10-708 23/06/2010

L12-review

(*) core focus areas:-
i) loopy B.P. + Bethe approximation (L11a)
ii) conjugate duality, marginal polytope, convex polytope (L11b)

L11a - loopy BP

(?) - why are factor graphs introduced?

(?) - what is nature of graph in slides?

(*) local and global consistency

- define:-
$\{\tau_c, c \in G\}$, $\{\tau_s, s \in S\}$ as set of functions associated with cliques and separator sets.

- These sets of functions are locally consistent if the following properties hold:-

$$\sum_{x'_s} \tau_s(x'_s) = 1 \quad \forall s \in S \quad \text{(normalisation)} \qquad (I)$$

$$\sum_{x'_c | x'_s = x_s} \tau_c(x'_c) = \tau_s(x_s) \quad \forall c \in G, s \subset c \quad \text{(marginalisation)} \qquad (II)$$

- Essentially; is a condition that ensures we obtain valid marginals; in the sense of consistent probability distri.
- not fully sure about these conditions

(O/S 1) → clarity ↓

- Essentially states that calibrated clique and sepset beliefs give valid marginal probabilities.

(I) Functions associated with sepsets ⟶ proper marginals
(II) Summing any clique function τ_c over all variables in clique c which are not in the sepset, obtain $\tau_s(x_s)$

(*) Global consistency ⟶ τ_c and τ_s are valid marginals

(*) Example quoted
↳ pathological case
- see Jordan (2007) ch17.
- The situation amounts to a violation of the junction tree property
(*) Illustration of general rule that local consistency $\not\Rightarrow$ global consistency

---

(*) Approx. inf. as alternative:-
- For junction trees, local consistency is equivalent to global consistency.
- Why not convert all GMs → junction trees? (to perform exact inference)
- tree width and comp. complexity may still be too high to be tractable

---

(*) Belief propagation (message-update) equations
- Apply standard belief propagation/message passing to a loopy graph
- context is a UGM
- More specifically, an MN/UGM with N nodes, pairwise potentials (Yedidia 2002, 2001)
- 0152 An ambiguity (mild) → slides specify pairwise and singleton potentials.
- (Yedidia 2001, 2002) specify that the singleton potentials are local evidence nodes
- we have:-

Joint prob / Gibbs distri (?) of pairwise MRF:

$$p(x_1, \dots, x_N) = \frac{1}{Z} \prod_{i,j} \psi_{ij}(x_i, x_j) \prod_i \psi_i(x_i) \qquad (1)$$

- BP updates:-

$$M_{i\to j}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i, x_j)\, \psi_i(x_i) \prod_{k \in N(i)\setminus j} M_{k\to i}(x_i) \qquad \text{(messages)}$$

$$= \alpha \sum_{x_i} \psi_{ij}(x_i, x_j)\, \psi_i(x_i) \prod_{k \in N(i)\setminus j} M_{k\to i}(x_i)$$

$b_i(x_i) \propto \psi_i(x_i) \prod_{k \in N(i)} M_{k \to i}(x_i)$ (beliefs)

$= \alpha \psi_i(x_i) \prod_{k \in N(i)} M_{k \to i}(x_i)$

- α - normalisation constant
- $N(i) \backslash j$ - all nodes neighbouring i, except j
- $M_{k \to i}$ - message that node k sends to node i
- b_i - belief (marginal posterior probability at node i) (*)
- $\psi_{ij}(x_i, x_j)$ - pairwise potential / compatibility
- $\phi(x_i)$ - external evidence / local evidence for node i.

(*) Note: - messages recursively computed from incoming messages to that node (hence $\prod_{k \in N(i) \backslash j}$)

- beliefs computed from all incoming messages (and potential) to that node (hence $\prod_{k \in N(i)}$) (notice no $\backslash j$)

(*): when pairwise MRF is singly connected i.e. no loops; the beliefs are exact; BP is exact

when pairwise MRF is loopy; both exactness, convergence not guaranteed.

(*) Yedidia (2002):-
- empirical study by Murphy, Weiss, Jordan (1997)
- Pearl noted outcome where messages continually circulate with no convergence to stable equilibrium (1988)
- But loopy BP also performed well (i.e. gave good approx.) in certain situations e.g. turbo codes (near Shannon limit performance) or computer vision

(*) Yedidia (2001):-
theoretical clarification on what approximation B.P. represents on non-tree structured graphs (but also including tree-structured GMS as an incidental consequence)

(*) Approximating intract distri (recap) (see Yedidia 2002)
- distri Q to approx intractable distri P.
- KL defines distance between 2 prob distri. (recap) $P(X) = \frac{1}{Z} \prod_{f_a \in F} f_a(X_a)$

$$KL(Q\|P) = \sum_X Q(X) \log\left(\frac{Q(X)}{P(X)}\right)$$

$$= \sum_X Q(X) \log Q(X) - \sum_X Q(X) \log P(X)$$

$$= -H_Q(X) - \mathbb{E}_Q[\log P(X)]$$

$$= -H_Q(X) - \mathbb{E}_Q\left[\log \frac{1}{Z} \prod_{f_a \in F} f_a(X_a)\right]$$

$$= -H_Q(X) - \log\left(\frac{1}{Z}\right) - \sum_{f_a \in F} \mathbb{E}_Q[\log f_a(X_a)]$$

$$= -H_Q(X) - \sum_{f_a \in F} \mathbb{E}_Q[\log f_a(X_a)] + \log Z$$

$$= F(P,Q) + \log Z$$

- $F(P,Q) = -H_Q(X) - \sum_{f_a \in F} \mathbb{E}_Q[\log f_a(X_a)]$ - free energy

Q/S 3: connect this with presentation in L6

- $F(P,Q)$ - free energy
- $H_Q(X)$ - entropy

(*) Free energy, entropy for tree structured G-MS

- Tree structured G-MS have joint prob.:-

Yedidia

$$p(x_1,\ldots,x_8) = b(X) = \frac{\prod_a b_a(\underline{x}_a)}{\prod_i b_i(x_i)^{(d_i-1)}}$$

X - entire space of inputs (not a base)

(*) or in Yedidia's (2002) spec.:-

$$b(\{x\}) = \frac{\prod_{ij} b_{ij}(x_i,x_j)}{\prod_i b_i(x_i)^{q_i-1}}$$

- trying to convey that tie-ups notation should not obscure princip.

(*) entropy and free energy: Htree and Ftree:-

$$H_{tree} = H_b(X) = -\sum_X b(X) \ln b(X)$$

$$= -\sum_X \left(\frac{\prod_a b_a(\underline{x}_a)}{\prod_i b_i(x_i)^{(d_i-1)}}\right) \ln \left(\frac{\prod_a b_a(\underline{x}_a)}{\prod_i b_i(x_i)^{(d_i-1)}}\right)$$

$$= -\sum_X \left(\frac{\prod_a b_a(\underline{x}_a)}{\prod_i b_i(x_i)^{(d_i-1)}}\right) \left\{\sum_a \ln b(\underline{x}_a) - (d_i-1)\sum_i \ln b_i(x_i)\right\}$$

"match as"

$$= -\sum_X \left(\sum_a b_a(\underline{x}_a) \ln b(\underline{x}_a) - (d_i-1)\sum_i \left(\frac{1}{b_i(x_i)^{d_i-1}}\right) \ln b_i(x_i)\right\}$$

? ? p/S 4

- don't understand how to get this
- stackexchange

= → come back to this

(*) note that:- There is a decomposition of total entropy over a tree to a function of entropy of $b_a(\underline{x}_a)$ and $b_i(x_i)$

(*) There are some notes in Wainwright & Jordan (2008) Weller, Sontag et al. (...)

- But no explicit deriv. in:-
Yedidia (2001, 2002)
Koller (2009)
- Probably elementary.

18

(*) will have to take the results on <u>faith (not ideal)</u>

(*) $H_{tree} = -\sum_a \sum_{x_a} b(x_a) + \sum_i (d_i - 1) \sum_{x_i} b_i(x_i) \ln b(x_i)$

$$F_{tree} = -\sum_a \sum_{x_a} b_a(x_a) \ln \frac{b_a(x_a)}{f_a(x_a)} + \sum_i (1-d_i) \sum_{x_i} b_i(x_i) \ln b(x_i)$$

$$= (F_{12} + F_{23} + F_{34} + F_{15} + F_{56} + F_{67} + F_{78})$$

$$-(F_1 - F_2 - F_3 - F_5 - F_6 - F_7)$$

(*) sum of pairwise free energies - sum of singleton free energies.

(*) An excerpt from Wainwright & Jordan (2008):-
(illuminating):-

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

$$H(p_\mu) = -A^*(\mu) = E_\mu[-\log p_\mu(X)]$$

$$= -\underbrace{\sum_{s \in V} H_s(\mu_s)}_{(i)} - \underbrace{\sum_{(s,t) \in E} I_{st}(\mu_{st})}_{(ii)} \qquad (4.11).$$

(i) <u>Singleton entropy</u>:

- For each node $s \in V$:- $H_s(\mu_s) := -\sum_{x_s \in \mathcal{X}_s} \mu_s(x_s) \log \mu_s(x_s)$

(ii) <u>Mutual information</u>:-

- For each edge $(s,t) \in E$:

$$I_{st}(\mu_{st}) := \sum_{(x_s, x_t) \in \mathcal{X}_s \times \mathcal{X}_t} \mu_{st}(x_s, x_t) \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}$$

- For tree-structured graphs; dual A^* can be expressed as explicit and easily computable function of mean parameters μ.

(*) (G.1): with this in mind, the Bethe approx. to the entropy of an MRF with cycles is easily described.

(*) (G.1): It simply assumes that decomposition (4.11) is approx. valid for graphs with cycles.

(*) This assumption yields the Bethe entropy approximation :-

$$-A^*(\tau) \approx H_{Bethe}(\tau) := \sum_{s \in V} H_s(\tau_s) - \sum_{s,t \in E} I_{st}(\tau_{st}) \qquad (4.14)$$

(*) Note that comparing (4.11) and (4.14) $\longrightarrow$ replacement of exact marginals μ with τ for pseudomarginals

(G.1): Yedidia et al. used an alternative form of the Bethe entropy approx (4.14)

(*) that is obtained by relation:-

$$I_{st}(\tau_{st}) = H_s(\tau_s) + H_t(\tau_t) - H_{st}(\tau_{st})$$

where H_{st} is the joint entropy defined by the pseudomarginal τ_{st}.

(*) This and algebraic manipulation :-

$$H_{Bethe}(\tau) = -\sum_{s \in V} (d_s - 1) H_s(\tau_s) + \sum_{(s,t) \in E} H_{st}(\tau_{st}) \qquad (4.15)$$

where d_s corresponds to no. of neighbours of s. (degree).

(*) Bethe approx. to Gibbs free energy

- For general graphs, including non-tree structured GMs; the Bethe approximation, that is the Bethe approximation of the free energy (and entropy) uses the free-energy functional form which is derived for tree-structured GMs.

- we select: $\hat{F}(P,Q) = F_{Bethe}$

where:-

$$H_{Bethe} = -\sum_a \sum_{x_a} b_a(x_a) \ln b(x_a) + \sum_i (d_i - 1) \sum_{x_i} b_i(x_i) \ln b_i(x_i)$$

$$F_{Bethe} = -\sum_a \sum_{x_a} \ln \frac{b_a(x_a)}{f_a(x_a)} + \sum_i (1-d_i) \sum_{x_i} \ln b_i(x_i) = -\langle f_a(x_a) \rangle - H_{Bethe}$$

(57): equivalent to the Gibbs free energy when factor graph is a tree

- loopy graph:

$$F_{Bethe} = F_{12} + F_{23} + \dots + F_{67} + F_{78} - F_1 - F_5 \dots - 2F_2 - 2F_6 - F8$$

?

O/S 5 Not sure how to fill the gaps

(*) in general; $H_{Bethe} \neq H_{tree}$

(*) constrained minimisation of Bethe free energy

$$\min F_{Bethe}(b_i(x_i), b_a(x_a))$$

$$\text{s.t. } \sum_{x_i} b_i(x_i) = 1 \; ; \; \sum_{x_a \setminus x_i} b_a(x_a) = b_i(x_i)$$

- define a lagrangian:-

$$L(b_i(x_i), b_a(x_a), \lambda) = F_{Bethe} + \sum_i \gamma_i \left\{ \sum_{x_i} b_i(x_i) - 1 \right\}$$

$$+ \sum_a \sum_{i \in N(a)} \sum_{x_i} \lambda_{ai}(x_i) \left\{ \sum_{x_a \setminus x_i} b_a(x_a) - b_i(x_i) \right\}$$

- objective: F_{Bethe}
- constraints: local consistency constraints (normalisation; marginalisation).

$$\frac{\partial L}{\partial b_i(x_i)} = 0 \Rightarrow b_i(x_i) \propto \exp\left(\frac{1}{d_i - 1} \sum_{a \in N(i)} \lambda_{ai}(x_i) \right\}$$

$$\frac{\partial L}{\partial b_a(x_a)} = 0 \Rightarrow b_a(x_a) \propto \exp\left(-E_a(x_a) + \sum_{i \in N(a)} \lambda_{ai}(x_i) \right\}$$

Reparametrisation

- identify $\lambda_{ai}(x_i) = \log(M_{i \to a}(x_i))$

$$= \log \prod_{b \in N(i) \setminus a} M_{b \to i}(x_i)$$

- get B.P. equations:- $b_i(x_i) \propto f_i(x_i) \prod_{a \in N(i)} M_{a \to i}(x_i)$

$$b_a(X_a) \propto f_a(X_a) \prod_{i \in N(a)} \prod_{c \in N(i) \setminus a} M_{c \to i}(x_i)$$

O/S 6: Fill in gaps of deriv when you have time to conclude → some resources ready for you at hand.

(*) Some high-level key points from readings:-

Yedidia (2002, 2001):- Fixed points of the BP algorithm correspond to stationary points of the Bethe free energy

— " — :- Bethe approx, for which energy and entropy are approximated by <u>at most</u> pairs of nodes is the simplest form of the Kikuchi cluster variational method. (a general approximation).

(*) generalised belief propagation algorithms (GBP) minimise an <u>arbitrary</u> <u>Kikuchi free energy</u> (approx)

(of which Bethe is one of them).

(*) This viewpoint of generating tractable approx. to Gibbs free energy also motivates the use <u>variational</u> pespective on the MF approx. to free energy and entropy (functions of <u>one node beliefs</u>)

(*) This therefore establishes a connection/variational pespective on MF and Bethe approximations.