(*) Both imply a factorisation of $p(x|\theta)$

(*) Neyman factorisation $\longrightarrow$ frequentist definition of sufficiency

(*) Sufficiency in this context means $T(x)$ is sufficient for $\theta$ if:-

$$\theta \perp\!\!\!\perp x \mid T(x)$$

---

Jordan (2003)

- 8.1.8. ML and KL divergence

- A general rel. between ML and KL divergence (not spec. to exp.)
- Necessary for late lec. material 16, 17
- Statistical interp of KL divergence to illus. rel. between KL and exp. family

(*) Empirical distn: $\hat{p}(\underline{x})$

- Places a point mass at each data point $\underline{x}_n$ in $\mathcal{D}$ (dataset)

(discrete)

(*) Empirical distn: $\hat{p}(\underline{x}) := \frac{1}{N}\sum_{n=1}^{N} \delta(\underline{x}, \underline{x}_n) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(\underline{x} = \underline{x}_n)$

(*) Sum/integrate $\hat{p}(\underline{x})$ against a function of $\underline{x}$, i.e. evaluate at each point $\underline{x}_n$

$\delta$ - Kronecker delta in cont. case.

(*) Log likelihood: (also, cross entropy of $\hat{p}(\underline{x})$ and $p(\underline{x}|\underline{\theta})$)

scaled

$$\sum_{\underline{x}} \hat{p}(\underline{x}) \log p(\underline{x}|\underline{\theta}) = \sum_{\underline{x}} \left( \frac{1}{N}\sum_{n=1}^{N} \delta(\underline{x}, \underline{x}_n) \log p(\underline{x}_n|\underline{\theta}) \right)$$

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

✓ de-nickifying

$$= \frac{1}{N}\sum_{n=1}^{N} \log p(\underline{x}_n|\underline{\theta})$$

$$= \frac{1}{N} l(\underline{\theta}|\mathcal{D})$$

(*) Note; the scaled log likelihood (by factor $\frac{1}{N}$) is equivalent to the
(IX) cross-entropy between the empirical distri and the model
$(\hat{p}(\underline{x}))$ $\&$ $p(\underline{x}|\underline{\theta})$

- see result for continuous

- KL divergence between empirical; model :-

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

$$= \sum_{\underline{x}} \hat{p}(\underline{x}) \log \hat{p}(\underline{x}) - \sum_{\underline{x}} \hat{p}(\underline{x}) \log p(\underline{x}|\underline{\theta})$$

$$= \sum_{\underline{x}} \hat{p}(\underline{x}) \log \hat{p}(\underline{x}) - \frac{1}{N} \ell(\underline{\theta}|D)$$

(I) (II)

(I) - independent of $\underline{\theta}$

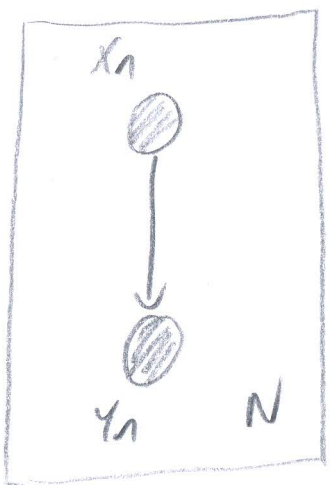
- value of $\underline{\theta}$ that minimises LHS is the value of $\underline{\theta}$ that maximises the RHS

(X): minimising KL divergence between the empirical distri and model distri is equivalent to maximising the likelihood

---

(*) Generalised linear models (GLM) - linear regression/classificatn

* covers linear reg. / discriminative linear classification.




(*) Both LR/LC → both assume a rep. for conditional expectation of Y.

(*) $\mu = f(\underline{\theta}^T \underline{x}) = \mathbb{E}_{y \sim p(y|f(x))}[y]$

(*) LR: $f(\cdot)$ - identity
LC: $f(\cdot)$ - sigmoid (logistic)

(*) Also: endow $Y$ with a particular cond. prob. distribution, with $\mu$ as a parameter.

(W): Remember JP → ColumbiaX ML (prob. intro of ML for LR!)

(*) LR - Gaussian LC - Bernoulli / Multinomial.

---

(*) generalised linear model Framework

- 3 assumptions on $p(y|\underline{x})$ :-

1. observed input $\underline{x}$ enters into model via linear comb. $\xi = \underline{\theta}^T \underline{x}$
2. conditional mean $\mu$ rep as a function $f(\xi)$ of the linear combination $\xi$ where $f$ is known as the response function
3. observed output $y$ is assumed to be characterised by exp. family with conditional mean $\mu$.

(*)