What is "relative" entropy?

- **Definition:** given p, q, p and q both supported on $\mathcal{X}$ and $\mathrm{Supp}(q) = \{x \text{ s.t. } q(x) > 0\} \subseteq \mathrm{Supp}(p) = \{x \text{ s.t. } p(x) > 0\}$ the relative entropy:

$$D(p \| q) = \mathbb{E}_{X \sim p}\left[\log\left(\frac{p(X)}{q(X)}\right)\right].$$

  - also called: "Kullback-Leibler Divergence (KL)"

  - motivating questions:
    (i) how can we interpret $D(p\|q)$? What does it mean? (separate from a convenient object arising in analysis)
    (ii) in particular, why is the asymmetry $D(p\|q) \neq D(q\|p)$ a feature of the definition rather than a bug?

- What we already know:

1. Properties:

(i) Nonnegativity: $D(p\|q) \geq 0$ and $= 0$ if and only if $p = q$ (almost surely for $X \sim p$)
  - loosely, $D(p\|q)$ is a "distance" between p & q — we'll qualify this

(ii) Asymmetry: $D(p\|q) \neq D(q\|p)$
  - apparent from definition. p and q are not exchangeable ... average over p not q
  - true distances (metrics) should be symmetric
  - not a distance, if we treat as a psuedo-metric then this is just a bug
  - why is this substantive/useful?

(iii) Joint Convexity (see T&C): if $\lambda \in [0,1]$ then

$$D(\lambda p + (1-\lambda)p' \| \lambda q + (1-\lambda)q') \leq \lambda D(p\|q) + (1-\lambda) D(p'\|q').$$

2. Example uses (so far):

(i) excess description length (cost of misspecification in compression):

(a) given code C, $\ell(x) = |C(x)|$, $L(C) = \mathbb{E}_{x\sim p}[\ell(x)]$

- if we can only send n characters (d-ary) per unit time, then, using C, the channel capacity (max-average X sent per unit time) is $\propto 1/L(C)$

- wanted to minimize $L(C)$ (maximizes capacity for noiseless channel)

- saw $L(C) = \underbrace{H[X]}_{\text{theoretical lower bound}} + D(p\|q) - \log(z)$ ← inefficiency from mismatch in code design and distribution of X

where $q(x) = \frac{1}{z} D^{-\ell(x)}$, $z = \sum_x D^{-\ell(x)}$.

(b) if we designed the code C using the assumption $X\sim q$ but $X\sim p$ then (ex: Shannon code)

$$L(C) \in H[X] + \underbrace{D(p\|q)}_{\text{excess description length}} + [0, 1]$$

- interpretation: $D(p\|q)$ is the cost of misspecification
  - true for other tasks
  - asymmetric since p is the true dist, generates samples } average over p not q
  - q is the approximation/proposal, does not generate samples }

(ii) differences in entropies ("relative"):

- Ex: $H[X] - \log(|\mathcal{X}|) \overset{*}{=} -D(p\,\|\,\text{Uniform}[\mathcal{X}])$

* recall from limiting construction of differential entropy

$$\left(H[X] - \log(|\mathcal{X}|) = H[X] + \log(|\mathcal{X}|^{-1}) = \mathbb{E}_{x\sim p}\left[-\log(p(x)) + \log(|\mathcal{X}|^{-1})\right] = -\mathbb{E}_{x\sim p}\left[\log\left(\frac{p(x)}{|\mathcal{X}|^{-1}}\right)\right] = -D(p\,\|\,\text{Uniform}[\mathcal{X}])\right)$$

so: $H[X] = H[\text{Uniform}(X)] - D(p \,\|\, \text{Uniform}(X))$

difference between max possible entropy (uniform) and entropy of p ... information needed to go from Uniform(X) to p

- this algebra does not work for generic q ($D(p\|q) \neq H(q) - H(p)$ generically)
- works when misspecified...
  - Suppose $X \sim p$ but we think $X \sim q$
  - measure our suprise on sampling X as $-\log(q(X))$
  - our expected surprise is:

$$-\mathbb{E}_{X\sim p}[\log(q(X))] \neq H(q) \text{ or } H(p)$$

if we have many samples could estimate
let $\tilde{H}(p\|q)$ be our estimate to $H[X]$
using q for surprise

- Compare: $\tilde{H}(p\|q) = -\mathbb{E}_{X\sim p}[\log(q(X))]$ ("cross entropy", usually $H(q,p)$) w/ $H[X] = H(p)$ ...
- $\mathbb{E}$ error in estimate: $\tilde{H}(p\|q) - H(p) = \mathbb{E}_X[-\log(q(X)) + \log(p(X))]$

$$= D(p\|q) \geq 0$$

- so, if $X \sim p$, think $X \sim q$, and estimate $H[X]$ using samples then the expected error in our estimate is $D(p\|q)$.
- the error is >0 if $p \neq q$. Always too/overly surprised (overestimate $H[X]$)
- $D(p\|q)$ is our excess surprise (when misspecified)

3. Relation to Mutual Info: $I[X;Y] = D(p_{X,Y} \,\|\, p_X p_Y)$

so:
(a) mutual info is the expected excess description length of joint draws X, Y if we ignored their dependence

(b) mutual info is the excess surprise when observing joint samples X, Y if we thought they were independent

} Contextualize I via D not D...

(c) we will see... if X is an unknown w/ prior p, observe Y, posterior $X|Y=y$, then $D(p_X \,\|\, p_{X|Y=y})$ = the information gained about X via observation.