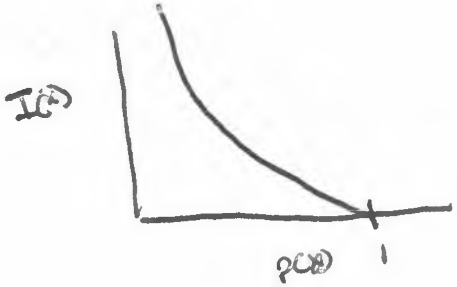
# Information theory

- info should be high when an unlikely event has occured
- info should be zero when a garaunteed event $(P(X)=1)$ has occured
- the information from 2 independent events should be additive:

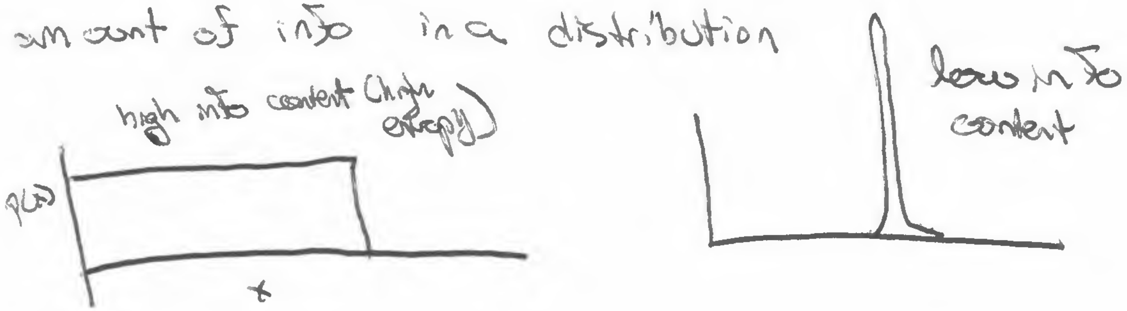
$I(X=x) = -\log[P(X=x)]$ (captures these intuitions)



## entropy

- average amount of info in a distribution

eg:



= avg. info contained w/in a dist: $H[X] = E_{x\sim p}[I[X]] = -E_{x\sim p}[P(X)]$

## K-L divergence

$D_{KL}(P\|Q) = \mathbb{E}_{x\sim p}\left[\frac{\log P(X)}{\log Q(X)}\right]$ ≈ a measure of difference between P and Q (but, $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$)

- One can show that the MLE minimizes the KL divergence between the empirical distribution and the model distribution. This is another, nice, way of interpreting an MLE.