(*) we have:-

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

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

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

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

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

-Drop v subscript

notat. of empirical marginals

-substitute feat. rep of joint prob. (condit. on param)

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

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

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

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

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

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

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

0/58

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

0/59

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

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

(*) At ML estimate:-

- Expectations of the sufficient statistics under the model must match the empirical feature average
- Presumably former is $\sum_{x} p(x|\theta) f_i(x)$; and latter is $\sum_{x} \hat{p}(x|\theta) f_i(x)$.

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

- Have to review (LS) on exponentials and GLMS

(*) - links with 36-705:

- sufficiency, exponentials

Pittman-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 sharply restricts possible forms of distn"

(i): leads to some interesting discussions by Persi Diaconis

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

(and to why exponential family comes up)

- EX makes point for modelling:-
- (*) (W) Can view exponential family as a solution to a constrained, variational optims. problem in which objective is entropy or KL divergence, and constraints are that expectations under the distn are matched to expectations under the empirical distn.

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

ii) ~~obj~~

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

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

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

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

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

(*) more generally,

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

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 distr p^* whose Shannon entropy is maximal

Amounts to choosing distr with maximal uncertainty, as defined by the entropy functional

(*) For KL divergence

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

(*) Additional details on info-theory - stat mech → Jaynes (1967)

Jordan (2008) - Foundations & Trends ch 3.