Variational Inference: Sum Product
10708, Fall 2020
Pradeep Ravikumar

1 Pairwise Discrete UGMs

Recall the overcomplete exponential family representation of discrete UGMs. Suppose we have variables $X = (X_1, \ldots, X_p)$ each taking values in $\mathcal{X} = \otimes_{s=1}^{p} \mathcal{X}_s$. Suppose we have an undirected graph $G = (V, E)$ with nodes V associated with each of the random variables $\{X_i\}_{i \in [p]}$. Then the exponential family correponding to the set sufficient statistics:

$$\phi(x) = \{\mathbb{I}[x_s = j\}_{s \in V, j \in \mathcal{X}_s} \cup \{\mathbb{I}[(x_s, x_t) = (j, k)]\}_{(s,t) \in E,\, j \in \mathcal{X}_s, k \in \mathcal{X}_t},$$

is given as:

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

where we use the shorthand:

$$\theta_s(x_s) = \sum_{j \in \mathcal{X}_s} \theta_{s;j} \mathbb{I}[x_s = j]$$
$$\theta_{st}(x_s, x_t) = \sum_{j \in \mathcal{X}_s, k \in \mathcal{X}_t} \theta_{st;jk} \mathbb{I}[x_s = j, x_t = k].$$

Associated with these overcomplete exponential family canonical parameters are also mean parameters for which we will use the shorthand:

$$\mu_s(x_s) = \sum_{j \in \mathcal{X}_s} \mu_{s;j} \mathbb{I}[x_s = j]$$
$$\mu_{st}(x_s, x_t) = \sum_{j \in \mathcal{X}_s, k \in \mathcal{X}_t} \mu_{st;jk} \mathbb{I}[x_s = j, x_t = k].$$

Thus, $\{\mu_s\}$ and $\{\mu_{st}\}$ are the nodewise and pairwise marginals of the exponential family distribution. These lie in the marginal polytope:

$$\mathcal{M}(G) = \{\mu \in \mathbb{R}^d \,|\, \exists\ p \text{ with nodewise marginals } \{\mu_s\} \text{ and pairwise marginals } \{\mu_{st}\}\}.$$

Recall the variational characterization of the log-partition function:

$$A(\theta) = \sup_{\mu \in \mathcal{M}} (\langle \theta, \mu \rangle - A^*(\mu)). \tag{1}$$

This is a convex optimization problem: the constraint set $\mathcal{M}$ is convex, specifically a convex polytope in the case of discrete overcomplete UGMs. While the negative entropy $A^*(\mu)$ is convex as well. However, both of these objects do not have a closed form expression. Characterizing the marginal polytope would require exponentially many halfspaces in general, and computing the entropy is in turn intractable. In what follows, we will be approximating both of these components with tractable alternatives.

2 Tree-based Outer Bound to $\mathcal{M}(G)$

As discussed earlier, $\mathcal{M}(G)$ is a convex polytope, and can be expressed as a convex hull of the sufficient statistics $\{\phi(x)\}_{x\in\mathcal{X}}$. So it has a finite number of vertices, but clearly an exponential number so that this is a not a tractable characterization. One could also characterize the polytope via half-space constraints (also known as facets), but even this can be shown to require an exponential number of constraints. But what if we only specify a subset of polynomially many constraints? This would then specify an outer bound of the marginal polytope.

Consider a candidate set of node-wise functions $\{\tau_s\}_{s\in V}$ and edge-wise functions $\{\tau_{st}\}_{(s,t)\in E}$. What are some constraints that will entail that they be marginals of some joint distribution p? One condition is that these functions be non-negative. Second, that the nodewise functions normalize to one:

$$\sum_{x_s} \tau_s(x_s) = 1,\ \forall s \in V, x_s \in \mathcal{X}_s. \tag{2}$$

And moreover, that the edgewise and nodewise candidate-marginal functions be consistent, so that $\forall (s,t) \in E$:

$$\sum_{x'_t} \tau_{st}(x_s, x'_t) = \tau_s(x_s),\ \forall x_s \in \mathcal{X}_s \tag{3}$$

$$\sum_{x'_s} \tau_{st}(x'_s, x_t) = \tau_t(x_t),\ \forall x_t \in \mathcal{X}_t. \tag{4}$$

Note that these local consistency constraints together with node-wise function normalization constraints also entail that the edgewise functions also normalize to one. These constraints together specify the polytope:

$$\begin{aligned}\mathcal{L}(G) = \{\tau \in \mathbb{R}_+^d \mid &\text{ condition (2) holds for all } s \in V,\\ &\text{and conditions (3), and (4) hold for all } (s,t) \in E\}.\end{aligned} \tag{5}$$

The set $\mathcal{L}(G)$ could be viewed as a set of locally consistent *pseudo-marginals*.

Theorem 1 *The set $\mathcal{L}(G)$ of pseudo-marginals defined in* (5) *satisfies:*

$$\mathcal{M}(G) \subseteq \mathcal{L}(G),$$

for any graph G, with equality when the graph G is a tree.

Proof. The inclusion follows from the observation that any set of true node and edgewise marginals definitely satisfy the normalization and local cconsistency constraints in $\mathcal{L}(G)$. Now suppose the graph is a tree, $G = T$. Now for any $\mu \in \mathcal{L}(T)$, consider the distribution

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

This can be seen to be a valid distribution just given the property that μ satisfies the local consistency properties in $\mathcal{L}(G)$. And moreover, by a simple induction argument, it can be seen that for this distribution, the node and edgewise marginals are precisely μ, so that it follows that $\mu \in \mathcal{M}(G)$ (since there exists a distribution that has μ as its node and edgewise marginals). $\square$

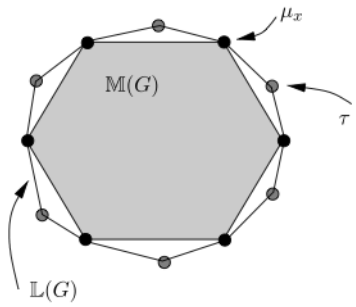

Figure 1: Outer Bound Local Polytope $\mathcal{L}(G)$

Thus, for tree-structured graphs, $\mathcal{L}(G)$ is an exact characterization of the marginal polytope, whereas for more general graphs, it is likely a strict upper bound.

3 Bethe Entropy Approximation

It turns out that for tree-structured graphs, one can also derive a closed-form expression for the negative entropy $A^*(\mu)$. Specifically, if the graph is a tree, then we know that we could write the distribution in reparameterized form:

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

Its entropy can then be written as:

$$\begin{aligned} H(p_\mu) &= -A^*(\mu) = \mathbb{E}_\mu[-\log p_\mu(X)] \\ &= \sum_{s\in V} \mathbb{E}_\mu[-\log \mu_s(X_s)] + \sum_{(s,t)\in E} \mathbb{E}_\mu[-\log \frac{\mu_{st}(X_s, X_t)}{\mu_s(X_s)\mu_t(X_t)}] \\ &= \sum_{s\in V} H_s(\mu_s) - \sum_{(s,t)\in E} I_{st}(\mu_{st}), \end{aligned}$$

where the first set of terms are the nodewise entropies:

$$H_s(\mu_s) = -\sum_{x_s} \mu_s(x_s) \log \mu_s(x_s), \tag{6}$$

and the second set of terms are the edgewise mutual information:

$$I_{st}(\mu_{st}) = \sum_{x_s\in\mathcal{X}_s, x_t\in\mathcal{X}_t} \mu_{st}(x_s, x_t) \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}. \tag{7}$$

Thus for a tree-structured graph, $A^*(\mu)$ has a simple closed form expression in terms of node-wise entropies and edgewise mutual information. The **Bethe approximation** to A^* is simply to use the closed form expression above which is exact for trees, as an inexact approximation for more general graphs:

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

4 Sum-Product as Variational Approximation

So far, we have discussed approximations to both of the two intractable components in the variational form for the log-partition function:

- The outer bound $\mathcal{L}(G)$ to the marginal polytope $\mathcal{M}(G)$ comprising locally consistent pseudo-marginals
- The Bethe entropy $H_{\text{bethe}}(\tau)$ as an approximation of the exact entropy $-A^*(\mu)$

Plugging these into the variational form for the log-partition function $A(\theta)$, we get the following so-called **Bethe variational**optimization problem:

$$\max_{\tau\in\mathcal{L}(G)} \left\{ \langle\theta, \tau\rangle + \sum_{s\in V} H_s(\tau_s) - \sum_{(s,t)\in E} I_{st}(\tau_{st}) \right\}. \tag{8}$$

We have the following natural proposition.

Proposition 2 *For tree-structured graphs G, the solution τ^* of the Bethe variational optimization problem in* (8) *corresponds to the exact node-wise and edge-wise marginals of the discrete exponential family UGM.*

Let us now consider how to solve this constrained optimization problem. Denote the normalization constraint expression as:

$$C_{ss}(\tau) = 1 - \sum_{x_s} \tau_s(x_s), \tag{9}$$

so that the normalization constraint is simply $C_{ss}(\tau) = 0$. Similarly, denote the local consistency constraint expression as:

$$C_{ts}(x_s, \tau) = \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s, x_t), \tag{10}$$

so that the local consistency constraints $C_{ts}(x_s, \tau) = 0$, and $C_{st}(x_t, \tau) = 0$. Consider the Lagrangian of the constrained optimization problem above, with Lagrange parameters λ_{ss} for the C_{ss} constraints, and parameters $\lambda_{ts}(x_s)$ and $\lambda_{st}(x_t)$ for the constraints C_{ts} and C_{st} respectively. We then get the Lagrangian:

$$\begin{aligned}\mathcal{L}(\tau, \lambda, \theta) = \langle \theta, \tau \rangle + H_{\text{bethe}}(\tau) + \sum_{s \in V} \lambda_{ss} C_{ss}(\tau) \\ + \sum_{(s,t) \in E} \left[\sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s, \tau) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t, \tau) \right].\end{aligned} \tag{11}$$

We have the following interesting theorem, due to Yedidia et al.

Theorem 3 *The sum-product message-passing updates are fixed point updates derived from stationary conditions on the Lagrangian in* (11)*, where the messages $M_{ts}(x_s)$ are related to the Lagrangian parameters $\lambda_{ts}(x_s)$ as $M_{ts}(x_s) = \exp(\lambda_{ts}(x_s))$. Any fixed point of the sum-product updates specifies a pair (τ^*, λ^*) that is a stationary point of the Lagrangian so that:*

$$\begin{aligned}\nabla_\tau \mathcal{L}(\tau^*, \lambda^*, \theta) = 0 \\ \nabla_\lambda \mathcal{L}(\tau^*, \lambda^*, \theta) = 0.\end{aligned}$$

A key consequence of this theorem is that it provides a principled basis for sum-product message passing even when the graph G is not a tree. It is simply solving for the Lagrangian of the Bethe variational optimization problem, which is an approximation of the exact variational form for the log-partition function. However, note that it is a simple fixed point update derived from the stationary condition of the Lagrangian of the Bethe variational optimization problem. It is thus not even guaranteed to converge. There have been some

approaches developed that convergently optimize the constrained problem, but these are typically not as fast "in practice" as sum-product, nor as simple, so the latter still remains the method of choice. Moreover, there is no guarantee that the pseudo-marginal stationary points of the Bethe variational optimization problem is close to the actual marginals. Indeed, note that the Bethe variational problem is not even convex, so that local optima need not even be global optima of the Bethe variational problem itself.

4.1 Reparameterization

We know that if the graph is a tree T, then the distribution p_θ admits a reparameterization in terms of its node and edgewise marginals as

$$p(x) = \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}.$$

This does not of course hold when the graph is not a tree. But interestingly, any local optimum pseudomarginals of the Bethe variational problem admits a reparameterization as shown in the following theorem.

Theorem 4 *Suppose* $(\tau * _s, s \in V;\ \tau^*_{st}, (s,t) \in E)$ *are any local optimum of the Bethe variational problem in* (8) *corresponding to a discrete exponential family UGM* p_θ*. It then holds that:*

$$p_\theta(x) = p_{\tau^*}(x) := \frac{1}{Z(\tau^*)} \prod_{s \in V} \tau^*_s(x_s) \prod_{(s,t) \in E} \frac{\tau^*_{st}(x_s, x_t)}{\tau^*_s(x_s)\tau^*_t(x_t)}.$$

Note that unlike in the case where the graph is a tree, the normalization constant $Z(\tau^*)$ for the sum-product pseudo-marginal based reparameterization is not in general equal to one.

5 Beyond Trees: Junction Tree Extensions

When the PGM has factors of size greater than two, and/or the graph is not a tree, we could then construct a junction tree T where the nodes correspond to cliques $\{C_i\}_{i \in V(T)}$ of the augmented chordal graph. Suppose the separator sets are denoted as $\{S_{ij}\}_{(i,j) \in E(T)}$.

It follows that the PGM distribution belongs to an exponential family with sufficient statistics:

$$\phi(x) = (\mathbb{I}[C_i = c_i], i \in V(T), c_i \in \mathrm{Val}(C_i);\ \mathbb{I}[S_{ij} = s_{ij}], (i,j) \in E(T), s_{ij} \in \mathrm{Val}(S_{ij})),$$

with the distribution specified as:

$$p_\theta(x) = \exp\left\{\langle\theta, \phi(x)\rangle - A(\theta)\right\},$$

for which as before, we could use shorthand to write:

$$p_\theta(x) = \exp\left\{\sum_{i\in V(T)} \theta_i(c_i) + \sum_{(i,j)\in E(T)} \theta_{ij}(s_{ij}) - A(\theta)\right\}.$$

As before, we could define the corresponding set of mean parameters with shorthand:

$$\mu_i(c_i) = \sum_{v\in \text{Val}(C_i)} \mu_{i;v}\mathbb{I}[c_i = v]$$
$$\mu_{ij}(s_{ij}) = \sum_{v\in \text{Val}} (S_{ij})\mu_{ij;v}\mathbb{I}[s_{ij} = v].$$

These define the marginal polytope:

$$\mathcal{M}(G) = \{\mu \,|\, \exists \text{ dist. } p \text{ with clique and separator set marginals } \mu\}.$$

Suppose we have candidate pseudomarginals τ_{C_i} over clique nodes C_i as well as those $\tau_{S_{ij}}$ over the separator sets S_{ij}.

$$\sum_{s_{ij}} \tau_{S_{ij}}(s_{ij}) = 1, \ \forall (i,j) \in E(T). \tag{12}$$

And moreover, that the clique and separate set pseudo-marginals be consistent, so that $\forall (i,j) \in E(T)$:

$$\sum_{c_i \setminus s_{ij}} \tau_{C_i}(c_i) = \tau_{S_{ij}}(s_{ij}), \ \forall s_{ij} \in \text{Val}(S_{ij}) \tag{13}$$
$$\sum_{c_j \setminus s_{ij}} \tau_{C_j}(c_j) = \tau_{S_{ij}}(s_{ij}), \ \forall s_{ij} \in \text{Val}(S_{ij}). \tag{14}$$

Note that these local consistency constraints together with separator-set function normalization constraints also entail that the clique-wise functions also normalize to one. These constraints together specify the polytope:

$$\mathcal{L}_T(G) = \{\tau \,|\, \text{ condition (12) holds for all } (i,j) \in E(T),$$
$$\text{and conditions (13), and (14) hold for all } (i,j) \in E(T)\}. \tag{15}$$

Along the lines of the junction tree consistency theorem, where local consistency entails global consistency, it can be shown that this is an exact characterization of the marginal polytope, so that:

$$\mathcal{L}_T(G) = \mathcal{M}(G).$$

Similarly, the RIP in the junction tree T ensures the reparameterization of p_θ in terms of its clique and separator set marginals μ:

$$p_\theta(X) = \frac{\prod_{i \in V_T} \mu_i(c_i)}{\prod_{(i,j) \in E(T)} \mu_{ij}(s_{ij})}.$$

This thus suggests the extension of the Bethe entropy for the negative entropy of p_θ:

$$H_{\text{bethe};T}(\mu) = \sum_{i \in V_T} H_i(\mu_i) - \sum_{(i,j) \in E(T)} H_{ij}(\mu_{ij}).$$

We thus get the following characterization of the log-partition function:

$$A(\theta) = \sup_{\mu \in \mathcal{L}_T(G)} \{\langle \theta, \mu \rangle - H_{\text{bethe};T}(\mu)\}.$$

Unlike in the sum-product case, this is an exact rather than approximate characterization of the log-partition function. Moreover, along similar lines as sum-product, it can presumably be shown that junction tree message passing updates solve for the stationary points of the Lagrangian of the above variational optimization problem. The caveat of course is that the complexity of this characterization scales exponentially with the size of the largest clique, which if we are lucky is the tree-width of the original graph G. This is thus not as scalable as simple sum-product as an approximate inference procedure for a general loopy graph G.

6 Converting Higher order UGMs to Pairwise UGMs

But what if we wish to just perform sum-product rather than full junction tree message passing, but that the UGM is not necessarily pairwise i.e. its local factors have size larger than two. In such a case, we can first augment the PGM to an extended pairwise PGM as follows. For each factor $\psi_C(x_C)$, we create a separate node Z_C, associated with a corresponding RV, which takes values in $\otimes_{s \in C} \mathcal{X}_s$. Thus, any value $z := (z_s, s \in C) \in \otimes_{s \in C} \mathcal{X}_s$ can be associated with values of the RVs $\{X_s\}_{s \in C}$ in X_C. We connect Z_C and each of X_s for $s \in C$, with pairwise factors $\psi_{C,s}(z, x_s) := \psi_C(z)^{1/|C|} \mathbb{I}[z_s = x_s]$. Consider the augmented graph $G = (\bar{V}, \bar{E})$ with nodes $\bar{V} = V \cup \{Z_C\}_{C \in \mathcal{C}}$, and edges $\bar{E} = E \cup \{(Z_C, X_s)\}_{C \in \mathcal{C}, s \in C}$, and with just the pairwise factors defined above.

It can then be seen that:

$$\psi_C(x_C) = \sum_{z \in \otimes_{s \in C} \mathcal{X}_s} \prod_{s \in C} \psi_{C,s}(z, x_s),$$

so the marginalization over the augmented variables in the augmented PGM results in the original PGM. On the other hand, the augmented PGM is pairwise, so that we can perform inference e.g. compute marginals in this augmented PGM to also get appropriate marginals in the original PGM.