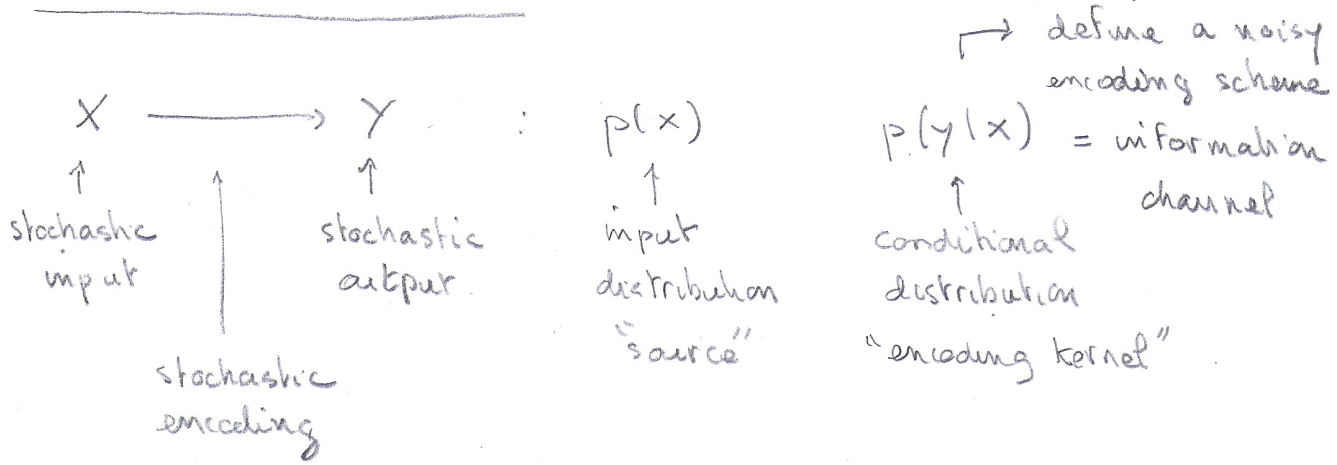
Information channel



In practice:
* $p(x)$ can be thought of the probability of some stimulus feature that are relevant to behavior.
* $p(y|x)$ represents the noisy processing of stimulus information by neural network: for example, y is a vector of neuronal activity.

Efficient coding hypothesis: (debated) Neural system have evolved to optimize the flow/processing of information given biophysical constraints on the organism resource (range of firing rate, mean level of activity...)

If relevant, the most important part would probably be to identify the nature of relevant constraints.

Problem : $\max I(X,Y)$ given some constraints on $p(x)$, $p(y|x)$ ...

Mutual information

↳ measure the dependence between X, Y or rather how well can X be described by reading out Y, ie. how many yes/no question can be answered from knowing Y.

↳ convexity properties: $I(X,Y) = \sum_x p(x) \sum p(y|x) \log \frac{p(y|x)}{p(x)}$

$p(x)$: ↑ concave; $p(y|x)$: ↑ convex

↳ assuming X, Y discrete variable for simplicity

A] $\max_{p(x)} I(X,Y) = C[p(y|x)]$ → channel capacity

↕ constrained or not

= maximum amount of information carried by the channel.

B] $\min_{p(y|x)} I(X,Y) = 0$ → unconstrained (uninteresting)

> 0 → under some constrained (interesting)

e.g. we impose a minimum level of average accuracy:

$\mathbb{E}[d(x,y)] \leq \delta$ ← error bound

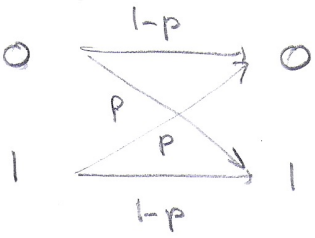
↑ distortion function

$\min_{p(y|x)} I(X,Y)$ given $\mathbb{E}[d(x,y)] \leq \delta$: rate distortion theory.

Channel capacity

Example: binary symmetric channel

p: error probability



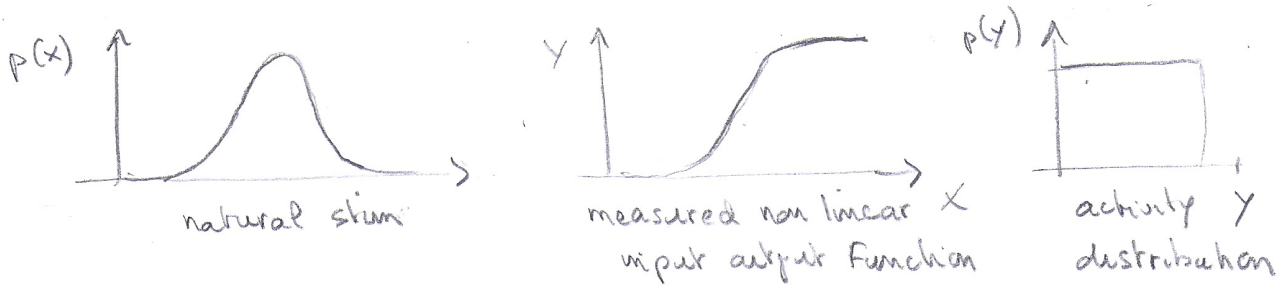
$$I(X,Y) = H(Y) - H(Y|X)$$

$$= H(Y) - \sum_x p(x)\, H(Y|X=x) \rightarrow H(p) \leftarrow \text{binary variable of parameter } p$$

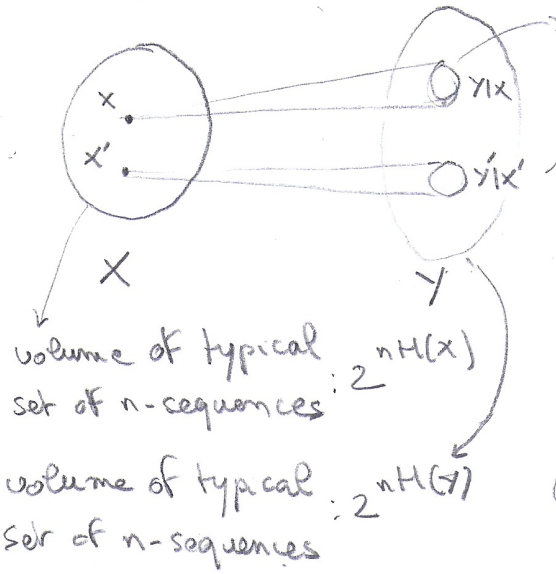
$$= H(Y) - H(p) \leq 1 - H(p) \leftarrow \text{maximum entropy of a binary variable is 1}$$

General principle: for encoding channel with $H(Y|X) = cte$ channel capacity corresponds to uniformely distributed output Y

↳ in neuroscience: Laughlin's histogram equalization



Shannon's Theorem:



average volume of typical set of n-sequence conditioned to a given output sequence: $2^{nH(Y|X)}$

# of reliably transmittable sequences

$$= \frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nH(Y)-H(Y|X)} = 2^{nI(Y|X)}$$

Capacity = optimal "soft" partitioning

Arimoto algorithm

↳ compute capacity and optimal input distribution.

↳ idea: solve a bigger optimization problem with identical solution to find a monotically converging method (MI increases each step).

$$I(X,Y) = \sum_x \sum_y p(x)p(y|x) \log \frac{p(y|x)}{\sum_{y'} p(x)p(y'|x)} = I[p]$$

(functional of p; $\leftarrow$ problem: non local term)

Variational problem under constraint $\sum_x p(x) = 1$, $p(x) \geq 0 \ \forall x$

Localizing the expression to get exploitable KKT conditions?

$$I(X,Y) = \sum_x \sum_y p(x)\, p(y|x) \log \frac{p(x|y)}{p(x)} \leftarrow \text{local}$$

but $p(x|y) = \dfrac{p(x)\,p(y|x)}{\sum_{y'} p(x)\,p(y'|x)}$ still non local $\Rightarrow$ assume $p(x|y)$ unknown kernel denoted $q(x|y)$.

$$J[p,q] = \sum_x \sum_y p(x)p(y|x) \log \frac{q(x|y)}{p(x)} \leftarrow \text{concave in } p \text{ and } q!$$

Fact 1: $\max_p \max_q J[p,q] = C$

Proof: immediate if $\max_q J[p,q] = I[p]$

KKT: $\dfrac{\delta}{\delta q[x|y]} \left[J[p,q] - \sum_y \mu(y) \left(\sum_x q(x|y) - 1 \right) \right] = 0$

$$\Rightarrow \frac{p(y,x)}{q(x|y)} - \mu(y) = 0 \Rightarrow q(x|y) = \frac{p(y,x)}{\mu(y)} = p(x|y)$$

($\mu(y)$: normalization)

✓

Fact 2: $\arg\max_p J[p,q] = p^*$, $p^*(x) = \dfrac{e^{\sum_y p(y|x)\log q(x|y)}}{\sum_x e^{\sum_y p(y|x)\log q(x|y)}}$

Proof: KKT: $\dfrac{\delta}{\delta p[x]}\left\{J[p,q] - \mu\left(\sum p(x) - 1\right)\right\} = 0$

$$\sum_y p(y|x)\left[\log\left(\frac{q(x|y)}{p(x)}\right) - 1\right] - \mu = 0$$

$$\log p(x) = \sum_y p(y|x)\log q(x|y) - \mu + 1 \quad \checkmark$$

Algorithm: iterates two optimization steps:

geometric interpretation?

$$\begin{cases} q^{(n)} \leftarrow q^{(n+1)} \;/\; q^{(n+1)}(x) \propto p^{(n)}(x)\, p(y|x) \\ p^{(n)} \leftarrow p^{(n+1)} \;/\; p^{(n+1)}(x) \propto \exp\left[\sum_y p(y|x)\log q^{(n+1)}(x|y)\right] \end{cases}$$

it can be shown that iterating these steps leads to a strictly increasing sequence of MI values converging to the capacity C.

At capacity $p(x) \propto \exp\left[\sum_y p(y|x)\log\left(\dfrac{p(x)\,p(y|x)}{\sum_y p(x)\,p(y|x)}\right)\right]$

$$p(x) \propto p(x)\, D_{KL}\left(p(y|x)\,\|\,p(y)\right) \quad \longrightarrow 1 = \text{constant}$$

$p^*(x)$ is such that $p^*(y)$ is equally distinct from each the conditional distribution $p(y|x)$ as measured by Kullback-Leibler divergence

Minimax interpretation

↳ setting from game theory

P_θ = $p(x|\theta)$

1- nature picks θ from a statistical model $p(x|\theta), \theta \in \Theta$

2- statistician guesses p_θ by choosing some distribution q

Loss function is measured by $D_{KL}(p(x|\theta) \| q(x))$

Game: nature wants to find a prior $p(\theta)$ that maximize the statistician loss, while the statistician want a decision strategy that minimizes his loss.

Statistician: $\arg\min_q \int D_{KL}(p(x|\theta)|q(x))\, p(\theta) d\theta = q^*$

$q^*(x) = \int p(x|\theta)\, p(\theta)\, d\theta$ → Bayes strategy

Nature: $\max_p \underbrace{\left(\min_q \left(\int D_{KL}(p(x|\theta)|q(x))\, p(\theta)\, d\theta\right)\right)}_{I(x,\theta)} = C$ ↑ capacity

maximin strategy ↑

What about minimax strategies?

↳ It turns out that the min and max commute (Haussler 1995)

$$C = \min_q \left(\max_p \left(\int D_{KL}(p(x|\theta)\|q(x))\, p(\theta)\, d\theta\right)\right)$$

↑ linear in $p(\theta)$!

Rate-distortion

↳ achieving high information rate is "costly" in terms of coding (redundant codes to average noise)

↳ idea: looking for least informative encoding given some constraints on the faithfulness of the encoding, as measure by a distortion metric, e.g. $d(x,y) \geq 0 \quad d(x,x) = 0$

Rate-distortion function:

non local

$$R(\delta) = \min_{p(y|x)} I(X,Y) = \min_{p(y|x)} \sum_x \sum_y p(x)\, p(y|x) \log \frac{p(y|x)}{\sum_x p(y|x)\, p(x)}$$

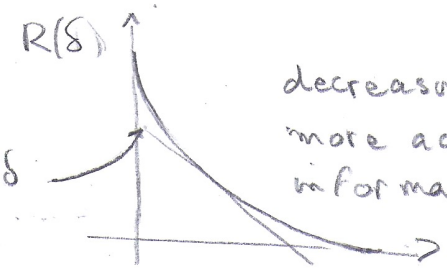
under average distortion constraint: $\sum_x \sum_y p(x) p(y|x) d(x,y) \leq \delta$

Lagrange multiplier s associated to constraint δ

$$R(\delta) = \min_{p(y|x)} \left[I(X,Y) - s \sum_x \sum_y p(x) p(y|x) d(x,y) \right]$$

Intuitively

$s \leq 0$ = slope of R at δ



decreasing function of δ: more accuracy ($\delta \downarrow$) requires more information to be transmitted

Same localization idea as for the capacity introducing

$$F[p|q] = \sum p(x)\, p(y|x) \log \frac{p(y|x)}{q(y)} - s \sum_x \sum_y p(x) p(y|x) d(x,y)$$

$$R[\delta] = \min_p \min_q F[p|q] \Rightarrow q^{(n)}(y) \leftarrow q^{(n+1)}(y) = \sum p(x)\, p^{(n)}(y|x)$$

Blahut algorithm →

$$p^{(n)}(y|x) \leftarrow p^{(n+1)}(y|x) \propto q^{(n+1)}(y)\, e^{s\, d(x,y)}$$