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

(A1)

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

Joint prob: $p(x_V|\underline{\theta}) = \frac{1}{Z} \prod_C \psi_C(x_C)$

$\underline{\theta} = \{\psi_C(x_C), C \in \mathcal{C}\}$ (param. collec.)

Z = normalisation

$$Z = \sum_{x_V} \prod_C \psi_C(x_C)$$

(*): Z - obtained via summing (or integrating) over all configurations x_V

The total counts

obs. no. of times that configuration x_V is observed in a dataset $\mathcal{D}$

$$m(x_V) := \sum_{n=1}^{N} \delta(x_V, x_{V,n})$$

Marginal counts (clique counts) for clique C

$$m(x_C) := \sum_{x_{V \setminus C}} m(x_V)$$

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

---

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

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

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

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

- OK - 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:-

$$p(D|\underline{\theta}) = \prod_{n=1}^{N} p(\underline{x}_{V,n} | \underline{\theta})$$

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

log-likelihood in terms of marginal counts

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

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

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

factor out terms without sum. arg.

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

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

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

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

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

- nuance mle
- see earlier 9.2
- Assume true and investigate

@ 0/S 1

· $m(x_c)$ - neg. counts → suff. statistics

· $N \log Z$ - appears (not in DGM)

- MLE of ll:

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

$$l(\theta; D) = \sum_C \sum_{x_C} m(x_C) \log \underbrace{\psi_C(x_C)}_{(i)} - N \underbrace{\log Z}_{(ii)} \qquad (9.43)$$

(*) 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 $\psi_C(x_C)$

$$= \sum_{\hat{x}} \delta(\hat{x}_C, x_C) \frac{1}{\psi_C(\hat{x}_C)} \frac{1}{Z} \prod_D \psi_D(\hat{x}_D)$$

Q S 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 $$\frac{\partial l}{\partial \psi_C(x_C)} = \frac{m(x_C)}{\psi_C(x_C)} - N \frac{p(x_C)}{\psi_C(x_C)}$$

w/ $\log \psi_C(x_C) > 0$

- set $\frac{\partial l}{\partial \psi_c(\underline{x}_c)} = 0 \Rightarrow m(\underline{x}_c) - N p(\underline{x}_c)$

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

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

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

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

---

- (*) Iterative prop. fitting (IPF)
- IPF is not only a fixed point algo, but coordinate ascent algo

(*) use:- $\frac{\hat{p}(\underline{x}_c)}{\psi_c(\underline{x}_c)} = \frac{p(\underline{x}_c)}{\psi_c(\underline{x}_c)}$

$\Rightarrow \frac{\psi_c(\underline{x}_c)}{\hat{p}(\underline{x}_c)} = \frac{\psi_c(\underline{x}_c)}{p(\underline{x}_c)}$

$\Rightarrow \psi_c(\underline{x}_c) = \psi_c(\underline{x}_c) \frac{\hat{p}(\underline{x}_c)}{p(\underline{x}_c)}$

- Hold $\psi_c(\underline{x}_c)$ fixed in ~~As~~ numerator, denom of RHS.
- why? As $\psi_c(\underline{x}_c)$ appears implicitly through $p(\underline{x}_c)$
- introduce time indexing;
  - params at iter t $\psi_c^{t}(\underline{x}_c)$
  - joint prob based on estimates $p^{(t)}(\underline{x})$ at iter t

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

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

(*) Properties of IPF

- Summarised from Jordan (2003) ch 9.3.

1) marginal $p^{(t+1)}(\underline{x}_c)$ is equal to empirical marginal $\tilde{p}(\underline{x}_c)$

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 marg/cond:

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

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

(*)(B): 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 $\longrightarrow$ -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 course; lecture also helps with exposition.

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

$\Rightarrow$ (II) unaltered 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. marginal to empirical marginal

(*) This what IPF achieves $\longrightarrow$ (511) coordinate ascent on log-like
coordinate descent on KL-div

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

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

(A6)

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

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

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

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

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

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

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

(*) Features and micropotentials

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

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

(5T) we instead use domain knowledge (e.g. a vocabulary/dictionary) to reduce representational granularity

(5T) Achieved by assigning a score to common three letter steams; or scores to 'significant' occurences e.g. fing = 10, fted = 9 etc. (X)

(5T) remainder $\rightarrow$ assign an arbitrarily low score to reflect low likelihood of occurrence.

(5T) scale 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. $\theta_{ingfing}$
- Assign each of K features a weight θ_R

- A indiv. micropotential is then achieved by exponentiating $e^{\theta_R f_R(c_1,c_2,c_3)}$

(*) 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

(5T) 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) $\rightarrow$ 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 $\longrightarrow$ overlapping features, unconstr. model; any function of subset of clique variables, overlapping windows do not alter anatomy

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

$$\psi_c(\underline{x}_c) \overset{\text{def}}{=} \exp\left\{\sum_{i \in I_c} \theta_k f_k(\underline{x}_{c_i})\right\}$$

(*) Features based model

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

- 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
$Z(\underline{\theta})$ - norm. factor

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

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

(*): Estimate θ_i from data D.
(param)

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

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

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

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

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

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

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

O/S 3: log(·) is not convex; it's concave

O/S 4: what conseq for presentation?

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

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

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

O/S 5

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. $Z(\underline{\theta})$)

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

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

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

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

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

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

As e(·) is convex, invoke Jensen's inequal.

O/S 5a

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

(*) 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)

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

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

· we then take derivatives with respect to lower bound $\Lambda(\underline{\theta})$:–

0/56

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

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

0/57

· 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 in Jordan or Xing not clear (lecture notes + slides) from recon; attempted to clarify

(*) $\theta_i^{(t+1)} = \theta_i^{(t)} + \Delta\theta_i^{(t)} \Rightarrow p^{(t+1)}(\underline{x}) = p^{(t)}\underline{x}\left(\prod_i e^{\Delta\theta_i^{(t)} f_i(\underline{x})}\right)$ (?) 0/57

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 →