10-708 22/06/2020

L12 - Theory of V.I : Inner and Outer

- 1st half of lecture - L11b - MF and loopy BP

- Approximate inference (L11b)
- wider theoretical picture

Inference problems

(W): Are you familiar with these as inference problems

- Brief review of VE as message passing
- Turn global → series of local computations for query.

(*) From elim to message passing

- Explicitly shows intermediate (product) factors
- Elimination → message passing on a clique tree.
- can conceive of messages in 2 directions; use these to compute statistics of interest
- Utility of message passing protocol → 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

- (i.e. clique tree), we have a sum-product algorithm

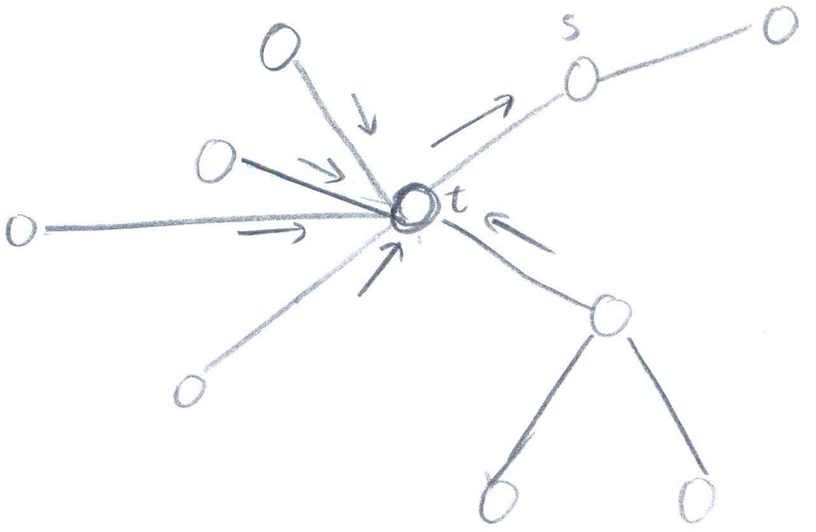
(PM): tree structured GM:-

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

(PM) clarity

(GM) Message passing on trees:-

$$M_{t \to s}(x_s) \leftarrow K \sum_{x_t'} \{ \psi_{st}(x_s, x_t') \psi_t(x_t') \prod_{u \in N(t) \setminus s} M_{u \to t}(x_t') \}$$



(*) Generic form of message-passing

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

- computed via multiplying factors at t and containing t $\psi(x_t')$ $\psi(x_s, x_t')$ - need clarity here (41)

and all incoming messages into t.

- After computing; it is sent to s as an outgoing message

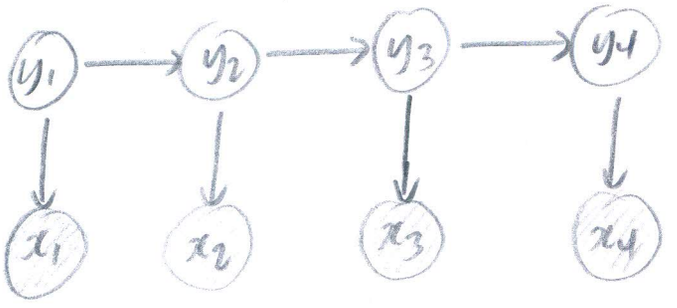
(*) Message passing protocol

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

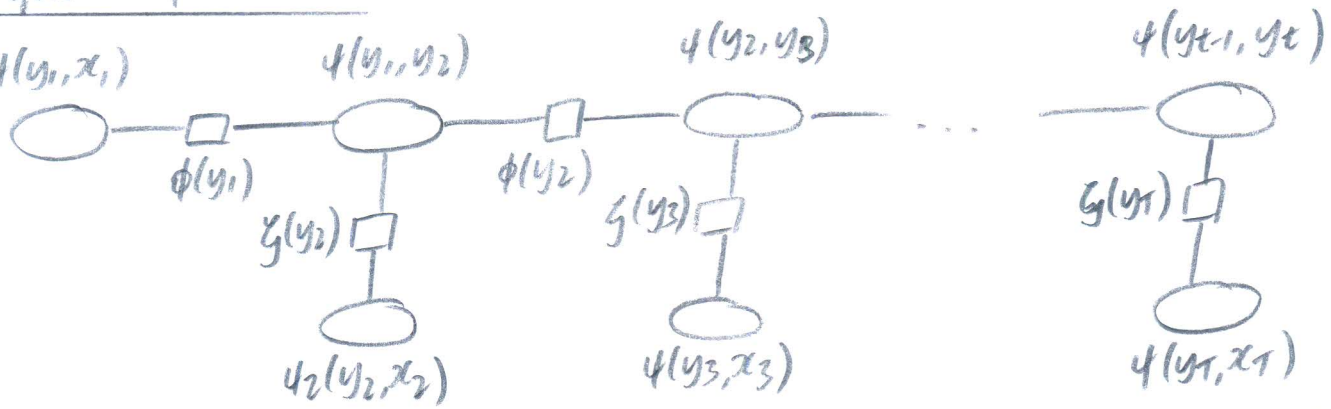
(*) Message passing for HMMs

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

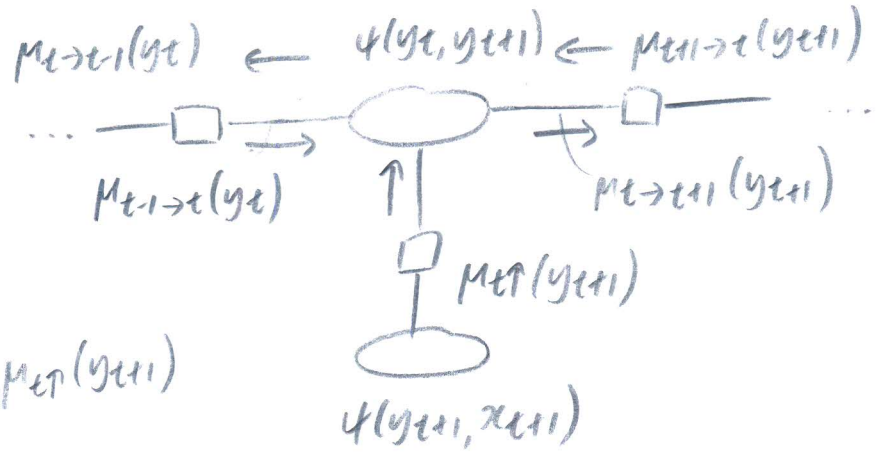
- x_t, y_t can be scalar or vector (doesn't matter much)



(*) Chain tree for HMMs



- rightward pass:



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

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

$= \sum_{y_t} p(y_{t+1}|y_t) \mu_{t-1 \to t}(y_t) p(x_{t+1}|y_{t+1})$ (A2) (2) - fill

$= p(x_{t+1}|y_{t+1}) \sum_{(?)} a_{y_t, y_{t+1}} \mu_{t-1 \to t}(y_t)$ (forward algorithm)

Upward pass

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

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

- Generic message passing applied to HMM tree structure
- messages are recursively computed using initial factor/pot. at node and incoming 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 tighter understanding of forward - backward.
-

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

(*) Correctness of BP on a tree

- correctness: message passing based on V.E. & 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

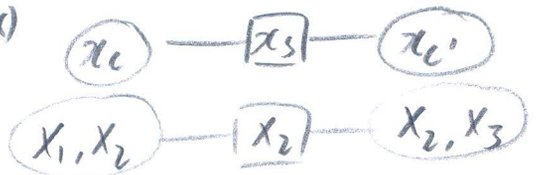
- message passing guarantees obtaining all marginals on tree:- (theorem)

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

- what is special about tree vs other graphs?
  - uniqueness of messages (necessary cond. of correctness)
  - one path through which messages are passed.

(*) Local → global consistency

(*)



$\phi(X_1, X_2)$ (?) $\phi(X_2, X_3)$ potentials

- message passing property that can be proved by M.P.P. → 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 distr. → summing $\sum_{x_2} \phi(x_1, x_2) \Rightarrow \phi(x_1)$

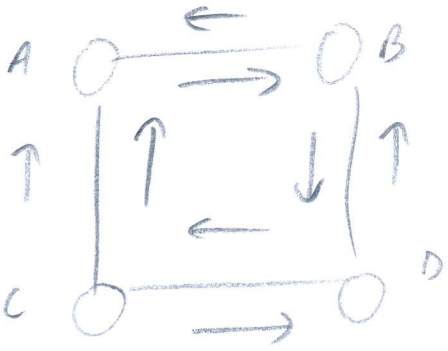
(14) - iron out precise details

(*) Junction tree: local ⇒ global consistency (as only one path along)
(i.e. correct marginal on every node on the tree)

- every node/edge gets a single contingent 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

- C occurs in 2 sides of clique tree
- no mechanism for ensuring consistency

Jordan (2003)

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

(*) Either:- i) Analogy to turn graphical model $\longrightarrow$ junction tree (to make to run M.P.)

ii) Approx. inference - large tree width $\rightarrow$ issues

(*) Why Approx. Inference

(AS) - Review, reinforce

- tree width is a determinant of cc.
- may be large

(*) What if graph is loopy?

- loopy graph $\longrightarrow$ 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
- no directionality 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 ⟶ they checked solutions via brute force marginalization to get ground truth, compared multiple message passing results
   - so sometimes solutions were good.

2) Did not converge ⟶ oscillates

(*) Loopy Belief Propagation

- Loopy BP is a 'hack' ; run M.P.P on loopy graph regardless of weights regarding convergence.
(*) see Loopy Belief Propagation, Murphy, Weiss, Jordan (1999)
- Kevin Murphy (aka my 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

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

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

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

- classify messages into 2 classes :-
  i) node incoming
  ii) factor outgoing
- define factors (i.e. potential fns) on graphical model and nodes explicitly :-
- messages $a \to i$: factor to r.v.
- messages $i \to a$: 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 → 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 ⇒ convergence not guaranteed.

(SM) what is being approximated?
and how good is the approximation?

- Previously; we saw an example of needing to approximate an intract. distri.
  (v.i.)
- Actual distri:- $p(X) = \frac{1}{Z} \prod_{f_a \in F} f_a(x_a)$ (or $p(x_a) = \sum_{X \setminus A} p(X)$)

(*) Wish to find distri Q
such that Q is a 'good' approx to P.

- find a $Q(x)$: $p(x_A) \approx \sum_{X \setminus A} Q(x)$)

(*) refine KL-divergence → get a free energy term (A6) KL-lower-bound-free energy - tighter.

(*) lower bound on log-likelihood → -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 → define Q to be a product of marginals (which are param by variational params)

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

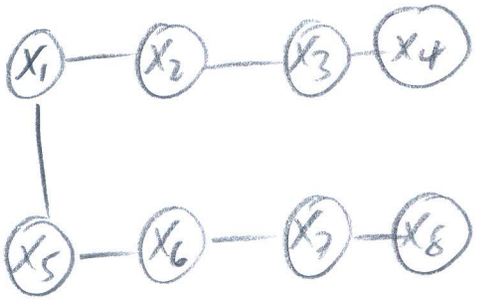
- Suppose we have convergence of approximation Q to true P.
- Then solve inference problem $p(x_A)$ directly by specifying $q(x_A)$
- A little unclear here (A6) -

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

- way to get optimised q^* → minimise KL

(*) Tree Energy Functionals

- use a different technique to above
- see how free energy term appears in the graph in slides



- local maginals 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 $p(x_i), p(x_i, x_j)$?

$p(x_1, \dots, x_8) =$

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

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

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

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

- x_4, x_8 do not appear in denom
- adjust using $(d_i - 1)$

- d_i - degree of every node in 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 $1 \Rightarrow (d_i - 1) = 0$
- Hence their marginals vanish from denom.

(*) provides a useful way of defining joint distri on trees

(*) Product of pairwise marginals / (Product of singleton)$^{\text{degree}-1}$ (Bethe)

$$b(\underline{x}) = \prod_a b_a(\underline{x}_a) \prod_i b_i(x_i)^{1-d_i}$$

(A7) pair of Htree, Ftree. ; -review

- (*) Result with sum of pairwise free energies subtract singleton free energies.

(*) Bethe approx. to Gibbs free energy

- loopy graph → how to define free energy, entropy or this?
- Review the mathematics here.

$$p(x_1, \ldots, x_8) \neq \frac{\prod_a b_a(\underline{x}_a)}{\prod_i b_i(x_i)^{d_i - 1}} \quad ; \text{however}$$

(*) we can define Bethe free energy Fbethe as an approx/surrogate of 'true' free energy of loopy graph (more naive than this, see rev notes) unknown due to intractability of summation (eq on f.) Ftrue(P, Q) (loopy graph)

- Choose $\hat{F}(P, Q) = F_{Bethe}$

$$F_{bethe} = -\langle f_a(\underline{x}_A) \rangle - H_{bethe}$$

- Treat $b_a(\underline{x}_A)$ and $b_i(\underline{x}_i)$ as unknown r.vs.; if known can define Fbethe

(?) - A little unclear here → review.

(*) High-level: solve for singleton and pairwise maginals $b_a(\underline{x}_A), b_i(\underline{x}_i)$ 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

(*) (A7) :- $\min F_{Bethe}(b_i(x_i), b_A(x_A))$ s.t. $\sum b_i(x_i) = 1$
Σ

(A7) (:) - complete/ review.

(*) Review material here. that leads to
(*) Hardcore optimisation problem ∩ sol to marg. queries; use loss function that is Bethe free energy (correct when graph is tree
nearest — " — 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

(*) Commonality
of MF, LBP → inference → optim.

10-708 - L12 22/06/2020

L12 - Theory of V.I - Inner and Outer

- 2nd half of (L12) covers this
- Theory of V.I. (via convex opt.)
- Relation of loopy BP and mean-field approx.
- convex optimisation will be used → conjugate duality

(*) Variational inference

(*) variat. → represent quantity of interest as solution to an optim. problem.
↳ approx. desired sol. by relaxing/approx. intractable opt. problem.

- use local, incremental, iterative methods on vari. problem.

(*) Inf. in GMs

Q: How to represent quantities of interest in a variational form

(*) Exponential Families

- Express joint distn as exp. family
- canonical parm:

$$p_\theta(x_1, \ldots, x_m) = \exp\{\theta^T \phi(x) - A(\theta)\}$$

- log-norm. const.

$$A(\theta) = \log \int \exp\{\theta^T \phi(x)\} dx$$

- Effective can. parm

$$\Omega := \{\theta \in \mathbb{R}^d \mid A(\theta) < +\infty\}$$

(*) G.M. as Exp Fam

- UGM/MRF: $p(x; \theta) = \frac{1}{Z(\theta)} \prod_{c \in C} \psi(x_c; \theta_c)$

(*) MRF (exp. form) :- $p(x;\theta) = \exp\left\{\sum_{c \in C} \log \psi_c(x_c;\theta_c) - \log Z(\theta)\right\}$

(*) Review examples (41)

(*) Why exp. families?

- nice properties intrinsic to exp. family.

(I) :- expect. of suff. statistics (mean param) given canonical param
yields marginals → query of interest.

(II) - compute of norm constant → part. function.

(*) computing mean param.

- looks like connection between inference and param estim.
  - computing mean param ⇔ compute of singleton (or other) marginals.

(*) classical approach → summation/marg.

(*) undesirable → expectation step is expensive with many r.vs.
compute.

- want incremental/gradient based approach.

(*) Conjugate duality

- can always define conjugate dual function for a function.

conjugate dual:

- Given any $f(\theta)$, conjugate dual :-

$$f^*(\mu) = \sup_{\theta} \{\langle \theta, \mu \rangle - f(\theta)\}$$

$\langle \theta, \mu \rangle$ ~ inn. prod. dual pair.

$f^*(\mu)$ - conjugate dual
- convex as pointwise supremum of class of lin functions (↑)

(42) - Geometric explanation.

(Q) For which $\mu \in \mathbb{R}^d$ do we have sol. $\theta(\mu)$?

(*) comp. of conjugate dual

- assume above i.e. $\theta(\mu)$: $\mu = E_{\theta(\mu)}[\phi(x)]$
- conjugate dual of log-partition is entropy of orig. distri. when above holds (μ in feasible space).

(*) Complexity of conj dual

- start with mean param.
- If 'exists', use stat distri. to mk canonical param.
- If 'exist.', use canonical param to define entropy fn. of orig distri
- Yields conjugate dual of log-part.

(*) Strategically, and in abstract, seems straightforward.
- But tractability wise, there are issues here

(*) Difficulty of intractability in summation reflected also in this transposed problem.

(Q): For which $\mu \in \mathbb{R}^d$ does it have sol. $\theta(\mu)$

(I) find <u>domain</u> of μ such that $\theta(\mu)$ <u>exists</u>

(II) once we know this <u>domain</u>, we can approx. of solving
e.g. exp. of marginals / log-partit.
given this domain in an <u>explicit way</u>.

(*) e.g. amend domain / loss funct. for tractability.

(*) <u>marginal polytope</u>

- the domain of μ (where μ is defined by exp. of sufficient statistic) is the <u>marginal polytope</u>

(*) Dual of dual is orig. (A3) - details

- under technical cond. satisfied by smooth functions
- can use conjugate duality twice to recover orig. function.

- for log-part. function:-
  - can define it as sol. to dual of dual (rather than sim.)

$$A(\theta) = \sup_{\mu} \{ \langle \theta, \mu \rangle - A^*(\mu) \} \quad \theta \in \Omega$$

- dual variable μ → interp as mean param.

(*) Comp. mean param. Bernoulli

- obtain stationary cond → $\mu = \frac{e^\theta}{1+e^\theta}$ $\mu = \nabla A(\theta)$
- consider range of μ i.e. find limit as $e^\theta \to \infty$, $e^\theta \to 0$
- that is consider $\mu \in [0,1]$ (well-behaved)
  $\mu \notin [0,1]$ (pathological)
- well behaved ⇒ correspondence between μ and θ valid
  - infer θ as function of μ; compute conjugate

(*) eventually have an explicit form on conjugate dual $A^*(\mu)$

(*) Then can define variational form as sol. to optimisation over the dual of the dual (!)

(A4) - Review example - crucial, simple illus. of variational methods and conjugate duality (interp. of V.I.)

(*) Comp. of conjugate dual

- Examine more generic exp. family distri.
- Generalisation of earlier Bern example.

(*) stationary cond ⟶ $\mu = \mathbb{E}_\theta[\phi(X)]$

(*) see the convex set of all realisab. mean params
- marginal polytope

(*) why is it convex?
- convex comb. avg. → convex hull

(*) marginal polytope is convex comb. of extreme points of sufficient statistic functions.

(*) convex polytope (A4) - overse - review

- more explicit way of specifying marginal polytope (convex hull rep.)

(*) end part of slide → half-space rep.

(*) 2-node Ising example

(A5) - Review example to grow abstr. concepts.

(A7): every point within convex hull (half-plane (corresp to a realisable μ) ⟶ yields valid θ.
- marginal polytope

(*) marginal polytope for general graphs

- difficult, complicated concept
- marginal polytope is difficult to characterise for general graphs.
- next lecture → we find out how to approx. this and also obj fn (conjugate of log part.) in order to directly solve optimisation and get marginals of data. (likelihoods)