Notation

X, Y are finite sets

$P, P^* \in \Delta(X)$ [P, P^* distributions on X]

$R \in \Delta(X \times Y)$ [joint distribution]

$P(x) = \sum_{y \in Y} R(x, y)$ [marginal on X]

$Q(y) = \sum_{x \in X} R(x, y)$ [marginal on Y]

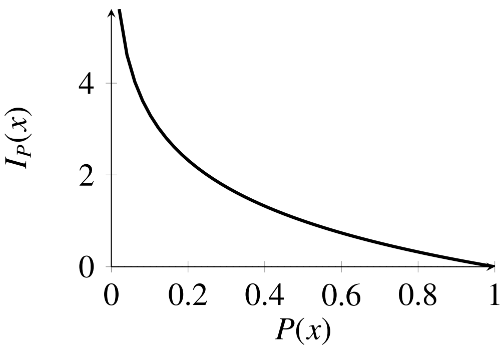
Info content (subjectivist version)

Info content $I_P(x)$ ("surprisal") measures the perplexity of an agent with beliefs $P \in \Delta(X)$ when observing $x \in X$.

think of: neural activity in a predictive brain

Definition: $I_P(x) = -\log_b P(x)$

base $b > 1$; common choice $b = 2$ (bits)



Justification: Negative log ($b > 1$) is the only function satisfying constraints:

if everything exactly as expected, zero perplexity

If $P(x) = 1$, $I_P(x) = 0$

less expected, more perplexing

If $P(x_1) > P(x_2)$, then $I_P(x_1) < I_P(x_2)$

perplexity adds up

$I_P(x_1 \& x_2) = I_P(x_1) + I_P(x_2)$

if x_1, x_2 stochastically independent

General template for all measures

Definitions below are all expected values of the form:

$$\sum_{x \in X} P_{GT}(x)\, F(x)$$

P_{GT} is the assumed ground-truth

F is some function related to perplexity

Logarithm rules

change of base

$$\log_a x = \frac{\log_b x}{\log_b a}$$

division-to-subtraction rule

$$\log_b \frac{x}{y} = \log_b x - \log_b y$$

		$\mathbf{P_{GT}}$	**F**	**definition**	
entropy	$\mathcal{H}(P)$	P	I_P	$-\sum_{x \in X} P(x)\, \log_b P(x)$	
average perplexity of an agent with beliefs P when the ground truth is P					
cross-entropy	$\mathcal{H}(P^*, P)$	P^*	I_P	$-\sum_{x \in X} P^*(x)\, \log_b P(x)$	
average perplexity of an agent with beliefs P when the ground truth is P^*					
joint entropy	$\mathcal{H}(P, Q)$	R	$I_{R'}$	$-\sum_{z \in X \times Y} R(z)\, \log_b R(z)$	
just entropy applied to a joint probability distribution; slightly boring but useful for the "fun facts" below NB: cross-entropy compares distributions on the same X, joint entropy looks at the joint distribution over product of space $X \times Y$					
conditional entropy	$\mathcal{H}(P \mid Q)$	Q	$\mathcal{H}(R^{\mid y})$	$-\sum_{y \in Y} Q(y)\, \sum_{x \in X} R(x \mid y)\, \log_b R(x \mid y)$	where $R^{\mid y}(x) = R(x \mid y)$
		R	I_S	$-\sum_{\langle x,y \rangle \in X \times Y} R(x, y)\, \log_b R(x \mid y)$	where $S(\langle x, y \rangle) = R(x \mid y)$
average entropy of an agent's conditional beliefs about X after observing events from Y; how uncertain is the agent about X when they observe Y two equivalent formulations here: the second is the usual (compact) definition; the first is easier to interpret					
relative entropy	$D_{\mathrm{KL}}(P \parallel Q)$	P	$I_Q - I_P$	$\sum_{x \in X} P(x)\, \log_b \frac{P(x)}{Q(x)}$	
also known as **Kullback-Leibler divergence**; average difference in perplexity when agent believes Q instead of true P "excess surprisal" or "unnecessary perplexity" on top of the minimum (when having "true beliefs" P)					
mutual information	$I(P, Q)$	R	$I_{R^{\perp\!\!\!\perp}} - I_R$	$\sum_{\langle x,y \rangle \in X \times Y} R(x, y)\, \log_b \frac{R(x,y)}{P(x)\, Q(x)}$	where $R^{\perp\!\!\!\perp}(x, y) = P(x)\, Q(x)$
excess perplexity of an agent believing that X and Y are independent, when in truth they might not be alternatively: how much learning about Y reduces uncertainty about X (and vice versa; see facts below) special case of KL-divergence for joint distributions, one treating X and Y as independent					

Fun facts

$P^* = \arg\min_P \mathcal{H}(P^*, P)$

$\mathcal{H}(P, P) = \mathcal{H}(P)$

$D_{KL}(P \parallel Q) = \mathcal{H}(P, Q) - \mathcal{H}(P)$

$I(P, Q) = I(Q, P)$

$I(P, Q) = \mathcal{H}(P) - \mathcal{H}(P \mid Q)$

$I(P, Q) = \mathcal{H}(P) + \mathcal{H}(Q) - \mathcal{H}(P, Q)$