Exponential Families: Marginal Polytopes

\[ p(x \mid \theta) = \exp\{\theta^T \phi(x) - \Phi(\theta)\} \quad \Phi(\theta) = \log \sum_x \exp\{\theta^T \phi(x)\} \]

- Any joint distribution has a unique set of expected statistics (moments):
  \[ \mu = \nabla_\theta \Phi(\theta) = \mathbb{E}_\theta \left[ \phi(x) \right] = \sum_x \phi(x) p(x \mid \theta) \]

- The convex marginal polytope is the set of feasible expected statistics:
  \[ \mathcal{M} \triangleq \{ \mu \in \mathbb{R}^d \mid \exists p \text{ such that } \mathbb{E}_p[\phi(x)] = \mu \} \subseteq [0, 1]^d \]

- Vertex representation: Convex hull of feature vectors for (exponentially large) number of joint states:
  \[ \mathcal{M} = \text{conv}\{\phi(x) \mid x \in \mathcal{X}\} \]

- Facet representation: Linear moment constraints (how many?) defining boundaries of polytope:
  \[ \mathcal{M} = \{ \mu \in \mathbb{R}^d \mid \langle a_j, \mu \rangle \geq b_j \ \forall j \in \mathcal{J} \} \]
Exponential Families: Variational Inference

\[ p(x \mid \theta) = \exp\{\theta^T \phi(x) - \Phi(\theta)\} \quad \Phi(\theta) = \log \sum_x \exp\{\theta^T \phi(x)\} \]

- Canonical parameters & moments:
  \[ \Theta \triangleq \{\theta \in \mathbb{R}^d \mid \Phi(\theta) < +\infty\} \]
  \[ \mathcal{M} \triangleq \{\mu \in \mathbb{R}^d \mid \exists p \text{ such that } \mathbb{E}_p[\phi(x)] = \mu\} \]

- Inference: Find moments of model with known parameters (joint distribution)
  \[ \mu = \nabla_\theta \Phi(\theta) = \mathbb{E}_\theta[\phi(x)] = \sum_x \phi(x)p(x \mid \theta) \]

- From KLD, moments are solution to a variational optimization problem:
  \[ \Phi(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \theta^T \mu + H(p(x \mid \theta(\mu)) \right\} \]

For learning from data, both log-normalizer and moments are important.
Variational Methods: Naïve and Structured Mean Field
Mean Field assumes Markov with respect to sub-graph $F$ of original graph $G$:
- Equivalent to constraining some exponential family parameters to equal zero
- Sub-graph picked so that entropy is “simple”, and thus optimization tractable

Mean field provides lower bound on true log-normalizer:
- Optimize over smaller set where true objective can be evaluated
- No approximation to any terms in variational objective, only to constraint set

Mean field optimization has local optima:
- Inner approx has all vertices but not full marginal polytope, so never convex
Naïve Mean Field Approximations

- Express pairwise MRF in exponential family ("energy") form:
\[
p(x) = \frac{1}{Z} \exp \left\{ - \sum_{(s,t) \in \mathcal{E}} \phi_{st}(x_s, x_t) - \sum_{s \in \mathcal{V}} \phi_s(x_s) \right\}
\]
\[
\phi_{st}(x_s, x_t) = - \log \psi_{st}(x_s, x_t)
\]
\[
\phi_s(x_s) = - \log \psi_s(x_s)
\]

- A naïve mean field method approximates distribution as fully factorized:
\[
q(x) = \prod_{s \in \mathcal{V}} q_s(x_s)
\]
\[
q_s(x_s = k) = \mu_{sk} \geq 0, \quad \sum_{k=1}^{K_s} \mu_{sk} = 1.
\]

Free parameters to be optimized:
Naïve Mean Field Updates

Free Energy: \( F(\mu) = -H(\mu) + E(\mu) \) \( q(x) = \prod_{s \in V} q_s(x_s) \)

Entropy: \( H(\mu) = -\sum_{s \in V} \sum_k \mu_{sk} \log \mu_{sk} \) \( q_s(x_s = k) = \mu_{sk} \)

Average Energy: \( E(\mu) = \sum_{(s,t) \in E} \sum_{k,\ell} \mu_{sk} \mu_{t\ell} \phi_{st}(k, \ell) + \sum_{s \in V} \sum_k \mu_{sk} \phi_s(k) \)

• Constraints which any feasible solution must satisfy:
\[ \sum_k \mu_{sk} = 1 \text{ for all nodes } s \in V. \]

• Lagrangian encoding objective and constraints:
\[ \mathcal{L}(\mu) = -H(\mu) + E(\mu) + \sum_{s \in V} \lambda_s (1 - \sum_k \mu_{sk}) \]

• Coordinate descent: Optimize one node marginal fixing others, iterate.
Mean Field as Message Passing

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

\[ \phi_{st}(x_s, x_t) = -\log \psi_{st}(x_s, x_t) \]

\[ q_i(x_i) \propto \psi_i(x_i) \prod_{j \in \Gamma(i)} m_{ji}(x_i) \]

\[ m_{ji}(x_i) \propto \exp \left\{ -\sum_{x_j} \phi_{ij}(x_i, x_j) q_j(x_j) \right\} \]

- Compared to belief propagation, has identical formula for estimating marginals from messages, but a different message update equation
- If neighboring marginals degenerate to single state, recover Gibbs sampling message
Any subgraph for which inference is tractable leads to a mean field approximation for which the update equations are tractable.
Tree-Structured Distributions

- Distributions that are Markov with respect to tree factorize, with parameters:

\[ q_s(x_s), s \in V \quad q_{st}(x_s, x_t), (s, t) \in E \quad \sum_{x_t} q_{st}(x_s, x_t) = q_s(x_s) \]

- The entropy of the tree-structured distribution then decomposes as:

\[ H(q) = \sum_{s \in V} H_s(q_s) - \sum_{(s, t) \in E} I_{st}(q_{st}) \]

\[ H_s(q_s) = -\sum_{x_s} q_s(x_s) \log q_s(x_s) \quad \text{Node Entropies} \]

\[ I_{st}(q_{st}) = \sum_{x_s, x_t} q_{st}(x_s, x_t) \log \frac{q_{st}(x_s, x_t)}{q_s(x_s)q_t(x_t)} \quad \text{Edge Mutual Info} \]

\[ q(x) = \prod_{s \in V} q_s(x_s) \prod_{(s, t) \in E} \frac{q_{st}(x_s, x_t)}{q_s(x_s)q_t(x_t)} \]
Partition the graph edges into two sets:

- $\mathcal{E}_c \rightarrow \textit{core}$ edges, dependence directly modeled: $q_{st}(x_s, x_t)$
- $\mathcal{E}_r \rightarrow \textit{residual}$ edges, assume nodes factorize: $q_s(x_s)q_t(x_t)$
Straightforward but Intimidating Objective

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

\[ \phi_{st}(x_s, x_t) = -\log \psi_{st}(x_s, x_t) \]

\[ \phi_s(x_s) = -\log \psi_s(x_s) \]

\[ \mathcal{L}(q, \lambda) = \]

\[ + \sum_{s \in \mathcal{V}} \sum_{x_s} q_s(x_s) (\phi_s(x_s) + \log q_s(x_s)) \]

\[ + \sum_{(s,t) \in \mathcal{E}_r} \sum_{x_s, x_t} q_s(x_s) q_t(x_t) \phi_{st}(x_s, x_t) \]

\[ + \sum_{(s,t) \in \mathcal{E}_c} \sum_{x_s, x_t} q_{st}(x_s, x_t) \left( \phi_{st}(x_s, x_t) + \log \frac{q_{st}(x_s, x_t)}{q_s(x_s) q_t(x_t)} \right) \]

\[ + \sum_{s \in \mathcal{V}} \lambda_{ss} \left( 1 - \sum_{x_s} q_s(x_s) \right) \]

\[ + \sum_{(s,t) \in \mathcal{E}_c} \left[ \sum_{x_s} \lambda_{ts}(x_s) \left( q_s(x_s) - \sum_{x_t} q_{st}(x_s, x_t) \right) + \sum_{x_t} \lambda_{st}(x_t) \left( q_t(x_t) - \sum_{x_s} q_{st}(x_s, x_t) \right) \right] \]
MF & BP: Message Passing

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

\[ \phi_{st}(x_s, x_t) = -\log \psi_{st}(x_s, x_t) \]

\[ \phi_s(x_s) = -\log \psi_s(x_s) \]

Beliefs:

**pseudo-marginals**

\[ q_t(x_t) = \frac{1}{Z_t} \psi_t(x_t) \prod_{u \in \Gamma(t)} m_{ut}(x_t) \]

**MF:**

\[ m_{ts}(x_s) \propto \exp \left\{ -\sum_{x_t} \phi_{st}(x_s, x_t) q_t(x_t) \right\} \]

**residual**

**BP:**

\[ m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t) \frac{q_t(x_t)}{m_{st}(x_t)} \]

- **Naïve mean field:** All edges in residual, guaranteed convergent
- **Structured mean field:** Acyclic subset of edges in core, remainder in residual, guaranteed convergent and strictly more expressive
- **Loopy belief propagation:** All edges in core, captures most direct dependences, but approximation uncontrolled and may not converge
Variational Methods:
Bethe Approximations & Loopy BP

Fig. 4.2 Highly idealized illustration of the relation between the marginal polytope $\mathcal{M}(G)$ and the outer bound $\mathcal{L}(G)$. The set $\mathcal{L}(G)$ is always an outer bound on $\mathcal{M}(G)$, and the inclusion $\mathcal{M}(G) \subset \mathcal{L}(G)$ is strict whenever $G$ has cycles. Both sets are polytopes and so can be represented either as the convex hull of a finite number of extreme points, or as the intersection of a finite number of half-spaces, known as facets.

Letting $\phi$ be a shorthand for the full vector of indicator functions in the standard overcomplete representation (3.34), the marginal polytope has the convex hull representation $\mathcal{M}(G) = \text{conv}\{\phi(x) | x \in X\}$. Since the indicator functions are $\{0, 1\}$-valued, all of its extreme points consist of $\{0, 1\}$ elements, of the form $\mu_x := \phi(x)$ for some $x \in X$; there are $|X|_m$ such extreme points. However, with the exception of tree-structured graphs, the number of facets for $\mathcal{M}(G)$ is not known in general, even for relatively simple cases like the Ising model; see the book [69] for background on the cut or correlation polytope, which is equivalent to the marginal polytope for an Ising model. However, the growth must be super-polynomial in the graph size, unless certain widely believed conjectures in computational complexity are false.

On the other hand, the polytope $\mathcal{L}(G)$ has a polynomial number of facets, upper bounded by any graph by $O(r_m + r^2|E|)$. It has more extreme points than $\mathcal{M}(G)$, since in addition to all the integral extreme points $\{\mu_x, x \in X\}$, it includes other extreme points $\tau \in \mathcal{L}(G) \setminus \mathcal{M}(G)$ that contain fractional elements; see Section 8.4 for further discussion of integral versus fractional extreme points. With the exception of trees and small instances, the total number of extreme points of $\mathcal{L}(G)$ is not known in general.
Tree-Based Outer Approximations

• For some graph $G$, denote true marginal polytope by $\mathbb{M}(G)$

• Given marginals for nodes & edges, impose local consistency constraints $\mathbb{L}(G)$

$$\sum_{x_s} \mu_s(x_s) = 1, \quad s \in V$$

$$\mu_s(x_s) \geq 0, \mu_{st}(x_s, x_t) \geq 0$$

$$\sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s), \quad (s, t) \in E, x_s \in X_s$$

• For any graph, this is a convex outer bound: $\mathbb{M}(G) \subseteq \mathbb{L}(G)$

• For any tree-structured graph $T$, we have $\mathbb{M}(T) = \mathbb{L}(T)$
Marginals and Pseudo-Marginals

**Local Constraints Exactly Represent Trees:**
Construct joint consistent with any marginals

\[ p_\mu(x) = \prod_{(s,t) \in \mathcal{E}} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s) \mu_t(x_t)} \prod_{s \in \mathcal{V}} \mu_s(x_s) \]

**For Any Graph with Cycles, Local Constraints are Loose:**

Consider three binary variables and restrict \( \mu_1 = \mu_2 = \mu_3 = 0.5 \) denote potentially invalid pseudo-marginals by \( \tau_s, \tau_{st} \)
Properties of Local Constraint Polytope

\[
\sum_{x_s} \mu_s(x_s) = 1, \quad s \in \mathcal{V} \quad \mu_s(x_s) \geq 0, \mu_{st}(x_s, x_t) \geq 0
\]

\[
\sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s), \quad (s, t) \in \mathcal{E}, x_s \in \mathcal{X}_s
\]

- Number of faces upper bounded by \( O(KN + K^2 E) \) for graphs with \( N \) nodes, \( E \) edges, \( K \) discrete states per node
- Contains all of the degenerate vertices of true marginal polytope, as well as additional \textit{fractional} vertices (total number unknown in general)
Bethe Variational Methods

\[ \Phi(\theta) \approx \sup_{\tau \in \mathbb{L}(G)} \left\{ \theta^T \tau + H_B(\tau) \right\} \]

\[ H_B(\tau) = \sum_{s \in \mathcal{V}} H_s(\tau_s) - \sum_{(s,t) \in \mathcal{E}} I_{st}(\tau_{st}) \]

- Local consistency constraints are convex, but allow globally inconsistent pseudo-marginals on graphs with cycles
- Bethe entropy approximation may be not be concave, and may not even be a valid (non-negative) entropy

**Example:** Four binary variables

\[ p_\mu(0,0,0,0) = p_\mu(1,1,1,1) = 0.5 \]

\[ \mu_s(x_s) = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \text{ for } s = 1,2,3,4 \]

\[ \mu_{st}(x_s,x_t) = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \forall (s,t) \in E. \]

\[ H_B(\mu) = 4 \log 2 - 6 \log 2 = -2 \log 2 \]

\[ H(\mu) = \log 2 \]
Tree-Based Entropy Bounds

\[ p(x) = \frac{1}{Z} \exp \left\{ - \sum_{(s,t) \in \mathcal{E}} \phi_{st}(x_s, x_t) - \sum_{s \in \mathcal{V}} \phi_s(x_s) \right\} \]

\[ H(\mu(T)) = \sum_{s \in \mathcal{V}} H_s(\mu_s) - \sum_{(s,t) \in \mathcal{E}(T)} I_{st}(\mu_{st}) \]

\[ H(\mu) \leq H(\mu(T)) \quad \text{for any tree } T \]

Maximum entropy property of exponential families:

- Original distribution maximizes entropy subject to constraints:
  \[ \mathbb{E}_p[\phi_{st}(x_s, x_t)] = \mu(x_s, x_t), \quad (s, t) \in \mathcal{E} \]

- Tree-structured distribution maximizes subject to a *subset* of the full constraints (those corresponding to edges in tree):
  \[ \mathbb{E}_p[\phi_{st}(x_s, x_t)] = \mu(x_s, x_t), \quad (s, t) \in \mathcal{E}(T) \]
Tree-Based Entropy Bounds

\[ p(x) = \frac{1}{Z} \exp \left\{ - \sum_{(s,t) \in E} \phi_{st}(x_s, x_t) - \sum_{s \in V} \phi_s(x_s) \right\} \]

\[ H(\mu(T)) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \]

\[ H(\mu) \leq H(\mu(T)) \quad \text{for any tree } T \]

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

- Family of bounds depends on edge appearance probabilities (one number per edge) from some distribution over subtrees in the original graph:

\[ H(\mu) \leq \sum_{T} \rho(T) H(\mu(T)) \quad \rho_{st} = \mathbb{E}_{\rho} \left[ \mathbb{I}[(s,t) \in E(T)] \right] \]
Reweighted Bethe Variational Methods

\[ \Phi(\theta) \leq \sup_{\tau \in \mathbb{L}(G)} \left\{ \theta^T \tau + H_\rho(\tau) \right\} \]

\[ H_\rho(\tau) = \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}) \]

• Local consistency constraints are convex, but allow globally inconsistent pseudo-marginals on graphs with cycles

• Assuming we pick weights corresponding to some distribution on acyclic sub-graphs, have upper bound on true entropy

• This defines a convex surrogate to true variational problem

Issues to resolve:

• Given edge weights, how can we efficiently find the best pseudo-marginals? A message-passing algorithm?

• There are many distributions over spanning trees. How can we find the best edge appearance probabilities?
Spanning Tree Polytope

Fig. 7.1 Illustration of valid edge appearance probabilities. Original graph is shown in panel (a). Probability \( \frac{1}{3} \) is assigned to each of the three spanning trees \( \{T_i | i = 1, 2, 3\} \) 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 \( \rho_e = \frac{2}{3} \) and \( \rho_f = \frac{1}{3} \).

must belong to the so-called spanning tree polytope \([54, 73]\) associated with \( G \). Note that these edge appearance probabilities must satisfy various constraints, depending on the structure of the graph. A simple example should help to provide intuition.

Example 7.1 (Edge Appearance Probabilities). Figure 7.1(a) shows a graph, and panels (b) through (d) show three of its spanning trees \( \{T_1, T_2, T_3\} \). Suppose that we form a uniform distribution \( \rho \) over these trees by assigning probability \( \rho(T_i) = \frac{1}{3} \) to each \( T_i \), \( i = 1, 2, 3 \).

Consider the edge with label \( f \); notice that it appears in \( T_1 \), but in neither of \( T_2 \) and \( T_3 \). Therefore, under the uniform distribution \( \rho \), the associated edge appearance probability is \( \rho_f = \frac{1}{3} \). Since edge \( e \) appears in two of the three spanning trees, similar reasoning establishes that \( \rho_e = \frac{2}{3} \). Finally, observe that edge \( b \) appears in any spanning tree (i.e., it is a bridge), so that it must have edge appearance probability \( \rho_b = 1 \).

In their work on fractional belief propagation, Wiegerinck and Heskes \([261]\) examined the class of reweighted Bethe problems of the form (7.11), but without the requirement that the weights \( \rho_{st} \) belong to the spanning tree polytope. Although loosening this requirement does yield a richer family of variational problems, in general one loses

\[
\Phi(\theta) \leq \sup_{\tau \in \mathbb{L}(G)} \left\{ \theta^T \tau + H_\rho(\tau) \right\} \\
H_\rho(\tau) = \sum_{s \in \mathcal{V}} H_s(\tau_s) - \sum_{(s,t) \in \mathcal{E}} \rho_{st} I_{st}(\tau_{st})
\]

• Bound holds assuming edge weights lie in the spanning tree polytope (generated by some valid distribution on trees)
• Optimize via conditional gradient method:
  - Find descent direction by maximizing gradient (linear) over constraints
  - For spanning tree polytope, this is a max-weight spanning tree problem
  - Iteratively tightens bound on the true partition function

Bertsekas 1999
\[ p(x) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s \in \mathcal{V}} \psi_s(x_s) \]

**Beliefs:** pseudo-marginals

\[ q_t(x_t) = \frac{1}{Z_t} \psi_t(x_t) \prod_{u \in \Gamma(t)} m_{ut}(x_t) \]

**Mean Field**

\[ m_{ts}(x_s) \propto \exp \left\{ \sum_{x_t} q_t(x_t) \log \psi_{st}(x_s, x_t) \right\} \]

**Loopy BP**

\[ m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t) \frac{q_t(x_t)}{m_{st}(x_t)} \]

**Reweighted BP**

\[ m_{ts}(x_s) \propto \left[ \sum_{x_t} \psi_{st}(x_s, x_t)^{1/\rho_{st}} \frac{q_t(x_t)}{m_{st}(x_t)^{1/\rho_{st}}} \right]^{\rho_{st}} \]

- **Reweighted BP** becomes loopy BP when \( \rho_{st} = 1 \)
- **Reweighted BP** approaches mean field as \( \rho_{st} \to \infty \)