Relative Entropy is a Divergence (not a distance)

- roughly, a divergence $p \| q$ measures how distinguishable q is from p when we draw data from p
  - If $X \sim p$, but I thought $X \sim q$, how hard would it be to tell?
  - How much data would I need to find out?

... easiest to motivate divergences via hypothesis testing ...

- **Problem:** We observe a sequence of samples $\{X_i\}_{i=1}^{n} \overset{iid}{\sim} p$ where $p = p_0$ or p_1, but we don't know which...

- **hypotheses:** $H_0 : p = p_0$, $H_1 : p = p_1$

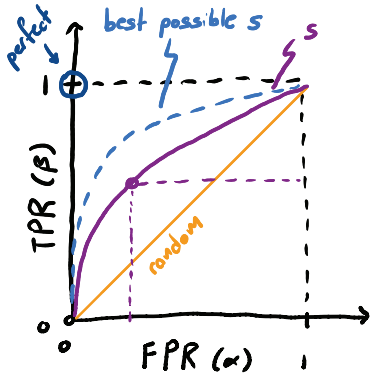
- **Test:**
  1. pick a test statistic $s(X_1, X_2, \dots X_n) : \mathcal{X}^n \to \mathbb{R}$
  2. estimate the sampling distribution of $s(X_1, X_2, \dots X_n)$ under $X \sim p_0$ and $X \sim p_1$
  3. compare observed s to its sampling distribution under p_0 and p_1
  4. define accept (H_1) and reject (H_0) regions using 2. (e.g. a threshold on s for a one-sided test)
  5. reject/accept based on where the observed s is

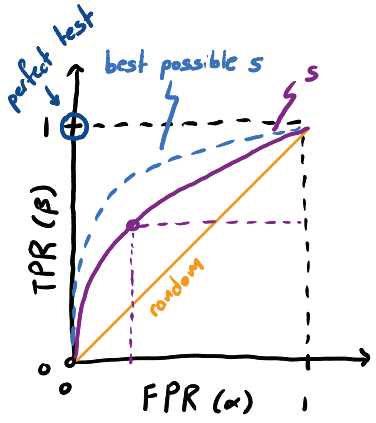
  (e.g. accept if $\Pr(\text{sampling } s(X_1, X_2, \dots X_n) \text{ less likely than observed } s(x_1, x_2, \dots x_n) \text{ under } H_0) \leq \alpha$ for some $\alpha \Rightarrow \alpha$ controls the FPR...
  i.e. chance we pick H_1 when H_0 is true
  i.e. controls the significance of the test)

  could exchange the roles of H_0 and H_1 here, convention is arbitrary...

- Choosing s and the accept/reject region (α) determines the test (fixes decision thresholds on s)
  - thus, also fixes the FNR, i.e. prob. of choosing H_0 when H_1 is true
    i.e. controls power/sensitivity

- generally want power & significance (sensitivity & specificity)
- for a given s, visualize trade-off w/ ROC
  - plot best TPR (power) possible for a given FPR (significance)
    equivalently, best TNR for a given FNR





- Want to choose our statistic s to:
  - (·) maximize power (sensitivity) for all significances (specificity) when using H_0 to control (fix accept/reject to control FPR = control significance
  - (··) maximize significance (specificity) for all powers (sensitivity) when using H_1 to control (fix accept/reject to control FNR = control sensitivity

- What s should we choose?

- Neyman-Pearson: if $f: \mathbb{R}^+ \to \mathbb{R}$ is injective (one-to-one) and monotonic and we define:

"likelihood ratio" ↘ $$LR(x_1, x_2, \dots x_n) = \frac{Pr(X_1 = x_1, X_2 = x_2, \dots X_n = x_n \mid H_1)}{Pr(X_1 = x_1, X_2 = x_2, \dots X_n = x_n \mid H_0)}$$

if large, data more likely under H_1, choose H_1
if small, ... under H_0 choose H_0

then $s(x_1, x_2, \dots x_n) = f(LR(x_1, x_2, \dots x_n))$ will achieve (·) and (··).

- idea: the best way to choose H_0 vs. H_1 is by comparing the likelihood of the data under each model.

- Ex: (i) LR test (ii) log(LR) test (iii) $H_0: X \sim \exp(\lambda_0)$, $H_1: X \sim \exp(\lambda_1)$ can use $s = \frac{1}{n}\sum_{j=1}^{n} x_j$.

- to understand how distinguishable P_0 and P_1 are, we should study the sampling dist. of $s(X_1 \dots X_n)$
  - at least for large n...

Q: What is $\mathbb{E}[s(X_1, \dots X_n)]$? What is the sampling dist. for large n?

- Single Sample:
$$\mathbb{E}[s(X)] = \begin{cases} \mathbb{E}_{X \sim P_0}[f(LR(X))] & \text{if } H_0 \text{ true} \\ \mathbb{E}_{X \sim P_1}[f(LR(X))] & \text{if } H_1 \text{ true} \end{cases}$$

- in particular: log(LR):
$$\mathbb{E}[s(X)] = \begin{cases} \mathbb{E}_{X \sim P_0}\left[\log\left(\frac{P_1(X)}{P_0(X)}\right)\right] = -D(P_0 \| P_1) & \text{if } H_0 \text{ true} \\ \mathbb{E}_{X \sim P_1}\left[\log\left(\frac{P_1(X)}{P_0(X)}\right)\right] = +D(P_1 \| P_0) & \text{if } H_1 \text{ true} \end{cases}$$

↘ sign flips if use P_0/P_1 for LR

↙ order depends which is true.

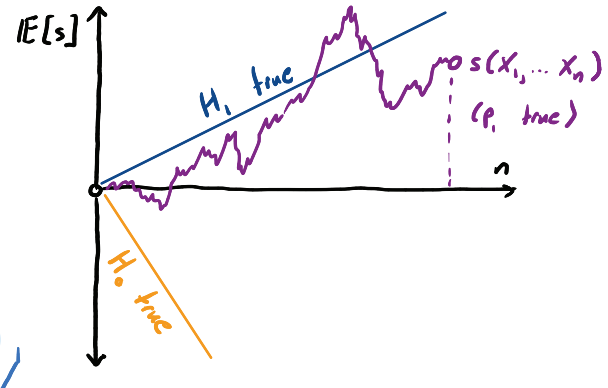
- **interpretation:** $D(p\|q)$ = expected value of log likelihood ratio (best test stat for distinguishing p & q) on a single draw, when p is true

- **multi-sample:** $\mathbb{E}[s(X_1,\dots,X_n)] = \mathop{\mathbb{E}}_{X\sim P_0 \text{ or } P_1}\left[f\left(\frac{P_r(X_1,X_2,\dots X_n\,|\,H_1)}{P_r(X_1,X_2,\dots X_n\,|\,H_0)}\right)\right] = \mathop{\mathbb{E}}_{X\sim P_0 \text{ or } P_1}\left[f\left(\prod_{j=1}^{n}\frac{P_r(X_j\,|\,H_1)}{P_r(X_j\,|\,H_0)}\right)\right]$

  - ideally, choose f s.t. the terms here separate ... log is natural (separate product over each sample)

- **log(LR):** $\mathbb{E}[s(X_1,\dots X_n)] = \mathop{\mathbb{E}}_{X\sim P_0 \text{ or } P_1}\left[\log\left(\prod_{j=1}^{n} LR(X_j)\right)\right] = \sum_{j=1}^{n}\mathop{\mathbb{E}}_{X\sim P_0 \text{ or } P_1}\left[\log(LR(X_j))\right]$

$$= \begin{cases} -n\,D(P_0\|P_1) & \text{if } H_0 \\ n\,D(P_1\|P_0) & \text{if } H_1 \end{cases}$$



- the larger $D(P_1\|P_0)$ the faster we expect to acquire evidence distinguishing P_1 and P_0 when P_1 is true
- ... $D(P_0\|P_1)$ ... when P_0 is true

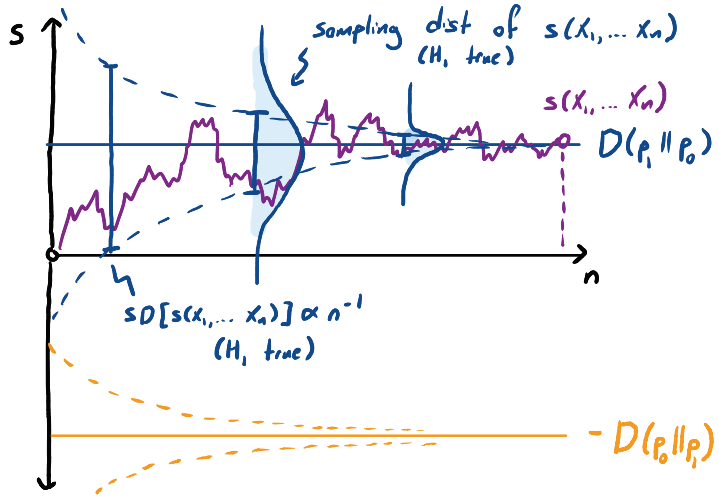
$D(p\|q)$ is the rate we gain evidence to distinguish p & q when p is true

- If we use $f^{(n)}(t) = \frac{1}{n}\log(t) = \log(t^{1/n})$

then:

$$\mathbb{E}[s(X_1,\dots X_n)] = \begin{cases} -D(P_0\|P_1) & \text{if } H_0 \text{ true} \\ +D(P_1\|P_0) & \text{if } H_1 \text{ true} \end{cases}$$

and, $s(X_1,\dots X_n) = \frac{1}{n}\sum_{j=1}^{n}\log(LR(X_j))$ (i.i.d. draws)



so, by the CLT, as $n\to\infty$, the sampling distribution of $s(X_1,\dots X_n)$ converges to a $\mathcal{N}$ distribution w/ $\mathbb{E} = D(\cdot\|\cdot)$ and variance $= \mathcal{O}(n^{-1})$.

- asymptotically, $D(p\|q)$ and $D(q\|p)$ measure how distinguishable p is from q when drawing from p, or when drawing from q

- **Interpretation**: $D(p\|q) = \mathbb{E}$ of test statistic $(s(X_1,\dots X_n) = \log(LR(X_1,\dots X_n)^{1/n}))$ for the $\log(LR)$ test for distinguishing p & q when p is true.
  - asymmetric
  - loosely, "distinguishability given data $\sim p$"

- Changing f in $s(X) = f(LR(X))$ induces a family of "divergences" generated by f... (see next section)

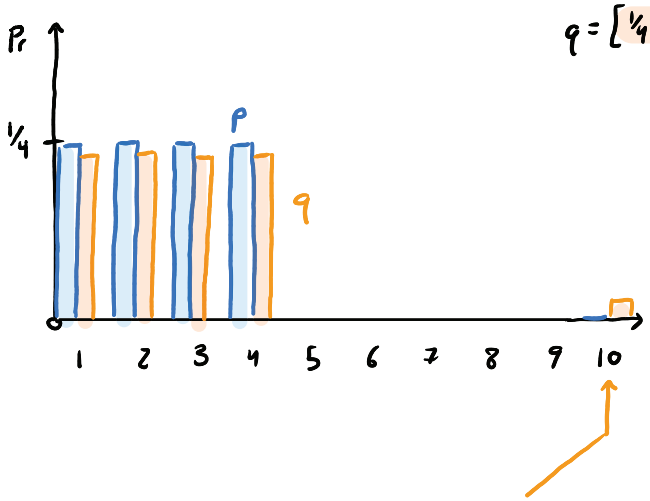
- This construction/interpretation explains the asymmetry of D...

- Ex: $\mathcal{X} = \{1, 2, \dots, 10\}$

$p = [1/4, 1/4, 1/4, 1/4, 0, 0, 0, 0, 0, 0]$

$q = [1/4 \cdot (1-\varepsilon)\ (\times 4), 0, 0, 0, 0, 0, \varepsilon]$

} only differ by a rare event that is possible under q



rare event: $X=10$, if $X=10$, we know w/ 100% certainty $X \sim q$

we can build examples where

$D(p\|q) \to 0$ but

$D(q\|p) \to \infty$!!!

very asymmetric.

- if p is true:

$$D(p\|q) = \mathbb{E}_{X\sim p}\left[\log\left(\frac{p(x)}{q(x)}\right)\right] = \sum_{x=1}^{4} \log\left(\frac{1/4}{1/4(1-\varepsilon)}\right)\cdot\frac{1}{4}$$

$$= -\log(1-\varepsilon) \simeq \varepsilon \xrightarrow[\text{as } \varepsilon\to 0]{} 0$$

so, when p is true, it's hard to distinguish q from p

- if q is true,

$$D(q\|p) = \mathbb{E}_{X\sim q}\left[\log\left(\frac{q(x)}{p(x)}\right)\right] = \sum_{x=1}^{4} \log(\dots)\cdot\frac{1}{4}(1-\varepsilon)$$

$$+ \varepsilon \underbrace{\log(\varepsilon/0)}_{\log(\infty) = +\infty}$$

$$= +\infty$$

so, when q is true p & q are easy to distinguish...

- Why? if q is true, we could sample $X=10$ (prob. ε) if we sample $X=10$ we know w/ complete certainty that $X \sim q$ not p since prob $X=10$ if $X\sim p$ is zero!