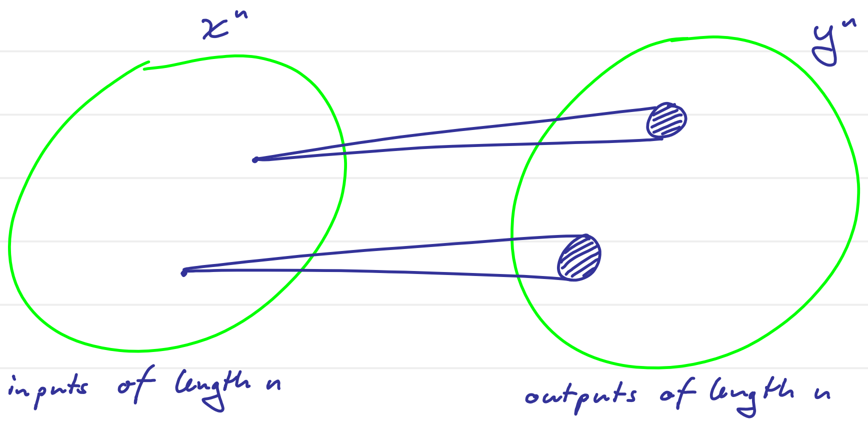
IV.6. Heuristic view on Shannon's noisy channel coding theorem



(i) for every input we obtain $\sim 2^{H(Y|X)}$ outputs with roughly equal prob. The other outputs may be possible, but they are not typical and thus ε unlikely.

(ii) the total number of typical sequences at the output is $\sim 2^{nH(Y)}$

(iii) to be able to distinguish different inputs at the output, these images must not overlap.

$\longrightarrow$ there are about $\dfrac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nI(X:Y)}$ such inputs corresponding to a rate of $\underline{\underline{I(X:Y)}}$

remark: strictly speaking, in (i) it should be $2^{nH(Y|X=x)}$ for an input $x \in \mathcal{X}^n$

IV.7. Properties of the channel capacity

Proposition: (1) $C \geq 0$

(2) $C \leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$

proof: both are rather obvious from the operational definition, but they also follow easily from properties of the mutual information:

(1) $I(X:Y) \geq 0$

(2) $I(X:Y) = \begin{cases} H(X) - H(X|Y) \leq H(X) \leq \log|\mathcal{X}| \\ H(Y) - H(Y|X) \leq H(Y) \leq \log|\mathcal{Y}| \end{cases}$ $\square$

Thm.: (additivity) Let $S_1, S_2, S_1 \otimes S_2$ be stochastic matrices describing two discrete memoryless channels and their product, respectively. Then

$$\boxed{C(S_1 \otimes S_2) = C(S_1) + C(S_2)}$$

proof: $C(S_1 \otimes S_2) = \max_{p(x^2)} I(X^2 : Y^2)$ where $X^2 := (X_1, X_2)$, $x^2 \in \mathcal{X} \times \mathcal{X}$

(i) $\geq \max_{p(x_1)p(x_2)} I(X^2 : Y^2) = C(S_1) + C(S_2)$

(ii) $\leq \max_{p(x^2)} \left(I(X_1 : Y_1) + I(X_2 : Y_2) \right)$ due to additivity Lemma

$= \max_{p(x_1)p(x_2)} (\text{—"—} + \text{—"—}) = C(S_1) + C(S_2)$

$\square$

Proposition: $I(X:Y)$ w.r.t. $p(x,y) = \overbrace{p(y|x)}^{=:S_{yx}} p(x)$ is a convex functional of $p(y|x)$ and a concave functional of $p(x)$.

proof:

- $$I(X:Y) = \underbrace{H(Y)}_{\text{concave in } p(x)} - H(Y|X) = H(Y) - \underbrace{\sum_x p(x) \overbrace{H(Y|X=x)}^{\text{depends only on } S \text{ not on } p(x)}}_{\text{linear in } p(x)}$$

  concave in $p(x)$ since $p(y) = \sum_x S_{yx} p(x)$ depends linearly on $p(x)$ & H is concave

  $\Rightarrow$ concavity in $p(x)$

- for the proof of convexity w.r.t. S fix $p(x)$ and define

  $$p_\lambda(y|x) := \lambda p_1(y|x) + (1-\lambda) p_2(y|x), \quad p_\lambda(x,y) := p_\lambda(y|x) p(x)$$

  Then $I_\lambda(X:Y) = D\left(p_\lambda(x,y) \,\|\, p_\lambda(y) p(x)\right)$ and convexity follows from joint convexity of the relative entropy. $\square$

Corollary:

- The channel capacity is a convex functional of the channel:

  $$C(\lambda S_1 + (1-\lambda) S_2) \leq \lambda C(S_1) + (1-\lambda) C(S_2)$$

- For $\max_{p(x)} I(X:Y)$ any local maximum is a global one.

$\rightarrow$ efficient algorithms for computing the capacity (e.g. Arimoto-Blahut)

IV.8. Computing some capacities

Prop.: Let S with $S_{yx} = p(y|x)$ be a stochastic matrix where all columns are permutations of a probability vector q. Then

$$C(S) = \left(\max_{p(x)} H(Y)\right) - H(q) \qquad \left(\text{where } Y \text{ is distributed according to } \sum_x p(y|x)p(x) \text{ and } y \in \mathcal{Y}\right)$$

$$\leq \log|\mathcal{Y}| - H(q)$$

with equality iff there is an input distribution $\tilde{p}$ s.t. $(S\tilde{p})_y = \frac{1}{|\mathcal{Y}|}$ $\forall y \in \mathcal{Y}$.

proof: $C = \max_{p(x)} I(X;Y) = \sup_{p(x)} H(Y) - H(Y|X)$

$$= \sup_{p(x)} \underbrace{H(Y)}_{\leq \log|\mathcal{Y}|} - \sum_x p(x) \underbrace{H(Y|X=x)}_{= H(q)}$$

and $H(Y) = \log|\mathcal{Y}|$ iff distribution is uniform. $\square$

Examples: ① binary symmetric channel:

$$S = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \qquad \text{(bit flipped with prob. } p\text{)}$$

$$\boxed{C(S) = \log 2 - h(p)} \qquad \text{for } \tilde{p} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$$

② noisy typewriter channel:

$$S = \begin{pmatrix} 1-2p & p & & & p \\ p & 1-2p & p & & \\ & p & \ddots & \ddots & \\ & & \ddots & \ddots & p \\ p & & & p & 1-2p \end{pmatrix} \qquad \text{"circulant matrix"}$$

$$\boxed{C(S) = \log|\mathcal{Y}| - H(q)}, \quad q = (1-2p, p, p) \qquad \text{for } \tilde{p} \text{ uniform}$$

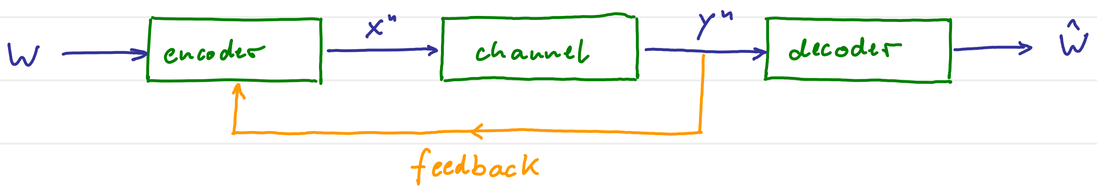
③ binary erasure channel

$$S = \begin{pmatrix} 1-p & 0 \\ p & p \\ 0 & 1-p \end{pmatrix}$$ (: bit erased with prob. p)

$$\boxed{C(S) = 1-p}$$

(proven in the exercise. A uniform output distribution is in this case not possible. Uniform input $\tilde{p}$ is optimal, though.)

IV.9. Feedback capacity



Consider sequential uses of the channel where in each encoding step the output of all previous transmissions can be used.

Def.:
- An (M,n)-code with feedback for a discrete memoryless channel with input & output alphabets $\mathcal{X}$ & $\mathcal{Y}$ is defined via
  - encoding functions $f_i : \{1,\dots,M\} \times \mathcal{Y}^{i-1} \to \mathcal{X}$, $i=1,\dots,n$
  - a decoding function $g : \mathcal{Y}^n \to \{1,\dots,M\}$
- Let $Y^i := (Y_1,\dots,Y_i)$ and $X_i := f_i(W, Y^{i-1})$
- R is a "rate achievable with feedback" iff $\forall \varepsilon > 0\ \exists\ (2^{nR}, n)$-code with feedback s.t. $\lambda^{(n)} < \varepsilon$.
- The "feedback capacity" C_{FB} is the supremum over all such rates.

Thm.: $$\boxed{C_{FB} = C}$$