Only when both sub-seq of $S_i=0$, $S_i=1$ look as if H_0 is true but actually H_1 is true

$$\Rightarrow \text{Prob}(①) \approx e^{-n\alpha D(P_0\|P_1)} \cdot e^{-n(1-\alpha)D(Q_0\|Q_1)}$$

② <u>Without the side info $S_1, \dots, S_n$</u>

$$\text{Prob}(②) \approx e^{-nD(\alpha P_0+(1-\alpha)Q_0\|\alpha P_1+(1-\alpha)Q_1)}$$

$\therefore \text{Prob}(①) \leq \text{Prob}(②)$ ✓

Application of $D(P_0\|P_1)$

* An alternative way of deriving the entropy formula:

① Entropy: $H(X) = E_X\left(\log\left(\frac{1}{P_X(X)}\right)\right)$

Physical meaning: Compression:

If we have a string of n

i.i.d, X_i. then the n-dim vectors $(X_1, \cdots, X_n)$ can be compressed to $nH(X)$ (bits/nats)

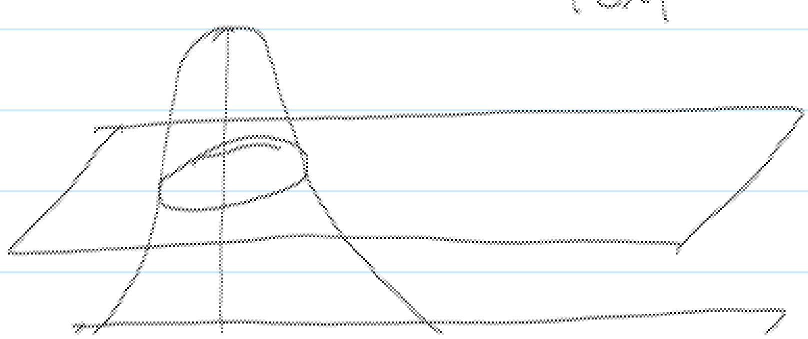
Namely: there are $e^{nH(X)}$ different values of the most likely (typical) $(X_1, \cdots, X_n)$

What is the connection between the divergence & the entropy?

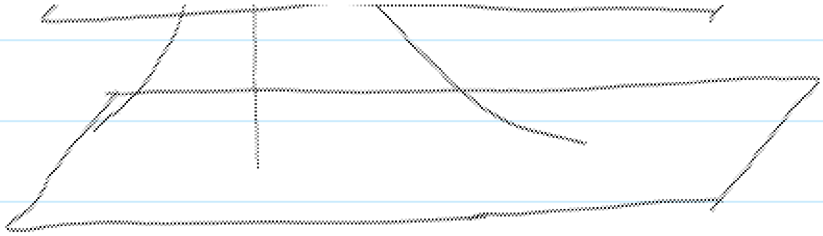
Let $\bigcirc$ denote the most likely outcomes of $(X_1, \cdots, X_n)$ under P_X.

Let U_X denote uniform distribution

$$P_{U_X}(X=x) = \frac{1}{|S_X|}$$



$(U_X)^n$



P_X

$$\left(\frac{1}{|S_X|}\right)^n \cdot |O| = e^{-nD(P_X \| U_X)}$$

$$= e^{-n\left(E_X\left(\log \frac{P_X(X)}{1/|S_X|}\right)\right)}$$

$$= e^{n\left(E_X\left(\log \frac{1}{|S_X|} + \log \frac{1}{P_X(X)}\right)\right)}$$

$$= e^{nE_X\left(\log \frac{1}{P_X(X)}\right)}$$

$$\Leftrightarrow |O| = e^{nH(X)}$$

The second application of $D(P_0 \| P_1)$ is to derive the channel capacity or mutual info. (The limit of error control codes

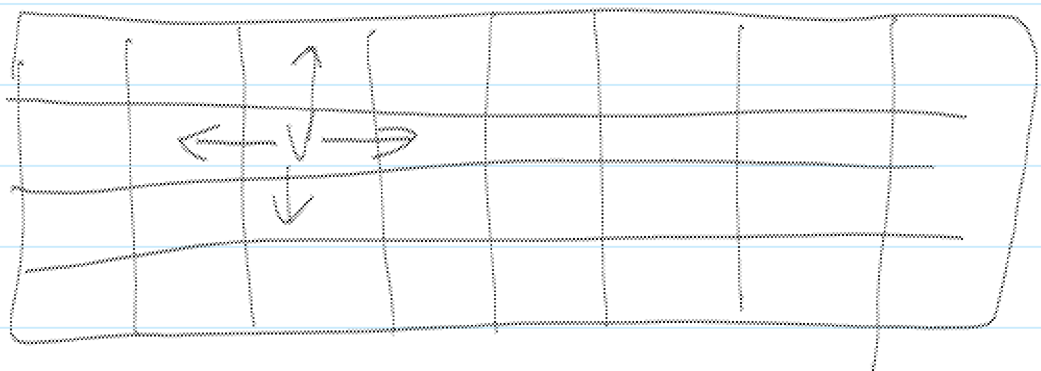
---

First look at the error control codes.

Ex:

A simple error control coding example:

* 32 check boxes that can be passed from A to B.
* One of them is checked



* In a noiseless environment, B can tell (out of 32 boxes) which one is checked.

⟹ We can convey 32 possible choices

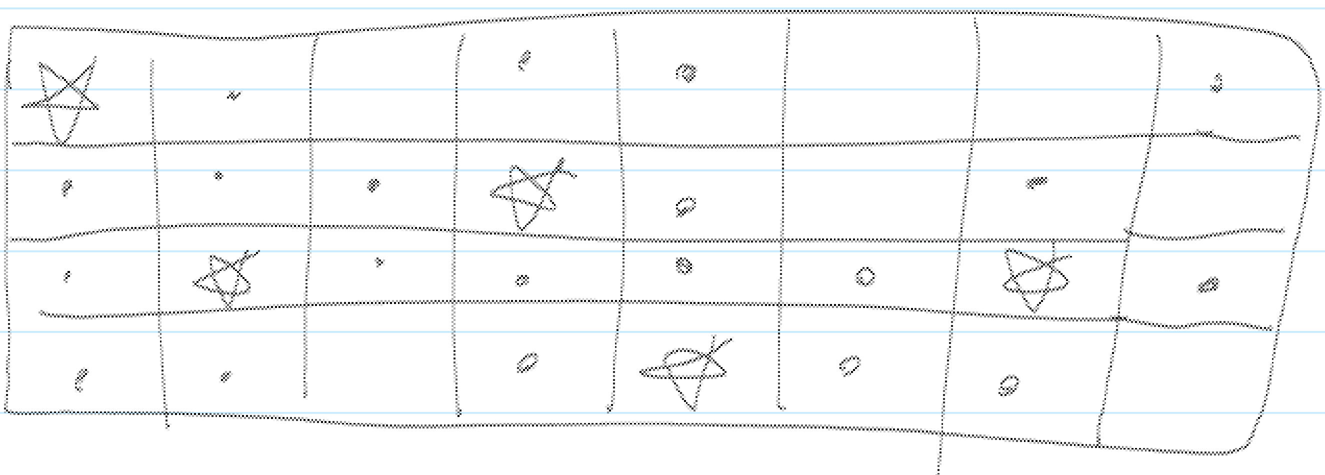
thus $\log_2(32) = 5$ bit info.

* In a noisy environment, the check box may shift one position either horizontally or

vertically (or stay in the same position)

Q: How to transmit error-free info at a reduced rate?

(We must not use all positions, but only some of them.)



reduced info $= \log_2(5) = 2.3219$ bits

$$\text{code rate} = \frac{2.3219}{\log_2(32)} \doteq 0.46$$

→ for the noisy env. (2.3219)

→ for the noiseless env. ($\log_2(32)$)

Can we do better? (hard to check)

But can have an easier upper bound:

$$\underline{X} \cdot 5 \le 32$$

$\hookrightarrow$ the number of distinct (error-free) choices

$$\Rightarrow X \le \frac{32}{5} = 6.4$$

$$\log_2(X) \le 2.678 \text{ bits.}$$

This simple upper bound is called the sphere-packing bound or the Hamming bound. In many cases, this sphere-packing bound is achievable & equal to the <u>channel capacity</u>.