Multinomial example

- constrained parameters in Multinomial (degrees of freedom) [sum to unity]
- Hence (K-1) in summation

Exponential family rep.

Q: Why do we go for exponential family reps?

① - Data and parameters clearly grouped into 2 terms — through η / through $T(\cdot)$

② Moment generating function.

$$\frac{dA}{d\eta} = \mathbb{E}[T(x)] \qquad \frac{d^2A}{d\eta^2} = Var[T(x)]$$

- Gives standard operator that yield n^{th} order moments (from derivatives of log-normaliser)

EX: Moments important → characterise

③: Relationship between moment and natural parameters

(*) key - Review moment and canonical param relation.

MLE for exponential family

- only differences between distri in exp. family are form of η and $T(\cdot)$ i.e. canonical param and suffic. stat.
- IID data → log-likelihood → optimise (set 1^{st} order moment to 0)
- moment matching

EX: Exponential family exposes the relationship between data (through $T(\cdot)$) and parameters ~~through~~ (through η) (in a linearly dependent fashion)

- Gives info about transformations of data or forms of data we need to worry about to preserve uniqueness and identity of distri
- E.g. only store sufficient statistic of data

- 3 ways of conceptualising relation between: (dependencies)

X (data) $T(x)$ (suff. statistic) θ (param)

- Bayesian: - Draw conclusions on parameters given data
  - dependency of parameters from data
  - use posterior $p(\theta|T(x), x)$

Frequentist - Data generated from unknown true value of param.
  - parameters impact data only through suff. stat $p(x|T(x), \theta)$

- ~~At~~ Influence flows through $T(\cdot)$ for both Bayesian, frequentist (due to exponential family)

· Neyman factorisation theorem: W A3 → WIKI - check you understand eq.

- exposes sufficiency of $T(x)$ for parameter θ.
- $T(x)$ d-separates X and θ

(*) Density estimation for single r.v. for many diff distr family
  - use sufficient statistic, moment matching, exp. family

· Move onto 2 nodes

- Generalised instance → GLIM

Ex: Builds on knowledge of exponential family

- Discrim - logistic regression; SVMs

  LDA → No it's generative

- logistic regression - $p(y=1|x) = \dfrac{1}{1+e^{-\theta^T x}}$

- These are GLIMS
  - But sigmoid → non linearity (taken care of by blanket function)
  - contains linear rel. ; so we use linear techniques with above.

Commonality (*): - $\mathbb{E}_p(Y) = \mu = f(\theta^T x)$