10-708 16/05/2020

L7 review - Parameter est. UGMs

- (*) check calculations/derivations
(*) use Jordan 9.3-9.5 + 20 to assist → skim
(*) skim Lafferty or tutorial

sufficient statistics from Jordan (2003) Ch 9.3

- total count and clique counts
- $\underline{X}_V$ - random vector associated with graph (A1)
- $\underline{X}_C$ $C \subseteq V$ - random vectors — " subsets of nodes
- IID replicates of obs.
- $\underline{x}_{C,n}$ - n^{th} replicate of subset C
- $\mathcal{D} = \{\underline{x}_{V,1}, \underline{x}_{V,2}, \ldots, \underline{x}_{V,N}\}$ - completely observed
- Parametrise UGM via clique potentials $\psi_C(\underline{x}_C)$ for $C \in \mathcal{C}$
- $\mathcal{C}$ - set of cliques.
- via Hammersley-Clifford:

Joint prob: $p(\underline{x}_V | \underline{\theta}) = \frac{1}{Z} \prod_C \psi_C(\underline{x}_C)$ $\underline{\theta} = \{\psi_C(\underline{x}_C), C \in \mathcal{C}\}$ (collec. param.)

Z = normalisation

$$Z = \sum_{\underline{x}_V} \prod_C \psi_C(\underline{x}_C)$$

(*) Z - obtained via summing (or integrating) over all configurations $\underline{x}_V$

The total counts

- No. of times that configuration $\underline{x}_V$ is observed in a dataset $\mathcal{D}$

$$m(\underline{x}_V) := \sum_{n=1}^{N} \delta(\underline{x}_V, \underline{x}_{V,n})$$

Marginal counts (clique counts)

for clique C

$$m(\underline{x}_C) := \sum_{\underline{x}_{V \setminus C}} m(\underline{x}_V)$$

- $N = \sum_{\underline{x}_V} m(\underline{x}_V)$ - total no. of observations

---

(*) log-likelihood for MRFs (tabular clique pot.) (A3)

(*) express log-likelihood in terms of counts (sufficient statistics for discrete models)

- introduce dummy var $\underline{x}_V$ (?)

$$p(\underline{x}_V, n \mid \underline{\theta}) = \prod_{\underline{x}_V} p(\underline{x}_V \mid \underline{\theta})^{\delta(\underline{x}_V, \underline{x}_V, n)}$$

- OIC - indicator trick that switches on/off depending on distri.

(**) $\delta(\underline{x}_V, \underline{x}_V, n) = \mathbb{I}\{\underline{x}_V = \underline{x}_V, n\}$

(**) dummy variable $\underline{x}_V$ ranges across configurations of nodes rather than across data points

- standard for multinomials (remember Bishop?)

Probability of observed data :- $p(\mathcal{D} \mid \underline{\theta}) = \prod_{n=1}^{N} p(\underline{x}_V, n \mid \underline{\theta})$

$$= \prod_{n=1}^{N} \prod_{\underline{x}_V} p(\underline{x}_V \mid \underline{\theta})^{\delta(\underline{x}_V, \underline{x}_V, n)}$$

Log-likelihood in terms of marginal counts

$$\ell(\underline{\theta}; \mathcal{D}) = \log p(\mathcal{D} \mid \underline{\theta})$$

$$= \sum_{n=1}^{N} \sum_{\underline{x}_V} \delta(\underline{x}_V, \underline{x}_V, n) \log p(\underline{x}_V \mid \underline{\theta})$$

$$= \sum_{\underline{x}_V} \sum_{n} \delta(\underline{x}_V, \underline{x}_V, n) \log p(\underline{x}_V \mid \underline{\theta})$$

factor out terms without sum. arg.

$$= \sum_{\underline{x}_V} m(\underline{x}_V) \log p(\underline{x}_V \mid \underline{\theta})$$

subs. $p(\underline{x}_V \mid \underline{\theta})$

$$= \sum_{\underline{x}_V} m(\underline{x}_V) \log \left(\frac{1}{Z} \prod_{C} \psi_C(\underline{x}_C) \right)$$

$$= \sum_{\underline{x}_V} m(\underline{x}_V) \sum_{C} \log \psi_C(\underline{x}_C) - \sum_{\underline{x}_V} m(\underline{x}_V) \log Z$$

(?) $\mathcal{L} = \sum_c \sum_{x_c} m(x_c) \log \psi_c(x_c) - N \log Z$ (*) (9.43)

- name me (Q) (0/5 1)
- see earlier 9.2
- Assume true and investigate

- $m(x_c)$ - neg. counts → suff. statistics
- $N \log Z$ - appears (not in DGM)

- MLE of LL:

(*) $N \log Z$ - coupled, nonlinear set of eq. with implicit perm app.

$$\ell(\theta; D) = \sum_c \sum_{x_c} m(x_c) \log \psi_c(x_c) - N \log Z \qquad (9.43)$$

(i) (ii)

(*) Partial deriv wrt $\psi_c(x_c)$; with clique C, config x_c held fixed.

- (i) $\frac{\partial}{\partial \psi_c(x_c)} m(x_c) \log \psi_c(x_c) = \frac{m(x_c)}{\psi_c(x_c)}$

- Note that D and $\hat{x}$ are just diff dummy indexing variables

(ii) $$\frac{\partial}{\partial \psi_c(x_c)} \log Z = \frac{1}{Z} \frac{\partial}{\partial \psi_c(x_c)} \left\{ \sum_{\hat{x}} \prod_D \psi_D(\hat{x}_D) \right\}$$

$$= \frac{1}{Z} \sum_{\hat{x}} \delta(\hat{x}_c, x_c) \frac{\partial}{\partial \psi_c(x_c)} \left(\prod_D \psi_D(\hat{x}_D) \right)$$

$$= \frac{1}{Z} \sum_{\hat{x}} \delta(\hat{x}_c, x_c) \prod_{D \neq C} \psi_D(\hat{x}_D)$$

- Note all $\psi_D(\hat{x}_D)$ where $D \neq C$ are assumed constant wrt $\psi_c(x_c)$

$$= \sum_{\hat{x}} \delta(\hat{x}_c, x_c) \frac{1}{\psi_c(\hat{x}_c)} \frac{1}{Z} \prod_D \psi_D(\hat{x}_D)$$

(9.2)

- not entirely clear crystal

$$= \frac{1}{\psi_c(\hat{x}_c)} \sum_{\hat{x}} \delta(\hat{x}_c, x_c) p(\hat{x})$$

$$= \frac{p_c(x_c)}{\psi_c(x_c)}$$

Yielding $$\frac{\partial \ell}{\partial \psi_c(x_c)} = \frac{m(x_c)}{\psi_c(x_c)} - N \frac{p(x_c)}{\psi_c(x_c)}$$

w/ $\log \psi_c(x_c) > 0$

- set $\frac{\partial \mathcal{L}}{\partial \psi_c(x_c)} = 0 \Rightarrow m(x_c) = N p(x_c)$

(*) define empirical distn $\hat{p}(x) = \frac{m(x)}{N}$, $\hat{p}(x_c) = \frac{m(x_c)}{N}$ is a marg. wrt emp. (?)

- $\Rightarrow \frac{m(x_c)}{N} = p(x_c) \Rightarrow \hat{p}_{ML}(x_c) = p(x_c)$

(*) For each clique $c \in C$, the model marginals $p(x_c)$ must be equal to empirical marginals $\hat{p}_{ML}(x_c)$

Jordan 2003 9.3.3. → obt MLE by inspection for decomposable graphs

---

- (*) Iterative prop. fitting (IPF)

- IPF is not only a fixed point algo, but coordinate ascent algo

(*) use:- $\frac{\tilde{p}(x_c)}{\psi_c(x_c)} = \frac{p(x_c)}{\psi_c(x_c)}$

- Hold $\psi_c(x_c)$ fixed in numerator, denom of RHS.
- ~~As~~
- Why? As $\psi_c(x_c)$ appears implicitly through $p(x_c)$
- Introduce the updating parameter at iter t $\psi_c^{(t)}(x_c)$
- Joint prob based on estimates $p^{(t)}(x)$ at iter t

$\Rightarrow \frac{\psi_c(x_c)}{\hat{p}(x_c)} = \frac{\psi_c(x_c)}{p(x_c)}$

$\Rightarrow \psi_c(x_c) = \psi_c(x_c) \frac{\hat{p}(x_c)}{p(x_c)}$

$\Rightarrow \psi_c^{(t+1)}(x_c) = \psi_c^{(t)}(x_c) \frac{\hat{p}(x_c)}{p^{(t)}(x_c)}$ (9.61) (IPF algorithm)

(*) Cycle through all cliques $c \in C$, applying (9.61); one cycle - one iter.

(*) Properties of IPF

- Summarised from Jordan (2003) ch 9.3.

1) Marginal $p^{(t+1)}(x_c)$ is equal to empirical marginal $\tilde{p}(x_c)$
2) Normalisation constant Z - constant across updates

which yields: $p^{(t+1)}(x_V) = p^{(t)}(x_{V\setminus C}|x_C)\hat{p}(x_C)$

- each IPF iteration retains previous cond. prob $p^{(t)}(x_{V\setminus C}|x_C)$ and replaces prev. marg prob $p^{(t)}(x_C)$ with new marginal $\hat{p}(x_C)$.

(*) IPF as co-ordinate ascent

- Additional derivations in Jordan show this sets gradient of log likelihood to 0; and is coordinate ascent (in log-likelihood)

(*) KL divergence view of IPF (A5)

- A few key equations:

KL-divergence with marginal:

- Decompose $p(x_A, x_B) = p(x_B|x_A)p(x_A)$ and $q(x_A, x_B) = q(x_B|x_A)q(x_A)$

- (A) KL div: $D(p(x_A,x_B)\|q(x_A,x_B)) = D(p(x_A)\|q(x_A)) + \sum_{x_A} p(x_A) D(p(x_B|x_A)\|q(x_B|x_A))$

(*)(☆): maximising likelihood equivalent to minimising following KL-div:-

$$D(\hat{p}(x)\|p(x|\theta)) = \sum_x \hat{p}(x) \log \frac{\hat{p}(x)}{p(x|\theta)}$$

$\hat{p}(x)$ - empir. distn.

(*) Equivalence ⟶ -ve log likelihood and KL div differ by $\sum_x \hat{p}(x)\log\hat{p}(x)$ which is independent of θ

(*) coordinate descent in $D(\hat{p}(x)\|p(x|\theta))$ (coordinates - parameters of clique potentials)

- Pick a clique C
- Adjust clique potential $\psi_C(x_C)$ to min KL-div. (☆)

$$\Rightarrow D(\hat{p}(x)\|p(x|\theta)) = \underbrace{D(\hat{p}(x_C)\|p(x_C|\theta))}_{(I)} + \sum_{x_C} \hat{p}(x_C) \underbrace{D(\tilde{p}(x_{V\setminus C}|x_C)\|p(x_{V\setminus C}|x_C,\theta))}_{(II)}$$

(☆): A little abuse; lecture also helps with exposition.

- changes in clique pot $\psi_C(x_C)$ does not affect $p(x_{V\setminus C}|x_C,\theta)$

⇒ (II) unaffected by changes to $\psi_C(x_C)$ and minimising KL div $D(\hat{p}(x)\|p(x|\theta))$ amounts to minimising (I)

(*) Minimising (I) achieved by setting $p(x_c|\theta) = \hat{p}(x_c)$ i.e. <u>marginal</u> to <u>empirical marginal</u>

(*) This what IPF achieves → (SII) coordinate ascent on log-like coordinate descent on KL-div

(**): Note IPF takes form of scaling algorithm in which <u>potentials</u> are multiplied by a <u>ratio of marginals</u>.

(**) $\max l \iff \min KL(\hat{p}(x) \| p(x|\theta)) = \sum_x \hat{p}(x) \log \frac{\hat{p}(x)}{p(x|\theta)}$ (*)

(*) Verifying identity (A):- (KL div. decomp.) (A6)

$$D(p(x_A, x_B) \| q(x_A, x_B)) = \sum_{x_A, x_B} p(x_A) p(x_B|x_A) \log \frac{p(x_A) p(x_B|x_A)}{q(x_A) q(x_B|x_A)}$$

$$= \sum_{x_A, x_B} p(x_A) p(x_B|x_A) \log \frac{p(x_A)}{q(x_A)} + \sum_{x_A, x_B} p(x_A) p(x_B|x_A) \log \frac{p(x_B|x_A)}{q(x_B|x_A)}$$

$$= \sum_{x_A, x_B} p(x_A, x_B) \log \frac{p(x_A)}{q(x_A)} + \sum_{x_A} p(x_A) \sum_{x_B} p(x_B|x_A) \log \frac{p(x_B|x_A)}{q(x_B|x_A)}$$

$$= \sum_{x_A} \log \frac{p(x_A)}{q(x_A)} \sum_{x_B} p(x_A, x_B) + \sum_{x_A} p(x_A) \sum_{x_B} p(x_B|x_A) \log \frac{p(x_B|x_A)}{q(x_B|x_A)}$$

$$= \sum_{x_A} p(x_A) \log \frac{p(x_A)}{q(x_A)} + \sum_{x_A} p(x_A) D(p(x_B|x_A) \| q(x_B|x_A))$$

$$= D(p(x_A) \| q(x_A)) + \sum_{x_A} p(x_A) D(p(x_B|x_A) \| q(x_B|x_A))$$

(*) Features and micropotentials

(N): lecture notes (S2019) are a little clearer on feature design.

(SII): Rather than define a table of values to encompass the mapping of $\psi_c(x_c) = \psi_c(x_1, x_2, x_3)$ on e.g. 26^3 possibilities

(VII) we instead use domain knowledge (e.g. a vocabulary/dictionary) to reduce representational granularity

(VII) Achieved by assigning a score to common three letter stems; or scores to 'significant' occurrences e.g. fing = 10, freq = 9 etc. (X)

(VII) Remainder → assign an arbitrarily low score to reflect low likelihood of occurrence.

(VII) Score in which 'feature' is a function which is 'vacuous' over all joint settings except a few part. ones.

(?) or use indicators?
- to avoid simult. estim. / setting θ_R AND f_R

Mathematical details

- Define K features $f_R(c_1, c_2, c_3)$ e.g. $\theta_{ing} f_{ing}$
- Assign each of K features a weight θ_R
- A micropotential is then achieved by exponentiating $e^{\theta_R f_R(c_1,c_2,c_3)}$ (indiv.)
- (*) A clique potential is formed by multiplying together micropotentials

- Yielding: $\psi_c(x_c) = \psi_c(x_1, x_2, x_3) = e^{\theta_1 f_1} \cdot e^{\theta_2 f_2} \cdot \ldots \cdot e^{\theta_K f_K}$

$$= \exp\left\{\sum_{R=1}^{K} \theta_R f_R\right\}$$

- Potential is still over 26^3 settings, but only have K parameters if K features are used
- (VII) we then can estimate the weights θ_R to deal with context info
- (X) θ_R - 'strength' of feature and whether it increases or decreases clique probability

---

(*) Combining features

- Jordan (2003) → suggests f_R is chosen to be indicator
- Marginal probability over a clique: $p(c_1, c_2, c_3) \propto \exp\left\{\sum_{R=1}^{K} \theta_R f_R\right\}$
- Addit. complexities → overlapping features, memory model, any function of subset of clique variables, overlapping windows do not alter anatomy

(IX) clique potential as weighted sum of exponen. features.

$$\psi_c(\underline{x}_c) = \exp\left\{\sum_{i \in \mathcal{I}_c} \theta_k f_k(\underline{x}_{c_i})\right\}$$

(X) Features based model

- usually, $p(\underline{x}_v|\underline{\theta}) = \frac{1}{Z(\underline{\theta})} \prod_c \psi_c(\underline{x}_c) = \frac{1}{Z(\underline{\theta})} \exp\left\{\sum_c \sum_{i \in \mathcal{I}_c} \theta_k f_k(\underline{x}_{c_i})\right\}$

- But drop explicit assoc. of features and cliques (??) and use

$$p(\underline{x}|\underline{\theta}) = \frac{1}{Z(\underline{\theta})} \exp\left\{\sum_i \theta_i f_i(\underline{x})\right\}$$

where $f_i(\cdot)$ - features
θ_i - param
$Z(\theta)$ - norm. factor

(XI) - Exponential family model with features as sufficient statistics

$$Z(\underline{\theta}) = \sum_{\underline{x}} \exp\left\{\sum_i \theta_i f_i(\underline{x})\right\}$$

(XII): Estimate θ_i from data $\mathcal{D}$. (param)

MLE of feature based UGMs: (see Jordan 2003) (problem setup)

- ~~IX~~ (X)

- Scaled log-likelihood: $\tilde{\ell}(\underline{\theta}; \mathcal{D}) = \frac{\ell(\underline{\theta}; \mathcal{D})}{N} = \frac{1}{N} \sum_{n=1}^{N} \log p(\underline{x}_n|\underline{\theta})$ (Q (?)

(X) $\tilde{\ell}(\underline{\theta}; \mathcal{D}) = \sum_{\underline{x}} \hat{p}(\underline{x}) \log p(\underline{x}|\underline{\theta})$

$$= \sum_{\underline{x}} \hat{p}(\underline{x}) \log\left[\frac{1}{Z(\underline{\theta})} \left\{\exp \sum_i \theta_i f_i(\underline{x})\right\}\right]$$

$$= \sum_{\underline{x}} \hat{p}(\underline{x}) \left\{\sum_i \theta_i f_i(\underline{x})\right\} - \log Z(\underline{\theta})$$

- Jordan (2003) states we use convexity of $\log(\cdot)$ to bound $\log z(\underline{\theta})$ (?)
- we then get:-

$$\log z(\underline{\theta}) \leq \mu z(\underline{\theta}) - \log \mu - 1$$

(0/53) $\log(\cdot)$ is not convex; it's concave

(0/54) : what conseq for presentation?

(*) for now, assume this is okay → add to overspill and investigate at the end.

(*) Bound holds for all μ and $\mu = z^{-1}(\underline{\theta}^{(t)})$

$$\Rightarrow \hat{\ell}(\underline{\theta}|D) \geq \sum_{\underline{x}} \tilde{p}(\underline{x}) \sum_i \theta_i f_i(\underline{x}) - \frac{z(\underline{\theta})}{z(\underline{\theta}^{(t)})} - \log z(\underline{\theta}^{(t)}) + 1$$

(0/55)

with equality at $z(\underline{\theta}^{(t)})$

(*) further manipulation of scaled log-line.

- define $\Delta\theta_i^{(t)} := \theta_i - \theta_i^{(t)}$, then (subs. $z(\underline{\theta})$)

$$\hat{\ell}(\underline{\theta}|D) \geq \sum_{\underline{x}} \tilde{p}(\underline{x}) \sum_i \theta_i f_i(\underline{x}) - \frac{1}{z(\underline{\theta}^{(t)})} \sum_{\underline{x}} \exp\left\{\sum_i \theta_i f_i(\underline{x})\right\} - \log z(\underline{\theta}^{(t)}) + 1$$

$$= \sum_{\underline{x}} \tilde{p}(\underline{x}) \sum_i \theta_i f_i(\underline{x}) - \frac{1}{z(\underline{\theta}^{(t)})} \sum_{\underline{x}} \exp\left\{\sum_i (\Delta\theta_i^{(t)} + \theta_i^{(t)}) f_i(\underline{x})\right\} - \log z(\underline{\theta}^{(t)}) + 1$$

$$= \sum_{\underline{x}} \tilde{p}(\underline{x}) \sum_i \theta_i f_i(\underline{x}) - \frac{1}{z(\underline{\theta}^{(t)})} \sum_{\underline{x}} \exp\left\{\sum_i \theta_i^{(t)} f_i(\underline{x})\right\} \exp\left\{\sum_i \Delta\theta_i^{(t)} f_i(\underline{x})\right\} - \log z(\underline{\theta}^{(t)}) + 1$$

$$= \sum_{\underline{x}} \tilde{p}(\underline{x}) \sum_i \theta_i f_i(\underline{x}) - \sum_{\underline{x}} \underbrace{\frac{1}{z(\underline{\theta}^{(t)})} \exp\left\{\sum_i \theta_i^{(t)} f_i(\underline{x})\right\}}_{= p(\underline{x}|\underline{\theta}^{(t)})} \exp\left\{\sum_i \Delta\theta_i^{(t)} f_i(\underline{x})\right\} - \log z(\underline{\theta}^{(t)}) + 1$$

$$= \sum_i \theta_i \sum_{\underline{x}} \tilde{p}(\underline{x}) f_i(\underline{x}) - \sum_{\underline{x}} p(\underline{x}|\underline{\theta}^{(t)}) \exp\left\{\sum_i \Delta\theta_i^{(t)} f_i(\underline{x})\right\} - \log z(\underline{\theta}^{(t)}) + 1$$

coupled $f_i(\underline{x})$ and $\Delta\theta_i^{(t)}$

assume $f_i(\underline{x}) \geq 0$ and $\sum_i f_i(\underline{x}) = 1$

* $e(\cdot)$ is convex, invoke Jensen's inequal.

(0/55a)

$$\exp\left(\sum_i \pi_i x_i\right) \leq \sum_i \pi_i \exp(x_i) \quad \text{for } \sum_i \pi_i = 1$$

(*) Note; f_i play the role of π_i as they are positive and sum to 1.

(*) cross-ref with lecture 7 notes

↳ f_i are being treated as 'weights' and $\Delta\theta_i^{(t)}$ as arguments

Yielding the following lower bound on scaled log-likelihood:- (with params decoupled)

$$\hat{\ell}(\underline{\theta};D) \geq \sum_i \theta_i \sum_{\underline{x}} \hat{p}(\underline{x}) f_i(\underline{x}) - \sum_{\underline{x}} p(\underline{x}|\underline{\theta}^{(t)}) \sum_i f_i(\underline{x}) \exp(\Delta\theta_i^{(t)}) - \log z(\underline{\theta}^{(t)}) + 1$$

$$= \Lambda(\underline{\theta})$$

- we then take derivatives with respect to low bound $\Lambda(\underline{\theta})$:- (0156)

$$\frac{\partial \Lambda}{\partial \theta_i} = \sum_{\underline{x}} \hat{p}(\underline{x}) f_i(\underline{x}) - \exp(\Delta\theta_i^{(t)}) \sum_{\underline{x}} p(\underline{x}|\underline{\theta}^{(t)}) f_i(\underline{x}) = 0$$

$$\Rightarrow \exp(\Delta\theta_i^{(t)}) = \frac{\sum_{\underline{x}} \hat{p}(\underline{x}) f_i(\underline{x})}{\sum_{\underline{x}} p(\underline{x}|\underline{\theta}^{(t)}) f_i(\underline{x})}$$

(0157)

- defining (or because?) $p^t(\underline{x})$ is an unnormalised version of $p(\underline{x}|\underline{\theta}^{(t)})$

$$\Rightarrow \exp(\Delta\theta_i^{(t)}) = \frac{\sum_{\underline{x}} \hat{p}(\underline{x}) f_i(\underline{x})}{\sum_{\underline{x}} p^t(\underline{x}) f_i(\underline{x})} z(\underline{\theta}^{(t)})$$

(?): next bit on Jordan a xing not clear from recon; attempted to clarify (lecture notes + slides)

(*) $\theta_i^{(t+1)} = \theta_i^{(t)} + \Delta\theta_i^{(t)} \Rightarrow p^{(t+1)}(\underline{x}) = p^{(t)}(\underline{x}) \left(\prod_i e^{\Delta\theta_i^{(t)} f_i(\underline{x})} \right)$ (?) (0157)

Jordan (2003):

(*) To update parameters from $\underline{\theta}^{(t)}$ to $\underline{\theta}^{(t+1)}$, multiply $p(\underline{x}|\underline{\theta}^{(t)})$ by $e^{\Delta\theta_i^{(t)} f_i(\underline{x})}$ $\forall i$

(*) Note: $\frac{p^{(t)}(\underline{x})}{z(\underline{\theta}^{(t)})} = p(\underline{x}|\underline{\theta}^{(t)})$ and →

(*) we have:-

$$p^{(t+1)}(\underline{x}) = p(\underline{x}|\underline{\theta}^{(t)}) \left(\prod_i e^{\Delta\theta_i^{(t)}} f_i(\underline{x}) \right)$$

- subs for $p(\underline{x}|\underline{\theta}^{(t)})$ and $e^{\Delta\theta_i^{(t)}}$

$$= \frac{p^{(t)}(\underline{x})}{z(\underline{\theta}^{(t)})} \prod_i \left(\frac{\sum_{\underline{x}} \hat{p}(\underline{x}) f_i(\underline{x})}{\sum_{\underline{x}} p^{(t)}(\underline{x}) f_i(\underline{x})} z(\underline{\theta}^{(t)}) \right)^{f_i(\underline{x})}$$

$$= \frac{p^{(t)}(\underline{x})}{z(\underline{\theta}^{(t)})} \prod_i \left(\frac{\sum_{\underline{x}} \hat{p}(\underline{x}) f_i(\underline{x})}{\sum_{\underline{x}} p^{(t)}(\underline{x}) f_i(\underline{x})} \right)^{f_i(\underline{x})} z(\underline{\theta}^{(t)})^{\sum_i f_i(\underline{x})}$$

As $\sum_i f_i(\underline{x}) = 1$

$$= p^{(t)}(\underline{x}) \prod_i \left(\frac{\sum_{\underline{x}} \hat{p}(\underline{x}) f_i(\underline{x})}{\sum_{\underline{x}} p^{(t)}(\underline{x}) f_i(\underline{x})} \right)^{f_i(\underline{x})}$$

(GIS Algorithm)

(*) Exponential family, MLE, sufficient statistics (**) - import. for understanding

- prop v subscript notion of empirical marginals
- substitute feat. rep of joint prob. (condit. or param)

$$l(\underline{\theta}; D) = \sum_{\underline{x}} m(\underline{x}) \log p(\underline{x}|\underline{\theta})$$

$$= \sum_{\underline{x}} m(\underline{x}) \log \left[\frac{1}{z(\underline{\theta})} \exp\left\{ \sum_i \theta_i f_i(\underline{x}) \right\} \right]$$

$$= \sum_{\underline{x}} m(\underline{x}) \left\{ \sum_i \theta_i f_i(\underline{x}) - \log z(\underline{\theta}) \right\}$$

$$= \sum_{\underline{x}} m(\underline{x}) \sum_i \theta_i f_i(\underline{x}) - \log z(\underline{\theta}) \sum_{\underline{x}} m(\underline{x})$$

- $\sum_{\underline{x}} m(\underline{x}) = N$

$$= \sum_{\underline{x}} m(\underline{x}) \sum_i \theta_i f_i(\underline{x}) - N \log z(\underline{\theta})$$

(*) Taking derivatives wrt θ_i, setting to 0, i.e. going for an MLE est. :- ↓ (0158)

$$\frac{\partial}{\partial \theta_i} l(\underline{\theta}; D) = \sum_{\underline{x}} m(\underline{x}) f_i(\underline{x}) - N \frac{\partial}{\partial \theta_i} \log z(\underline{\theta}) \quad (*)$$

↓ (0159) (*)

$$= \sum_{\underline{x}} m(\underline{x}) f_i(\underline{x}) - N \sum_{\underline{x}} p(\underline{x}|\underline{\theta}) f_i(\underline{x}) = 0$$

$$\Rightarrow \sum_{\underline{x}} p(\underline{x}|\underline{\theta}) f_i(\underline{x}) = \sum_{\underline{x}} \frac{m(\underline{x})}{N} f_i(\underline{x}) = \sum_{\underline{x}} \hat{p}(\underline{x}|\underline{\theta}) f_i(\underline{x})$$

(*) At ML estimate:-

- Expectations of the sufficient statistics under the model must match the empirical feature average

- Presumably former is $\sum_{\underline{x}} p(\underline{x}|\underline{\theta}) f_i(\underline{x})$; and latter is $\sum_{\underline{x}} \hat{p}(\underline{x}|\underline{\theta}) f_i(\underline{x})$.

(*)(☆)- Information theoretic and statistical physics pov. in ML

- Have to review (LS) on exponentials and GLMS

(☆)- Links with 36-705:

- Sufficiency, exponentials

Pitman-Koopman-Darmois Theorem

(*) "Among families of probability distributions whose domain does not vary with the parameter being estimated, only in exponential families is there a sufficient statistic whose dimension remains bounded as sample size ↑"

(*) "Sufficiency simply restricts possible forms of distri"

(w): Leads to some interesting discussions by Persi Diaconis

(*) Info-theory / statistical physics principles in ML

- Ex makes point for modelling:- (and to why exponential family comes up)

- (w)(☆) To view exponential family as a solution to a constrained, variational optimn. problem in which objective is entropy or KL divergence; and constraints are that expectations under the distri are matched to expectations under the empirical distri.

i) variational in the sense of min/max a functional by choosing fn (distn)

ii) ~~obj~~

- Ex: Begin with fixed feature exp. $\sum_{\underline{x}} p(\underline{x}) f_i(\underline{x}) = \alpha_i$

(*) Assuming consistent exp; choose a distri:-

$$\max_p H(p(\underline{x})) = -\sum_{\underline{x}} p(\underline{x}) \log p(\underline{x})$$

$$\text{s.t. } \sum_{\underline{x}} p(\underline{x}) f_i(\underline{x}) = \alpha_i \quad \Rightarrow p^*(\underline{x}|\underline{\theta}) = \frac{1}{Z(\underline{\theta})} \exp\left\{\sum_i \theta_i f_i(\underline{x})\right\}$$

$$\sum_{\underline{x}} p(\underline{x}) = 1$$

(*) more generally;

$$\min_p KL(p(\underline{x}) \| h(\underline{x})) = \min_p \sum_{\underline{x}} p(\underline{x}) \log \frac{p(\underline{x})}{h(\underline{x})} = \min_p -H(p) - \sum_{\underline{x}} p(\underline{x}) \log h(\underline{x})$$

$$\text{s.t.} \sum_{\underline{x}} p(\underline{x}) f_i(\underline{x}) = \alpha_i$$

$$\sum_{\underline{x}} p(\underline{x}) = 1$$

$$\Rightarrow p^*(\underline{x}|\underline{\theta}) = \frac{1}{Z(\underline{\theta})} h(\underline{x}) \exp\left\{\sum_i \theta_i f_i(\underline{x})\right\}$$

(*) interpretation

(*) maximum entropy principle:-

- From amongst distributions consistent with the data, select the distribution p^* whose shannon entropy is maximal
- Amounts to choosing distn with maximal uncertainty, as defined by the entropy functional

(*) for KL divergence

- incorporates prior $h(\underline{x})$
- choose distn that contains 'least additional' above priors.

(*) Additional details on

info-theory – stat mech → Jaynes (1967)

Jordan (2008) - Foundations & Trends ch 3.