Arithmetic Coding Summary from L7

"Language model": often given by conditional probability distributions:

$$P(x_n \mid \underbrace{x_1, \ldots, x_{n-1}}_{x^{n-1}}) \quad \text{for } n = 1, 2, \ldots, N$$

equivalent

w/ joint distribution

$$P(x^n) = P(x_1)\, P(x_2 \mid x_1)\, P(x_3 \mid x_1 x_2) \cdots P(x_n \mid x^{n-1})$$

↳ see last lecture notes + exercise class

Arithmetic coding:

Input: $x^N \in \mathcal{A}^N$ to compress

Algo:

* $Q \leftarrow 0$, $R \leftarrow 1$, $p \leftarrow 1$
* For $n = 1, 2, \ldots, N$:
  1. $R \leftarrow Q + p\, R(x_n \mid x_1, \ldots, x_{n-1})$
     $Q \leftarrow Q + p\, Q(x_n \mid x_1, \ldots, x_{n-1})$
  2. While $R \leq \frac{1}{2}$ or $Q \geq \frac{1}{2}$:
     $b \leftarrow \begin{cases} 0 & R \leq \frac{1}{2} \\ 1 & Q \geq \frac{1}{2} \end{cases}$
     Write b
     $R \leftarrow 2R - b$
     $Q \leftarrow 2Q - b$
  3. $p \leftarrow R - Q$
* Write $\lceil \log \frac{2}{p} \rceil$ bits of binary expansion of $\frac{Q+R}{2}$

$\sum_{y \leq x_n} P(y \mid x_1 \cdots x_n)$ upper cumulative prob

lower cumulative prob $\sum_{y < x_n} P(y \mid x_1 \cdots x_n)$

Average rate: $\approx \dfrac{H(X^N)}{N}$ for large N

Joint Entropies (§8)

Joint distribution $P(x,y) \to H(X,Y)$

* marginal distributions: $P(x), P(y) \to H(X), H(Y)$

  $\hookrightarrow H(X)+H(Y) \geq H(X,Y)$, $=$ iff X,Y independent $\boxed{\text{HW 3}}$

* Conditional distributions: $P(y|x)$, $P(x|y)$

  $\hookrightarrow H(Y|X=x) = \sum_y P(y|x)\cdot \log \frac{1}{P(y|x)}$ & similarly $H(X|Y=y)$

__Conditional entropy__:

$$H(Y|X) := \sum_x P(x)\, H(Y|X=x)$$

* $H(Y|X) \geq 0$, $=0$ iff $Y=f(X)$ for some function f

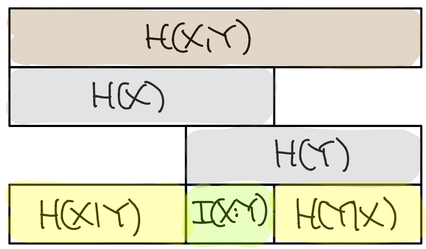
  __Pf__: $=0$ iff $H(Y|X=x)=0$ $\forall x$ iff $\forall x\ \exists y$: $P(y|x)=1$ $\square$

  with $P(x)>0$

* $H(Y|X) = H(X,Y) - H(X)$

  __Pf__: $H(Y|X) = \sum_{x,y} P(x)P(y|x) \log \frac{1}{P(y|x)}$

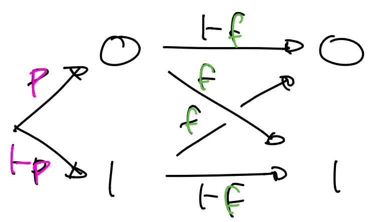
  $= \sum_{x,y} P(x,y) \log \frac{P(x)}{P(x,y)} = H(X,Y) - H(X)$. $\square$



* $H(Y|X) \leq H(Y)$, $=$ iff X,Y independent (*)

  __Pf__: eqv to $H(X,Y) \leq H(X) + H(Y)$! $\square$

__ex__:



$\Longrightarrow$ $H(Y|X) = p\cdot H(\{1-f,f\}) + (1-p)\cdot H(\{f,1-f\})$

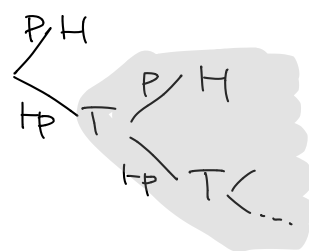
$= H(\{f,1-f\}) = \begin{cases} 0 & \text{if } f=0 \text{ OR } f=1 \\ 1 & \text{if } f=\frac{1}{2} \end{cases}$

independent of p!

__ex__: $N = \#$coin flips of biased coin until 1st heads

$H(N) = ?$

$X = \begin{cases} 1 & \text{if 1st outcome is heads } (N=1) \\ 0 & \text{otherwise } (N>1) \end{cases}$



$\Rightarrow H(N) = H(N,X) = H(X) + H(N|X)$

$X = f(N)$

$$= \underbrace{H(X)}_{=H(\{p,1-p\})} + p\cdot \underbrace{H(N|X=1)}_{=0 \text{ since } N= \text{ if } X=1} + (1-p)\cdot \underbrace{H(N|X=0)}_{=H(N)}$$

$$\Rightarrow H(N) = \frac{H(\{p,1-p\})}{p}$$

Mutual information:

$I(X:Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

* $I(X:Y) \geq 0$, $=0$ iff X, Y independent
* $I(X:Y) \leq H(X), H(Y)$

} reformulations of facts for $H(Y|X)$, $H(Y|X)$ from above

* $I(X:Y) = D(P_{xy} \| Q_{xy})$, where $Q(x,y) = P(x)\,P(y)$

EX CLASS

Recall: Relative entropy:

$0 \cdot \log 0 = 0$

$$D(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \in [0,\infty]$$

* $D(P\|Q) < \infty \Leftrightarrow \forall x: Q(x)=0 \Rightarrow P(x)=0$
* Gibbs inequality: $D(P\|Q) \geq 0$, $=0$ iff $P=Q$

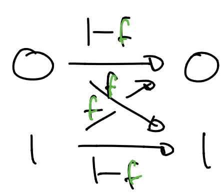
Communicating over Noisy Channels (§9)

we will... mathematical model ... design these

(Discrete memoryless) channel: $Q(y|x)$ cond. probability dist.

where $x \in \mathcal{A}_X$ input alphabet, $y \in \mathcal{A}_Y$ output alphabet

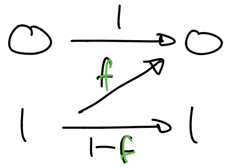
e.g. ① Binary symmetric channel: (our old friend)



$Q(0|0) = Q(1|1) = 1-f$

$Q(1|0) = Q(0|1) = f$

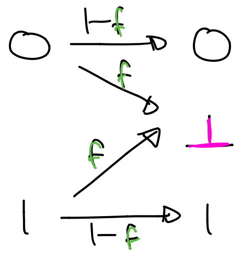
② Binary asymmetric channel: (from HW1)



$Q(0|0)=1,\ Q(1|0)=0$

$Q(0|1)=f,\ Q(1|1)=1-f$

③ Binary erasure channel:



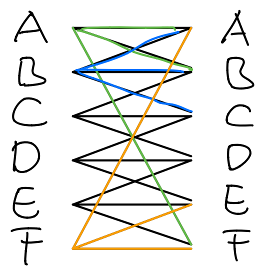
$Q(\perp|0)=Q(\perp|1)=f$

$Q(0|0)=Q(1|1)=1-f$

$Q(1|0)=Q(0|1)=0$

$\mathcal{A}_X=\{0,1\}\quad \mathcal{A}_Y=\{0,1,\perp\}$

④ Noisy typewriter:



$Q(A|A)=Q(B|A)=Q(F|A)=\frac{1}{3}$

$Q(B|B)=Q(C|B)=Q(A|B)=\frac{1}{3}$

$\vdots$

$Q(F|F)=Q(A|F)=Q(E|F)=\frac{1}{3}$

How well can we communicate over each of them?

* If we allow no errors at all: ① ↯ any y could come from either x

② ↯ y=0 can come from any x (+ sending 0 all the time is not informative)

③ ↯ $y=\perp$ can come from either x "zero error commun."

④ ☺ encode $0 \mapsto B$, $1 \mapsto E$ decode $A,B,C \mapsto 0$, $D,E,F \mapsto 1$ → $\boxed{\text{EX CLASS}}$

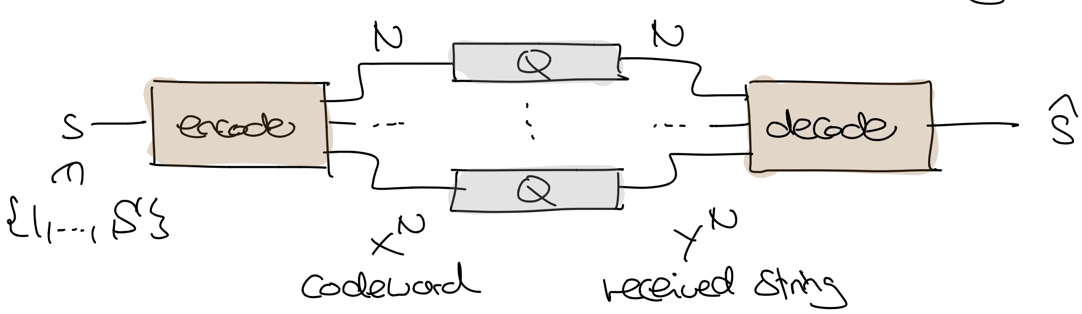
* If we allow error: Can use Bayes' theorem to infer most likely x:

$$P(x|y)=\frac{Q(y|x)\,P(x)}{\sum_{\tilde{x}} Q(y|\tilde{x})\,P(\tilde{x})}$$

assuming x come from some ensemble

↳ Lecture 1 & 2

For reliable communication, consider block encodings:



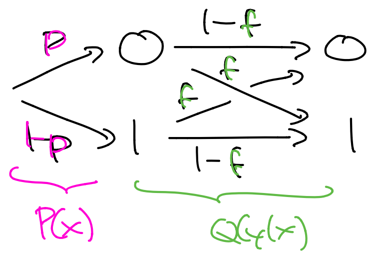
WANT: $S=\hat{S}$ with high probability

Rate $R = \frac{\log(S)}{N}$ bits per channel use   e.g. $R = \frac{\log \# A}{N}$ for N-fold repetition code

$\boxed{\text{Shannon's Noisy Coding Theorem (informal):}}$ The "optimal" rate at which we can communicate "reliably" is given by the Capacity of the channel $Q(y|x)$:

$$\boxed{C(Q) = \max_{P(x)} I(X:Y)} \quad \text{— for } P(x,y) = P(x) \cdot Q(y|x)$$

e.g. for the binary symmetric channel:



$$* \; I(X:Y) = H(Y) - H(Y|X)$$

$$= H(Y) - \underbrace{H(\{f, 1-f\})}_{\text{indep of } p} \quad \text{// see above!}$$

$$* \; \max_{p} H(Y) = 1 \quad \text{since } P(Y=0) = p(1-f) + (1-p)f = \frac{1}{2} \text{ if } p = \frac{1}{2}$$

$$\Longrightarrow C(Q) = \max_{p} I(X:Y) = 1 - H(\{f, 1-f\}) = \begin{cases} 0 & \text{if } f = \frac{1}{2} \\ 1 & \text{if } f = 0 \text{ OR } f = 1 \end{cases}$$

Intuitive!