Arithmetic Coding Summary from L7

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

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

w/ joint distribution

$$P(x^n) = P(x_1)\,P(x_2|x_1)\,P(x_3|x_1x_2)\cdots P(x_n|x^{n-1})$$

equivalent

↳ 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 $n = 1, 2, \dots, N$:
    ① $r \leftarrow q + p\,R(x_n | x_1,\dots,x_{n-1})$
      $q \leftarrow q + p\,Q(x_n | x_1,\dots,x_{n-1})$
    ② While $r \le \frac{1}{2}$ or $q \ge \frac{1}{2}$:
        $b \leftarrow \begin{cases} 0 & \text{if } r \le \frac{1}{2} \\ 1 & \text{if } q \ge \frac{1}{2} \end{cases}$
        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 $\frac{q+r}{2}$
```

$\sum_{y \le x_n} P(y|x_1 \cdots x_{n-1})$

upper cumulative prob

lower cumulative prob

$\sum_{y < x_n} P(y|x_1 \cdots x_{n-1})$

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

Joint Entropies (§8)

Joint distribution $P(x,y) \rightarrow H(XY)$

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

$\hookrightarrow H(X) + H(Y) \geq H(XY)$, $=$ iff X, Y independent 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

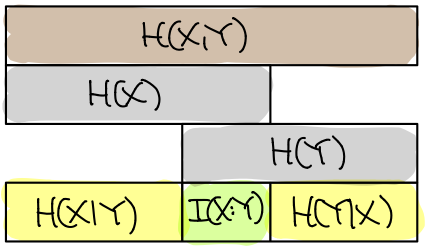
Pf: $=0$ iff $H(Y|X=x) = 0$ $\forall x$ iff $\forall x\ \exists y: P(y|x) = 1$ ☐

with $P(x) > 0$

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

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

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



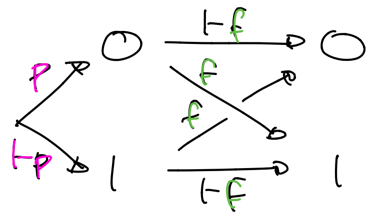
* $H(Y|X) \leq H(Y)$, $=$ iff X, Y independent ⊛

Pf: eqv to $H(XY) \leq H(X) + H(Y)$! ☐

* Chain rule: $H(YZ|X) = H(Y|X) + H(Z|XY)$

Pf: RHS $= H(YX) - H(X) + H(XYZ) - H(XY) =$ LHS. ☐

ex:



$\Longrightarrow$

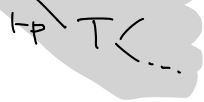
$H(Y|X) = p \cdot H(\{1-f, f\}) + (1-p) \cdot H(\{f, 1-f\})$

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

independent of p!

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

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



$$\Rightarrow H(N) \underset{X=f(N)}{=} H(N,X) = H(X) + H(N|X)$$

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

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

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:

$\frac{0}{0}$ etc!? let's discuss?

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

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

Communicating over Noisy Channels (§9)

we will... ... design these

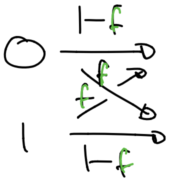
mathematical model

(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)

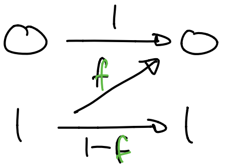


$Q(0|0)=Q(1|1)=1-f$

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

② Binary asymmetric channel:

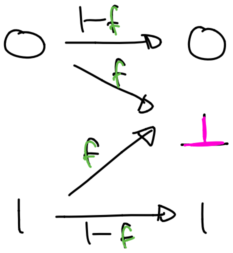
(from HW 1)



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

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

③ Binary erasure channel:



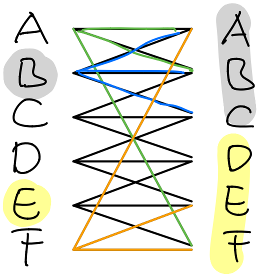
$Q(\perp|0)=Q(\perp|1)=f$

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

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

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

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=⊥ can come from either x

"zero error commun."

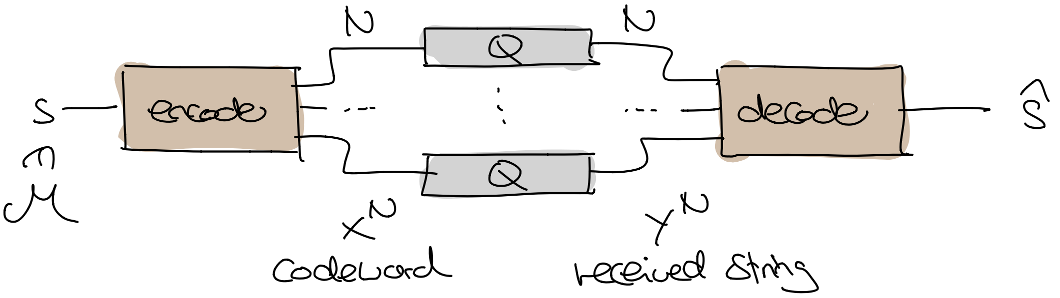
④ ☺ encode 0 ↦ B, 1 ↦ E decode A,B,C ↦ 0, D,E,F ↦ 1 → $\boxed{\text{EX CLASS}}$

* 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

e.g. for S uniform

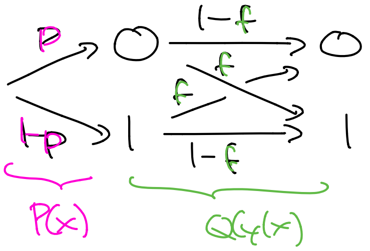
Rate $R = \frac{\log \# \mathcal{M}}{N}$ bits per channel use   e.g. $R = \frac{\log \# \mathcal{M}}{N}$ for N-fold repetition code

the larger the better

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:



$$* \quad I(X:Y) = H(Y) - \underbrace{H(Y|X)}_{\text{|| see above !}}$$

$$= H(Y) - \underbrace{H(\{f, 1-f\})}_{\text{indep of } p}$$

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

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

Intuitive !