Distances and Similarities

- **Distances**
  - What properties do they have?

- **Similarities**
  - How have we computed them?
KL-Divergence

A) Distance
B) Similarity
C) Neither

\[ D_{KL}(P \parallel Q) = - \sum_i P(i) \log \frac{Q(i)}{P(i)}, \]
KL divergence properties

- Non-negative: \( D(P||Q) \geq 0 \)

- Divergence 0 if and only if \( P \) and \( Q \) are equal:
  - \( D(P||Q) = 0 \) iff \( P = Q \)

- Non-symmetric: \( D(P||Q) \neq D(Q||P) \)

- Does not satisfy triangle inequality
  - \( D(P||Q) \leq D(P||R) + D(R||Q) \)

Not a distance metric
Kullback Leibler divergence

- \( P \) = true distribution;
- \( Q \) = alternative distribution that is used to encode data
- KL divergence is the expected extra message length per datum that must be transmitted using \( Q \)

\[
D_{\text{KL}}(P \mid\mid Q) = \sum_i P(x_i) \log (P(x_i)/Q(x_i))
\]

\[
= - \sum_i P(x_i) \log Q(x_i) + \sum_i P(x_i) \log P(x_i)
\]

\[
= H(P,Q) - H(P)
\]

\[
= \text{Cross-entropy} - \text{entropy}
\]

- Measures how different the two distributions are
KL divergence as info gain

- The KL divergence of the posteriors measures the information gain expected from query \((x')\):
  \[
  D\left( p(\theta \mid x, x') \mid\mid p(\theta \mid x) \right)
  \]

- Goal: choose a query that maximizes the KL divergence between the updated posterior probability and the current posterior probability
  - This represents the largest expected information gain