Notes

Andrew Kontaxis, Chris White

November 4, 2017

1 Background

1.1 Sum Product Algorithm

Suppose we have a graphical model which is a connected undirected tree; in that case we can choose an arbitrary ordering for the nodes and write:

$$p(x_1, x_2, .., x_n) = \prod_i \psi_i(x_i) \prod_{i,j\in E} \psi_{i,j}(x_i, x_j)$$

Suspend disbelief about probabilistic interpretations for a moment and suppose we simply want to compute the quantity

$$p(x_s) := \sum_{i\neq s} p(x_1, x_2, .., x_s, ...x_n)$$

where we interpret the sum as being over the *state space* of the corresponding variables. We can write

$$\begin{aligned} p(x_1, x_2, .., x_s, ...x_n) &= \psi_s(x_s) \prod_{i\neq s} \psi_i(x_i) \prod_{i,j\in E} \psi_{i,j}(x_i, x_j) \\ &= \psi_s(x_s) \prod_{i\in\mathcal{N}(s)} \psi_i(x_i)\psi_{i,s}(x_i, x_s)\omega(T_i) \end{aligned}$$

where $\mathcal{N}(s)$ denotes the set of *neighbors* of node s and $\omega(T_i)$ is a *weighting* of the subtree containing node i formed by removing node s; this weighting is a function of all variables in T_i. Let us focus our attention on a single $i \in \mathcal{N}(s)$ for a moment, and imagine marginalizing out only the nodes in T_i first:

$$\sum_{j\in T_i} \psi_s(x_s) \prod_{k\in\mathcal{N}(s)} \psi_k(x_k)\psi_{k,s}(x_k, x_s)\omega(T_k) = \kappa(x_s, x_{V\setminus T_i}) \sum_{j\in T_i} \psi_k(x_i)\psi_{i,s}(x_i, x_s)\omega(T_i)$$

We should now see that in order to complete this computation, the weighting need only be given to us as a function of x_i alone. I.e., we can write

$$\kappa(x_s, x_{V\setminus T_i}) \sum_{x_i} \psi_i(x_i)\psi_{i,s}(x_i, x_s)\omega(x_i)$$

where $\omega(x_i)$ is given by

$$\omega(x_i) := \sum_{T_i\setminus i} \prod_{k\in\mathcal{N}(i)\setminus s} \psi_k(x_k)\psi_{k,s}(x_k, x_i)\omega(T_k)$$

and now we begin to see the recursive nature of our task. We can then proceed to marginalize out the variables in each of the other subtrees resulting in an expression of the form

$$\psi_s(x_s) \prod_{i\in\mathcal{N}(s)} \left(\sum_{x_i} \psi_i(x_i)\psi_{i,s}(x_i, x_s)\omega(x_i) \right)$$

Consequently, to compute the marginal $p(x_s)$, each neighbor i of s needs to pass a “message” to node s which is a function purely of x_s, specifying the ”weighting” of the subtree T_i *conditional on* the value x_s:

$$\mu_{i\to s}(x_s) := \sum_{x_i} \psi_i(x_i)\psi_{i,s}(x_i, x_s) \prod_{k\in\mathcal{N}(i)\setminus s} \mu_{k\to i}(x_i)$$

Note that the messages μ are proxies for the “true” weightings ω and that whenever $\mu \equiv \omega$ we have a fixed point and can compute

$$p(x_s) = \psi_s(x_s) \prod_{i \in \mathcal{N}(s)} \mu_{i \to s}(x_s).$$

This distinction is particularly important in the case when we want to apply this algorithm to graphs with cycles; otherwise we can iteratively updating the messages by beginning at the leaves with

$$\mu_{\ell \to q}(x_q) = \sum_{x_\ell} \psi_\ell(x_\ell)\psi_{\ell,q}(x_\ell, x_q)$$

1.1.1 An Example: Independent Sets on Graphs

An *independent set* of vertices on a graph is a subset of vertices such that no two are adjacent. This example will be particularly fruitful in future discussions; consider a binary tree whose nodes are assigned weights. We imagine this as an undirected graphical model by putting a binary random variable at each node, with potential functions given by:

$$\psi_i(x_i) := \exp(w_i x_i)$$
$$\psi_{ij}(x_i, x_j) := \chi(x_i + x_j <= 1)$$

Here $\chi(\cdot)$ is the indicator function for whether the condition holds. Let us see what the sum-product algorithm looks like for this model; first imagine a parent of two leaf nodes. Applying the above formulas, each leaf (here denoted ℓ and r) sends the message

$$\mu_{\ell \to p}(x_p) = \exp(w_\ell)\chi(x_p = 0) + 1$$
$$\mu_{r \to p}(x_p) = \exp(w_r)\chi(x_p = 0) + 1$$

and consequently the message the x_p sends to its parent x_P is given by

$$\begin{aligned}\mu_{p \to P}(x_P) &= \sum_{x_p} \left(\exp(w_\ell)\chi(x_p = 0) + 1\right)\left(\exp(w_r)\chi(x_p = 0) + 1\right)\exp(w_p x_p)\chi(x_p + x_P <= 1) \\ &= \left(\exp(w_\ell) + 1\right)\left(\exp(w_r) + 1\right) + \exp(w_p)\chi(x_P = 0)\end{aligned}$$

If we pause here and suppose x_p is the only node connected to x_P, we see that the marginal for x_P would be proportional to

$$\exp(w_P x_P) \cdot \left[\left(\exp(w_\ell) + 1\right)\left(\exp(w_r) + 1\right) + \exp(w_p)\chi(x_P = 0)\right]$$

Observe that each term of this summation corresponds to an allowable configuration for an independent set. For example, when $x_P = 1$ the leading term $\exp(w_P + w_\ell + w_r)$ corresponds to the independent set $\{P, \ell, r\}$. $\exp(w_\ell)$ occurs in precisely 2 terms because regardless of whether P is chosen, ℓ can be chosen in two possible independent sets. Lastly we note that the constant term corresponds to the null set.

2 Exponential Families

2.1 Exponential Families as Max Entropy Distributions

Suppose we are given a collection of *sufficient statistics* or *potentials* $\phi_\alpha : \mathcal{X} \to \mathbb{R}$ and a collection of *mean parameters* for each statistic, μ_α. There are possibly multiple probability distributions over $\mathcal{X}$ satisfying

$$\mathbb{E}[\phi_\alpha(X)] = \mu_\alpha \quad \forall \alpha \in \mathcal{I}$$

or there are none. For example, if $\phi_1(X) = X$ and $\phi_2(X) = X^2$ then by Jensen's inequality we must have

$$\mu_2 \geq \mu_1^2.$$

Thus not every combination of μ and ϕ will admit an admissible p.

Supposing there exists at least one distribution with mean parameters $\{\mu_\alpha\}_{\alpha \in \mathcal{I}}$, it is natural to ask for the distribution with the *maximal* amount of uncertainty satisfying the above mean parameter constraint (we want to encode the least amount of additional information above and beyond the constraints). Using *Shannon entropy* as our measure of "uncertainty" we are led to the optimization problem:

$$p^* := \arg\max_p -\mathbb{E}_p[\log(p)] \qquad \text{s.t.} \quad \mathbb{E}[\phi_\alpha(X)] = \mu_\alpha \quad \forall \alpha \in \mathcal{I}$$

2.1.1 Discrete Case

First suppose that $\mathcal{X}$ is discrete. In this case p is simply a non-negative vector which sums to 1 and we form the Lagrangian

$$\mathcal{L}(p, \lambda, \tau) := -\sum_{i=1}^{n} \log(p_i) p_i + \sum_{\alpha \in \mathcal{I}} \lambda_\alpha \left(\mathbb{E}[\phi_\alpha(X)] - \mu_\alpha\right) + \tau(\sum_{i=1}^{n} p_i - 1).$$

We have the necessary stationarity conditions

$$\frac{\partial \mathcal{L}}{\partial p_i} = -\log(p_i) - 1 + \sum_{\alpha \in \mathcal{I}} \lambda_\alpha \phi_\alpha(X_i) + \tau = 0$$

which implies that p^* is of the form

$$p^* = \exp\left(\sum_{\alpha \in \mathcal{I}} \lambda_\alpha \phi_\alpha + \tau - 1\right),$$

i.e., it is in the exponential family with *canonical* or *exponential* parameters λ_α and τ enforces the appropriate normalization.

2.1.2 Continuous Case

Something more subtle happens in the continuous case; for example, with ϕ_1 and ϕ_2 as above, if $\mu_1 = 1$ and $\mu_2 = 1$, the distribution is a degenerate delta.

2.2 Example: Gaussian