Lecture 04

Note Title 1/24/2012

* Chernoff Bound

Misdetection $P_1(T(Y) \geq \tau^*)$

where $T(\cdot) = \log\left(\frac{P_0(\cdot)}{P_1(\cdot)}\right)$

* $\tau^* = D(P_0 \| P_1) = E_0(T(Y))$

$$\triangleq E_0\left(\log \frac{P_0(Y)}{P_1(Y)}\right)$$

Divergence, Kullback-Leibler info number.

* Prob ($Y_1, \ldots, Y_n$ look like H_0 is true $|$ H_1 is true)

$$\approx e^{-nD(P_0 \| P_1)}$$

* The building foundation of large deviation theory

How to use $D(P_0 \| P_1)$?

Ex: X_i is i.i.d Bernoulli with uniform distribution

$$\bar{X}_{1000} = \frac{1}{1000} \sum_{i=1}^{1000} X_i$$

Q: $P(\bar{X}_{1000} > 0.75) = ?$

Ans: The actual distribution is $P_1 = \frac{1}{2} / \frac{1}{2}$

but the outcomes look like $P_0 = 0.75/0.25$

$$\Rightarrow P(\bar{X}_{1000} > 0.75) \approx e^{-1000 D(P_0 \| P_1)}$$

$$= e^{-1000\left(0.75 \times \log \frac{0.75}{0.5} + 0.25 \log \frac{0.25}{0.5}\right)}$$

$$= \left(2 \times 3^{-0.75}\right)^{1000}$$

Note: The actual distribution is P_1 (H_1 is true) but

$$D(P_0 \| P_1) = E_0\left(\log \frac{P_0(Y)}{P_1(Y)}\right)$$ is evaluated by P_0

Properties of $D(P_0 \| P_1)$.

③ $D(P_0 \| P_1) < \infty$ if $\forall x$ we have $P_1(X=x)=0 \Rightarrow P_0(X=x)=0$.

Mathematically

Suppose $E_0\left(\log \frac{P_0(Y)}{P_1(Y)}\right) < \infty$

So if $P_1(y)=0 \Rightarrow P_0(y)=0$

Intuition: $D(P_0 \| P_1) < \infty$

$\Longleftrightarrow$ When H_1 is true, the "prob of $Y_1, \cdots, Y_n$ looking like H_0 is true" is non-zero

$\Longleftrightarrow$ Given H_1 is true, for $Y_1, \cdots, Y_n$ to mimic H_0 is true, we must have

$P_1(y)=0 \Rightarrow P_0(y)=0$

④ $D(P_0 \| P_1)$ is a convex function with respect to P_0, & P_1

Namely for any pair of P_0 P_1, Q_0, Q_1.

We have

$$D(\alpha P_0 + (1-\alpha) Q_0 \| \alpha P_1 + (1-\alpha) Q_1)$$
$$\leq \alpha D(P_0 \| P_1) + (1-\alpha) D(Q_0 \| Q_1)$$

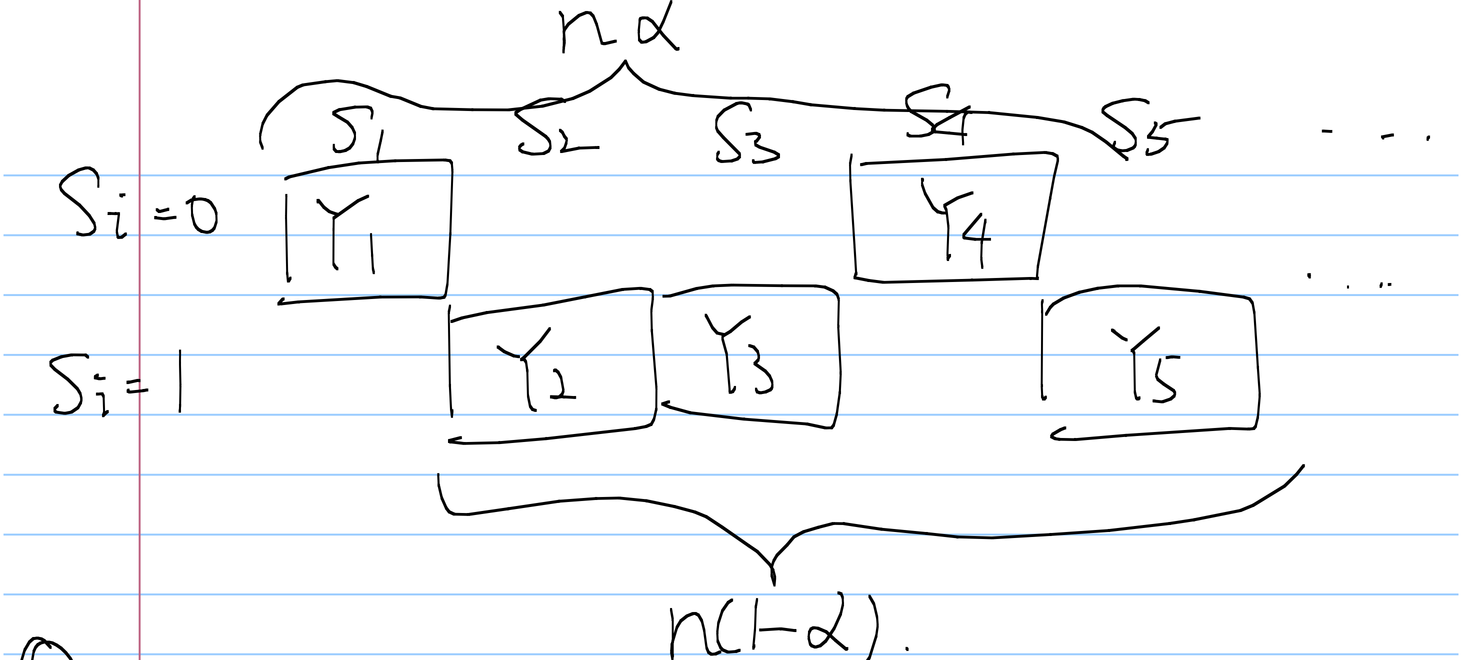
(Note the LHS is

$$\sum_y \left(\alpha P_0(y) + (1-\alpha) Q_0(y)\right) \log \frac{\alpha P_0(y) + (1-\alpha) Q_0(y)}{\alpha P_1(y) + (1-\alpha) Q_1(y)}\Bigg)$$

pf: Exercise (Hint: By Jensen's inequality)

Intuition: Consider side information S

$$P_0(y) = P(Y=y \mid X=0, S=0)$$
$$Q_0(y) = P(Y=y \mid X=0, S=1)$$
$$P_1(y) = P(Y=y \mid X=1, S=0)$$
$$Q_1(y) = P(Y=y \mid X=1, S=1)$$
$$P(S=0) = \alpha \qquad P(S=1) = 1-\alpha$$



① With the side info $S_1, \dots, S_n$ when will $Y_1, \dots, Y_n$ look like H_0 is true but actually H_1 is true?

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 Without the side info $S_1, \dots, S_n$

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

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

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)

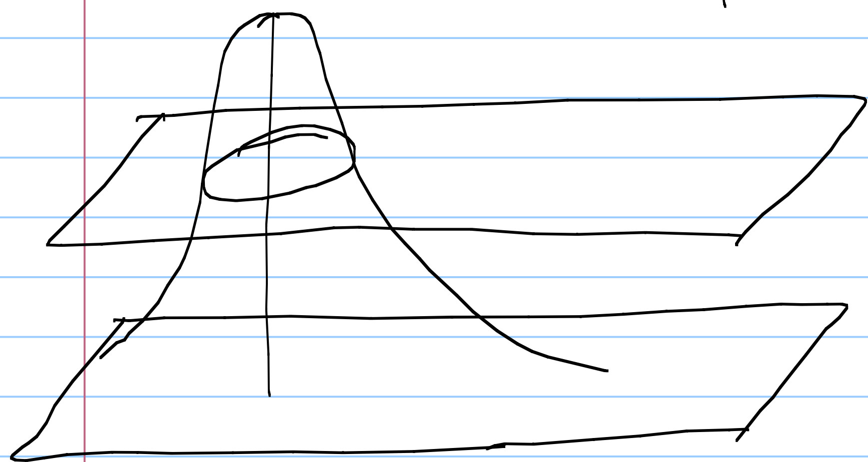
Namely: there are $e^{nH(X)}$ different values of the most likely (typical) $(X_1, \cdots, X_n)$

What is the connection between the divergence & the entropy?

Let O 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

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

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

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

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

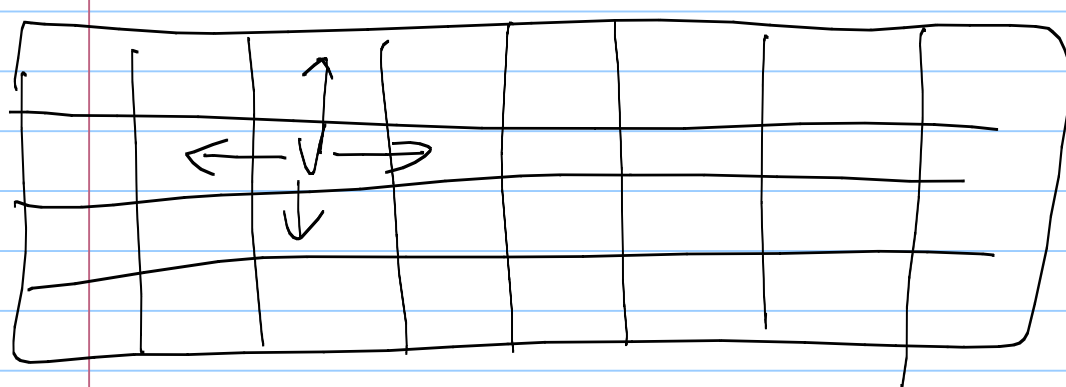
$$\Leftrightarrow |O| = e^{nH(X)}$$

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.

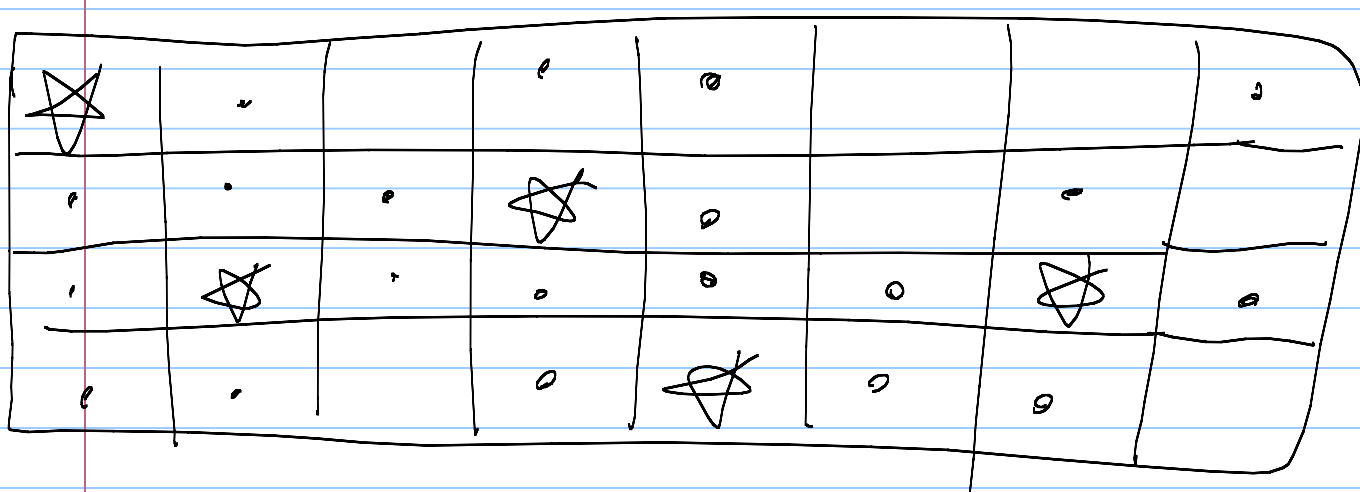
$\Rightarrow$ We can convey 32 possible choices

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

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

$\to$ for the noisy env.

$\to$ for the noiseless env.

Can we do better? (hard to check)

But can have an easier upper bound:

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

$\hookrightarrow$ 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>.