CG136: Introduction to Computational Linguistics

Mark Johnson

Brown University

February 2006

Talk overview

- Introduction to Information Theory

Why information theory?

- *Entropy* quantifies the amount of information in a probability distribution
  - The entropy of a distribution is a measure of the randomness of the distribution
    (high entropy $\Rightarrow$very uncertain, low entropy $\Rightarrow$predictable)
  - Measured in *bits* (number of binary choices)
- *Cross-entropy* and *mutual information* quantify the amount of information one distribution provides about another one
  - Cross entropy measures how useful knowing one distribution is in order to predict another
    * ideal for evaluating probabilistic models

Optimal coding

- Suppose we want to encode in binary a signal consisting of samples X distributed according to $\mathrm{P}(X = x) = p(x)$. An *optimal code* assigns each x a *code word* of length $-\log_2 p(x)$ bits
- Why $\log_2$? Each additional bit in a code word doubles the possible values we can describe, so it's plausible that they can be half as probable

Encoding a sequence of flips of a fair coin:

$p(\text{heads}) = p(\text{tails}) = \frac{1}{2}$, so an optimal code might be
$C(\text{heads}) = 1, C(\text{tails}) = 0$

Encoding a sequence of rolls of a biased 3-sided die:

$p(\text{a}) = \frac{1}{2}, p(\text{b}) = p(\text{c}) = \frac{1}{4}$, so an optimal code might be
$C(\text{a}) = 1, C(\text{b}) = 00, C(\text{c}) = 01$

Entropy

- The *entropy* $\mathrm{H}(p)$ (in bits) of a random variable X where $\mathrm{P}(X = x) = p(x)$ is the *expected length of an optimal encoding of X*

$$\begin{aligned}\mathrm{H}(p) &= \mathrm{E}[-\log_2 p] \\ &= -\sum_x p(x) \log_2 p(x)\end{aligned}$$

(Hint: remember $\log_2(y) = \log_b(y)/\log_b(2)$ for any base b)

Entropy of a fair coin: $p(\text{heads}) = p(\text{tails}) = \frac{1}{2}$, so

$\mathrm{H} = -\frac{1}{2} \times \log_2(\frac{1}{2}) - \frac{1}{2} \times \log_2(\frac{1}{2}) = 1\,\text{bit}$

Entropy of a fair die: $p(1) = \ldots = p(6) = \frac{1}{6}$, so

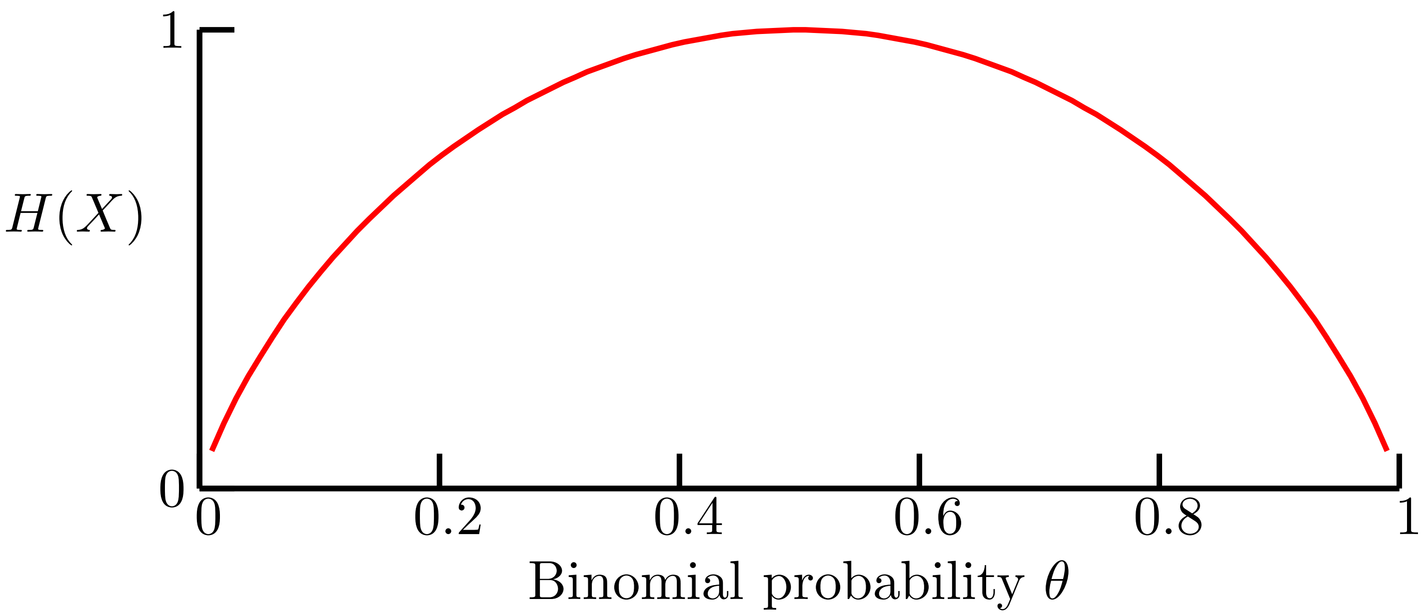
$\mathrm{H} = -6 \times \frac{1}{6} \times \log_2(\frac{1}{6}) \approx 2.58\,\text{bits}$

- $\mathrm{H}(p) \geq 0$, and $\mathrm{H}(p) = 0$ iff $p(x) = 1$ for some x

Entropy of a biased coin

- $p(\text{heads}) = \frac{3}{4}, p(\text{tails}) = \frac{1}{4}$, so $\mathrm{H} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.81$ bits
- Suppose $p(\text{heads}) = \theta$ and $p(\text{tails}) = 1-\theta$. Then:

$$\begin{aligned}
\mathrm{H}(p) &= \sum_x -p(x)\log_2 p(x) \\
&= -\theta\log_2\theta - (1-\theta)\log_2(1-\theta)
\end{aligned}$$



Cross Entropy

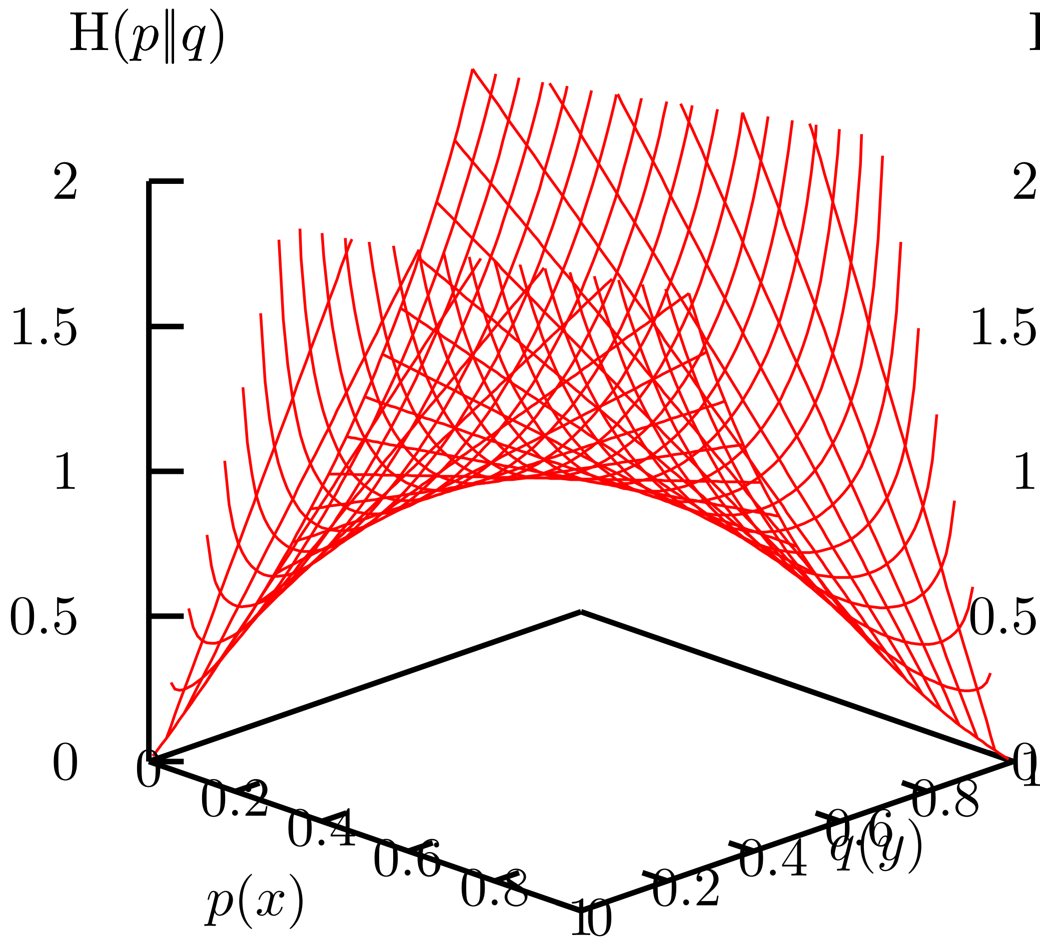
- The *cross entropy* of a pair of random variables X, Y where $\mathrm{P}(X = x) = p(x)$ and $\mathrm{P}(Y = y) = q(y)$ is the *expected number of bits needed to encode X using an optimal code for Y*:

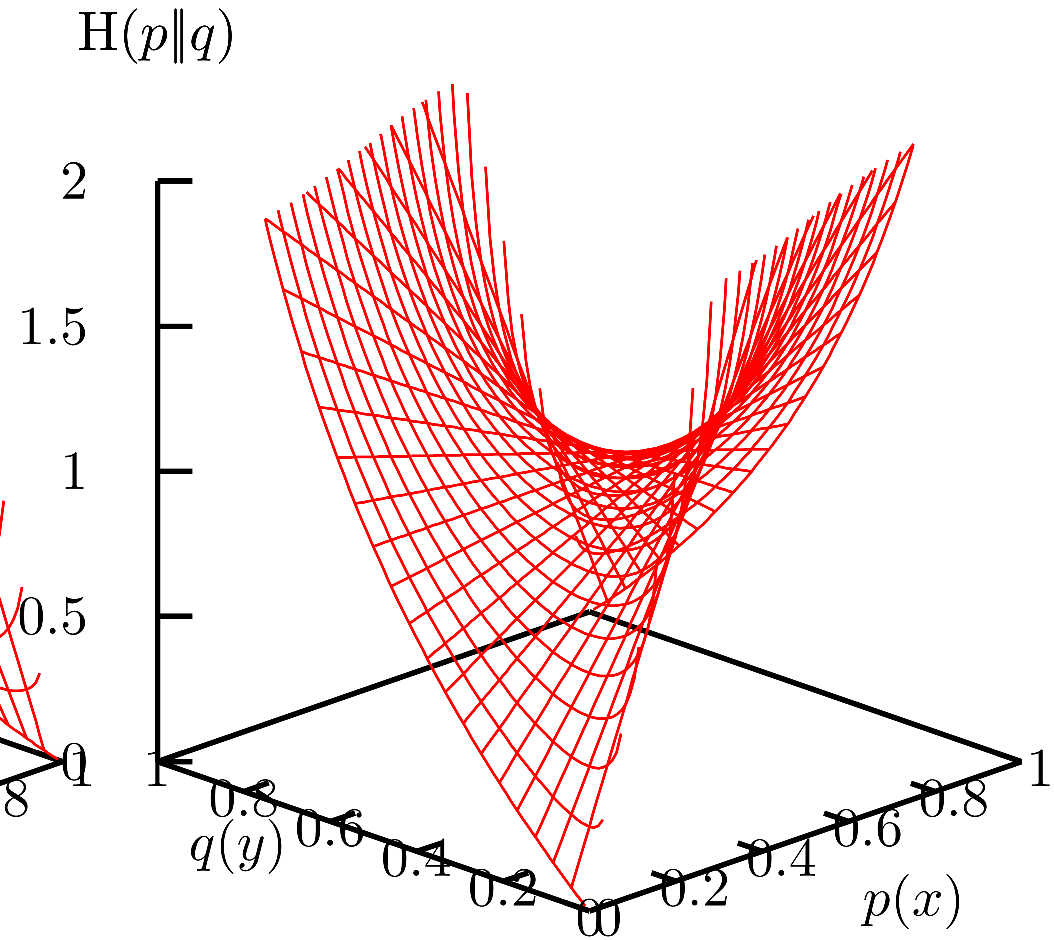
$$\begin{aligned} \mathrm{H}(p\|q) &= \mathrm{E}_p[-\log_2 q] \\ &= -\sum_x p(x)\log_2 q(x) \end{aligned}$$

- $\mathrm{H}(p\|q) \geq \mathrm{H}(p)$, with $\mathrm{H}(p\|q) = \mathrm{H}(p)$ iff for all x $p(x) = q(x)$, i.e., the optimal code for X is the one based on X

- In general $\mathrm{H}(p\|q) \neq \mathrm{H}(q\|p)$

Cross-entropy between binomials





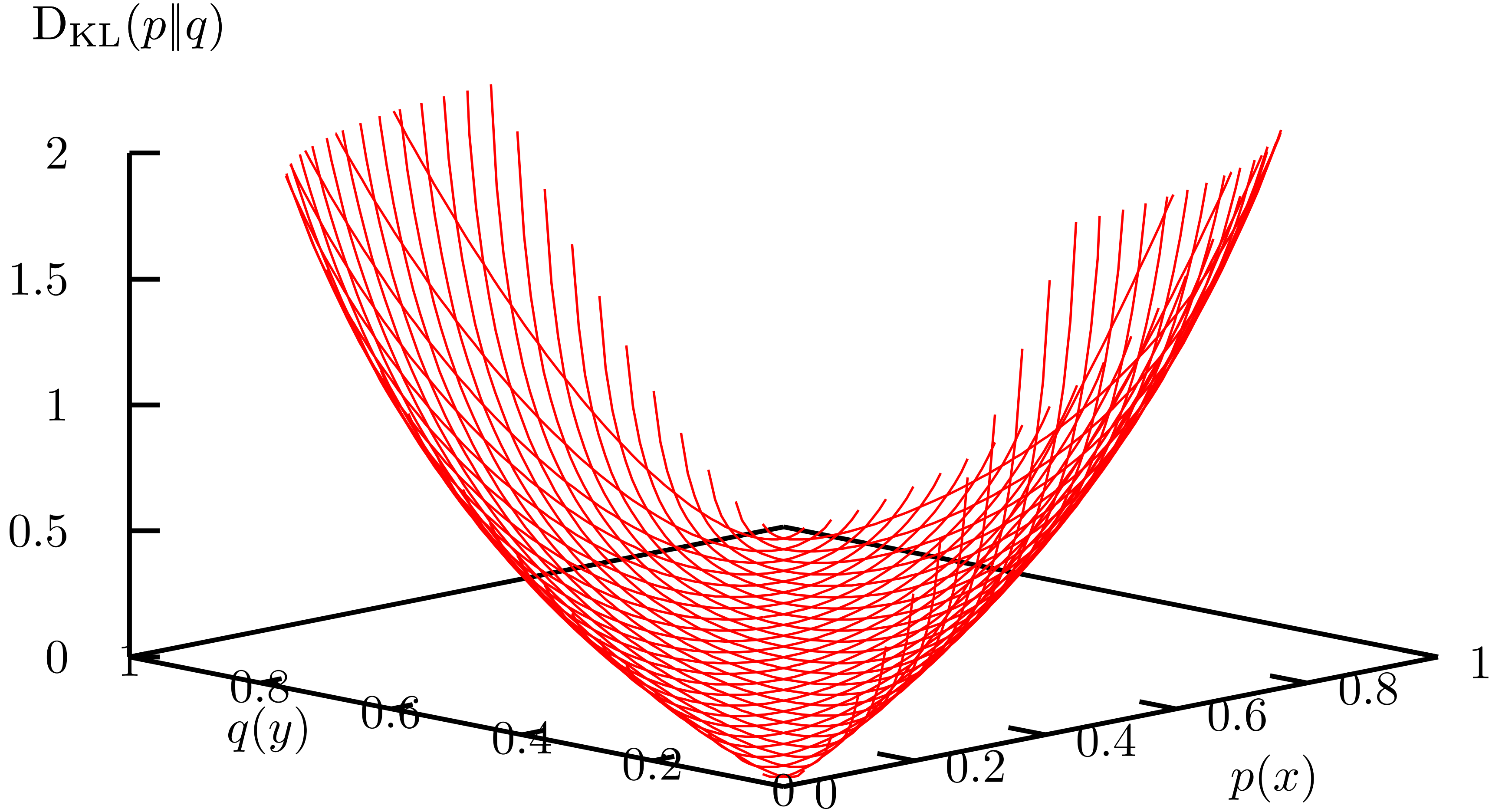
The Kullback-Leibler divergence

- The *Kullback-Leibler divergence* or KL-divergence between X and Y where $\mathrm{P}(X = x) = p(x)$ and $\mathrm{P}(Y = y) = q(y)$ is the *expected number of bits "lost" in encoding X using an optimal code for Y*

$$
\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(p \| q) &= \mathrm{E}_p[-\log_2 q] - \mathrm{E}_p[-\log_2 p] \\
&= \mathrm{E}_p[\log_2 \frac{p}{q}] \\
&= \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \\
&= \mathrm{H}(p \| q) - \mathrm{H}(p)
\end{aligned}
$$

- $\mathrm{D}_{\mathrm{KL}}(p \| q) \geq 0$, with $\mathrm{D}_{\mathrm{KL}}(p \| q) = 0$ iff for all x $p(x) = q(x)$
- $\mathrm{D}_{\mathrm{KL}}(\cdot \| \cdot)$ is *not a distance metric*! In general $\mathrm{D}_{\mathrm{KL}}(p \| q) \neq \mathrm{D}_{\mathrm{KL}}(q \| p)$

KL-divergence between binomials



Evaluating models with KL-divergence

- Suppose we have constructed a probabilistic model for Y, where $\mathrm{P}(Y = y) = q(y)$, and we want to see how well it predicts some empirical data

- Treat the empirical data as another variable X, where $\mathrm{P}(X = x) = p(x)$ is the relative frequency of x in the data

- Then $\mathrm{D}_{\mathrm{KL}}(p \| q)$ is the number of bits lost by modeling X with Y
  - $\mathrm{D}_{\mathrm{KL}}(p \| q) = 0$ if the model q is exactly the same as empirical data p

- Is defined so long as $\mathrm{Support}(q) \supseteq \mathrm{Support}(p)$ where $\mathrm{Support}(q) = \{x : q(x) > 0\}$.

Joint Entropy

- The *joint entropy* of a pair of random variables X, Y is just the entropy of their joint distribution $Z = (X, Y)$, where $\mathrm{P}(Z = (x, y)) = \mathrm{P}(X = x, Y = y) = r(x, y)$

$$
\begin{aligned}
\mathrm{H}(X, Y) &= \mathrm{H}(Z) = \mathrm{H}(r) \\
&= -\mathrm{E}_r[\log_2 r(x, y)] \\
&= -\sum_{x,y} r(x, y) \log_2 r(x, y)
\end{aligned}
$$

- $\mathrm{H}(X) + \mathrm{H}(Y) \geq \mathrm{H}(X, Y) \geq \mathrm{H}(X)$ and $\mathrm{H}(Y)$
- If X and Y are *independent*, then $\mathrm{H}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y)$

Conditional Entropy

- The *conditonal entropy* of a pair of random variables X, Y where $\mathrm{P}(X = x, Y = y) = r(x, y)$ is the amount of extra information needed to identify Y given X

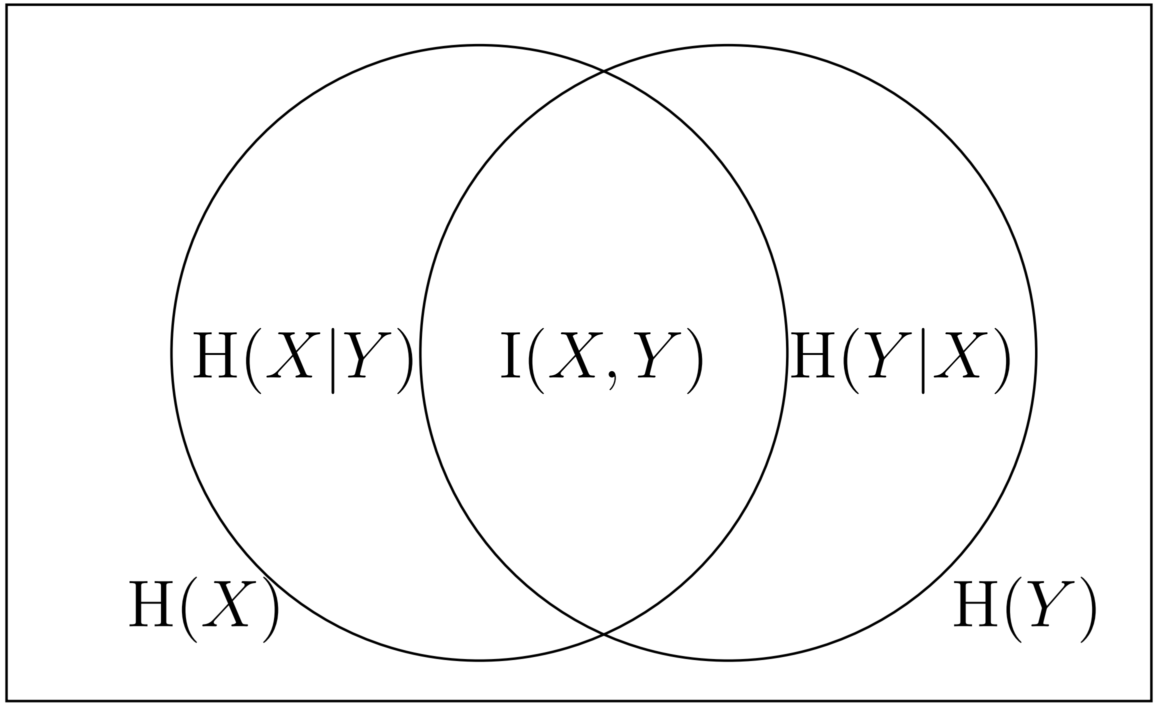
$$\begin{aligned}
\mathrm{H}(Y|X) &= \mathrm{H}(X,Y) - \mathrm{H}(X) \\
&= \mathrm{H}(r) - \mathrm{H}(p) \quad \text{where} \\
p(x) &= \mathrm{P}(X = x) \; = \; \sum_{y} r(x, y)
\end{aligned}$$

Mutual information

- The mutual information $\mathrm{I}(X, Y)$ between random variables X and Y is the amount of shared information in X and Y

$$
\begin{aligned}
\mathrm{I}(X, Y) &= \mathrm{H}(X) - \mathrm{H}(X|Y) \;=\; \mathrm{H}(Y) - \mathrm{H}(Y|X) \\
&= \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X, Y) \\
&= \sum_{x,y} r(x, y) \log_2 \frac{r(x, y)}{p(x)q(y)} \quad \text{where} \\
\mathrm{P}(X = x, Y = y) &= r(x, y) \\
\mathrm{P}(X = x) &= p(x) \;=\; \sum_{y} r(x, y) \\
\mathrm{P}(Y = y) &= q(y) \;=\; \sum_{x} r(x, y)
\end{aligned}
$$

Mutual information (2)



Homework for next Tuesday

- Please add the English word None to each English sentence in the IBM model 1 and rerun. How does it affect the alignments?
- We need to do more serious evaluation!
  - Please read documentation on HLT-NAACL 2003 word alignment workshop (link on class web page)
  - Please read Bob Moore's (2004) article "Improving IBM Word Alignment Model 1".
- We should be discussing textbook chapter 8 ...