

U2 - Theory of V.I: inner and outer

- 1st half of lecture - U11b - MF and loopy BP

- Approximate inference (U11b)

- wide theoretical picture

Inference problems

Q: Are you familiar with these as inference problems.

- brief review of VE as message passing.

- turn global \rightarrow series of local computations for query.

(*) From elim to message passing

- explicitly shows intermediate (product) factors.

- elimination \rightarrow message passing on a clique tree.

- can combine messages in 2 directions; use these to compute statistics of interest.

- utility of message passing protocol \rightarrow gives info on comput. complexity.

- size of largest clique - info on bottleneck of inference algorithm

(*) Complexity

- use heuristics to find good elimination ordering

- e.g. eliminating from leaves/periphery better than elim from center.

(*) Sum-product algorithm

- on clique tree; we have a sum-product algorithm

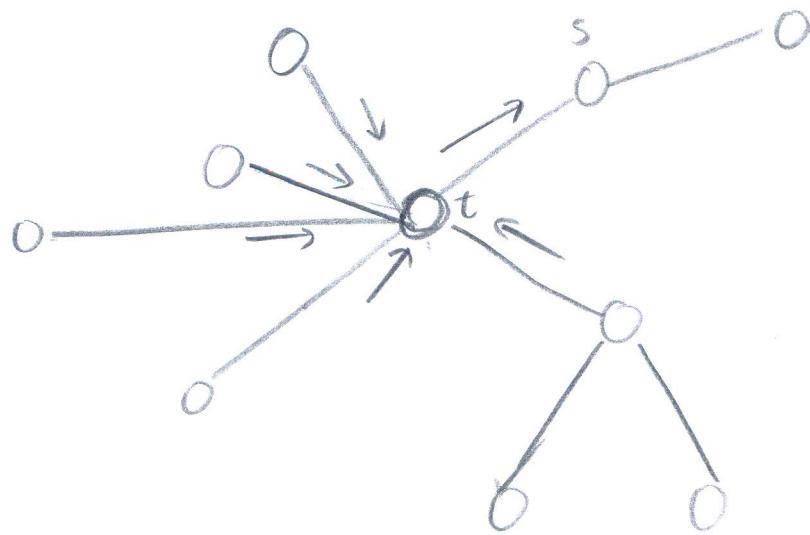
Q: tree structured GM:-

(*) daily

$$p(x_1, \dots, x_m) = \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t)$$

Q: message passing on trees:-

$$M_{t \rightarrow s}(x_s) \leftarrow \kappa \sum_{x'_t} \{ \psi_{st}(x_s, x'_t) \psi_t(x'_t) \prod_{u \in M(t) \setminus s} M_{u \rightarrow t}(x'_u) \}$$



(*) Generic form of message-passing

$M_{t \rightarrow s}(x_s)$ is message from $t \rightarrow s$ (outgoing)

- computed via multiplying factors at ~~at~~ t and containing t
memory $q(x'_i)$ $q(x_s, y'_i)$ Ⓐ need clarity here
- and all messages into t .
- After computing; it is set to s as an outgoing message

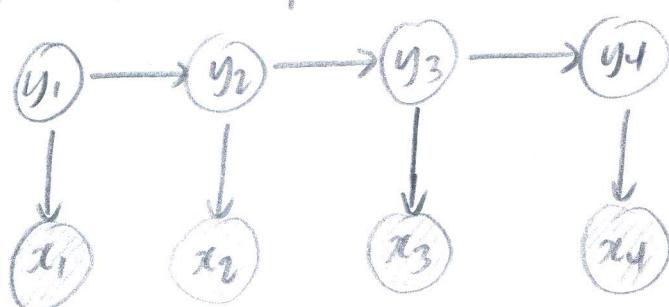
(*) Message passing protocol

- slides illustrate repeated execution of a single upward pass of message passing, with 1 root clique selected for each execution
- there exists a fixed convergence point for which relevant statistics can be computed.

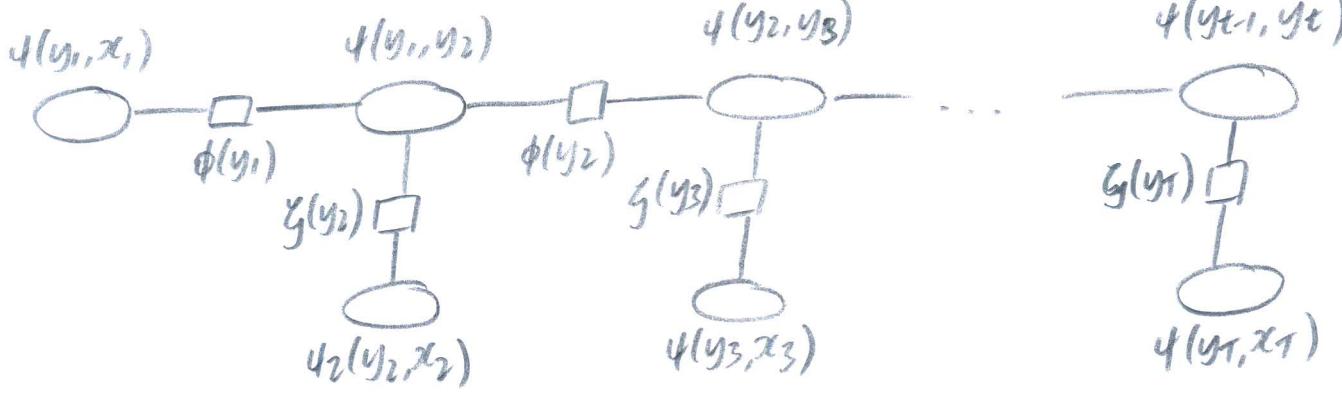
(*) Message passing for HMMs

- HMMs are trees.
- (*) Simply - write down cliques of HMMs (maximal?)
- list HMM; write equations (as important)

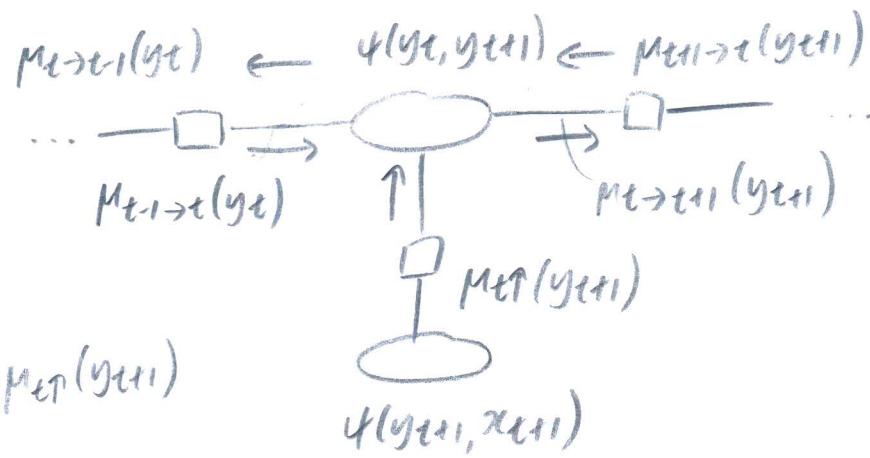
- x_i, y_i can be scalar or vector
(doesn't matter much)



(*) Unique tree for HMMs



- rightward pass:



$\mu_{t|t}(y_{t+1})$

$$= \sum_{y_t} \psi(y_t, y_{t+1}) \mu_{t-1|t}(y_t) \mu_{t|t}(y_{t+1})$$

$$= \sum_{y_t} p(y_t|y_t) \mu_{t-1|t}(y_t) p(x_{t+1}|y_{t+1})$$

$$= p(x_{t+1}|y_{t+1}) \sum_{y_t} \psi_{t|t}(y_t, y_{t+1}) \mu_{t-1|t}(y_t) \quad (\text{forward algorithm})$$

(A2) (2)
- fill

leftward pass

$$\mu_{t|t-1}(y_t) = \sum_{y_{t+1}} \psi(y_t, y_{t+1}) \mu_{t+1|t}(y_{t+1}) \mu_{t|t}(y_{t+1})$$

$$= \sum_{y_{t+1}} p(y_t|y_t) \mu_{t+1|t}(y_{t+1}) p(x_{t+1}|y_{t+1}) \quad (\text{backward algo})$$

- generic message passing applied to HMM tree structure

- messages are recursively computed using initial factor / pot. at node and measuring messages to that node.

(*) statistical meaning of new message - summed out intermediate factor

- not quite
- each is $p(x_{t+1}|y_{1:t})$, $p(x_{t+2}|y_{1:t+1})$ (?) (rightward pass)

(A3)
- need a tight
- undirected
- enforced
- vectorized

(*) These give exact answers to inference
for a query.

(*) Correctness of BP on a tree

- correctness: message passing based on V.E. it will yield correct answers
i.e. correct way to do V.E. gives you to a marginal
distn of subset of r.v.s given evidences

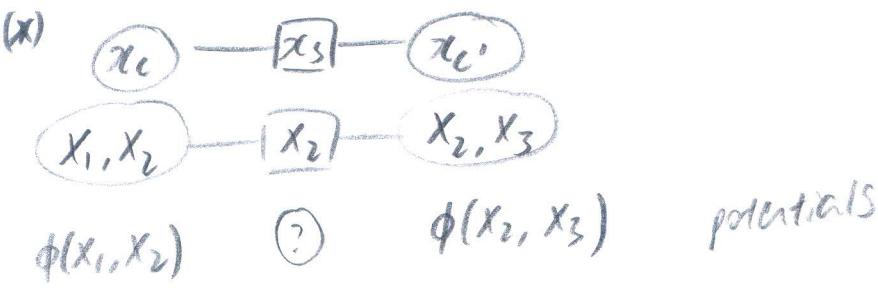
- message passing guarantees obtaining all marginals in tree: - (Theorem)

$$m_{ji}(x_i) = \sum_{x_j} \left(\psi(x_j) \psi(x_i, x_j) \prod_{k \in N(j) \setminus i} m_{kj}(x_j) \right)$$

- what is special about tree vs other graphs?

- uniqueness of messages (necessary cond. of correctness)
- one path through which messages are passed.

(*) Local \rightarrow Global consistency



- message passing property that can be proved by M.P.P. \rightarrow local consistency

(*) local consistency ensures that we have proper marginals

for a single marginal.

(*) And also consistency of pair of marginals; in the sense of coming

from the same distn. \rightarrow summing $\sum_{x_2} \phi(x_1, x_2) \Rightarrow \phi(x_1)$

(*) iron out precise details

(*) Junction tree: local \Rightarrow global consistency

as only one path
along

(i.e. correct marginal on every node of the tree)

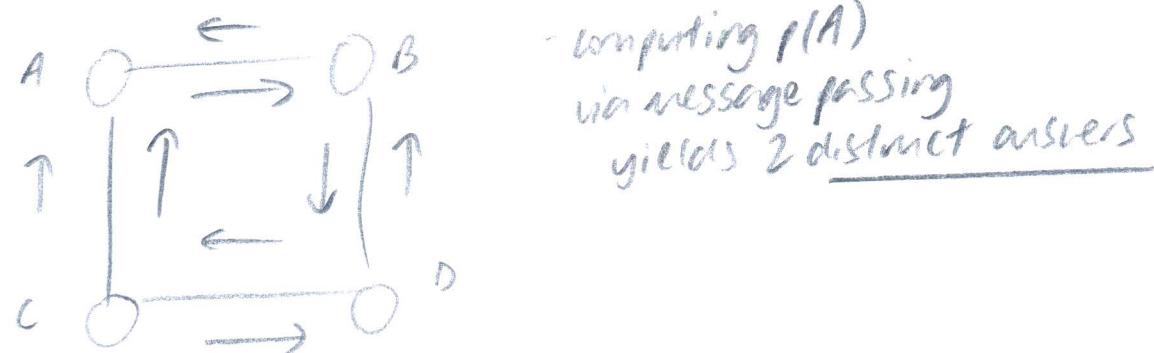
- every node/edge gets a single convergent point

of singleton and pairwise potentials

(*) Issues

- consider graph and clique tree

- consider 2 different ways of passing messages on this pathological graph



- computing $P(A)$
via message passing
yields 2 distinct answers

- occurs in 2 sides of clique tree

- no mechanism for ensuring consistency

Jordan (2003)

(*) In general, not true that L.C. \Rightarrow G.C. (non about if junction tree property is satisfied?)

(*) either:- i) An algo to turn graphical model \rightarrow junction tree (to be able to run M.P.)

ii) Approx. inference

- large tree width \rightarrow issues

(*) Why Approx. Inference

Q5 - Review, reinforce

- tree width is a determinant of cl.

- may be large

(*) What if graph is loopy?

- loopy graph \rightarrow non-tree structured graphs.

(*) Belief propagation on loopy graphs

- Precepts / wisdom on B.P. to sum-product message passing on trees is recent, contemporary result.

- Previously in 70s/80s, practitioners were not hindered by these insights.

- nondirectionality in loopy graphs; 'roots' can also be 'leaves'

- message passing due to this loopiness / cyclicity can continue indefinitely

in contrast to calibration/convergence of beliefs after 2 passes on a tree.

Practitioners discovered:-

that running this indefinitely either:-

- 1) converged \rightarrow they checked solutions via brute force marginalisation to get ground truth; compared infinite message passing results.
- so sometimes solutions were good.
- 2) did not converge \rightarrow oscillates

(*) Loopy Belief Propagation

- loopy BP is a 'black'; no M.P.P. on loopy graph regardless of insights regarding convergence.

(*) see loopy Belief Propagation, Murphy, Weiss, Jordan (1999)

- Kevin Murphy (aka in textbook) just experimented with M.P.P. (20 years ago) on non-tree structures.
- An empirical paper that inspired theoreticians

(*) Beliefs and messages in factored graphs

$$b_i(x_i) \propto f_i(x_i) \prod_{a \in N(i)} m_{a \rightarrow i}(x_i)$$

$$m_{i \rightarrow a}(x_i) = \prod_{c \in N(i) \setminus a} m_{c \rightarrow i}(x_i)$$

$$b_a(x_a) \propto f_a(x_a) \prod_{i \in N(a)} m_{i \rightarrow a}(x_i) \quad m_{a \rightarrow i}(x_i) = \sum_{x_a \setminus x_i} f_a(x_a) \prod_{j \in N(a) \setminus i} m_{j \rightarrow a}(x_j)$$

- classify messages into 2 classes :-

- i) node incoming
- ii) factor outgoing

- define factors (i.e. potential fns) on graphical model and nodes explicitly:-

- messages $a \rightarrow i$: factor to r.v.

- messages $i \rightarrow a$: r.v. to factor

- once messages are converging; compute singleton marginals and cluster marginals using a formula. (i.e. $b(x_i)$ and $b(x_A)$)
- generic formula is just M.P.P. with product of either messages or factors.
- combine together \rightarrow get a more global def. of message passing

Approx Inference

- spend time on what is going on in loopy B.P.

Ex: goal of inference: - $p(x_A)$ i.e. marginal prob over a subset of r.v.s.

- use above equations; but why is this approx. inference?

- That is because we know empirically and theoretically that running M.P. on loopy graph / non-tree structured \Rightarrow convergence not guaranteed.

⑥ what is being approximated?

and how good is the approximation?

- Previously; we saw an example of needing to approximate an intract. distri. (V.I.)

$$\text{- actual distri: } p(x) = \frac{1}{Z} \prod_{a \in F} p_a(x_a) \quad (\text{or } p(x_A) = \sum_{x|A} p(x))$$

$$\begin{aligned} &\text{- find a } Q(x) \\ &: p(x_A) \geq \sum_{x|A} Q(x) \end{aligned}$$

(*) wish to find distri Q

such that Q is a 'good' approx to P .

(*) refine KL-divergence \rightarrow get a free energy term

⑥ KL-diver-bound-
free energy
- tighten.

(*) lower bound on log. likelihood \rightarrow -ve free energy

(*) free energy is surrogate of an objective we can optimise to find a Q that is close to P .

(*) recall MF approx in topic models \rightarrow define Q to be a product of marginals

(which are param by variational params)

$$Q = \prod q(x_a; \phi)$$

- suppose we have convergence of approximation Q to true P .

- then solve inference problem $p(x_A)$ directly by specifying $q(x_A)$

- A little unclear here ⑥ -

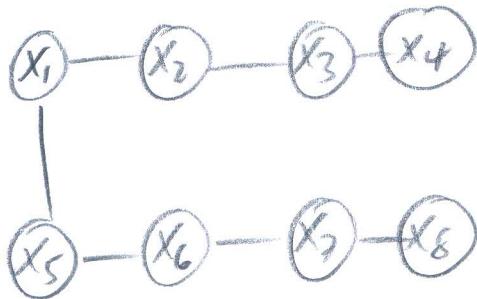
(*) at a high-level; defining q appropriately precludes necessity for further marginalis.

- way to get optimised q^* \rightarrow minimise KL

(*) free energy functionals

- use a different technique to above

- see how free energy term appears on the graph in slides



- local marginals of interest:- $p(x_i)$; $p(x_i, x_j)$
singleton adjacent nodes/pairwise.

- what is relation between:-

$p(x_1, \dots, x_8)$ and $p(x_i), p(x_i, x_j)$?

$$p(x_1, \dots, x_8) = \\ (\text{surely}) \quad = p(x_7, x_8) p(x_6|x_7) p(x_5|x_6) p(x_4|x_5) \dots p(x_1|x_3)$$

$$\begin{matrix} \text{via chain} \\ \text{rule, MB} \\ \text{prop.} \end{matrix} = p(x_7, x_8) \frac{p(x_6, x_7)}{p(x_7)} \frac{p(x_5, x_6)}{p(x_6)} \dots \frac{p(x_4, x_3)}{p(x_3)}$$

(*) i.e. use chain rule to rewrite conditionals in terms of joint, Marg.

- x_4, x_8 do not appear on bottom

- adjust using $(d_i - 1)$

$$= \frac{\prod p(x_i, x_j)}{\left\{ \prod p(x_i)^{d_i - 1} \right\}}$$

- d_i - degree of every node on graph.

- All nodes except x_4, x_8 have a degree of 2 $\Rightarrow (d_i - 1) = 1$

- nodes x_4, x_8 have a degree of 1 $\Rightarrow b_{i-1} = 0$
 - hence their marginals vanish from denom.
 - (*) provides a useful way of defining joint distri on trees
 - (*) product of pairwise marginals (Bethe)
 - (Product of singleton) $\text{degree} - 1$
- $$b(\underline{x}) = \prod_a b_a(x_a) \prod_i b_i(x_i)^{1-d_i}$$
- Ⓐ review of Htree, Ftree. i
- review
- (*) result with sum of pairwise free energies subtract singleton free energies.

(*) Bethe approx. to Gibbs free energy

- copy graph \rightarrow how to define free energy, entropy on this?
- review the mathematics here.

$$p(x_1, \dots, x_d) \neq \frac{\prod_a b_a(x_a)}{\prod_i b_i(x_i)^{d_i-1}} \quad ; \text{ however}$$

- (*) we can define Bethe free energy F_{Bethe} as an approx/surrogate of 'true' free energy of loopy graph (more robust from this see review) unknown due to intractability of summation (or int.)
- choose $\hat{F}(P, Q) = F_{\text{Bethe}}$ $F_{\text{true}}(P, Q)$ (loopy graph)

$$F_{\text{Bethe}} = -\langle f_a(x_a) \rangle - H_{\text{Bethe}}$$

- treat $b_a(x_a)$ and $b_i(x_i)$ as unknown r.v.s.; if known can define F_{true}

③ - A little unclear here \rightarrow review.

- (*) high-level: solve for singleton and pairwise marginals

$b_a(x_a), b_i(x_i)$ via minimising the Bethe free energy

- (*) Assuming this works due to Bethe free energy being 'similar' to true free energy

- (*) Bethe free energy approx. can be defined for any G.M.

(*) conduct constrained minimisation of Bethe free energy

(a function of singleton and pairwise marginals)

by imposing normalisation constraints on these + local consistency

(*) (Q) :- $\min F_{\text{Bethe}}(b_i(x_i), b_A(x_A)) \text{ s.t. } \sum b_i(x_i) = 1$

Σ

Ⓐ
① - complete review.

(*) Review material here. that leads to

(*) Handcraft optimisation problem n. sol to mag. queries; use loss

function that is Bethe free energy (correct when graph is tree
incorrect — → is not tree)

- directly solve constrained opt (lagrange) with normal. constraints,
or local consistency constraints

(*) solve for $b_i(x_i)^*$ and $b_A(x_A)^*$

(*) Reparametrise λ

- use local consistency constraints between cluster and node

- define λ as 'message'; 'rotate'

(*) Then sol. to constrained opt. takes same form as M.P.P. on loopy graph.

(*) It was a novel insight that an ad-hoc app. of M.P.P. out of context
to loopy graphs could be so contextualised theoretically.
- i.e. constrained optim. with B.F.E.

(*) theoretical paper by physics researchers

- connection between optim.; message passing

(*) connectedness

of MF, GBP → inference → optim.