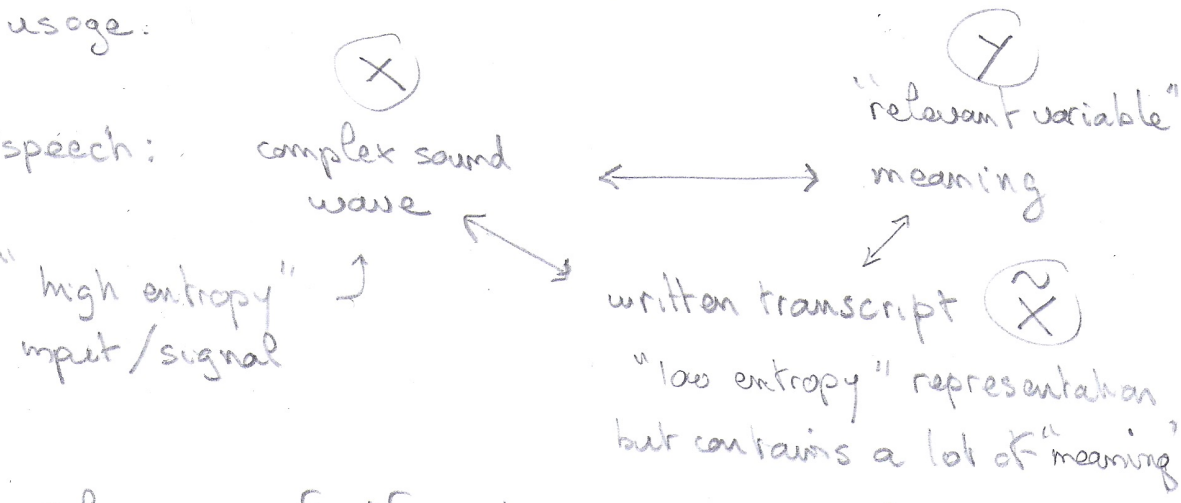
Information relevance

The original formulation of information theory by Shannon purposedly ignored the problem of information relevance, i.e. judging the value – or meaning – of information to the recipient. Only concern: transmission efficiency

Problem of information relevance is a priori not well-posed in the absence of additional knowledge about information structure / usage:



Example: speech:

Idea: The relevance of information conveyed about X in the representation $\tilde{X}$ can be measured via the introduction of an additional relevance variable Y

Tradeoff: There is a natural tradeoff between the representation size and the expected distortion of the meaningful content. ⟷ akin to rate-distortion theory

| Choice of relevant variable ⟷ Feature selection |

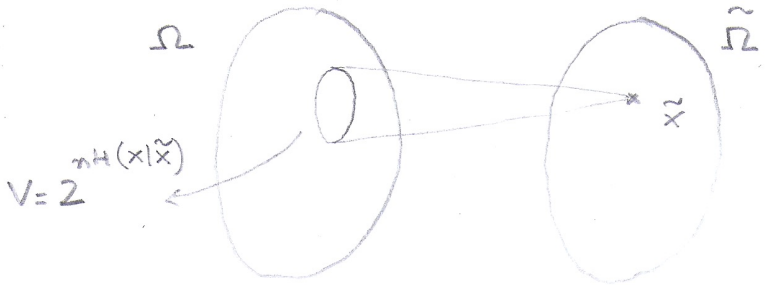
Rate distortion theory

$X \longrightarrow \tilde{X}$: encoding channel

X: input / signal / stimulus

$\tilde{X}$: code / pattern / representation

↑ stochastic mapping $p(\tilde{x}|x)$ = soft partitioning / quantization



V = average volume of input space that get mapped to the same code.

$vol(\Omega) = 2^{nH(x)}$ $\quad vol(\tilde{\Omega}) = 2^{nH(\tilde{x})}$

distinguishible states $= 2^{nH(x)} / 2^{nH(\tilde{x}|x)} = 2^{nI(x,\tilde{x})}$

Quality of the quantization is measured by the rate of bits needed to specify input without confusion, which is bounded below by $I(x,\tilde{x})$.

Trade-off between rate and quality of the reconstruction as measured via a distortion function:

$d: X \times \tilde{X} \longrightarrow \mathbb{R}^+$: d small when x can be faithfully reconstructed from $\tilde{x}$.

$\mathbb{E}_{p(x,\tilde{x})}[d(x,\tilde{x})]$: expected distortion given a code

↳ the larger the rate, the smaller the achievable distortion

Convex optimization setting:

$R(D) = \min_{p(\tilde{x}|x) \text{ s.t. } \mathbb{E}[d] \leq D} I(x,\tilde{x})$ ← minimal achievable rate under constraint of expected distortion

Variational problem and iterative algorithm

$$F[p(\tilde{x}|x)] = I(X,\tilde{X}) + \beta \, \mathbb{E}[d(x,\tilde{x})] + \lambda(x) \sum p(\tilde{x}|x)$$

↑ Lagrangian — ↑ Lagrangian multiplier — ↑ normalization

$$F[p(\tilde{x}|x)] = \sum_{x,x'} p(x)\,p(\tilde{x}|x) \log p(\tilde{x}|x) - \sum_{\tilde{x}} p(\tilde{x}) \log p(\tilde{x}) + \beta \sum_x p(x) \sum_{\tilde{x}} p(\tilde{x}|x)\, d(x,\tilde{x}) + \lambda(x) \sum_{\tilde{x}} p(\tilde{x}|x)$$

↑ non-local term

$$\frac{\delta F}{\delta p(\tilde{x}|x)} = p(x)\left(1 + \log p(\tilde{x}|x)\right) - \sum_{\tilde{x}'} \frac{\delta p(\tilde{x}')}{\delta p(\tilde{x}|x)}\left(1 + \log p(\tilde{x}')\right) + \beta\, p(x)\, d(x,\tilde{x}) + \lambda(x)$$

$$\frac{\delta p(\tilde{x}')}{\delta p(\tilde{x}|x)} = \delta_{\tilde{x}\tilde{x}'}\, p(x')$$

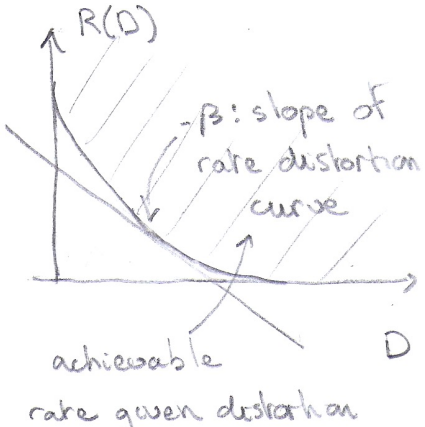
$$= p(x)\left(1 + \log p(\tilde{x}|x) - 1 - \log p(\tilde{x}) + \beta\, d(x,\tilde{x}) + \frac{\lambda(x)}{p(x)}\right)$$

assumed non-zero

$$\frac{\delta F}{\delta p(\tilde{x}|x)} = 0 \quad \Rightarrow \quad \boxed{p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)}\, e^{-\beta\, d(x,\tilde{x})}}$$

$Z(x,\beta)$ ← normalization constant; "free energy"

Moreover, at optimum: $\delta F = \delta I + \beta\, \delta \mathbb{E}[d] = 0 \Rightarrow \boxed{\beta = -\frac{\delta R}{\delta D}}$



self consistent equations

(*) $$p(\tilde{x}) = \sum_x p(\tilde{x}|x)\, p(x)$$

(**) $$p(\tilde{x}|x) = \frac{p(\tilde{x})\, e^{-\beta(d(x,\tilde{x}))}}{Z(x,\beta)}$$

Iterative Blahut algorithm:

(*) $p_{n+1}(\tilde{x}) = \sum_x p_n(\tilde{x}|x)\, p(x)$

(step $n+1$ on the left, step n on the right)

(**) $p_{n+1}(\tilde{x}|x) = \dfrac{p_{n+1}(\tilde{x})\, e^{-\beta d(x,\tilde{x})}}{Z(x,\beta)}$

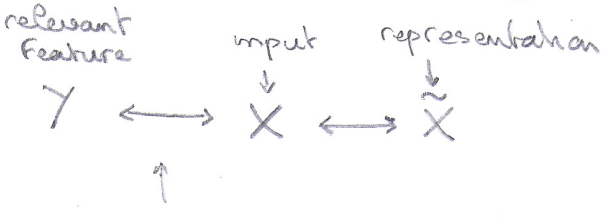
Justification:

$$R(D) = \min_{p(\tilde{x}|x) \text{ s.t. } \mathbb{E}[d] \le D} I(x,y) = \min_{p(\tilde{x}|x)} \min_{p(\tilde{x})} F(p(\tilde{x}|x), p(\tilde{x}))$$
$$\text{s.t. } \mathbb{E}[d] \le D$$

$$F(p(\tilde{x}|x), p(\tilde{x})) = \sum_{\tilde{x},x} p(x)\, p(\tilde{x}|x) \log \frac{p(\tilde{x}|x)}{p(\tilde{x})}$$

↑ minimization of F can be performed independently on the convex sets of probabilities for $p(\tilde{x})$ and $p(\tilde{x}|x), x \in X$

Information bottleneck



$\tilde{X} \perp Y \mid X$ and $Y \perp \tilde{X} \mid X$ are independent

↳ Markov information channel

⚠ double arrow notation to emphasize lack of directionality

$$Y \rightarrow X \rightarrow \tilde{X} \quad \simeq \quad Y \leftarrow X \leftarrow \tilde{X}$$

* Relevant representations are those for which $I(\tilde{X}, Y)$ is high. Remember that $I(\tilde{X}, Y) \leq I(X, Y)$ by the data processing inequality.
* There is a tradeoff between preserving relevant information and compressing the representation. The goal is to pass as much information about Y via X, but through the "bottleneck" formed by compact representations $\tilde{X}$.
* Variational formulation: $L[p(\tilde{x}|x)] = I(X, \tilde{X}) - \beta I(X, Y)$

  ↑ Lagrangian multiplier

  $\beta = 0$: most sketchy representation

  $\beta \to \infty$: arbitrary detailed quantization

  ↳ no explicit distortion function!

  ↳ nonlinear convex optimization problem!

Self consistent equations

$$L[p(\tilde{x}|x)] = \sum_{x,\tilde{x}} p(x)\, p(\tilde{x}|x) \log p(\tilde{x}|x) - \sum_{\tilde{x}} p(\tilde{x}) \log p(\tilde{x}) \leftarrow \text{non-local}$$

$$- \beta \sum_{y,\tilde{x}} p(y)\, p(\tilde{x}|y) \log p(\tilde{x}|y) + \beta \sum_{y} p(\tilde{x}) \log p(\tilde{x}) \leftarrow$$

$$- \sum_{x} \lambda(x) \sum_{\tilde{x}} p(\tilde{x}|x) \leftarrow \text{normalization constraints}$$

$$\frac{\delta L}{\delta p(\tilde{x}|x)} = p(x)\left(1 + \log p(\tilde{x}|x)\right) - \frac{\delta p(\tilde{x})}{\delta p(\tilde{x}|x)} \left(1 + \log p(\tilde{x})\right)$$

$$- \beta \sum_{y} p(y) \frac{\delta p(\tilde{x}|y)}{\delta p(\tilde{x}|x)} \left(1 + \log p(\tilde{x}|y)\right) + \beta \frac{\delta p(\tilde{x})}{\delta p(\tilde{x}|x)} \left(1 + \log p(\tilde{x})\right)$$

$$- \lambda(x).$$

Markov chain $Y \leftrightarrow X \leftrightarrow \tilde{X}$:

$$p(\tilde{x}) = \sum_{\tilde{x}} p(\tilde{x}|x)\, p(x)$$

$$p(\tilde{x}|y) = \sum_{x} p(\tilde{x}|x)\, p(x|y)$$

$$\frac{\delta p(\tilde{x})}{\delta p(\tilde{x}|x)} = p(x) \quad \text{and} \quad \frac{\delta p(\tilde{x}|y)}{\delta p(\tilde{x}|x)} = p(x|y)$$

$$\frac{\delta L}{\delta p(\tilde{x}|x)} = p(x) \Big\{ \cancel{1} + \log p(\tilde{x}|x) - \cancel{1} - \log p(\tilde{x}) - \beta \sum_{y} p(y|x) \left(\cancel{1} + \log p(\tilde{x}|y)\right)$$

assumed $\neq 0$ ↑ (pointing to $p(x)$)

$$+ \beta \left(\cancel{1} + \log p(\tilde{x})\right) - \frac{\lambda(x)}{p(x)} \Big\}$$

$$= p(x) \left\{ \log \frac{p(\tilde{x}|x)}{p(\tilde{x})} - \beta \sum_{y} p(y|x) \log \frac{p(y|\tilde{x})}{p(y)} - \frac{\lambda(x)}{p(x)} \right\}$$

$$= p(x) \Big\{ \log \frac{p(\tilde{x}|x)}{p(\tilde{x})} + \beta \sum_{y} p(y|x) \log \frac{p(y|x)}{p(y|\tilde{x})}$$

↗ $D_{KL}(p(y|x) \| p(y|\tilde{x}))$ "def" $d(x\|\tilde{x})$

$$- \left(\frac{\lambda(x)}{p(x)} + \beta \sum_{y} p(y|x) \log \frac{p(y|x)}{p(y)} \right) \Big\}$$

new multiplier: $\tilde{\lambda}(x)$ ← (pointing to the bracketed term above)

$$\frac{\delta L}{\delta p(\tilde{x}|x)} = 0 \Rightarrow \boxed{p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} e^{-\beta d(x,\tilde{x})}}$$

$Z(x,\beta)$: normalization constant; $d(x,\tilde{x})$: distortion function

The Kullback-Leibler divergence $d(x,\tilde{x}) = D_{KL}(p(y|x) \| p(y|\tilde{x}))$ emerges as the relevant distortion function.

At optimum: self consistent equations

$$(*) \quad p(\tilde{x}) = \sum_x p(\tilde{x}|x)\, p(x)$$

$$(**) \quad p(y|\tilde{x}) = \sum p(y|x)\, p(x|\tilde{x})$$

$$(***) \quad p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} e^{-\beta d(x,\tilde{x})}$$

Markov chain $Y \leftrightarrow X \leftrightarrow \tilde{X}$

Iterative algorithm

1. $p_n(\tilde{x}|x) = \frac{p_n(\tilde{x})}{Z_n(x,\beta)} e^{-\beta D_{KL}(p(y|x)\|p_n(y|\tilde{x}))}$ ← (***)

2. $p_{n+1}(\tilde{x}) = \sum_x p_n(\tilde{x}|x)\, p(x)$ ← (*)

2'. $p_{n+1}(y|\tilde{x}) = \sum_x p(y|x)\, p_n(x|\tilde{x})$ ← (**)

Interpretation

$$\min_{p(\tilde{x}|x) \text{ s.t. } I(\tilde{X},Y)\geq D} I(X,\tilde{X}) = \min_{p(y|\tilde{x})} \min_{p(\tilde{x})} \min_{p(\tilde{x}|x) \text{ s.t. } I(\tilde{X},Y)\geq D} F\left[p(\tilde{x}|x), p(\tilde{x}), p(y|\tilde{x})\right]$$

$$F\left[p(\tilde{x}|x), p(\tilde{x}), p(y|\tilde{x})\right] = I(X,\tilde{X}) + \beta\, \mathbb{E}_{p(x,\tilde{x})}\left[D_{KL}(p(y|x)\|p(y|\tilde{x}))\right] + K[X,Y] \leftarrow \text{constant}$$

$I(X,\tilde{X})$: regular rate-distortion function minimized when (*) and (***) hold at fixed $p(y|\tilde{x})$

$\mathbb{E}_{p(x,\tilde{x})}\left[D_{KL}(p(y|x)\|p(y|\tilde{x}))\right]$: functional of $p(y|\tilde{x})$ minimized when (**) holds at fixed $p(x,\tilde{x})$