

U2 review

U2 - Theory of V.I.: Info / outer 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.M.S.

- n IID observations X^1, \dots, X^N

empirical exp: $\hat{\mu}_\alpha := \frac{1}{n} \sum_{i=1}^N \phi_\alpha(X^i) \quad \forall \alpha \in \mathbb{I}$ (3.1)

of prob. function

\mathbb{I} - index set.
mapping
 $\phi_\alpha: \mathcal{X} \rightarrow \mathbb{R}$

- stack $\hat{\mu}_\alpha$ into a vector

- giving $|\mathbb{I}|$ -dim vector of emp. exp. $\hat{\mu}$

- using μ , goal is to infer full prob. distri over r.v. X .

- represent prob. distri as density p , absolutely continuous wrt measure ν

- consider expectations: - (ass. existence)

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

(*) expectations under distri p matched to exp. under empirical distri

(*) underdetermined → many distris p that are consistent with observations

(*) require principle for selection:-

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

$$H(p) := \int_X p(x) \log p(x) \nu dx \quad (3.2)$$

(*) Principle of max entropy:

- from among distris consistent with data, select p^* whose Shannon entropy is maximal

(*) Formally;

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

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

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

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

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

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

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

3.2 Basics of Exp Families

(*) Exponential family \rightarrow parameterised family of densities taken with respect to some underlying measure ν .

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

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

(*) ϕ_α - potential function/sufficient statistic

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

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

(*) 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 in \mathbb{R}^d of two vectors θ and $\phi(x)$

(***) Exponential family associated with ϕ :- parameterised coll of districs:-
 $\text{wt } dv$

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

$$\begin{aligned} (\text{****}) \text{ log partition/cumulant: } A(\theta) &= \log \int_{\mathcal{X}^m} \exp \{ \langle \theta, \phi(x) \rangle \} v dx \\ &= \log \int \exp \cdot \theta^T \phi(x) dx \end{aligned} \quad (3.6)$$

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

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

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

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

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

(*****) Log-partition function A - prominent

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

Wainwright, Jordan (2003) (2.2)

restrict attention to case where SL is open \rightarrow 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 .

(b) overcomplete exp. family: there exist linear combinations $(\alpha, \phi(x))$ equal to a constant

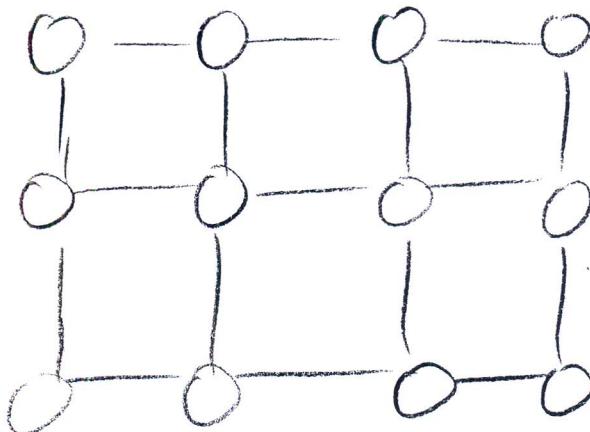
- there exists an affine subset of parameter vectors Θ , each associated with same distri.
- identifiability of param vector θ lost

(*) Examples:-

Ising model

$$G = (V, E) \quad \text{nodes}$$

$$x_s \in \{-1, +1\} \quad x_s \sim \text{Bern}(p) \text{ s.e.v}$$



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

$$p_0(x) = \exp \left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - A(\theta) \right\} \quad (3.8)$$

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

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

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

(*) 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 $S\Omega$ is full space \mathbb{R}^d , family regular

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

$n=m \rightarrow$ extreme, connected

all nodes, represent any distri over a binary random vector

(*) Example 3.2

Potts Model: R.V x_s at each node $s \in V$.

(gen. of Ising) values in discrete space $\mathcal{X} := \{0, \dots, r-1\}$ $r \geq 2$ 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: $I_{s,j}(x_s) = \begin{cases} 1 & \text{if } x_s=j \\ 0 & \text{otherwise} \end{cases}$ (3.10)

Associated cov. param: $\theta_s = \{\theta_{s,j}, j=0, 1, \dots, r-1\}$

Sufficient statistic: $I_{s,t,j,k}(x_s, x_t) = \begin{cases} 1 & \text{if } x_s=j, x_t=k \\ 0 & \text{otherwise} \end{cases}$ (3.11)

Assoc. param: $\theta_{s,t,j,k} \in \mathbb{R}$

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

(?) are complete/regular?

(*) Example 3.3

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

MVG / vec (x_1, \dots, x_m) respects Markov prop. of G

represented in exponential form using ^{coll.} 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

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

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

(i) - negative of inverse covariance / precision matrix

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

$d = 2m + |E|$

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

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

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

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

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

(*) Heterogeneous comb. of exp family \rightarrow mixture models. (2 level hierarchy)
 \rightarrow LDA (3 level - - -)

Hard constraints \rightarrow convex theory.

0/5 1

(*) 3.4 Mean Param. and Inference

- characterize exp family p_θ by vector of mean param $\theta \in \mathbb{R}$

- Any exponential family has alternative param using vector mean param

(*) conditioning \rightarrow treat case of regularization and ignore

3.4.1. mean param spaces; moment polytope

- p - arbitrary density defined w.r.t underlying base measure. ✓

(*) do not assume that p is a member of an exp fam not ✓

(*) mean param μ associated with sufficient statistic $\phi_x: \mathcal{X}^m \rightarrow \mathbb{R}$ is defn!.

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

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

(*) Define a vector of mean parameters $\mu = (\mu_1, \dots, \mu_d)$, one for each of the $|I| = d$ sufficient statistics ϕ_α with resp. to arbit. density p .

(*) Consider the set of all such vectors $\mu \in \mathbb{R}^d$ traced as underlying density p is varied.

(*) Formally:-

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

(*) corresponds to set of all realisable mean vectors

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

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

(*) For discrete exponential families $\rightarrow M$ is the marginal polytope

(*) convex-hull rep; convex polytope (minkowski-weyl rep.)

(*) Discrete r.v.s. yields M with special properties

- for my random vec (X_1, \dots, X_m) such that state space X^m is finite

(*) Representation:
(convex hull rep)
$$M = \left\{ \mu \in \mathbb{R}^d \mid \mu = \sum_{x \in X^m} \phi(x) p(x) \text{ for some } p(x) \geq 0, \right.$$

$$\left. \sum_{x \in X^m} p(x) = 1 \right\}$$

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

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

(*) Half-plane rep

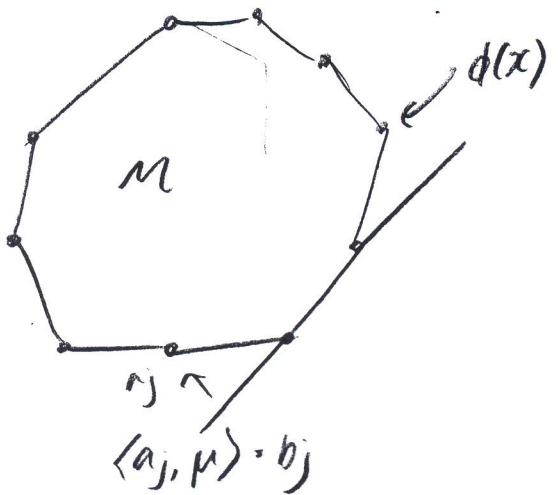
- via minkowski-weyl, alt. description of convex polytope characterised using

- finite-collection-linear-inequality-constraints

- For any polytope 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.

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

- geometrically; intersection of finite coll. of half-spaces



M for discrete r.v. $|X^m|$ finite
 M - convex polytope
 - convex hull of $\{\phi(x) | x \in X^m\}$
 Minkowski-hull:- intersection
 of finite no. halfspaces
 each of $\{x \in \mathbb{R}^d | \langle a_j, x \rangle \geq b_j\}$
 for some $(a_j, b_j) \in \mathbb{R}^d \times \mathbb{R}$

(*) Example 3.8. Ising (2 node)

(*) marginal polytope, suff stats, mean param

(*) marginal polytope for discrete GMs.

(*) marginal polytope - any GM - multinomial r.v. $X_S \in \mathcal{X}_S := \{0, 1, \dots, r_S - 1\}$
 - at each vertex $s \in V$.

(*) cardinality $|\mathcal{X}_S| = r_S$ can differ.

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

$$\forall s \in V, j \in \mathcal{X}_S \quad I_{S,j}(x_S) := \begin{cases} 1 & \text{if } x_S = j \\ 0 & \text{otherwise} \end{cases}$$

$$\forall (s, t) \in E, (j, k) \quad I_{S,t;j,k}(x_S, x_t) := \begin{cases} 1 & \text{if } (x_S, x_t) = (j, k) \\ 0 & \text{otherwise} \end{cases} \quad (3.34)$$

(*) Sufficient stats - standard overcomplete rep.

(*) mean params :-

$$\text{for each node } s \in V \quad \mu_{S,j} := \mathbb{E}_p[I_j(x_S)] = P[x_S = j] \quad \forall j \in \mathcal{X}_S \quad (3.35)$$

$$\text{edge } (s, t) \in E \quad \mu_{S,t;j,k} = \mathbb{E}_p[I_{S,t;j,k}(x_S, x_t)] = P[x_S = j, x_t = k] \quad \forall (j, k) \in \mathcal{X}_S \times \mathcal{X}_t \quad (3.36)$$

(*) mean param \rightarrow singleton marginal dists μ_S (nodes)
 pairwise \longrightarrow $\mu_{S,t}$ (edges)

i) In this case; set M is the marginal polytope-associated-graph- G

$$IM(G) := \{ \mu \in \mathbb{R}^d \mid \exists p \text{ s.t. (3.35) holds } V(s, j) \text{ and } (3.36) \text{ --- } V(st; jk) \}$$

characterisation of $IM(G)$:

- How many half-space constraints to characterise $IM(G)$?
- trivial for early examples.
- depends on graph structure
- trees: $IM(G)$ char by local constraints, pairs of r.v.s. on edges - linear growth in graph size

General G : nonlocal constraints, growth is fast

(*) Intractability of representing marginal polytopes - compact - manner

- one underlying cause - complexity - statistical computation.

3.4.2. Role of mean params in inference

(*) Mean params - central role - marginalis. problem

(*) Forward mapping: canon $\theta \in \Omega \rightarrow$ 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 $\mu \in M \rightarrow$ canon $\theta \in \Omega$

- Statistical interp:

- given samples $X_i^n := \{x_1^n, \dots, x_n^n\}$ indep. from exp-fam. $p_\theta(x)$

- θ unknown

- goal: estimate $\theta \rightarrow$ use ML to obtain $\hat{\theta}$

- equivalently: maximise $\ell(\theta; X_i^n) := \frac{1}{n} \sum_{i=1}^n \log p_\theta(x^i) = \langle \theta, \hat{\mu} \rangle - A(\theta)$ (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_i^n .

- ML estimate $\hat{\theta}$ chosen to maximise obj.

- computing $\hat{\theta}$ may be challenging due to $A(\theta)$ (log-part.)

(*) under suitable cond; MLE is unique is specified by $\partial_{\theta} \ell(\theta; X) = \hat{\mu}$

(*) finding unique solution - equivalent - computing - backed mapping

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

(*) inverse mapping \rightarrow comp. intensive

3.5 Properties of A

(*) (i) convexity of A \rightarrow convex analysis \rightarrow 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 \rightarrow 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 :-

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y) \quad \text{for all } \lambda \in (0,1) \quad (3.39)$$

P.3.1.

(*) Cumulant / log part. :-

$$\begin{aligned} A(\theta) &:= \log \int_{\mathcal{X}^m} \exp \langle \theta, \phi(x) \rangle v dx \\ &= \log \int \exp \langle \theta, \phi(x) \rangle dx \end{aligned} \quad (3.40)$$

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 / vec $\phi(x)$:-

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

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 ϕ 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

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

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

Q2) when does the range of Ω under mapping ∇A that is $\nabla A(\Omega)$, fully cover set M ?

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

Q2) - delicate

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

(*) Extra 'freedom' does not really enlarge set M

(*) Theorem 3.3 - suitable cond :-

- All mean params in M can be realised by an exp family distri

(*) ∇A defines 1-1 mapping?

Prop 3.2: gradient mapping $\nabla A: \Omega \rightarrow M$ is 1-1 iff exponential representation is minimal

(*) Proof \rightarrow WdJ(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

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

(ii) image $\nabla A(\Omega)$ of the domain of valid conc. 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 $E_\theta[\phi(X)] = \mu$?

sol: the image of $\nabla A(\Omega)$ is the interior M° (on)

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

(iv) 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 $E_{P(\theta|\mu)}[\phi(X)] = \mu$ (moment matching)
identical to ML problem.

(v) theorem 3.3: in a minimal exp. family, gradient map ∇A is onto interior of M , denoted M° .

(vi) For each $\mu \in M^\circ$, there exists some $\theta = \theta(\mu)$ such that

$$E_\theta[\phi(X)] = \mu.$$

(vii) minimal exp. family, each $\mu \in M^\circ$ (mean param) uniquely realised by some density $P(\theta|\mu)$ in exponential family

(viii) note:- typical exponential family $\{\rho(\theta) e^{S(\theta)}\}$ describes strict subset of all possible densities

(ix) there exist some other density p (not in exp. family) that realises μ .

(x) the distinguishing property of exp. distri. $P(\theta|\mu)$ 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 10-208 presents this in an arbitrary way)
the conjugate dual A^* :-

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

i) $\mu \in \mathbb{R}^d$ - fixed vector of dual variables of some dim as θ .
ii) deliberate notational choice, \rightarrow dual variables are naturally the mean param

iii) statistical MEP of (3.42) (ML) :-

- RHS is optimise of rescaled log-likelihood.
- only makes sense when $\mu \in M$ e.g. vector of empirical moments

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^n \phi(x^i) \text{ induced by set of data } X_i^n = \{x_1^n, \dots, x_n^n\}$$

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

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

Connection of conjugate dual and entropy

(3.42), (3.2) closely connected.

Next theorem:-

(*) When $\mu \in M^0$:-

vi) 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

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

(*)
if $\mu \notin M^0$:-

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

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

(**) - In general, this underpins variational methods:-

(**) optimisation problem over dual can be reduced to an optimisation problem over M

(*) we can view optimisation over dual (i.e. param est'n) through the structure of M , and for more complex GMS, approximations to M .

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

a) for any $\mu \in M^0$, denote by $\theta(\mu)$ the unique canonical param satisfying dual matching condition (3.43).

conjugate dual A^* takes the form

$$A^*(\mu) = \begin{cases} -H(\theta(\mu)) & \text{if } \mu \in M^0 \\ +\infty & \text{if } \mu \notin \bar{M} \end{cases} \quad (3.44)$$

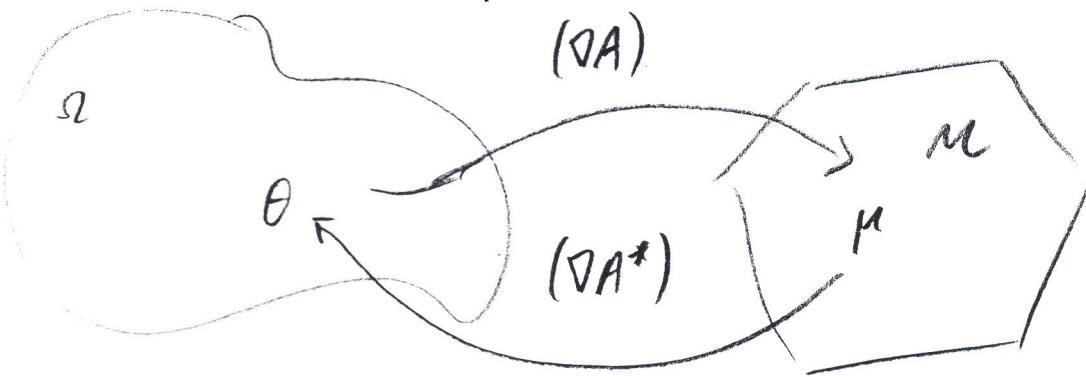
for any boundary point $\mu \in \bar{M} \setminus M^0$, we have $A^*(\mu) = \lim_{n \rightarrow \infty} A^*(\mu^n)$ taken over any sequence $\{\mu^n\} \subset M^0$ converging to μ .

b) In terms of this dual, log partition has variational representation

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

c) for all $\theta \in \Omega$, the supremum in (3.45) is obtained uniquely at the vector $\mu \in M^0$ specified by moment matching conditions

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



REMARKS:

- $H(\mu)$ maps density functions $\rightarrow \mathbb{R}$
- A^* is an extended real-valued function on \mathbb{R}^d ; infinite for only valid mean points $\mu \in M$
- (1) Maximisation over \mathbb{R}^d is reexpressed as maximisation over M .
- (2) Nature of M plays crucial role in complexity of computing cumulant $A(\theta)$.
- (3) Bijection between Ω and M° , for minimal exponent. family
- (4) $\partial A: \Omega \rightarrow M^\circ; \partial A^*: M^\circ \rightarrow \Omega$: bijective correspondence.
(inverse mapping)

Table 3.2 \rightarrow intuition ✓

* Example 3.10 (conj. duality for Bernoulli)

Bernoulli r.v. $X \in \{0, 1\}$ $p(x) =$

exp family rep:-

$$p(x; \theta) = \exp \{ \langle \theta, \phi(x) \rangle - A(\theta) \}$$

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

$$\phi(x) = x \quad A(\theta) = \log(1 + e^\theta)$$

$$\Omega = \mathbb{R}$$

$$\text{Inference} \rightarrow p(\theta) = \mathbb{E}_\theta[X] = 1 \cdot p(X=1; \theta) + 0 \cdot p(X=0; \theta) = \frac{e^\theta}{1 + e^\theta}$$

to this step in a variational manner.

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

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

(3.4.7)

(H) take derivatives \rightarrow stationary conditions

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

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

• $\mu \in (0, 1)$, $\partial(\mu) = \log\left(\frac{M}{1-\mu}\right)$ (unique sol via P.3.3 a P3.2)

• strictly convex objective \Rightarrow subs. $\partial(\mu) \rightarrow (3.47)$

$$A^*(\mu) = \mu \log \mu + (1-\mu) \log(1-\mu) = H(\text{Ber}(\partial(\mu))) \quad \underline{\mu \in (0, 1) = M^0}$$

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

• consider limiting behavior of supremum as $\theta \rightarrow \pm\infty$

(-)

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

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

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

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

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

(concave max.)

(claim 3.45)

$$\text{Attained at } \mu(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)}$$

(T.3.4c)

(P/S): need to process what is going on at a high-level

3.7 Computational challenges with n.gh-dim models

6) important features:-

- variational representation of log-partition:-

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

- optimum uniquely achieved at mean parameters $\mu = E_\theta[\phi(X)]$
- we can compute A and μ by solving (3.45)

7) In practice, two challenges:-

- i) constraint set M of realisable mean params is extremely difficult to characterise.
- ii) negative entropy function A^* is defined indirectly; lacks explicit form.

(*) comput. challenges:

i) constraint set M (complexity):-

- MRPs for discrete r.v.s. $\pi \in \{0, 1, \dots, r-1\}^M$; we have marginal polytope
- finite no. linear inequality constraints
- general graphs \rightarrow no. of inequalities grows with graph size.
- (*) not possible to optimise a linear function over M for a general discrete MRP.

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

- evaluating cost function at single point $\mu \in M$, i.e. done opt. difficult.

$$\mu \xrightarrow{(i)} \boxed{(\nabla A)^{-1}} \xrightarrow{\theta(\mu)} \boxed{-H(\rho_\theta(\mu))} \xrightarrow{(ii)} A^*(\mu)$$

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

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

- $(\nabla A)^{-1}: M^0 \rightarrow \Omega$: μ maps to $\theta(\mu)$; cores to exp family

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

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

iii) These mappings (use $\nabla(A)^{-1}(\mu)$, $-H(\rho_{\theta(\mu)})$) are
often intractable \rightarrow use approx. methods.