10-708 23/06/2020

U2-review

(*) core focus areas:-
i) loopy B.P. + Bethe approximation (U11a)
ii) conjugate duality, marginal polytope, convex polytope (U11b)

U11a - 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

(0/5/1) → 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 variable sm 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 ⇏ 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 MGM
- more specifically, an MN/UGM with N nodes, pairwise potentials (Yedidia 2002, 2001)
- (0152) An ambiguity (mild) → slides specify pairwise & singleton potentials.
- (Yedidia 2001, 2002) specify that the singleton potentials are local evidence nodes
- we have:-
Joint prob (Gibbs distri (?) of pairwise MRF:

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

- BP updates:-

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

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

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

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

- α - normalisation constant
- $N(i) \setminus j$ - 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 $\prod_{k \in N(i) \setminus j}$)

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

(*): 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)
- Recd 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 eg. 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

(DS 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

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

X - entire space of inputs (not. abuse)

(*) 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:-

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

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

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

"match as"

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

(2) (7) 0/54

- 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 $b_a(x_a)$ 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.

(X)

(*) will have to take the results on faith (not ideal)

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

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

$$= (F_{12} + F_{23} + F_{34} + F_{15} + F_{56} + F_{67} + F_{28}) - (F_1 - F_2 - F_3 - F_5 - F_6 - F_7)$$

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

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

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

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

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

(i) Singleton entropy:

- for each node $s \in V$:- $H_s(\mu_s) := -\sum_{x_s \in X_s} \mu_s(x_s) \log \mu_s(x_s)$

(ii) Mutual information:-

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

$$I_{st}(\mu_{st}) := \sum_{(x_s, x_t) \in X_s \times X_t} \mu_{st}(x_s, x_t) \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}$$

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

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

(*)(69): 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

(69): 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 pseudomagnal 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:-

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

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

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

- Loopy graph:

$$F_{Bethe} = F_{12} + F_{23} + \ldots + F_{67} + F_{78} - F_1 - F_5 \ldots - 2F_2 - 2F_6 - F_8$$

(?)

(OIS 5) not sure how to fill in 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:- $n_i(x_i) \propto f_i(x_i) \prod_{a \in N(i)} M_{a\to i}(x_i)$

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

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 minimise an <u>arbitrary Kikuchi free energy</u> (approx) (GBP)

(of which Bethe is one of them).

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

(*) this therefore establishes a connection/variational perspective on MF and Bethe approximations.

10-100 [illegible]

(12) review

L12 - Theory of V.I: Inner route approx.

- Resources: - Murphy (2012) ch 22
  - Wainwright, Jordan (2003)
  - ——— " ——— (2008) (*) → focus on this (ref. point).

(*) 3.1 Exponential Representations via Maximum Entropy

(not covered in lectures; but seems import.)

- Principle of maximum entropy (M.E) to motivate G.MS.
- n IID observations $X^1, \ldots, X^N$

empirical exp: of rob function

$$\hat{\mu}_\alpha := \frac{1}{N} \sum_{i=1}^{N} \phi_\alpha(X^i) \qquad \forall \alpha \in \mathcal{I} \qquad (3.1)$$

$\mathcal{I}$ - index set. indexing

$\phi_\alpha : \mathcal{X} \to \mathbb{R}$

- Stack $\hat{\mu}_\alpha$ into a vector
- Giving $|\mathcal{I}|$-dim vector of emp. exp. $\hat{\mu}$
- using $\hat{\mu}$, goal is to infer full prob. distn over r.v. X.
- Represent prob. distn as density p, absolutely cont. wrt measure ν
- Consider expectations: - (ass. existence)

$$\mathbb{E}_p[\phi_\alpha(X)] := \hat{\mu}_\alpha \qquad \forall \alpha \in \mathcal{I}$$

(*) Expectations under distn p matched to exp. under empirical distn

(*) Underdetermined → many distns p that are consistent with observations

(*) Require principle for selection: -

(*) define functional of density p, i.e. Shannon entropy:

$$H(p) := \int_{\mathcal{X}} p(x) \log p(x) \, \nu(dx) \qquad (3.2)$$

(*) Principle of max entropy:

- From among distns consistent with data, select p^* whose Shannon entropy is maximal

(*) formally;

- $\mathcal{P}$ - set of all probability distrs over r.v. X
- maximum entropy sol p^* solves:-

$$p^* := \underset{p \in \mathcal{P}}{\operatorname{argmax}}\ H(p) \quad \text{s.t.} \quad \mathbb{E}_p[\phi_\alpha(X)] = \hat{\mu}_\alpha \quad \forall\, \alpha \in I \qquad (3.3)$$

(*) interpretation:- select distri with maximal uncertainty, measured by entropy functional, while remaining faithful to data.

(*) can show that optimal p^* takes following form:-

$$p_\theta(x) \propto \exp\left\{ \sum_{\alpha \in I} \theta_\alpha \phi_\alpha(x) \right\} \qquad (3.4)$$

where $\theta \in \mathbb{R}^d$ represents parametrisation of distrn exponential family form

(*) ME perspective:- parameters θ have interpretations as lagrange multip. associated with constraints specified by empirical moments $\hat{\mu}$.

3.2 Basics of Exp Families

(*) Exponential family ⟶ parametrised family of densities taken with respect to some underlying measure ✓.

(*) Random vector:- $(x_1, x_2, \dots, x_m) \in \mathcal{X}^m$ $\quad \mathcal{X}^m = \bigotimes_{s=1}^m \mathcal{X}_s$ (cartesian product)

(*) collection of potential functions $\quad \phi = (\phi_\alpha, \alpha \in I) \quad \phi_\alpha : \mathcal{X}^m \to \mathbb{R}$

(*) ϕ_α - potential function / sufficient statistic

(*) I - index set $d = |I|$ elements

(*) ϕ - vector valued mapping from $\mathcal{X}^m \to \mathbb{R}^d$

(*) ϕ - vector of sufficient statistics

(*) Let $\theta = (\theta_\alpha, \alpha \in I)$ - vector of canonical/exp. params

(*) For $x \in \mathcal{X}^m$, denote:-

$\langle \theta, \phi(x) \rangle$ - Euclidean inner product on $\mathbb{R}^d$ of two vectors θ and $\phi(x)$

(*) Exponential family associated with ϕ :- parameterised coll of densities:- wrt $d\nu$

$$p_\theta(x_1, \dots, x_m) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\} \tag{3.5}$$

(*) Log partition/cumulant:- $A(\theta) = \log \int_{\mathcal{X}^m} \exp\langle \theta, \phi(x) \rangle \, \nu \, dx$ (3.6)

$$= \log \int \exp \theta^T \phi(x) \, dx$$

(*) p_θ is properly norm.: $\int_{\mathcal{X}^m} p_\theta(x) \, \nu \, dx = \int p_\theta(x) \, dx = 1$

(*) Set of potentials ϕ fixed; each param vec θ indexes a particular member p_θ of family.

(*) Canonical param of interest, θ belong to set:

$$\Omega := \{\theta \in \mathbb{R}^d \mid A(\theta) < +\infty\} \tag{3.7}$$

(*) A is a convex function of $\theta \Rightarrow \Omega$ is a convex set.

(*) Log-partition function A - prominent

use Wai (2003) nice as next section is technical.

Wainwright, Jordan (2003) (2.2)

restrict attention to case where Ω is open $\longrightarrow$ regular exponential family

Minimal exp family :- no affine combo of potential functions $\phi = \{\phi_\alpha, \alpha \in I\}$ that is equal to a constant

(**) :- one-to-one corresp. between exponential/canon param θ and distributions $p(x; \theta)$

- log-part. function strictly convex on θ.

(*) Overcomplete exp. family: there exist linear combinations $(a, \phi(x))$ equal to a constant

- there exists an affine subset of parameter vectors Θ, each associated with same distri.
- identifiability of param vector θ lost

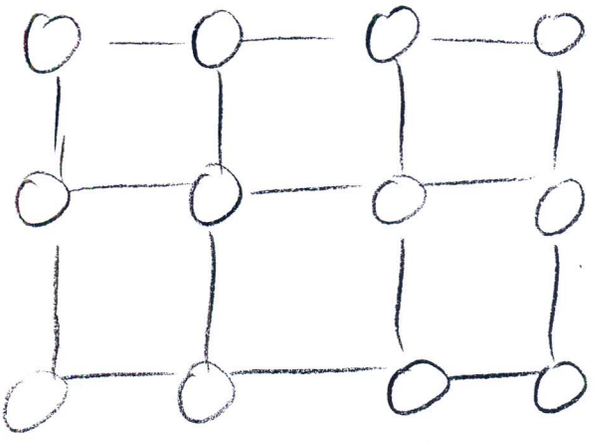
(*) examples:-

Ising model

$G = (V, E)$ nodes

$x_s \in \{-1, +1\}$ $x_s \sim \text{Ber}(p)$ $s \in V$

- x_s and x_t nodes of full random vector only interact if s and t are joined by geoges



$$p_\theta(x) = \exp\left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta) \right\} \quad (3.8)$$

(*) $\theta_{st} \in \mathbb{R}$ - strength of edge (s,t)

$\theta_s \in \mathbb{R}$ - potential for node s.

(*) Ising index set $I = V \cup E$ $d = m + |E|$

(*) log-part:-

$$A(\theta) = \log \sum_{x \in \{0,1\}^m} \exp\left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \right\} \quad (3.9)$$

(*) $\sum$ is finite for all choices of $\theta \in \mathbb{R}^d$; domain Ω is full space $\mathbb{R}^d$, family regular

(*) generalisations - higher order interactions e.g. k-cliques

$k = m \rightarrow$ extreme, connect k_2 all nodes; represent any distri over a binary random vector.

(*) Example 3.2

Potts model: (gen. of Ising) — r.v x_s at each node $s \in V$; values in discrete space $\mathcal{X} := \{0, \ldots, r-1\}$ $r > 2$ integer.

- state $j \in \mathcal{X}$ as a label (e.g. image seg.)
- Pairing of node $s \in V$, state $j \in \mathcal{X}$:-

Sufficient statistic: $$\mathbb{I}_{s;j}(x_s) = \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{otherwise} \end{cases} \qquad (3.10)$$

Associated canon. param: $\theta_s = \{\theta_{s;j}, j = 0, 1, \ldots, r-1\}$

Sufficient statistic: $$I_{st;jk}(x_s, x_t) = \begin{cases} 1 \text{ if } x_s = j \;\; x_t = k \\ 0 \quad \text{otherwise} \end{cases} \qquad (3.11)$$

Assoc. param: $\theta_{st;jk} \in \mathbb{R}$

*) Chosen suff stats :- exp family $d = r|V| + r^2|E|$ (?)

(?) overcomplete/regular?

(*) Example 3.3

Gaussian MRF: • undirected G with vertex set $V = \{1, \ldots, m\}$.

• MVG rvec $(X_1, \ldots, X_m)$ respects Markov prop. of G

represented in exponential form using all sufficient statistics :-

$(x_s, x_s^2, s \in V, x_s x_t, (s,t) \in E)$

m-vector θ of param: associated with m-vector of suff stats

$x = (x_1, \ldots, x_m)$

symmetric matrix $\Theta \in \mathbb{R}^{m \times m}$ assoc. with matrix xx^T

Θ - negative of inverse covariance/precision matrix

$\Theta_{st} = 0$ if $(s,t) \notin E$

$d = 2m + |E|$

exp family: $p_\theta(x) = \exp\left\{\langle \theta, x\rangle + \frac{1}{2}\langle \Theta, xx^T\rangle - A(\theta, \Theta)\right\}$ (3.12)

$\langle \theta, x\rangle := \sum_{i=1}^m \theta_i x_i$ - Euclidean product on $\mathbb{R}^m$

$\langle\langle \Theta, xx^T\rangle\rangle := \text{trace}(\Theta xx^T) = \sum_{i=1}^m \sum_{j=1}^m \Theta_{ij} x_i x_j$ Frobenius inner product on matrices

· integral defining $A(\theta, \Theta)$ finite iff $\Theta \prec 0$ so:-

$$\Omega = \{(\theta, \Theta) \in \mathbb{R}^m \times \mathbb{R}^{m\times m} \mid \Theta \prec 0 \quad \Theta = \Theta^T\} \tag{3.14}$$

(*) Heterogenous comb. of exp family → mixture models. (2 level hierarchy)
→ LDA (3 level —11—)

Hard constraints → convex theory. (O/S 1)

(*) 3.4 Mean param. and inference

- Recall exp family p_θ by vector of canon. param $\theta \in \Omega$
- Any exponential family has alternative param using vector mean param
(*) conditioning → treat case of marginalization and ignore
3.4.1 - Mean param spaces; marginal polytope

- p - arbitrary density defined wrt underlying base measure. ✓
(*) do not assume that p is a member of an exp fam rest ✓
(*) mean param μ_α associated with sufficient statistic $\phi_\alpha : \mathcal{X}^m \to \mathbb{R}$ is defined

$$\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) \nu dx, \quad \alpha \in \mathcal{I} \tag{3.25}$$

$$= \int \phi_\alpha(x) p(x) dx$$

(*) define a vector of mean parameters $\mu = (\mu_1, \ldots, \mu_d)$, one for each of the $|\mathcal{I}| = d$ sufficient statistics ϕ_α with resp. to arbitr. density p.
(*) consider the set of all such vectors $\mu \in \mathbb{R}^d$ traced as underlying density p is varied.

(*) Formally:-

$$\mathcal{M} := \{\mu \in \mathbb{R}^d \mid \exists p \text{ s.t. } \mathbb{E}_p[\phi_\alpha(x)] = \mu_\alpha \ \forall \alpha \in I\}$$

(*) corresponds to set of all realisable mean params

(*) No restriction on density p to exp. family associated with suff stats ϕ and base measure ν

(*) $\mathcal{M}$ is always a convex subset of $\mathbb{R}^d$

(*) For discrete exponential families $\longrightarrow$ $\mathcal{M}$ is the marginal polytope

(*) convex-null rep; convex polytope (Minkowski-Weyl rep.

(*) Discrete r.v.s. yields $\mathcal{M}$ with special properties

- For any random vec $(x_1, \dots, x_m)$ such that state space $\mathcal{X}^m$ is finite

(*) Representation: (convex null rep)

$$\mathcal{M} = \left\{\mu \in \mathbb{R}^d \mid \mu = \sum_{x \in \mathcal{X}^m} \phi(x) p(x) \text{ for some } p(x) \geq 0, \ \sum_{x \in \mathcal{X}^m} p(x) = 1\right\}$$

$$= \text{conv}\{\phi(x), x \in \mathcal{X}^m\}$$

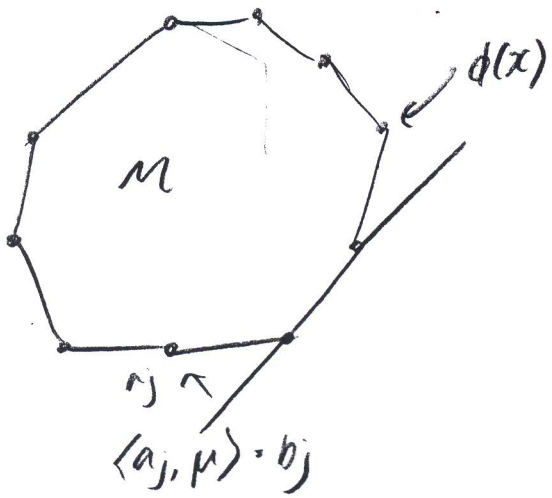
(*) $|\mathcal{X}^m|$ finite $\rightarrow$ $\mathcal{M}$ is a convex polytope.

(*) Half-plane rep

- Via Minkowski-Weyl, alt description of convex polytope characterised using
- finite-collection-linear-inequality-constraints
- For any polytope $\mathcal{M}$, there exists a collection $\{(a_j, b_j) \in \mathbb{R}^d \times \mathbb{R} \mid j \in J\}$ with $|J|$ finite:- s.t.

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \mid \langle a_j, \mu \rangle \geq b_j \ \forall j \in J\} \tag{3.29}$$

- geometrically; intersection of finite coll. of half spaces



M for discrete r.v. $|\mathcal{X}^m|$ finite

M - convex polytope

- convex hull of $\{\phi(x) \mid x \in \mathcal{X}^m\}$

Minkow-Weyl :- intersection of finite no. halfspaces each of form $\{\mu \in \mathbb{R}^d \mid \langle a_j, \mu \rangle \geq b_j\}$ for some $(a_j, b_j) \in \mathbb{R}^{d} \times \mathbb{R}$

(*) Example 3.8. Ising (2 node)

(*) magnal polytope, suff stats, mean param

(*) magnal polytope for discrete GMs.

(*) magnal polytope - any GM - multinomial r.v. $x_s \in \mathcal{X}_s := \{0, 1, \ldots, r_s - 1\}$ - at each - vertex $s \in V$.

(*) cardinality $|\mathcal{X}_s| = r_s$ can differ.

(*) Exp. family - $\{0,1\}$ indicators :-

$$\forall s \in V, j \in \mathcal{X}_s \quad \mathbb{I}_{s;j}(x_s) := \begin{cases} 1 \text{ if } x_s = j \\ 0 \text{ otw} \end{cases}$$

$$\forall (s,t) \in E, (j,k) \quad \mathbb{I}_{st;jk}(x_s, x_t) := \begin{cases} 1 \text{ if } (x_s, x_t) = (j,k) \\ 0 \text{ otherwise} \end{cases} \qquad (3.34)$$

(*) Sufficient stats - standard overcomplete rep.

(*) mean params :-

for each node $s \in V$ $\quad \mu_{s;j} := \mathbb{E}_p[\mathbb{I}_j(x_s)] = P[x_s = j] \quad \forall j \in \mathcal{X}_s \qquad (3.35)$

edge $(s,t) \in E$ $\quad \mu_{st;jk} = \mathbb{E}_p[\mathbb{I}_{s,t;jk}(x_s, x_t)] = P[x_s = j, x_t = k] \quad \forall (j,k) \in \mathcal{X}_s \times \mathcal{X}_t \qquad (3.36)$

(*) : mean param $\longrightarrow$ singleton magnal distris μ_s (nodes)
pairwise —"— μ_{st} (edges)

(*) In this case, set M is the marginal polytope - associated - graph - G

$$\mathbb{M}(G) := \{\mu \in \mathbb{R}^d \mid \exists p \text{ s.t. } (3.35) \text{ holds } \forall (s;j) \text{ and } (3.36) \text{ -''- } \forall (st;jk)\}$$

characterisation of $\mathbb{M}(G)$:

- How many half-space constraints to characterise $\mathbb{M}(G)$?
- trivial for easie examples.
- depends on graph structure
- trees: $\mathbb{M}(G)$ char by local constraints, pairs of r.vs. on edges - linear growth in graph size

General G: nonlocal constraints, growth is fast

(!!!) Intractability of representing marginal polytopes - compact - inverse
- one underlying cause - complexity - statistical computation.

3.4.2. Role of mean parameters

(*) mean params - central role - marginals. problem

(*) forward mapping: canon $\theta \in \Omega \longrightarrow$ mean $\mu \in M$

- fundamental class of inference problems in exp. family
- tractable for low-dim; extremely diffi for high-dim exp family

(*) Backward mapping: mean $\mu \in M \longrightarrow$ canon $\theta \in \Omega$

- Statistical interp:
  - Given samples $X_1^N := \{x^1, \dots, x^N\}$ indep. from exp-fam. $p_\theta(x)$
  - θ unknown
  - Goal: estimate $\theta \rightarrow$ use ML to obtain $\hat{\theta}$
  - Equivalently: Maximise $\ell(\theta; X_1^N) := \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(x^i) = \langle \theta, \hat{\mu} \rangle - A(\theta)$ (3.38)

- $\hat{\mu} := \hat{\mathbb{E}}[\phi(X)] = \frac{1}{N} \sum_{i=1}^{N} \phi(x^i)$ - vector of empirical mean params defined by data X_1^N.

- ML estimate $\hat{\theta}$ chosen to maximise obj
- computing $\hat{\theta}$ is may be challenging due to $A(\theta)$ (log-part.)

(*) under suitable cond; MLE is unique ispecified by $\mathbb{E}_{\hat{\theta}}[\phi(X)] = \hat{\mu}$

(*) Finding unique solution - equivalent - computing - inverse mapping

$\mu \mapsto \theta$ (from mean → canonical)

(*) Inverse mapping → comp. intensive

3.5 Properties of A

(*) (i) convexity of A ⟶ convex analysis ⟶ conjugate duality.

(ii): Suitable cond - function A - and - conjugate dual - A* (derivatives) define a - one-to-one and surjective mapping between μ and θ.

(iii): mapping between canonical, mean → core challenge underlying stat comp in high dim GM.

(*) Derivatives convexity:-

(*) Real valued g convex if for any two x, y in domain of g:- scalar $\lambda \in (0,1)$

$$g(\lambda x + (1-\lambda) y) \leq \lambda g(x) + (1-\lambda) g(y) \qquad (3.39)$$

P.3.1.

(*) cumulant / log part.:-

$$A(\theta) := \log \int_{x^m} \exp\langle \theta, \phi(x) \rangle \nu dx \qquad (3.40)$$

$$= \log \int \exp\langle \theta, \phi(x) \rangle dx$$

for regular exp. family has prop:-

a) Derivatives of all order on domain Ω

- First two deriv (only 1st shown as slides) yield 'cumulants' of r.vec $\phi(X)$:-

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha(X)] := \int \phi_\alpha(x) p_\theta(x) \nu dx \qquad (3.41.a)$$

b) A is a convex function of θ on its domain Ω; strictly so if representation is minimal

3.5.2 - Forward mapping ($\theta \mapsto \mu$)

(*) Forward mapping $\theta \mapsto \mu$

Given param $\theta \in \Omega$ defining p_θ to mean param vec $\mu \in \mathbb{R}^d$

(*) Gradient ∇A is mapping from Ω to $\mathbb{R}^d$.

P 3.1 : Range of mapping contained within realisable set of mean param

$$\mathcal{M} := \{\mu \in \mathbb{R}^d \mid \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\}$$

(*) : Q1) When does ∇A define 1-1 mapping?

Q2) When does the image of Ω under mapping ∇A that is $\nabla A(\Omega)$, fully cover set $\mathcal{M}$?

(*) Q1) - depends on whether exp family is minimal

Q2) - Delicate

(*) Observe:- $\mathcal{M}$'s definition → mean param $\mu \in \mathbb{R}^d$ gen. by any poss distri ; not just distris p_θ in exp family defined by sufficient statistics

(*) Extra 'freedom' does not really enlarge set $\mathcal{M}$

(*) Theorem 3.3 - Suitable cond :-

- All mean params in $\mathcal{M}$ can be realised by an exp family distri

---

(*) ∇A defines 1-1 mapping?

R1

Prop 3.2 : Gradient mapping $\nabla A : \Omega \to \mathcal{M}$ is 1-1 iff exponential representation is minimal

(*) Proof → WaJ (2008)

(*) overcomplete representation ; ∇A is not 1-1.

(*) BUT, 1-1 between each element of $\nabla A(\Omega)$ and an affine subset of Ω.

(*) Affine subset contains all those canonical parameters θ that are mapped to same mean parameter

(iv): for minimal or overcomplete representations; a pair (θ, μ) is dually coupled if $\mu = \nabla A(\theta)$.

(Q2): image $\nabla A(\Omega)$ of the domain of valid canonical params Ω under gradient mapping ∇A

- Goal: Determine for which mean parameters $\mu \in M$ does there exist a vector $\theta = \theta(\mu) \in \Omega$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$?

Sol: The image of $\nabla A(\Omega)$ is the interior M° (Q1) (iv)

(Q1) All mean parameters M that are realisable by some distribution are realised by a member of the exponential family

(i) The connection between ML and ME principles for estimation is due to primal-dual relationship

- MEP:- p^* is exponential family i.e. $P_\theta(\mu)$
  - must satisfy $\mathbb{E}_{\theta(\mu)}[\phi(X)] = \mu$ (moment matching) identical to ML problem.

(iii) Theorem 3.3: In a minimal exp. family, gradient map ∇A is onto interior of M, denoted M°.

(*) For each $\mu \in M^\circ$, there exists some $\theta = \theta(\mu)$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$.

(iii): Minimal exp family, each $\mu \in M^\circ$ (mean param) uniquely realised by some density $P_\theta(\mu)$ in exponential family

(i) Note:- Typical exponential family $\{P_\theta | \theta \in \Omega\}$ describes strict subset of all possible densities

(iii): There exist some other density p (not in exp family) that realises μ.

(iii): The distinguishing property of exp distr $P_\theta(\mu)$ is that amongst all set of all distributions that realise μ; it has maximum entropy

*) connection between A and MEP formalised via conjugate dual A^*.

3.6. Conjugate Duality, ML, MEP

*) Relationship between log partition A and its conjugate dual A^* is central to the variational principle underlying many exact and approx inference algorithms

3.6.1. Conjugate Dual (General)

- Given a function A (lecture 10-708 presents this in an arbitrary way) the conjugate dual A^*:-

$$A^*(\mu) := \sup_{\theta \in \Omega} \{\langle \mu, \theta \rangle - A(\theta)\} \tag{3.42}$$

*) $\mu \in \mathbb{R}^d$ - fixed vector of dual variables of same dim as θ.

*) Deliberate notational choice. $\longrightarrow$ dual variables are naturally the mean param

*) Statistical interp of (3.42) (ML):-

- RHS is optimise of rescaled log-likelihood.
- only makes sense when $\mu \in \mathcal{M}$ e.g. vector of empirical moments $\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} \phi(x^i)$ induced by set of data $X_1^n = \{x^1, \ldots, x^n\}$

*) Inside (3.42) more broadly $\rightarrow$ any vector $\mu \in \mathbb{R}^d$.

*) View A^* as fn with values on extended real line:- $\mathbb{R}^* = \mathbb{R} \cup \{+\infty\}$.

Connection of conjugate dual and entropy

- (3.42), (3.2) closely connected.

Next theorem:-

*) When $\mu \in \mathcal{M}^o$:-

) value of dual $A^*(\mu)$ is negative entropy of exp family distri $P_{\theta(\mu)}$, with $\theta(\mu)$ being unique vector of canon params satisfying

$$\mathbb{E}_{\theta(\mu)}[\phi(x)] = \nabla A(\theta(\mu)) = \mu \tag{3.43}$$

(*)(on): $\mu \notin \mathcal{M}^\circ$:-

(*) impossible to find canonical params satisfying (3.43)

(*) when $\mu \notin \bar{\mathcal{M}}$ ($\bar{\mathcal{M}}$ is closure of $\mathcal{M}$), then $A^*(\mu) = +\infty$

(on):- in general, this underpins variational methods:-

(on) optimisation problem over dual can be reduced to an optimisation problem over $\mathcal{M}$

(*) we can view optimisation over dual (i.e. param est^n) through the structure of $\mathcal{M}$, and for more complex GMs, approximations to $\mathcal{M}$.

Theorem 3.4: (A and conjugate dual A^*)

a) For any $\mu \in \mathcal{M}^\circ$, denote by $\theta(\mu)$ the unique canonical param satisfying dual matching condition (3.43)

conjugate dual A^* takes the form

$$A^*(\mu) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \bar{\mathcal{M}} \end{cases} \qquad (3.44)$$

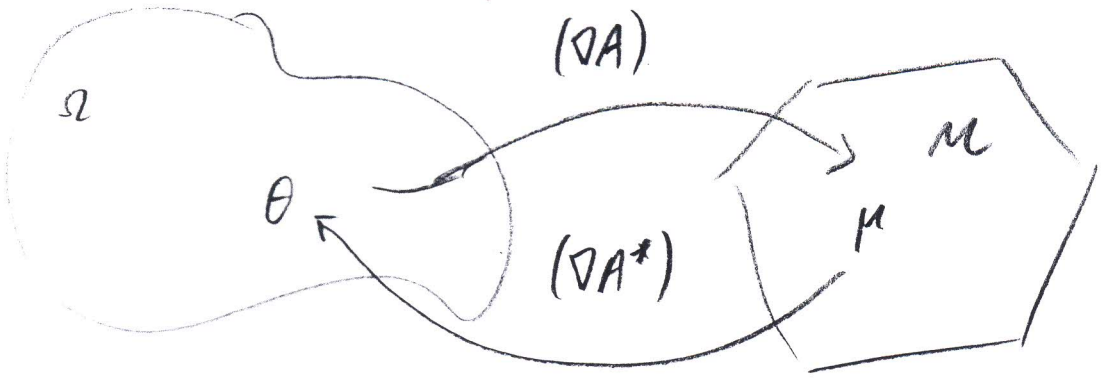
for any boundary point $\mu \in \bar{\mathcal{M}} \setminus \mathcal{M}^\circ$, we have $A^*(\mu) = \lim_{n \to +\infty} A^*(\mu^n)$ taken over any sequence $\{\mu^n\} \subset \mathcal{M}^\circ$ converging to μ.

(on)
b) in terms of this dual, log partition has variational representation

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \qquad (3.45)$$

(*)
c) for all $\theta \in \Omega$, the supremum in (3.45) is obtained uniquely at the vector $\mu \in \mathcal{M}^\circ$ specified by moment matching conditions

(*) see W&J (2008) for informative discussion of nuances



Remarks:

- $H(p)$ maps density functions $\to \mathbb{R}$
- A^* is an extended real-valued function on $\mathbb{R}^d$; finite for only valid mean values $\mu \in M$
- (i) Maximisation over $\mathbb{R}^d$ is re-expressed as maximisation over M.
- (ii) : Nature of M plays crucial role in complexity of computing cumulant $A(\theta)$.
- (iii) Bijection between Ω and M^o, for minimal exponent. family
- (*) $\nabla A : \Omega \to M^o$; $\nabla A^* : M^o \to \Omega$: bijective correspondence. (inverse mapping)
- (*) Table 3.2 → intuition ✓

---

*) Example 3.10 (conj. duality for Bernoulli)

- Bernoulli r.v. $X \in \{0,1\}$ $p(x) =$
- exp family rep:-

$$p(x;\theta) = \exp\{\langle \theta, \phi(x)\rangle - A(\theta)\} \qquad \phi(x) = x \quad A(\theta) = \log(1+e^\theta) \quad \Omega = \mathbb{R}$$

$$= \exp\{\theta x - \log(1+e^\theta)\}$$

Inference → $\mu(\theta) = \mathbb{E}_\theta[X] = 1 \cdot p(X=1;\theta) + 0 \cdot p(X=0;\theta) = \dfrac{e^\theta}{1+e^\theta}$

Do this step in a variational manner.

- For any fixed $\mu \in \mathbb{R}$

$$A^*(\mu) = \sup_{\theta \in \mathbb{R}} \{\theta \cdot \mu - \log(1+\exp\theta)\} \qquad (3.47)$$

(*) take derivatives ⟶ stationary conditions

$\Rightarrow \mu = \frac{\exp(\theta)}{1+\exp(\theta)}$ (moment matching cond.)

(Q): When is there a solution to the stationary condition?

- $\mu \in (0,1)$, $\theta(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ (unique sol via T.3.3 a P3.2)
- strictly convex objective ⇒ subs. $\theta(\mu)$ → (3.47)

$A^*(\mu) = \mu \log \mu + (1-\mu)\log(1-\mu) = -H(\text{Ber}(\theta(\mu)))$ $\mu \in (0,1) = \mathcal{M}^\circ$

$\mu \notin [0,1] = \bar{\mathcal{M}}$

$\mu \notin \bar{\mathcal{M}} = [0,1]$

- consider limiting behaviour of supremum as $\theta \to \pm\infty$

(..)

- $\mu > 1$, $\mu < 0$; $A^*(\mu) = +\infty$ for $\mu \notin \bar{\mathcal{M}}$

unless $\mu \in [0,1]$, $A^*(\mu) = +\infty$

⇒ optimis. problem $A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$ becomes

$$\max_{\mu \in [0,1]} \{\mu \cdot \theta - \mu \log \mu - (1-\mu)\log(1-\mu)\}$$

(*) Solving this explicitly yields optimal value at $\log(1+\exp(\theta))$ (claim 3.45)

(concave max.)

Attained at $\mu(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)}$ (T.3.4c)

(IS!): Need to process what is going on at a high-level

3.7 Computational challenges with high-dim models

(5m): important features:-

- variational representation of log-partition:-

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\} \qquad (3.45)$$

- optimum uniquely achieved at mean parameters $\mu = \mathbb{E}_\theta[\phi(X)]$
- (5m): we can compute A and μ by solving (3.45)

(5m): in practice, two challenges:-

i) constraint set $\mathcal{M}$ of realisable mean params is extremely difficult to characterize

ii) negative entropy function A^* is defined indirectly; lacks explicit form.

(*) comput. challenges:

i) constraint set $\mathcal{M}$ (complexity):-

. MRFs for discrete r.v.s $x \in \{0, 1, \ldots, r-1\}^m$, we have marginal polytope

- finite no. linear inequality constraints
- general graphs $\longrightarrow$ no. of inequalities grows with graph size.

(*) not possible to optimise a linear function over $\mathcal{M}$ for a general discrete MRF.

ii) indirectly specified $A^*(\mu)$:-

- evaluating cost function at single point $\mu \in \mathcal{M}$, let alone opt. difficult.

$$\mu \longrightarrow \boxed{(\nabla A)^{-1}} \longrightarrow \theta(\mu) \longrightarrow \boxed{-H(p_{\theta(\mu)})} \longrightarrow A^*(\mu)$$

(i) (ii)

(5m): complexity of evaluating dual value $A^*(\mu)$:-

- 1.3.4 - implicit comput. of $A^*(\mu)$ via comp. of mappings. :-

- $(\nabla A)^{-1}: \mathcal{M}^\circ \to \Omega$: μ maps to $\theta(\mu)$; corres. to exp family

ii) $-H(p_{\theta(\mu)}) : \Omega \to \mathbb{R}$: $\theta(\mu)$ maps to $A^*(\mu)$

- mapping from $\theta(\mu)$ to negative entropy $-H(p_{\theta(\mu)})$ of assoc. exp family density.

(or): These mappings (these mappings $\nabla(A)^{-1}(\mu)$, $-H(p_{\theta(\mu)})$) are often intractable $\longrightarrow$ use approx. methods.