
  \[ p(x; \theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \psi(x_C; \theta_C) \]

- MRF in an exponential form:
  \[ p(x; \theta) = \exp \left\{ \sum_{C \in \mathcal{C}} \log \psi(x_C; \theta_C) - \log Z(\theta) \right\} \]

  - \( \log \psi(x_C; \theta_C) \) can be written in a linear form after some parameterization
Example: Gaussian MRF

- Consider a zero-mean multivariate Gaussian distribution that respects the Markov property of a graph
  - Hammersley-Clifford theorem states that the precision matrix $\Lambda = \Sigma^{-1}$ also respects the graph structure

![Graph Example](image1)

- Gaussian MRF in the exponential form
  $$p(x) = \exp \left\{ \frac{1}{2} \langle \Theta, xx^T \rangle - A(\Theta) \right\}, \text{ where } \Theta = -\Lambda$$

- Sufficient statistics are $\{x_s^2, s \in V; x_s x_t, (s, t) \in E\}$

Example: Discrete MRF

- In exponential form
  $$p(x; \theta) \propto \exp \left\{ \sum_{s \in V} \sum_j \theta_{s;j} I_j(x_s) + \sum_{(s,t) \in E} \theta_{st;jk} I_j(x_s) I_k(x_t) \right\}$$
Why Exponential Families?

- Computing the expectation of sufficient statistics (mean parameters) given the canonical parameters yields the marginals

\[ \mu_{s:j} = \mathbb{E}_p[\mathbb{I}_j(X_s)] = \mathbb{P}[X_s = j] \quad \forall j \in \mathcal{X}_s, \]

\[ \mu_{st:jk} = \mathbb{E}_p[\mathbb{I}_{st:jk}(X_s, X_t)] = \mathbb{P}[X_s = j, X_t = k] \quad \forall (j, k) \in \mathcal{X}_s \times \mathcal{X}_t. \]

- Computing the normalizer yields the log partition function

\[ \log Z(\theta) = A(\theta) \]

Computing Mean Parameter: Bernoulli

- A single Bernoulli random variable

\[ p(x; \theta) = \exp \{\theta x - A(\theta)\}, x \in \{0, 1\}, A(\theta) = \log(1 + e^\theta) \]

- Inference = Computing the mean parameter

\[ \mu(\theta) = \mathbb{E}_\theta[X] = 1 \cdot p(X = 1; \theta) + 0 \cdot p(X = 0; \theta) = \frac{e^\theta}{1 + e^\theta} \]

- Want to do it in a variational manner: cast the procedure of computing mean (summation) in an optimization-based formulation
Conjugate Dual Function

- Given any function \( f(\theta) \), its conjugate dual function is:

\[
f^*(\mu) = \sup_{\theta} \{ \langle \theta, \mu \rangle - f(\theta) \}\]

- Conjugate dual is always a convex function: point-wise supremum of a class of linear functions

Dual of the Dual is the Original

- Under some technical condition on \( f \) (convex and lower semi-continuous), the dual of dual is itself:

\[
f = (f^*)^*
\]

\[
f(\theta) = \sup_{\mu} \{ \langle \theta, \mu \rangle - f^*(\mu) \}\]

- For log partition function

\[
A(\theta) = \sup_{\mu} \{ \langle \theta, \mu \rangle - A^*(\mu) \}, \quad \theta \in \Omega
\]

- The dual variable \( \mu \) has a natural interpretation as the mean parameters
Computing Mean Parameter: Bernoulli

- The conjugate \( A^*(\mu) := \sup_{\theta \in \mathbb{R}} \{\mu \theta - \log[1 + \exp(\theta)]\} \)
- Stationary condition \( \mu = \frac{e^\theta}{1 + e^\theta} \) (\( \mu = \nabla A(\theta) \))
- If \( \mu \in (0, 1) \), \( \theta(\mu) = \log \left( \frac{\mu}{1 - \mu} \right) \), \( A^*(\mu) = \mu \log(\mu) + (1 - \mu) \log(1 - \mu) \)
- If \( \mu \notin [0, 1] \), \( A^*(\mu) = +\infty \)
- We have \( A^*(\mu) = \begin{cases} \mu \log \mu + (1 - \mu) \log(1 - \mu) & \text{if } \mu \in [0, 1] \\ +\infty & \text{otherwise} \end{cases} \).
- The variational form: \( A(\theta) = \max_{\mu \in [0, 1]} \{\mu \cdot \theta - A^*(\mu)\} \).
- The optimum is achieved at \( \mu(\theta) = \frac{e^\theta}{1 + e^\theta} \). This is the mean!

Remark

- The last few identities are not coincidental but rely on a deep theory in general exponential family.
  - The dual function is the negative entropy function
  - The mean parameter is restricted
  - Solving the optimization returns the mean parameter and log partition function
- Next step: develop this framework for general exponential families/graphical models.
- However,
  - Computing the conjugate dual (entropy) is in general intractable
  - The constrain set of mean parameter is hard to characterize
  - Hence we need approximation
Computation of Conjugate Dual

- Given an exponential family
  \[ p(x_1, \ldots, x_m; \theta) = \exp \left\{ \sum_{i=1}^{d} \theta_i \phi_i(x) - A(\theta) \right\} \]

- The dual function
  \[ A^*(\mu) := \sup_{\theta \in \Omega} \{ \langle \mu, \theta \rangle - A(\theta) \} \]

- The stationary condition: \( \mu - \nabla A(\theta) = 0 \)

- Derivatives of \( A \) yields mean parameters
  \[ \frac{\partial A}{\partial \theta_i}(\theta) = \mathbb{E}_\theta[\phi_i(X)] = \int \phi_i(x)p(x; \theta) \, dx \]

- The stationary condition becomes \( \mu = \mathbb{E}_\theta[\phi(X)] \)

- Question: for which \( \mu \in \mathbb{R}^d \) does it have a solution \( \theta(\mu) \) ?

Computation of Conjugate Dual

- Let’s assume there is a solution \( \theta(\mu) \) such that \( \mu = \mathbb{E}_{\theta(\mu)}[\phi(X)] \)

- The dual has the form
  \[ A^*(\mu) = \langle \theta(\mu), \mu \rangle - A(\theta(\mu)) \]
  \[ = \mathbb{E}_{\theta(\mu)}[\langle \theta(\mu), \phi(X) \rangle] - A(\theta(\mu)) \]
  \[ = \mathbb{E}_{\theta(\mu)}[\log p(X; \theta(\mu))] \]

- The entropy is defined as
  \[ H(p(x)) = -\int p(x) \log p(x) \, dx \]

- So the dual is \( A^*(\mu) = -H(p(x; \theta(\mu))) \) when there is a solution \( \theta(\mu) \)
Complexity of Computing Conjugate Dual

- The dual function is implicitly defined:

\[ \mu \xrightarrow{(\nabla A)^{-1}} \theta(\mu) \xrightarrow{-H(p_\theta(\mu))} A^*(\mu) \]

- Solving the inverse mapping \( \mu = \mathbb{E}_\theta[\phi(X)] \) for canonical parameters \( \theta(\mu) \) is nontrivial

- Evaluating the negative entropy requires high-dimensional integration (summation)

- Question: for which \( \mu \in \mathbb{R}^d \) does it have a solution \( \theta(\mu) \)? i.e., the domain of \( A^*(\mu) \).
  - the ones in marginal polytope!

Marginal Polytope

- For any distribution \( p(x) \) and a set of sufficient statistics \( \phi(x) \), define a vector of mean parameters

\[ \mu_i = \mathbb{E}_p[\phi_i(X)] = \int \phi_i(x)p(x)\,dx \]

\( p(x) \) is not necessarily an exponential family

- The set of all realizable mean parameters

\[ \mathcal{M} := \{ \mu \in \mathbb{R}^d \mid \exists \ p \ s.t. \ \mathbb{E}_p[\phi(X)] = \mu \} \]

- It is a convex set

- For discrete exponential families, this is called marginal polytope
Convex Polytope

- Convex hull representation
  \[ \mathcal{M} = \left\{ \mu \in \mathbb{R}^d \mid \sum_{x \in X^m} \phi(x)p(x) = \mu, \text{ for some } p(x) \geq 0, \sum_{x \in X^m} p(x) = 1 \right\} \]
  \[ \triangleq \text{conv} \left\{ \phi(x), x \in X^m \right\} \]

- Half-plane representation
  - Minkowski-Weyl Theorem: any non-empty convex polytope can be characterized by a finite collection of linear inequality constraints
  \[ \mathcal{M} = \left\{ \mu \in \mathbb{R}^d \mid \mu^T \mu \geq b_j, \forall j \in J \right\} \]
  where \(|J|\) is finite.

Example: Two-node Ising Model

- Sufficient statistics:
  \[ \phi(x) := (x_s, s \in V; x_s x_t, (s, t) \in E) \in \mathbb{R}^{|V| + |E|}. \]

- Mean parameters:
  \[ \mu_s = \mathbb{E}_\phi[X_s] = \mathbb{P}[X_s = 1] \text{ for all } s \in V, \text{ and} \]
  \[ \mu_{st} = \mathbb{E}_\phi[X_s X_t] = \mathbb{P}[(X_s, X_t) = (1,1)] \text{ for all } (s,t) \in E. \]

- Two-node Ising model
  - Convex hull representation
    \[ \text{conv}\{0,0,0\},(1,0,0),(0,1,0),(1,1,1) \}\]
  - Half-plane representation
    \[ \mu_1 \geq \mu_{12} \]
    \[ \mu_2 \geq \mu_{12} \]
    \[ \mu_{12} \geq 0 \]
    \[ 1 + \mu_{12} \geq \mu_1 + \mu_2 \]
Marginal Polytope for General Graphs

- Still doable for connected binary graphs with 3 nodes: 16 constraints
- For tree graphical models, the number of half-planes (facet complexity) grows only linearly in the graph size
- General graphs?
  - extremely hard to characterize the marginal polytope

Variational Principle (Theorem 3.4)

- The dual function takes the form
  \[ A^*(\mu) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^o \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases} \]
- \( \theta(\mu) \) satisfies \( \mu = \mathbb{E}_{\theta(u)}[\phi(X)] \)
- The log partition function has the variational form
  \[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \theta^T \mu - A^*(\mu) \} \]
- For all \( \theta \in \Omega \), the above optimization problem is attained uniquely at \( \mu(\theta) \in \mathcal{M}^o \) that satisfies
  \[ \mu(\theta) = \mathbb{E}_{\theta}[\phi(X)] \]
Example: Two-node Ising Model

- The distribution \( p(x; \theta) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2) \)
- Sufficient statistics \( \phi(x) = \{x_1, x_2, x_1 x_2\} \)
- The marginal polytope is characterized by:
  \[
  \begin{align*}
  \mu_1 &\geq \mu_{12} \\
  \mu_2 &\geq \mu_{12} \\
  \mu_{12} &\geq 0 \\
  1 + \mu_{12} &\geq \mu_1 + \mu_2
  \end{align*}
  \]
- The dual has an explicit form:
  \[
  A^*(\mu) = \mu_{12} \log \mu_{12} + (\mu_1 - \mu_{12}) \log(\mu_1 - \mu_{12}) + (\mu_2 - \mu_{12}) \log(\mu_2 - \mu_{12}) \\
  + (1 + \mu_{12} - \mu_1 - \mu_2) \log(1 + \mu_{12} - \mu_1 - \mu_2)
  \]
- The variational problem:
  \[
  A(\theta) = \max_{(\mu_1, \mu_2, \mu_{12}) \in \mathcal{M}} \{\theta_1 \mu_1 + \theta_2 \mu_2 + \theta_{12} \mu_{12} - A^*(\mu)\}
  \]
- The optimum is attained at:
  \[
  \mu_1(\theta) = \frac{\exp(\theta_1) + \exp(\theta_2 + \theta_{12})}{1 + \exp(\theta_1) + \exp(\theta_2) + \exp(\theta_1 + \theta_2 + \theta_{12})}
  \]

Variational Principle

- Exact variational formulation:
  \[
  A(\theta) = \sup_{\mu \in \mathcal{M}} \{\theta^T \mu - A^*(\mu)\}
  \]
  
  - \( \mathcal{M} \): the marginal polytope, difficult to characterize
  - \( A^* \): the negative entropy function, no explicit form

- Mean field method: non-convex inner bound and exact form of entropy

- Bethe approximation and loopy belief propagation: polyhedral outer bound and non-convex Bethe approximation
Mean Field Approximation

Tractable Subgraphs

- Definition: A subgraph F of the graph G is **tractable** if it is feasible to perform exact inference.

- Example:

\[ \Omega := \{ \theta \in \mathbb{R}^d | A(\theta) < +\infty \} \]

\[ \Omega(F_0) := \{ \theta \in \Omega | \theta_{(s,t)} = 0, \forall (s,t) \in E \} \]

\[ \Omega(T) := \{ \theta \in \Omega | \theta_{(s,t)} = 0 \ \forall (s,t) \notin E(T) \} \]
Mean Field Methods

- For an exponential family with sufficient statistics $\phi$ defined on graph $G$, the set of realizable mean parameter set
  \[ \mathcal{M}(G; \phi) := \{ \mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu \} \]
- For a given tractable subgraph $F$, a subset of mean parameters of interest
  \[ \mathcal{M}(F; \phi) := \{ \tau \in \mathbb{R}^d | \tau = \mathbb{E}_\theta[\phi(X)] \text{ for some } \theta \in \Omega(F) \} \]
- Inner approximation $\mathcal{M}(F; \phi)^o \subseteq \mathcal{M}(G; \phi)^o$
- Mean field solves the relaxed problem
  \[ \max_{\tau \in \mathcal{M}_F(G)} \{ \langle \tau, \theta \rangle - A^*_F(\tau) \} \]
  - $A^*_F = A^*|_{\mathcal{M}_F(G)}$ is the exact dual function restricted to $\mathcal{M}_F(G)$

Example: Naïve Mean Field for Ising Model

- Ising model in $\{0,1\}$ representation
  \[ p(x) \propto \exp \left\{ \sum_{s \in V} x_s \theta_s + \sum_{(s,t) \in E} x_s x_t \theta_{st} \right\} \]
- Mean parameters
  \[ \mu_s = \mathbb{E}_p[X_s] = \mathbb{P}[X_s = 1] \text{ for all } s \in V, \text{ and } \mu_{st} = \mathbb{E}_p[X_s X_t] = \mathbb{P}[X_s, X_t] = (1,1) \text{ for all } (s,t) \in E. \]
- For fully disconnected graph $F$,
  \[ \mathcal{M}_F(G) := \{ \tau \in \mathbb{R}^{|V|+|E|} | 0 \leq \tau_s \leq 1, \forall s \in V, \tau_{st} = \tau_s \tau_t, \forall (s,t) \in E \} \]
- The dual decomposes into sum, one for each node
  \[ A^*_F(\tau) = \sum_{s \in V} [\tau_s \log \tau_s + (1 - \tau_s) \log(1 - \tau_s)] \]
Example: Naïve Mean Field for Ising Model

- Mean field problem
  \[ A(\theta) \geq \max_{(\tau_1, \ldots, \tau_m) \in [0,1]^m} \left\{ \sum_{s \in V} \theta_s \tau_s + \sum_{(s,t) \in E} \theta_{st} \tau_s \tau_t - A^*_F(\tau) \right\} \]

- The same objective function as in free energy based approach

- The naïve mean field update equations
  \[ \tau_s \leftarrow \sigma \left( \theta_s + \sum_{t \in N(s)} \theta_s \tau_t \right) \]

- Also yields lower bound on log partition function

Geometry of Mean Field

- Mean field optimization is always non-convex for any exponential family in which the state space \( \mathcal{X}^m \) is finite

- Recall the marginal polytope is a convex hull
  \[ \mathcal{M}(G) = \text{conv} \{ \phi(e); e \in \mathcal{X}^m \} \]

- \( \mathcal{M}_F(G) \) contains all the extreme points
  - If it is a strict subset, then it must be non-convex

- Example: two-node Ising model
  \[ \mathcal{M}_F(G) = \{ 0 \leq \tau_1 \leq 1, 0 \leq \tau_2 \leq 1, \tau_{12} = \tau_1 \tau_2 \} \]
  - It has a parabolic cross section along \( \tau_1 = \tau_2 \), hence non-convex
Bethe Approximation and Sum-Product

Sum-Product/Belief Propagation Algorithm

- Message passing rule:
  \[ M_{ts}(x_s) \leftarrow \kappa \sum_{x'_t} \left\{ \psi_{st}(x_s, x'_t) \psi_t(x'_t) \prod_{u \in N(t)/s} M_{ut}(x'_t) \right\} \]

- Marginals:
  \[ \mu_s(x_s) = \kappa \psi_s(x_s) \prod_{t \in N(s)} M^*_ts(x_s) \]

- Exact for trees, but approximate for loopy graphs (so called loopy belief propagation)

- Question:
  - How is the algorithm on trees related to variational principle?
  - What is the algorithm doing for graphs with cycles?
Tree Graphical Models

- Discrete variables $X_s \in \{0, 1, \ldots, m_s - 1\}$ on a tree $T = (V, E)$
- Sufficient statistics:
  \[
  I_j(x_s) \quad \text{for } s = 1, \ldots, n, \quad j \in X_s \\
  \mathbb{I}_{jk}(x_s, x_t) \quad \text{for } (s, t) \in E, \quad (j, k) \in X_s \times X_t
  \]
- Exponential representation of distribution:
  \[
  p(x; \theta) \propto \exp \left\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s, t) \in E} \theta_{st}(x_s, x_t) \right\}
  \]
  where $\theta_s(x_s) := \sum_{j \in X_s} \theta_{s,j} \mathbb{I}_j(x_s)$ (and similarly for $\theta_{st}(x_s, x_t)$)
- Mean parameters are marginal probabilities:
  \[
  \mu_{s,j} = \mathbb{E}_p[\mathbb{I}_j(X_s)] = \mathbb{P}[X_s = j] \quad \forall j \in X_s, \quad \mu_s(x_s) = \sum_{j \in X_s} \mu_{s,j} \mathbb{I}_j(x_s) = \mathbb{P}(X_s = x_s)
  \]
  \[
  \mu_{st,jk} = \mathbb{E}_p[\mathbb{I}_{st,jk}(X_s, X_t)] = \mathbb{P}[X_s = j, X_t = k] \quad \forall (j, k) \in X_s \times X_t.
  \]
  \[
  \mu_{st}(x_s, x_t) = \sum_{(j, k) \in X_s \times X_t} \mu_{st,jk} \mathbb{I}_{jk}(x_s, x_t) = \mathbb{P}(X_s = x_s, X_t = x_t)
  \]

Marginal Polytope for Trees

- Recall marginal polytope for general graphs
  \[
  \mathcal{M}(G) = \{ \mu \in \mathbb{R}^d \mid \exists p \text{ with marginals } \mu_{s,j}, \mu_{st,jk} \}
  \]
- By junction tree theorem (see Prop. 2.1 & Prop. 4.1)
  \[
  \mathcal{M}(T) = \left\{ \mu \geq 0 \mid \sum_{x_s} \mu_s(x_s) = 1, \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s) \right\}
  \]
- In particular, if $\mu \in \mathcal{M}(T)$, then
  \[
  p_\mu(x) := \prod_{s \in V} \mu_s(x_s) \prod_{(s, t) \in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s) \mu_t(x_t)},
  \]
  has the corresponding marginals
Decomposition of Entropy for Trees

- For trees, the entropy decomposes as
  \[ H(p(x; \mu)) = - \sum_x p(x; \mu) \log p(x; \mu) \]
  \[ = \sum_{s \in V} \left( - \sum_{x_s} \mu_s(x_s) \log \mu_s(x_s) \right) - \sum_{(s, t) \in E} \left( \sum_{x_s, x_t} \mu_{st}(x_s, x_t) \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s) \mu_t(x_t)} \right) \]
  \[ = \sum_{s \in V} H_s(\mu_s) - \sum_{(s, t) \in E} I_{st}(\mu_{st}) \]

- The dual function has an explicit form \( A^*(\mu) = -H(p(x; \mu)) \)

Exact Variational Principle for Trees

- Variational formulation
  \[ A(\theta) = \max_{\mu \in \mathcal{M}(T)} \left\{ (\theta, \mu) + \sum_{s \in V} H_s(\mu_s) - \sum_{(s, t) \in E} I_{st}(\mu_{st}) \right\} \]

- Assign Lagrange multiplier \( \lambda_{ss} \) for the normalization constraint \( C_{ss}(\mu) := 1 - \sum x_s \mu_s(x_s) = 0 \); and \( \lambda_{ts}(x_s, x_t) \) for each marginalization constraint \( C_{ts}(x_s; \mu) := \mu_s(x_s) - x_t \mu_{st}(x_s, x_t) = 0 \)

- The Lagrangian has the form
  \[ \mathcal{L}(\mu, \lambda) = (\theta, \mu) + \sum_{s \in V} H_s(\mu_s) - \sum_{(s, t) \in E} I_{st}(\mu_{st}) + \sum_{s \in V} \lambda_{ss} C_{ss}(\mu) \]
  \[ + \sum_{(s, t) \in E} \left[ \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t) + \sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s) \right] \]
Lagrangian Derivation

- Taking the derivatives of the Lagrangian w.r.t. $\mu_s$ and $\mu_{st}$

\[
\frac{\partial L}{\partial \mu_s(x_s)} = \theta_s(x_s) - \log \mu_s(x_s) + \sum_{t \in N(s)} \lambda_{ts}(x_s) + C
\]

\[
\frac{\partial L}{\partial \mu_{st}(x_s, x_t)} = \theta_{st}(x_s, x_t) - \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s) \mu_t(x_t)} - \lambda_{ts}(x_s) - \lambda_{st}(x_t) + C'
\]

- Setting them to zeros yields

\[
\mu_s(x_s) \propto \exp\{\theta_s(x_s)\} \prod_{t \in N(s)} \exp\{\lambda_{ts}(x_s)\}
\]

\[
\mu_s(x_s, x_t) \propto \exp\{\theta_s(x_s) + \theta_t(x_t) + \theta_{st}(x_s, x_t)\} \times \prod_{u \in N(s) \setminus t} \exp\{\lambda_{us}(x_s)\} \prod_{v \in N(t) \setminus s} \exp\{\lambda_{vt}(x_t)\}
\]

Lagrangian Derivation (continued)

- Adjusting the Lagrange multipliers or messages to enforce

\[
C_{ts}(x_s; \mu) := \mu_s(x_s) - \sum_{x_t} \mu_{st}(x_s, x_t) = 0
\]

yields

\[
M_{ts}(x_s) \leftarrow \sum_{x_t} \exp\{\theta_t(x_t) + \theta_{st}(x_s, x_t)\} \prod_{u \in N(t) \setminus s} M_{ut}(x_t)
\]

- Conclusion: the message passing updates are a Lagrange method to solve the stationary condition of the variational formulation
BP on Arbitrary Graphs

- Two main difficulties of the variational formulation

\[ A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \theta^T \mu - A^*(\mu) \} \]

- The marginal polytope \( \mathcal{M} \) is hard to characterize, so let’s use the tree-based outer bound

\[ \mathcal{L}(G) = \left\{ \tau \geq 0 \mid \sum_{x_s} \tau_s(x_s) = 1, \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s) \right\} \]

These locally consistent vectors \( \tau \) are called pseudo-marginals.

- Exact entropy \(-A^*(\mu)\) lacks explicit form, so let’s approximate it by the exact expression for trees

\[ -A^*(\tau) \approx H_{\text{Bethe}}(\tau) := \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}). \]

Bethe Variational Problem (BVP)

- Combining these two ingredient leads to the Bethe variational problem (BVP):

\[
\max_{\tau \in \mathcal{L}(G)} \left\{ \langle \theta, \tau \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}) \right\}. 
\]

- A simple structured problem (differentiable & constraint set is a simple convex polytope)

- Loopy BP can be derived as an iterative method for solving a Lagrangian formulation of the BVP (Theorem 4.2); similar proof as for tree graphs
Geometry of BP

- Consider the following assignment of pseudo-marginals
  - Can easily verify \( \tau \in \mathbb{L}(G) \)
  - However, \( \tau \not\in \mathcal{M}(G) \) (need a bit more work)

- Tree-based outer bound
  - For any graph, \( \mathcal{M}(G) \subseteq \mathbb{L}(G) \)
  - Equality holds if and only if the graph is a tree

- Question: does solution to the BVP ever fall into the gap?
  - Yes, for any element of outer bound \( \mathbb{L}(G) \), it is possible to construct a distribution with it as a BP fixed point (Wainwright et al. 2003)

Inexactness of Bethe Entropy Approximation

- Consider a fully connected graph with
  \[
  \mu_s(x_s) = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \quad \text{for } s = 1, 2, 3, 4
  \]
  \[
  \mu_{st}(x_s, x_t) = \begin{bmatrix} 0.5 & 0 & 0.5 \end{bmatrix} \quad \forall (s, t) \in E.
  \]

- It is globally valid: \( \tau \in \mathcal{M}(G) \); realized by the distribution that places mass 1/2 on each of configuration (0,0,0,0) and (1,1,1,1)

- \( H_{\text{Bethe}}(\mu) = 4\log 2 - 6\log 2 = -2\log 2 < 0 \),
- \( -A^*(\mu) = \log 2 > 0 \).
Remark

- This connection provides a principled basis for applying the sum-product algorithm for loopy graphs.

- However,:
  - Although there is always a fixed point of loopy BP, there is no guarantees on the convergence of the algorithm on loopy graphs.
  - The Bethe variational problem is usually non-convex. Therefore, there are no guarantees on the global optimum.
  - Generally, no guarantees that $A_{\text{Bethe}}(\theta)$ is a lower bound of $A(\theta)$.

- Nevertheless:
  - The connection and understanding suggest a number of avenues for improving upon the ordinary sum-product algorithm, via progressively better approximations to the entropy function and outer bounds on the marginal polytope (Kikuchi clustering).

Summary

- Variational methods in general turn inference into an optimization problem via exponential families and convex duality.

- The exact variational principle is intractable to solve; there are two distinct components for approximations:
  - Either inner or outer bound to the marginal polytope.
  - Various approximation to the entropy function.

- Mean field: non-convex inner bound and exact form of entropy.
- BP: polyhedral outer bound and non-convex Bethe approximation.
- Kikuchi and variants: tighter polyhedral outer bounds and better entropy approximations (Yedidia et. al. 2002).